SIAM J. MATH. ANAL.
Vol. 40, No. 1, pp. 1–20

THE ∞-WASSERSTEIN DISTANCE: LOCAL SOLUTIONS AND EXISTENCE OF OPTIMAL TRANSPORT MAPS*

THIERRY CHAMPION†, LUIGI DE PASCALE‡, AND PETRI JUUTINEN§

Abstract. We consider the non-nonlinear optimal transportation problem of minimizing the cost functional $\mathcal{C}_\infty(\lambda) = \lambda\text{-ess sup}_{(x,y)\in\Omega^2} |y-x|$ in the set of probability measures on Ω^2 having prescribed marginals. This corresponds to the question of characterizing the measures that realize the infinite Wasserstein distance. We establish the existence of "local" solutions and characterize this class with the aid of an adequate version of cyclical monotonicity. Moreover, under natural assumptions, we show that local solutions are induced by transport maps.

Key words. infinite Wasserstein distance, restrictable solutions, infinite cyclical monotonicity

AMS subject classifications. 49Q20, 49K30

DOI. 10.1137/07069938X

1. Introduction.

In this paper, we consider the non-nonlinear optimal transportation problem that can be mathematically stated as the problem of minimizing the cost functional

$$\mathcal{C}_\infty(\lambda) := \lambda\text{-ess sup}_{(x,y)\in\Omega^2} |y-x| \tag{1.1}$$

in the set of probability measures on Ω^2 having prescribed marginals. Here, and throughout the paper, we assume that Ω is a compact subset of $\mathbb{R}^d$, $d \geq 1$, $|\cdot|$ denotes the usual Euclidean norm in $\mathbb{R}^d$, μ, ν are the two (given) Borel probability measures on Ω, and $\Pi(\mu,\nu)$ denotes the set of admissible transport plans, i.e., the set of Borel probability measures λ on $\Omega^2 := \Omega \times \Omega$ with first marginal $\pi_{1\,\#}\lambda = \mu$ and second marginal $\pi_{2\,\#}\lambda = \nu$. Informally, if λ is induced by a transport map $T : \Omega \to \Omega$, i.e., $\lambda = (id \times T)_\# \mu$, then $\mathcal{C}_\infty(\lambda)$ is simply the maximum of the transport distances $|T(x) - x|$.

The problem formulated above corresponds to the question of characterizing the measures that realize the infinite Wasserstein distance

$$(P_\infty) \qquad W_\infty(\mu,\nu) \;=\; \inf\Big\{\mathcal{C}_\infty(\lambda) \;=\; \lambda\text{-ess sup}_{(x,y)\in\Omega^2} |y-x| \;:\; \lambda \in \Pi(\mu,\nu)\Big\}$$

*Received by the editors August 6, 2007; accepted for publication (in revised form) November 16, 2007; published electronically March 26, 2008. Part of this research was conducted while the authors were visiting each other at their respective institutions.

http://www.siam.org/journals/sima/40-1/69938.html

†Laboratoire d'Analyse Non Linéaire Appliquée, U.F.R. des Sciences et Techniques, Université du Sud Toulon-Var, Avenue de l'Université, BP 20132, 83957 La Garde cedex, France (champion@univ-tln.fr). This author's visit to Pisa in November, 2006, was supported by the I.N.D.A.M project "Traffic flows and optimization on complex networks" and by the G.N.A.M.P.A project "Fenomeni di evoluzione non lineari suggeriti dalla Fisica e dalla Biologia", and his visit to Jyväskylä in June, 2007, was supported by the ESF program "Global and geometric properties of solutions of nonlinear partial differential equations".

‡Dipartimento di Matematica Applicata, Universitá di Pisa, Via Buonarroti 1/c, 56127 Pisa, Italy (depascal@dm.unipi.it). This author's research was partially supported by the Italian M.I.U.R. project "Metodi variazionali nella teoria del trasporto ottimo di massa e nella teoria geometrica della misura" and by the "Fondo di ateneo per la ricerca" of the University of Pisa.

§Department of Mathematics and Statistics, P.O. Box 35 (MaD), FI-40014 University of Jyväskylä, Finland (peanju@maths.jyu.fi). This author's research was supported by Academy of Finland project 108374.

between μ and ν. Clearly, this is the limiting case, as $p \to \infty$, of the more familiar (see, e.g., [1, 2, 32]) p-Wasserstein distance problem

$$(P_p) \quad W_p(\mu,\nu) \;=\; \inf\left\{ \left(\int_{\Omega^2} |y-x|^p d\lambda(x,y) \right)^{\frac{1}{p}} \;:\; \lambda \in \Pi(\mu,\nu) \right\}, \qquad 1 \le p < \infty,$$

which is a model example of a Monge–Kantorovich-type optimal transport problem. Despite the close relationship, there are fundamental differences between these two problems. Most importantly, while (P_p) is linear in λ (removing the $1/p$-power does not change the solution set), the mapping $\lambda \mapsto \mathcal{C}_\infty(\lambda)$ is not even convex. In particular, the problem (P_∞) is not additive, which implies that, unlike in the case of (P_p), a restriction of an optimal transport plan need not be optimal for its own marginals.

In view of simple examples, it turns out that only imposing this "local optimality" property can lead to a satisfactory class of solutions. Hence we introduce in this paper the notion of *restrictable solutions*; this subclass of minimizers of (1.1) is characterized by the property that every portion of μ is transported onto its target in an optimal way; see Definition 4.1 below. The existence of a restrictable solution is obtained with the aid of approximating (P_∞) by the problems (P_p). The same strategy also provides us with the notion of *infinite cyclical monotonicity*, which is derived from the standard c-cyclical monotonicity by applying it to the sequence of costs $c_p(x,y) = |x-y|^p$ and then taking the limit as $p \to \infty$. A reader familiar with the theory of infinity Laplacian and related problems [5] should recognize the analogy between restrictable solutions and absolute minimizers of supremum functionals.

It is one of the main results of this paper that restrictable and infinitely cyclically monotone solutions coincide; see Theorems 3.4 and 4.4 below. We would like to emphasize that although both of these notions are derived via an approximation argument, the proof for their equivalence is completely independent of the derivation. Moreover, this result holds without any further assumptions on the marginals μ and ν. The second principal question we address in this paper is existence and uniqueness of an optimal transport map. Our main result in this direction, Theorem 5.5, states that if $\mu \ll \mathcal{L}^d$, then any infinitely cyclically monotone solution γ to (P_∞) is induced by a map $T : \Omega \to \Omega$, i.e., $\gamma = (id \times T)_\# \mu$. Regarding the question of uniqueness, we are able to show that if, in addition to the previous assumptions, the second marginal ν is discrete, then the infinitely cyclically monotone solution to (P_∞) is unique.

A major technical difficulty that we are facing in the proofs is the absence of a useful duality theory, which is due to the nonconvexity of the objective functional (see section 5.4). As a consequence, we must rely on ad hoc arguments designed for the problem at hand. On the other hand, it is quite clear from the proofs that the machinery we are developing applies to more general problems than just (P_∞). In fact, we could have just as well considered a functional $\lambda \mapsto \lambda\text{-ess}\sup_{(x,y)\in\Omega^2} c(x,y)$, where $c(x,y)$ is, say, nonnegative and lower semicontinuous to begin with. In this work we concentrate on the model case $c(x,y) = |y-x|$ so as to identify the useful tools and notions without coping with the additional technical difficulties required by a more general cost c, which seems to be the natural next step in this study.

Let us finish this introduction by discussing some applications in which the infinite Wasserstein distance W_∞ appears. First on our list is the optimal design problem

$$\sup\left\{ \frac{W_\infty(\mu,\nu)^{p+d}}{W_p(\mu,\nu)^p \| (\frac{d\mu}{dx})^{-1} \|_{L^\infty(U)}} \;:\; (\mu,\nu) \in \mathcal{P}_{a.c.}(U) \times \mathcal{P}(\overline{U}) \right\} \tag{1.2}$$

that appears in [9] in connection with stability estimates for optimal transport maps; here $\mathcal{P}$ and $\mathcal{P}_{a.c.}$ denote the spaces of probability measures and absolutely continuous probability measures, respectively, and $\frac{d\mu}{dx}$ is the Radon–Nikodym derivative of μ with respect to the Lebesgue measure. In [9], the authors prove that if $U \subset \mathbb{R}^d$ is a bounded Lipschitz domain, then the estimate

$$W_\infty(\mu,\nu)^{p+d} \le C_{p,d}(U)\|(\tfrac{d\mu}{dx})^{-1}\|_{L^\infty(U)} W_p(\mu,\nu)^p \tag{1.3}$$

holds for every $p > 1$. The inequality (1.3) is an intrinsic counterpart of a beautiful uniform estimate for the optimal transport maps proved in [9] under stronger regularity assumptions. The optimal constant in (1.3) is given by the supremum in (1.2), and it is conjectured (based on 1-dimensional examples and remarks on increasing transport maps) that it does not blow up when $p \to 1$.

Second, during the last few years, models of branching processes using in one way or another tools from the optimal transportation theory have been proposed by several authors; see, for example, [8, 10, 23, 34]. Roughly speaking, these models favor joint transportation, which in many real world situations, such as in the design of communication or irrigation networks, is more economical than individualized transportation. In particular, in [10] the authors propose minimizing a certain cost functional (which penalizes diffused measures) on the p-Wasserstein space of probability measures. It is remarked in [29] that this model is somewhat less realistic than the others cited above, but it has the advantage of being mathematically simpler. Moreover, as pointed out in Remark 6.2.7, Chapter 4, and section 0.2 of [29], the use of the infinite Wasserstein distance W_∞ in the model of [10] produces results which are closer to the ones derived from the other models.

Then there are applications to PDEs. The metric structure associated with the ∞-Wasserstein distance is a crucial tool in proving the existence of stable solutions for a compressible fluid model of rotating binary stars in [24]. The same metric was used also to bound the growth of the wetted regions in the porous medium flow [13] and to study the long time asymptotics of nonlinear scalar conservation laws [12]. Moreover, the ∞-Wasserstein distance is being used in some N-particle approximations of the Vlaslov equations [21, 22].

Finally, in [11] the authors have considered a mathematical model of the optimal pricing policy for the use of a public transportation network. This model assumes that the price of a ticket (for the use of the network) is a function of the distance traveled. This seems reasonable in the case when each citizen is associated with a single journey, but it is not so realistic if we allow multiple journeys and an inexpensive season ticket is available. In the latter case, the price of a season ticket could be assumed to be a piecewise constant function of the maximal distance traveled, and hence it might be a good idea to insert a component similar to the functional we have considered into the model. It is also quite easy to imagine that in many other transportation problems a significant portion of the total cost is in one way or another connected with the maximal transportation distance. For example, if we assume that the physical transportation device (airplane, car, etc.) is the same for all distances, then it has to be chosen so that the longest transportation can be handled.

2. Existence of global solutions. As pointed out in the introduction, the objective functional

$$\begin{aligned}\lambda \mapsto \mathcal{C}_\infty(\lambda) &:= \operatorname*{\lambda\text{-ess sup}}_{(x,y)\in\Omega^2} |y-x| \\ &= \inf\Big\{t \ge 0 : \lambda(\{(x,y)\in\Omega^2 : |y-x| > t\}) = 0\Big\}\end{aligned}$$

is not linear (and not even convex) in λ, contrary to what is usually the case in classical optimal transport problems. However, it is, quite interestingly, level convex in the sense that if $\lambda_1, \lambda_2 \in \Pi(\mu,\nu)$, then

$$\mathcal{C}_\infty(t\lambda_1 + (1-t)\lambda_2) \le \max\{\mathcal{C}_\infty(\lambda_1), \mathcal{C}_\infty(\lambda_2)\} \qquad \text{for all } t \in (0,1).$$

Note that this implies that the set of solutions to (P_∞) is convex. Moreover, it should be observed that $\mathcal{C}_\infty(\lambda)$ depends on the measure λ only via its support. More precisely, one has

$$\mathcal{C}_\infty(\lambda) \;=\; \sup\{|y-x| \,:\, (x,y) \in \mathrm{supp}(\lambda)\}. \tag{2.1}$$

Thanks to this last property, we are in position to give the following existence result.

PROPOSITION 2.1. *Assume that Ω is a compact subset of $\mathbb{R}^d$ and μ, ν are two probability measures on Ω. Then the problem*

$$W_\infty(\mu,\nu) \;=\; \inf\Big\{\mathcal{C}_\infty(\lambda) := \operatorname*{\lambda\text{-ess\,sup}}_{(x,y)\in\Omega^2} |y-x| \,:\, \lambda \in \Pi(\mu,\nu)\Big\} \tag{P_∞}$$

admits at least one solution $\lambda \in \Pi(\mu,\nu)$.

The optimal set of (P_∞) may be very large thanks to (2.1).

Example 2.2. Let $\mu := \frac{1}{2}\,\mathcal{L}^2\lfloor_{[0,1]^2\cup[2,3]^2}$ and $\nu := \frac{1}{2}(\delta_{(2,1)} + \delta_{(1,2)})$. Then it is clear that the value of (P_∞) is $\sqrt{5}$ and that any admissible transport plan $\lambda \in \Pi(\mu,\nu)$ is a solution of (P_∞).

In the proof of Proposition 2.1, we shall use the following lemma.

LEMMA 2.3. *If the sequence $(\lambda_n)_n$ converges weakly to λ in $\Pi(\mu,\nu)$, then for any $(x,y) \in \mathrm{supp}(\lambda)$ there exists a sequence $((x_n,y_n))_{n\in\mathbb{N}}$ such that*

$$(x_n,y_n) \to (x,y) \text{ as } n\to\infty \quad \text{and} \quad (x_n,y_n) \in \mathrm{supp}(\lambda_n) \quad \text{for all } n \in \mathbb{N}. \tag{2.2}$$

Proof. Suppose $(x,y) \in \Omega$ is such that (2.2) does not hold. Then we may assume without loss of generality that there exists $r > 0$ such that $B((x,y),r) \cap \mathrm{supp}(\lambda_n) = \emptyset$ for any $n \in \mathbb{N}$. It then obviously follows from the weak convergence that $B((x,y),r) \cap \mathrm{supp}(\lambda) = \emptyset$, which concludes the proof. ∎

Proof of Proposition 2.1. Since the set Ω is a compact subset of $\mathbb{R}^d$ and the measures μ and ν are probability measures on Ω, the nonempty set $\Pi(\mu,\nu)$ is compact for the weak convergence of measures (cf. [32, p. 49]). To apply the direct method of the calculus of variations, it remains to notice that $\lambda \mapsto \mathcal{C}_\infty(\lambda)$ is lower semicontinuous for this topology: this is a direct consequence of (2.1) and Lemma 2.3. ∎

3. Infinitely cyclically monotone solutions. The proof of the existence of a solution to (P_∞) given in Proposition 2.1 is intrinsic, but one may obtain this result also via an approximation argument involving the family of problems $(P_p)_{p\ge1}$ given by

$$W_p(\mu,\nu) \;=\; \inf\left\{\mathcal{C}_p(\lambda) := \left(\int_{\Omega^2} |y-x|^p d\lambda(x,y)\right)^{\frac{1}{p}} \,:\, \lambda \in \Pi(\mu,\nu)\right\}; \tag{P_p}$$

that is, the functional $\lambda \mapsto \mathcal{C}_p(\lambda)$ is being minimized over the set $\Pi(\mu,\nu)$.

Alternative Proof of Proposition 2.1. Under the assumptions made on Ω, μ, and ν, for any $p \ge 1$ the problem (P_p) admits at least one solution $\gamma_p \in \Pi(\mu,\nu)$; see, e.g.,

[32, Theorem 1.3]. Since $\Pi(\mu,\nu)$ is compact, we infer that $(\gamma_p)_{p\geq 1}$ converges weakly (up to a subsequence) to some $\gamma_\infty \in \Pi(\mu,\nu)$ as $p\to\infty$. Then, for any $\lambda\in\Pi(\mu,\nu)$, we have by the optimality of γ_p and Hölder's inequality that

$$\mathcal{C}_q(\gamma_p) \;=\; \left(\int_{\Omega^2} |y-x|^q d\gamma_p(x,y)\right)^{\frac{1}{q}} \;\leq\; \mathcal{C}_p(\gamma_p) \;\leq\; \mathcal{C}_p(\lambda)$$

for any $p\geq q\geq 1$. For a fixed $q\geq 1$, since the function $(x,y)\mapsto |y-x|^q$ is continuous and bounded on Ω^2 one has $\mathcal{C}_q(\gamma_p)\to\mathcal{C}_q(\gamma_\infty)$ as $p\to\infty$. Therefore, taking the limit in p and then in q in the above inequality we obtain $\mathcal{C}_\infty(\gamma_\infty)\leq\mathcal{C}_\infty(\lambda)$. Since this holds for any $\lambda\in\Pi(\mu,\nu)$, γ_∞ is a minimizer of (P_∞). ∎

The reason for considering the problems (P_p) in this context is not merely the fact that they provide an alternative route to the existence. Namely, it is known that an element $\gamma_p\in\Pi(\mu,\nu)$ is a solution of (P_p) if and only if its support is p-cyclically monotone; that is,

$$\sum_{i=1}^n |y_i-x_i|^p \;\leq\; \sum_{i=1}^n |y_{\sigma(i)}-x_i|^p \tag{3.1}$$

for every $n\geq 2$, $(x_1,y_1),\dots,(x_n,y_n)\in\operatorname{supp}(\gamma_p)$, and for every permutation $\sigma\in\mathcal{S}_n$. We refer the reader, for example, to Theorem 3.2 in [2] (or to [20, 28, 33]).

By analogy with the p-cyclical monotonicity when $1\leq p<\infty$, we introduce the corresponding notion for the case $p=\infty$ obtained by taking the limit in (3.1).

Definition 3.1. *A transport plan $\gamma\in\Pi(\mu,\nu)$ is* infinitely cyclically monotone *if*

$$\max_{1\leq i\leq n} |y_i-x_i| \;\leq\; \max_{1\leq i\leq n} |y_{\sigma(i)}-x_i|$$

for every $n\geq 2$, $(x_1,y_1),\dots,(x_n,y_n)\in\operatorname{supp}(\gamma)$, and $\sigma\in\mathcal{S}_n$.

Using again the approximation of (P_∞) by the problems (P_p), we obtain the existence of an infinitely cyclically monotone solution to (P_∞).

Theorem 3.2. *For $1\leq p<\infty$, let $\gamma_p\in\Pi(\mu,\nu)$ be a solution to (P_p). Then any cluster point γ_∞ of $(\gamma_p)_{p\geq 1}$ in $\Pi(\mu,\nu)$ as $p\to\infty$ is an infinitely cyclically monotone solution to (P_∞).*

Proof. For simplicity, let us assume that the entire family $(\gamma_p)_{p\geq 1}$ converges weakly to $\gamma_\infty\in\Pi(\mu,\nu)$. It suffices to show that γ_∞ is infinitely cyclically monotone. To this end, let $n\geq 2$, $(x_1,y_1),\dots,(x_n,y_n)\in\operatorname{supp}(\gamma)$, and $\sigma\in\mathcal{S}_n$. We apply Lemma 2.3 to each pair (x_i,y_i) to obtain the existence of sequences $(x_1^p,y_1^p),\dots,(x_n^p,y_n^p)$ such that $(x_i^p,y_i^p)\to(x_i,y_i)$ for any i as $p\to\infty$, and $(x_i^p,y_i^p)\in\operatorname{supp}(\gamma_p)$ for all $1\leq p<\infty$ and $i=1,\dots,n$. Since the support of γ_p is p-cyclically monotone, one has

$$\sum_{i=1}^n |y_i^p-x_i^p|^p \;\leq\; \sum_{i=1}^n |y_{\sigma(i)}^p-x_i^p|^p \qquad \text{for all } 1<p<\infty.$$

Taking the $1/p$-power on both sides and letting p go to ∞, one obtains the desired inequality. ∎

Since for any $1\leq p<\infty$ an admissible transport plan $\lambda\in\Pi(\mu,\nu)$ is a minimizer of (P_p) if and only if it is p-cyclically monotone, it is natural to ask whether this still holds for $p=\infty$. It is, however, quite clear that a generic minimizer of (P_∞) need

not be infinitely cyclically monotone; the following gives a counterexample for this implication.

Example 3.3. As an admissible transport plan for the measures μ and ν in Example 2.2 one may take

$$\lambda := \frac{1}{2}\left(\mathcal{L}^2\lfloor_{[0,1]^2} \times \delta_{(2,1)} + \mathcal{L}^2\lfloor_{[2,3]^2} \times \delta_{(1,2)}\right).$$

Then λ is a minimizer of (P_∞), but it is not infinitely cyclically monotone: for example, $((0,1),(2,1))$ and $((3,2),(1,2))$ belong to $\mathrm{supp}(\lambda)$, and

$$\max\{|(0,1)-(1,2)|,\, |(3,2)-(2,1)|\} \;<\; \max\{|(0,1)-(2,1)|,\, |(3,2)-(1,2)|\}.$$

Notice that in this case, for any $1 < p < \infty$, problem (P_p) admits a unique solution (up to a μ-negligible set) (see [20]), which in fact does not depend on p.

On the other hand, the following result shows that the reverse implication does hold: infinite cyclical monotonicity is indeed a sufficient condition for an admissible plan to be a minimizer of (P_∞).

THEOREM 3.4. *Any infinitely cyclically monotone transport plan $\gamma \in \Pi(\mu,\nu)$ is a solution of the problem (P_∞).*

Proof. We make a proof by contradiction. Let $\gamma \in \Pi(\mu,\nu)$ be infinitely cyclically monotone, and assume that

$$\gamma\text{-}\operatorname*{ess\,sup}_{(x,y)\in\Omega^2} |y-x| \;\geq\; 10\,\varepsilon + \tilde{\gamma}\text{-}\operatorname*{ess\,sup}_{(x,y)\in\Omega^2} |y-x| \tag{3.2}$$

for some $\tilde{\gamma} \in \Pi(\mu,\nu)$ and $\varepsilon > 0$.

Since Ω is compact, there exists a finite family $(c_i)_{1\leq i\leq k}$ such that $\Omega \subset \bigcup_{i=1}^k B(c_i,\varepsilon)$. We shall denote $C := \{c_1,\dots,c_k\}$ and $V_1 := B(c_1,\varepsilon)$, and for any $i \in \{2,\dots,k\}$ we set $V_i := B(c_i,\varepsilon) \setminus \bigcup_{j=1}^{i-1} V_j$; without loss of generality, we assume that $V_i \neq \emptyset$ for all $i \in \{1,\dots,k\}$.

Next we define two discrete measures γ^ε and $\tilde{\gamma}^\varepsilon$ on Ω^2 by

$$\gamma^\varepsilon \;:=\; \sum_{1\leq i,j\leq k} \gamma(V_i\times V_j)\,\delta_{(c_i,c_j)}$$

and

$$\tilde{\gamma}^\varepsilon \;:=\; \sum_{1\leq i,j\leq k} \tilde{\gamma}(V_i\times V_j)\,\delta_{(c_i,c_j)}.$$

Notice that since γ and $\tilde{\gamma}$ have the same marginals, the same holds for γ^ε and $\tilde{\gamma}^\varepsilon$. In particular, one has

$$(x,y) \in \mathrm{supp}(\gamma^\varepsilon) \quad\Rightarrow\quad \text{there exists } \tilde{x} \in C \text{ such that } (\tilde{x},y) \in \mathrm{supp}(\tilde{\gamma}^\varepsilon) \tag{3.3}$$

and

$$(\tilde{x},\tilde{y}) \in \mathrm{supp}(\tilde{\gamma}^\varepsilon) \quad\Rightarrow\quad \text{there exists } y \in C \text{ such that } (\tilde{x},y) \in \mathrm{supp}(\gamma^\varepsilon). \tag{3.4}$$

The following properties will also be useful in our argument.

Claim 1. There exists (x_0,y_0) in the support of γ^ε such that

$$|y_0-x_0| \;\geq\; 5\,\varepsilon + \max\left\{|y-x| \,:\, (x,y) \in \mathrm{supp}(\tilde{\gamma}^\varepsilon)\right\}.$$

Claim 2. For any $n \geq 1$, $(x_1, y_1), \dots, (x_n, y_n) \in \mathrm{supp}(\gamma^\varepsilon)$, and $\sigma \in \mathcal{S}_n$,

$$\max_{1 \leq i \leq n} |y_i - x_i| \ \leq \ 4\,\varepsilon + \max_{1 \leq i \leq n} |y_{\sigma(i)} - x_i|.$$

Above, the first claim is simply a counterpart of the antithesis (3.2) for the discretized measures, while the second says that γ^ε is "almost" infinitely cyclically monotone. We postpone the verification of these two claims until the end of this proof.

Let $(x_0, y_0) \in \mathrm{supp}(\gamma^\varepsilon)$ be given by Claim 1. Owing to (3.3) and (3.4), we can recursively define two sequences $(D_m)_{m \geq 1}$ and $(E_m)_{m \geq 0}$ of subsets of C by setting $E_0 := \{y_0\}$, and for $m \geq 1$,

$$D_m \ := \ \{\tilde{x} \ : \ \text{there exists } y \in E_{m-1} \text{ such that } (\tilde{x}, y) \in \mathrm{supp}(\tilde{\gamma}^\varepsilon)\}$$

and

$$E_m \ := \ \{y \ : \ \text{there exists } \tilde{x} \in D_m \text{ such that } (\tilde{x}, y) \in \mathrm{supp}(\gamma^\varepsilon)\}\,.$$

We then set $D := \bigcup_{m \geq 1} D_m$ and $E := \bigcup_{m \geq 0} E_m$.

There are now two alternatives: either x_0 belongs to D or not.

First case: $x_0 \in D$. In this case, there exists $m \geq 1$ such that $x_0 \in D_m$, and by going backwards from D_m to E_0 it is possible to define two finite families $(x_i)_{0 \leq i \leq m}$ and $(y_i)_{0 \leq i \leq m-1}$ such that

$$\text{for all } i \in \{0, \dots, m-1\}, \quad (x_i, y_i) \in \mathrm{supp}(\gamma^\varepsilon) \quad \text{and} \quad (x_{i+1}, y_i) \in \mathrm{supp}(\tilde{\gamma}^\varepsilon),$$

where we have set $x_m := x_0$. Claim 2 then yields

$$\max_{0 \leq i \leq m-1} |y_i - x_i| - 4\,\varepsilon \ \leq \ \max_{0 \leq i \leq m-1} |y_i - x_{i+1}|.$$

Since $\max_{0 \leq i \leq m-1} |y_i - x_i| \ \geq \ |y_0 - x_0|$, we infer from Claim 1 and the previous inequality that

$$\max\left\{|y - x| \ : \ (x, y) \in \mathrm{supp}(\tilde{\gamma}^\varepsilon)\right\} + \varepsilon \ \leq \ \max_{0 \leq i \leq m-1} |y_i - x_{i+1}|.$$

Since $(x_{i+1}, y_i) \in \mathrm{supp}(\tilde{\gamma}^\varepsilon)$ for any $i \in \{0, \dots, m-1\}$, this yields a contradiction.

Second case: $x_0 \notin D$. From the definitions of D and E, we notice the following two facts:

$$x \in D,\, (x, y) \in \mathrm{supp}(\gamma^\varepsilon) \quad \Rightarrow \quad y \in E \tag{3.5}$$

and

$$\tilde{y} \in E,\, (\tilde{x}, \tilde{y}) \in \mathrm{supp}(\tilde{\gamma}^\varepsilon) \quad \Rightarrow \quad \tilde{x} \in D. \tag{3.6}$$

As a consequence of (3.5) and since γ^ε and $\tilde{\gamma}^\varepsilon$ have the same marginals, one has

$$\gamma^\varepsilon(D \times E) \ = \ \gamma^\varepsilon(D \times C) \ = \ \tilde{\gamma}^\varepsilon(D \times C).$$

Similarly, one has

$$\tilde{\gamma}^\varepsilon(D \times E) \ = \ \tilde{\gamma}^\varepsilon(C \times E) \ = \ \gamma^\varepsilon(C \times E).$$

We then obtain

$$\gamma^\varepsilon(D\times E) \;=\; \tilde\gamma^\varepsilon(D\times C) \;\geq\; \tilde\gamma^\varepsilon(D\times E) \;=\; \gamma^\varepsilon(C\times E).$$

This implies that $\gamma^\varepsilon((C\setminus D)\times E)=0$, whereas by hypothesis one has $(x_0,y_0)\in (C\setminus D)\times E$ and $\gamma^\varepsilon(\{(x_0,y_0)\})>0$ since (x_0,y_0) belongs to the support of the discrete measure γ^ε. This yields a contradiction.

To complete the proof of Theorem 3.4, it remains to prove Claims 1 and 2.

Proof of Claim 1. We infer from (3.2) that

$$\gamma\left(\left\{(x,y)\,:\,|y-x|\geq 9\,\varepsilon+\tilde\gamma\text{-}\operatorname*{ess\,sup}_{(x,y)\in\Omega^2}|y-x|\right\}\right)>0.$$

As a consequence, there exist $i_1,i_2\in\{1,\dots,k\}$ such that

$$\gamma\left((V_{i_1}\times V_{i_2})\cap\left\{(x,y)\,:\,|y-x|\geq 9\,\varepsilon+\tilde\gamma\text{-}\operatorname*{ess\,sup}_{(x,y)\in\Omega^2}|y-x|\right\}\right)>0.$$

Since $V_m\subset B(c_m,\varepsilon)$ for $m=i_1,i_2$, one then has

$$|c_{i_2}-c_{i_1}| \;\geq\; 7\,\varepsilon+\tilde\gamma\text{-}\operatorname*{ess\,sup}_{(x,y)\in\Omega^2}|y-x|, \tag{3.7}$$

and (c_{i_1},c_{i_2}) belongs to the support of γ^ε. On the other hand, if (c_{j_1},c_{j_2}) belongs to the support of $\tilde\gamma^\varepsilon$, then $\tilde\gamma(V_{j_1}\times V_{j_2})>0$, and thus

$$|c_{j_2}-c_{j_1}| \;\leq\; 2\,\varepsilon+\tilde\gamma\text{-}\operatorname*{ess\,sup}_{(x,y)\in\Omega^2}|y-x|.$$

Since this inequality holds whenever $(c_{j_1},c_{j_2})\in\operatorname{supp}(\tilde\gamma^\varepsilon)$, one has

$$\max\{|y-x|\,:\,(x,y)\in\operatorname{supp}(\tilde\gamma^\varepsilon)\} \;\leq\; 2\,\varepsilon+\tilde\gamma\text{-}\operatorname*{ess\,sup}_{(x,y)\in\Omega^2}|y-x|.$$

This together with (3.7) shows that $(x_0,y_0):=(c_{i_1},c_{i_2})$ has the desired property.

Proof of Claim 2. Let $n\geq 1$, $\sigma\in\mathcal{S}_n$, and (x_i,y_i) belong to the support of γ^ε for $i\in\{1,\dots,n\}$. For any $i\in\{1,\dots,n\}$, one has $(x_i,y_i)=(c_{j_1},c_{j_2})$ for some $j_1,j_2\in\{1,\dots,k\}$ with $\gamma(V_{j_1}\times V_{j_2})>0$, and thus there exists $(x_i',y_i')\in(V_{j_1}\times V_{j_2})\cap\operatorname{supp}(\gamma)$. As a consequence,

$$\big|\,|y_r'-x_s'|-|y_r-x_s|\,\big| \;\leq\; 2\,\varepsilon \qquad \text{for all } r,s\in\{1,\dots,n\}.$$

Since γ is infinitely cyclically monotone, we have

$$\max_{1\leq i\leq n}|y_{\sigma(i)}'-x_i'| \;\geq\; \max_{1\leq i\leq n}|y_i'-x_i'|.$$

It follows that

$$4\,\varepsilon+\max_{1\leq i\leq n}|y_{\sigma(i)}-x_i| \;\geq\; \max_{1\leq i\leq n}|y_i-x_i|,$$

which proves the claim. □

Remark 3.5. Observe that in the proof above, we in fact always have $x_0\in D$; that is, the *first case* always occurs. This is a consequence of the conservation of the masses: all the mass transported to E by $\tilde\gamma^\varepsilon$ originates from D, while all the mass in

D is transported to E by γ^ε. Since γ^ε and $\tilde{\gamma}^\varepsilon$ have the same marginals, this implies that both plans transport D exactly to E. The fact that the *second case* never occurs in the above proof also underlies the recursive construction of *Step* II in the proof of Theorem A in [25], even if the arguments are different. That paper deals with the sufficiency of cyclical monotonicity for optimality in the classical case (see also [31]).

Remark 3.6. There is a variation of the self-contained proof given above that relies directly on the fact that the total cost $\mathcal{C}_\infty(\lambda) = \lambda\text{-ess sup}_{(x,y)\in\Omega^2} |y-x|$ depends only on the support of λ and not on its density. In the discrete case this means that, as far as the total cost is concerned, the exact amount of mass transferred from any given point to another is irrelevant: it matters only whether the amount is positive or not. Hence in the proof we are allowed to change the transport plans γ_ε and $\tilde{\gamma}_\varepsilon$, along with their marginals, as long as we do not change their supports and make sure that the marginals of the new transport plans agree with each other. Now assuming that we can change the measures in such a way that all the point masses are of integer size, the problem can be interpreted as a pairing problem in which the infinite cyclical monotonicity is both a necessary and sufficient condition for optimality.

The existence of the required integer transport plans with given supports is non-trivial and follows from Dines' algorithm [17], which provides positive solutions for a system of linear equations.

4. Restrictable solutions. In the previous section, we derived the notion of infinitely cyclically monotone plans from the approximation of the problem (P_∞) by the family of problems (P_p). Another interesting notion may be derived in the same way: let $1 \le p < \infty$ and $\gamma_p \in \Pi(\mu,\nu)$ be a solution to (P_p). Then it follows from the linearity of the functional

$$\gamma \mapsto \mathcal{C}_p(\gamma)^p \;=\; \int_{\Omega^2} |y-x|^p d\lambda(x,y)$$

that any nonzero measure γ' that is majorized by γ_p, i.e., $\gamma'(B) \le \gamma_p(B)$ for all Borel sets $B \subset \Omega\times\Omega$, is an optimal transport plan for the problem

$$\mathcal{C}_p(\gamma') = \inf\left\{\mathcal{C}_p(\gamma) : \gamma \in \Pi(\mu',\nu')\right\},$$

where $\mu' := \pi_1{}_\#\gamma'$ and $\nu' := \pi_2{}_\#\gamma'$. In other words, optimality is automatically inherited by restriction, and hence we may say that a solution $\gamma_p \in \Pi(\mu,\nu)$ of (P_p) is a *restrictable solution* of this problem. By analogy, we may define a similar notion of restrictable solutions for problem (P_∞) as follows.

Definition 4.1. *A transport plan $\gamma \in \Pi(\mu,\nu)$ is a* restrictable solution *of (P_∞) if any nonzero Borel measure γ' in $\Omega\times\Omega$ that is majorized by γ is a solution to the problem*

$$(P'_\infty) \qquad \inf\left\{\lambda\text{-ess sup}_{(x,y)\in\Omega^2} |y-x| \;:\; \lambda \in \Pi(\mu',\nu')\right\},$$

where $\mu' := \pi_1{}_\#\gamma'$ and $\nu' := \pi_2{}_\#\gamma'$.

The reader should notice the obvious abuse of notation above, as the measures μ' and ν' in (P'_∞) are not, in general, probability measures. However, $\mu'(\Omega) = \nu'(\Omega) > 0$, and that is really all that is needed.

Example 4.2. It is quite clear that not every solution of (P_∞) is restrictable. Indeed, the optimal plan λ considered in Example 3.3 admits the following restriction:

$$\lambda' := \frac{1}{2}\left(\mathcal{L}^2\lfloor_{S_1} \times \tfrac{1}{2}\delta_{(2,1)} + \mathcal{L}^2\lfloor_{S_2} \times \tfrac{1}{2}\delta_{(1,2)}\right),$$

where $S_1 := [0, \frac{1}{2}] \times [\frac{1}{2}, 1]$ and $S_2 := [\frac{5}{2}, 3] \times [2, \frac{5}{2}]$. But λ' is not optimal for its own marginals: a better transport plan is the one that takes all the mass that lies in S_1 to $(1, 2)$ and the mass in S_2 to $(2, 1)$.

Remark 4.3. The notion of a restrictable solution bears a strong resemblance to that of an *absolute minimizer* used in connection with the L^∞ variational problems. We recall that a locally Lipschitz continuous function $u : D \to \mathbb{R}^m$, $m \geq 1$, is called an absolute minimizer of the functional $S(\varphi, D) := \operatorname{ess\,sup}_{x \in D} H(x, \varphi(x), D\varphi(x))$ if

$$S(u, V) \leq S(v, V)$$

for every open $V \subset\subset D$ and $v \in W^{1,\infty}(V) \cap C(\overline{V}))$ such that $v|_{\partial V} = u|_{\partial V}$. Absolute minimizers were introduced by Aronsson [3, 4]. It has turned out that this is the proper notion of a solution for this type of minimization problem in the sense that important properties such as uniqueness, regularity, and characterization via an Euler–Lagrange equation can be obtained for this class of functions. Absolute minimizers were introduced by Aronsson [3]; see also, e.g., [6, 5, 30, 14] for further details and background.

It is natural to ask whether any cluster point γ_∞ of $(\gamma_p)_{p \geq 1}$ in $\Pi(\mu, \nu)$ as $p \to \infty$ is a restrictable solution of (P_∞). In view of Theorem 3.2, this can be established by showing that the restrictable solutions coincide with the class of infinitely cyclically monotone solutions.

THEOREM 4.4. *A transport plan $\gamma \in \Pi(\mu, \nu)$ is infinitely cyclically monotone if and only if it is a restrictable solution of the problem (P_∞).*

Notice that Theorem 3.2 provides the existence of restrictable solutions to (P_∞).

Proof. If $\gamma \in \Pi(\mu, \nu)$ is infinitely cyclically monotone, then the same holds for any restriction $\gamma' \leq \gamma$, and Theorem 3.4 then yields that such a restriction γ' is a solution of the corresponding problem (P'_∞).

We now turn to the proof of the sufficiency. Let $\gamma \in \Pi(\mu, \nu)$ be a restrictable solution to (P_∞), and let us fix points $(x_1, y_1), \ldots, (x_m, y_m) \in \operatorname{supp}(\gamma)$, $m \geq 2$, and a permutation σ of $\{1, \ldots, m\}$. Without loss of generality, we may assume that $(x_i, y_i) \neq (x_j, y_j)$ whenever $i \neq j$. Then there is $\varepsilon_0 > 0$ such that for all $0 < \varepsilon \leq \varepsilon_0$, the sets $B_i := B(x_i, \varepsilon) \times B(y_i, \varepsilon)$ are pairwise disjoint and $\gamma(B_i) > 0$ for all $i = 1, \ldots, m$.

We define two measures γ' and γ_σ by setting

$$\gamma' := \sum_{i=1}^m c_i \gamma \lfloor_{B_i} \qquad \text{and} \qquad \gamma_\sigma := \sum_{i=1}^m c_i T^i_\# \gamma \lfloor_{B_i}.$$

Here $T^i_\# \gamma \lfloor_{B_i}$ is the push-forward of $\gamma \lfloor_{B_i}$ by the mapping $T^i(x, y) := (x, y + y_{\sigma(i)} - y_i)$, and the positive numbers

$$c_i := \frac{\min_k \gamma(B_k)}{\gamma(B_i)}$$

are chosen so that $\gamma'(B_i) = \min_k \gamma(B_k) > 0$ is independent of i. Observe that the support of $T^i_\# \gamma \lfloor_{B_i}$ lies in $B^\sigma_i := B(x_i, \varepsilon) \times B(y_{\sigma(i)}, \varepsilon)$ and $\gamma_\sigma(B^\sigma_i) = \gamma'(B_i)$. Moreover, the first marginals $\mu' = \pi_1 \#\gamma'$ and $\mu_\sigma = \pi_1 \#\gamma_\sigma$ are equal.

Since γ' is majorized by the restrictable solution γ, we have

$$\gamma'\text{-}\operatorname*{ess\,sup}_{(x,y) \in \Omega^2} |x - y| = W_\infty(\mu', \nu') = \inf \left\{ \tilde{\gamma}\text{-}\operatorname*{ess\,sup}_{(x,y) \in \Omega^2} |x - y| \, : \, \tilde{\gamma} \in \Pi(\mu', \nu') \right\},$$

where $\nu' = \pi_2{}_\#\gamma'$. On the other hand, since the supports of both ν' and $\nu_\sigma = \pi_2{}_\#\gamma_\sigma$ are contained in the union of the balls $B(y_i, \varepsilon)$ and $\nu'(B(y_i, \varepsilon)) = \nu_\sigma(B(y_i, \varepsilon))$ for every i by the construction of γ' and γ_σ, we can rearrange ν' to ν_σ by transporting mass only within the balls $B(y_i, \varepsilon)$. Thereby we obtain that $W_\infty(\nu', \nu_\sigma) \leq 2\varepsilon$, and hence

$$\begin{aligned}\gamma'\text{-}\operatorname*{ess\,sup}_{(x,y)\in\Omega^2} |x-y| = W_\infty(\mu', \nu') &\leq W_\infty(\mu_\sigma, \nu_\sigma) + W_\infty(\nu_\sigma, \nu') \\ &\leq 2\varepsilon + \gamma_\sigma\text{-}\operatorname*{ess\,sup}_{(x,y)\in\Omega^2} |x-y|.\end{aligned}$$

Now clearly

$$\gamma'\text{-}\operatorname*{ess\,sup}_{(x,y)\in\Omega^2} |x-y| \geq \max_{1\leq i\leq m} |x_i - y_i|$$

and

$$\gamma_\sigma\text{-}\operatorname*{ess\,sup}_{(x,y)\in\Omega^2} |x-y| \leq \max_{1\leq i\leq m} |x_i - y_{\sigma(i)}| + 2\varepsilon,$$

and thus the preceding inequality yields

$$\max_{1\leq i\leq m} |x_i - y_i| \leq \max_{1\leq i\leq m} |x_i - y_{\sigma(i)}| + 4\varepsilon.$$

Since this holds for all $\varepsilon > 0$ small enough we are done. ∎

Observe that for any $\gamma \in \Pi(\mu, \nu)$ and any Borel set $B \subset \Omega\times\Omega$ such that $\gamma(B) > 0$, the measure $\gamma\lfloor_B$ is majorized by γ. Thus, if γ is a restrictable solution of (P_∞), then each such measure $\gamma\lfloor_B$ is an optimal transport plan for its own marginals. It turns out that in general the converse is false; that is, this family of measures alone does not suffice for characterizing the restrictable solutions, as shown in Example 4.5. Notice that this is another difference with the case of integral costs functionals (like the costs $\mathcal{C}_p$), since for those functionals the converse would be true: if $\gamma\lfloor_B$ is an optimal transport plan for its own marginals for any $B \subset \Omega \times \Omega$ with $\gamma(B) > 0$, then γ is a restrictable solution.

Example 4.5. Take $\Omega = [0, 1]$, and let

$$\mu := \frac{1}{3}\delta_0 + \frac{2}{3}\delta_1 \qquad \text{and} \qquad \nu := \frac{2}{3}\delta_0 + \frac{1}{3}\delta_1.$$

Then the plan $\gamma := \frac{1}{3}\delta_{(0,1)} + \frac{2}{3}\delta_{(1,0)}$ is not a restrictable solution since $\gamma' := \frac{1}{3}\delta_{(0,1)} + \frac{1}{3}\delta_{(1,0)}$ is majorized by γ, and it is clearly not an optimal transport plan for its own marginals $\mu' = \frac{1}{3}\delta_0 + \frac{1}{3}\delta_1$ and $\nu' = \frac{1}{3}\delta_0 + \frac{1}{3}\delta_1$. On the other hand, one can check that $\gamma\lfloor_B$ is an optimal transport plan, for its own marginals, for each Borel set $B \subset \Omega^2$. Notice that the only restrictable solution of (P_∞) in this case (which is also the unique solution to (P_p) when $p \geq 1$) is $\gamma_\infty = \frac{1}{3}\delta_{(0,0)} + \frac{1}{3}\delta_{(1,0)} + \frac{1}{3}\delta_{(1,1)}$.

In view of the above example, and under some regularity assumption on the measures μ and ν, it is possible to obtain the following refined version of Theorem 4.4.

PROPOSITION 4.6. *Let $\gamma \in \Pi(\mu, \nu)$ be an optimal transport plan, and assume that neither μ nor ν concentrates on sets of dimension $d-1$. Then the following are equivalent:*

(1) *γ is infinitely cyclically monotone;*
(2) *for each Borel set $B \subset \Omega \times \Omega$, $\gamma\lfloor_B$ is optimal between its projections.*

Proof. We need only prove that under these assumptions (2) implies (1). Assume by contradiction that γ is not infinitely cyclically monotone. Then there exist a family $\{(x_i, y_i)\}_{i=1,\dots,n}$ in $\mathrm{supp}(\gamma)$ and a permutation $\sigma \in \mathcal{S}_n$ such that

$$\max\{|x_1 - y_{\sigma(1)}|, \dots, |x_n - y_{\sigma(n)}|\} < \max\{|x_1 - y_1|, \dots, |x_n - y_n|\}.$$

By continuity, the same inequality holds true for any family $\{(x_i', y_i')\}_{i=1,\dots,n}$ for which $(x_i', y_i') \in B(x_i, \varepsilon) \times B(y_i, \varepsilon)$ for all $i = 1, \dots, n$ and for some small enough $\varepsilon > 0$. Notice that $\gamma(B(x_i, \varepsilon) \times B(y_i, \varepsilon))$ is positive for any i.

For each i we define $g_i : [0, \varepsilon] \to \mathbb{R}^+$ by $g_i(r) := \gamma(B(x_i, r) \times B(y_i, r))$. Since μ and ν do not concentrate on sets of dimension $d-1$, the function g_i is continuous. Let $\alpha = \min_i \gamma(B(x_i, \varepsilon) \times B(y_i, \varepsilon))$, and choose $0 < \tilde{\varepsilon}_i \leq \varepsilon$ so that $g_i(\tilde{\varepsilon}_i) = \alpha$ for all i; this is possible since $\lim_{r \to 0} g_i(r) = 0$.

Following now the proof of sufficiency of the previous theorem we obtain that $\gamma \lfloor_B$ for $B = \bigcup_{i=1}^n B(x_i, \tilde{\varepsilon}_i) \times B(y_i, \tilde{\varepsilon}_i)$ and for $\varepsilon > 0$ small enough is not an optimal transport between its marginals, which contradicts (2). ∎

Remark 4.7. It is natural to ask what happens if $|x - y|$ is replaced by a more general (real-valued) cost function $c(x, y)$; that is, we consider the functional

$$\gamma \mapsto \gamma\text{-ess sup}_{(x,y) \in \Omega^2} c(x, y).$$

As expected, the basic existence results, Proposition 2.1 and Theorem 3.2, remain valid, provided that c is nonnegative and lower semicontinuous with all the relevant concepts appropriately redefined: in particular, in the definition of *infinite cyclical monotonicity* one should replace the support of γ with some appropriate set on which γ is concentrated. Moreover, the proof of the equivalence of restrictable and infinitely cyclically monotone solutions works in this generality if $c(x, y)$ is (uniformly) continuous.

5. Existence and uniqueness of an optimal transport map. In this section, we prove that under reasonably weak assumptions an infinitely cyclically monotone transport plan is induced by a transport map. Moreover, we start the analysis of the uniqueness of such transport maps and then comment on our method of proof in light of the duality issue for problem (P_∞).

5.1. Properties of transport plans. We begin by considering some generic properties of transport plans. This subsection is largely independent of the cost, and the technique detailed below has applications also in the framework of classical transportation problems involving cost functionals in integral form; see [15].

Definition 5.1. *Let $y \in \Omega$ and $r > 0$, and let $\gamma \in \Pi(\mu, \nu)$ be a transport plan. We define*

$$\gamma^{-1}(B(y, r)) := \pi^1((\Omega \times B(y, r)) \cap \mathrm{supp}\, \gamma).$$

In other words, $\gamma^{-1}(B(y, r))$ is the set of points whose mass is partially or completely transported to $B(y, r)$ by γ. We recognize the slight abuse of notation, but if γ is thought of as a device that transports mass, then this seems justifiable. Notice also that $\gamma^{-1}(B(y, r))$ is a Borel set. In fact, it is a countable union of compact sets as shown by the equation

$$\pi^1((\Omega\times B(y,r))\cap\operatorname{supp}\gamma)=\pi^1\left(\bigcup_n(\Omega\times\overline{B(y,r-1/n)})\cap\operatorname{supp}\gamma\right)$$
$$=\bigcup_n\pi^1((\Omega\times\overline{B(y,r-1/n)})\cap\operatorname{supp}\gamma).$$

Since this notion is important in what follows, we recall that when A is $\mathcal{L}^d$-measurable, one has

$$\lim_{r\to 0}\frac{\mathcal{L}^d(A\cap B(x,r))}{\mathcal{L}^d(B(x,r))}\ =\ 1$$

for almost every x in A: we shall call such a point x a Lebesgue point of A, this terminology deriving from the fact that such a point may also be considered as a Lebesgue point of χ_A.

The following lemma, although quite simple, is the cornerstone of the proof of Theorem 5.5 below.

LEMMA 5.2. *Let $\gamma\in\Pi(\mu,\nu)$, and assume that $\mu\ll\mathcal{L}^d$. Then γ is concentrated on a σ-compact set $R(\gamma)$ such that for all $(x,y)\in R(\gamma)$ the point x is a Lebesgue point of $\gamma^{-1}(B(y,r))$ for all $r>0$.*

Proof. In the following, we shall denote by $\mathrm{Leb}(E)$ the set of points $x\in E$ which are Lebesgue points of E. Let

$$A:=\{(x,y)\in\operatorname{supp}(\gamma)\ :\ \ x\notin\mathrm{Leb}(\gamma^{-1}(B(y,r)))\text{ for some }r>0\};$$

we first intend to show that $\gamma(A)=0$. To this end, for each positive integer n we consider a finite covering $\Omega\subset\bigcup_{i\in I(n)}B(y_i^n,\frac{1}{2n})$ by balls of radius $\frac{1}{2n}$. We notice that if $(x,y)\in\operatorname{supp}(\gamma)$ and x is not a Lebesgue point of $\gamma^{-1}(B(y,r))$ for some $r>0$, then for any $n\geq\frac{1}{r}$ and y_i^n such that $|y_i^n-y|<\frac{1}{2n}$ the point x belongs to $\gamma^{-1}(B(y_i^n,\frac{1}{2n}))$ but is not a Lebesgue point of this set. Then

$$\pi^1(A)\ \subset\ \bigcup_{n\geq1}\bigcup_{i\in I(n)}\left(\gamma^{-1}\left(B\left(y_i^n,\frac{1}{2n}\right)\right)\setminus\mathrm{Leb}\left(\gamma^{-1}\left(B_\gamma\left(y_i^n,\frac{1}{2n}\right)\right)\right)\right).$$

Notice that the set on the right-hand side has Lebesgue measure 0 and thus μ-measure 0. It follows that $\gamma(A)\leq\gamma(\pi^1(A)\times\Omega)=\mu(\pi^1(A))=0$.

Finally, since $\mathcal{L}^d(\pi^1(A))=0$, there exists a sequence $(U_k)_{k\geq0}$ of open sets such that

$$\text{for all }k\geq0,\qquad\pi^1(A)\subset U_k\qquad\text{and}\qquad\lim_{k\to\infty}\mathcal{L}^d(U_k)=0.$$

Then the set $R(\gamma):=\operatorname{supp}(\gamma)\cap(\bigcup_{k\geq0}(\Omega\setminus U_k)\times\Omega)$ has the desired properties. ∎

The above lemma yields the introduction of the following notion.

DEFINITION 5.3. *The couple $(x,y)\in\Omega\times\Omega$ is a* γ-regular point *if $x\in\gamma^{-1}(B(y,r))$ is a Lebesgue point of this set for any positive r.*

Notice that any element of the set $R(\gamma)$ of Lemma 5.2 is a γ-regular point.

For future use, we introduce a suitable notation to indicate a cone: let $x_0,\xi\in\mathbb{R}^d$ with $|\xi|=1$, and let $\delta\in[0,2]$. Then we define

$$C(x_0,\xi,\delta):=\left\{x\in\mathbb{R}^d\setminus\{x_0\}\ :\ \frac{x-x_0}{|x-x_0|}\cdot\xi\geq1-\delta\right\}.$$

Notice that if $\delta = 0$, $C(x_0, \xi, 0)$ degenerates to a half-line, while in the case $\delta = 2$, $C(x_0, \xi, 2)$ is $\mathbb{R}^d \setminus \{x_0\}$.

We now remark the following property for the regular points of a transport plan.

PROPOSITION 5.4. *Let (x_0, y_0) be a γ-regular point, $r > 0$, $\alpha \in (0,1)$, and $\delta > 0$. Then for $\varepsilon > 0$ sufficiently small the set of points $x \in \gamma^{-1}(B(y_0, r))$ such that $x \in C(x_0, \xi, \delta) \cap (B(x_0, \varepsilon) \setminus B(x_0, \alpha\varepsilon))$ has positive $\mathcal{L}^d$ measure.*

Proof. It is enough to remark that $x_0 \in \mathrm{Leb}(\gamma^{-1}(B(y_0, r)))$ and then

$$\lim_{\varepsilon \to 0} \frac{\mathcal{L}^d((B(x_0, \varepsilon) \setminus B(x_0, \alpha\varepsilon)) \cap C(x_0, \xi, \delta) \cap \gamma^{-1}(B(y_0, r)))}{\mathcal{L}^d(B(x_0, \varepsilon))} \;=\; c(\alpha, \delta),$$

where $c(\alpha, \delta) := \frac{\mathcal{L}^d((B(x_0,1) \setminus B(x_0,\alpha)) \cap C(x_0,\xi,\delta))}{\mathcal{L}^d(B(x_0,1))} > 0$. ∎

5.2. Existence of an optimal transport map. Our main result in this subsection is the following theorem, which states that under the hypothesis that μ is absolutely continuous with respect to the Lebesgue measure $\mathcal{L}^d$, any optimal infinitely cyclically monotone transport plan for (P_∞) is induced by a transport map. This generalizes the corresponding result for the problem (P_p) when $p \in]1, \infty[$: if one assumes that

$$\mu(B) = 0 \quad \text{whenever} \quad \mathcal{H}^{d-1}(B) < \infty, \tag{5.1}$$

then any p-cyclically monotone transport plan is induced by a transport map (see Remark 4.7 in [20]).

It is in doubt whether (5.1) is sufficient to ensure that the conclusion of Theorem 5.5 holds. In the case $p = 1$, the hypothesis (5.1) is not sufficient to ensure that any 1-cyclically monotone transport plan is induced by a transport map. Even worse, for $\mu = \mathcal{L}^1\lfloor_{[0,1]}$ and $\nu = \frac{1}{2}\mathcal{L}^1\lfloor_{[0,2]}$ there exists an optimal (and hence 1-cyclically monotone) transport plan which is not induced by a transport map; see [2, p. 125].

We now state the main result of this section and refer the reader to section 5.4 for further comments.

THEOREM 5.5. *Assume that $\mu \ll \mathcal{L}^d$, and let $\gamma \in \Pi(\mu, \nu)$ be an infinitely cyclically monotone transport plan. Then there exists a Borel transport map $T : \Omega \to \Omega$ such that $\gamma = (id \times T)_{\#}\mu$.*

Proof. By Proposition 2.1 in [1], it is sufficient to prove that γ is concentrated on a γ-measurable graph. In view of Lemma 5.2, it is then sufficient to prove that $R(\gamma)$ is included in a graph, or more generally that if (x_0, y_0) and (x_0, y_0') are both γ-regular points, then $y_0 = y_0'$.

We divide the proof into two parts and first show that $|x_0 - y_0| = |x_0 - y_0'|$. Arguing by contradiction, we assume that $|x_0 - y_0| < |x_0 - y_0'|$ and suppose for the time being that $x_0 \neq y_0$. Let $\xi' = \frac{y_0' - x_0}{|y_0' - x_0|}$, $0 < \varepsilon < |x_0 - y_0|$, and $0 < r < |y_0' - x_0| - |y_0 - x_0|$. We claim that for $\delta := 1 - \frac{|x_0 - y_0|}{|x_0 - y_0'|}$ one has

$$\max\{|x - y_0'|, |x_0 - y|\} < \max\{|x - y|, |x_0 - y_0'|\} \tag{5.2}$$

for any (x, y) such that $x \in C(x_0, \xi', \delta) \cap B(x_0, \varepsilon) \setminus B(x_0, \frac{1}{2}\varepsilon)$ and $y \in B(y_0, r)$. Indeed, take (x, y) as above; it then follows from the choice of r that $|x_0 - y| < |x_0 - y_0'|$, while on the other hand

$$\begin{aligned} |x - y_0'|^2 &= |x - x_0| \left(|x - x_0| - 2\frac{x - x_0}{|x - x_0|} \cdot (y_0' - x_0) \right) + |x_0 - y_0'|^2 \\ &\leq |x - x_0| \left(|x_0 - y_0| - 2(1 - \delta)|x_0 - y_0'| \right) + |x_0 - y_0'|^2 \;<\; |x_0 - y_0'|^2. \end{aligned}$$

This proves the claim. We now infer from Proposition 5.4 that the set of points $x \in \gamma^{-1}(B(y_0,r))$ such that $x \in C(x_0,\xi',\delta) \cap B(x_0,\varepsilon) \setminus B(x_0,\frac{1}{2}\varepsilon)$ has positive measure when ε is small enough. In particular, this set is nonempty for small ε, and (5.2) then clearly contradicts the infinite cyclical monotonicity of γ. As a consequence, $|x_0 - y_0| = |x_0 - y_0'|$ in the case $x_0 \neq y_0$.

If $x_0 = y_0$, we repeat the argument above with the choices $0 < \varepsilon < \frac{1}{4}|x_0 - y_0'|$, $0 < r < \frac{1}{4}|y_0' - x_0|$, and $\delta = \frac{1}{2}$. Then for any (x,y) such that $x \in C(x_0,\xi',\delta) \cap B(x_0,\varepsilon) \setminus B(x_0,\frac{1}{2}\varepsilon)$ and $y \in B(y_0,r)$, we clearly have $|x_0 - y| < |x_0 - y_0'|$, and the other inequality $|x - y_0'|^2 < |x_0 - y_0'|^2$ follows as above.

We now prove by contradiction that $y_0 = y_0'$. Note that since we already know that $|x_0 - y_0| = |x_0 - y_0'|$, we may assume that $|x_0 - y_0| = |x_0 - y_0'| > 0$. If $y_0 \neq y_0'$, we can find $\xi \in \mathbb{R}^d$ such that

$$\xi \cdot \frac{x_0 - y_0}{|x_0 - y_0|} \, < \, 0 \quad \text{and} \quad \xi \cdot \frac{x_0 - y_0'}{|x_0 - y_0'|} \, > \, 0.$$

Next we choose $r > 0$ such that

$$\sup \left\{ \xi \cdot \frac{x_0 - y}{|x_0 - y|} \, : \, y \in B(y_0,r) \right\} \, < \, 0$$

and $\delta > 0$ such that

$$\alpha \, := \, \inf \left\{ \frac{x_0 - x}{|x_0 - x|} \cdot \frac{x_0 - y_0'}{|x_0 - y_0'|} \, : \, x \in C(x_0,\xi,\delta) \right\} \, > \, 0 \tag{5.3}$$

as well as

$$\sup \left\{ \frac{x_0 - x}{|x_0 - x|} \cdot \frac{x_0 - y}{|x_0 - y|} \, : \, x \in C(x_0,\xi,\delta),\, y \in B(y_0,r) \right\} \, < \, 0. \tag{5.4}$$

We now claim that for $\varepsilon > 0$ small enough, (5.2) holds for any (x,y) such that $x \in C(x_0,\xi,\delta) \cap B(x_0,\varepsilon) \setminus B(x_0,\frac{1}{2}\varepsilon)$ and $y \in B(y_0,r)$. Notice that this claim concludes the proof of $y_0 = y_0'$ modulo applying Proposition 5.4 as before. To verify that (5.2) holds, we first notice that (5.4) implies that

$$\text{for all } x \in C(x_0,\xi,\delta),\, y \in B(y_0,r), \qquad |x_0 - y| \, < \, |x - y| \tag{5.5}$$

since $|x-y|^2 = |x-x_0|^2 - 2(x_0 - x)\cdot(x_0 - y) + |x_0 - y|^2$. We can also infer from (5.3) that

$$|x - y_0'|^2 \, \leq \, |x - x_0| \, (|x_0 - x| - 2\alpha |x_0 - y_0'|) + |x_0 - y_0'|^2$$

for any $x \in C(x_0,\xi,\delta)$. It follows that

$$\text{for all } x \in C(x_0,\xi,\delta) \cap B(x_0,\varepsilon), \qquad |x - y_0'| \, < \, |x_0 - y_0'| \tag{5.6}$$

whenever $0 < \varepsilon < 2\alpha|x_0 - y_0'|$. We then get (5.2) from (5.5) and (5.6), which concludes the proof. □

5.3. Uniqueness of the infinitely cyclically monotone transport map. We now consider the question of the uniqueness of the infinitely cyclically monotone transport map obtained in the preceding section. We recall that when (5.1) holds and $p \in]1,\infty[$, problem (P_p) admits a unique (up to a μ negligible set) p-cyclically

monotone transport map (see, for example, [20] and section 5.4). Notice that in contrast this result does not hold for $p = 1$, not even under the stronger hypothesis that $\mu \ll \mathcal{L}^d$, as shown by Example 1.3 in [1]: when $\mu = \mathcal{L}^1\lfloor_{[0,1]}$ and $\nu = \mathcal{L}^1\lfloor_{[1,2]}$, both transport maps $t \mapsto t+1$ and $t \mapsto 2-t$ are optimal.

In the case of problem (P_∞), the question of uniqueness is largely open. At the moment, we have only the following partial result stating the uniqueness of the infinitely cyclically monotone transport map under the hypothesis that ν is purely atomic with finite support.

THEOREM 5.6. *Suppose that $\mu \ll \mathcal{L}^d$ and $\nu = \sum_{i=0}^k a_i\,\delta_{y_i}$ for some $(y_i)_{0\le i\le k} \subset \Omega$ and positive numbers $a_0, \dots, a_k$. Then there exists a unique (up to a μ-negligible set) infinitely cyclically monotone Borel transport map T from μ to ν.*

Proof. Suppose that there are two distinct infinitely cyclically monotone Borel transport maps T and $\tilde{T}$, and let us introduce the sets

$$U_j^i = T^{-1}(y_j) \cap \tilde{T}^{-1}(y_i).$$

We first claim that it is possible to define a sequence of integers $(i(p))_{p\ge 0}$ such that

$$\text{for all} \quad p \ge 0, \qquad i(p) \neq i(p+1) \quad \text{and} \quad \mu\left(U_{i(p)}^{i(p+1)}\right) > 0.$$

Indeed, the fact that the two transport maps are distinct means that it is possible to choose two indices $i(0) \neq i(1)$ such that $\mu\left(U_{i(0)}^{i(1)}\right) > 0$. Next we notice that since $\tilde{T}$ maps μ to ν,

$$\nu(\{y_{i(1)}\}) \;=\; \sum_{p=0}^{k} \mu\left(U_p^{i(1)}\right) \;\ge\; \mu\left(U_{i(0)}^{i(1)}\right) + \mu\left(U_{i(1)}^{i(1)}\right) \;>\; \mu\left(U_{i(1)}^{i(1)}\right).$$

Since T also maps μ to ν, we infer from the above inequality that there exists $p \neq i(1)$ such that $\mu\left(U_{i(1)}^{p}\right) > 0$: we then set $i(2) = p$ and start again from $U_{i(1)}^{i(2)}$. By repeating the above argument we can build recursively the sequence $(i(p))_{p\ge 0}$ with the desired properties.

Since the sequence $(i(p))_{p\ge 0}$ takes its values in the finite set $\{0, \dots, k\}$, we may assume that there exists some $m \ge 2$ such that $i(m) = i(0)$. For any $p \in \{0, \dots, m-1\}$, the set $U_{i(p)}^{i(p+1)}$ has nonzero Lebesgue measure, so we may choose a Lebesgue point x_p of $U_{i(p)}^{i(p+1)}$ for which $|y_{i(p+1)} - x_p| \neq |y_{i(p)} - x_p|$ and then set $x_m = x_0$. By definition,

$$T(x_p) = y_{i(p)} \quad \text{and} \quad \tilde{T}(x_p) = y_{i(p+1)} \qquad \text{for all } p \in \{0, \dots, m-1\}.$$

Since T and $\tilde{T}$ are infinitely cyclically monotone, we have

$$\max_{0\le p\le m-1} |y_{i(p)} - x_p| \;\le\; \max_{0\le p\le m-1} |y_{i(p+1)} - x_p| \;\le\; \max_{0\le p\le m-1} |y_{i(p)} - x_p|,$$

so that

$$\max_{0\le p\le m-1} |y_{i(p)} - x_p| = \max_{0\le p\le m-1} |y_{i(p+1)} - x_p|. \tag{5.7}$$

Then let $I := \{q : |y_{i(q)} - x_q| = \max_{0\le p\le m-1} |y_{i(p)} - x_p|\}$. We infer from (5.7) and the choice of the points x_p that for any $q \in I$ one has $|y_{i(q)} - x_q| > |y_{i(q+1)} - x_q|$. Since x_q is a Lebesgue point of $U_{i(q)}^{i(q+1)}$, there exists $\tilde{x} \in U_{i(q)}^{i(q+1)}$ arbitrarily close to x_q for which

$$|y_{i(q)} - \tilde{x}| \ > \ |y_{i(q)} - x_q|,$$

and thus we can choose $\tilde{x} \in U_{i(q)}^{i(q+1)}$ such that

$$|y_{i(q)} - \tilde{x}| > \max\{|y_{i(q+1)} - \tilde{x}|, |y_{i(q)} - x_q|\}. \tag{5.8}$$

We now set $\tilde{x}_q := \tilde{x}$ as well as $\tilde{x}_p := x_p$ for $p \neq q$ and notice that (5.7) should in fact also hold for this new choice of elements $\tilde{x}_p$ in $U_{i(p)}^{i(p+1)}$. This leads to a contradiction since we can infer from the definition of I, the choice of q, and (5.8) that

$$\begin{aligned}\max_{0\le p\le m-1} |y_{i(p)} - \tilde{x}_p| &= |y_{i(q)} - \tilde{x}| \\ &> \max\{|y_{i(q+1)} - \tilde{x}|, |y_{i(q)} - x_q|\} \ \ge \ \max_{0\le p\le m-1} |y_{i(p+1)} - \tilde{x}_p|.\end{aligned}$$

This is in contradiction with (5.7) and concludes the proof of the theorem. □

Remark 5.7. The construction of the sequence $(x_p)_{0\le p\le m}$ is close to that proposed in the proof of Theorem 3.4, but it is easier since we do not need that it loops at x_0. Indeed, in the above proof we do not really need that $x_m = x_0$, and we assume this only for convenience of notation, while in the course of the proof of Theorem 3.4 we intended to use Claim 1 and then had to start from the special x_0 found there.

At the moment we are not able to generalize this result to the case where ν is any probability measure on Ω. On the other hand, it is clear that the above uniqueness theorem requires that μ does not concentrate, as the following example shows.

Example 5.8. Assume that $\mu := \mathcal{H}^1\lfloor_{[0,1]\times\{0\}}$, while $\nu := \frac{1}{2}\left(\delta_{(0,-1)} + \delta_{(0,1)}\right)$. Then any transport map T (i.e., any μ-measurable function for which $T_\#\mu = \nu$) is an infinitely cyclically monotone optimal transport map from μ to ν.

5.4. Comments around duality. For $1 \le p < \infty$, the mass transport problem (P_p) may be rewritten as

$$(P_p) \qquad W_p^p(\mu,\nu) \ = \ \inf\left\{\mathcal{C}_p^p(\lambda) \ = \ \int_{\Omega^2} |y-x|^p d\lambda(x,y) \ : \ \lambda \in \Pi(\mu,\nu)\right\}.$$

In this form, the objective functional $\lambda \mapsto \mathcal{C}_p^p(\lambda)$ is linear over the compact convex set $\Pi(\mu,\nu)$, and it is then quite natural to associate with (P_p) its dual problem

$$(D_p) \qquad \sup\left\{\int_\Omega \phi(x)d\mu(x) + \int_\Omega \psi(y)d\nu(y) \ : \ \phi(x)+\psi(y) \le |y-x|^p\right\},$$

where $\phi \in L^1(d\mu)$, $\psi \in L^1(d\nu)$, and the constraint is required to hold for μ a.e. x and ν a.e. y. Due to the regularity of the integrand $c_p(x,y) := |y-x|^p$, the supremum of (D_p) is achieved for a couple (φ, φ^{c_p}) where the Kantorovich potential φ is continuous and c_p-concave. We refer the reader, for example, to section 3 of [2], Part I of [20], section 3.3 of [26], or section 2.4 of [32] for more on the related concepts and results.

The Kantorovich dual problem (D_p) appears to be a fundamental tool in understanding and solving the problems of the characterization, existence, and uniqueness

for an optimal transport map for (P_p). For example, the notion of p-cyclical monotonicity (3.1) naturally appears via the equality $\sup(D_p) = \inf(P_p)$; see, e.g., the proof of Theorem 3.2 in [2] (an alternative and direct proofis, for example, that of Theorem 2.3 in [20]). Moreover, a key point in the construction of an optimal transport map $T_p : \mathrm{supp}(\mu) \to \mathrm{supp}(\nu)$ for (P_p) is to identify the directions of transportation (known as *transport rays* for $p = 1$), that is, to associate with μ a.e. $x \in \mathrm{supp}(\mu)$ the direction $\frac{T_p(x)-x}{|T_p(x)-x|}$ to which the mass present at x is transferred. It is now well understood that this direction may be obtained as the adequate c_p-gradient of an optimal Kantorovich potential φ (see, e.g., [20, 19, 27] or section 2.4 of [32]). In fact, the definition as well as the regularity properties of the transport rays are deeply linked with the fact that the support of an optimal transport plan γ_p for (P_p) is p-cyclically monotone and thus inherits good properties from being included in the subdifferential of a c_p-concave function (which, in turn, turns out to be a Kantorovich potential; see, e.g., section 2.4 of [32]).

In light of the preceding discussion, it is natural to try to develop a duality theory for the problem (P_∞) as well. We hereafter informally discuss this issue.

First, in view of the Example 3.3, it does not seem realistic that one could obtain useful information from an intrinsic approach. Indeed, the optimal transport plan λ proposed in Example 3.3 is not induced by any transport map, so we cannot expect that a dual problem directly associated with (P_∞) via some general construction gives information on the geometry of the optimal transport plans. On the other hand, notice that Theorems 5.5 and 5.6 do indicate that there exists a unique infinitely cyclically monotone optimal transport map for the particular problem of Example 3.3.

In view of the results of the two preceding sections, and since the notion of infinitely cyclically monotone plan was at first obtained via a limiting argument, one is led to study the asymptotic behavior of the family of dual problems (D_p) as $p \to \infty$. But as mentioned above, (D_p) is not directly related to (P_p) but to a reformulation of (P_p) which requires taking the p-power of the objective functional $\mathcal{C}_p$. As a consequence, one should in fact take the $\frac{1}{p}$-power of the objective functional of (D_p) and then study the limiting problem as $p \to \infty$; unfortunately, our research in this direction has been unfruitful up to now. Finally, since the functional $\mathcal{C}_p$ is not convex in λ, the convex duality theory does not apply directly to (P_p). But one may wonder whether it is possible to overcome this difficulty and associate with (P_p) a dual problem with a structure similar to that of (D_p): this is a quite involved question known as Dudley's problem (see, e.g., equation (1.1.10) and Remark 2.6.2 in [26]), and it is out of the scope of the present study.

Despite the above difficulties, we believe that developing a duality theory for the problem (P_∞) is an important issue since it would yield a deeper understanding of the problem of the existence and uniqueness for a particular optimal transport map. Further explorations in this direction could follow the methods of [7, 16, 18].

Acknowledgments. The authors would like to thank Guy Bouchitté for proposing the problem and for several fruitful discussions. The authors also thank Jouni Parkkonen for bringing [17] to their attention.

REFERENCES

[1] L. Ambrosio, *Lecture notes on optimal transportation problems*, in Mathematical Aspects of Evolving Interfaces (Funchal, 2000), Lecture Notes in Math. 1812, Springer, Berlin, 2003, pp. 1–52.

[2] L. Ambrosio and A. Pratelli, *Existence and stability results in the L^1 theory of optimal transportation*, in Optimal Transportation and Applications (Martina Franca, 2001), Lecture Notes in Math. 1813, Springer, Berlin, 2003, pp. 123–160.

[3] G. Aronsson, *Minimization problems for the functional* $\sup_x F(x, f(x), f'(x))$, Ark. Mat., 6 (1965), pp. 33–53.

[4] G. Aronsson, *Extension of functions satisfying Lipschitz conditions*, Ark. Mat., 6 (1967), pp. 551–561.

[5] G. Aronsson, M. G. Crandall, and P. Juutinen, *A tour of the theory of absolutely minimizing functions*, Bull. Amer. Math. Soc. (N.S.), 41 (2004), pp. 439–505.

[6] E. N. Barron, R. R. Jensen, and C. Y. Wang, *The Euler equation and absolute minimizers of L^∞ functionals*, Arch. Ration. Mech. Anal., 157 (2001), pp. 255–283.

[7] J. D. Benamou and Y. Brenier, *A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem*, Numer. Math., 84 (2000), pp. 375–393.

[8] M. Bernot, V. Caselles, and J. M. Morel, *Are there infinite irrigation trees?*, J. Math. Fluid Mech., 8 (2006), pp. 311–332.

[9] G. Bouchitté, C. Jimenez, and M. Rajesh, *A new L^∞ estimate in optimal mass transport*, Proc. Amer. Math. Soc., 135 (2007), pp. 3525–3535.

[10] A. Brancolini, G. Buttazzo, and F. Santambrogio, *Path functionals over Wasserstein spaces*, J. Eur. Math. Soc. (JEMS), 8 (2006), pp. 415–434.

[11] G. Buttazzo, A. Pratelli, and E. Stepanov, *Optimal pricing policies for public transportation networks*, SIAM J. Optim., 16 (2006), pp. 826–853.

[12] J. A. Carrillo, M. Di Francesco, and C. Lattanzio, *Contractivity and asymptotics in Wasserstein metrics for viscous nonlinear scalar conservation laws*, Boll. Unione Mat. Ital. Sez. B Artic. Ric. Mat. (8), 10 (2007), pp. 277–292.

[13] J. A. Carrillo, M. P. Gualdani, and G. Toscani, *Finite speed of propagation in porous media by mass transportation methods*, C. R. Math. Acad. Sci. Paris, 338 (2004), pp. 815–818.

[14] T. Champion and L. De Pascale, *A principle of comparison with distance functions for absolute minimizers*, J. Convex Anal., 14 (2007), pp. 515–541.

[15] T. Champion and L. De Pascale, in preparation.

[16] L. De Pascale and A. Pratelli, *Regularity properties for Monge transport density and for solutions of some shape optimization problem*, Calc. Var. Partial Differential Equations, 14 (2002), pp. 249–274.

[17] L. L. Dines, *On positive solutions of a system of linear equations*, Ann. of Math. (2), 28 (1926/27), pp. 386–392.

[18] L. C. Evans, *Partial differential equations and Monge–Kantorovich mass transfer*, in Current Developments in Mathematics, 1997, International Press, Boston, MA, 1999, pp. 65–126.

[19] L. C. Evans and W. Gangbo, *Differential equations methods for the Monge–Kantorovich mass transfer problem*, Mem. Amer. Math. Soc., 137 (1999).

[20] W. Gangbo and R. J. McCann, *The geometry of optimal transportation*, Acta Math., 177 (1996), pp. 113–161.

[21] M. Hauray and P.-E. Jabin, *N-particles approximation of the Vlasov equations with singular potential*, Arch. Ration. Mech. Anal., 183 (2007), pp. 489–524.

[22] M. Hauray, *private communication*.

[23] F. Maddalena, J. M. Morel, and S. Solimini, *A variational model of irrigation patterns*, Interfaces Free Bound., 5 (2003), pp. 391–415.

[24] R. J. McCann, *Stable rotating binary stars and fluid in a tube*, Houston J. Math., 32 (2006), pp. 603–631.

[25] A. Pratelli, *On the sufficiency of c-cyclical monotonicity for optimality of transport plans*, Math. Z., 258 (2008), pp. 677–690.

[26] S. Rachev and L. Rüschendorf, *Mass Transportation Problems. Vol.* I. *Theory*, Springer-Verlag, New York, 1998.

[27] L. Rüschendorf, *Optimal solutions of multivariate coupling problems*, Appl. Math. (Warsaw), 23 (1995), pp. 325–338.

[28] L. Rüschendorf, *On c-optimal random variables*, Statist. Probab. Lett., 27 (1996), pp. 267–270.

[29] F. Santambrogio, *Variational Problems in Transport Theory with Mass Concentration*, Ph.D. thesis, Scuola Normale Superiore, Pisa, Italy, 2007 (available at http://cvgmt.sns.it).

[30] O. Savin, *C^1 regularity for infinity harmonic functions in two dimensions*, Arch. Ration. Mech. Anal., 176 (2005), pp. 351–361.

[31] W. Schachermayer and J. Teichmann, *Characterization of optimal transport plans for the Monge-Kantorovich-problem*, Proc. Amer. Math. Soc., to appear.

[32] C. Villani, *Topics in Optimal Transportation*, Grad. Stud. Math. 58, AMS, Providence, RI, 2003.
[33] C. Villani, *Optimal Transport, Old and New*, http://www.umpa.ens-lyon.fr/~cvillani/surveys.html#oldnew.
[34] Q. Xia, *Optimal paths related to transport problems*, Commun. Contemp. Math., 5 (2003), pp. 251–279.

SIAM J. MATH. ANAL.
Vol. 40, No. 1, pp. 21–43

REGULARITY UP TO THE BOUNDARY FOR NONLINEAR ELLIPTIC SYSTEMS ARISING IN TIME-INCREMENTAL INFINITESIMAL ELASTO-PLASTICITY*

PATRIZIO NEFF† AND DOROTHEE KNEES‡

Abstract. In this paper we investigate the question of higher regularity up to the boundary for quasi-linear elliptic systems which originate from the time discretization of models from infinitesimal elasto-plasticity. Our main focus lies on an elasto-plastic Cosserat model. More specifically we show that the time discretization renders H^2-regularity of the displacement and H^1-regularity for the symmetric plastic strain ε_p up to the boundary, provided that the plastic strain of the previous time step is in H^1 as well. This result contrasts with classical Hencky and Prandtl–Reuss formulations where it is known not to hold due to the occurrence of slip lines and shear bands. Similar regularity statements are obtained for other regularizations of ideal plasticity such as viscosity or isotropic hardening. In the first part we recall the time continuous Cosserat elasto-plasticity problem, provide the update functional for one time step, and show various preliminary results for the update functional (Legendre–Hadamard/monotonicity). Using nonstandard difference quotient techniques we are able to show the higher global regularity. Higher regularity is crucial for qualitative statements of finite element convergence. As a result we may obtain estimates linear in the mesh-width h in error estimates.

Key words. polar materials, perfect plasticity, higher global regularity, quasi-linear elliptic systems, error estimates, time increments

AMS subject classifications. 35B65, 49N60, 74A35, 74C05, 74G40

DOI. 10.1137/070695824

1. Introduction.

1.1. Plasticity and Cosserat models. This article addresses the regularity question for time-incremental formulations of *geometrically linear* elasto-plasticity. As a representative model problem we consider generalized continua of *Cosserat-micropolar* type.

The basic difference of a Cosserat model as compared with classical continuum models is the appearance of a nonsymmetric stress tensor which is augmented by a generalized balance of angular momentum equation allowing one to model interaction of particles not only by surface forces (classical Cauchy continuum) but also through surface couples (Cosserat continuum). General continuum models involving *independent rotations* as additional degrees of freedom were first introduced by the Cosserat brothers in [15]. For an introduction to the theory of Cosserat and micropolar models we refer the reader to the introduction in [49, 43, 45, 44, 48]; see also [22, 9].

There are a great many proposals for extensions of the elastic Cosserat framework to infinitesimal elasto-plasticity. We mention only [17, 19, 31, 55]. Recently the finite-strain formulation has been put into focus; see, e.g., [56, 62, 23] and the references therein.

*Received by the editors June 29, 2007; accepted for publication (in revised form) November 16, 2007; published electronically March 26, 2008.
http://www.siam.org/journals/sima/40-1/69582.html
†Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstrasse 7, 64289 Darmstadt, Germany (neff@mathematik.tu-darmstadt.de).
‡Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, 10117 Berlin, Germany (knees@wias-berlin.de).

The first author has also proposed an elasto-plastic Cosserat model [45, 44] in a finite-strain framework. A geometrical linearization of this model has been investigated in [46, 48] and is shown to be well-posed also in the rate-independent limit for both quasi-static and dynamic processes.

When it comes to numerically solving problems in elasto-plasticity, then it is common practice to discretize the time evolution in the flow rule for the plastic variable with a backward Euler method and to consider a sequence of discrete-in-time problems [50]. Provided that the elasto-plastic model has certain variational features (hyperelasticity of the elastic response, associative flow rule) it is possible to recast the problem for one time step (called the update problem in the following) itself into a variational framework: the updated displacement is obtained as a minimizer of some update functional; see, e.g., [61, 60, 2, 66, 67]. This line of thought can be nicely extended to finite-strain multiplicative plasticity; see [52, 37, 36, 38] and the references therein. In the geometrically linear setting the resulting variational update problem usually has the form of a quasi-linear elliptic system whose corresponding energy has only linear growth (in case of perfect plasticity).

For qualitative statements on the rate of convergence of finite element methods it is necessary to know precisely the regularity of the function to be approximated. This then is the question on the regularity of the solution of the quasi-linear elliptic system constituting the update problem.

As far as classical rate-independent (perfect) elasto-plasticity is concerned we remark that global existence for the displacement has been shown only in a very weak, measure-valued sense, while the stresses could be shown to remain in $L^2(\Omega)$, provided that a safe load condition is assumed. For these results we refer the reader, for example, to [3, 13, 64]. If hardening or viscosity is added, then global H^1-displacement solutions are found (see, e.g., [1, 12, 11]), already without safe load assumption. A complete theory for the classical rate-independent case remains, however, elusive; see also the remarks in [13].

Since classical perfect plasticity is, therefore, notoriously ill-posed (the updated displacements have derivatives only in a measure-valued sense) we focus in our investigation of higher regularity on certain modified update functionals which might allow for more regular updates. The Cosserat elasto-plastic model in [46] is our basic candidate. Based on this time-continuous model we investigate the time-incremental formulation and study the global regularity of minimizers of the corresponding update functional. In [49] this time-incremental formulation is the basis of a finite element approximation.

Our focus on Cosserat models is justified by the fact that the Cosserat-type models are today increasingly advocated as a means to regularize the pathological mesh size dependence of localization computations where shear failure mechanisms [14, 40, 4] play a dominant role; for applications in plasticity, see the nonexhaustive list [31, 19, 55, 17].

1.2. Outline of this contribution. Our contribution is organized as follows: first, we recall the time-continuous geometrically linear elasto-plastic Cosserat model as introduced in [45, 44] and investigated mathematically in [46, 48, 47].

Referring to the development in [49] we provide in section 2 the corresponding time-discretized formulation based on a fully implicit backward Euler discretization of the plastic flow rule in time. It is shown in [49] that at each time step t_n the updated displacement field u^n and the updated "Cosserat-microrotation-matrix" A^n can equivalently be obtained from a convex minimization problem which involves only

data from the previous time step. The plastic strain ε_p^n is then derived from A^n and u^n via a simple update formula. Furthermore, in [49] it has been shown that the update problem admits unique minimizers $u^n \in H^1(\Omega, \mathbb{R}^3)$, $A^n \in H^1(\Omega, \mathfrak{so}(3))$, and $\varepsilon_p^n \in L^2(\Omega, \mathrm{Sym}(3))$, provided that the data coming from the previous time step are smooth enough. In order to quantify the rate of convergence of corresponding finite element methods for the update problem we investigate the regularity of the displacements u^n by studying the corresponding weak Euler–Lagrange equations. These equations form a quasi-linear elliptic system of partial differential equations. The main result of this paper is Theorem 5.2 in section 5, where we formulate a global regularity result for weak solutions of a rather general class of quasi-linear elliptic systems of second order. The time-incremental Cosserat plasticity formulation satisfies all the necessary assumptions of the regularity result, which allows us to show higher regularity to the extent that for all $n \in \mathbb{N} : u^n \in H^2(\Omega, \mathbb{R}^3)$, $A^n \in H^2(\Omega, \mathfrak{so}(3))$, and $\varepsilon_p^n \in H^1(\Omega, \mathrm{Sym}(3))$ if pure Dirichlet data are assumed. Let us remark that it remains an open problem whether a similar regularity result is also valid for the time-continuous Cosserat model or other regularized time-continuous plasticity formulations.

The general quasi-linear elliptic systems, which we study in section 5, are of the following type: Find $u \in H_0^1(\Omega)$ such that for every $v \in H_0^1(\Omega)$

$$\int_\Omega \langle \mathcal{M}(x, \nabla u(x), z(x)), \nabla v(x) \rangle \,\mathrm{dx} = \int_\Omega \langle f, v \rangle \,\mathrm{dx}.$$

Here, $z \in L^2(\Omega, \mathbb{R}^N)$ and $f \in L^2(\Omega, \mathbb{R}^3)$ are the given data. For the Cosserat model, z is identified with (ε_p^n, A^n); the explicit structure of $\mathcal{M} = \mathcal{M}_C$ is given in section 2.4. It is shown that $\mathcal{M}_C$ is rank-one monotone in ∇u and Lipschitz continuous but not differentiable. Consequently, we assume in the general case that the function $\mathcal{M} : \Omega \times \mathbb{M}^{m\times d} \times \mathbb{R}^N \to \mathbb{M}^{m\times d}$ is Lipschitz continuous, is rank-one monotone in ∇u, and induces a Gårding inequality. The precise conditions on $\mathcal{M}$ are formulated as R1–R3 in section 5.1. Our main result is Theorem 5.2, where we prove for smooth domains that $u \in H^2(\Omega)$, provided that $z \in H^1(\Omega)$ and $f \in L^2(\Omega)$. We emphasize that we do not need the differentiability of $\mathcal{M}$ and that we require $\mathcal{M}$ to be rank-one monotone, only, instead of uniformly or strongly monotone. A further new aspect compared to systems studied in the literature is the presence of the function z in the definition of the differential operator.

Let us give a short overview on global regularity results for quasi-linear second order systems. Systems with quadratic growth or, more generally, with p growth are studied by several authors. We mention here the books [42, 39, 6] and the paper [53], where global regularity results for systems of the type

$$\mathrm{Div}\, \mathcal{M}(x, \nabla u(x)) + f(x) = 0, \quad u\big|_{\partial\Omega} = g_{\mathrm{D}},$$

are shown for smooth domains assuming that $\mathcal{M}$ is differentiable and strongly monotone. Further results for Lipschitz domains were obtained in [21, 20, 57], again assuming that $\mathcal{M}$ is strongly monotone (or uniformly monotone if $p \neq 2$), that it is differentiable, and that there is a function W such that $\mathcal{M} = DW$. These results are proved with a difference quotient technique which relies on the standard finite differences $\delta_h u(x) := u(x + h) - u(x)$.

In [16] the authors study systems where $\mathcal{M}(x, u, \nabla u) = B(x)\nabla u + h(x, u, \nabla u)$. The main assumption in [16] is that B is uniformly positive definite, h is Hölder-continuous with respect to ∇u, and $h(x, u, \cdot)$ is uniformly monotone in zero. They prove that the gradient of solutions belongs locally to certain Campanato–Spanne

spaces. With our main result we can treat the case where h does not depend on u, where it is Lipschitz continuous and monotone but not necessarily uniformly monotone, and where B induces a rank-one positive quadratic form. We obtain $u \in H^2(\Omega)$ globally.

In [58] a nonlinear elliptic system is studied which is more related to our Cosserat model. There, $\mathcal{M}$ is chosen as $\mathcal{M}(\nabla u) = \frac{h(|\varepsilon(u)|)}{|\varepsilon(u)|}\,\varepsilon(u)$, where $\varepsilon(u)$ is the linearized strain tensor, and it is assumed that h is differentiable except for a finite number of points and that h is strongly monotone. It is shown for smooth domains that $u \in H^2(\Omega)$ by investigating the regularity of functions u_δ with $\mathrm{Div}(\delta\varepsilon(u_\delta) + \mathcal{M}(\varepsilon(u_\delta))) + f = 0$ for $\delta \searrow 0$. The results for u_δ are obtained with standard finite differences. Further results for related models were obtained in [54, 7]. Let us remark that the quasi-linear system we are interested in contains the above described systems as special cases (if $p = 2$) and that our main result is not covered by the above references. The local and global regularity of the stress fields of a class of degenerated quasi-linear elliptic systems is investigated in the papers [10, 33].

Note that higher regularity is not known to hold for the displacements of the classical limit of our formulation, which is the classical time-incremental Prandtl–Reuss model. In the first update step this model in turn is nothing else than the total deformation Hencky plasticity model. The Hencky model does not allow for regular displacements. Here, it is known that the displacement $u \in L^{\frac{3}{2}}(\Omega, \mathbb{R}^3)$ (see, e.g., [6, p. 423]), while the classical symmetric stresses satisfy $\sigma \in H^1_{\mathrm{loc}}(\Omega, \mathrm{Sym}(3)) \cap H^{\frac{1}{2}-\delta}(\Omega)$ for every $\delta > 0$ if the data are sufficiently regular and if Ω is a Lipschitz domain. See [59, 24, 5, 51, 18] for the local and [32] for the global result.

The proof of our own regularity result is split into the three classical steps. In the first step we investigate the tangential regularity of weak solutions in the case where Ω is a cube. Since we assumed rank-one monotonicity, only, we cannot apply the standard difference quotient technique in this step. Instead, we use finite differences which are based on inner variations: $\triangle_h u(x) = u(\tau_h(x)) - u(x)$, where $\tau_h(x) = x + \varphi^2(x)h$ for $h \in \mathbb{R}^d$ and a cut-off function φ. This will be explained in more detail in Remark 5.5. Let us note that these nonstandard differences were recently applied by Nesenenko [51] in order to obtain higher local regularity for models from elasto-plasticity with linear hardening. In the second step we prove higher regularity in directions normal to the boundary. Due to the lack of differentiability of $\mathcal{M}$ we cannot apply the usual arguments (i.e., solving the equation for the normal derivatives) to obtain the differentiability of ∇u in the normal direction. Instead, we exploit the rank-one monotonicity of $\mathcal{M}$ in order to get more information on the missing derivative. In the final step we prove the result for arbitrary bounded $\mathcal{C}^{1,1}$-smooth domains by the usual localization procedure. The notation is found in the appendix.

2. The infinitesimal elasto-plastic Cosserat model. In this section we recall the specific isotropic infinitesimal elasto-plastic Cosserat model which has been proposed in a finite-strain setting in [44] and which was analyzed in [46]. Moreover, we derive a discrete formulation. This section does not contain new results; it serves to provide the clear definition of the problem and to introduce some of the notation.

2.1. Time-continuous infinitesimal elasto-plastic Cosserat model. The geometrically linear time continuous system in variational form with nondissipative Cosserat effects reads as follows: for given body forces $f(t) \in L^2(\Omega, \mathbb{R}^3)$ and given Dirichlet data find the *displacement* $u(t) \in H^1(\Omega, \mathbb{R}^3)$, the *skew-symmetric microrotation* $A(t) \in H^1(\Omega, \mathfrak{so}(3))$, and the *symmetric plastic strain* $\varepsilon_p(t) \in L^2(\Omega, \mathrm{Sym}(3))$

with

$$
\begin{aligned}
&\int_\Omega W(\nabla u, A, \varepsilon_p(t)) - \langle f(t), u\rangle \,\mathrm{d}x \mapsto \min \quad \text{w.r.t. } (u, A) \text{ at fixed } \varepsilon_p(t),\\
&W(\nabla u, A, \varepsilon_p) = \mu \left\|\operatorname{sym}\nabla u - \varepsilon_p\right\|^2\\
&\qquad\qquad + \mu_c \left\|\operatorname{skew}(\nabla u - A)\right\|^2 + \frac{\lambda}{2}\operatorname{tr}\left[\nabla u\right]^2 + 2\mu\, L_c^2 \left\|\nabla \operatorname{axl}(A)\right\|^2,\\
&\dot{\varepsilon_p}(t) \in \partial\chi(T_E(t)), \qquad T_E = 2\mu\,(\varepsilon - \varepsilon_p), \quad \varepsilon_p \in \operatorname{Sym}(3)\cap\mathfrak{sl}(3), \quad \varepsilon_p(0) = \varepsilon_p^0,\\
&u_{|\Gamma_D} = g_{\mathrm{D}}(t,x) - x, \quad A_{|\Gamma_D} = \operatorname{skew}(\nabla g_{\mathrm{D}}(t,x))_{|\Gamma_D}.
\end{aligned}
\tag{2.1}
$$

Here, $\Omega \subset \mathbb{R}^3$ is a bounded smooth domain and $\Gamma_D \subset \partial\Omega$ is that part of the boundary where Dirichlet data are prescribed. The parameters $\mu, \lambda > 0$ are the Lamé constants of isotropic linear elasticity, $\mu_c > 0$ is the Cosserat couple modulus, and $L_c > 0$ is an internal length parameter.[1] The classical symmetric elastic strain $\operatorname{sym}\nabla u$ is denoted by ε. The linear operator $\operatorname{axl} : \mathfrak{so}(3) \to \mathbb{R}^3$ provides the canonical identification between the Lie algebra $\mathfrak{so}(3)$ of skew-symmetric matrices and vectors in $\mathbb{R}^3$. The Lie algebra of trace-free matrices is denoted by $\mathfrak{sl}(3)$, and $\operatorname{dev} : \mathbb{M}^{3\times 3} \to \mathfrak{sl}(3)$, $\operatorname{dev} X = X - \frac{1}{3}\mathbb{1}$ is the orthogonal projection onto $\mathfrak{sl}(3)$. As regards the plastic flow rule, $\partial\chi$ is the subdifferential of a convex flow potential $\chi : \mathbb{M}^{3\times 3} \to \mathbb{R}^+$ acting on the generalized conjugate forces, i.e., the Eshelby stress tensor $T_E = -\partial_{\varepsilon_p} W(\nabla u, A, \varepsilon_p)$, where W is the free energy used in (2.1).[2]

The corresponding system of partial differential equations coupled with the flow rule is given by (note that $\|A\|^2_{\mathbb{M}^{3\times 3}} = 2\left\|\operatorname{axl}(A)\right\|^2_{\mathbb{R}^3}$ for $A \in \mathfrak{so}(3,\mathbb{R})$)

$$
\begin{aligned}
\operatorname{Div}\sigma &= -f, \quad x \in \Omega, \quad \text{balance of forces},\\
\sigma &= 2\mu\,(\varepsilon - \varepsilon_p) + 2\,\mu_c\,(\operatorname{skew}(\nabla u) - A) + \lambda \operatorname{tr}[\varepsilon]\cdot\mathbb{1},\\
-\mu\, L_c^2\,\Delta \operatorname{axl}(A) &= \mu_c \operatorname{axl}(\operatorname{skew}(\nabla u) - A), \quad \text{balance of angular momentum},\\
\dot{\varepsilon_p}(t) &\in \partial\chi(T_E), \quad T_E = 2\mu\,(\varepsilon - \varepsilon_p),\\
u_{|\Gamma_D}(t,x) &= g_{\mathrm{D}}(t,x) - x, \quad A_{|\Gamma_D} = \operatorname{skew}(\nabla g_{\mathrm{D}}(t,x))_{|\Gamma_D},\\
\sigma.\vec{n}|_{\partial\Omega\setminus\Gamma_D}(t,x) &= 0, \quad \mu\, L_c^2 \nabla \operatorname{axl}(A).\vec{n}|_{\partial\Omega\setminus\Gamma_D}(t,x) = 0,\\
\varepsilon_p(0) &\in \operatorname{Sym}(3)\cap\mathfrak{sl}(3).
\end{aligned}
$$

Note that in this model the force stresses σ need not be symmetric and that the Cosserat effects, active through the microrotations A, appear only in the balance equations but not in the plastic flow rule since T_E does not depend on A. It is worth noting that this model is intrinsically thermodynamically correct. If $\Gamma_D = \partial\Omega$, then the model admits global weak solutions with the regularity [46]:

$$
\begin{aligned}
&u \in L^\infty([0,T], H^1(\Omega,\mathbb{R}^3)), \quad A \in L^\infty([0,T], H^1(\Omega, \mathfrak{so}(3))),\\
&\varepsilon_p \in L^\infty([0,T], L^2(\Omega, \operatorname{Sym}(3)\cap\mathfrak{sl}(3))).
\end{aligned}
$$

2.2. Backward Euler time discretization of the flow rule. For a numerical treatment we consider the time discretization of the flow rule with the fully implicit backward Euler scheme. Let $0 = t_0 < t_1 < \cdots < t_N = T$ be a subdivision of the

[1] Observe that for $\mu_c = 0$ or $L_c = 0$ one recovers the classical Prandtl–Reuss formulation for the displacement u.

[2] The specification $\chi = I_K$ as indicator function of some elastic domain is not necessary at this point.

time interval $[0,T]$ with $t_j - t_{j-1} = \Delta t$. Let $f^n(x) = f(x,t_n)$, and assume that at time t_{n-1} a sufficiently regular plastic strain field $\varepsilon_p^{n-1} \in \operatorname{Sym}(3) \cap \mathfrak{sl}(3)$ is given. We want to determine the *updated displacement* $u^n \in H^1(\Omega,\mathbb{R}^3)$, the *updated skew-symmetric microrotation* $A^n \in H^1(\Omega,\mathfrak{so}(3))$ and the *updated symmetric plastic strain* $\varepsilon_p^n \in L^2(\Omega,\operatorname{Sym}(3)\cap\mathfrak{sl}(3))$ satisfying

$$
\begin{aligned}
\operatorname{Div}\sigma^n &= -f^n, \quad x\in\Omega,\\
\sigma^n &= 2\mu\,(\varepsilon^n-\varepsilon_p^n) + 2\,\mu_c\,(\operatorname{skew}(\nabla u^n) - A^n) + \lambda\operatorname{tr}[\varepsilon^n]\cdot\mathbb{1},\\
-\mu\,L_c^2\,\Delta\operatorname{axl}(A^n) &= \mu_c\operatorname{axl}(\operatorname{skew}(\nabla u^n) - A^n),\\
\frac{\varepsilon_p^n-\varepsilon_p^{n-1}}{\Delta t} &\in \partial\chi(T_E^n), \quad T_E^n = 2\mu\,(\varepsilon^n-\varepsilon_p^n),\\
u^n_{|\Gamma_D}(x) &= g_{\mathrm{D}}^n(x) - x, \quad A^n_{|\Gamma_D} = \operatorname{skew}(\nabla g_{\mathrm{D}}^n(x)),\\
\sigma^n.\vec{n}|_{\partial\Omega\setminus\Gamma_D}(x) &= 0, \quad \mu\,L_c^2\nabla\operatorname{axl}(A)^n.\vec{n}|_{\partial\Omega\setminus\Gamma_D}(x) = 0,\\
\varepsilon_p^{n-1} &\in L^2(\Omega,\operatorname{Sym}(3)\cap\mathfrak{sl}(3)).
\end{aligned}
\tag{2.2}
$$

It is possible to explicitly solve the discretized flow rule $(2.2)_4$ for ε_p^n in terms of ε_p^{n-1}, ε^n, and Δt. To see this, consider

$$
\begin{aligned}
\frac{\varepsilon_p^n-\varepsilon_p^{n-1}}{\Delta t} \in \partial\chi(2\mu\,(\varepsilon^n-\varepsilon_p^n)) \;&\Leftrightarrow\; 0 \in \partial\chi(2\mu\,(\varepsilon^n-\varepsilon_p^n)) - \frac{\varepsilon_p^n-\varepsilon_p^{n-1}}{\Delta t}\\
&\Leftrightarrow\; 0 \in \partial_{\varepsilon_p^n}\left(\mu\left\|\varepsilon_p^n-\varepsilon_p^{n-1}\right\|^2 + \Delta t\,\chi(2\mu\,(\varepsilon^n-\varepsilon_p^n))\right).
\end{aligned}
\tag{2.3}
$$

Thus we can define the local potential for the local flow rule

$$
V^{\mathrm{time}}(\varepsilon^n,\varepsilon_p^n,\varepsilon_p^{n-1},\Delta t) := \mu\left\|\varepsilon_p^n-\varepsilon_p^{n-1}\right\|^2 + \Delta t\,\chi(2\mu\,(\varepsilon^n-\varepsilon_p^n)).
\tag{2.4}
$$

It is easy to see that V^{time} is strictly convex in ε_p^n; thus V^{time} admits a unique minimizer satisfying $(2.3)_3$. Moreover, we have

$$
\begin{aligned}
V^{\mathrm{time}}(\varepsilon^n,\varepsilon_p^n,\varepsilon_p^{n-1},\Delta t) &= \mu\left\|\varepsilon_p^n-\varepsilon_p^{n-1}\right\|^2 + \Delta t\,\chi(2\mu\,(\varepsilon^n-\varepsilon_p^n))\\
&= \frac{1}{4\mu}\left\|2\mu(\varepsilon_p^n-\varepsilon^n+\varepsilon^n-\varepsilon_p^{n-1})\right\|^2 + \Delta t\,\chi(2\mu\,(\varepsilon^n-\varepsilon_p^n))\\
&= \frac{1}{4\mu}\left\|\Sigma^n-\Sigma^n_{\mathrm{trial}}\right\|^2 + \Delta t\,\chi(\Sigma^n) = \tilde{V}(\Sigma^n,\Sigma^n_{\mathrm{trial}}),
\end{aligned}
$$

where $\Sigma^n = 2\mu\,(\varepsilon^n-\varepsilon_p^n)$ and the so-called trial stresses $\Sigma^n_{\mathrm{trial}} = 2\mu\,(\varepsilon^n-\varepsilon_p^{n-1})$. Minimizing V^{time} with respect to ε_p^n is equivalent to minimizing $\tilde{V}$ with respect to Σ^n. Proceeding further, we specialize χ. Let us define the elastic domain in stress space

$$
K := \{\Sigma\in\mathbb{M}^{3\times3} \mid \|\operatorname{dev}\Sigma\| \le \sigma_{\mathrm{y}}\},
$$

with initial yield stress σ_{y}, $[\sigma_{\mathrm{y}}] = [\mathrm{MPa}]$, and corresponding indicator function

$$
I_K(\Sigma) = \begin{cases} 0, & \|\operatorname{dev}\Sigma\| \le \sigma_{\mathrm{y}},\\ \infty, & \|\operatorname{dev}\Sigma\| > \sigma_{\mathrm{y}},\end{cases}
$$

and let $\chi = I_K$. We have therefore $\partial\chi = \partial I_K$ in the sense of the subdifferential. With this choice, the unique minimizer of $\tilde{V}$ is simply characterized by

$$
\inf_{\Sigma^n\in K}\left\|\Sigma^n-\Sigma^n_{\mathrm{trial}}\right\|^2,
$$

independent of Δt. The solution is the orthogonal projection of Σ^n_{trial} onto the convex set K, denoted by

$$\Sigma^n = P_K(\Sigma^n_{\text{trial}}) \;\Rightarrow\; 2\mu\,(\varepsilon^n - \varepsilon^n_p) = P_K(2\mu\,(\varepsilon^n - \varepsilon^{n-1}_p)).$$

Reintroducing the last result into the balance of forces equation $(2.2)_1$ delivers

$$\begin{aligned} &\operatorname{Div}\sigma^n = -f^n, \quad x \in \Omega, \\ &\sigma^n = P_K(2\mu\,(\varepsilon^n - \varepsilon^{n-1}_p)) + 2\,\mu_c\,(\operatorname{skew}(\nabla u^n) - A^n) + \lambda \operatorname{tr}[\varepsilon^n]\cdot \mathbb{1}. \end{aligned} \tag{2.5}$$

This step is called *return mapping* [61, 60] in an engineering context of classical plasticity. At the given plastic strain of the previous time step ε^{n-1}_p this equation is the strong form of the update problem for the force-balance equation.

Gathering the previous development the formal problem for the update consists of determining $u^n \in H^1(\Omega, \mathbb{R}^3)$, $A^n \in H^1(\Omega, \mathfrak{so}(3))$, and $\varepsilon^n_p \in L^2(\Omega, \operatorname{Sym}(3) \cap \mathfrak{sl}(3))$ satisfying

$$\begin{aligned} &\operatorname{Div}\sigma^n = -f^n, \quad x \in \Omega, \\ &\sigma^n = P_K(2\mu\,(\varepsilon^n - \varepsilon^{n-1}_p)) + 2\,\mu_c\,(\operatorname{skew}(\nabla u^n) - A^n) + \lambda \operatorname{tr}[\varepsilon^n]\cdot \mathbb{1}, \\ &-\mu\, L_c^2\, \Delta \operatorname{axl}(A^n) = \mu_c \operatorname{axl}(\operatorname{skew}(\nabla u^n) - A^n). \end{aligned} \tag{2.6}$$

The updated plastic strain field is then given by

$$\varepsilon^n_p = \varepsilon^n - \frac{1}{2\mu}\, P_K(2\mu\,(\varepsilon^n - \varepsilon^{n-1}_p)). \tag{2.7}$$

For the precise formulation of this system we need the projection operator onto the yield surface which we recall in the following.

2.3. The projection onto the yield surface. Let K be a convex domain in stress space defined as

$$K := \left\{\Sigma \in \mathbb{M}^{3\times 3} \mid \|\operatorname{dev}\Sigma\| \le \sigma_{\mathrm{y}}\right\}. \tag{2.8}$$

The orthogonal projection $P_K : \mathbb{M}^{3\times 3} \to K$ onto this set is uniquely given by (see, e.g., [29, 30])

$$\begin{aligned} P_K(\Sigma) &= \begin{cases} \Sigma, & \Sigma \in K, \\ \Sigma - (\|\operatorname{dev}\Sigma\| - \sigma_{\mathrm{y}})\, \frac{\operatorname{dev}\Sigma}{\|\operatorname{dev}\Sigma\|}, & \Sigma \notin K \end{cases} \\ &= \begin{cases} \Sigma, & \|\operatorname{dev}\Sigma\| \le \sigma_{\mathrm{y}}, \\ \frac{1}{3}\operatorname{tr}[\Sigma]\,\mathbb{1} + \frac{\sigma_{\mathrm{y}}}{\|\operatorname{dev}\Sigma\|}\operatorname{dev}\Sigma, & \|\operatorname{dev}\Sigma\| > \sigma_{\mathrm{y}}. \end{cases} \end{aligned}$$

It is easy to see that P_K is Lipschitz continuous but not differentiable at Σ with $\|\operatorname{dev}\Sigma\| = \sigma_{\mathrm{y}}$.[3] From convex analysis it is clear that P_K represents a monotone operator which is nonexpansive. Therefore, P_K has Lipschitz constant 1. Observe also that

$$P_K(\Sigma) = \frac{1}{3}\operatorname{tr}[\Sigma]\,\mathbb{1} + P_K(\operatorname{dev}\Sigma). \tag{2.10}$$

[3] Consider the simple example $p : \mathbb{R} \to \mathbb{R}$,

$$p(x) = \begin{cases} x, & |x| \le \sigma_{\mathrm{y}}, \\ \sigma_{\mathrm{y}}\, \frac{x}{|x|}, & |x| > \sigma_{\mathrm{y}}. \end{cases} \tag{2.9}$$

For future reference we calculate also

$$\begin{aligned}\Sigma - P_K(\Sigma) &= \begin{cases} 0, & \|\operatorname{dev}\Sigma\| \le \sigma_{\rm y}, \\ \operatorname{dev}\Sigma\left(1-\frac{\sigma_{\rm y}}{\|\operatorname{dev}\Sigma\|}\right), & \|\operatorname{dev}\Sigma\| > \sigma_{\rm y} \end{cases} \\ &= [\|\operatorname{dev}\Sigma\| - \sigma_{\rm y}]_+ \frac{\operatorname{dev}\Sigma}{\|\operatorname{dev}\Sigma\|}, \qquad (2.11) \\ \|\Sigma - P_K(\Sigma)\|^2 &= [\|\operatorname{dev}\Sigma\| - \sigma_{\rm y}]_+^2 ,\end{aligned}$$

where $[x]_+ := \max\{0, x\}$.

2.4. Weak form of the reduced update problem. From now onwards we take $\Gamma_D = \partial\Omega$ and assume $g_{\rm D} = x$; i.e., the body is fixed everywhere on its boundary and subject only to body forces. This assumption allows us to confine attention to the simpler setting in $H^1_0(\Omega)$. We introduce the nonlinear mapping

$$\begin{aligned}&\mathcal{M}_C : \mathbb{M}^{3\times 3}\times \operatorname{Sym}(3)\times \mathfrak{so}(3) \to \mathbb{M}^{3\times 3}, \\ &\mathcal{M}_C(X,\varepsilon_p,A) := P_K(2\mu(\operatorname{sym}X - \varepsilon_p)) + \lambda \operatorname{tr}[X]\ \mathbb{1} + 2\mu_c(\operatorname{skew}(X) - A). \qquad (2.12)\end{aligned}$$

The weak form of the update problem (2.6) now reads as follows: for given $f^n \in L^2(\Omega,\mathbb{R}^3)$ and $\varepsilon_p^{n-1} \in L^2(\Omega,\operatorname{Sym}(3)\cap\mathfrak{sl}(3))$ find $(u^n, A^n) \in H^1_0(\Omega,\mathbb{R}^3)\times H^1_0(\Omega,\mathfrak{so}(3))$ satisfying for all $v \in H^1_0(\Omega,\mathbb{R}^3)$ and all $B \in H^1_0(\Omega,\mathfrak{so}(3))$

$$\int_\Omega \langle \mathcal{M}_C(\nabla u^n, \varepsilon_p^n, A^n), \nabla v\rangle \,\mathrm{dx} = \int_\Omega \langle f^n, v\rangle \,\mathrm{dx}, \qquad (2.13)$$

$$\mu\, L_c^2 \int_\Omega \langle DA^n, DB\rangle \,\mathrm{dx} = \mu_c \int_\Omega \langle \operatorname{skew}\nabla u^n - A^n, B\rangle \,\mathrm{dx}. \qquad (2.14)$$

The updated plastic strain field ε_p^n is then obtained by (2.7). It is shown in [49] that for every n the system (2.13)–(2.14) admits a unique weak solution $u^n \in H^1_0(\Omega,\mathbb{R}^3)$ and $A^n \in H^1_0(\Omega,\mathfrak{so}(3))$. Equation (2.13) represents the quasi-linear elliptic system for determining u^n, which will be discussed with respect to regularity. Together with $\varepsilon_p^{n-1}, \varepsilon^n \in H^1(\Omega,\operatorname{Sym}(3))$, which we will obtain from the regularity result to be proven below, using (2.7) we see that $\varepsilon_p^n \in H^1(\Omega,\operatorname{Sym}(3))$.

LEMMA 2.1 (strong Legendre–Hadamard ellipticity). *Let $\mu > 0$, $2\mu + 3\lambda > 0$, and $0 < \mu_c$. Then the matrix-valued function $\mathcal{M}_C$ is strongly rank-one monotone; i.e., there exists a constant $c^+_{LH} > 0$ such that for every $X \in \mathbb{M}^{3\times 3}$, $\varepsilon_p \in \operatorname{Sym}(3)$, $A \in \mathfrak{so}(3)$ and for all $\xi,\eta \in \mathbb{R}^3$ we have*

$$\langle \mathcal{M}_C(X + \xi\otimes\eta, \varepsilon_p, A) - \mathcal{M}_C(X,\varepsilon_p,A), \xi\otimes\eta\rangle \ge c^+_{LH}\ \|\xi\|^2\|\eta\|^2 . \qquad (2.15)$$

Proof. The projection P_K itself is monotone, and for $\mu > 0$ there is no sign change. Thus the map $X \to P_K(2\mu(\operatorname{sym}X - \varepsilon_p))$ is also monotone in X. Since (2.10) holds we have

$$\langle P_K(2\mu(\operatorname{sym}X + \xi\otimes\eta - \varepsilon_p)) - P_K(2\mu(\operatorname{sym}X - \varepsilon_p)), \xi\otimes\eta\rangle \ge \frac{2\mu}{3}\operatorname{tr}[\xi\otimes\eta]^2 .$$

For the remaining linear contribution we have

$$\begin{aligned}&\langle \lambda\operatorname{tr}[X+\xi\otimes\eta]\,\mathbb{1} + 2\mu_c\operatorname{skew}(X+\xi\otimes\eta - A) - [\lambda\operatorname{tr}[X]\,\mathbb{1} + 2\mu_c\operatorname{skew}(X-A)], \xi\otimes\eta\rangle \\ &= \lambda\operatorname{tr}[\xi\otimes\eta]^2 + 2\mu_c\|\operatorname{skew}(\xi\otimes\eta)\|^2 .\end{aligned}$$

Thus

(2.16)

$$\begin{aligned}
&\langle \mathcal{M}_C(X+\xi\otimes\eta,\varepsilon_p,A)-\mathcal{M}_C(X,\varepsilon_p,A),\xi\otimes\eta\rangle\\
&\quad\geq \tfrac{2\mu+3\lambda}{3}\operatorname{tr}[\xi\otimes\eta]^2+2\mu_c\left\|\operatorname{skew}(\xi\otimes\eta)\right\|^2=\tfrac{2\mu+3\lambda}{3}\langle\xi,\eta\rangle^2+\mu_c\left(\|\xi\|^2\|\eta\|^2-\langle\xi,\eta\rangle^2\right)\\
&\quad\text{split } \mu_c^1+\mu_c^2=\mu_c\\
&\quad=\left(\tfrac{2\mu+3\lambda}{3}-\mu_c^1\right)\langle\xi,\eta\rangle^2+\mu_c^1\|\xi\|^2\|\eta\|^2+\mu_c^2\underbrace{\left(\|\xi\|^2\|\eta\|^2-\langle\xi,\eta\rangle^2\right)}_{\geq 0}\\
&\quad\geq\left(\tfrac{2\mu+3\lambda}{3}-\mu_c^1\right)\langle\xi,\eta\rangle^2+\mu_c^1\|\xi\|^2\|\eta\|^2\geq\mu_c^1\|\xi\|^2\|\eta\|^2
\end{aligned}$$

if $0<\mu_c^1<\frac{3\lambda+2\mu}{3}$. Thus $\mathcal{M}_C$ generates a strongly Legendre–Hadamard elliptic operator with ellipticity constant $c_{LH}^+=\min(\mu_c,\frac{2\mu+3\lambda}{3})$. ∎

Obviously, $\mathcal{M}$ is Lipschitz continuous: for every $X_i\in\mathbb{M}^{3\times3}$, $P_i\in\operatorname{Sym}(3)$, $A_i\in\mathfrak{so}(3)$ we have

$$\|\mathcal{M}_C(X_1,P_1,A_1)-\mathcal{M}_C(X_2,P_2,A_2)\|\leq L_{\mathcal{M}_C}\left(\|X_1-X_2\|+\|P_1-P_2\|+\|A_1-A_2\|\right).$$

LEMMA 2.2. *Let $\mu>0$, $2\mu+3\lambda>0$, and $\mu_c>0$. The operator $\mathcal{M}_C$ generates a strongly monotone operator on $H_0^1(\Omega,\mathbb{R}^3)$; that is, there exists a constant $c_{\mathcal{M}_C}>0$ such that for every $v_1,v_2\in H_0^1(\Omega,\mathbb{R}^3)$ and for all $\varepsilon_p\in L^2(\Omega,\operatorname{Sym}(3))$ and $A\in L^2(\Omega,\mathfrak{so}(3))$ we have*

(2.17)

$$\int_\Omega\langle\mathcal{M}_C(\nabla v_1,\varepsilon_p,A)-\mathcal{M}_C(\nabla v_2,\varepsilon_p,A),\nabla v_1-\nabla v_2\rangle\,\mathrm{dx}\geq c_{\mathcal{M}_C}\|v_1-v_2\|^2_{H_0^1(\Omega,\mathbb{R}^3)}.$$

Proof. The same calculation as in the proof of Lemma 2.1 yields the estimate

$$\begin{aligned}
&\langle\mathcal{M}_C(\nabla v_1,\varepsilon_p,A)-\mathcal{M}_C(\nabla v_2,\varepsilon_p,A),\nabla v_1-\nabla v_2\rangle\\
&\quad\geq\frac{2\mu+3\lambda}{3}\operatorname{tr}[\nabla v_1-\nabla v_2]^2+2\mu_c\left\|\operatorname{skew}(\nabla v_1-\nabla v_2)\right\|^2.
\end{aligned}$$

Set $u=v_1-v_2$ and consider

$$\frac{2\mu+3\lambda}{3}\operatorname{tr}[\nabla u]^2+2\mu_c\|\operatorname{skew}\nabla u\|^2=\frac{2\mu+3\lambda}{3}|\operatorname{Div}u|^2+\mu_c\|\operatorname{curl}u\|^2. \tag{2.18}$$

The Div/Curl inequality on the space $H_0^1(\Omega)$ guarantees that there exists $C^+>0$ such that

$$\forall\,u\in H_0^1(\Omega,\mathbb{R}^3):\quad\int_\Omega|\operatorname{Div}u|^2+\|\operatorname{curl}u\|^2\,\mathrm{dx}\geq C^+\|u\|^2_{H_0^1(\Omega,\mathbb{R}^3)}; \tag{2.19}$$

see, for example, [28]. Applying this inequality to (2.18) implies finally (2.17). ∎

It is instructive to realize that although the quadratic form (2.18) is formally positive in the sense of Nečas [41] and strongly Legendre–Hadamard elliptic with constant coefficients it is impossible to extend the analysis to Dirichlet boundary conditions given only on a part of the boundary $\partial\Omega$. We observe that

$$\left\|\sqrt{\mu_c}\operatorname{skew}X+\sqrt{\tfrac{\lambda}{2\cdot3}}\operatorname{tr}[X]\,\mathbb{1}\right\|^2=\frac{\lambda}{2}\operatorname{tr}[X]^2+\mu_c\|\operatorname{skew}X\|^2. \tag{2.20}$$

Let $\widehat{\mathcal{A}}$ be the constant-coefficient first order differential operator

$$\widehat{\mathcal{A}}.\nabla u = \sqrt{\mu_c}\,\mathrm{skew}(\nabla u) + \sqrt{\tfrac{\lambda}{2\cdot 3}}\,\mathrm{tr}\,[\nabla u]\,\mathbb{1}.$$

The corresponding Fourier symbol is given as a linear operator $\mathcal{A}(\xi) : \mathbb{C}^3 \to \mathbb{C}^{3\times 3}$ with

$$\mathcal{A}(\xi).\hat{u} := \sqrt{\mu_c}\,\mathrm{skew}(\xi\otimes\hat{u}) + \sqrt{\tfrac{\lambda}{2\cdot 3}}\,\mathrm{tr}\,[\xi\otimes\hat{u}]\,\mathbb{1}. \tag{2.21}$$

From (2.20) it follows that

$$\|\mathcal{A}(\xi).\hat{u}\|^2 = \frac{\lambda}{2}\,\mathrm{tr}\,[\xi\otimes\hat{u}]^2 + \mu_c\,\|\mathrm{skew}\,\xi\otimes\hat{u}\|^2\,.$$

By algebraic completeness of the symbol $\mathcal{A}(\xi) : \mathbb{C}^3 \to \mathbb{C}^{3\times 3}$ it is meant

$$\forall\,\xi\in\mathbb{C}^3,\,\xi\neq 0: \quad \mathcal{A}(\xi).\hat{u} = 0_{\mathbb{C}^{3\times 3}} \;\Rightarrow\; \hat{u} = 0_{\mathbb{C}^3}.$$

Recall that the corresponding statement for real ξ, i.e.,

$$\forall\,\xi\in\mathbb{R}^3,\,\xi\neq 0: \quad \mathcal{A}(\xi).\hat{u} = 0_{\mathbb{R}^{3\times 3}} \;\Rightarrow\; \hat{u} = 0_{\mathbb{R}^3},$$

is a consequence of strict Legendre–Hadamard ellipticity of $\widehat{\mathcal{A}}$. If the symbol is algebraically complete, then, using the result in Nečas [41] the induced quadratic form

$$\int_\Omega \left\|\widehat{\mathcal{A}}.\nabla u\right\|^2 + \|u\|^2\,\mathrm{dx}$$

is an equivalent norm on $H^1(\Omega,\mathbb{R}^3)$. However, we proceed to show that $\mathcal{A}$ as defined in (2.21) corresponding to our quadratic form (2.18) is not algebraically complete.

Proof. To this end we write

$$\mathcal{A}(\xi).\hat{u} = 0 \;\Rightarrow\; \mathrm{tr}\,[\xi\otimes\hat{u}] = 0, \;\text{ and }\; \mathrm{skew}(\xi\otimes\hat{u}) = 0 \;\Rightarrow\; \xi = \hat{u}, \quad \mathrm{tr}\,[\xi\otimes\xi] = 0.$$

Consider for simplicity the two-dimensional case:

$$\xi = \begin{pmatrix}\alpha_1 + i\,\beta_1\\ \alpha_2 + i\,\beta_2\end{pmatrix}, \quad \xi\otimes\xi = \begin{pmatrix}\xi_1\,\xi_1 & \xi_1\,\xi_2\\ \xi_2\,\xi_1 & \xi_2\,\xi_2\end{pmatrix},$$
$$\mathrm{tr}\,[\xi\otimes\xi] = \xi_1\xi_1 + \xi_2\xi_2 = \alpha_1^2 + \alpha_2^2 - (\beta_1^2+\beta_2^2) + 2i(\alpha_1\beta_1 + \alpha_2\beta_2) = 0.$$

Choosing $\xi = (i,1)^T$ shows that $\mathrm{tr}\,[\xi\otimes\xi] = 0$, which proves the claim. ∎

Thence, the quadratic form is not algebraically complete, and this excludes the treatment of mixed boundary conditions on u in the following: we are forced to assume $\Gamma_D = \partial\Omega$. However, inhomogeneous Dirichlet conditions may be prescribed as far as the use of the Div/Curl estimate is concerned.

2.5. Variational form of the update problem. Due to the underlying variational formulation, the weak form (2.13) of the time-incremental Cosserat problem still has a variational structure. In [49] it is shown that solving (2.13)–(2.14) is equivalent to the following minimization problem: find $(u^n, A^n) \in H^1_0(\Omega,\mathbb{R}^3)\times H^1_0(\Omega,\mathfrak{so}(3))$ which minimize the functional

$$I^n_{\mathrm{incr}}(u,A) = \mathcal{E}_{\mathrm{incr}}(u,A,\varepsilon_p^{n-1}) - \int_\Omega \langle f^n, u\rangle\,\mathrm{dx} \tag{2.22}$$

in $H^1_0(\Omega,\mathbb{R}^3)\times H^1_0(\Omega,\mathfrak{so}(3))$. Here, $\mathcal{E}_{\mathrm{incr}}$ denotes the free energy of the incremental problem defined by

$$\begin{aligned}\mathcal{E}_{\mathrm{incr}}(u,A,\varepsilon_p) = \frac{1}{2\mu}\int_\Omega \Psi\big(2\mu(\mathrm{sym}(\nabla u)-\varepsilon_p)\big)\,\mathrm{dx} + \frac{\lambda}{2}\int_\Omega \mathrm{tr}\,[\nabla u]^2\,\mathrm{dx}\\ + \mu_c\int_\Omega \|\mathrm{skew}(\nabla u)-A\|^2\,\mathrm{dx} + \mu L_c^2\int_\Omega \|DA\|^2\,\mathrm{dx},\end{aligned} \tag{2.23}$$

with a potential function $\Psi:\mathbb{M}^{3\times 3}\to\mathbb{R}^+$ of the form

$$\begin{aligned}\Psi(X) &:= \begin{cases}\frac{1}{2}\|X\|^2, & \|\mathrm{dev}\,X\|\le\sigma_{\mathrm{y}},\\ \frac{1}{2}\left(\frac{1}{3}\,\mathrm{tr}\,[X]^2+2\sigma_{\mathrm{y}}\|\mathrm{dev}\,X\|-\sigma_{\mathrm{y}}^2\right), & \|\mathrm{dev}\,X\|>\sigma_{\mathrm{y}}\end{cases}\\ &= \frac{1}{2}\|X\|^2-\frac{1}{2}\left[\|\mathrm{dev}\,X\|-\sigma_{\mathrm{y}}\right]_+^2.\end{aligned} \tag{2.24}$$

Clearly, Ψ is convex but not strongly convex outside the yield surface. Moreover, it has only linear growth outside the yield surface. Note that for the first time step $n=1$ and $\varepsilon_p^0=0$, $\mu_c=0$, $L_c=0$ the functional $I^1_{\mathrm{incr}}(u,0)$ reduces to the primal plastic functional of static perfect plasticity (Hencky plasticity) [35, 63, 24, 25, 6].

Calculating the subdifferential of the convex potential shows that

$$\begin{aligned}\partial\Psi(\Sigma).H &= \begin{cases}\langle\Sigma,H\rangle, & \|\mathrm{dev}\,\Sigma\|\le\sigma_{\mathrm{y}},\\ \frac{1}{3}\,\mathrm{tr}\,[\Sigma]\,\mathrm{tr}\,[H]+\frac{\sigma_{\mathrm{y}}}{\|\mathrm{dev}\,\Sigma\|}\langle\mathrm{dev}\,\Sigma,\mathrm{dev}\,H\rangle, & \|\mathrm{dev}\,\Sigma\|>\sigma_{\mathrm{y}}\end{cases}\\ &= \langle P_K(\Sigma),H\rangle.\end{aligned} \tag{2.25}$$

Hence $\partial\Psi(\Sigma)=P_K(\Sigma)$, motivating the variational structure. The following relationship between the potential Ψ and the projection P_K is also valid:

$$\Psi(X)=\tfrac{1}{2}\|X\|^2-\tfrac{1}{2}\|X-P_K(X)\|^2.$$

For future reference the second differential of the potential Ψ can be calculated in those points where the potential is differentiable. It holds that

$$D^2_X\Psi(X).(H,H) = \begin{cases}\|H\|^2, & \|\mathrm{dev}\,X\|<\sigma_{\mathrm{y}},\\ \text{does not exist}, & \|\mathrm{dev}\,X\|=\sigma_{\mathrm{y}},\\ \frac{1}{3}\,\mathrm{tr}\,[H]^2+\sigma_{\mathrm{y}}\left(\frac{\|\mathrm{dev}\,H\|^2}{\|\mathrm{dev}\,X\|}-\frac{\langle\mathrm{dev}\,X,H\rangle^2}{\|\mathrm{dev}\,X\|^3}\right), & \|\mathrm{dev}\,X\|>\sigma_{\mathrm{y}}.\end{cases} \tag{2.26}$$

The potential Ψ is not strictly rank-one convex in X, since taking $H=\xi\otimes\eta$ with $\langle\xi,\eta\rangle=0$ yields

$$D^2_X\Psi(X).(\xi\otimes\eta,\xi\otimes\eta) = \begin{cases}\|\xi\|^2\|\eta\|^2, & \|\mathrm{dev}\,X\|\le\sigma_{\mathrm{y}},\\ \sigma_{\mathrm{y}}\left(\frac{\|\mathrm{dev}\,\xi\otimes\eta\|^2}{\|\mathrm{dev}\,X\|}-\frac{\langle\mathrm{dev}\,X,\xi\otimes\eta\rangle^2}{\|\mathrm{dev}\,X\|^3}\right), & \|\mathrm{dev}\,X\|>\sigma_{\mathrm{y}}.\end{cases}$$

Taking $X=\xi\otimes\eta$ shows finally

$$D^2_X\Psi(X).(\xi\otimes\eta,\xi\otimes\eta) = \begin{cases}\|\xi\|^2\|\eta\|^2, & \|\mathrm{dev}\,X\|\le\sigma_{\mathrm{y}},\\ 0, & \|\mathrm{dev}\,X\|>\sigma_{\mathrm{y}}.\end{cases}$$

3. Improved error estimates for Cosserat plasticity. Let $h > 0$ be the mesh size of a finite element method, and let $V_h \subset H^1_0(\Omega, \mathbb{R}^3)$ be a corresponding discrete finite element space. Let us concentrate on the displacement approximation only. In [49, Thm. 8] the following error estimate for the discrete solution $u_h^{\mu_c,n} \in V_h$ of the Galerkin approximation of (2.23) in V_h has been shown:

$$\|u^{\mu_c,n} - u_h^{\mu_c,n}\|_{H^1_0(\Omega)} \leq \frac{C_1}{\mu_c} \inf_{v_h \in V_h} \|u^{\mu_c,n} - v_h\|_{H^1_0(\Omega)}\,, \tag{3.1}$$

with a constant $C_1 > 0$. Here, $u^{\mu_c,n} = u^n$ is the exact solution of (2.13).

Using our regularity result from section 5, i.e., $u^{\mu_c,n} \in H^2(\Omega, \mathbb{R}^3)$, the right-hand side can be estimated qualitatively. If V_h is chosen to be the space of piecewise linear finite elements, then it holds [8, p. 107] that

$$\|u^{\mu_c,n} - u_h^{\mu_c,n}\|_{H^1_0(\Omega)} \leq \frac{C_2}{\mu_c}\, h\, \|u^{\mu_c,n}\|_{H^2(\Omega)}\,. \tag{3.2}$$

In [49] it has also been shown that for $\mu_c \to 0$ the classical Prandtl–Reuss symmetric Cauchy stresses σ^0 are approximated by the sequence of nonsymmetric stresses σ^{μ_c} whenever a safe load condition is satisfied. The estimate (3.2) strongly suggests therefore to balance h against μ_c to obtain optimal rates of convergence to the classical solution as in [54], where hardening-type approximations have been considered.

4. Higher regularity for alternative regularized update potentials. Our regularity result can also be applied to many other problems arising in the context of infinitesimal plasticity. There exist several other possibilities to regularize the classical update problem for the Prandtl–Reuss model. We recall the classical update problem: find a minimizer $u^n \in BD(\Omega, \mathbb{R}^3)$ of the functional

$$I^{\text{class}}_{\text{incr}}(u) = \mathcal{E}^{\text{class}}_{\text{incr}}(u, \varepsilon_p^{n-1}) - \int_\Omega \langle f^n, u\rangle \,\mathrm{dx}, \tag{4.1}$$

where $\mathcal{E}^{\text{class}}_{\text{incr}}$ denotes the free energy of the classical incremental problem defined by

$$\mathcal{E}^{\text{class}}_{\text{incr}}(u, \varepsilon_p) = \frac{1}{2\mu} \int_\Omega \Psi\big(2\mu(\operatorname{sym}(\nabla u) - \varepsilon_p)\big)\,\mathrm{dx} + \int_\Omega \frac{\lambda}{2} \operatorname{tr}[\nabla u]^2\,\mathrm{dx}, \tag{4.2}$$

with the potential Ψ as in (2.24). There is a vast literature on this Prandtl–Reuss update problem, mostly for the first time step $n = 1$, in which case it is the classical Hencky problem of total deformation plasticity [63, 54, 24, 25]. In this case, the plastic strain field ε_p is a symmetric bounded measure [63, 6]. The classical symmetric Cauchy stresses $\sigma = 2\mu\,(\operatorname{sym}\nabla u - \varepsilon_p) + \lambda \operatorname{tr}[\nabla u]\,\mathbb{1}$ satisfy $\sigma \in L^2(\Omega, \operatorname{Sym}(3))$; indeed higher regularity for the stresses can be shown in the sense that $\sigma \in H^1_{\text{loc}}(\Omega, \operatorname{Sym}(3)) \cap H^{\frac{1}{2}-\delta}(\Omega)$.

For regularization purposes the following proposals are usually made:

$$\mathcal{E}^{\text{reg}}_{\text{incr}}(u, \varepsilon_p) = \frac{1}{2\mu} \int_\Omega \Psi\big(2\mu(\operatorname{sym}(\nabla u) - \varepsilon_p)\big)\,\mathrm{dx} + \int_\Omega \frac{\lambda}{2} \operatorname{tr}[\nabla u]^2 + \operatorname{Reg}(\nabla u, \varepsilon_p)\,\mathrm{dx}, \tag{4.3}$$

with the function Reg in the form

$$\operatorname{Reg}(\nabla u, \varepsilon_p) = \frac{\mu\,\delta}{2}\,\|\operatorname{dev}\operatorname{sym}\nabla u - \varepsilon_p\|^2\,, \quad \text{Fuchs and Seregin [24, p. 60]}, \tag{4.4}$$

$$\mathrm{Reg}(\nabla u, \varepsilon_p) = \frac{1}{2\,\mu\,(1+\frac{\Delta t}{\eta})}\left[\|\mu(\operatorname{dev}\operatorname{sym}\nabla u - \varepsilon_p)\| - \sigma_{\mathrm{y}}\right]_+^2, \quad \text{linear viscosity } \eta,$$

$$\mathrm{Reg}(\nabla u, \varepsilon_p) = \frac{\mu\,\delta}{2}\left\|\nabla u - \varepsilon_p\right\|^2, \quad \text{locally strictly convex in } \nabla u.$$

In each case, for $\delta > 0$ the density of the update problem is then uniformly convex in the symmetric strain $\varepsilon = \operatorname{sym}\nabla u$. Moreover,

$$\mathrm{Reg}(\nabla u, \varepsilon_p) + \frac{\lambda}{2}\operatorname{tr}\left[\nabla u\right]^2 \geq c^+\left\|\varepsilon - \varepsilon_p\right\|^2, \tag{4.5}$$

Korn's first inequality establishes quadratic growth, and we have uniform convexity for the regularized problem. Our main regularity result applies therefore also to these models.

In the case with linear hardening it is simpler to write the update potential directly. We consider as an example isotropic hardening with the hardening variable $\alpha \geq 0$ (a measure for the accumulated plastic strain in the previous time step). Here, the energy $\mathcal{E}_{\mathrm{incr}}$ can be expressed as (cf. [60, p. 124])

$$\mathcal{E}^{\mathrm{hard}}_{\mathrm{incr}}(u, \varepsilon_p, \alpha) = \frac{1}{2\mu}\int_\Omega \Psi_{\mathrm{hard}}\big(2\mu(\operatorname{sym}(\nabla u) - \varepsilon_p), \alpha\big)\,\mathrm{dx} + \int_\Omega \frac{\lambda}{2}\operatorname{tr}\left[\nabla u\right]^2\mathrm{dx}, \tag{4.6}$$

with (cf. (2.24))

$$\Psi_{\mathrm{hard}}(X,\alpha) = \begin{cases} \frac{1}{2}\left\|X\right\|^2, & \|\operatorname{dev} X\| \leq \sigma_{\mathrm{y}} + H\,\alpha, \\ \frac{1}{2\,(1+\frac{H}{1[\mathrm{MPa}]})}\Big(\frac{H}{1[\mathrm{MPa}]}\left\|X\right\|^2 + \frac{1}{3}\operatorname{tr}\left[X\right]^2 & \\ \quad + 2\,(\sigma_{\mathrm{y}} + H\,\alpha)\,\|\operatorname{dev} X\| - (\sigma_{\mathrm{y}} + H\,\alpha)^2\Big), & \|\operatorname{dev} X\| > \sigma_{\mathrm{y}} + H\,\alpha \end{cases}$$

$$= \frac{1}{2}\left\|X\right\|^2 - \frac{1}{2\,(1+\frac{H}{1[\mathrm{MPa}]})}\left[\|\operatorname{dev} X\| - (\sigma_{\mathrm{y}} + H\,\alpha)\right]_+^2, \tag{4.7}$$

whose second derivative coincides with the consistent tangent method introduced already in [61]. The constant $H > 0$ is the hardening modulus with dimension [MPa]. In this form it is easy to see that for positive hardening modulus $H > 0$ the isotropic hardening update potential is uniformly convex in $\operatorname{sym}\nabla u$ with quadratic growth and has a Lipschitz continuous derivative. Therefore, our main regularity result applies also to this functional.[4] The relative merits of each individual regularization scheme depend on their ability to balance regularization and approximation. Linear viscosity and hardening can be justified on physical grounds, but the (small) viscosity parameter $\eta > 0$ is difficult to estimate, as is the linear hardening modulus $H > 0$. The physically motivated regularization terms have the property to control only the symmetric part of the displacement gradient. The regularization $(4.4)_3$, however, does not satisfy the linearized frame-indifference condition.

All alternative regularization procedures thus establish local coercivity in the strains. In contrast, the Cosserat regularization is weaker in the sense that only strong Legendre–Hadamard ellipticity is reestablished, which, provided that displacement boundary data are prescribed, suffices for existence, uniqueness, and higher regularity. Thus the Cosserat approach appears as the weakest regularization among the ones considered.

[4] Repin [54, eq. (2.3)] calls $(4.4)_2$ linear hardening and shows the regularity $u^\delta \in H^2_{\mathrm{loc}}(\Omega, \mathbb{R}^3)$, while for the planar case $n = 2$ he obtains $u^\delta \in H^2(\Omega, \mathbb{R}^2)$ if $\Gamma = \partial\Omega$ is smooth.

5. The regularity theorem. We know already that problem (2.22) has solutions $u^n \in H^1(\Omega,\mathbb{R}^3)$. Looking at the system for the microrotations A^n at given $\nabla u^n \in L^2(\Omega,\mathbb{M}^{3\times 3})$ we realize at once that the linearity in A^n together with the Laplacian structure allows us to use standard elliptic regularity results for linear systems, which yields higher regularity for the microrotations: $A^n \in H^2(\Omega,\mathfrak{so}(3))$. In this section we study the regularity of the displacement field u^n, which is determined through (2.13).

5.1. Higher regularity for a quasi-linear elliptic system. The quasi-linear elliptic system introduced in section 2.4 is a special case of the systems which we define here below. For $d,m,N\geq 1$ and $\Omega\subset\mathbb{R}^d$ let $\mathcal{M}:\Omega\times\mathbb{M}^{m\times d}\times\mathbb{R}^N\to\mathbb{M}^{m\times d}$ be a matrix-valued function with the following properties.

R1. The mapping $\mathcal{M}:\Omega\times\mathbb{M}^{m\times d}\times\mathbb{R}^N\to\mathbb{M}^{m\times d}$ is a Carathéodory function which is Lipschitz continuous in the following sense: there exist constants $L_1,L_2>0$ such that for every $x,x_i\in\Omega$, $a,a_i\in\mathbb{M}^{m\times d}$, and $z,z_i\in\mathbb{R}^N$ we have

$$\begin{aligned}\|\mathcal{M}(x_1,a,z)-\mathcal{M}(x_2,a,z)\|&\leq L_1(\|a\|+\|z\|)\,\|x_1-x_2\|\,,\\ \|\mathcal{M}(x,a_1,z_1)-\mathcal{M}(x,a_2,z_2)\|&\leq L_2(\|a_1-a_2\|+\|z_1-z_2\|),\\ \mathcal{M}(x,0,0)&=0.\end{aligned}$$

Assumption R1 implies the useful estimate

$$\begin{aligned}&\|\mathcal{M}(x_1,a_1,z_1)-\mathcal{M}(x_2,a_2,z_2)\|\\ &\qquad\leq L_1(\|a_1\|+\|z_1\|)\,\|x_1-x_2\|+L_2(\|a_1-a_2\|+\|z_1-z_2\|).\end{aligned}\tag{5.1}$$

R2. The mapping $\mathcal{M}$ is strongly rank-one monotone. That means that there exists a constant $c_{LH}>0$ such that for every $x\in\overline{\Omega}$, $a\in\mathbb{M}^{m\times d}$, $z\in\mathbb{R}^N$, $\xi\in\mathbb{R}^m$, and $\eta\in\mathbb{R}^d$ we have

$$\langle\mathcal{M}(x,a+\xi\otimes\eta,z)-\mathcal{M}(x,a,z),\xi\otimes\eta\rangle\geq c_{LH}\,\|\xi\|^2\,\|\eta\|^2\,.\tag{5.2}$$

R3. The Gårding inequality shall be satisfied: there exist constants $C_G>0$, $c_G\in\mathbb{R}$ such that for every $u_1,u_2\in H^1(\Omega)$ with $u_1-u_2\in H^1_0(\Omega)$ and for every $z\in L^2(\Omega)$ the following inequality is valid:

$$\begin{aligned}\int_\Omega\langle\mathcal{M}(x,\nabla u_1,z)-\mathcal{M}(x,\nabla u_2,z),\nabla(u_1-u_2)\rangle\,\mathrm{dx}&\\ \geq C_G\,\|\nabla(u_1-u_2)\|^2_{L^2(\Omega)}-c_G\,\|u_1-u_2\|^2_{L^2(\Omega)}\,.&\end{aligned}$$

Remark 5.1. If $\mathcal{M}$ is differentiable, then the Gårding inequality already implies that $\mathcal{M}$ is rank-one monotone; see, for example, [65, Thm. 6.1].

We investigate the regularity properties of weak solutions to the following quasi-linear elliptic boundary value problem. For given $g\in H^{\frac12}(\partial\Omega)$, $z\in L^2(\Omega,\mathbb{R}^N)$, and $f\in L^2(\Omega,\mathbb{R}^m)$ find $u\in H^1(\Omega,\mathbb{R}^m)$ with $u|_{\partial\Omega}=g$ such that for every $v\in H^1_0(\Omega,\mathbb{R}^m)$ we have

$$\int_\Omega\langle\mathcal{M}(x,\nabla u(x),z(x)),\nabla v(x)\rangle\,\mathrm{dx}=\int_\Omega\langle f,v\rangle\,\mathrm{dx}.\tag{5.3}$$

THEOREM 5.2. *Let $\Omega\subset\mathbb{R}^d$ be a bounded $\mathcal{C}^{1,1}$-smooth domain, $m\geq 1$, and $N\geq 1$, and assume that $\mathcal{M}:\Omega\times\mathbb{M}^{m\times d}\times\mathbb{R}^N\to\mathbb{M}^{m\times d}$ satisfies* R1–R3. *Furthermore, let*

$g \in H^{\frac{3}{2}}(\partial\Omega)$, $z \in H^1(\Omega)$, and $f \in L^2(\Omega)$. Every weak solution $u \in H^1(\Omega)$ of (5.3) with $u|_{\partial\Omega} = g$ is an element of $H^2(\Omega)$ and satisfies

$$\|u\|_{H^2(\Omega)} \le c\left(\|g\|_{H^{\frac{3}{2}}(\partial\Omega)} + \|z\|_{H^1(\Omega)} + \|f\|_{L^2(\Omega)} + \|u\|_{H^1(\Omega)}\right).$$

Before we prove Theorem 5.2, we apply it to the situation described in section 2.4. There, $m = d = 3$ and $\mathbb{R}^N$ is identified with $\mathrm{Sym}(3) \times \mathfrak{so}(3)$ so that $z = (\varepsilon_p, A)$. Moreover,

$$\begin{aligned}\mathcal{M}(x, \nabla u, z) &= \mathcal{M}_C(\nabla u, \varepsilon_p, A)\\ &= P_K(2\mu(\mathrm{sym}\,\nabla u - \varepsilon_p)) + \lambda(\mathrm{tr}\,[\nabla u])\mathbb{1} + 2\mu_c(\mathrm{skew}(\nabla u) - A).\end{aligned}$$

Since P_K is a Lipschitz continuous mapping, we see immediately that assumption R1 is satisfied. R2 is proved in Lemma 2.1, and the Gårding inequality is satisfied since $\mathcal{M}_C$ generates a strongly monotone operator on $H^1_0(\Omega)$; see Lemma 2.2. Therefore, we have the following result for the reduced update problem (2.6).

THEOREM 5.3. *Let Ω be $\mathcal{C}^{1,1}$-smooth, $f^n \in L^2(\Omega)$, and $\varepsilon_p^{n-1} \in H^1(\Omega)$. Then $u_n \in H^2(\Omega)$, $A^n \in H^2(\Omega)$, and $\varepsilon_p^n \in H^1(\Omega)$.*

The proof of Theorem 5.2 is carried out with a difference quotient technique. We cover the boundary of Ω with a finite number of domains and map each of these domains with a $C^{1,1}$ diffeomorphism onto the unit cube in such a way that the image of the boundary of Ω lies on the midplane of the unit cube. We first prove higher regularity in directions tangential to the midplane by estimating difference quotients. The regularity in the normal direction is then obtained on the basis of the tangential regularity and by using the differential equation together with the rank-one monotonicity of $\mathcal{M}$.

Since $\mathcal{M}$ is nonlinear and since we assumed rank-one monotonicity instead of strong monotonicity, we cannot use as test functions the usual finite differences of the type $h^{-1}\varphi^2(x)(u(x+h) - u(x))$, where φ is a cut-off function. Instead, we use differences which are based on inner variations. We begin the proof of Theorem 5.2 by studying a model problem on a half cube.

5.2. A model problem on a half cube. Let $C_r = \{x \in \mathbb{R}^d; |x_i| < r, 1 \le i \le d\}$ be a cube with side length $2r$, $C_r^{\pm}$ the upper and lower half cube, respectively, and $M_r = \{x \in C_r;\, x_d = 0\}$ the midplane.

LEMMA 5.4. *Let $\Omega = C_1^-$, $f \in L^2(C_1^-)$, and $z \in H^1(C_1^-)$, and assume that $u \in H^1(C_1^-)$ with $u|_{M_1} = 0$ satisfies (5.3). Then for every $r \in (0,1)$ and for $1 \le i \le d-1$ we have $\partial_i u \in H^1(C_r^-)$. Moreover, there is a constant $c_r > 0$ such that*

$$\|\partial_i u\|_{H^1(C_r^-)} \le c_r\big(\|u\|_{H^1(C_1^-)} + \|z\|_{H^1(C_1^-)} + \|f\|_{L^2(C_1^-)}\big). \tag{5.4}$$

Proof. Let $r \in (0,1)$ and $\varphi \in \mathcal{C}_0^\infty(C_1)$ with $\varphi(x) = 1$ on C_r. For $h \in \mathbb{R}^d$ we introduce the mapping

$$\tau_h : C_1 \to \mathbb{R}^d : x \to \tau_h(x) = x + \varphi(x)h.$$

Let $h_0 = \|\varphi\|^{-1}_{W^{1,\infty}(C_1)} \min\big\{1, \mathrm{dist}(\mathrm{supp}\,\varphi, \partial C_1)\big\}$. For every $h \in \mathbb{R}^d$ with $|h| < h_0$ and h parallel to the plane M_1, the mapping τ_h is a diffeomorphism from C_1 onto itself with $\tau_h(C_1^{\pm}) = C_1^{\pm}$, $\tau_h(M_1) = M_1$, and $\tau_h(x) = x$ for every $x \in \partial C_1$; see, e.g., [26]. Moreover, τ_h has the following properties (if $|h| < h_0$):

$$\nabla\tau_h(x) = \big(\mathbb{1} + h \otimes \nabla\varphi(x)\big), \quad \det[\nabla\tau_h(x)] = 1 + \langle h, \nabla\varphi(x)\rangle,$$
$$\nabla_y\tau_h^{-1}(y) = \big(\mathbb{1} + h \otimes \nabla\varphi\big)^{-1}\big|_{\tau_h^{-1}(y)} = \mathbb{1} - \big((1 + \langle h, \nabla\varphi\rangle)^{-1} h \otimes \nabla\varphi\big)\big|_{\tau_h^{-1}(y)}.$$

For a function $v : C_1^- \to \mathbb{R}^s$ we introduce

$$\triangle_h v = v \circ \tau_h - v, \qquad \triangle^h v = v - v \circ \tau_h^{-1}.$$

For $f, g \in L^2(C_1^-)$, $|h| < h_0$ and $h \parallel M_1$ the following product rule is valid:

$$\int_{C_1^-} f \triangle^h g \,\mathrm{dx} = -\int_{C_1^-} g \triangle_h f \,\mathrm{dx} - \int_{C_1^-} (f \circ \tau_h \, g) \,\langle h, \nabla\varphi\rangle \,\mathrm{dx}. \tag{5.5}$$

This identity can be shown by a transformation of coordinates $y = \tau_h(x)$ in the term $(g \circ \tau_h^{-1})f$. Let $u \in H_0^1(C_1^-)$ be a solution of (5.3). For $h \in \mathbb{R}^d$ with $|h| < h_0$ and $h \parallel M_1$ we define the double difference $v_h(x) = \triangle^h(\triangle_h u(x))$. In view of the assumptions on φ, h_0, and h it follows that $v_h \in H_0^1(C_1^-)$. Inserting v_h into (5.3) yields

$$\int_{C_1^-} \langle \mathcal{M}(x, \nabla u, z), \nabla v_h\rangle \,\mathrm{dx} = \int_{C_1^-} \langle f, v_h\rangle \,\mathrm{dx}. \tag{5.6}$$

Note that $\nabla v_h = \triangle^h \nabla(\triangle_h u) + [(\det[\nabla\tau_h])^{-1}(\nabla\triangle_h u)\, h \otimes \nabla\varphi] \circ \tau_h^{-1}$, and therefore (5.6) is equivalent to

$$\begin{aligned}
&\int_{C_1^-} \langle \mathcal{M}(x, \nabla u, z), \triangle^h \nabla(\triangle_h u)\rangle \,\mathrm{dx} \\
&\quad = -\int_{C_1^-} \big\langle \mathcal{M}(x, \nabla u, z), \big(\det[\nabla\tau_h]^{-1}(\nabla\triangle_h u)\, h \otimes \nabla\varphi\big) \circ \tau_h^{-1}\big\rangle \,\mathrm{dx} + \int_{C_1^-} \langle f, \triangle^h\triangle_h u\rangle \,\mathrm{dx}.
\end{aligned}$$

Furthermore, the product rule (5.5) entails

$$\begin{aligned}
\int_{C_1^-} \langle \triangle_h \mathcal{M}(x, \nabla u, z), \nabla\triangle_h u\rangle \,\mathrm{dx} &= -\int_{C_1^-} \langle \mathcal{M}(x, \nabla u, z) \circ \tau_h, \nabla\triangle_h u\rangle \langle h, \nabla\varphi\rangle \,\mathrm{dx} \\
&\quad + \int_{C_1^-} \big\langle \mathcal{M}(x, \nabla u, z), \big((\det[\nabla\tau_h])^{-1}(\nabla\triangle_h u)\, h \otimes \nabla\varphi\big) \circ \tau_h^{-1}\big\rangle \,\mathrm{dx} \\
&\quad + \int_{C_1^-} \langle f, \triangle^h\triangle_h u\rangle \,\mathrm{dx} \\
&=: S_1 + S_2 + S_3.
\end{aligned} \tag{5.7}$$

Finally we have

$$\begin{aligned}
&\int_{C_1^-} \langle \mathcal{M}(x, \nabla(u \circ \tau_h), z) - \mathcal{M}(x, \nabla u, z), \nabla\triangle_h u\rangle \,\mathrm{dx} \\
&\qquad = \int_{C_1^-} \langle \triangle_h \mathcal{M}(x, \nabla u, z), \nabla\triangle_h u\rangle \,\mathrm{dx} \\
&\qquad\quad + \int_{C_1^-} \langle \mathcal{M}(x, \nabla(u \circ \tau_h), z) - \mathcal{M}(x, \nabla u, z) \circ \tau_h, \nabla\triangle_h u\rangle \,\mathrm{dx} \\
&\qquad \overset{(5.7)}{=} S_1 + S_2 + S_3 + \int_{C_1^-} \langle \mathcal{M}(x, \nabla(u \circ \tau_h), z) - \mathcal{M}(x, \nabla u, z) \circ \tau_h, \nabla\triangle_h u\rangle \,\mathrm{dx} \\
&\qquad = S_1 + \cdots + S_4.
\end{aligned} \tag{5.8}$$

The next task is to show that there is a constant $c > 0$, which does not depend on h, such that

$$|S_1 + \cdots + S_4| \le c\,|h| \left(\|u\|_{H^1(C_1^-)} + \|z\|_{H^1(C_1^-)} + \|f\|_{L^2(C_1^-)}\right) \|\triangle_h u\|_{H^1(C_1^-)}. \tag{5.9}$$

Due to the Lipschitz assumptions on $\mathcal{M}$ we have

$$\begin{aligned}|S_1| + |S_2| &\le c\,|h|\,\|\mathcal{M}(\cdot,\nabla u,z)\|_{L^2(C_1^-)}\,\|\triangle_h u\|_{H^1(C_1^-)}\\ &\le c\,|h|\,\big(\|u\|_{H^1(C_1^-)} + \|z\|_{L^2(C_1^-)}\big)\,\|\triangle_h u\|_{H^1(C_1^-)}\,.\end{aligned}$$

Moreover, since $f \in L^2(C_1^-)$, the term S_3 can be estimated as

$$|S_3| \le c\,|h|\,\|f\|_{L^2(C_1^-)}\,\|\triangle_h u\|_{H^1(C_1^-)}\,.$$

By inequality (5.1) we see that

$$\begin{aligned}|S_4| &\le cL_1\,|h|\,\big(\|u\|_{H^1(C_1^-)} + \|z\|_{H^1(C_1^-)}\big)\,\|\triangle_h u\|_{H^1(C_1^-)}\\ &\quad + cL_2\big(\|\nabla(u\circ\tau_h) - (\nabla u)\circ\tau_h\|_{L^2(C_1^-)} + c\,|h|\,\|z\|_{H^1(C_1^-)}\big)\,\|\triangle_h u\|_{H^1(C_1^-)}\,.\end{aligned}$$

The identity $\nabla(u\circ\tau_h) - (\nabla u)\circ\tau_h = (\nabla u)\circ\tau_h\,(h\otimes\nabla\varphi)$ leads to

$$|S_4| \le c\,|h|\,\big(\|u\|_{H^1(C_1^-)} + \|z\|_{H^1(C_1^-)}\big)\,\|\triangle_h u\|_{H^1(C_1^-)}\,.$$

Collecting all the above estimates we finally arrive at inequality (5.9). Gårding's inequality (see R3) and Poincaré's inequality imply that

$$\begin{aligned}\int_{C_1^-} &\langle \mathcal{M}(x,\nabla(u\circ\tau_h),z) - \mathcal{M}(x,\nabla u,z), \nabla\triangle_h u\rangle\,\mathrm{dx}\\ &\ge C_G\,\|\nabla\triangle_h u\|^2_{L^2(C_1^-)} - c_G\,\|\triangle_h u\|^2_{L^2(C_1^-)}\\ &\ge c\big(\|\triangle_h u\|^2_{H^1(C_1^-)} - |h|^2\,\|u\|^2_{H^1(C_1^-)}\big).\end{aligned}$$

Combining the above estimates with (5.8) and (5.9) results finally in

$$\begin{aligned}\|\triangle_h u\|^2_{H^1(C_1^-)} &\le c\,|h|\,\big(\|u\|_{H^1(C_1^-)} + \|z\|_{H^1(C_1^-)} + \|f\|_{L^2(C_1^-)}\big)\,\|\triangle_h u\|_{H^1(C_1^-)}\\ &\quad + c\,|h|^2\,\|u\|^2_{H^1(C_1^-)}\,,\end{aligned}$$

and the constant c is independent of h. From Young's inequality we obtain

$$|h|^{-1}\,\|\triangle_h u\|_{H^1(C_1^-)} \le c\big(\|u\|_{H^1(C_1^-)} + \|z\|_{H^1(C_1^-)} + \|f\|_{L^2(C_1^-)}\big). \tag{5.10}$$

It follows from this inequality that $\partial_i u \in H^1(C_r^-)$ for $1 \le i \le d-1$ and that $\|\partial_i u\|_{H^1(C_r^-)}$ is bounded by the right-hand side in (5.10); see, e.g., [34]. □

Remark 5.5. If we choose the usual finite differences as test functions, i.e., $\tilde{v}_h(x) = \delta_{-h}(\varphi^2\delta_h u)$, where $\delta_h u = u(x+h) - u(x)$, then similar calculations as those for v_h lead to the estimate

$$\begin{aligned}\int_{C_1^-} &\varphi^2(x)\langle \mathcal{M}(x,\nabla u(x+h),z(x)) - \mathcal{M}(x,\nabla u(x),z(x)), \delta_h\nabla u\rangle\,\mathrm{dx}\\ &\le c\,|h|\,\big\|\varphi^2\delta_h u\big\|_{H^1(C_1^-)}\,;\end{aligned} \tag{5.11}$$

compare also (5.8) and (5.9). But now neither R2 nor R3 helps us to find a lower bound for the left-hand side of (5.11) in terms of $\|\varphi^2\delta_h\nabla u\|^2_{L^2(C_1^-)}$, since in general $\delta_h\nabla u$ is not a rank-one matrix, and since we cannot interchange φ and $\mathcal{M}$ due to the nonlinearity of $\mathcal{M}$.

LEMMA 5.6 (regularity in the normal direction). *With the same assumptions as in Lemma 5.4 it follows for every $r \in (0,1)$ that $\partial_d u \in H^1(C_r^-)$. Furthermore, there exists a constant $c_r > 0$ such that*

$$\|u\|_{H^2(C_r^-)} \leq c_r\big(\|z\|_{H^1(C_1^-)} + \|f\|_{L^2(C_1^-)} + \|u\|_{H^1(C_1^-)}\big). \tag{5.12}$$

Proof. Let $r \in (0,1)$. Equation (5.3) implies that

$$\operatorname{Div} \mathcal{M}(x, \nabla u(x), z(x)) + f(x) = 0 \tag{5.13}$$

for almost every $x \in C_1^-$. Let $\mathcal{M}_i$ denote the columns of the matrix-valued function $\mathcal{M}$, i.e., $\mathcal{M}_i(x,a,z) = (\mathcal{M}_i^\alpha(x,a,z))_{1\leq\alpha\leq m} \in \mathbb{R}^m$ for $1 \leq i \leq d$. The Lipschitz continuity of $\mathcal{M}$ and the tangential regularity proved in Lemma 5.4 guarantee that $\partial_i \mathcal{M}_i(\cdot, \nabla u, z) \in L^2(C_r^-)$ for $1 \leq i \leq d-1$ and is bounded by the right-hand side in (5.4). Together with (5.13) we obtain therefore

$$\partial_d \mathcal{M}_d(\cdot, \nabla u, z) = -f - \sum_{i=1}^{d-1} \partial_i \mathcal{M}_i(\cdot, \nabla u, z) \in L^2(C_r^-).$$

By Lemma 7.23 in [27] the derivative ∂_d can be replaced with a finite difference in the following way: For every $\Omega' \subset\subset C_r^-$ and every $h \in \mathbb{R}^d$ with $|h| < \operatorname{dist}(\Omega', \partial C_r^-)$ and $h \perp M_1$ we have

$$\begin{aligned} \|\delta_h \mathcal{M}_d(\cdot, \nabla u, z)\|_{L^2(\Omega')} &\leq \left(\|f\|_{L^2(C_r^-)} + \sum_{i=1}^{d-1} \|\partial_i \mathcal{M}_i(\cdot, \nabla u, z)\|_{L^2(C_r^-)} \right) |h| \\ &=: c_0 \, |h| \,. \end{aligned} \tag{5.14}$$

Here, $\delta_h v(x) := v(x+h) - v(x)$ for $h \in \mathbb{R}^d$. Thus, for every $h \perp M_1$ with $|h| < \operatorname{dist}(\Omega', \partial C_r^-)$ we have

$$\int_{\Omega'} \langle \delta_h \mathcal{M}_d(x, \nabla u, z), \delta_h \partial_d u \rangle \, \mathrm{dx} \leq c_0 \, |h| \, \|\delta_h \partial_d u\|_{L^2(\Omega')} \,, \tag{5.15}$$

where c_0 is the constant from (5.14). We now split the left-hand side into a term which can be estimated from below due to the rank-one monotonicity of $\mathcal{M}$ and into terms which may be estimated from above using the Lipschitz continuity of $\mathcal{M}$ and the regularity results from Lemma 5.4. For functions $v : C_1^- \to \mathbb{R}^m$ we define $\tilde{\nabla} v(x) = (\partial_1 v(x), \dots, \partial_{d-1} v(x), 0) \in \mathbb{M}^{m\times d}$. Furthermore, $v_h(x) := v(x+h)$ and $e_d = (0, \dots, 0, 1)^\top \in \mathbb{R}^d$. With these notations we have

$$\begin{aligned} &\int_{\Omega'} \langle \mathcal{M}_d(x, \tilde{\nabla} u + \partial_d u_h \otimes e_d, z) - \mathcal{M}_d(x, \nabla u, z), \delta_h \partial_d u \rangle \, \mathrm{dx} \\ &\quad = \int_{\Omega'} \langle \delta_h \mathcal{M}_d(x, \nabla u, z), \delta_h \partial_d u \rangle \, \mathrm{dx} \\ &\qquad + \int_{\Omega'} \langle \mathcal{M}_d(x, \tilde{\nabla} u + \partial_d u_h \otimes e_d, z) - \mathcal{M}_d(x+h, \nabla u_h, z_h), \delta_h \partial_d u \rangle \, \mathrm{dx} \\ &\quad = S_1 + S_2. \end{aligned} \tag{5.16}$$

The term S_1 is already estimated in (5.15). From the Lipschitz continuity of $\mathcal{M}$ (see (5.1)) and the regularity results of Lemma 5.4 we obtain by straightforward

calculations

(5.17)
$$\begin{aligned}|S_2| &\le c\,\|\delta_h\partial_d u\|_{L^2(\Omega')}\left(\left(\left\|\tilde\nabla u+\partial_d u_h\otimes e_d\right\|_{L^2(\Omega')}+\|z\|_{H^1(C_1^-)}\right)|h|+\left\|\delta_h\tilde\nabla u\right\|_{L^2(\Omega')}\right)\\ &\le c\,|h|\left(\|u\|_{H^1(C_1^-)}+\|z\|_{H^1(C_1^-)}+\left\|\partial_d\tilde\nabla u\right\|_{L^2(C_r^-)}\right)\|\delta_h\partial_d u\|_{L^2(\Omega')},\end{aligned}$$

and the constant c is independent of Ω' and h. Moreover, choosing $\xi=\partial_d u_h$ and $\eta=e_d$ in (5.2), we obtain for the left-hand side in (5.16) from the rank-one monotonicity of $\mathcal{M}$ that

$$\begin{aligned}&\int_{\Omega'}\langle\mathcal{M}_d(x,\tilde\nabla u+\partial_d u_h\otimes e_d,z)-\mathcal{M}_d(x,\nabla u,z),\delta_h\partial_d u\rangle\,\mathrm{dx}\\ &\qquad\qquad\ge c_{LH}\,\|\delta_h\partial_d u\|^2_{L^2(\Omega')}.\end{aligned}\tag{5.18}$$

Estimates (5.15)–(5.18) together with Young's inequality finally imply that

$$|h|^{-1}\,\|\delta_h\partial_d u\|_{L^2(\Omega')}\le c\big(\|u\|_{H^1(C_1^-)}+\|z\|_{H^1(C_1^-)}+\left\|\partial_d\tilde\nabla u\right\|_{L^2(C_r^-)}\big)\tag{5.19}$$

for every $h\perp M_1$. The constant c is independent of h and $\Omega'\Subset C_r^-$. This implies that $\partial_d^2 u\in L^2(C_r^-)$ and that $\|\partial_d^2 u\|_{L^2(c_r^-)}$ is bounded by the right-hand side in (5.19). Estimate (5.12) is a combination of (5.19) and (5.4). ∎

5.3. Proof of Theorem 5.2. Let the assumptions of Theorem 5.2 be valid, and assume that $g=0$. Choose $x_0\in\partial\Omega$, and let U_{x_0} be a neighborhood of x_0 such that there exists a $\mathcal{C}^{1,1}$ diffeomorphism $\Phi_{x_0}:U_{x_0}\to C_1$, where C_1 is the unit cube in $\mathbb{R}^d$, with the following properties (we omit the index x_0): $\Phi(U)=C_1$, $\Phi(U\cap\Omega)=C_1^-$, $\Phi(U\setminus\overline{\Omega})=C_1^+$, $\Phi(U\cap\partial\Omega)=M_1$, and $\Phi(x_0)=0$. Let $u\in H_0^1(\Omega)$ be a solution for (5.3) with the data $f\in L^2(\Omega)$ and $z\in H^1(\Omega)$. It follows that

$$\int_{U\cap\Omega}\langle\mathcal{M}(x,\nabla u,z),\nabla v\rangle\,\mathrm{dx}=\int_{U\cap\Omega}\langle f,v\rangle\,\mathrm{dx}$$

for every $v\in H_0^1(\Omega\cap U)$. After a transformation of coordinates with $y=\Phi(x)$ and $\Psi:=\Phi^{-1}$, the previous equation can be written as follows: Let $\tilde u(y)=u(\Psi(y))$. For every $v\in H_0^1(C_1^-)$ we have

$$\int_{C_1^-}\langle\tilde{\mathcal{M}}(y,\nabla\tilde u,\tilde z),\nabla v\rangle\,\mathrm{dy}=\int_{C_1^-}\langle\tilde f,v\rangle\,\mathrm{dy}.$$

Here, we use the abbreviations

$$\tilde{\mathcal{M}}(y,a,\zeta)=|\det[\nabla\Psi(y)]|\,\mathcal{M}(\Psi(y),a(\nabla\Psi(y))^{-1},\zeta)(\nabla\Psi(y))^{-\top},\tag{5.20}$$

$$\tilde f(y)=|\det[\nabla\Psi(y)]|\,f(\Psi(y)),\tag{5.21}$$

$$\tilde z(y)=z(\Psi(y))\tag{5.22}$$

for $y\in C_1^-$, $a\in\mathbb{M}^{m\times d}$, and $\zeta\in\mathbb{R}^N$. It follows immediately from the properties of the diffeomorphism Φ and from those of $\mathcal{M}$ that $\tilde{\mathcal{M}}$ satisfies R1–R3 with respect to C_1^-. Furthermore, $\tilde f$ and $\tilde z$ have the smoothness required in Lemma 5.4. Thus, Lemmata 5.4 and 5.6 guarantee that $\tilde u\in H^2(C_r^-)$ for every $r<1$ and that estimate (5.12) is valid. After applying the inverse transformation $\Psi:C_1^-\to U\cap\Omega$, we have finally shown the following: For every $x_0\in\overline{\Omega}$ there exists an open neighborhood $\tilde U_{x_0}$ such

that $u|_{\tilde{U}_{x_0}\cap\Omega} \in H^2(\tilde{U}_{x_0}\cap\Omega)$ and estimate (5.12) is valid with respect to $\tilde{U}_{x_0}\cap\Omega$. The constants may depend on x_0. Since Ω is assumed to be bounded, we can cover $\overline{\Omega}$ by a finite number of the domains $\tilde{U}_{x_0}$ and obtain finally that $u \in H^2(\Omega)$ with

$$\|u\|_{H^2(\Omega)} \leq c\big(\|z\|_{H^1(\Omega)} + \|f\|_{L^2(\Omega)} + \|u\|_{H^1(\Omega)}\big). \tag{5.23}$$

This proves Theorem 5.2 for the case of vanishing Dirichlet conditions. The general case can be seen as follows. There exists a linear and continuous extension operator $F : H^{\frac{3}{2}}(\partial\Omega) \to H^2(\Omega)$ with $(F(g))\big|_{\partial\Omega} = g$ for every $g \in H^{\frac{3}{2}}(\Omega)$; see, for example, [68]. Then $u \in H^1(\Omega)$ with $u|_{\partial\Omega} = g$ for some $g \in H^{\frac{3}{2}}(\partial\Omega)$ is a solution to (5.3) if and only if there exists an element $\tilde{u} \in H^1_0(\Omega)$ with $u = \tilde{u} + F(g)$ and for every $v \in H^1_0(\Omega)$, $\tilde{u}$ satisfies

$$\int_\Omega \langle \hat{\mathcal{M}}(x, \nabla\tilde{u}, \tilde{z}), \nabla v\rangle \,\mathrm{dx} = \int_\Omega \langle f, v\rangle \,\mathrm{dx},$$

where $\tilde{z} = (F(g), z)$ and $\hat{\mathcal{M}}(x, a, \tilde{z}) = \mathcal{M}(x, a + F(g)(x), z)$. Clearly, $\hat{\mathcal{M}}$ satisfies R1–R3 as well, and by the first part of this proof it follows that $\tilde{u} \in H^2(\Omega)$. This finishes the proof of Theorem 5.2.

6. Discussion. We have shown that the time-incremental Cosserat elasto-plasticity problem admits $H^1(\Omega)$-regular updates of the symmetric plastic strain ε_p^n, provided that the previous plastic strain ε_p^{n-1} is in $H^1(\Omega)$ and the domain and data are suitably regular. Altogether, the time-incremental problem allows the regularity for all $n \in \mathbb{N} : u^n \in H^2(\Omega, \mathbb{R}^3)$, $\varepsilon_p^n \in H^1(\Omega, \mathrm{Sym}(3))$, and $A^n \in H^2(\Omega, \mathfrak{so}(3))$. Uniform bounds in time are missing, and it is an open question whether a similar result holds for the time-continuous problem.

The presented method of proof for higher regularity uses a difference quotient method which is based on inner variations and can be extended to more general problems. This will be the subject of further investigations.

Appendix. Notation. We denote by $\mathbb{M}^{3\times3}$ the set of real 3×3 second order tensors, written with capital letters. The standard Euclidean scalar product on $\mathbb{M}^{3\times3}$ is given by $\langle X, Y\rangle_{\mathbb{M}^{3\times3}} = \mathrm{tr}\,[XY^T]$, and thus the Frobenius tensor norm is $\|X\|^2 = \langle X, X\rangle_{\mathbb{M}^{3\times3}}$ (we use these symbols indifferently for tensors and vectors). The identity tensor on $\mathbb{M}^{3\times3}$ will be denoted by $\mathbb{1}$, so that $\mathrm{tr}\,[X] = \langle X, \mathbb{1}\rangle$. We let Sym and PSym denote the symmetric and positive definite symmetric tensors, respectively. We adopt the usual abbreviations of Lie algebra theory; i.e., $\mathfrak{so}(3) := \{X \in \mathbb{M}^{3\times3} \mid X^T = -X\}$ are skew symmetric second order tensors, and $\mathfrak{sl}(3) := \{X \in \mathbb{M}^{3\times3} \mid \mathrm{tr}\,[X] = 0\}$ are traceless tensors. We set $\mathrm{sym}(X) = \frac{1}{2}(X^T + X)$ and $\mathrm{skew}(X) = \frac{1}{2}(X - X^T)$ such that $X = \mathrm{sym}(X) + \mathrm{skew}(X)$. For $X \in \mathbb{M}^{3\times3}$ we set for the deviatoric part $\mathrm{dev}\,X = X - \frac{1}{3}\,\mathrm{tr}\,[X]\,\mathbb{1} \in \mathfrak{sl}(3)$.

For a second order tensor X we let $X.e_i$ be the application of the tensor X to the column vector e_i. The first and second differential of a scalar-valued function $W(F)$ are written $D_F W(F).H$ and $D^2_F W(F).(H, H)$, respectively. Sometimes we use also $\partial_X W(X)$ to denote the first derivative of W with respect to X. We employ the standard notation of Sobolev spaces, i.e., $L^2(\Omega)$, $H^{1,2}(\Omega)$, $H^{1,2}_\circ(\Omega)$, which we use indifferently for scalar-valued functions as well as for vector-valued and tensor-valued functions.

Acknowledgments. The idea for this work was conceived during a visit of P.N. at the Weierstrass Institute, Berlin. The kind hospitality of A. Mielke is gratefully

acknowledged. P.N. has also profited from the exchange with M. Costabel regarding Div/Curl systems. P.N. and D.K. both acknowledge the help of inspiring discussions on regularity of nonlinear elliptic problems with J. Frehse.

REFERENCES

[1] H. D. Alber, *Materials with Memory. Initial-Boundary Value Problems for Constitutive Equations with Internal Variables*, Lecture Notes in Math. 1682, Springer, Berlin, 1998.

[2] J. Alberty, C. Carstensen, and D. Zarrabi, *Adaptive numerical analysis in primal elastoplasticity with hardening*, Comput. Methods Appl. Mech. Engrg., 171 (1999), pp. 175–204.

[3] G. Anzellotti and S. Luckhaus, *Dynamical evolution of elasto-perfectly plastic bodies*, Appl. Math. Optim., 15 (1987), pp. 121–140.

[4] J. P. Bardet, *Observations on the effects of particle rotations on the failure of idealized granular materials*, Mech. Mater., 18 (1994), pp. 159–182.

[5] A. Bensoussan and J. Frehse, *Asymptotic behaviour of Norton-Hoff's law in plasticity theory and H^1 regularity*, in Boundary Value Problems for Partial Differential Equations and Applications, J.-L. Lions and C. Baiocchi, eds., RMA Res. Notes Appl. Math. 29, Masson, Paris, 1993, pp. 3–25.

[6] A. Bensoussan and J. Frehse, *Regularity Results for Nonlinear Elliptic Systems and Applications*, Appl. Math. Sci. 151, Springer, Berlin, 2002.

[7] M. Bildhauer and M. Fuchs, *Smoothness of weak solutions of the Ramberg/Osgood equations on plane domains*, ZAMM Z. Angew. Math. Mech., 87 (2007), pp. 70–76.

[8] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*, Springer, Heidelberg, 1994.

[9] G. Capriz, *Continua with Microstructure*, Springer, Heidelberg, 1989.

[10] C. Carstensen and S. Müller, *Local stress regularity in scalar nonconvex variational problems*, SIAM J. Math. Anal., 34 (2002), pp. 495–509.

[11] K. Chełmiński, *Coercive approximation of viscoplasticity and plasticity*, Asymptot. Anal., 26 (2001), pp. 105–133.

[12] K. Chełmiński, *Perfect plasticity as a zero relaxation limit of plasticity with isotropic hardening*, Math. Methods Appl. Sci., 24 (2001), pp. 117–136.

[13] K. Chełmiński, *Global existence of weak-type solutions for models of monotone type in the theory of inelastic deformations*, Math. Methods Appl. Sci., 25 (2002), pp. 1195–1230.

[14] B. D. Coleman and M. L. Hodgdon, *On shear bands in ductile materials*, Arch. Rational Mech. Anal., 90 (1985), pp. 219–247.

[15] E. Cosserat and F. Cosserat, *Théorie des Corps Déformables*, Librairie Scientifique A. Hermann et Fils (Translation: *Theory of Deformable Bodies*, NASA TT F-11 561, Washington, DC, 1968), Paris, 1909.

[16] J. Daněček and E. Viszus, *$\mathcal{L}^{2,\Phi}$ regularity for nonlinear elliptic systems of second order*, Electron. J. Differential Equations, 2002, article 20.

[17] R. de Borst, *A generalization of J_2-flow theory for polar continua*, Comput. Methods Appl. Mech. Engrg., 103 (1992), pp. 347–362.

[18] A. Demyanov, *Regularity in Prandtl–Reuss perfect plasticity*, in Workshop: Analysis and Numerics for Rate-Independent Processes, G. Dal Maso, G. Francfort, A. Mielke, and T. Roubíček, eds., Oberwolfach Report, Vol. 4, Oberwolfach, Germany, 2007, pp. 591–666.

[19] A. Dietsche, P. Steinmann, and K. Willam, *Micropolar elastoplasticity and its role in localization*, Int. J. Plasticity, 9 (1993), pp. 813–831.

[20] C. Ebmeyer, *Mixed boundary value problems for nonlinear elliptic systems with p-structure in polyhedral domains*, Math. Nachr., 236 (2002), pp. 91–108.

[21] C. Ebmeyer and J. Frehse, *Mixed boundary value problems for nonlinear elliptic equations in multidimensional non-smooth domains*, Math. Nachr., 203 (1999), pp. 47–74.

[22] A. C. Eringen, *Microcontinuum Field Theories*, Springer, Heidelberg, 1999.

[23] S. Forest, G. Cailletaud, and R. Sievert, *A Cosserat theory for elastoviscoplastic single crystals at finite deformation*, Arch. Mech., 49 (1997), pp. 705–736.

[24] M. Fuchs and G. Seregin, *Variational Methods for Problems from Plasticity Theory and for Generalized Newtonian Fluids*, Ann. Univ. Sarav. Ser. Math. 10, FB Mathematik, University of Saarbrücken, Saarbrücken, Germany, 1999.

[25] M. Fuchs and G. Seregin, *Variational Methods for Problems from Plasticity Theory and for Generalized Newtonian Fluids*, Lecture Notes in Math. 1749, Springer, Berlin, 2000.

[26] M. Giaquinta and S. Hildebrandt, *Calculus of Variations* I, Springer, Berlin, Heidelberg, 1996.

[27] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Grundlehren Math. Wiss. 224, Springer, Berlin, 1977.
[28] V. Girault and P. A. Raviart, *Finite Element Approximation of the Navier-Stokes Equations*, Lecture Notes in Math. 749, Springer, Heidelberg, 1979.
[29] W. Han and B. D. Reddy, *Plasticity. Mathematical Theory and Numerical Analysis*, Springer, Berlin, 1999.
[30] I. R. Ionescu and M. Sofonea, *Functional and Numerical Methods in Viscoplasticity*, 1st ed., Oxford University Press, Oxford, UK, 1993.
[31] M. M. Iordache and K. Willam, *Localized failure analysis in elastoplastic Cosserat continua*, Comput. Methods Appl. Mech. Engrg., 151 (1998), pp. 559–586.
[32] D. Knees, *Global regularity of the elastic fields of a power-law model on Lipschitz domains*, Math. Methods Appl. Sci., 29 (2006), pp. 1363–1391.
[33] D. Knees, *Global stress regularity of convex and some nonconvex variational problems*, Ann. Mat. Pura Appl. (4), 187 (2008), pp. 157–184.
[34] D. Knees and A. Mielke, *Energy release rate for cracks in finite-strain elasticity*, Math. Methods Appl. Sci., 31 (2008), pp. 501–528.
[35] R. V. Kohn and R. Temam, *Dual spaces of stresses and strains, with applications to Hencky plasticity*, Appl. Math. Optim., 10 (1983), pp. 1–35.
[36] A. Mielke, *Deriving new evolution equations for microstructures via relaxation of variational incremental problems*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 5095–5127.
[37] A. Mielke, *Existence of minimizers in incremental elasto-plasticity with finite strains*, SIAM J. Math. Anal., 36 (2004), pp. 384–404.
[38] A. Mielke and S. Müller, *Lower semicontinuity and existence of minimizers in incremental finite-strain elastoplasticity*, ZAMM Z. Angew. Math. Mech., 86 (2006), pp. 233–250.
[39] C. B. Morrey, *Multiple Integrals in the Calculus of Variations*, Springer, New York, 1966.
[40] H. B. Mühlhaus, *Shear band analysis for granular materials within the framework of Cosserat theory*, Ing. Archiv., 56 (1989), pp. 389–399.
[41] J. Nečas, *Les Méthodes Directes en Théorie des Équations Elliptiques*, Masson et Cie, Éditeurs, Paris, 1967.
[42] J. Nečas, *Introduction to the Theory of Nonlinear Elliptic Equations*, Teubner Verlagsgesellschaft, Leipzig, 1983.
[43] P. Neff, *Existence of minimizers for a finite-strain micromorphic elastic solid*, Proc. Roy. Soc. Edinburgh Sect. A, 136 (2006), pp. 997–1012.
[44] P. Neff, *A finite-strain elastic-plastic Cosserat theory for polycrystals with grain rotations*, Internat. J. Engrg. Sci., 44 (2006), pp. 574–594.
[45] P. Neff, *Finite Multiplicative Elastic-Viscoplastic Cosserat Micropolar Theory for Polycrystals with Grain Rotations. Modelling and Mathematical Analysis*, Preprint 2297; http://wwwbib.mathematik.tu-darmstadt.de/Math-Net/Preprints/Listen/pp03.html.
[46] P. Neff and K. Chełmiński, *Infinitesimal elastic-plastic Cosserat micropolar theory. Modelling and global existence in the rate independent case*, Proc. Roy. Soc. Edinburgh Sect. A, 135 (2005), pp. 1017–1039.
[47] P. Neff and K. Chełmiński, *A note on approximation of Prandtl-Reuss plasticity through Cosserat plasticity*, Quart. Appl. Math., to appear.
[48] P. Neff and K. Chełmiński, *Well-posedness of dynamic Cosserat plasticity*, Appl. Math. Optim., 56 (2007), pp. 19–35.
[49] P. Neff, K. Chełmiński, W. Müller, and C. Wieners, *Numerical solution method for an infinitesimal elastic-plastic Cosserat model*, Math. Models Methods Appl. Sci., 17 (2007), pp. 1211–1239.
[50] P. Neff and C. Wieners, *Comparison of models for finite plasticity. A numerical study*, Comput. Vis. Sci., 6 (2003), pp. 23–35.
[51] S. Nesenenko, *Homogenization and Regularity in Viscoplasticity*. Ph.D. thesis, TU Darmstadt, Logos Verlag, Berlin, 2006.
[52] M. Ortiz and L. Stainier, *The variational formulation of viscoplastic constitutive updates*, Comput. Methods Appl. Mech. Engrg., 171 (1999), pp. 419–444.
[53] J.-P. Raymond, *Régularité globale des solutions de systèmes elliptiques non linéaires*, Rev. Mat. Univ. Complut. Madrid, 2 (1989), pp. 241–270.
[54] S. I. Repin, *Errors of finite element method for perfectly elasto-plastic problems*, Math. Models Methods Appl. Sci., 6 (1996), pp. 587–604.
[55] M. Ristinmaa and M. Vecchi, *Use of couple-stress theory in elasto-plasticity*, Comput. Methods Appl. Mech. Engrg., 136 (1996), pp. 205–224.
[56] C. Sansour, *A theory of the elastic-viscoplastic Cosserat continuum*, Arch. Mech., 50 (1998), pp. 577–597.

[57] G. Savaré, *Regularity results for elliptic equations in Lipschitz domains*, J. Funct. Anal., 152 (1998), pp. 176–201.
[58] G. A. Seregin and T. N. Shilkin, *Regularity for minimizers of some variational problems in plasticity theory*, J. Math. Sci., 99 (2000), pp. 969–988.
[59] G. A. Seregin, *Differentiability of extremals of variational problems in the mechanics of elastic perfectly plastic media*, Differential Equations, 23 (1987), pp. 1349–1358.
[60] J. C. Simo and J. R. Hughes, *Computational Inelasticity*, Interdiscip. Appl. Math. 7, Springer, Berlin, 1998.
[61] J. C. Simo and R. L. Taylor, *Consistent tangent operators for rate-independent elasto-plasticity*, Comput. Methods Appl. Mech. Engrg., 48 (1985), pp. 101–118.
[62] P. Steinmann, *A micropolar theory of finite deformation and finite rotation multiplicative elastoplasticity*, Internat. J. Solids Structures, 31 (1994), pp. 1063–1084.
[63] R. Temam, *Mathematical Problems in Plasticity*, Gauthier–Villars, New York, 1985.
[64] R. Temam, *A generalized Norton-Hoff model and the Prandtl-Reuss law of plasticity*, Arch. Rational Mech. Anal., 95 (1986), pp. 137–183.
[65] T. Valent, *Boundary Value Problems of Finite Elasticity*, Springer, Berlin, 1988.
[66] C. Wieners, *Multigrid methods for Prandtl-Reuss plasticity*, Numer. Linear Algebra Appl., 6 (1999), pp. 457–478.
[67] C. Wieners, *Efficient elasto-plastic simulation*, in Multifield Problems. State of the Art, A. M. Sändig, W. Schiehlen, and W. L. Wendland, eds., Springer, Berlin, 2000, pp. 209–216.
[68] J. Wloka, *Partielle Differentialgleichungen*, Teubner Verlag, Stuttgart, 1982.

SIAM J. MATH. ANAL.
Vol. 40, No. 1, pp. 44–69

GLOBAL EXISTENCE RESULTS AND UNIQUENESS FOR DISLOCATION EQUATIONS*

GUY BARLES†, PIERRE CARDALIAGUET‡, OLIVIER LEY†, AND RÉGIS MONNEAU§

Abstract. We are interested in nonlocal eikonal equations arising in the study of the dynamics of dislocation lines in crystals. For these nonlocal but also nonmonotone equations, only the existence and uniqueness of Lipschitz and local-in-time solutions were available in some particular cases. In this paper, we propose a definition of weak solutions for which we are able to prove the existence for all time. Then we discuss the uniqueness of such solutions in several situations, both in the monotone and the nonmonotone case.

Key words. nonlocal Hamilton–Jacobi equations, dislocation dynamics, nonlocal front propagation, level-set approach, geometrical properties, lower-bound gradient estimate, viscosity solutions, eikonal equation, L^1-dependence in time

AMS subject classifications. 49L25, 35F25, 35A05, 35D05, 35B50, 45G10

DOI. 10.1137/070682083

1. Introduction.

In this article we are interested in the dynamics of defects in crystals, called dislocations. The dynamics of these dislocations is the main microscopic explanation of the macroscopic behavior of metallic crystals (see, for instance, the physical monographs of Nabarro [24], Hirth and Lothe [19], or Lardner [21] for a mathematical presentation). A dislocation is a line moving in a crystallographic plane, called a slip plane. The typical length of such a dislocation line is of the order of 10^{-6} m. Its dynamics is given by a normal velocity proportional to the Peach–Koehler force acting on this line.

This Peach–Koehler force may have two possible contributions: the first one is the self-force created by the elastic field generated by the dislocation line itself (i.e., this self-force is a nonlocal function of the shape of the dislocation line); the second one is the force created by everything exterior to the dislocation line, such as the exterior stress applied on the material, or the force created by other defects. In this paper, we study a particular model introduced in Rodney, Le Bouar, and Finel [27].

More precisely, if, at time t, the dislocation line is the boundary of an open set $\Omega_t \subset \mathbb{R}^N$ with $N = 2$ for the physical application, the normal velocity to the set Ω_t is given by

$$V_n = c_0 \star \mathbb{1}_{\overline{\Omega}_t} + c_1, \tag{1}$$

where $\mathbb{1}_{\overline{\Omega}_t}(x)$ is the indicator function of the set $\overline{\Omega}_t$, which is equal to 1 if $x \in \overline{\Omega}_t$ and equal to 0 otherwise. The function $c_0(x,t)$ is a kernel which depends only on the physical properties of the crystal and on the choice of the dislocation line, whose

*Received by the editors February 6, 2007; accepted for publication (in revised form) November 16, 2007; published electronically March 26, 2008. This work was supported by contracts ACI JC 1025 (2003-2005) and ACI JC 1041 (2002-2004) of the French Ministry of Research.
http://www.siam.org/journals/sima/40-1/68208.html
†Laboratoire de Mathématiques et Physique Théorique, Fédération Denis Poisson, Université de Tours, Parc de Grandmont, 37200 Tours, France (barles@lmpt.univ-tours.fr, ley@lmpt.univ-tours.fr).
‡Université de Bretagne Occidentale, UFR des Sciences et Techniques, 6 Av. Le Gorgeu, BP 809, 29285 Brest, France (pierre.cardaliaguet@univ-brest.fr).
§CERMICS, Ecole Nationale des Ponts et Chaussées, 6 et 8 avenue Blaise Pascal, Cité Descartes, Champs-sur-Marne, 77455 Marne-la-Vallée Cedex 2, France (monneau@cermics.enpc.fr).

evolution we follow. In the special case of application to dislocations, the kernel c_0 does not depend on time, but to keep a general setting we allow here a dependence on the time variable. Here $\star$ denotes the convolution is space, namely

$$(c_0(\cdot,t) \star \mathbb{1}_{\overline{\Omega}_t})(x) = \int_{\mathbb{R}^N} c_0(x-y,t)\mathbb{1}_{\overline{\Omega}_t}(y)dy, \tag{2}$$

and this term appears to be the Peach–Koehler self-force created by the dislocation itself, while $c_1(x,t)$ is an additional contribution to the velocity, created by everything exterior to the dislocation line. We refer the reader to Alvarez et al. [3] for a detailed presentation and a derivation of this model.

We proceed as in the level-set approach to derive an equation for the dislocation line. We replace the evolution of a set Ω_t (the strong solution) by the evolution of a function u such that $\Omega_t = \{u(\cdot,t) > 0\}$. Roughly speaking the dislocation line is represented by the zero level-set of the function u which solves the following equation:

$$\begin{cases} \dfrac{\partial u}{\partial t} = (c_0(\cdot,t) \star \mathbb{1}_{\{u(\cdot,t)\geq 0\}}(x) + c_1(x,t))|Du| & \text{in } \mathbb{R}^N \times (0,T), \\ u(\cdot,0) = u_0 \quad \text{in } \mathbb{R}^N, \end{cases} \tag{3}$$

where (2) now reads

$$c_0(\cdot,t) \star \mathbb{1}_{\{u(\cdot,t)\geq 0\}}(x) = \int_{\mathbb{R}^N} c_0(x-y,t)\mathbb{1}_{\{u(\cdot,t)\geq 0\}}(y)dy. \tag{4}$$

Note that (3) is not really a level-set equation since it is not invariant under non-decreasing changes of functions $u \to \varphi(u)$, where φ is nondecreasing. As noticed by Slepčev [28], the natural level-set equation should be (11); see section 1.2.

Although (3) seems very simple, there are only a few known results. Under suitable assumptions on the initial data and on c_0, c_1, the existence and uniqueness of the solution is known in two particular cases: either for short time (see [3]), or for all time under the additional assumption that $V_n \geq 0$, which is, for instance, always satisfied for c_1 satisfying $c_1(x,t) \geq |c_0(\cdot,t)|_{L^1(\mathbb{R}^N)}$ (see [2, 12, 5] for a level-set formulation).

In the general case, the existence for all time of solutions to (3) is not known and, in particular, in the case when the kernel c_0 has negative values; indeed, in this case, the front propagation problem (3) does not satisfy any monotonicity property (preservation of inclusions), and therefore, even if a level-set-type equation can be derived, viscosity solution theory cannot be readily used. At this point, it is worth pointing out that a key property in the level-set approach is the comparison principle for viscosity solutions which is almost equivalent to this monotonicity property (see, for instance, Giga's monograph [18]). On the other hand, one may try to partly use viscosity solution theory together with some other approximation and/or compactness arguments to prove at least the existence of weak solutions (in a suitable sense). But here also the bad sign of the kernel creates difficulties since one cannot readily use the classical half-relaxed limits techniques to pass to the limit in the approximate problems. Additional arguments are needed to obtain weak solutions.

The aim of this paper is to describe a general approach of these dislocations' dynamics, based on the level-set approach, which allows us to introduce a suitable notion of weak solutions, to prove the existence of these weak solutions for all time, and to analyze the uniqueness (or nonuniqueness) of these solutions.

1.1. Weak solutions of the dislocation equation. We introduce the following definition of weak solutions, which itself uses the definition of L^1-viscosity solutions, recalled in Appendix A.

DEFINITION 1.1 (classical and weak solutions). *For any $T > 0$, we say that a function $u \in W^{1,\infty}(\mathbb{R}^N \times [0,T))$ is a weak solution of* (3) *on the time interval $[0,T)$ if there is some measurable map $\chi : \mathbb{R}^N \times (0,T) \to [0,1]$ such that u is an L^1-viscosity solution of*

$$\begin{cases} \dfrac{\partial u}{\partial t} = \bar{c}(x,t)|Du| \quad \text{in } \mathbb{R}^N \times (0,T), \\ u(\cdot,0) = u_0 \quad \text{in } \mathbb{R}^N, \end{cases} \tag{5}$$

where

$$\bar{c}(x,t) = c_0(\cdot,t) \star \chi(\cdot,t)(x) + c_1(x,t) \tag{6}$$

and

$$\mathbb{1}_{\{u(\cdot,t)>0\}}(x) \;\le\; \chi(x,t) \;\le \mathbb{1}_{\{u(\cdot,t)\ge 0\}}(x) \tag{7}$$

for almost all $(x,t) \in \mathbb{R}^N \times [0,T]$. We say that u is a classical solution of (3) *if u is a weak solution to* (5) *and if*

$$\mathbb{1}_{\{u(\cdot,t)>0\}}(x) = \mathbb{1}_{\{u(\cdot,t)\ge 0\}}(x) \tag{8}$$

for almost all $(x,t) \in \mathbb{R}^N \times [0,T]$.

Note that we have $\chi(x,t) = \mathbb{1}_{\{u(\cdot,t)>0\}}(x) = \mathbb{1}_{\{u(\cdot,t)\ge 0\}}(x)$ for almost all $(x,t) \in \mathbb{R}^N \times [0,T]$ for classical solutions.

To state our first existence result, we use the following assumptions.

(H0) $u_0 \in W^{1,\infty}(\mathbb{R}^N)$, $-1 \le u_0 \le 1$, and there exists $R_0 > 0$ such that $u_0(x) \equiv -1$ for $|x| \ge R_0$,

(H1) $c_0 \in C([0,T); L^1(\mathbb{R}^N))$, $D_x c_0 \in L^\infty([0,T); L^1(\mathbb{R}^N))$, $c_1 \in C(\mathbb{R}^N \times [0,T))$, and there exist constants M_1, L_1 such that, for any $x, y \in \mathbb{R}^N$ and $t \in [0,T]$,

$$|c_1(x,t)| \le M_1 \quad \text{and} \quad |c_1(x,t) - c_1(y,t)| \le L_1|x-y|.$$

In what follows, we denote by M_0, L_0 constants such that, for any (or almost every) $t \in [0,T)$, we have

$$|c_0(\cdot,t)|_{L^1(\mathbb{R}^N)} \le M_0 \quad \text{and} \quad |D_x c_0(\cdot,t)|_{L^1(\mathbb{R}^N)} \le L_0.$$

Our first main result is the following.

THEOREM 1.2 (existence of weak solutions). *Under assumptions* (H0)–(H1)*, for any $T > 0$ and for any initial data u_0, there exists a weak solution of* (3) *on the time interval $[0,T)$ in the sense of Definition* 1.1.

Our second main result states that a weak solution is a classical one if the evolving set is expanding and if the following additional condition is fulfilled

(H2) c_1 and c_0 satisfy (H1), and there exist constants m_0, N_1 and a positive function $N_0 \in L^1(\mathbb{R}^N)$ such that, for any $x, h \in \mathbb{R}^N$, $t \in [0,T)$, we have

$$|c_0(x,t)| \le m_0,$$

$$|c_1(x+h,t) + c_1(x-h,t) - 2c_1(x,t)| \le N_1|h|^2,$$

$$|c_0(x+h,t) + c_0(x-h,t) - 2c_0(x,t)| \le N_0(x)|h|^2.$$

THEOREM 1.3 (some links between weak solutions and classical continuous viscosity solutions and uniqueness results). *Assume* (H0)–(H1)*, and suppose that there is some* $\delta \geq 0$ *such that, for all measurable maps* $\chi : \mathbb{R}^N \times (0,T) \to [0,1]$,

$$\text{for all } (x,t) \in \mathbb{R}^N \times [0,T], \quad c_0(\cdot,t) \star \chi(\cdot,t)(x) + c_1(x,t) \geq \delta, \tag{9}$$

and that the initial data u_0 *satisfies (in the viscosity sense)*

$$-|u_0| - |Du_0| \leq -\eta_0 \quad \text{in } \mathbb{R}^N \tag{10}$$

for some $\eta_0 > 0$*. Then any weak solution* u *of* (3) *in the sense of Definition* 1.1 *is a classical continuous viscosity solution of* (3)*. This solution is unique if* (H2) *holds and*

(i) *either* $\delta > 0$,

(ii) *or* $\delta = 0$ *and* u_0 *is semiconvex, i.e., satisfies for some constant* $C > 0$

$$u_0(x+h) + u_0(x-h) - 2u_0(x) \geq -C|h|^2 \quad \text{for all } x, h \in \mathbb{R}^N.$$

Assumption (9) ensures that the velocity V_n in (1) is positive for positive δ. Of course, we can state similar results in the case of negative velocity. Assumption (10) means that u_0 is a viscosity subsolution of $-|v(x)| - |Dv(x)| + \eta_0 \leq 0$. When u_0 is C^1, it follows that the gradient of u_0 does not vanish on the set $\{u_0 = 0\}$ (see [22] for details). Point (ii) of the theorem is the main result of [2, 5]. We also point out that, with adapted proofs, only a bound from below could be required in (H2) on $c_1(x+h,t) + c_1(x-h,t) - 2c_1(x,t)$ and $c_0(x+h,t) + c_0(x-h,t) - 2c_0(x,t)$.

Remark 1.1. In particular Theorem 1.3 implies uniqueness in the case $c_0 \geq 0$ and $c_1 \equiv 0$. The general study of nonnegative kernels is provided below (see Theorem 1.5 and Remark 1.3).

1.2. Nonnegative kernel $c_0 \geq 0$. In the special case where the kernel c_0 is nonnegative, an inclusion principle for the dislocation lines, or equivalently a comparison principle for the functions of the level-set formulations, is expected (cf. Cardaliaguet [11] and Slepčev [28]).

Moreover, in the classical level-set approach, all the level-sets of u should have the same type of normal velocity, and Slepčev [28] remarked that a formulation with a nonlocal term of the form $\{u(\cdot,t) \geq u(x,t)\}$ is more appropriate. Therefore it is natural to start studying the following equation (which replaces (3)):

$$\begin{cases} \dfrac{\partial u}{\partial t} = (c_0(\cdot,t) \star \mathbb{1}_{\{u(\cdot,t) \geq u(x,t)\}}(x) + c_1(x,t))|Du| & \text{in } \mathbb{R}^N \times (0,T)\,, \\ u(\cdot,0) = u_0 & \text{in } \mathbb{R}^N\,, \end{cases} \tag{11}$$

where $\star$ denotes the convolution in space as in (4).

The precise meaning of a viscosity solution of (11) is given in Definition 5.1.

In this context, assumption (H0) can be weakened into the following condition, which allows us to consider unbounded evolving sets:

(H0′) $u_0 \in BUC(\mathbb{R}^N)$.

Our main result for this equation is the following.

THEOREM 1.4 (existence and uniqueness). *Assume that* $c_0 \geq 0$ *on* $\mathbb{R}^N \times [0,T]$ *and that* (H0′)–(H1) *hold. Then there exists a unique viscosity solution* u *of* (11).

Remark 1.2. The comparison principle for this equation (see Theorem 5.2) is a generalization of [28, Theorem 2.3]: indeed, in [28], everything takes place in a fixed

bounded set, whereas here one has to deal with unbounded sets. See also [15] for related results.

Now we turn to the connections with weak solutions. To do so, if u is the unique continuous solution of (11) given by Theorem 1.4, we introduce the functions $\rho^+, \rho^- : \mathbb{R}^N \times [0,T] \to \mathbb{R}$ defined by

$$\rho^+ := \mathbb{1}_{\{u \geq 0\}} \quad \text{and} \quad \rho^- := \mathbb{1}_{\{u > 0\}} .$$

Our result is the following.

THEOREM 1.5 (maximal and minimal weak solutions). *Under the assumptions of Theorem* 1.4, *the maximal and minimal weak solutions of* (3) *are the continuous functions v^+, v^- which are the unique L^1-viscosity solutions of the equations*

$$\begin{cases} \dfrac{\partial v^\pm}{\partial t} = c[\rho^\pm](x,t)|Dv^\pm| & \text{in } \mathbb{R}^N \times (0,T), \\ v^\pm(x,0) = u_0(x) & \text{in } \mathbb{R}^N, \end{cases} \tag{12}$$

where

$$c[\rho](x,t) := c_0(\cdot,t) \star \rho(\cdot,t)(x) + c_1(x,t) \quad \text{in} \quad \mathbb{R}^N \times (0,T).$$

The functions $v^\pm$ satisfy $\{v^+(\cdot,t) \geq 0\} = \{u(\cdot,t) \geq 0\}$ and $\{v^-(\cdot,t) > 0\} = \{u(\cdot,t) > 0\}$, where u is the solution of (11).

Moreover, if the set $\{u(\cdot,t) = 0\}$ has a zero-Lebesgue measure for almost all $t \in (0,T)$, then problem (3) *has a unique weak solution which is also a classical one.*

Remark 1.3. 1. Theorem 1.5 shows that, in the case when $c_0 \geq 0$, Slepčev's approach allows us to identify the maximal and minimal weak solutions as being associated with $\rho^\pm$.

2. Equalities $\{v^-(\cdot,t) \geq 0\} = \{u(\cdot,t) \geq 0\}$ and $\{v^+(\cdot,t) > 0\} = \{u(\cdot,t) > 0\}$ do not hold in general (see, for instance, Example 3.1 in section 3).

3. If the set $\{v^\pm(\cdot,t) = 0\}$ develops an interior, a dramatic loss of uniqueness for the weak solution of (3) may occur. This is illustrated by Example 3.1 below, where we are able to build infinitely many solutions after the onset of fattening.

4. We have uniqueness for (3) if $\{u(\cdot,t) = 0\}$ has a zero-Lebesgue measure for almost all $t \in (0,T)$. This condition is fulfilled when, for instance, $c[\rho] \geq 0$ holds for any indicator function ρ and (10) holds (see also Remark 1.1).

1.3. Organization of the paper. In section 2, we recall basic results for the classical eikonal equation which are used throughout the paper. In section 3, we prove the existence of weak solutions for (3), namely Theorem 1.2, and give a counterexample to the uniqueness in general. Let us mention that this first part of the paper, even if it requires rather deep results of viscosity solution theory, is of a general interest for a wide audience and can be read without having an expertise in this theory since one just needs to apply the results which, anyway, are rather natural. In section 4, we prove Theorem 1.3 in the case of expanding dislocations. The arguments we use here are far more involved from a technical point of view: in particular we need some fine estimates of the perimeter of the evolving sets. In section 5, we study the Slepčev formulation in the case of nonnegative kernels and prove Theorems 1.4 and 1.5. In spirit, this section is closely related to the classical level-set approach but is more technical. Finally, for sake of completeness, we recall in Appendix A the definition of L^1-viscosity solutions and a new stability result proved by Barles in [4].

2. Some basic results for the classical (local) eikonal equation. We want to recall in this section some basic results on the level-set equation

$$\begin{cases} \dfrac{\partial v}{\partial t} = a(x,t)|Dv| & \text{in } \mathbb{R}^N \times (0,T), \\ v(\cdot,0) = u_0 & \text{in } \mathbb{R}^N, \end{cases} \tag{13}$$

where $T > 0$ and $a : \mathbb{R}^N \times [0,T] \to \mathbb{R}$ is, at least, a continuous function.

We provide some classical estimates on the solutions to (13) when a satisfies suitable assumptions. Our result is the following.

THEOREM 2.1. *If u_0 satisfies* (H0) *and a satisfies the assumptions of c_1 in* (H1)*, then* (13) *has a unique continuous solution v which is Lipschitz continuous in $\mathbb{R}^N \times [0,T]$ and which satisfies*

(i) $-1 \le v \le 1$ *in* $\mathbb{R}^N \times (0,T)$*,* $v(x,t) \equiv -1$ *for* $|x| \ge R_0 + M_1 t$*,*

(ii) $|Dv(\cdot,t)|_\infty \le |Du_0|_\infty \, \mathrm{e}^{L_1 t}$*,*

(iii) $|v_t(\cdot,t)|_\infty \le M_1 |Du_0|_\infty \, \mathrm{e}^{L_1 t}$*.*

We skip the very classical proof of Theorem 2.1; we just point out that the first point comes from the comparison result for (13) and the "finite speed of propagation property" (see Crandall and Lions [14]), while the second one is a basic gradient estimate (see, for example, Ley [22]), and the last one comes directly from the fact that the equation is satisfied almost everywhere.

The main consequence of this result is that the solution remains in a compact subset of the Banach space $(C(\mathbb{R}^N \times [0,T]), |\cdot|_\infty)$ as long as u_0 and a satisfy (H0)–(H1) with fixed constants.

Let us introduce the following.

DEFINITION 2.2 (interior ball property). *We say that a closed set $K \subset \mathbb{R}^N$ has an interior ball property of radius $r > 0$ if, for any $x \in K$, there exists $p \in \mathbb{R}^N \setminus \{0\}$ such that $B(x - r\frac{p}{|p|}, r) \subset K$.*

We will also use the following result, due to Cannarsa and Frankowska [10], the proof of which is given in Appendix B for the sake of completeness.

LEMMA 2.3 (interior ball regularization). *Suppose* (H0) *and that a satisfies the assumptions of c_1 in* (H1)–(H2) *and there exists a constant $\delta > 0$ such that*

$$c_1 \ge \delta > 0 \quad \text{on} \quad \mathbb{R}^N \times [0,T].$$

Then there exists a constant γ (depending in particular on $\delta > 0$ and T and on the other constants of the problem) such that for the solution v of (13)*, the set $\{v(\cdot,t) \ge 0\}$ has an interior ball property of radius $r_t \ge \gamma t$ for $t \in (0,T)$.*

3. Existence of weak solutions for (3). We aim to solve (3), i.e.,

$$\begin{cases} \dfrac{\partial u}{\partial t} = (c_0(\cdot,t) \star \mathbb{1}_{\{u(\cdot,t)\ge 0\}}(x) + c_1(x,t))|Du| & \text{in } \mathbb{R}^N \times (0,T), \\ u(\cdot,0) = u_0 & \text{in } \mathbb{R}^N, \end{cases}$$

proving Theorem 1.2, which states the existence of weak solutions as introduced in Definition 1.1.

A key difficulty in solving (3) comes from the fact that, in this kind of level-set equation, one may face the so-called nonempty interior difficulty, i.e., that the 0–level-set of the solution is "fat," which may mean that it has either a nonempty interior or a nonzero Lebesgue measure. Clearly, in both cases, $\mathbb{1}_{\{u(\cdot,t)\ge 0\}}$ is different from

$\mathbb{1}_{\{u(\cdot,t)>0\}}$, and this leads to rather bad stability properties for (3) and therefore to difficulties in proving the existence of a solution (and even more for the uniqueness). The notion of weak solution (5)–(7) emphasizes this difficulty. On the contrary, if $\bar{c}(x,t) \geq 0$ in $\mathbb{R}^N \times [0,T]$, it is known that the "nonempty interior difficulty" cannot happen (see Barles, Soner, and Souganidis [6] and Ley [22]), and we recover a more classical formulation. We discuss this question in the next section as well as some uniqueness issues for our weak solutions. Let us finally note that weak solutions for (3) satisfy the following inequalities.

PROPOSITION 3.1. *Let u be a weak solution to* (3). *Then u also satisfies in the L^1-sense*

$$\frac{\partial u}{\partial t} \leq \left(c_0^+(\cdot,t) \star \mathbb{1}_{\{u(\cdot,t)\geq 0\}}(x) - c_0^-(\cdot,t) \star \mathbb{1}_{\{u(\cdot,t)>0\}}(x) + c_1(x,t)\right)|Du|, \tag{14}$$

$$\frac{\partial u}{\partial t} \geq \left(c_0^+(\cdot,t) \star \mathbb{1}_{\{u(\cdot,t)>0\}}(x) - c_0^-(\cdot,t) \star \mathbb{1}_{\{u(\cdot,t)\geq 0\}}(x) + c_1(x,t)\right)|Du| \tag{15}$$

in $\mathbb{R}^N \times (0,T)$, where $c_0^+ = \max(0,c_0)$ and $c_0^- = \max(0,-c_0)$.

Proof of Proposition 3.1. Let $\bar{c}$ be associated with u as in (5)–(7). Then we have

$$\bar{c}(x,t) \geq c_0^+(\cdot,t) \star \mathbb{1}_{\{u(\cdot,t)>0\}}(x) - c_0^-(\cdot,t) \star \mathbb{1}_{\{u(\cdot,t)\geq 0\}}(x) + c_1(x,t)$$

for every $x \in \mathbb{R}^N$ and almost every $t \in (0,T)$. We note that the right-hand side of the inequality is lower semicontinuous. Following Lions and Perthame [23], u then solves (15) in the usual viscosity sense. The proof of (14) can be achieved in a similar way. □

Proof of Theorem 1.2. 1. *Introduction of a perturbated equation.* First we are going to solve the equation

$$\frac{\partial u}{\partial t} = \left(c_0(\cdot,t) \star \psi_\varepsilon(u(\cdot,t))(x) + c_1(x,t)\right)|Du| \quad \text{in } \mathbb{R}^N \times (0,T)\,, \tag{16}$$

where $\psi_\varepsilon : \mathbb{R} \to \mathbb{R}$ is a sequence of continuous functions such that $\psi_\varepsilon(t) \equiv 0$ for $t \leq -\varepsilon$, $\psi_\varepsilon(t) \equiv 1$ for $t \geq 0$, and ψ_ε is an affine function on $[-\varepsilon, 0]$.

We aim at applying Schauder's fixed point theorem to a suitable map. We note that an alternative proof could be given by using techniques developed by Alibaud in [1].

2. *Definition of a map $\mathcal{T}$.* We introduce the convex and compact (by Ascoli's theorem) subset

$$X = \{u \in C(\mathbb{R}^N \times [0,T]) : u \equiv -1 \text{ in } \mathbb{R}^N \backslash B(0, R_0 + MT),$$
$$|Du|, |u_t|/M \leq |Du_0|_\infty e^{LT}\}$$

of $(C(\mathbb{R}^N \times [0,T]), |\cdot|_\infty)$, for $M = M_0 + M_1$ and $L = L_0 + L_1$, and the map $\mathcal{T} : X \to X$ defined as follows: if $u \in C(\mathbb{R}^N \times [0,T])$, then $\mathcal{T}(u)$ is the unique solution v of (13) for

$$c_\varepsilon(x,t) = c_0(\cdot,t) \star \psi_\varepsilon(u(\cdot,t))(x) + c_1(x,t)$$
$$= \int_{\mathbb{R}^N} c_0(x-z,t)\psi_\varepsilon(u(z,t))dz + c_1(x,t).$$

This definition is justified by the fact that, under assumption (H1) on c_1 and c_0, c_ε satisfies (H1) with fixed constants $M = M_0 + M_1$ and $L = L_0 + L_1$; indeed M is a bound on $\sup_{[0,T]} |c_0(\cdot,t)|_{L^1} + M_1$, while L is estimated by the following calculation: for all $x, y \in \mathbb{R}^N$, $t \in [0,T]$, and $u \in X$, we have

$$
\begin{aligned}
(17)\qquad & c_\varepsilon(x,t) - c_\varepsilon(y,t) \\
&= \int_{\mathbb{R}^N} (c_0(x-z,t) - c_0(y-z,t))\psi_\varepsilon(u(z,t))dz + c_1(x,t) - c_1(y,t) \\
&\le \int_{\mathbb{R}^N} |c_0(x-z,t) - c_0(y-z,t)|dz + |c_1(x,t) - c_1(y,t)| \\
&\le (L_0 + L_1)|x-y|,
\end{aligned}
$$

since $0 \le \psi_\varepsilon \le 1$.

Finally, under assumptions (H0)–(H1), for any $u \in X$, the results of Theorem 2.1 apply to (16), which imply that $\mathcal{T}(u) \in X$. It follows that $\mathcal{T}$ is well defined.

3. *Application of Schauder's fixed point theorem to* $\mathcal{T}$. The map $\mathcal{T}$ is continuous since ψ_ε is continuous by using the classical stability result for viscosity solutions (see, for instance, (30) in section 4). Therefore $\mathcal{T}$ has a fixed point u_ε which is bounded in $W^{1,\infty}(\mathbb{R}^N \times [0,T])$ uniformly with respect to ε (since M and L are independent of ε).

4. *Convergence of the fixed point when* $\varepsilon \to 0$. From Ascoli's theorem, we extract a subsequence $(u_{\varepsilon'})_{\varepsilon'}$ which converges locally uniformly to a function denoted by u (in fact globally since the $u_{\varepsilon'}$ are equal to -1 outside a fixed compact subset).

The functions $\chi_{\varepsilon'} := \psi_{\varepsilon'}(u_{\varepsilon'})$ satisfy $0 \le \chi_{\varepsilon'} \le 1$. Therefore we can extract a subsequence—still denoted $(\chi_{\varepsilon'})$—which converges weakly$-*$ in $L^\infty_{\mathrm{loc}}(\mathbb{R}^N \times [0,T])$ to some function $\chi : \mathbb{R}^N \times (0,T) \to [0,1]$. Therefore, for all $\varphi \in L^1_{\mathrm{loc}}(\mathbb{R}^N \times [0,T])$,

$$
(18)\qquad \int_0^T \int_{\mathbb{R}^N} \varphi \chi_{\varepsilon'} dxdt \to \int_0^T \int_{\mathbb{R}^N} \varphi \chi dxdt.
$$

From Fatou's lemma, if φ is nonnegative, it follows that

$$
\begin{aligned}
\int_0^T \int_{\mathbb{R}^N} \varphi(x,t)\chi(x,t)dxdt &\le \int_0^T \int_{\mathbb{R}^N} \varphi(x,t) \limsup_{\varepsilon' \to 0} \chi_{\varepsilon'}(x,t)dxdt \\
&\le \int_0^T \int_{\mathbb{R}^N} \varphi(x,t) \limsup_{\varepsilon' \to 0, x' \to x, t' \to t} \chi_{\varepsilon'}(x',t')dxdt \\
&\le \int_0^T \int_{\mathbb{R}^N} \varphi(x,t) \mathbb{1}_{\{u(\cdot,t) \ge 0\}}(x)dxdt.
\end{aligned}
$$

Since the previous inequalities hold for any nonnegative $\varphi \in L^1_{\mathrm{loc}}(\mathbb{R}^N \times [0,T])$, we obtain that, for almost every $(x,t) \in \mathbb{R}^N \times (0,T)$,

$$
\chi(x,t) \le \mathbb{1}_{\{u(\cdot,t) \ge 0\}}(x).
$$

Similarly we get

$$
\mathbb{1}_{\{u(\cdot,t) > 0\}}(x) \le \chi(x,t).
$$

Furthermore, setting $c_{\varepsilon'} = c_0 \star \chi_{\varepsilon'} + c_1$, from (18), we have, for all $(x,t) \in \mathbb{R}^N \times [0,T]$,

$$\begin{aligned}\int_0^t c_{\varepsilon'}(x,s)ds &= \int_0^t \int_{\mathbb{R}^N} c_0(x-y,s)\chi_{\varepsilon'}(y,s)dyds + \int_0^t c_1(x,s)ds \\ &\to \int_0^t \bar{c}(x,s)ds,\end{aligned}$$

where $\bar{c}(x,t) = c_0(\cdot,t) \star \chi(\cdot,t)(x) + c_1(x,t)$. The above convergence is pointwise, but, noticing that $c_{\varepsilon'}$ satisfies (H3) (with $M := M_0 + M_1$ and $L := L_0 + L_1$) and using Remark A.1, we can apply the stability theorem (Theorem A.3) given in Appendix A. We obtain that u is an L^1-viscosity solution to (5) with $\bar{c}$ satisfying (6)–(7). ∎

The following example is inspired from [6].

Example 3.1 (counterexample to the uniqueness of weak solutions). Let us consider, in dimension $N = 1$, the following equation of type (3),

$$\begin{cases} \dfrac{\partial U}{\partial t} = (1 \star \mathbb{1}_{\{U(\cdot,t)\geq 0\}}(x) + c_1(t))|DU| & \text{in } \mathbb{R} \times (0,2], \\ U(\cdot,0) = u_0 & \text{in } \mathbb{R}, \end{cases} \tag{19}$$

where we set $c_0(x,t) := 1$, $c_1(x,t) := c_1(t) = 2(t-1)(2-t)$, and $u_0(x) = 1 - |x|$. Note that $1 \star \mathbb{1}_A = \mathcal{L}^1(A)$ for any measurable set $A \subset \mathbb{R}$, where $\mathcal{L}^1(A)$ is the Lebesgue measure on $\mathbb{R}$.

We start by solving auxiliary problems for time in $[0,1]$ and $[1,2]$ in order to produce a family of solutions for the original problem in $[0,2]$.

1. *Construction of a solution for* $0 \leq t \leq 1$. The function $x_1(t) = (t-1)^2$ is the solution of the ODE

$$\dot{x}_1(t) = c_1(t) + 2x_1(t) \text{ for } 0 \leq t \leq 1, \quad \text{and } x(0) = 1$$

(note that $\dot{x}_1 \leq 0$ in $[0,1]$). Consider

$$\begin{cases} \dfrac{\partial u}{\partial t} = \dot{x}_1(t) \left|\dfrac{\partial u}{\partial x}\right| & \text{in } \mathbb{R} \times (0,1], \\ u(\cdot,0) = u_0 & \text{in } \mathbb{R}. \end{cases} \tag{20}$$

There exists a unique continuous viscosity solution u of (20). Looking for u under the form $u(x,t) = v(x,\Gamma(t))$ with $\Gamma(0) = 0$, we obtain that v satisfies

$$\frac{\partial v}{\partial t}\,\dot{\Gamma}(t) = \dot{x}_1(t) \left|\frac{\partial v}{\partial x}\right|.$$

Choosing $\Gamma(t) = -x_1(t) + 1$, we get that v is the solution of

$$\begin{cases} \dfrac{\partial v}{\partial t} = -\left|\dfrac{\partial v}{\partial x}\right| & \text{in } \mathbb{R} \times (0,1], \\ v(\cdot,0) = u_0 & \text{in } \mathbb{R}. \end{cases}$$

By the Oleinik–Lax formula, $v(x,t) = \inf_{|x-y|\leq t} u_0(y)$. Since u_0 is even, we have, for all $(x,t) \in \mathbb{R} \times [0,1]$,

$$u(x,t) = \inf_{|x-y|\leq \Gamma(t)} u_0(y) = u_0(|x| + \Gamma(t)) = u_0(|x| - x_1(t) + 1).$$

Therefore, for $0 \le t \le 1$,

$$\{u(\cdot,t) > 0\} = (-x_1(t), x_1(t)) \quad \text{and} \quad \{u(\cdot,t) \ge 0\} = [-x_1(t), x_1(t)]. \tag{21}$$

We will see in step 3 that u is a solution of (19) in $[0,1]$.

2. *Construction of solutions for* $1 \le t \le 2$. Consider now, for any measurable function $0 \le \gamma(t) \le 1$, the unique solution y_γ of the ODE

$$\dot{y}_\gamma(t) = c_1(t) + 2\gamma(t) y_\gamma(t) \text{ for } 1 \le t \le 2, \quad \text{and } y_\gamma(1) = 0. \tag{22}$$

By comparison, we have $0 \le y_0(t) \le y_\gamma(t) \le y_1(t)$ for $1 \le t \le 2$, where y_0, y_1 are the solutions of (22) obtained with $\gamma(t) \equiv 0, 1$. In particular, it follows that $\dot{y}_\gamma \ge 0$ in $[1,2]$. Consider

$$\begin{cases} \dfrac{\partial u_\gamma}{\partial t} = \dot{y}_\gamma(t) \left| \dfrac{\partial u_\gamma}{\partial x} \right| & \text{in } \mathbb{R} \times (1,2], \\ u_\gamma(\cdot,1) = u(\cdot,1) & \text{in } \mathbb{R}, \end{cases}$$

where u is the solution of (20). Again, this problem has a unique continuous viscosity solution u_γ, and setting $\Gamma_\gamma(t) = y_\gamma(t) \ge 0$ for $1 \le t \le 2$, we obtain that v_γ defined by $v_\gamma(x, \Gamma_\gamma(t)) = u_\gamma(x,t)$ is the unique continuous viscosity solution of

$$\begin{cases} \dfrac{\partial v_\gamma}{\partial t} = \left| \dfrac{\partial v_\gamma}{\partial x} \right| & \text{in } \mathbb{R} \times (0, \Gamma_\gamma(2)], \\ v_\gamma(\cdot,0) = u(\cdot,1) & \text{in } \mathbb{R}. \end{cases}$$

Therefore, for all $(x,t) \in \mathbb{R} \times [1,2]$, we have

$$u_\gamma(x,t) = \sup_{|x-y| \le y_\gamma(t)} u(y,1) = \begin{cases} 0 & \text{if } |x| \le y_\gamma(t), \\ u(|x| - y_\gamma(t), 1) & \text{otherwise.} \end{cases}$$

(Note that $u(-x,t) = u(x,t)$ since u_0 is even and, since $u(\cdot,1) \le 0$, by the maximum principle, we have $u_\gamma \le 0$ in $\mathbb{R} \times [1,2]$.) It follows that, for all $1 \le t \le 2$,

$$\{u_\gamma(\cdot,t) > 0\} = \emptyset \quad \text{and} \quad \{u_\gamma(\cdot,t) \ge 0\} = \{u_\gamma(\cdot,t) = 0\} = [-y_\gamma(t), y_\gamma(t)]. \tag{23}$$

3. *There are several weak solutions of* (19). Set, for $0 \le \gamma(t) \le 1$,

$$\begin{aligned} &c_\gamma(t) = c_1(t) + 2x_1(t), && U_\gamma(x,t) = u(x,t) && \text{if } (x,t) \in \mathbb{R} \times [0,1], \\ &c_\gamma(t) = c_1(t) + 2\gamma(t) y_\gamma(t), && U_\gamma(x,t) = u_\gamma(x,t) && \text{if } (x,t) \in \mathbb{R} \times [1,2]. \end{aligned}$$

Then, from steps 1 and 2, U_γ is the unique continuous viscosity solution of

$$\begin{cases} \dfrac{\partial U_\gamma}{\partial t} = c_\gamma(t) \left| \dfrac{\partial U_\gamma}{\partial x} \right| & \text{in } \mathbb{R} \times (0,2], \\ U_\gamma(\cdot,0) = u_0 & \text{in } \mathbb{R}. \end{cases} \tag{24}$$

Taking $\chi_\gamma(\cdot,t) = \gamma(t) \mathbb{1}_{[-y_\gamma(t), y_\gamma(t)]}$ for $1 \le t \le 2$, from (21) and (23), we have

$$\mathbb{1}_{\{U_\gamma(\cdot,t) > 0\}} \le \chi_\gamma(\cdot,t) \le \mathbb{1}_{\{U_\gamma(\cdot,t) \ge 0\}}$$

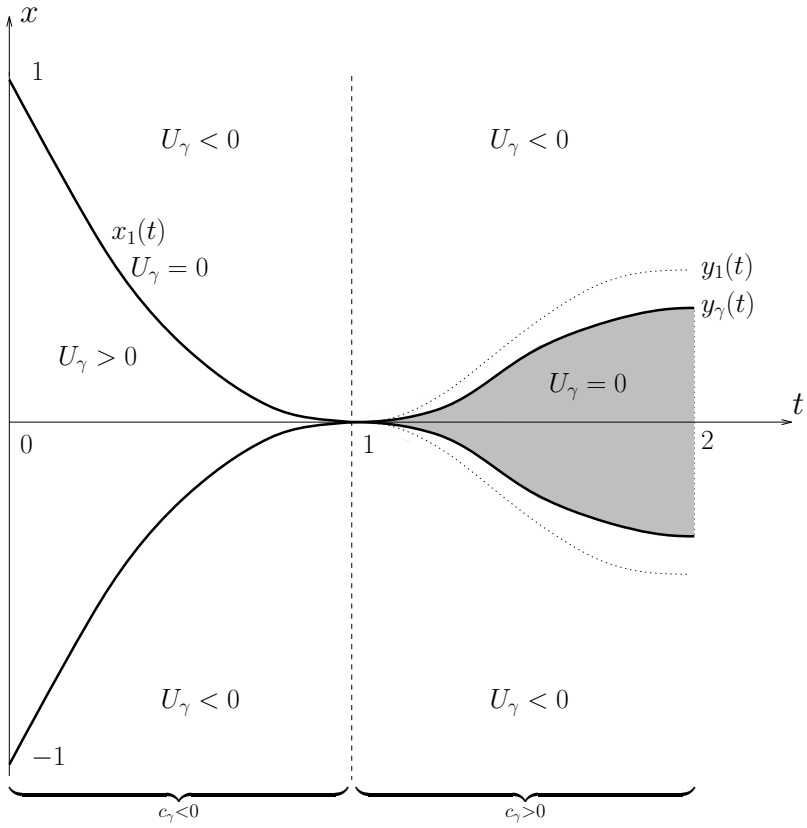


FIG. 1. *Fattening phenomenon for the functions* U_γ.

(see Figure 1). It follows that all the U_γ's, for measurable $0 \le \gamma(t) \le 1$, are all weak solutions of (19), so we do not have uniqueness and the set of solutions is quite large.

Let us complete this counterexample by pointing out the following:

(i) As in [6], nonuniqueness comes from the fattening phenomenon for the front which is due to the fact that c_γ in (24) changes its sign at $t = 1$. It is even possible to build an autonomous counterexample up to start with a front with several connected components.

(ii) $c_0 \ge 0$, and therefore it also complements the results of section 5; indeed the unique solution u of (11) has the same 0–level-set as U_1 (obtained with $\gamma(t) \equiv 1$), and, with the notation of Theorem 1.5, $\rho^+ = \mathbb{1}_{\{U^1(\cdot,t)\ge 0\}}$ and $\rho^- = \mathbb{1}_{\{U^1(\cdot,t)>0\}}$. In particular, for $t \ge 1$,

$$\{v^+(\cdot,t) \ge 0\} = \{U^1(\cdot,t) \ge 0\} = [-y_1(t), y_1(t)]$$

and

$$\{v^-(\cdot,t) > 0\} = \{U^1(\cdot,t) > 0\} = \emptyset.$$

Finally, we note that there are no strong solutions since (8) is obviously never satisfied.

(iii) $c_0 = 1$ does not satisfies (H1), but because of the finite speed of propagation property, it is possible to keep the same solution on a large ball in space and for $t \in (0,T)$ if we replace c_0 by a function with compact support in space such that

$c_0(x,t) = 1$ for $|x| \le R$ with R large enough. In this way, it is possible for c_0 to satisfy (H1).

4. Uniqueness results for weak solutions of (3). Uniqueness of weak solutions of (3) is false in general, as shown in the counterexample of the previous section for sign changing velocities c_1. This is in particular related to the "fattening phenomenon." In [2] and [5] the authors proved that there is a unique "classical" viscosity solution for (3) under the assumptions that the initial set $\{u(\cdot,0) \ge 0\}$ has the "interior sphere property" and that $c_1(\cdot,t) \ge |c_0(\cdot,t)|_{L^1}$ for any $t \ge 0$—a condition which ensures that the velocity $\bar{c}$ is nonnegative. By "classical" continuous viscosity solutions we mean that $t \to \mathbb{1}_{\{u(\cdot,t)\ge 0\}}$ is continuous in L^1, which entails that $(x,t) \mapsto \bar{c}(x,t)$ is continuous, and that (3) holds in the usual viscosity sense.

Here we prove Theorem 1.3. If the condition $c_1(\cdot,t) \ge |c_0(\cdot,t)|_{L^1}$ is satisfied, then weak solutions are viscosity solutions. We also prove that the weak solution is unique if we suppose, moreover, either that the initial condition has the interior sphere property or that the strict inequality $c_1(\cdot,t) > |c_0(\cdot,t)|_{L^1}$ holds for any $t \ge 0$.

Proof of Theorem 1.3. 1. *Weak solutions are classical continuous viscosity solutions.* Let u be a weak solution, and let $\bar{c}$ be associated with u as in Definition 1.1. Then, for any $x \in \mathbb{R}^N$ and for almost all $t \in [0,T]$, we have

$$\begin{aligned}\bar{c}(x,t) &\ge c_1(x,t) + c_0^+(\cdot,t) \star \mathbb{1}_{\{u(\cdot,t)>0\}}(x) - c_0^-(\cdot,t) \star \mathbb{1}_{\{u(\cdot,t)\ge 0\}}(x)\\ &\ge c_1(x,t) - |c_0^-(\cdot,t)|_{L^1}\\ &\ge \delta \ge 0\,.\end{aligned}$$

From [22, Theorem 4.2], there exists a constant η which depends on T such that (10) implies

$$-|u| - |Du| \le -\eta \quad \text{on } \mathbb{R}^N \times (0,T) \tag{25}$$

when we assume, moreover, that $\bar{c}$ is continuous. In our case, where $\bar{c}$ is not assumed continuous in time, (10) follows from the L^1-stability result, Theorem A.3, where we approximate $\bar{c}$ by a continuous function, and from the usual stability for L^1-viscosity subsolutions.

Let us note that from the proof of [5, Corollary 2.5] we have in the viscosity sense

$$-|u(\cdot,t)| - |Du(\cdot,t)| \le -\eta \quad \text{on } \mathbb{R}^N \text{ for every } t \in (0,T)\,. \tag{26}$$

Following [5, Corollary 2.5], we get that, for every $t \in (0,T)$, the 0–level-set of $u(\cdot,t)$ has a zero-Lebesgue measure. Then we deduce that

$$\chi(x,t) = \mathbb{1}_{\{u(\cdot,t)\ge 0\}}(x) \quad \text{for a.e. } x \in \mathbb{R}^N \quad \text{and} \quad \text{for all } t \in (0,T),$$

which (with (7)) entails that

$$\bar{c}(x,t) = c_1 + c_0 \star \mathbb{1}_{\{u(\cdot,t)\ge 0\}}$$

for any (x,t). Moreover, $t \mapsto \mathbb{1}_{\{u(\cdot,t)\ge 0\}}$ is also continuous in L^1, and then $\bar{c}$ is continuous. Therefore u is a classical viscosity solution of (3).

2. *Uniqueness when u_0 is semiconvex (part* (ii)*).* If we assume that (9) and (10) hold and that u_0 is semiconvex, then weak solutions are viscosity solutions, and we can apply the uniqueness result for viscosity solutions given in [5], namely Theorem 4.2

(which remains true under our assumptions), which requires, in particular, semiconvexity of the velocity; see assumption (H2).

3. *A Gronwall-type inequality (part* (i)*).* From now on we assume that $\delta > 0$, and we aim to prove that the solution to (3) is unique. Let u_1, u_2 be two solutions. We set

$$\rho_i = \mathbb{1}_{\{u_i(\cdot,t)\geq 0\}} \qquad \text{and} \qquad \bar{c}_i(x,t) = c_0 \star \rho_i + c_1 \qquad \text{for } i = 1,2.$$

We want to prove in a first step the following Gronwall-type inequality for any t sufficiently small:

$$\begin{aligned} &|\rho_1(\cdot,t) - \rho_2(\cdot,t)|_{L^1} \\ &\leq\ C\left[\mathrm{per}(\{u_1(\cdot,t)\geq 0\}) + \mathrm{per}(\{u_2(\cdot,t)\geq 0\})\right] \int_0^t |\rho_1(\cdot,s) - \rho_2(\cdot,s)|_{L^1} ds, \end{aligned} \tag{27}$$

where C is a constant depending on the constants of the problem, where $\mathrm{per}(\{u_i(\cdot,t)\geq 0\})$ is the $\mathcal{H}^{N-1}$ measure of the set $\partial\{u_i(\cdot,t)\geq 0\}$ (for $i = 1,2$).

We have

$$|\rho_1(\cdot,t) - \rho_2(\cdot,t)|_{L^1} \leq \mathcal{L}^N(\{-\alpha_t \leq u_1(\cdot,t) < 0\}) + \mathcal{L}^N(\{-\alpha_t \leq u_2(\cdot,t) < 0\})\ , \tag{28}$$

where $\mathcal{L}^N$ is the Lebesgue measure in $\mathbb{R}^N$ and

$$\alpha_t = \sup_{s\in[0,t]} |(u_1 - u_2)(\cdot,s)|_\infty \qquad \text{for any } t \in (0,T). \tag{29}$$

In order to estimate the right-hand side of inequality (28), as in the proof of [5, Theorem 4.2], we need a lower-gradient bound as well as a semiconvexity property for u_1 and u_2. We already know from step 1 that $\bar{c}_i$ is continuous for $i = 1,2$.

Let us start to estimate the right-hand side of (28). From the "stability estimates" on the solutions with respect to variations of the velocity (see [5, Lemma 2.2]), we have

$$\alpha_t \leq |Du_0|_\infty e^{Lt} \int_0^t |(\bar{c}_1 - \bar{c}_2)(\cdot,s)|_\infty ds\ , \tag{30}$$

where $L = L_0 + L_1$. Therefore

$$\alpha_t \leq m_0\ |Du_0|_\infty e^{Lt} \int_0^t |(\rho_1 - \rho_2)(\cdot,s)|_{L^1} ds\ . \tag{31}$$

where the constant m_0 is given in (H2). In particular, since the $\rho_i(\cdot,t)$ are continuous in L^1 and equal at time $t = 0$ for $i = 1,2$, we have $\alpha_t/t \to 0$ as $t \to 0^+$.

From now on, we mimic the proof of [5, Proposition 4.5]. Using the lower-gradient bound (25) for $u_i(\cdot,t)$ combined with the increase principle (see [5, Lemma 2.3]), we obtain for $\alpha_t < \eta/2$ that

$$\{-\alpha_t \leq u_i(\cdot,t) < 0\} \subset \{u_i(\cdot,t)\geq 0\} + (2\alpha_t/\eta)\overline{B}(0,1)$$

for $i = 1,2$. From the interior ball regularization lemma (Lemma 2.3), the set $\{u_i(\cdot,t)\geq 0\}$ satisfies for $t \in (0,T)$ the interior ball property of radius $r_t = \eta t/C_0$.

Applying [2, Lemmas 2.5 and 2.6], we obtain for $\sigma_t = 2\alpha_t/\eta$ that

$$\begin{aligned}\mathcal{L}^N(\{-\alpha_t \le u_i(\cdot,t) < 0\}) \le\ & \mathcal{L}^N\left((\{u_i(\cdot,t) \ge 0\} + \sigma_t\,\overline{B}(0,1))\backslash\{u_i(\cdot,t) \ge 0\}\right)\\ &\le \frac{r_t}{N}\left[\left(1+\frac{\sigma_t}{r_t}\right)^N - 1\right]\mathrm{per}(\{u_i(\cdot,t)\ge 0\})\\ &\le \frac{2^N\alpha_t}{\eta}\mathrm{per}(\{u_i(\cdot,t)\ge 0\})\end{aligned}$$

(using $(1+a)^N - 1 \le aN(1+a)^{N-1}$ for $a \ge 0$) for $t \in [0,\tau]$, where $0 < \tau \le T$ is defined by

$$\tau = \sup\left\{t > 0 : \alpha_t < \frac{\eta}{2} \text{ and } 2\frac{C_0}{\eta^2}\frac{\alpha_t}{t} \le 1\right\}. \tag{32}$$

Putting together (31), (28), and the previous inequality proves (27).

4. *Uniqueness when $\delta > 0$ (part* (i)*).* We now complete the uniqueness proof under the assumption $\delta > 0$. For this we first show that $\rho_1 = \rho_2$ in $[0,\tau]$. In order to apply the Gronwall lemma to the L^1-estimate (27) obtained in step 3, it is enough to prove that the functions $t \mapsto \mathrm{per}(\{u_i(\cdot,t)\ge 0\})$ belong to L^1. For this, let us set

$$w_i(x) = \inf\{t \ge 0\ :\ u_i(x,t) \ge 0\}\ .$$

Since u_i solves the eikonal equation $(u_i)_t = \bar{c}_i(x,t)|Du_i|$, from classical representation formulae, we have

$$\begin{aligned}\{u_i(\cdot,t)\ge 0\} = \{x\ :\ &\exists y(\cdot),\ |\dot y(s)| \le c(y(s),s),\ 0\le s\le t,\\ &u_0(y(0)) \ge 0 \text{ and } y(t) = x\}.\end{aligned}$$

Therefore

$$\begin{aligned}w_i(x) = \inf\{t \ge 0\ :\ &\exists y(\cdot),\ |\dot y(s)| \le c(y(s),s),\ 0\le s\le t,\\ &u_0(y(0)) \ge 0 \text{ and } y(t) = x\}.\end{aligned}$$

Applying the dynamic programming principle, since $\bar{c}_i \ge \delta > 0$, we obtain that w_i is Lipschitz continuous and is a viscosity solution of the autonomous equation $\bar{c}_i(x,w_i(x))|Dw_i(x)| = 1$. Note that $\{u_i(\cdot,t)\ge 0\} = \{w_i \le t\}$. In particular, by Theorem 2.1(i), $\{w_i \le t\} \subset B(0, R_0 + (M_0+M_1)T)$ is bounded for any t. From the coarea formula, we have

$$\begin{aligned}\int_0^t \mathrm{per}(\{u_i(\cdot,s)\ge 0\})ds\ &=\ \int_0^t \mathrm{per}(\{w_i \le s\})ds\\ &=\ \int_{\{w_i\le t\}} |Dw_i(x)|dx,\end{aligned}$$

which is finite since w_i is Lipschitz continuous. Therefore we have proved that $t \mapsto \mathrm{per}(\{u_i(\cdot,t)\ge 0\})$ belongs to $L^1([0,\tau])$, which entails from the Gronwall lemma that $\rho_1 = \rho_2$ in $[0,\tau]$ since $\rho_1(\cdot,0) = \rho_2(\cdot,0)$. Hence $\bar{c}_1 = \bar{c}_2$ and $u_1 = u_2$ in $[0,\tau]$. From the definition of α_t and τ (see (29) and (32)), necessarily $\tau = T$. It completes the proof. □

5. Nonnegative kernel c_0 and Slepčev formulation for the nonlocal term. In this section, we deal with nonnegative kernels $c_0 \geq 0$. In this monotone framework, inclusion principles for evolving sets and comparison for solutions to the dislocation equation are expected (see Cardaliaguet [11] for related results). We start by studying the right level-set equation using a Slepčev formulation with the convolution term using all the level-sets $\{u(\cdot,t) \geq u(x,t)\}$ instead of only one level-set $\{u(\cdot,t) \geq 0\}$. This choice is motivated by the good stability properties of the Slepčev formulation.

The equation we are concerned with is

$$\begin{cases} \dfrac{\partial u}{\partial t} = c^+[u](x,t)|Du| & \text{in } \mathbb{R}^N \times (0,T), \\ u(\cdot,0) = u_0 & \text{in } \mathbb{R}^N, \end{cases} \tag{33}$$

where the nonlocal velocity is

$$\begin{aligned} c^+[u](x,t) &= c_1(x,t) + c_0(\cdot,t) \star \mathbb{1}_{\{u(\cdot,t) \geq u(x,t)\}}(x) \\ &= c_1(x,t) + \int_{\mathbb{R}^N} c_0(x-z,t)\mathbb{1}_{\{u(\cdot,t) \geq u(x,t)\}}(z)dz \end{aligned} \tag{34}$$

and the additional velocity c_1 has no particular sign.

We denote

$$c^-[u](x,t) = c_1(x,t) + \int_{\mathbb{R}^N} c_0(x-z,t)\mathbb{1}_{\{u(\cdot,t) > u(x,t)\}}(z)dz.$$

We recall the notion of viscosity solutions for (33) as it appears in [28].

DEFINITION 5.1 (Slepčev viscosity solutions). *An upper semicontinuous function $u : \mathbb{R}^N \times [0,T] \to \mathbb{R}$ is a viscosity subsolution of* (33) *if, for any $\varphi \in C^1(\mathbb{R}^N \times [0,T])$, for any maximum point $(\bar{x},\bar{t})$ of $u - \varphi$, if $\bar{t} > 0$, then*

$$\frac{\partial \varphi}{\partial t}(\bar{x},\bar{t}) \leq c^+[u](\bar{x},\bar{t})|D\varphi(\bar{x},\bar{t})|$$

and $u(\bar{x},0) \leq u_0(\bar{x})$ if $\bar{t} = 0$.

A lower semicontinuous function $u : \mathbb{R}^N \times [0,T] \to \mathbb{R}$ is a viscosity supersolution of (33) *if, for any $\varphi \in C^1(\mathbb{R}^N \times [0,T])$, for any minimum point $(\bar{x},\bar{t})$ of $u - \varphi$, if $\bar{t} > 0$, then*

$$\frac{\partial \varphi}{\partial t}(\bar{x},\bar{t}) \geq c^-[u](\bar{x},\bar{t})|D\varphi(\bar{x},\bar{t})|$$

and $u(\bar{x},0) \geq u_0(\bar{x})$ if $\bar{t} = 0$.

A locally bounded function is a viscosity solution of (33) *if its upper semicontinuous envelope is a subsolution and its lower semicontinuous envelope is a supersolution of* (33).

Note that for the supersolution, we require the viscosity inequality with c^- instead of c^+. It is the definition providing the expected stability results (see [28]).

THEOREM 5.2 (comparison principle). *Assume* (H0′) *and that the kernels $c_0 \geq 0$ and c_1 satisfy* (H1). *Let u (respectively, v) be a bounded upper semicontinuous subsolution (respectively, a bounded lower semicontinuous supersolution) of* (33). *Then $u \leq v$ in $\mathbb{R}^N \times [0,T]$.*

Remark 5.1. We could deal with second-order terms in (33) (for instance we can add the mean curvature to the velocity (1)). See Forcadel [17] and Srour [29] for related results.

Before giving the proof of Theorem 5.2, let us note the following consequence.

Proof of Theorem 1.4. The uniqueness of a continuous viscosity solution to (33) is an immediate consequence of Theorem 5.2. Then existence is proved by Perron's method using classical arguments (see, for instance, [16, Theorem 1.2]), so we skip the details. □

Proof of Theorem 5.2. 1. *The test function.* Since $u - v$ is a bounded upper semicontinuous function, for any $\varepsilon, \eta, \alpha > 0$ and $K = 2(L_0 + L_1) \geq 0$, the supremum

$$M_{\varepsilon,\eta,\alpha} = \sup_{(x,y,t)\in(\mathbb{R}^N)^2\times[0,T]} \left\{ u(x,t) - v(y,t) - \mathrm{e}^{Kt}\left(\frac{|x-y|^2}{\varepsilon^2} + \alpha|x|^2 + \alpha|y|^2\right) - \eta t \right\}$$

is finite and achieved at a point $(\bar{x}, \bar{y}, \bar{t})$. Classical arguments show that

$$\liminf_{\varepsilon,\eta,\alpha\to 0} M_{\varepsilon,\eta,\alpha} = \sup_{\mathbb{R}^N\times[0,T]} \{u - v\}$$

and that

$$\frac{|\bar{x}-\bar{y}|^2}{\varepsilon^2}, \alpha|\bar{x}|^2, \alpha|\bar{y}|^2 \leq M_\infty, \tag{35}$$

where $M_\infty = |u|_\infty + |v|_\infty$.

2. *Viscosity inequalities when* $\bar{t} > 0$. Writing the viscosity inequalities for the subsolution u and the supersolution v, we obtain

$$K\mathrm{e}^{K\bar{t}}\left(\frac{|\bar{x}-\bar{y}|^2}{\varepsilon^2} + \alpha|\bar{x}|^2 + \alpha|\bar{y}|^2\right) + \eta \leq c^+[u](\bar{x},\bar{t})|\bar{p} + \bar{q}_x| - c^-[v](\bar{y},\bar{t})|\bar{p} - \bar{q}_y|, \tag{36}$$

where $\bar{p} = 2e^{K\bar{t}}(\bar{x} - \bar{y})/\varepsilon^2$, $\bar{q}_x = 2e^{K\bar{t}}\alpha\bar{x}$, and $\bar{q}_y = 2e^{K\bar{t}}\alpha\bar{y}$. We point out a difficulty in obtaining this inequality: in general, one gets it by doubling the time variable first and then by passing to the limit in the time penalization. This is not straightforward here because of the dependence with respect to time of the nonlocal terms. But the stability arguments of the Slepčev formulation take care of this difficulty.

3. *Difference between* $\{u(\cdot,\bar{t}) \geq u(\bar{x},\bar{t})\}$ *and* $\{v(\cdot,\bar{t}) > v(\bar{y},\bar{t})\}$. We have

$$\{u(\cdot,\bar{t}) \geq u(\bar{x},\bar{t})\} \subset \{v(\cdot,\bar{t}) > v(\bar{y},\bar{t})\} \cup \mathcal{E}, \tag{37}$$

where $\mathcal{E} = \{u(\cdot,\bar{t}) \geq u(\bar{x},\bar{t})\} \cap \{v(\cdot,\bar{t}) \leq v(\bar{y},\bar{t})\}$. If $x \in \mathcal{E}$, then $u(\bar{x},\bar{t}) - v(\bar{y},\bar{t}) \leq u(x,\bar{t}) - v(x,\bar{t})$. But from the definition of $M_{\varepsilon,\eta,\alpha}$,

$$\begin{aligned} &u(x,\bar{t}) - v(x,\bar{t}) - \mathrm{e}^{K\bar{t}}2\alpha|x|^2 - \eta\bar{t} \\ &\leq u(\bar{x},\bar{t}) - v(\bar{y},\bar{t}) - \mathrm{e}^{K\bar{t}}\left(\frac{|\bar{x}-\bar{y}|^2}{\varepsilon^2} + \alpha|\bar{x}|^2 + \alpha|\bar{y}|^2\right) - \eta\bar{t}. \end{aligned}$$

It follows that

$$\mathcal{E} \subset \left\{ x \in \mathbb{R}^N : |x|^2 \geq \frac{1}{2}(|\bar{x}|^2 + |\bar{y}|^2) + \frac{|\bar{x}-\bar{y}|^2}{2\alpha\varepsilon^2} \right\}.$$

4. *Upper bound for* $c^+[u](\bar{x},\bar{t})$. We have

$$c^+[u](\bar{x},\bar{t}) = \int_{\mathbb{R}^N} c_0(\bar{x}-z,\bar{t})\mathbb{1}_{\{u(\cdot,\bar{t})\geq u(\bar{x},\bar{t})\}}(z)dz + c_1(\bar{x},\bar{t}) \tag{38}$$

$$\leq \int_{\mathbb{R}^N} (c_0(\bar{x}-z,\bar{t}) - c_0(\bar{y}-z,\bar{t}))\mathbb{1}_{\{u(\cdot,\bar{t})\geq u(\bar{x},\bar{t})\}}(z)dz$$

$$+ \int_{\mathbb{R}^N} c_0(\bar{y}-z,\bar{t})\mathbb{1}_{\{u(\cdot,\bar{t})\geq u(\bar{x},\bar{t})\}}(z)dz + c_1(\bar{x},\bar{t}).$$

Using that $c_0 \geq 0$ and (37), we obtain

$$\int_{\mathbb{R}^N} c_0(\bar{y}-z,\bar{t})\mathbb{1}_{\{u(\cdot,\bar{t})\geq u(\bar{x},\bar{t})\}}(z)dz \leq \int_{\{v(\cdot,\bar{t})>v(\bar{y},\bar{t})\}\cup\mathcal{E}} c_0(\bar{y}-z,\bar{t})dz.$$

From (38), we get

$$c^+[u](\bar{x},\bar{t}) \leq c^-[v](\bar{y},\bar{t}) + \mathcal{I}_1 + \mathcal{I}_2 + c_1(\bar{x},\bar{t}) - c_1(\bar{y},\bar{t}), \tag{39}$$

where

$$\mathcal{I}_1 = \int_{\mathbb{R}^N} (c_0(\bar{x}-z,\bar{t}) - c_0(\bar{y}-z,\bar{t}))\mathbb{1}_{\{u(\cdot,\bar{t})\geq u(\bar{x},\bar{t})\}}(z)dz$$

and

$$\mathcal{I}_2 = \int_{\mathcal{E}} c_0(\bar{y}-z,\bar{t})dz.$$

5. *Estimate of* $\mathcal{I}_1$ *using* (H1). We have

$$c_0(\bar{x}-z,t) - c_0(\bar{y}-z,t) = \int_0^1 D_x c_0((1-\lambda)(\bar{y}-z) + \lambda(\bar{x}-z),\bar{t})(\bar{x}-\bar{y})d\lambda.$$

It follows that

$$|\mathcal{I}_1| \leq \int_{\mathbb{R}^N}\int_0^1 |D_x c_0((1-\lambda)(\bar{y}-z) + \lambda(\bar{x}-z),\bar{t})||\bar{x}-\bar{y}|d\lambda dz$$

$$\leq |D_x c_0(\cdot,\bar{t})|_{L^1}|\bar{x}-\bar{y}|$$

$$\leq L_0|\bar{x}-\bar{y}|. \tag{40}$$

6. *Estimate of the right-hand side of inequality* (36). Noticing that $|\bar{p}||\bar{x}-\bar{y}| = 2\mathrm{e}^{K\bar{t}}|\bar{x}-\bar{y}|^2/\varepsilon^2$ and using (40), (39), and (H1), we have

$$c^+[u](\bar{x},\bar{t})|\bar{p}+\bar{q}_x| - c^-[v](\bar{y},\bar{t})|\bar{p}-\bar{q}_y|$$

$$\leq (c^-[v](\bar{y},\bar{t}) + \mathcal{I}_1 + \mathcal{I}_2 + c_1(\bar{x},\bar{t}) - c_1(\bar{y},\bar{t}))|\bar{p}+\bar{q}_x| - c^-[v](\bar{y},\bar{t})|\bar{p}-\bar{q}_y|$$

$$\leq \left|c^-[v](\bar{y},\bar{t})\right||\bar{q}_x+\bar{q}_y| + (L_0+L_1)|\bar{x}-\bar{y}||\bar{p}+\bar{q}_x| + \mathcal{I}_2|\bar{p}+\bar{q}_x|$$

$$\leq (M_0+M_1)(|\bar{q}_x|+|\bar{q}_y|) + 2\mathrm{e}^{K\bar{t}}(L_0+L_1)\frac{|\bar{x}-\bar{y}|^2}{\varepsilon^2}$$

$$+(L_0+L_1)|\bar{x}-\bar{y}||\bar{q}_x| + \mathcal{I}_2\left(|\bar{q}_x| + 2\mathrm{e}^{K\bar{t}}\frac{|\bar{x}-\bar{y}|}{\varepsilon^2}\right).$$

Since $|\bar{q}_x|, |\bar{q}_y| \to 0$ as $\alpha \to 0$ (see (35)) and $\mathcal{I}_2$ is bounded by $|c_0(\cdot,\bar{t})|_{L^1} \leq L_0$, there exists a modulus $m_\varepsilon(\alpha) \to 0$ as $\alpha \to 0$ such that (36) becomes

$$K\mathrm{e}^{K\bar{t}}\left(\frac{|\bar{x}-\bar{y}|^2}{\varepsilon^2}+\alpha|\bar{x}|^2+\alpha|\bar{y}|^2\right)+\eta$$

$$\leq m_\varepsilon(\alpha)+2(L_0+L_1)\mathrm{e}^{K\bar{t}}\frac{|\bar{x}-\bar{y}|^2}{\varepsilon^2}+2\,\mathcal{I}_2\,\mathrm{e}^{K\bar{t}}\frac{|\bar{x}-\bar{y}|}{\varepsilon^2}.$$

Recalling that we chose $K \geq 2(L_0+L_1)$, we finally obtain

$$0<\eta\leq m_\varepsilon(\alpha)+2\,\mathcal{I}_2\,\mathrm{e}^{K\bar{t}}\frac{|\bar{x}-\bar{y}|}{\varepsilon^2}. \tag{41}$$

7. *Limit when* $\alpha \to 0$. First, suppose that

$$\frac{|\bar{x}-\bar{y}|^2}{\varepsilon^2}\to 0 \quad \text{as } \alpha\to 0. \tag{42}$$

It follows that $|\bar{x}-\bar{y}| \to 0$ as $\alpha \to 0$. Passing to the limit in (41), we obtain a contradiction. Therefore, (42) cannot hold, and, up to extracting a subsequence, there exists $\delta > 0$ such that

$$\frac{|\bar{x}-\bar{y}|^2}{\varepsilon^2}\geq\delta>0 \quad \text{for } \alpha>0 \text{ small enough.} \tag{43}$$

From (41) and (35), we get

$$\eta\leq\limsup_{\alpha\to0}2\,\mathcal{I}_2\,\mathrm{e}^{K\bar{t}}\frac{|\bar{x}-\bar{y}|}{\varepsilon^2}\leq\frac{2\mathrm{e}^{K\bar{t}}M_\infty^{1/2}}{\varepsilon}\limsup_{\alpha\to0}\mathcal{I}_2. \tag{44}$$

To obtain a contradiction, it suffices to show that $\limsup_{\alpha\to0}\mathcal{I}_2 = 0$.

8. *Convergence of* $\mathcal{I}_2$ *to* 0 *when* $\alpha \to 0$. By a change of variable, we have

$$\mathcal{I}_2=\int_{\mathcal{E}}c_0(\bar{y}-z,\bar{t})dz\leq\int_{\bar{\mathcal{E}}}c_0(z,\bar{t})dz,$$

where

$$\bar{\mathcal{E}}=\left\{x\in\mathbb{R}^N:|x-\bar{y}|^2\geq\frac{1}{2}(|\bar{x}|^2+|\bar{y}|^2)+\frac{|\bar{x}-\bar{y}|^2}{2\alpha\varepsilon^2}\right\}.$$

Since $|c_0(\cdot,\bar{t})|_{L^1} \leq L_0$, to prove that $\mathcal{I}_2 \to 0$, it suffices to show that $\bar{\mathcal{E}} \subset \mathbb{R}^N \backslash B(0,R_\alpha)$ with $R_\alpha \to +\infty$. From (43), if $x \in \bar{\mathcal{E}}$, then

$$|x|^2\geq-2|x||\bar{y}|+\frac{1}{2}|\bar{x}|^2-\frac{1}{2}|\bar{y}|^2+\frac{\delta}{2\alpha}$$

$$\geq-2|x||\bar{y}|-|\bar{y}||\bar{x}-\bar{y}|+\frac{\delta}{2\alpha}$$

$$\geq\frac{1}{2\alpha}(\delta-2(C^2+2C|x|)\sqrt{\alpha}),$$

since by (35), there exists $C > 0$ such that $|\bar{x}|, |\bar{y}| \leq C/\sqrt{\alpha}$ and $|\bar{x}-\bar{y}| \leq C$. It follows that

$$|x|\geq\frac{1}{\sqrt{\alpha}}\left(-C+\sqrt{C^2+\delta/2-C^2\sqrt{\alpha}}\right):=R_\alpha\underset{\alpha\to0}{\longrightarrow}+\infty.$$

9. *End of the proof.* Finally, for every ε, if $\alpha = \alpha_\varepsilon$ is small enough, the supremum $M_{\varepsilon,\eta,\alpha}$ is necessarily achieved for $\bar{t} = 0$. It follows that

$$M_{\varepsilon,\eta,\alpha} \leq u(\bar{x},0) - v(\bar{y},0) - \frac{|\bar{x}-\bar{y}|^2}{\varepsilon^2} \leq u_0(\bar{x},0) - u_0(\bar{y},0) - \frac{|\bar{x}-\bar{y}|^2}{\varepsilon^2}.$$

Since u_0 is uniformly continuous, for all $\rho > 0$, there exists $C_\rho > 0$ such that

$$M_{\varepsilon,\eta,\alpha} \leq \rho + C_\rho|\bar{x}-\bar{y}| - \frac{|\bar{x}-\bar{y}|^2}{\varepsilon^2} \leq \rho + \frac{C_\rho^2\varepsilon^2}{4}.$$

Passing to the limits $\varepsilon \to 0$ and then $\rho, \alpha, \eta \to 0$, we obtain that $\sup\{u - v\} \leq 0$. □

Now we turn to the connections with *discontinuous solutions* and weak solutions, which are closely connected. To do so, if u is the unique continuous solution of (33) given by Theorem 1.4, we recall that we use the notation

$$\rho^+ := \mathbb{1}_{\{u\geq 0\}}, \quad \rho^- := \mathbb{1}_{\{u>0\}}, \quad \text{and} \quad c[\rho](x,t) = c_0(\cdot,t) \star \rho(\cdot,t)(x) + c_1(x,t).$$

Proof of Theorem 1.5. 1. *Claim: Under the assumptions of Theorem* 1.5*, the functions ρ^+ and ρ^- are L^1-viscosity solutions of the equation*

$$\text{(45)} \qquad \begin{cases} \rho_t = c[\rho]|D\rho| & \text{in } \mathbb{R}^N \times [0,T)\,, \\ \rho(x,0) = \mathbb{1}_{\{u_0\geq 0\}} & \text{in } \mathbb{R}^N\,. \end{cases}$$

We consider two sequences of smooth nondecreasing functions $(\psi_\alpha)_\alpha$, $(\psi^\alpha)_\alpha$, taking values in $[0,1]$, such that, for any $s \in \mathbb{R}$,

$$\psi_\alpha(s) \leq \mathbb{1}_{\{x>0\}}(s) \leq \mathbb{1}_{\{x\geq 0\}}(s) \leq \psi^\alpha(s)\,,$$

and such that, as $\alpha \to 0$, $\psi_\alpha \uparrow \mathbb{1}_{\{x>0\}}$, $\psi^\alpha \downarrow \mathbb{1}_{\{x\geq 0\}}$.

We first remark that u satisfies, in the sense of Definition 5.1,

$$u_t \leq c[\psi^\alpha(u(\cdot,t) - u(x,t))]|Du| \quad \text{in } \mathbb{R}^N \times (0,T)\,,$$

$$u_t \geq c[\psi_\alpha(u(\cdot,t) - u(x,t))]|Du| \quad \text{in } \mathbb{R}^N \times (0,T)$$

since u is a continuous solution of (33), $c^+[u](x,t) \leq c[\psi^\alpha(u(\cdot,t) - u(x,t))]$, and $c^-[u](x,t) \geq c[\psi_\alpha(u(\cdot,t) - u(x,t))]$. The point for doing that is that the functions $c[\psi^\alpha(u(\cdot,t) - u(x,t))]$, $c[\psi_\alpha(u(\cdot,t) - u(x,t))]$ are now continuous in x and t.

Then we show that ρ^+, ρ^- satisfy the same inequalities, the functions $c[\psi^\alpha(u(\cdot,t) - u(x,t))]$, $c[\psi_\alpha(u(\cdot,t) - u(x,t))]$ being considered as fixed functions (in other words, we forget that they depend on u). In fact, we just provide the proof in detail for ρ^+, the one for ρ^- being analogous. Following the proof of [6], we set

$$u_\varepsilon(x,t) := \frac{1}{2}\left(1 + \tanh\left(\varepsilon^{-1}(u(x,t) + \varepsilon^{1/2})\right)\right)\,.$$

Noticing that $u_\varepsilon = \phi_\varepsilon(u)$ for an increasing function ϕ_ε, we have that the function u_ε still satisfies the two above inequalities. It is easy to see that

$$\rho^+ = \limsup{}^*\, u_\varepsilon \quad \text{and} \quad (\rho^+)_* = \liminf_*\, u_\varepsilon\,,$$

and the half-relaxed limits method indeed shows that

$$(\rho^+)^*_t \le c[\psi^\alpha(u(\cdot,t)-u(x,t))]|D(\rho^+)^*| \quad \text{in } \mathbb{R}^N\times(0,T)\ ,$$
$$((\rho^+)_*)_t \ge c[\psi_\alpha(u(\cdot,t)-u(x,t))]|D(\rho^+)_*| \quad \text{in } \mathbb{R}^N\times(0,T)\ .$$

The next step consists of remarking that the viscosity sub- and supersolution inequalities for ρ^+ are obviously satisfied in the complementary of $\partial\{u\ge 0\}$ since ρ^+ is locally constant there, and therefore it is a classical solution of the problem. The only nontrivial viscosity sub- and supersolution inequalities we have to check are at points $(x,t)\in\partial\{u\ge 0\}$, i.e., such that $u(x,t)=0$ since u is continuous. For such points, as $\alpha\to 0$,

$$c[\psi^\alpha(u(\cdot,t)-u(x,t))]\to c[\rho^+](x,t)=c[(\rho^+)^*](x,t)$$

since ρ^+ is upper semicontinuous, and

$$c[\psi_\alpha(u(\cdot,t)-u(x,t))]\to c[\rho^-](x,t).$$

The stability result for equations with an L^1-dependence in time yields the inequalities

$$(\rho^+)^*_t \le c[(\rho^+)^*]|D(\rho^+)^*| \quad \text{in } \mathbb{R}^N\times(0,T),$$
$$((\rho^+)_*)_t \ge c[\rho^-]|D(\rho^+)_*| \quad \text{in } \mathbb{R}^N\times(0,T). \tag{46}$$

The second inequality is weaker than the one we claim: to obtain $c[(\rho^+)_*]$ instead of $c[\rho^-]$, we have to play with the different level-sets of u: for $\beta>0$ small, we set $\rho^+_\beta=\mathbb{1}_{\{u\ge-\beta\}}$. Since u is a solution of (33) and $\psi_{\alpha,\beta}:=\psi_\alpha(\cdot+\beta)$ is nondecreasing, then $\psi_{\alpha,\beta}(u)$ is a (continuous) supersolution of

$$(\psi_{\alpha,\beta}(u))_t \ge c^-[\psi_{\alpha,\beta}(u)]|D\psi_{\alpha,\beta}(u)| \quad \text{in } \mathbb{R}^N\times(0,T).$$

By stability we get, as $\alpha\to 0$,

$$((\rho^+_\beta)_*)_t \ge c^-[(\rho^+_\beta)_*]|D(\rho^+_\beta)_*| \quad \text{in } \mathbb{R}^N\times(0,T).$$

But, for $(x,t)\in\partial\{u\ge-\beta\}$,

$$c^-[(\rho^+_\beta)_*](x,t)=c[\mathbb{1}_{\{(\rho^+_\beta)_*(\cdot,t)>0\}}](x,t)=c[(\rho^+_\beta)_*](x,t)\ge c[\rho^+](x,t). \tag{47}$$

It follows that $(\rho^+_\beta)_*$ is a supersolution of the eikonal equation with $c[\rho^+](x,t)$ (as before, the only nontrivial inequalities we have to check are on $\partial\{u\ge-\beta\}$, and they are true because of (47)). Letting β tend to 0 and using that $(\rho^+)_*=\liminf_*(\rho^+_\beta)_*$, we obtain the expected inequality (even something better since (46) holds actually with $c[\rho^+](x,t)$). In particular, we get that ρ^+ is a solution of

$$\begin{cases} \rho_t = c[\rho^+]|D\rho| & \text{in } \mathbb{R}^N\times[0,T),\\ \rho(x,0)=\mathbb{1}_{\{u_0\ge 0\}} & \text{in } \mathbb{R}^N. \end{cases} \tag{48}$$

The proof of the claim is complete.

2. *The functions $v^\pm$ are weak solutions of* (3). Let us start with the "+" case. We first remark that the existence and uniqueness of v^+ follows from the standard theory for equations with an L^1-dependence in time (see Appendix A).

To prove that v^+ is a weak solution of (3), it remains to prove that (7) holds. It is sufficient to show that

$$\{v^+(\cdot,t)>0\}\subset\{u(\cdot,t)\geq 0\}\subset\{v^+(\cdot,t)\geq 0\}\ . \tag{49}$$

We use again the functions ψ_α,ψ^α introduced above. We remark that

$$\psi_\alpha(u_0)\leq\rho^+(x,0)\leq\psi^\alpha(u_0)\quad\text{in }\mathbb{R}^N\ .$$

Moreover, v^+ and ρ^+ are solutions of the same equation, namely (48) with $c[\rho^+]$, which is considered as a fixed function, and so are $\psi_\alpha(v^+)$ and $\psi^\alpha(v^+)$ because the equation is geometric. Therefore, a standard comparison result implies

$$\psi_\alpha(v^+)\leq(\rho^+)_*\leq\rho^+=(\rho^+)^*\leq\psi^\alpha(v^+)\quad\text{in }\mathbb{R}^N\times[0,T)\ .$$

Letting α tend to 0, these inequalities imply (49).

We can prove the symmetric result for v^-, the only difference being that inclusion (49) has to be replaced by

$$\{v^-(\cdot,t)>0\}\subset\{u(\cdot,t)>0\}\subset\{v^-(\cdot,t)\geq 0\}\ . \tag{50}$$

3. *Claim: If v is a weak solution of* (3)*, then $\mathbb{1}_{\{v(\cdot,t)\geq 0\}}$ is an L^1-subsolution of* (45). From Proposition 3.1 and since $c_0\geq 0$, v satisfies in the L^1-sense

$$v_t\leq c[\mathbb{1}_{\{v(\cdot,t)\geq 0\}}]|Dv|\quad\text{in }\mathbb{R}^N\times[0,T)\ . \tag{51}$$

By similar arguments as we used above, the function $\mathbb{1}_{\{v(\cdot,t)\geq 0\}}$ satisfies the same inequality which gives the result.

4. *The function ρ^+ is the maximal L^1-subsolution of* (45). Let w be an L^1-(upper semicontinuous) subsolution of (45). First we have $w\leq 1$ in $\mathbb{R}^N\times[0,T)$ by comparison with the constant supersolution 1 for the equation with $c[w]$ fixed. By considering $\max(w,0)$ we may assume that $0\leq w\leq 1$ in $\mathbb{R}^N\times[0,T)$. By similar arguments as we already used in step 1, we can show that $\mathbb{1}_{\{w(\cdot,t)>0\}}$ is also an L^1-subsolution of (45); thus we can assume that w is a characteristic function.

Then we remark that w is also a subsolution of (33): indeed, again, the only nontrivial viscosity inequalities are on the boundary of the set $\{w=1\}$, and if (x,t) is such a point, we have $w(x,t)=1$ because w is upper semicontinuous and $w=\mathbb{1}_{\{w(\cdot,t)\geq w(x,t)\}}$. Since u is a solution of the geometric equation (33), $\psi^\alpha(u)$ is still a solution which satisfies $\psi^\alpha(u)(x,0)\geq\mathbb{1}_{\{u_0\geq 0\}}\geq w(x,0)$ in $\mathbb{R}^N$. By Theorem 5.2 we obtain

$$w\leq\psi^\alpha(u)\quad\text{in }\mathbb{R}^N\times[0,T)\ .$$

Letting α tend to 0 provides $w\leq\rho^+$, which proves that ρ^+ is the maximal subsolution of (45).

5. *The function v^+ is the maximal weak solution of* (3). Let v be a weak solution of (3). From steps 3 and 4, we get $\mathbb{1}_{\{v(\cdot,t)\geq 0\}}\leq\rho^+(\cdot,t)$ in $\mathbb{R}^N\times[0,T)$, and (51) implies

$$v_t\leq c[\rho^+]|Dv|\quad\text{in }\mathbb{R}^N\times[0,T)\ .$$

Therefore v is a subsolution of (12), and by a standard comparison result, this leads to $v\leq v^+$ in $\mathbb{R}^N\times[0,T)$, which proves the result.

6. *We have* $\{v^+(\cdot,t)\geq 0\}=\{u(\cdot,t)\geq 0\}$ *and* $\{v^-(\cdot,t)>0\}=\{u(\cdot,t)>0\}$. From step 5, we get $\mathbb{1}_{\{v^+(\cdot,t)\geq 0\}}\leq \rho^+(\cdot,t)=\mathbb{1}_{\{u(\cdot,t)\geq 0\}}$. The conclusion follows for v^+ using (49). The inclusion for v^- uses symmetric arguments.

7. *Uniqueness when* $\{u(\cdot,t)=0\}$ *has Lebesgue measure* 0. If $\mathcal{L}^N(\{u(\cdot,t)=0\})=0$, then $c[\rho^+]=c[\rho^-]$. Hence $v^+=v^-$ is the unique weak solution of (3), and it is obviously a classical one. ∎

Appendix A. A stability result for eikonal equations with L^1-dependence in time. The aim of this appendix is to provide a self-contained presentation of a stability result for viscosity solutions of eikonal equations with L^1-dependences in time which handles the case of weak convergence of the equations instead of the classical strong L^1-convergence. This stability result is a particular case of a general stability result proved by Barles in [4].

For $T>0$, we are interested in solutions of the following equation:

$$\begin{cases} \dfrac{\partial v}{\partial t}=\bar{c}(x,t)|Dv| & \text{in } \mathbb{R}^N\times(0,T),\\ v(\cdot,0)=u_0 & \text{in } \mathbb{R}^N, \end{cases} \tag{52}$$

where the velocity $\bar{c}:\mathbb{R}^N\times(0,T)\to\mathbb{R}$ is defined for almost every $t\in(0,T)$. We also assume that $\bar{c}$ satisfies the following.

(H3) The function $\bar{c}$ is continuous with respect to $x\in\mathbb{R}^N$ and measurable in t. For all $x,y\in\mathbb{R}^N$ and almost all $t\in[0,T]$,

$$|\bar{c}(x,t)|\leq M \quad\text{and}\quad |\bar{c}(x,t)-\bar{c}(y,t)|\leq L|x-y|.$$

Let us underline that we do not assume any continuity in time of $\bar{c}$. We recall the following (under assumption (H0)).

Definition A.1 (L^1-viscosity solutions). *An upper semicontinuous (respectively, lower semicontinuous) function v on $\mathbb{R}^N\times[0,T]$ is an L^1-viscosity subsolution (respectively, supersolution) of* (52) *if*

$$v(0,\cdot)\leq u_0 \quad (\textit{respectively,}\quad v(0,\cdot)\geq u_0)$$

and if for every $(x_0,t_0)\in\mathbb{R}^N\times[0,T]$, $b\in L^1(0,T)$, $\varphi\in C^\infty(\mathbb{R}^N\times(0,T))$, and continuous function $G:\mathbb{R}^N\times(0,T)\times\mathbb{R}^N\to\mathbb{R}$ such that

(i) *the function*

$$(x,t)\longmapsto v(x,t)-\int_0^t b(s)ds-\varphi(x,t)$$

has a local maximum (respectively, minimum) at (x_0,t_0) over $\mathbb{R}^N\times(0,T)$ and such that

(ii) *for almost every $t\in(0,T)$ in some neighborhood of t_0 and for every (x,p) in some neighborhood of (x_0,p_0) with $p_0=\nabla\varphi(x_0,t_0)$, we have*

$$\bar{c}(x,t)|p|-b(t)\leq G(x,t,p) \quad (\textit{respectively,}\quad \bar{c}(x,t)|p|-b(t)\geq G(x,t,p)),$$

then

$$\frac{\partial\varphi}{\partial t}(x_0,t_0)\leq G(x_0,t_0,p_0) \quad \left(\textit{respectively,}\quad \frac{\partial\varphi}{\partial t}(x_0,t_0)\geq G(x_0,t_0,p_0)\right).$$

Finally, we say that a locally bounded function v defined on $\mathbb{R}^N \times [0,T]$ is an L^1-viscosity solution of (52) *if its upper semicontinuous (respectively, lower semicontinuous) envelope is an L^1-viscosity subsolution (respectively, supersolution).*

Let us recall that viscosity solutions in the L^1-sense were introduced in Ishii's paper [20]. We refer the reader to Nunziante [25, 26] and Bourgoing [7, 8] for a complete presentation of the theory.

Then we have the following result.

THEOREM A.2 (existence and uniqueness). *For any $T > 0$, under assumptions* (H0) *and* (H3)*, there exists a unique L^1-viscosity solution to* (52).

Finally, let us consider the solutions v^ε to the following equation:

$$\begin{cases} \dfrac{\partial v^\varepsilon}{\partial t} = \bar{c}^\varepsilon(x,t)|Dv^\varepsilon| & \text{in } \mathbb{R}^N \times (0,T), \\ v^\varepsilon(\cdot,0) = u_0 & \text{in } \mathbb{R}^N. \end{cases} \tag{53}$$

Then we have the following.

THEOREM A.3 (L^1-stability [4]). *Under assumption* (H0)*, let us assume that the velocity $\bar{c}^\varepsilon$ satisfies* (H3) *(with some constants M, L independent of ε). Let us consider the L^1-viscosity solution v^ε to* (53)*. Assume that v^ε converges locally uniformly to a function v and, for all $x \in \mathbb{R}^N$,*

$$\int_0^t \bar{c}^\varepsilon(x,s)ds \to \int_0^t \bar{c}(x,s)ds \quad \textit{locally uniformly in } (0,T). \tag{54}$$

Then v is a L^1-viscosity solution of (52).

Remark A.1. Theorem A.3 is stated as in [4], but note that, under (H3), assumption (54) is automatically satisfied as soon as the convergence is merely pointwise. Indeed, since

$$\left|\int_0^t \bar{c}^\varepsilon(x,s)ds\right| \le MT \quad \text{and} \quad \left|\int_0^t \bar{c}^\varepsilon(x,s)ds - \int_0^{t'} \bar{c}^\varepsilon(x,s)ds\right| \le M|t-t'|,$$

from Ascoli's theorem, the convergence is uniform.

Appendix B. Interior ball regularization (proof of Lemma 2.3). The proof of this result can be adapted from those of Cannarsa and Frankowska [10] or [2] (see also [9] for related perimeter estimates for general equations). For the sake of completeness, we give a proof close to the one of [10] (this latter holds for much more general, but time-independent, dynamics). The unique (and small) contribution of this part amounts to explaining how this proof can be simplified in the particular case of dynamics of the form (55) and to point out that the time dependence is not an issue for the results of [10] to hold.

We first prove that the reachable set for controlled dynamics of the form

$$\dot{x}(t) = c(x(t),t)u(t), \quad u \in L^\infty([0,T], \overline{B}(0,1)), \tag{55}$$

enjoys the interior ball property for positive time. We assume that $c : [0,T] \times \mathbb{R}^N \to \mathbb{R}$ satisfies, for any $x, y \in \mathbb{R}^N$ and $t \in [0,T]$,

$$\begin{cases} \text{(i)} & c \text{ is Borel measurable,} \\ & \text{differentiable with respect to the space variable for a.e. time,} \\ \text{(ii)} & |c(x,t) - c(y,t)| \le L_1|x-y|, \\ \text{(iii)} & |D_x c(x,t) - D_x c(y,t)| \le N_1|x-y|, \\ \text{(iv)} & M_1 \ge c(x,t) \ge \delta > 0, \end{cases}$$

where $L_1, N_1 \geq 0$ and $M_1, \delta > 0$ are given constants. Let $K_0 \subset \mathbb{R}^N$ be the initial set. We define the reachable set $\mathcal{R}(t)$ from K_0 for (55) at time t by

$$\mathcal{R}(t) = \{x(t)\ ,\ x(\cdot) \text{ solution to (55) with } x(0) \in K_0\}.$$

It is known that $\mathcal{R}(t)$ is a closed subset of $\mathbb{R}^N$. Let y_0 be an extremal solution on the time interval $[0,T]$, i.e., a solution of (55) such that

$$y_0(0) \in K_0 \quad \text{and} \quad y_0(T) \in \partial\mathcal{R}(T).$$

From the Pontryagin maximum principle for extremal trajectories (see, for instance, [13]), there is some adjoint function $p_0 : [0,T] \to \mathbb{R}^N\backslash\{0\}$ such that (y_0, p_0) is a solution to

$$\begin{cases} \dot{y}_0(t) = c(y_0(t),t)\dfrac{p_0(t)}{|p_0(t)|}, \\ -\dot{p}_0(t) = D_x c(y_0(t),t)|p_0(t)|. \end{cases}$$

Since the system is positively homogeneous with respect to p, we can assume, without loss of generality, that $|p_0(T)| = 1$, and we set $\theta_0 := p_0(T)$.

Let P be the matrix-valued solution to

$$\begin{cases} \dot{P}(t) = \dfrac{p_0(t)}{|p_0(t)|}[D_x c(y_0(t),t)]^* P(t), \\ P(T) = Id. \end{cases}$$

A straightforward computation shows that $P^*(t)p_0(t) = \theta_0$ for any $t \in [0,T]$.

Let us fix some parameter $\gamma > 0$ to be chosen later, $\theta \in B(0,1)$ and let us set, for all $t \in [0,T]$,

$$y_\theta(t) = y_0(t) - \gamma t P(t)(\theta_0 - \theta).$$

Our aim is to show that y_θ is a solution to (55). Indeed we have

$$\begin{aligned} |\dot{y}_\theta|^2 &= \left| c(y_0,t)\frac{p_0}{|p_0|} - \gamma P(\theta_0-\theta) - \gamma t \frac{p_0}{|p_0|} D_x c^* P(\theta_0-\theta) \right|^2 \\ &= |c(y_\theta,t)|^2 - 2\gamma c(y_0,t)\left\langle \frac{p_0}{|p_0|}, P(\theta_0-\theta)\right\rangle \\ &\quad + |c(y_0,t)|^2 - |c(y_\theta,t)|^2 - 2\gamma t c(y_0,t)\left\langle \frac{p_0}{|p_0|}, \frac{p_0}{|p_0|} D_x c^* P(\theta_0-\theta)\right\rangle \\ &\quad + \gamma^2 \left| P(\theta_0-\theta) + t\frac{p_0}{|p_0|} D_x c^* P(\theta_0-\theta)\right|^2 \\ &\leq |c(y_\theta,t)|^2 - 2\gamma \frac{c(y_0,t)}{|p_0|}\langle \theta_0, \theta_0-\theta\rangle \quad (\text{because } P^* p_0 = \theta_0) \\ &\quad + c^2(y_0,t) - c^2(y_\theta,t) - \langle D_x(c^2)(y_0,t), y_0 - y_\theta\rangle + \gamma^2 M|\theta_0-\theta|^2 \\ &\leq |c(y_\theta,t)|^2 - \gamma\frac{\delta}{|p_0|}|\theta_0-\theta|^2 \ + \ M_1'|y_0-y_\theta|^2 + \gamma^2 M|\theta_0-\theta|^2 \\ &\leq |c(y_\theta,t)|^2 - \gamma\frac{\delta}{|p_0|}|\theta_0-\theta|^2 \ + \ \gamma^2 M'|\theta_0-\theta|^2, \end{aligned}$$

with $M_1' = L_1^2 + M_1 N_1$, and where M and M' depend only on T, L_1, N_1, M_1 because $|p_0(t)|$ is bounded from below by a constant depending only on T, L_1. Hence, for γ sufficiently small, y_θ is a solution of (55) starting from $y_0(0) \in K_0$ and therefore $y_\theta(T) \in \mathcal{R}(T)$.

Finally, $\mathcal{R}(T)$ contains all the $y_\theta(T)$ for $\theta \in B(0,1)$, i.e., the ball centered at $y_0(T) - \gamma T \theta_0$ and of radius γT (since $P(T) = Id$).

We apply the previous result with $c = c_1$ and $K_0 = \{v(\cdot, 0) \geq 0\}$. Then $\{v(\cdot, t) \geq 0\} = \mathcal{R}(t)$ for all $t > 0$.

We end with a remark: in the statement of Lemma 2.3, c_1 is assumed to be continuous in time. As we have seen, it is not necessary; c_1 can be merely measurable in time up to considering the L^1-solution v of (13), as recalled in Appendix A. ∎

REFERENCES

[1] N. Alibaud, *Existence, uniqueness and regularity for nonlinear degenerate parabolic equations with nonlocal terms*, NoDEA Nonlinear Differential Equations Appl., to appear.

[2] O. Alvarez, P. Cardaliaguet, and R. Monneau, *Existence and uniqueness for dislocation dynamics with nonnegative velocity*, Interfaces Free Bound., 7 (2005), pp. 415–434.

[3] O. Alvarez, P. Hoch, Y. Le Bouar, and R. Monneau, *Dislocation dynamics: Short-time existence and uniqueness of the solution*, Arch. Ration. Mech. Anal., 181 (2006), pp. 449–504.

[4] G. Barles, *A new stability result for viscosity solutions of nonlinear parabolic equations with weak convergence in time*, C. R. Math. Acad. Sci. Paris, 343 (2006), pp. 173–178.

[5] G. Barles and O. Ley, *Nonlocal first-order Hamilton-Jacobi equations modelling dislocations dynamics*, Comm. Partial Differential Equations, 31 (2006), pp. 1191–1208.

[6] G. Barles, H. M. Soner, and P. E. Souganidis, *Front propagation and phase field theory*, SIAM J. Control Optim., 31 (1993), pp. 439–469.

[7] M. Bourgoing, *Viscosity solutions of fully nonlinear second order parabolic equations with L1-time dependence and Neumann boundary conditions*, Discrete Contin. Dyn. Syst., to appear.

[8] M. Bourgoing, *Viscosity solutions of fully nonlinear second order parabolic equations with L1-time dependence and Neumann boundary conditions. Existence and applications to the level-set approach*, Discrete Contin. Dyn. Syst., to appear.

[9] P. Cannarsa and P. Cardaliaguet, *Perimeter estimates for reachable sets of control systems*, J. Convex Anal., 13 (2006), pp. 253–267.

[10] P. Cannarsa and H. Frankowska, *Interior sphere property of attainable sets and time optimal control problems*, ESAIM Control Optim. Calc. Var., 12 (2006), pp. 350–370.

[11] P. Cardaliaguet, *On front propagation problems with nonlocal terms*, Adv. Differential Equations, 5 (2000), pp. 213–268.

[12] P. Cardaliaguet and C. Marchi, *Regularity of the eikonal equation with Neumann boundary conditions in the plane: Application to fronts with nonlocal terms*, SIAM J. Control Optim., 45 (2006), pp. 1017–1038.

[13] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley and Sons, New York, 1983.

[14] M. G. Crandall and P.-L. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[15] F. Da Lio, N. Forcadel, and R. Monneau, *Convergence of a non-local eikonal equation to anisotropic mean curvature motion. Application to dislocations dynamics*, J. Eur. Math. Soc. (JEMS), to appear.

[16] F. Da Lio, C. I. Kim, and D. Slepčev, *Nonlocal front propagation problems in bounded domains with Neumann-type boundary conditions and applications*, Asymptot. Anal., 37 (2004), pp. 257–292.

[17] N. Forcadel, *Dislocation Dynamics with a Mean Curvature Term: Short Time Existence and Uniqueness*, preprint, 2005.

[18] Y. Giga, *Surface Evolution Equations: A Level-Set Method*, Monogr. Math. 99, Birkhäuser Verlag, Basel, 2006.

[19] J. R. Hirth and L. Lothe, *Theory of Dislocations*, 2nd ed., Krieger, Malabar, FL, 1992.

[20] H. Ishii, *Hamilton-Jacobi equations with discontinuous Hamiltonians on arbitrary open sets*, Bull. Fac. Sci. Engrg. Chuo Univ., 28 (1985), pp. 33–77.

[21] R.W. Lardner, *Mathematical Theory of Dislocations and Fracture*, Mathematical Expositions 17, University of Toronto Press, Toronto, ON, Canada, 1974.
[22] O. Ley, *Lower-bound gradient estimates for first-order Hamilton-Jacobi equations and applications to the regularity of propagating fronts*, Adv. Differential Equations, 6 (2001), pp. 547–576.
[23] P.-L. Lions and B. Perthame, *Remarks on Hamilton-Jacobi equations with measurable time-dependent Hamiltonians*, Nonlinear Anal., 11 (1987), pp. 613–621.
[24] F. R. N. Nabarro, *Theory of Crystal Dislocations*, Clarendon Press, Oxford, UK, 1969.
[25] D. Nunziante, *Uniqueness of viscosity solutions of fully nonlinear second order parabolic equations with discontinuous time-dependence*, Differential Integral Equations, 3 (1990), pp. 77–91.
[26] D. Nunziante, *Existence and uniqueness of unbounded viscosity solutions of parabolic equations with discontinuous time-dependence*, Nonlinear Anal., 18 (1992), pp. 1033–1062.
[27] D. Rodney, Y. Le Bouar, and A. Finel, *Phase field methods and dislocations*, Acta Mater., 51 (2003), pp. 17–30.
[28] D. Slepčev, *Approximation schemes for propagation of fronts with nonlocal velocities and Neumann boundary conditions*, Nonlinear Anal., 52 (2003), pp. 79–115.
[29] A. Srour, *Nonlocal Second-Order Hamilton-Jacobi Equations Arising in Tomographic Reconstruction*, preprint, 2007.

SIAM J. MATH. ANAL.
Vol. 40, No. 1, pp. 70–102

REFINABLE FUNCTIONS AND CASCADE ALGORITHMS IN WEIGHTED SPACES WITH HÖLDER CONTINUOUS MASKS*

BIN HAN†

Abstract. Refinable functions and cascade algorithms play a fundamental role in wavelet analysis, which is useful in many applications. In this paper we shall study several properties of refinable functions, cascade algorithms, and wavelets, associated with Hölder continuous masks, in the weighted subspaces $L_{2,p,\gamma}(\mathbb{R})$ of $L_2(\mathbb{R})$, where $1 \leqslant p \leqslant \infty$, $\gamma \geqslant 0$ and $f \in L_{2,p,\gamma}(\mathbb{R})$ means $\|f\|_{L_{2,p,\gamma}(\mathbb{R})} := \|\sum_{k\in\mathbb{Z}} |\widehat{e^{\gamma|\cdot|}f}(\cdot+2\pi k)|^2\|^{1/2}_{L_p(\mathbb{T})} < \infty$. In particular, $\|f\|_{L_{2,1,\gamma}(\mathbb{R})} = \|fe^{\gamma|\cdot|}\|_{L_2(\mathbb{R})}$ and $\|f\|_{L_{2,\infty,0}(\mathbb{R})} = \|\sum_{k\in\mathbb{Z}} |\hat{f}(\cdot+2\pi k)|^2\|^{1/2}_{L_\infty(\mathbb{T})}$. For a mask $\hat{a} \in C^\beta(\mathbb{T})$ with $\beta > 0$ and $\hat{a}(0) = 1$ (that is, $\hat{a}$ is a Hölder continuous mask with Hölder exponent β), we prove that the cascade algorithm associated with the mask $\hat{a}$ converges in the space $L_{2,\infty,0}(\mathbb{R})$ if and only if $\nu_2(\hat{a}) > 0$, where the quantity $\nu_2(\hat{a})$ will be defined in this paper and plays an important role in our study of refinable functions and cascade algorithms with Hölder continuous masks. In particular, if the shifts of a refinable function ϕ, satisfying $\hat{\phi}(2\cdot) = \hat{a}\hat{\phi}$, are stable in $L_2(\mathbb{R})$, then we must have $\nu_2(\hat{a}) > 0$, and therefore the cascade algorithm associated with mask $\hat{a}$ converges in the space $L_{2,\infty,0}(\mathbb{R})$. Based on this result, we are able to settle several problems on refinable functions, cascade algorithms, and wavelets associated with masks having infinitely many nonzero Fourier coefficients. As an application of the characterization of the convergence of a cascade algorithm in the space $L_{2,\infty,0}(\mathbb{R})$, we are able to show that for a mask $\hat{a}$ having exponential decay of order $r > 0$, the cascade algorithm associated with mask $\hat{a}$ converges in the weighted space $L_{2,1,\gamma}(\mathbb{R})$ for $0 < \gamma < 2r$ if and only if $\nu_2(\hat{a}) > 0$. Consequently, if a mask $\hat{a}$ has exponential decay of order $r > 0$ and $\nu_2(\hat{a}) > 0$, then its standard refinable function ϕ, defined by $\hat{\phi}(\xi) := \prod_{j=1}^\infty \hat{a}(2^{-j}\xi)$, must have exponential decay of order $2r$ in $L_2(\mathbb{R})$; that is, $\|\phi\|^2_{L_{2,1,\gamma}(\mathbb{R})} = \int_{\mathbb{R}} |\phi(x)|^2 e^{2\gamma|x|}\,dx < \infty$ for all $0 < \gamma < 2r$. As another application of the characterization of the convergence of a cascade algorithm in the space $L_{2,\infty,0}(\mathbb{R})$, we completely characterize biorthogonal wavelets and Riesz wavelets in $L_2(\mathbb{R})$, which are derived from refinable functions and whose involved wavelet filters in the frequency domain are Hölder continuous. We shall also investigate some basic properties of the quantity $\nu_2(\hat{a})$ and discuss how to calculate and estimate $\nu_2(\hat{a})$. Examples using fractional splines and the Butterworth filters will be given to illustrate the results in this paper.

Key words. refinable functions, cascade algorithms, Hölder continuous masks, masks having infinitely many nonzero Fourier coefficients, transition operator, weighted L_2 spaces, biorthogonal wavelets, Riesz wavelets, exponential decay

AMS subject classifications. 42C40, 41A15, 46B15, 47B37

DOI. 10.1137/060661016

1. Introduction and motivation.

A wavelet system is generally derived from a refinable function via a multiresolution analysis (MRA). We say that ϕ is *a refinable function* if it satisfies the refinement equation:

$$\hat{\phi}(2\xi) = \hat{a}(\xi)\hat{\phi}(\xi) \qquad \text{a.e. } \xi \in \mathbb{R}, \tag{1.1}$$

where $\hat{a}$ is a 2π-periodic function, called the *mask* (or filter) for ϕ, and the Fourier transform is defined to be $\hat{f}(\xi) := \int_{\mathbb{R}} f(x)e^{-ix\xi}\,dx$ for $f \in L_1(\mathbb{R})$.

*Received by the editors May 26, 2006; accepted for publication (in revised form) December 4, 2007; published electronically March 26, 2008. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) under grant RGP 228051.
http://www.siam.org/journals/sima/40-1/66101.html
†Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton T6G 2G1, AB, Canada (bhan@math.ualberta.ca).

For a 2π-periodic function $\hat{a}$, we say that $\hat{a}$ has *exponential decay of order* r if $\hat{a}$ is the restriction of a 2π-periodic function $\hat{a}(z)$ on the real line $\mathrm{Im}(z) = 0$ such that $\hat{a}(z)$ is holomorphic in the strip $\{z \in \mathbb{C} : |\mathrm{Im}(z)| < r\}$, where $\mathrm{Im}(z)$ denotes the imaginary part of the complex number z. Write $\hat{a}(\xi) = \sum_{k\in\mathbb{Z}} a_k e^{-ik\xi}$ in terms of its Fourier series. It is easy to check that $\hat{a}$ has exponential decay of order r if and only if for every $0 \leqslant \gamma < r$, there is a positive constant C_γ such that $|a_k| \leqslant C_\gamma e^{-\gamma|k|}$ for all $k \in \mathbb{Z}$. A particular family of masks with exponential decay consists of rational masks that can be written in the form of $\hat{b}(\xi)/\hat{c}(\xi)$ for some 2π-periodic trigonometric polynomials $\hat{b}$ and $\hat{c}$ such that $\hat{c}(\xi) \neq 0$ for all $\xi \in \mathbb{R}$. For example, the well-known Butterworth filters are rational masks [5, 11, 18, 25].

Masks with infinitely many nonzero Fourier coefficients, or equivalently, filters with infinite support, are called infinite impulse response (IIR) filters in electrical engineering. Due to some desirable properties, IIR filters, including masks with exponential decay and masks for bandlimited wavelets [7] and fractional splines [27], are of interest in some applications and have been extensively designed for various purposes in the area of digital signal processing in electrical engineering [3, 5, 18, 25, 27]. In contrast to the case of trigonometric polynomial masks whose various mathematical properties have been extensively studied and more or less well understood in the literature (see [1, 4, 6, 7, 8, 10, 15, 19, 26, 28] and the references therein), there are still several unsolved questions related to refinable functions and wavelets with masks having infinitely many nonzero Fourier coefficients. For example, Daubechies and Huang in [9] showed that if the standard refinable function ϕ, defined by

$$\hat{\phi}(\xi) := \prod_{j=1}^{\infty} \hat{a}(2^{-j}\xi), \qquad \xi \in \mathbb{R}, \tag{1.2}$$

lies in $L_1(\mathbb{R})$ with a mask $\hat{a}$ having an absolutely summable sequence of Fourier coefficients, and if ϕ has exponential decay, then the mask $\hat{a}$ must have exponential decay. But to the best of our knowledge, there are very few results on the converse direction, and it is widely believed that for a mask $\hat{a}$ with exponential decay, its standard refinable function ϕ in (1.2) should also have exponential decay in certain spaces in some sense. This is one of our motivations to study refinable functions and cascade algorithms, associated with masks having infinitely many nonzero Fourier coefficients, in some weighted subspaces of $L_2(\mathbb{R})$.

Before proceeding further, let us introduce some definitions and notation. Let $\mathbb{T} := \mathbb{R}/[2\pi\mathbb{Z}]$ and $L_p(\mathbb{T})$ denote the linear space of all 2π-periodic measurable functions $f : \mathbb{R} \mapsto \mathbb{C}$ such that $2\pi\|f\|^p_{L_p(\mathbb{T})} := \int_{-\pi}^{\pi} |f(x)|^p\, dx < \infty$ for $1 \leqslant p < \infty$ and $\|f\|_{L_\infty(\mathbb{T})}$ denotes its essential upper bound. For $\gamma \geqslant 0$ and $1 \leqslant p \leqslant \infty$, throughout the paper, the space $L_{2,p,\gamma}(\mathbb{R})$ denotes the subspace of all $f \in L_2(\mathbb{R})$ such that

$$\|f\|_{L_{2,p,\gamma}(\mathbb{R})} := \left\| \left[\widehat{e^{\gamma|\cdot|}f}, \widehat{e^{\gamma|\cdot|}f}\right] \right\|^{1/2}_{L_p(\mathbb{T})} < \infty, \tag{1.3}$$

where the *bracket product* [20] is defined to be

$$[f, g](\xi) := \sum_{k\in\mathbb{Z}} f(\xi + 2\pi k)\overline{g(\xi + 2\pi k)}, \qquad \xi \in \mathbb{R},\ f, g \in L_2(\mathbb{R}). \tag{1.4}$$

It is easy to verify that $[f, g] \in L_2(\mathbb{T})$ for $f, g \in L_2(\mathbb{R})$ and $L_{2,p,\gamma}(\mathbb{R})$ is a Banach space. In particular, by Plancherel's theorem, we have

$$\|f\|^2_{L_{2,1,\gamma}(\mathbb{R})} = \|e^{\gamma|\cdot|}f\|^2_{L_2(\mathbb{R})} = \int_{\mathbb{R}} |f(x)|^2 e^{2\gamma|x|}\, dx.$$

Therefore, $L_{2,1,\gamma}(\mathbb{R})$ is a weighted subspace of $L_2(\mathbb{R})$, and it is a natural candidate of subspaces to measure the exponential decay of a function in $L_2(\mathbb{R})$. Note that $L_{2,1,0}(\mathbb{R}) = L_2(\mathbb{R})$.

We say that the shifts of a function f are *stable* in $L_2(\mathbb{R})$ if there exists a positive constant C such that $C^{-1} \leqslant [\hat{f}, \hat{f}](\xi) \leqslant C$ for almost every $\xi \in \mathbb{R}$. It is well known that stability plays an important role in wavelet analysis [3, 4, 7]. If the shifts of a function f are stable in $L_2(\mathbb{R})$, then it is obvious that $f \in L_{2,\infty,0}(\mathbb{R})$, since $\|f\|_{L_{2,\infty,0}(\mathbb{R})} := \|[\hat{f}, \hat{f}]\|^{1/2}_{L_\infty(\mathbb{T})} < \infty$. On the other hand, we shall see in Proposition 6.1 that $L_{2,1,\gamma}(\mathbb{R}) \subseteq L_{2,\infty,0}(\mathbb{R})$ for any $\gamma > 0$. In particular, all compactly supported functions in $L_2(\mathbb{R})$ are included in $L_{2,\infty,0}(\mathbb{R})$. Therefore, the space $L_{2,\infty,0}(\mathbb{R})$ is a large enough subspace of $L_2(\mathbb{R})$ and includes most interesting functions in wavelet analysis. In this paper, we are particularly interested in the subspaces $L_{2,\infty,0}(\mathbb{R})$ and $L_{2,1,\gamma}(\mathbb{R})$ for $\gamma > 0$.

In the following, let us introduce a basic quantity $\nu_2(\hat{a})$, which plays a critical role in our study of refinable functions, cascade algorithms, and wavelets with masks having infinitely many nonzero Fourier coefficients. For 2π-periodic functions $\hat{a}$ and f, the *transition operator* $T_{\hat{a}}$ is defined to be

$$[T_{\hat{a}} f](\xi) := |\hat{a}(\xi/2)|^2 f(\xi/2) + |\hat{a}(\xi/2+\pi)|^2 f(\xi/2+\pi), \qquad \xi \in \mathbb{R}. \tag{1.5}$$

For $\tau \in \mathbb{R}$ and $1 \leqslant p \leqslant \infty$, we define a quantity

$$\rho_\tau(\hat{a}, p) := \limsup_{n\to\infty} \|T^n_{\hat{a}} \left(|\sin(\cdot/2)|^\tau\right)\|^{1/n}_{L_p(\mathbb{T})}. \tag{1.6}$$

Now we define the quantity $\nu_2(\hat{a})$ in this paper as follows:

$$\nu_2(\hat{a}) := -[\log_2 \rho(\hat{a})]/2, \tag{1.7}$$

where

$$\rho(\hat{a}) := \inf\{\rho_\tau(\hat{a}, \infty) \ : \ |\hat{a}(\cdot+\pi)|^2/|\sin(\cdot/2)|^\tau \in L_\infty(\mathbb{T}) \ \text{ and } \ \tau \geqslant 0\}. \tag{1.8}$$

When $\hat{a}$ is a 2π-periodic trigonometric polynomial, the quantity $\nu_2(\hat{a})$ in (1.7) agrees with the one in [13] and can be calculated by finding the spectral radius of an associated finite matrix generated by $\hat{a}$. See section 4 for details on calculating and estimating the quantity $\nu_2(\hat{a})$. The quantity $\nu_2(\hat{a})$ in (1.7), whose definition appears to be a little bit technical in its nature, plays a very important role in investigating many problems in wavelet analysis. See [13] for applications and the importance of $\nu_2(\hat{a})$ in wavelet analysis with 2π-periodic trigonometric polynomial masks $\hat{a}$.

We say that f belongs to the *Hölder class* $C^\beta(\mathbb{T})$ with $\beta > 0$ if f is a 2π-periodic continuous function such that $f \in C^n(\mathbb{T})$ and there exists a positive constant C satisfying $|f^{(n)}(x) - f^{(n)}(y)| \leqslant C|x-y|^{\beta-n}$ for all $x, y \in \mathbb{T}$, where n is the largest integer such that $n \leqslant \beta$ and $f^{(n)}$ denotes the nth derivative of f. Throughout the paper, we say that $\hat{a}$ is *a Hölder continuous mask* if $\hat{a} \in C^\beta(\mathbb{T})$ for some $\beta > 0$ and $\hat{a}(0) = 1$. That $\hat{a}$ is a Hölder continuous mask is a very natural and weak condition to guarantee that as the Fourier transform of the standard refinable function ϕ with mask $\hat{a}$, the function $\hat{\phi}$, which is defined through the infinite product in (1.2), is well defined.

In section 2, we shall present a necessary and sufficient condition in Theorem 2.1 for the convergence of a cascade algorithm in the space $L_{2,\infty,0}(\mathbb{R})$. Theorem 2.1 plays a central role in our study of refinable functions with exponential decay in $L_2(\mathbb{R})$ and

of MRA Riesz wavelet bases in $L_2(\mathbb{R})$. More precisely, we prove in Theorem 2.1 that for a mask $\hat{a} \in C^\beta(\mathbb{T})$ with $\hat{a}(0) = 1$ and $\beta > 0$ (that is, $\hat{a}$ is a Hölder continuous mask), the cascade algorithm associated with mask $\hat{a}$ converges in the space $L_{2,\infty,0}(\mathbb{R})$ if and only if $\nu_2(\hat{a}) > 0$. As a direct consequence of Theorem 2.1, we show in Corollary 2.2 that if the refinement equation $\hat{\phi}(2\cdot) = \hat{a}\hat{\phi}$ has a solution ϕ such that the shifts of ϕ are stable in $L_2(\mathbb{R})$, then we must have $\nu_2(\hat{a}) > 0$, and therefore the cascade algorithm associated with mask $\hat{a}$ converges in the space $L_{2,\infty,0}(\mathbb{R})$.

Cascade algorithms in $l_2(\mathbb{R})$ and other spaces with 2π-periodic trigonometric polynomial masks have been extensively studied in the literature. To cite only a few references here, we refer the reader to [1, 4, 6, 7, 8, 10, 15, 19, 22, 26, 28] and the references therein, where the property of the masks being 2π-periodic trigonometric polynomials plays a critical role. The study of cascade algorithms and refinable functions with Hölder continuous masks in this paper is not a trivial generalization of the known results in the literature, as we shall see in sections 4–6.

As an application of Theorem 2.1, we are able to prove in Theorem 2.3 that for a mask $\hat{a}$ having exponential decay of order $r > 0$, the cascade algorithm associated with the mask $\hat{a}$ converges in the space $L_{2,1,\gamma}(\mathbb{R})$ for $0 < \gamma < 2r$ if and only if $\nu_2(\hat{a}) > 0$. Consequently, the standard refinable function ϕ associated with mask $\hat{a}$ in (1.2) must have exponential decay of order $2r$, namely,

$$\|\phi\|^2_{L_{2,1,\gamma}(\mathbb{R})} = \int_{\mathbb{R}} |\phi(x)|^2 e^{2\gamma|x|}\, dx < \infty \qquad \forall\, 0 \leqslant \gamma < 2r. \tag{1.9}$$

An MRA wavelet function ψ is obtained from a refinable function ϕ with mask $\hat{a}$ via

$$\hat{\psi}(2\xi) := \hat{b}(\xi)\hat{\phi}(\xi), \qquad \xi \in \mathbb{R}, \tag{1.10}$$

for some 2π-periodic measurable function $\hat{b}$. We say that ψ generates *a Riesz wavelet basis* in $L_2(\mathbb{R})$ if $\operatorname{span}\{\psi_{j,k} := 2^{j/2}\psi(2^j \cdot -k) \,:\, j,k \in \mathbb{Z}\}$ is dense in $L_2(\mathbb{R})$ and there exists a positive constant C such that

$$C^{-1} \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} |c_{j,k}|^2 \leqslant \left\| \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} c_{j,k}\psi_{j,k} \right\|^2_{L_2(\mathbb{R})} \leqslant C \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} |c_{j,k}|^2 \tag{1.11}$$

for all finitely supported sequences $\{c_{j,k}\}_{j,k\in\mathbb{Z}}$. MRA Riesz wavelet bases in $L_2(\mathbb{R})$ are of interest in some applications [2, 5, 17, 21, 24]. A natural and important question here is when ψ generates a Riesz wavelet basis in $L_2(\mathbb{R})$. MRA Riesz wavelet bases with compact support have been investigated in [5, 14, 16, 17, 21, 24], where some necessary and sufficient conditions have been obtained for trigonometric polynomial masks. Most approaches in these papers rely largely on an interesting result of Cohen and Daubechies in [5] saying that for a mask $\hat{a}$ with exponential decay, the transition operator $T_{\hat{a}}$ acting on some weighted subspaces of $\ell_2(\mathbb{Z})$ is a compact operator. Built on this interesting result of [5], a characterization of Riesz wavelets with trigonometric polynomial masks is obtained in [16] (also cf. [5]) in terms of the spectrum of $T_{\hat{a}}$. However, the approach in [5, 16, 24] seems difficult, if not impossible, to be generalized to masks without exponential decay, since the compactness of the operator $T_{\hat{a}}$ may be lost.

In order to study biorthogonal wavelets and Riesz wavelets with Hölder continuous masks, using a quite different approach in this paper, as another application

of Theorem 2.1, we shall prove in Theorem 3.2 that for $\hat{a}, \hat{b} \in C^\beta(\mathbb{T})$ with $\beta > 0$, assuming that the shifts of the standard refinable function ϕ with mask $\hat{a}$ are stable in $L_2(\mathbb{R})$, then ψ in (1.10) generates a Riesz wavelet basis in $L_2(\mathbb{R})$ if and only if (i) $\hat{b}(0) = 0$ and $d(\xi) := \hat{a}(\xi)\overline{\hat{b}(\xi+\pi)} - \hat{a}(\xi+\pi)\hat{b}(\xi) \neq 0$ for all $\xi \in \mathbb{R}$; (ii) $\nu_2(\hat{a}) > 0$ and $\nu_2(\hat{\tilde{a}}) > 0$, where $\hat{\tilde{a}}(\xi) := \overline{\hat{b}(\xi+\pi)}/\overline{d(\xi)}$. Moreover, in the case that $\hat{a}$ is a 2π-periodic trigonometric polynomial, we prove that the shifts of $\phi \in L_2(\mathbb{R})$ must be stable in $L_2(\mathbb{R})$ if ψ generates a Riesz wavelet basis in $L_2(\mathbb{R})$. Even for the case that both $\hat{a}$ and $\hat{b}$ are 2π-periodic trigonometric polynomials, the mask $\hat{\tilde{a}}$ is generally not a 2π-periodic trigonometric polynomial, and consequently refinable functions with masks being nontrigonometric polynomials will naturally appear in our study of MRA Riesz wavelet bases in $L_2(\mathbb{R})$. This is another motivation for us to study refinable functions, cascade algorithms, and wavelets with masks having infinitely many nonzero Fourier coefficients.

To illustrate the results in this paper, we shall apply the results in sections 2 and 3 to a family of Hölder continuous masks $\widehat{a_{\beta_1,\beta_2,\beta_3}}$ and $|\widehat{a_{\beta_1,\beta_2,\beta_3}}|$, where

$$\widehat{a_{\beta_1,\beta_2,\beta_3}}(\xi) := \frac{2^{-2\beta_1}(1+e^{-i\xi})^{2\beta_1}}{\big(|\cos(\xi/2)|^{2\beta_2} + |\sin(\xi/2)|^{2\beta_2}\big)^{\beta_3}}, \qquad \beta_1, \beta_2, \beta_3 > 0. \tag{1.12}$$

In fact, the masks for the B-splines correspond to the case $\widehat{a_{\beta_1,\beta_2,\beta_3}}$ with $\beta_2 = 1$ and $2\beta_1 \in \mathbb{N}$. The masks for various types of fractional splines in [27] correspond to the case $\widehat{a_{\beta_1,\beta_2,\beta_3}}$ or $|\widehat{a_{\beta_1,\beta_2,\beta_3}}|$ with $\beta_2 = 1$. The classical Butterworth filters in [5, 11, 18, 25] correspond to the case $|\widehat{a_{\beta_1,\beta_2,\beta_3}}|$ with $\beta_3 = 1$ and $\beta_1 = \beta_2 \in \mathbb{N}$. To illustrate the main results in sections 2 and 3, we shall study the convergence of cascade algorithms and MRA Riesz wavelet bases associated with the masks given in (1.12).

Since the quantity $\nu_2(\hat{a})$ plays a very important role in our study of refinable functions, cascade algorithms, and wavelets with Hölder continuous masks, we shall investigate in section 4 some basic properties of the quantity $\nu_2(\hat{a})$ and discuss how to calculate and estimate the quantity $\nu_2(\hat{a})$ for a mask $\hat{a}$ being a general Lebesgue measurable 2π-periodic function. In section 4, we shall generalize a well-known result on $\nu_2(\hat{a})$, whose proof in the general case of Hölder continuous masks is nontrivial and will be presented in the last section of this paper. The general discussion on the quantity $\nu_2(\hat{a})$ in section 4 is of interest in its own right, and the results in section 4 may be useful elsewhere.

For simplicity of presentation and readability of this paper, the proofs of Theorems 2.1 and 2.3 in section 2, which are a little bit technical in their nature, will be postponed to sections 5 and 6, respectively. The results in this paper can be nontrivially generalized to high dimensions and multiwavelets, which we shall discuss elsewhere.

2. Convergence of cascade algorithms in subspaces of $L_2(\mathbb{R})$. In this section, we shall present the main results on the convergence of cascade algorithms in the subspaces $L_{2,\infty,0}(\mathbb{R})$ and $L_{2,1,\gamma}(\mathbb{R})$. For simplicity of presentation, the proofs of the main results in this section will be postponed to sections 5 and 6. To illustrate the results in this section, we shall apply these results to the masks in (1.12), which include the Butterworth filters in [25] and the masks for fractional splines in [27] as special cases.

For a quotient function f/g, throughout the paper, we use the convention that $(f/g)(\xi)$ is equal to $f(\xi)/g(\xi)$ if $g(\xi) \neq 0$, 1 if $f(\xi) = g(\xi) = 0$, or $+\infty$ if $g(\xi) = 0$ but

$f(\xi) \neq 0$. We say that a function $f \in L_2(\mathbb{R})$ is *admissible with respect to* $\hat{a}$ if there exists a positive number $\tau > 0$ such that

$$[\hat{a}(\cdot/2)\hat{f}(\cdot/2) - \hat{f}, \hat{a}(\cdot/2)\hat{f}(\cdot/2) - \hat{f}]/|\sin(\cdot/2)|^\tau \in L_\infty(\mathbb{T}). \tag{2.1}$$

Note that all compactly supported functions in $L_2(\mathbb{R})$ belong to $L_{2,1,\gamma}(\mathbb{R})$ for all $\gamma > 0$. For every $f \in L_{2,1,\gamma}(\mathbb{R})$ with $\gamma > 0$ such that $\hat{f}(2\pi k) = 0$ for all $k \in \mathbb{Z}\backslash\{0\}$, we shall show in Proposition 6.2 that f is admissible with respect to $\hat{a}$ for any $\hat{a} \in C^\beta(\mathbb{T})$ with $\hat{a}(0) = 1$, $\hat{a}(\pi) = 0$, and $\beta > 0$.

For $0 < \beta \leqslant 1$, we say that f belongs to the Lipschitz class $\Lambda^\beta(\mathbb{T})$ if there is a positive constant C such that $|f(x) - f(y)| \leqslant C|x - y|^\beta$ for all $x, y \in \mathbb{T}$.

Now we have the following result on the convergence of cascade algorithms in the space $L_{2,\infty,0}(\mathbb{R})$, which plays a central role in this paper and whose proof will be given in section 5.

THEOREM 2.1. *Let $\hat{a} \in C^\beta(\mathbb{T})$ with $\hat{a}(0) = 1$ and $\beta > 0$ (that is, $\hat{a}$ is a Hölder continuous mask). Then the following are equivalent:*

(i) *$\hat{a}(\pi) = 0$, and for every admissible function $f \in L_{2,\infty,0}(\mathbb{R})$ with respect to $\hat{a}$, $\{f_n\}_{n=1}^\infty$ is a Cauchy sequence in $L_{2,\infty,0}(\mathbb{R})$, where the functions f_n are defined by*

$$\widehat{f_n}(\xi) := \hat{a}(\xi/2)\widehat{f_{n-1}}(\xi/2) = \hat{f}(2^{-n}\xi)\prod_{j=1}^{n}\hat{a}(2^{-j}\xi), \qquad \xi \in \mathbb{R}, n \in \mathbb{N}. \tag{2.2}$$

(ii) *For one admissible function $f \in L_{2,\infty,0}(\mathbb{R})$ with respect to $\hat{a}$ such that the shifts of f are stable, $\{f_n\}_{n=1}^\infty$ is a Cauchy sequence in $L_{2,\infty,0}(\mathbb{R})$.*

(iii) *For every $\tau > 0$, $\rho_\tau(\hat{a}, \infty) < 1$, where $\rho_\tau(\hat{a}, \infty)$ is defined in* (1.6).

(iv) *$\nu_2(\hat{a}) > 0$; that is, $\rho(\hat{a}) < 1$, where $\nu_2(\hat{a})$ and $\rho(\hat{a})$ are defined in* (1.7) *and* (1.8)*, respectively.*

(v) *For at least one $\tau > 0$, $\rho_\tau(\hat{a}, \infty) < 1$ and $|\hat{a}(\cdot + \pi)|^2/|\sin(\cdot/2)|^\tau \in L_\infty(\mathbb{T})$.*

Let ϕ denote the standard refinable function associated with the mask $\hat{a}$ in (1.2)*. If $\nu_2(\hat{a}) > 0$, then $\hat{\phi}$ belongs to the Lipschitz class $\Lambda^{\min(1,\beta)}(\mathbb{R})$, $[\hat{\phi}, \hat{\phi}] \in \Lambda^{\min(1,\beta)}(\mathbb{T})$, and for any $0 \leqslant \nu < \nu_2(\hat{a})$,*

$$[\hat{\phi}, \hat{\phi}]_\nu := \sum_{k\in\mathbb{Z}}(1 + |\cdot + 2\pi k|^2)^\nu |\hat{\phi}(\cdot + 2\pi k)|^2 \in C(\mathbb{T}). \tag{2.3}$$

In fact, by a more complicated argument, we could show in Theorem 2.1 that $\hat{\phi} \in C^\beta(\mathbb{R})$ and $[\hat{\phi}, \hat{\phi}] \in C^\beta(\mathbb{T})$, which we shall address elsewhere. According to the proof of Theorem 2.1 in section 5, (ii) implies $\nu_2(\hat{a}) > 0$ without the admissibility condition on f; that is, if $\{f_n\}_{n=1}^\infty$ is a Cauchy sequence in $L_{2,\infty,0}(\mathbb{R})$ for a function $f \in L_{2,\infty,0}(\mathbb{R})$ with stability, then $\nu_2(\hat{a}) > 0$. We say that the cascade algorithm associated with a mask $\hat{a}$ converges in a given function space if for every admissible function f in that space with respect to $\hat{a}$, the sequence $\{f_n\}_{n=1}^\infty$ defined in (2.2) is a Cauchy sequence in that space.

As a direct consequence of Theorem 2.1, we have the following corollary.

COROLLARY 2.2. *Let $\hat{a} \in C^\beta(\mathbb{T})$ with $\hat{a}(0) = 1$ and $\beta > 0$. Suppose that ϕ is a (not necessarily the standard) refinable function such that $\hat{\phi}(2\xi) = \hat{a}(\xi)\hat{\phi}(\xi)$ a.e. $\xi \in \mathbb{R}$ and the shifts of ϕ are stable in $L_2(\mathbb{R})$. Then $\nu_2(\hat{a}) > 0$ and the cascade algorithm associated with mask $\hat{a}$ must converge in the space $L_{2,\infty,0}(\mathbb{R})$.*

Proof. Since the shifts of ϕ are stable in $L_2(\mathbb{R})$, we have $[\hat{\phi}, \hat{\phi}] \in L_\infty(\mathbb{T})$, and so $\phi \in L_{2,\infty,0}(\mathbb{R})$. Since $\hat{a}(\cdot/2)\hat{\phi}(\cdot/2) - \hat{\phi} = 0$, ϕ is an admissible function in $L_{2,\infty,0}(\mathbb{R})$

with respect to $\hat{a}$. So, for $f = \phi$, (ii) of Theorem 2.1 holds since $f_n = \phi$ for all $n \in \mathbb{N}$. Now by Theorem 2.1, $\nu_2(\hat{a}) > 0$. □

For a mask with exponential decay, using Theorem 2.1, we characterize the convergence of a cascade algorithm with an exponentially decaying mask in the spaces $L_{2,p,\gamma}(\mathbb{R})$ in the following result, whose proof will be given in section 6.

THEOREM 2.3. *Let $\hat{a}$ be a mask such that $\hat{a}(0) = 1$ and $\hat{a}$ has exponential decay of order r for some $r > 0$. Then the following are equivalent:*

(i) *$\hat{a}(\pi) = 0$, and for every $0 < \gamma < 2r$ and every admissible function $f \in L_{2,1,\gamma}(\mathbb{R})$ with respect to $\hat{a}$, $\{f_n\}_{n=1}^{\infty}$ is a Cauchy sequence in $L_{2,1,\gamma}(\mathbb{R})$, where f_n are defined in (2.2).*

(ii) *$\hat{a}(\pi) = 0$, and for every $0 < \gamma < 2r$, every $1 \leqslant p \leqslant \infty$, and every admissible function $f \in L_{2,p,\gamma}(\mathbb{R})$ with respect to $\hat{a}$, $\{f_n\}_{n=1}^{\infty}$ is a Cauchy sequence in $L_{2,p,\gamma}(\mathbb{R})$.*

(iii) *$\hat{a}(\pi) = 0$, and for some $0 < \gamma < 2r$, some $1 \leqslant p \leqslant \infty$, and every admissible function $f \in L_{2,p,\gamma}(\mathbb{R})$ with respect to $\hat{a}$, $\{f_n\}_{n=1}^{\infty}$ is a Cauchy sequence in $L_{2,p,\gamma}(\mathbb{R})$.*

(iv) *For some $0 < \gamma < 2r$, some $1 \leqslant p \leqslant \infty$, and one admissible function $f \in L_{2,p,\gamma}(\mathbb{R})$ with respect to $\hat{a}$ such that the shifts of f are stable in $L_2(\mathbb{R})$, the sequence $\{f_n\}_{n=1}^{\infty}$ is a Cauchy sequence in $L_{2,p,\gamma}(\mathbb{R})$.*

(v) *$\nu_2(\hat{a}) > 0$.*

In particular, if $\nu_2(\hat{a}) > 0$ and $\hat{a}$ has exponential decay of order r, then the standard refinable function ϕ with mask $\hat{a}$ in (1.2) must have exponential decay of order $2r$; that is, (1.9) holds.

In the following, we shall apply the above results to the masks in (1.12), which include both the classical Butterworth filters in [25] and the masks for the fractional splines in [27] as special cases.

Example 2.4. Let $\hat{a} = \widehat{a_{\beta_1,\beta_2,\beta_3}}$ or $\hat{a} = |\widehat{a_{\beta_1,\beta_2,\beta_3}}|$, where the masks $a_{\beta_1,\beta_2,\beta_3}$ are defined in (1.12). By Proposition 4.2, it is easy to check that $\hat{a} \in C^{\beta}(\mathbb{T})$ for some $\beta > 0$. Denote

$$\hat{B}(\xi) = (|\cos(\xi/2)|^{2\beta_2} + |\sin(\xi/2)|^{2\beta_2})^{\beta_3}.$$

Then $|\hat{a}(\xi)| = 2^{-2\beta_1}|1 + e^{-i\xi}|^{2\beta_1}/\hat{B}(\xi)$. By calculation, it is easy to deduce that

$$\min(1, 2^{1-\beta_2}) \leqslant |\cos(\xi/2)|^{2\beta_2} + |\sin(\xi/2)|^{2\beta_2} \leqslant \max(1, 2^{1-\beta_2}) \qquad \forall\, \xi \in \mathbb{R}. \tag{2.4}$$

Consequently, we have $\hat{B}(\xi) \geqslant \min(1, 2^{(1-\beta_2)\beta_3})$ for all $\xi \in \mathbb{R}$ and $\beta_2, \beta_3 > 0$.

For $0 < \beta_2 \leqslant 1$, by Theorem 4.1 and Lemma 4.3, it follows from $\hat{B}(\xi) \geqslant \min(1, 2^{(1-\beta_2)\beta_3}) = 1$ that

$$\rho(\hat{a}) = \rho_{4\beta_1}(\hat{a}, \infty) = \rho_0(2^{-2\beta_1}/\hat{B}, \infty) \leqslant \rho_0(2^{-2\beta_1}, \infty) = 2^{1-4\beta_1},$$

since for a constant c, $T_c^n 1 = 2^n |c|^{2n}$, and therefore

$$\rho_0(c, \infty) = \limsup_{n\to\infty} \|T_c^n 1\|_{L_\infty(\mathbb{T})}^{1/n} = 2|c|^2.$$

For $\beta_2 > 1$, by Theorem 4.1 and Lemma 4.3, it follows from $\hat{B}(\xi) \geqslant 2^{(1-\beta_2)\beta_3}$ that

$$\rho(\hat{a}) = \rho_{4\beta_1}(\hat{a}, \infty) = \rho_0(2^{-2\beta_1}/\hat{B}, \infty) \leqslant \rho_0(2^{-2\beta_1-(1-\beta_2)\beta_3}, \infty) = 2^{1-4\beta_1-2(1-\beta_2)\beta_3}.$$

Therefore, for

$$\beta_1, \beta_2, \beta_3 > 0 \quad \text{satisfying} \quad \beta_1 > 1/4 + \max(0, (\beta_2 - 1)\beta_3/2), \tag{2.5}$$

we have $\rho(\hat{a}) < 1$; that is, $\nu_2(\hat{a}) > 0$. By Theorem 2.1, the cascade algorithm associated with mask $\hat{a}$ converges in the space $L_{2,\infty,0}(\mathbb{R})$.

The classical Butterworth filters in [25] correspond to the case $\beta_3 = 1$ and $\beta_1 = \beta_2 \in \mathbb{N}$. The condition in (2.5) holds for all $\beta_3 = 1$ and $\beta_1 = \beta_2 \in \mathbb{N}$, that is, holds for all Butterworth filters.

For the fractional splines in [27], since $\beta_2 = 1$, the condition in (2.5) becomes $\beta_1 > 1/4$. Note that when $\beta_2 = 1$, the standard refinable function ϕ associated with mask $\hat{a}$ in (1.2) satisfies $|\hat{\phi}(2\xi)| = |(\sin\xi)/\xi|^{2\beta_1}$. Clearly, if $0 < \beta_1 \leqslant 1/4$, then $\hat{\phi} \notin L_2(\mathbb{R})$. So, the condition in (2.5), that is, $\beta_1 > 1/4$, is sharp for the case of fractional splines in [27].

Now we consider the case that $\hat{a}$ has exponential decay; that is, $\hat{a}(\xi) = \widehat{a_{\beta_1,\beta_2,\beta_3}}(\xi)$ with $2\beta_1 \in \mathbb{N}$ and $\beta_2 \in \mathbb{N}$. In this case, by the definition of the masks $\widehat{a_{\beta_1,\beta_2,\beta_3}}$ in (1.12), it is easy to see that $\hat{a}$ can be extended into a holomorphic function on some strip $\{z \in \mathbb{C} \ : \ |\mathrm{Im}(z)| < r\}$ for some $r > 0$ depending only on β_2. So, $\hat{a}$ has exponential decay of order r. Now by Theorem 2.3, if (2.5) holds with $2\beta_1 \in \mathbb{N}$ and $\beta_2 \in \mathbb{N}$, then the cascade algorithm associated with mask $\hat{a}$ converges in the spaces $L_{2,p,\gamma}(\mathbb{R})$ for all $0 \leqslant \gamma < 2r$ and the standard refinable function ϕ associated with mask $\hat{a}$ must have exponential decay of order $2r$ in $L_2(\mathbb{R})$.

3. Characterization of MRA biorthogonal wavelets and Riesz wavelets. In this section, using Theorem 2.1, we shall study MRA biorthogonal wavelets and Riesz wavelets with Hölder continuous masks.

Let us first recall the definition of biorthogonal wavelets in [6]. For two functions $\psi, \tilde{\psi} \in L_2(\mathbb{R})$, we say that $(\psi, \tilde{\psi})$ generates *a pair of biorthogonal wavelet bases* in $L_2(\mathbb{R})$ if each of ψ and $\tilde{\psi}$ generates a Riesz wavelet basis in $L_2(\mathbb{R})$ and the following biorthogonality relation holds:

$$\langle \psi_{j,k}, \tilde{\psi}_{j',k'} \rangle := \int_{\mathbb{R}} \psi_{j,k}(x)\overline{\tilde{\psi}_{j',k'}(x)}\,dx = \delta_{j-j'}\delta_{k-k'} \qquad \forall\, j, j', k, k' \in \mathbb{Z}, \tag{3.1}$$

where $\psi_{j,k} := 2^{j/2}\psi(2^j \cdot -k)$ and δ denotes the *Dirac sequence* such that $\delta_0 = 1$ and $\delta_k = 0$ for all $k \neq 0$. Compactly supported biorthogonal wavelets have been investigated in [4, 6, 7, 12] and other papers.

As an application of Theorem 2.1, we have the following result on biorthogonal wavelets with Hölder continuous masks.

THEOREM 3.1. *Let $\hat{a}, \hat{\tilde{a}} \in C^\beta(\mathbb{T})$ with $\hat{a}(0) = \hat{\tilde{a}}(0) = 1$ and $\beta > 0$. Define two refinable functions $\hat{\phi}$ and $\hat{\tilde{\phi}}$ associated with masks $\hat{a}$ and $\hat{\tilde{a}}$ by*

$$\hat{\phi}(\xi) := \prod_{j=1}^{\infty} \hat{a}(2^{-j}\xi) \quad \text{and} \quad \hat{\tilde{\phi}}(\xi) := \prod_{j=1}^{\infty} \hat{\tilde{a}}(2^{-j}\xi), \qquad \xi \in \mathbb{R}. \tag{3.2}$$

Then the following are equivalent:

(i) *$[\hat{\phi}, \hat{\phi}], [\hat{\tilde{\phi}}, \hat{\tilde{\phi}}] \in L_\infty(\mathbb{T})$ and $[\hat{\phi}, \hat{\tilde{\phi}}] = 1$; or equivalently, the shifts of both ϕ and $\tilde{\phi}$ are stable in $L_2(\mathbb{R})$ and the biorthogonality relation holds:*

$$\langle \phi, \tilde{\phi}(\cdot - k) \rangle = \int_{\mathbb{R}} \phi(x)\overline{\tilde{\phi}(x-k)}\,dx = \delta_k \qquad \forall\, k \in \mathbb{Z}. \tag{3.3}$$

(ii) $\nu_2(\hat{a})>0$, $\nu_2(\hat{\tilde{a}})>0$, *and* $\hat{\tilde{a}}$ *is a dual mask of* $\hat{a}$, *where we say that* $\hat{\tilde{a}}$ *is a dual mask of* $\hat{a}$ *if*

$$\overline{\hat{a}(\xi)}\hat{\tilde{a}}(\xi)+\overline{\hat{a}(\xi+\pi)}\hat{\tilde{a}}(\xi+\pi)=1. \tag{3.4}$$

Let $\hat{b},\hat{\tilde{b}}\in C^\beta(\mathbb{T})$ *such that* $\hat{b}(0)=\hat{\tilde{b}}(0)=0$. *Define two wavelet functions* ψ *and* $\tilde{\psi}$ *by*

$$\hat{\psi}(\xi):=\hat{b}(\xi/2)\hat{\phi}(\xi/2)\quad \textit{and}\quad \hat{\tilde{\psi}}(\xi):=\hat{\tilde{b}}(\xi/2)\hat{\tilde{\phi}}(\xi/2). \tag{3.5}$$

If $\nu_2(\hat{a})>0$, $\nu_2(\hat{\tilde{a}})>0$, *and*

$$\begin{bmatrix}\hat{a}(\xi) & \hat{a}(\xi+\pi)\\ \hat{b}(\xi) & \hat{b}(\xi+\pi)\end{bmatrix}\overline{\begin{bmatrix}\hat{\tilde{a}}(\xi) & \hat{\tilde{a}}(\xi+\pi)\\ \hat{\tilde{b}}(\xi) & \hat{\tilde{b}}(\xi+\pi)\end{bmatrix}}^T=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}, \tag{3.6}$$

then $(\psi,\tilde{\psi})$ *generates a pair of biorthogonal wavelet bases in* $L_2(\mathbb{R})$.

Proof. Suppose that (i) holds. By $1=|[\hat{\phi},\hat{\tilde{\phi}}]|^2\leqslant[\hat{\phi},\hat{\phi}][\hat{\tilde{\phi}},\hat{\tilde{\phi}}]$, we have $[\hat{\phi},\hat{\phi}]\geqslant\|[\hat{\tilde{\phi}},\hat{\tilde{\phi}}]\|^{-1}_{L_\infty(\mathbb{T})}$ and $[\hat{\tilde{\phi}},\hat{\tilde{\phi}}]\geqslant\|[\hat{\phi},\hat{\phi}]\|^{-1}_{L_\infty(\mathbb{T})}$. Therefore, the shifts of ϕ and $\tilde{\phi}$ are stable in $L_2(\mathbb{R})$. By Corollary 2.2, we have $\nu_2(\hat{a})>0$ and $\nu_2(\hat{\tilde{a}})>0$. Now by $[\hat{\phi},\hat{\tilde{\phi}}]=1$, it follows directly from the refinement equations $\hat{\phi}(2\xi)=\hat{a}(\xi)\hat{\phi}(\xi)$ and $\hat{\tilde{\phi}}(2\xi)=\hat{\tilde{a}}(\xi)\hat{\tilde{\phi}}(\xi)$ that

$$\begin{aligned}1=[\hat{\tilde{\phi}},\hat{\phi}](2\xi)&=\overline{\hat{a}(\xi)}\hat{\tilde{a}}(\xi)[\hat{\tilde{\phi}},\hat{\phi}](\xi)+\overline{\hat{a}(\xi+\pi)}\hat{\tilde{a}}(\xi+\pi)[\hat{\tilde{\phi}},\hat{\phi}](\xi+\pi)\\&=\overline{\hat{a}(\xi)}\hat{\tilde{a}}(\xi)+\overline{\hat{a}(\xi+\pi)}\hat{\tilde{a}}(\xi+\pi).\end{aligned}$$

So, $\hat{\tilde{a}}$ is a dual mask of $\hat{a}$. Therefore, (i)⇒(ii).

To prove (ii)⇒(i), since $\nu_2(\hat{a})>0$ and $\nu_2(\hat{\tilde{a}})>0$, by Theorem 2.1, we see that (i) of Theorem 2.1 holds for both $\hat{a}$ and $\hat{\tilde{a}}$. Moreover, by (2.3), we have $[\hat{\phi},\hat{\phi}],[\hat{\tilde{\phi}},\hat{\tilde{\phi}}]\in L_\infty(\mathbb{T})$.

Take $\hat{f}:=\chi_{[-\pi,\pi]}$, the characteristic function of the interval $[-\pi,\pi]$. By (i) of Theorem 2.1, we have $\hat{a}(\pi)=\hat{\tilde{a}}(\pi)=0$. Since $\hat{a},\hat{\tilde{a}}\in C^\beta(\mathbb{T})$ and $\hat{a}(0)=\hat{\tilde{a}}(0)=1$, it is easy to directly verify that f is an admissible function in $L_{2,\infty,0}(\mathbb{R})$ with respect to both $\hat{a}$ and $\hat{\tilde{a}}$. Define

$$\widehat{f_n}(\xi):=\hat{f}(2^{-n}\xi)\prod_{j=1}^n\hat{a}(2^{-j}\xi)\quad\text{and}\quad\widehat{\tilde{f}_n}(\xi):=\hat{f}(2^{-n}\xi)\prod_{j=1}^n\hat{\tilde{a}}(2^{-j}\xi),\qquad n\in\mathbb{N}.$$

Then by (i) of Theorem 2.1, both $\{f_n\}_{n=1}^\infty$ and $\{\tilde{f}_n\}_{n=1}^\infty$ are Cauchy sequences in $L_{2,\infty,0}(\mathbb{R})$. Note that $\lim_{n\to\infty}\widehat{f_n}(\xi)=\hat{\phi}(\xi)$ and $\lim_{n\to\infty}\widehat{\tilde{f}_n}(\xi)=\hat{\tilde{\phi}}(\xi)$. So, we must have

$$\lim_{n\to\infty}\|f_n-\phi\|_{L_{2,\infty,0}(\mathbb{R})}=\lim_{n\to\infty}\|\tilde{f}_n-\tilde{\phi}\|_{L_{2,\infty,0}(\mathbb{R})}=0.$$

Since $[\hat{f},\hat{f}]=1$ and the discrete biorthogonality relation in (3.4) holds, it is easy to show by induction that $[\widehat{f_n},\widehat{\tilde{f}_n}]=1$ for all $n\in\mathbb{N}$. Consequently, we must have $[\hat{\phi},\hat{\tilde{\phi}}]=1$. Therefore, (ii)⇒(i).

If $\nu_2(\hat{a})>0$, $\nu_2(\hat{\tilde{a}})>0$, and (3.6) holds, then all the conditions in (ii) hold. So, $[\hat{\phi},\hat{\tilde{\phi}}]=1$. Now it follows from (3.6) that $[\hat{\phi},\hat{\tilde{\psi}}]=0$, $[\hat{\psi},\hat{\tilde{\phi}}]=0$, and $[\hat{\psi},\hat{\tilde{\psi}}]=1$. By a

standard argument on MRA [4, 6], we deduce that (3.1) holds. Since $\hat{\phi}(0) = \hat{\tilde{\phi}}(0) = 1$ and both $\hat{\phi}$ and $\hat{\tilde{\phi}}$ are continuous, by the standard argument on MRA, we see that both $\{\psi_{j,k} : j,k \in \mathbb{Z}\}$ and $\{\tilde{\psi}_{j,k} : j,k \in \mathbb{Z}\}$ are dense in $L_2(\mathbb{R})$. To show that both ψ and $\tilde{\psi}$ generate Riesz wavelet bases in $L_2(\mathbb{R})$, we need to show that ψ and $\tilde{\psi}$ satisfy (1.11). By the biorthogonality relation in (3.1), it suffices to show that the right-hand inequality in (1.11) holds for both ψ and $\tilde{\psi}$ [4], which is equivalent to showing that there exists a positive constant C such that

$$\sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}}\big[|\langle f,\psi_{j,k}\rangle|^2 + |\langle f,\tilde{\psi}_{j,k}\rangle|^2\big] \leqslant C\|f\|^2 \qquad \forall\, f \in L_2(\mathbb{R}). \tag{3.7}$$

Since $\nu_2(\hat{a}) > 0$ and $\nu_2(\hat{\tilde{a}}) > 0$, by Theorem 2.1, we have $[\hat{\phi},\hat{\phi}]_\nu, [\hat{\tilde{\phi}},\hat{\tilde{\phi}}]_\nu \in L_\infty(\mathbb{T})$ for $0 < \nu < \min(\nu_2(\hat{a}), \nu_2(\hat{\tilde{a}}))$. Using Fourier transform and the Parseval's identity, by $\hat{\psi}(\xi) = \hat{b}(\xi/2)\hat{\phi}(\xi/2)$, we have

$$\begin{aligned}
2\pi\sum_{k\in\mathbb{Z}}|\langle f,\psi_{j,k}\rangle|^2 &= 2^j\int_{-\pi}^{\pi}|[\hat{f}(2^j\cdot),\hat{\psi}](\xi)|^2\,d\xi\\
&\leqslant 2^{j+1}\int_{-\pi}^{\pi}|\hat{b}(\xi)|^2|[\hat{f}(2^{j+1}\cdot),\hat{\phi}](\xi)|^2\,d\xi\\
&\leqslant 2^{j+1}\int_{-\pi}^{\pi}|\hat{b}(\xi)|^2[\hat{f}(2^{j+1}\cdot),\hat{f}(2^{j+1}\cdot)]_{-\nu}(\xi)[\hat{\phi},\hat{\phi}]_\nu(\xi)\,d\xi\\
&\leqslant \|[\hat{\phi},\hat{\phi}]_\nu\|_{L_\infty(\mathbb{T})}2^{j+1}\int_{-\pi}^{\pi}|\hat{b}(\xi)|^2[\hat{f}(2^{j+1}\cdot),\hat{f}(2^{j+1}\cdot)]_{-\nu}(\xi)\,d\xi\\
&= \|[\hat{\phi},\hat{\phi}]_\nu\|_{L_\infty(\mathbb{T})}\int_{\mathbb{R}}\frac{|\hat{b}(2^{-j-1}\xi)|^2}{(1+|2^{-j-1}\xi|^2)^\nu}|\hat{f}(\xi)|^2\,d\xi.
\end{aligned}$$

Note that $\hat{b}(0) = \hat{\tilde{b}}(0) = 0$ and $\hat{b},\hat{\tilde{b}} \in C^\beta(\mathbb{T})$ with $\beta > 0$. Consequently, we see that (3.7) holds with

$$\begin{aligned}
C = \|[\hat{\phi},\hat{\phi}]_\nu\|_{L_\infty(\mathbb{T})}&\left\|\sum_{j\in\mathbb{Z}}\frac{|\hat{b}(2^j\cdot)|^2}{(1+|2^j\cdot|^2)^\nu}\right\|_{L_\infty(\mathbb{R})}\\
&+\|[\hat{\tilde{\phi}},\hat{\tilde{\phi}}]_\nu\|_{L_\infty(\mathbb{T})}\left\|\sum_{j\in\mathbb{Z}}\frac{|\hat{\tilde{b}}(2^j\cdot)|^2}{(1+|2^j\cdot|^2)^\nu}\right\|_{L_\infty(\mathbb{R})} < \infty.
\end{aligned}$$

Therefore, $(\psi,\tilde{\psi})$ generates a pair of biorthogonal wavelet bases in $L_2(\mathbb{R})$. □

As an application of Theorems 2.1 and 3.1, we characterize MRA Riesz wavelet bases in $L_2(\mathbb{R})$ in the following result, which improves and generalizes [14, Theorem 6] and [16, Theorem 1.1] by taking a different approach.

THEOREM 3.2. *Let $\hat{a},\hat{b} \in C^\beta(\mathbb{T})$ with $\hat{a}(0) = 1$ and $\beta > 0$. Define ϕ and ψ by*

$$\hat{\phi}(\xi) := \prod_{j=1}^{\infty}\hat{a}(2^{-j}\xi) \quad \textit{and} \quad \hat{\psi}(\xi) := \hat{b}(\xi/2)\hat{\phi}(\xi/2), \qquad \xi \in \mathbb{R}. \tag{3.8}$$

Then the shifts of ϕ are stable in $L_2(\mathbb{R})$ and ψ generates a Riesz wavelet basis in $L_2(\mathbb{R})$ if and only if

(i) $\hat{b}(0) = 0$ *and* $d(\xi) := \hat{a}(\xi)\overline{\hat{b}(\xi+\pi)} - \hat{a}(\xi+\pi)\overline{\hat{b}(\xi)} \neq 0$ *for all* $\xi \in \mathbb{R}$,

(ii) $\nu_2(\hat{a}) > 0$ *and* $\nu_2(\hat{\tilde{a}}) > 0$, *where* $\hat{\tilde{a}}(\xi) := \overline{\hat{b}(\xi+\pi)}/\overline{d(\xi)}$.

In the case that $\hat{a}$ is a 2π-periodic trigonometric polynomial, then the shifts of $\phi \in L_2(\mathbb{R})$ must be stable in $L_2(\mathbb{R})$ if ψ generates a Riesz wavelet basis in $L_2(\mathbb{R})$.

Proof. Suppose that (i) and (ii) hold. Define $\hat{\tilde{b}}(\xi) := -\overline{\hat{a}(\xi+\pi)}/\overline{d(\xi)}$. Since $\hat{a}, \hat{b} \in C^\beta(\mathbb{T})$ and $d(\xi) \neq 0$ for all $\xi \in \mathbb{R}$, it is evident that $\hat{\tilde{a}}, \hat{\tilde{b}} \in C^\beta(\mathbb{T})$. By $\nu_2(\hat{a}) > 0$ and Theorem 2.1, $\hat{a}(\pi) = 0$. Since $\hat{a}(0) = 1$ and $\hat{b}(0) = 0$, we must have $\hat{\tilde{a}}(0) = 1$ and $\hat{\tilde{a}}(\pi) = 0$. Moreover, it is easy to check that (3.6) holds. Define $\tilde{\phi}$ and $\tilde{\psi}$ as in (3.2) and (3.5). Now it follows from Theorem 3.1 that $(\psi, \tilde{\psi})$ generates a pair of biorthogonal wavelet bases in $L_2(\mathbb{R})$. In particular, we conclude that ψ generates a Riesz wavelet basis in $L_2(\mathbb{R})$.

Conversely, suppose that the shifts of ϕ are stable in $L_2(\mathbb{R})$ and ψ generates a Riesz wavelet basis in $L_2(\mathbb{R})$. Since the shifts of ϕ are stable in $L_2(\mathbb{R})$, by Corollary 2.2, we have $\nu_2(\hat{a}) > 0$ and $\hat{a}(\pi) = 0$. Since $\hat{\psi}$ is continuous and ψ generates a Riesz wavelet basis in $L_2(\mathbb{R})$, we must have $\hat{\psi}(0) = 0$, and therefore $\hat{b}(0) = 0$ by $\hat{\phi}(0) = 1$. So, $\hat{\tilde{a}}(0) = \overline{\hat{b}(\pi)}/\overline{d(0)} = 1$. By [14, Lemma 1], we see that (i) must hold and there exists a function $\tilde{\phi} \in L_2(\mathbb{R})$ such that the shifts of $\tilde{\phi}$ are stable in $L_2(\mathbb{R})$ and $\hat{\tilde{\phi}}(2\xi) = \hat{\tilde{a}}(\xi)\hat{\tilde{\phi}}(\xi)$. Since $\hat{\tilde{a}} \in C^\beta(\mathbb{T})$ and $\hat{\tilde{a}}(0) = 1$, by Corollary 2.2, we have $\nu_2(\hat{\tilde{a}}) > 0$. Therefore, both (i) and (ii) hold.

Suppose that $\hat{a}$ is a 2π-periodic trigonometric polynomial and the shifts of $\phi \in L_2(\mathbb{R})$ are not stable in $L_2(\mathbb{R})$. Since ϕ is compactly supported, by [19, Theorem 5.3], there exists a compactly supported refinable function $\eta \in L_2(\mathbb{R})$ with stable shifts in $L_2(\mathbb{R})$ such that $\hat{\eta}(2\xi) = \hat{c}(\xi)\hat{\eta}(\xi)$ for some 2π-periodic trigonometric polynomial $\hat{c}$, and $\hat{\phi}(\xi) = \theta(\xi)\hat{\eta}(\xi)$ for some 2π-periodic trigonometric polynomial θ. Note that $[\hat{\phi}, \hat{\phi}](\xi) = |\theta(\xi)|^2[\hat{\eta}, \hat{\eta}](\xi)$ and $\hat{a}(\xi) = \theta(2\xi)\hat{c}(\xi)/\theta(\xi)$. Since the shifts of the compactly supported function ϕ are not stable in $L_2(\mathbb{R})$, there is $\xi_0 \in \mathbb{R}\backslash[2\pi\mathbb{Z}]$ such that $\theta(\xi_0) = 0$.

Since $\hat{\psi}(2\xi) = \hat{b}(\xi)\hat{\phi}(\xi) = \hat{b}(\xi)\theta(\xi)\hat{\eta}(\xi)$ and the shifts of η are stable in $L_2(\mathbb{R})$, if ψ generates a Riesz wavelet basis in $L_2(\mathbb{R})$, by what has been proved, then we must have

$$0 \neq \mathring{d}(\xi) := \hat{c}(\xi)\overline{\hat{b}(\xi+\pi)}\theta(\xi+\pi) - \hat{c}(\xi+\pi)\overline{\hat{b}(\xi)}\theta(\xi) = \frac{\theta(\xi)\theta(\xi+\pi)}{\theta(2\xi)}d(\xi) \quad \forall\, \xi \in \mathbb{R} \tag{3.9}$$

and $\nu_2(\hat{c}) > 0$, $\nu_2(\hat{\mathring{a}}) > 0$, where $\hat{\mathring{a}}(\xi) := \overline{\hat{b}(\xi+\pi)\theta(\xi+\pi)}/\overline{\mathring{d}(\xi)}$.

By (3.9), we conclude that if $\xi \in \mathbb{R}$ is a zero of θ, then 2ξ must also be a zero of θ. Since $\theta(\xi_0) = 0$, we now see that $\theta(2^j\xi_0) = 0$ for all $j \in \mathbb{N} \cup \{0\}$. By the definition of $\mathring{d}(\xi)$ in (3.9), for all $j \in \mathbb{N} \cup \{0\}$, we have

$$\begin{aligned}\mathring{d}(2^j\xi_0) &:= \hat{c}(2^j\xi_0)\overline{\hat{b}(2^j\xi_0+\pi)}\theta(2^j\xi_0+\pi) - \hat{c}(2^j\xi_0+\pi)\overline{\hat{b}(2^j\xi_0)}\theta(2^j\xi_0)\\ &= \hat{c}(2^j\xi_0)\overline{\hat{b}(2^j\xi_0+\pi)}\theta(2^j\xi_0+\pi).\end{aligned}$$

Now by the definition of $\hat{\hat{a}}$, we deduce that

$$\hat{\hat{a}}(2^j\xi_0) = \frac{\overline{\hat{b}(2^j\xi_0+\pi)\theta(2^j\xi_0+\pi)}}{\overline{\hat{c}(2^j\xi_0)\hat{b}(2^j\xi_0+\pi)\theta(2^j\xi_0+\pi)}} = \frac{1}{\overline{\hat{c}(2^j\xi_0)}} \qquad \forall\, j \in \mathbb{N}\cup\{0\},$$

from which we see that

$$\prod_{j=0}^{n} \hat{\hat{a}}(2^j\xi_0) = \frac{1}{\prod_{j=0}^{n}\overline{\hat{c}(2^j\xi_0)}} \qquad \forall\, n \in \mathbb{N}. \tag{3.10}$$

Since the shifts of η are stable, there exists $k_0 \in \mathbb{Z}$ such that $\hat{\eta}(\xi_0+2\pi k_0) \neq 0$. Since $\xi_0 \notin 2\pi\mathbb{Z}$, we have $\xi_0 + 2\pi k_0 \neq 0$, and therefore $\lim_{n\to\infty}\hat{\eta}(2^n(\xi_0+2\pi k_0)) = 0$ by $\eta \in L_1(\mathbb{R})\cap L_2(\mathbb{R})$. Now by (3.10), we have

$$\left|\prod_{j=0}^{n}\hat{\hat{a}}(2^j\xi_0)\right| = \frac{1}{\left|\prod_{j=0}^{n}\hat{c}(2^j\xi_0)\right|} = \frac{1}{\left|\prod_{j=0}^{n}\hat{c}(2^j(\xi_0+2\pi k_0))\right|}$$

$$= \frac{|\hat{\eta}(\xi_0+2\pi k_0)|}{|\hat{\eta}(2^{n+1}(\xi_0+2\pi k_0))|} \to \infty$$

as $n \to \infty$, which is a contradiction to [14, Lemma 1] (also see (5.19)), since $\hat{\hat{a}} \in C^\beta(\mathbb{T})$. Therefore, the shifts of $\phi \in L_2(\mathbb{R})$ must be stable in $L_2(\mathbb{R})$. □

To illustrate the results in this section, we consider MRA Riesz wavelet bases in $L_2(\mathbb{R})$ using the masks in (1.12). The following result generalizes [17, Theorem 2.2] for the case of B-splines.

THEOREM 3.3. *Let $\hat{a} = \widehat{a_{\beta_1,\beta_2,\beta_3}}$ or $\hat{a} = |\widehat{a_{\beta_1,\beta_2,\beta_3}}|$, where the masks $\widehat{a_{\beta_1,\beta_2,\beta_3}}$ are defined in* (1.12). *Let ϕ denote the standard refinable function associated with mask $\hat{a}$ in* (1.2) *and define a wavelet function ψ by*

$$\hat{\psi}(2\xi) := e^{-i\xi}\overline{\hat{a}(\xi+\pi)}\hat{\phi}(\xi), \qquad \xi \in \mathbb{R}. \tag{3.11}$$

Then the shifts of ϕ are stable in $L_2(\mathbb{R})$ and the wavelet function ψ generates a Riesz wavelet basis in $L_2(\mathbb{R})$, provided that $\beta_1, \beta_2, \beta_3 > 0$ satisfy

$$\beta_1 > 1/4 + |\beta_2 - 1|\beta_3/2,\ (\beta_2-1)\beta_3 > -1/2 \quad \textit{or} \quad \beta_1 \geqslant 1,\ (\beta_2-1)\beta_3 \leqslant -1/2. \tag{3.12}$$

Proof. Let $\hat{b}(\xi) := e^{-i\xi}\overline{\hat{a}(\xi+\pi)}$. Then $\hat{\psi}(2\xi) = \hat{b}(\xi)\hat{\phi}(\xi)$. Clearly, $\hat{a}, \hat{b} \in C^\beta(\mathbb{T})$ for some $\beta > 0$ and $\hat{b}(0) = 0$. By calculation, we have

$$d(\xi) = e^{-i(\xi+\pi)}\big(|\hat{a}(\xi)|^2 + |\hat{a}(\xi+\pi)|^2\big) = e^{-i(\xi+\pi)}\frac{|\cos(\xi/2)|^{4\beta_1} + |\sin(\xi/2)|^{4\beta_1}}{\big(|\cos(\xi/2)|^{2\beta_2} + |\sin(\xi/2)|^{2\beta_2}\big)^{2\beta_3}}.$$

So, $d(\xi) \neq 0$ for all $\xi \in \mathbb{R}$, and (i) of Theorem 3.2 holds.

By calculation, we have

$$|\hat{\tilde{a}}(\xi)| = |\overline{\hat{b}(\xi+\pi)}/\overline{d(\xi)}| = 2^{-2\beta_1}|1+e^{-i\xi}|^{2\beta_1}\hat{c}(\xi),$$

where

$$\hat{c}(\xi) := \frac{\big(|\cos(\xi/2)|^{2\beta_2} + |\sin(\xi/2)|^{2\beta_2}\big)^{\beta_3}}{|\cos(\xi/2)|^{4\beta_1} + |\sin(\xi/2)|^{4\beta_1}}.$$

Now it follows from the inequalities in (2.4) that

$$0 < \hat{c}(\xi) \leqslant c_{\beta_1,\beta_2,\beta_3} := \frac{\max(1, 2^{(1-\beta_2)\beta_3})}{\min(1, 2^{1-2\beta_1})} \qquad \forall\, \xi \in \mathbb{R}.$$

If (3.12) holds, then (2.5) must be true, and therefore, by Example 2.4, we have $\nu_2(\hat{a}) > 0$. By Theorem 4.1 and Lemma 4.3, if (3.12) holds, then we have

$$\rho(\hat{\tilde{a}}) = \rho_{4\beta_1}(\hat{\tilde{a}}, \infty) = \rho_0(2^{-2\beta_1}\hat{c}, \infty) \leqslant \rho_0(2^{-2\beta_1}c_{\beta_1,\beta_2,\beta_3}, \infty) = 2^{1-4\beta_1}|c_{\beta_1,\beta_2,\beta_3}|^2 < 1,$$

since the last inequality combined with (2.5) is equivalent to (3.12). Now by Theorem 3.2, the shifts of ϕ are stable in $L_2(\mathbb{R})$, and the function ψ generates a Riesz wavelet basis in $L_2(\mathbb{R})$. □

For the fractional splines in [27], we have $\beta_2 = 1$, and the condition in (3.12) becomes $\beta_1 > 1/4$. As discussed in Example 2.4, the standard refinable function associated with mask $\hat{a}$ and $\beta_2 = 1$ does not belong to $L_2(\mathbb{R})$ for $0 < \beta_1 \leqslant 1/4$. So, Theorem 3.3 is sharp for all the fractional splines in [27].

For the classical Butterworth filters, we have $\beta_3 = 1$ and $\beta_1 = \beta_2 \in \mathbb{N}$. Now the condition in (3.12) becomes $\beta_1 > 1/2$. Therefore, Theorem 3.3 holds for all the classical Butterworth filters in [25].

4. Some properties and estimate of the quantity $\nu_2(\hat{a})$. In this section, we shall investigate some properties of the quantity $\nu_2(\hat{a})$ in (1.7) and discuss how to estimate the quantity $\nu_2(\hat{a})$. Some results in this section will be needed in our study of refinable functions, cascade algorithms, and wavelets with Hölder continuous masks. The results in this section for the general case of Lebesgue measurable masks are also of interest in their own right and may be useful elsewhere.

The following result generalizes a well-known result for a univariate 2π-periodic trigonometric polynomial $\hat{a}$ and a positive integer τ in the wavelet literature. The proof of the following result for the general case of Hölder continuous masks is nontrivial and will be presented in section 7.

THEOREM 4.1. *Let $\hat{a}$ be a 2π-periodic measurable function such that $|\hat{a}|^2 \in C^\beta(\mathbb{T})$ with $|\hat{a}(0)|^2 \neq 0$ and $\beta > 0$. If $|\hat{a}(\xi)|^2 = |1+e^{-i\xi}|^{2\tau}|\hat{A}(\xi)|^2$ a.e. $\xi \in \mathbb{R}$ for some $\tau \geqslant 0$ such that $\hat{A} \in L_\infty(\mathbb{T})$, then*

$$\begin{aligned}\rho_{2\tau}(\hat{a}, \infty) &= \inf_{n\to\infty} \left\| \frac{T_{\hat{a}}^n(|\sin(\cdot/2)|^{2\tau})}{|\sin(\cdot/2)|^{2\tau}} \right\|_{L_\infty(\mathbb{T})}^{1/n} \\ &= \lim_{n\to\infty} \|T_{\hat{A}}^n 1\|_{L_\infty(\mathbb{T})}^{1/n} = \inf_{n\in\mathbb{N}} \|T_{\hat{A}}^n 1\|_{L_\infty(\mathbb{T})}^{1/n}.\end{aligned} \tag{4.1}$$

As in (1.7), we define a similar quantity as follows:

$$\nu_2(\hat{a}, p) := -[\log_2 \rho(\hat{a}, p)]/2, \qquad 1 \leqslant p \leqslant \infty, \tag{4.2}$$

where the quantity $\rho(\hat{a}, p)$, using $\rho_\tau(\hat{a}, p)$ in (1.6), is defined to be

$$\rho(\hat{a}, p) := \inf\{\rho_\tau(\hat{a}, p) \ : \ |\hat{a}(\cdot + \pi)|^2 / |\sin(\cdot/2)|^\tau \in L_\infty(\mathbb{T}) \ \text{ and } \ \tau \geqslant 0\}. \tag{4.3}$$

Clearly, $\nu_2(\hat{a}) = \nu_2(\hat{a}, \infty)$ and $\rho(\hat{a}) = \rho(\hat{a}, \infty)$. For a 2π-periodic trigonometric polynomial $\hat{a}$, we can write $\hat{a}(\xi) = (1 + e^{-i\xi})^\tau \hat{A}(\xi)$ for some nonnegative integer τ and some 2π-periodic trigonometric polynomial $\hat{A}$ with $\hat{A}(\pi) \neq 0$. Write $|\hat{A}(\xi)|^2 = \sum_{k=-K}^{K} c_k e^{-ik\xi}$. It is known [7, 13, 15] that $\nu_2(\hat{a}, p) = \nu_2(\hat{a}) = -1/2 - \log_2 \sqrt{\rho}$, where ρ is the spectral radius of the square matrix $(c_{2j-k})_{-K \leqslant j,k \leqslant K}$.

In the following, we shall investigate the mutual relations among the quantities $\nu_2(\hat{a}, p)$.

PROPOSITION 4.2. *The following statements hold:*

(1) *For $0 < \tau < 1$, $|\sin(\cdot/2)|^\tau \in C^\tau(\mathbb{T})$ and $|\cos(\cdot/2)|^\tau \in C^\tau(\mathbb{T})$.*

(2) *For a 2π-periodic measurable function $\hat{a}$ and $0 \leqslant \tau_1 \leqslant \tau_2$,*

$$\rho_{\tau_2}(\hat{a}, p) \leqslant \rho_{\tau_1}(\hat{a}, p) \leqslant \rho_{\tau_1}(\hat{a}, q) \quad \text{and} \quad \nu_2(\hat{a}, q) \leqslant \nu_2(\hat{a}, p) \quad \forall\, 1 \leqslant p \leqslant q \leqslant \infty. \tag{4.4}$$

(3) *If $\hat{a} \in C^\beta(\mathbb{T})$ with $\hat{a}(0) = 1$ and $\beta > 0$, then the condition*

$$\liminf_{n \to \infty} \|T_{\hat{a}}^n(|\sin(\cdot/2)|^\tau)\|_{L_1(\mathbb{T})} = 0 \quad \text{for some } \tau \geqslant 0 \tag{4.5}$$

implies $\hat{a}(\pi) = 0$. In particular, (4.5) holds if $\nu_2(\hat{a}, p) > 0$ for some $1 \leqslant p \leqslant \infty$.

Proof. It is easy to prove that $1 - x^\tau \leqslant (1 - x)^\tau$ for all $0 < \tau \leqslant 1$ and $0 \leqslant x \leqslant 1$. Consequently, we have

$$||\sin(x/2)|^\tau - |\sin(y/2)|^\tau| \leqslant |\sin(x/2) - \sin(y/2)|^\tau \leqslant |x - y|^\tau.$$

So, $|\sin(\cdot/2)|^\tau \in C^\tau(\mathbb{T})$, and it follows that $|\cos(\cdot/2)|^\tau \in C^\tau(\mathbb{T})$. So, (1) holds.

Since $0 \leqslant \tau_1 \leqslant \tau_2$, it is evident that $|\sin(\xi/2)|^{\tau_2} \leqslant |\sin(\xi/2)|^{\tau_1}$ for all $\xi \in \mathbb{R}$. Therefore,

$$T_{\hat{a}}^n\left(|\sin(\cdot/2)|^{\tau_2}\right) \leqslant T_{\hat{a}}^n\left(|\sin(\cdot/2)|^{\tau_1}\right).$$

Now the claim in (4.4) follows directly from the above inequality and the fact that $\|\cdot\|_{L_p(\mathbb{T})} \leqslant \|\cdot\|_{L_q(\mathbb{T})}$ for all $1 \leqslant p \leqslant q \leqslant \infty$. So, (2) holds.

To prove (3), we denote $\Phi(\xi) := \prod_{j=1}^{\infty} |\hat{a}(2^{-j}\xi)|^2$. Since $\hat{a} \in C^\beta(\mathbb{T})$ with $\hat{a}(0) = 1$ and $\beta > 0$, Φ is well defined and is continuous with $\Phi(0) = 1$. Suppose that $\hat{a}(\pi) \neq 0$. Then there exist $0 < \varepsilon < \pi/2$ and a positive constant C such that $|\hat{a}(\xi + \pi)|^2 |\sin(\xi/2 + \pi/2)|^\tau \geqslant C$ and $C \leqslant \Phi(\xi) \leqslant 1/C$ for all $\xi \in (-2\varepsilon, 2\varepsilon)$. By the

definition of $T_{\hat{a}}$, we deduce that

$$\begin{aligned}\int_0^{2\pi}[T_{\hat{a}}^n(|\sin(\cdot/2)|^\tau)](\xi)\,d\xi &= 2^n\int_0^{2\pi}|\sin(\xi/2)|^\tau|\hat{a}(\xi)|^2|\hat{a}(2\xi)|^2\cdots|\hat{a}(2^{n-1}\xi)|^2\,d\xi\\ &\geqslant 2^n\int_{\pi-\varepsilon}^{\pi+\varepsilon}|\sin(\xi/2)|^\tau|\hat{a}(\xi)|^2|\hat{a}(2\xi)|^2\cdots|\hat{a}(2^{n-1}\xi)|^2\,d\xi\\ &\geqslant 2^nC\int_{\pi-\varepsilon}^{\pi+\varepsilon}|\hat{a}(2\xi)|^2\cdots|\hat{a}(2^{n-1}\xi)|^2\,d\xi\\ &= 2^{n-1}C\int_{-2\varepsilon}^{2\varepsilon}|\hat{a}(\xi)|^2\cdots|\hat{a}(2^{n-2}\xi)|^2\,d\xi\\ &= 2^{n-1}C\int_{-2\varepsilon}^{2\varepsilon}\frac{\Phi(2^{n-1}\xi)}{\Phi(\xi)}\,d\xi\\ &\geqslant 2^{n-1}C^2\int_{-2\varepsilon}^{2\varepsilon}\Phi(2^{n-1}\xi)\,d\xi\\ &= C^2\int_{-2^n\varepsilon}^{2^n\varepsilon}\Phi(\xi)\,d\xi.\end{aligned}$$

Hence,

$$\liminf_{n\to\infty}\|T_{\hat{a}}^n(|\sin(\cdot/2)|^\tau)\|_{L_1(\mathbb{T})}\geqslant(2\pi)^{-1}C^2\int_{\mathbb{R}}\Phi(\xi)\,d\xi>0,$$

since $\Phi\geqslant 0$ is continuous and $\Phi(0)=1$. This is a contradiction to our assumption in (4.5). So, we must have $\hat{a}(\pi)=0$. If $\nu_2(\hat{a},p)>0$, then $\nu_2(\hat{a},1)\geqslant\nu_2(\hat{a},p)>0$, and therefore (4.5) holds. □

The following result will be needed later in this section.

LEMMA 4.3. *Let $\hat{a}$ and $\hat{c}$ be 2π-periodic measurable functions such that $|\hat{a}(\xi)|\leqslant|\hat{c}(\xi)|$ for almost every $\xi\in\mathbb{R}$. Then*

$$\rho_\tau(\hat{a},p)\leqslant\rho_\tau(\hat{c},p)\quad\text{and}\quad\nu_2(\hat{c},p)\leqslant\nu_2(\hat{a},p)\qquad\forall\,1\leqslant p\leqslant\infty,\tau\in\mathbb{R}.\tag{4.6}$$

Proof. To prove (4.6), since $|\hat{a}(\xi)|^2\leqslant|\hat{c}(\xi)|^2$, it is obvious that

$$0\leqslant[T_{\hat{a}}^n(|\sin(\cdot/2)|^\tau)](\xi)\leqslant[T_{\hat{c}}^n(|\sin(\cdot/2)|^\tau)](\xi)\qquad\text{a.e. }\xi\in\mathbb{R}.$$

Therefore, $\|T_{\hat{a}}^n(|\sin(\cdot/2)|^\tau)\|_{L_p(\mathbb{T})}\leqslant\|T_{\hat{c}}^n(|\sin(\cdot/2)|^\tau)\|_{L_p(\mathbb{T})}$, which implies the first part of (4.6).

If $|\hat{c}(\cdot+\pi)|^2/|\sin(\cdot/2)|^\tau\in L_\infty(\mathbb{T})$ for some τ, then we also have

$$|\hat{a}(\cdot+\pi)|^2/|\sin(\cdot/2)|^\tau\in L_\infty(\mathbb{T})$$

since $|\hat{a}(\xi)|\leqslant|\hat{c}(\xi)|$. Now by the definition of $\nu_2(\hat{a},p)$ in (4.2) and the first part of (4.6), it is easy to see that $\nu_2(\hat{c},p)\leqslant\nu_2(\hat{a},p)$. □

For a particular family of masks, the following result reveals the mutual relations among the quantities $\nu_2(\hat{a},p)$ for different $1\leqslant p\leqslant\infty$.

LEMMA 4.4. *Let $\hat{a}$ be a 2π-periodic measurable function such that $|\hat{a}(\xi)| = |1 + e^{-i\xi}|^\tau |\hat{A}(\xi)|$ for some $\tau \geqslant 0$ and some 2π-periodic trigonometric polynomial $\hat{A}$ with $\hat{A}(0) \neq 0$. Then*

$$\rho_{2\tau}(\hat{a}, p) = \lim_{n\to\infty} \left\| T_{\hat{a}}^n (|\sin(\cdot/2)|^{2\tau}) \right\|_{L_p(\mathbb{T})}^{1/n} = \rho_0(\hat{A}, \infty) = \inf_{n\in\mathbb{N}} \|T_{\hat{A}}^n 1\|_{L_\infty(\mathbb{T})}^{1/n} \quad \forall\, 1 \leqslant p \leqslant \infty. \tag{4.7}$$

In particular, $\nu_2(\hat{a}, p) = \nu_2(\hat{a}) = \nu_2(\hat{A})$ for all $1 \leqslant p \leqslant \infty$. That is, $\nu_2(\hat{a}, p)$ is independent of p.

Proof. Let N be an integer such that $N \geqslant \tau$. Define $\hat{c}(\xi) := (1 + e^{-i\xi})^N \hat{A}(\xi)$. By calculation, it follows from $|\hat{a}(\xi)|^2 = 2^{2\tau} |\cos(\xi/2)|^{2\tau} |\hat{A}(\xi)|^2$ that

$$\begin{aligned}
\left[T_{\hat{a}}(f(\cdot)|\sin(\cdot/2)|^{2\tau})\right](\xi) &= |\hat{a}(\xi/2)|^2 |\sin(\xi/4)|^{2\tau} f(\xi/2) \\
&\quad + |\hat{a}(\xi/2+\pi)|^2 |\sin(\xi/4+\pi/2)|^{2\tau} f(\xi/2+\pi) \\
&= |\sin(\xi/2)|^{2\tau} [|\hat{A}(\xi/2)|^2 f(\xi/2) + |\hat{A}(\xi/2+\pi)|^2 f(\xi/2+\pi)] \\
&= |\sin(\xi/2)|^{2\tau} [T_{\hat{A}} f](\xi).
\end{aligned}$$

That is, when $|\hat{a}(\xi)|^2 = 2^{2\tau} |\cos(\xi/2)|^{2\tau} |\hat{A}(\xi)|^2$, for any 2π-periodic function f, by induction, we have

$$\left[T_{\hat{a}}^n (f(\cdot)|\sin(\cdot/2)|^{2\tau})\right](\xi) = |\sin(\xi/2)|^{2\tau} [T_{\hat{A}}^n f](\xi), \qquad n \in \mathbb{N}. \tag{4.8}$$

Setting $f = 1$ in (4.8), we have

$$[T_{\hat{a}}^n (|\sin(\cdot/2)|^{2\tau})](\xi) = |\sin(\xi/2)|^{2\tau} [T_{\hat{A}}^n 1](\xi) \leqslant [T_{\hat{A}}^n 1](\xi). \tag{4.9}$$

Note that $2N - 2\tau \geqslant 0$. Similarly, by $|\hat{c}(\xi)|^2 = 2^{2N-2\tau} |\cos(\xi/2)|^{2N-2\tau} |\hat{a}(\xi)|^2$, setting $f(\xi) = |\sin(\xi/2)|^{2\tau}$ and replacing τ by $N - \tau$ in (4.8), we have

$$\begin{aligned}
[T_{\hat{c}}^n (|\sin(\cdot/2)|^{2N})](\xi) &= |\sin(\xi/2)|^{2N-2\tau} [T_{\hat{a}}^n (|\sin(\cdot/2)|^{2\tau})](\xi) \\
&\leqslant [T_{\hat{a}}^n (|\sin(\cdot/2)|^{2\tau})](\xi).
\end{aligned} \tag{4.10}$$

Thus, it follows from (4.9) and (4.10) that

$$\begin{aligned}
\|T_{\hat{c}}^n (\sin^{2N}(\cdot/2))\|_{L_p(\mathbb{T})} &\leqslant \|T_{\hat{a}}^n (|\sin(\cdot/2)|^{2\tau})\|_{L_p(\mathbb{T})} \leqslant \|T_{\hat{A}}^n 1\|_{L_p(\mathbb{T})} \\
&\leqslant \|T_{\hat{A}}^n 1\|_{L_\infty(\mathbb{T})}.
\end{aligned} \tag{4.11}$$

Since both $|\hat{c}|^2$ and $\sin^{2N}(\cdot/2)$ are 2π-periodic trigonometric polynomials, by induction, it is known [4, 7, 15] that $\{T_{\hat{c}}^n (\sin^{2N}(\cdot/2))\}_{n=1}^\infty$ spans a finite dimensional space and in fact the degrees of all trigonometric polynomials $T_{\hat{c}}^n (\sin^{2N}(\cdot/2))$ are uniformly bounded. Therefore, there exists a positive constant C, independent of all n, such that

$$C\|T_{\hat{c}}^n (\sin^{2N}(\cdot/2))\|_{L_\infty(\mathbb{T})} \leqslant \|T_{\hat{c}}^n (\sin^{2N}(\cdot/2))\|_{L_p(\mathbb{T})} \qquad \forall\, 1 \leqslant p \leqslant \infty, n \in \mathbb{N}.$$

Hence, we conclude from the above inequality and (4.11) that for $1 \leqslant p \leqslant \infty$,

$$C^{1/n}\|T_{\hat{c}}^n(\sin^{2N}(\cdot/2))\|_{L_\infty(\mathbb{T})}^{1/n} \leqslant \|T_{\hat{a}}^n(|\sin(\cdot/2)|^{2\tau})\|_{L_p(\mathbb{T})}^{1/n} \leqslant \|T_{\hat{A}}^n 1\|_{L_\infty(\mathbb{T})}^{1/n} \quad \forall\, n \in \mathbb{N}. \tag{4.12}$$

Since $\hat{c}(\xi) = (1+e^{-i\xi})^N \hat{A}(\xi)$ and $\hat{A}(0) \neq 0$, by Theorem 4.1, we have

$$\rho_{2N}(\hat{c},\infty) = \lim_{n\to\infty} \|T_{\hat{c}}^n(\sin^{2N}(\cdot/2))\|_{L_\infty(\mathbb{T})}^{1/n} = \lim_{n\to\infty} \|T_{\hat{A}}1\|_{L_\infty(\mathbb{T})}^{1/n} = \inf_{n\in\mathbb{N}} \|T_{\hat{A}}1\|_{L_\infty(\mathbb{T})}^{1/n}.$$

Now (4.7) follows directly from (4.12).

Since $\hat{A}$ is a 2π-periodic trigonometric polynomial, we can write $\hat{A}(\xi) = (1+e^{-i\xi})^k \hat{B}(\xi)$ for some nonnegative integer k and some 2π-periodic trigonometric polynomial $\hat{B}$ such that $\hat{B}(\pi) \neq 0$. So, $|\hat{a}(\xi)| = |1+e^{-i\xi}|^{k+\tau}|\hat{B}(\xi)|$. Now by the definition of $\nu_2(\hat{a},p)$ and Theorem 4.1, it follows from (4.7) that

$$\begin{aligned}\nu_2(\hat{a},p) &:= -[\log_2 \rho_{2k+2\tau}(\hat{a},p)]/2 = -[\log_2 \rho_0(\hat{B},\infty)]/2\\ &= -[\log_2 \rho_{2k}(\hat{A},\infty)]/2 = \nu_2(\hat{A}).\end{aligned}$$

Therefore, $\nu_2(\hat{a},p) = \nu_2(\hat{A})$ for all $1 \leqslant p \leqslant \infty$. □

In the following, we shall discuss how to approximate the quantity $\rho_0(\hat{A},\infty)$.

PROPOSITION 4.5. *Let $\hat{A}, \widehat{A_j} \in L_\infty(\mathbb{T})$, $j \in \mathbb{N}$ such that $\lim_{j\to\infty} \|\widehat{A_j} - \hat{A}\|_{L_\infty(\mathbb{T})} = 0$. Then*

$$\limsup_{j\to\infty} \rho_0(\widehat{A_j},\infty) \leqslant \rho_0(\hat{A},\infty). \tag{4.13}$$

If in addition $|\hat{A}(\xi)| \leqslant |\widehat{A_j}(\xi)|$ for almost every $\xi \in \mathbb{R}$ and for all $j \in \mathbb{N}$, then

$$\lim_{j\to\infty} \rho_0(\widehat{A_j},\infty) = \rho_0(\hat{A},\infty). \tag{4.14}$$

Proof. By Proposition 7.1, $\rho_0(\hat{A},\infty) = \inf_{n\in\mathbb{N}} \|T_{\hat{A}}^n 1\|_{L_\infty(\mathbb{T})}$. Therefore, for any $\varepsilon > 0$, there exists a positive integer N such that $\|T_{\hat{A}}^N 1\|_{L_\infty(\mathbb{T})}^{1/N} < \rho_0(\hat{A},\infty) + \varepsilon$. Now by $\lim_{j\to\infty} \|\widehat{A_j} - \hat{A}\|_{L_\infty(\mathbb{T})} = 0$, there exists a positive integer J such that

$$\|T_{\widehat{A_j}}^N 1\|_{L_\infty(\mathbb{T})}^{1/N} < \rho_0(\hat{A},\infty) + \varepsilon \qquad \forall\, j \geqslant J.$$

Consequently, by Proposition 7.1, we deduce that

$$\rho_0(\widehat{A_j},\infty) = \inf_{n\in\mathbb{N}} \|T_{\widehat{A_j}}^n 1\|_{L_\infty(\mathbb{T})}^{1/n} \leqslant \|T_{\widehat{A_j}}^N 1\|_{L_\infty(\mathbb{T})}^{1/N} < \rho_0(\hat{A},\infty) + \varepsilon \qquad \forall\, j \geqslant J.$$

Hence, we have $\limsup_{j\to\infty} \rho_0(\widehat{A_j},\infty) \leqslant \rho_0(\hat{A},\infty) + \varepsilon$. Taking $\varepsilon \to 0$, we conclude that (4.13) holds. If $|\hat{A}| \leqslant |\widehat{A_j}|$, then by Lemma 4.3, $\rho_0(\hat{A},\infty) \leqslant \rho_0(\widehat{A_j},\infty)$ for all $j \in \mathbb{N}$. Therefore, $\rho_0(\hat{A},\infty) \leqslant \liminf_{j\to\infty} \rho_0(\widehat{A_j},\infty)$. Now it follows from (4.13) that (4.14) holds. □

As a consequence of Proposition 4.5, we have the following corollary.

COROLLARY 4.6. *Let $\hat{A} \in C^\beta(\mathbb{T})$ with $\hat{A}(0) \neq 0$, $\hat{A}(\pi) \neq 0$, and $\beta > 0$. Suppose that there is a sequence $\{\widehat{A_j}\}_{j\in\mathbb{N}}$ in $C^\beta(\mathbb{T})$ such that $\lim_{j\to\infty} \|\widehat{A_j} - \hat{A}\|_{L_\infty(\mathbb{T})} = 0$, and*

$|\hat{A}(\xi)| \leqslant |\widehat{A_j}(\xi)|$ *for all* $\xi \in \mathbb{R}$ *and* $j \in \mathbb{N}$. *For* $\tau \geqslant 0$ *and a* 2π*-periodic trigonometric polynomial* $\hat{c}$ *with* $\hat{c}(0) \neq 0$, *let* $\hat{a}(\xi) := (1+e^{-i\xi})^\tau \hat{c}(\xi)\hat{A}(\xi)$ *and* $\widehat{a_j}(\xi) := (1+e^{-i\xi})^\tau \hat{c}(\xi)\widehat{A_j}(\xi)$. *Then* $\lim_{j\to\infty} \nu_2(\widehat{a_j}) = \nu_2(\hat{a})$ *and* $\nu_2(\widehat{a_j}) \leqslant \nu_2(\hat{a})$ *for all* $j \in \mathbb{N}$.

Proof. Write $\hat{c}(\xi) = (1+e^{-i\xi})^k \hat{B}(\xi)$ for some nonnegative integer k and some 2π-periodic trigonometric polynomial $\hat{B}$ with $\hat{B}(\pi) \neq 0$. Since $\hat{A}(\pi) \neq 0$ and $\lim_{j\to\infty} \|\widehat{A_j} - \hat{A}\|_{L_\infty(\mathbb{T})} = 0$, by $\widehat{A_j} \in C^\beta(\mathbb{T})$, we have $\widehat{A_j}(\pi) \neq 0$ for sufficiently large j. Now by Theorem 4.1 and Proposition 4.5, we have

$$\begin{aligned}\rho(\hat{a},\infty) = \rho_{2k+2\tau}(\hat{a},\infty) = \rho_0(\hat{A}\hat{B},\infty) &= \lim_{j\to\infty} \rho_0(\widehat{A_j}\hat{B},\infty)\\ &= \lim_{j\to\infty} \rho_{2k+2\tau}(\widehat{a_j},\infty) = \lim_{j\to\infty} \rho(\widehat{a_j},\infty).\end{aligned}$$

It follows from the definition of $\nu_2(\hat{a})$ that $\nu_2(\hat{a}) = \lim_{j\to\infty} \nu_2(\widehat{a_j})$. Now by $|\hat{a}(\xi)| \leqslant |\widehat{a_j}(\xi)|$, it follows from Lemma 4.3 that $\nu_2(\widehat{a_j}) \leqslant \nu_2(\hat{a})$. □

PROPOSITION 4.7. *Let* $\hat{a}$ *and* $\widehat{a_j}$, $j \in \mathbb{N}$, *be* 2π*-periodic measurable functions such that*

$$\lim_{j\to\infty} \|\widehat{a_j}/\hat{a}\|_{L_\infty(\mathbb{T})} = \lim_{j\to\infty} \|\hat{a}/\widehat{a_j}\|_{L_\infty(\mathbb{T})} = 1, \tag{4.15}$$

where by convention $(\widehat{a_j}/\hat{a})(\xi)$ *is equal to* $\widehat{a_j}(\xi)/\hat{a}(\xi)$ *if* $\hat{a}(\xi) \neq 0$, 1 *if* $\hat{a}(\xi) = \widehat{a_j}(\xi) = 0$, *or* $+\infty$ *if* $\hat{a}(\xi) = 0$ *but* $\widehat{a_j}(\xi) \neq 0$. *Then*

$$\lim_{j\to\infty} \nu_2(\widehat{a_j},p) = \nu_2(\hat{a},p) \qquad \forall\, 1 \leqslant p \leqslant \infty. \tag{4.16}$$

Moreover, if $\nu_2(\widehat{a_j},p) = \nu_2(\widehat{a_j})$ *for all* $j \in \mathbb{N}$, *then* $\nu_2(\hat{a},p) = \nu_2(\hat{a})$. *In particular, if a mask* $\hat{a}$ *has exponential decay, then* $\nu_2(\hat{a},p) = \nu_2(\hat{a})$ *for all* $1 \leqslant p \leqslant \infty$.

Proof. Denote $\widehat{c_j} = \widehat{a_j}/\hat{a}$. By convention and (4.15), $0 < |\widehat{c_j}(\xi)| < \infty$ for almost every $\xi \in \mathbb{R}$, and it is easy to check that $\widehat{a_j}(\xi) = \widehat{c_j}(\xi)\hat{a}(\xi)$ and $\hat{a}(\xi) = \widehat{a_j}(\xi)/\widehat{c_j}(\xi)$ for almost every $\xi \in \mathbb{R}$. For $\tau \geqslant 0$, we now have

$$\begin{aligned}\|1/\widehat{c_j}\|_{L_\infty(\mathbb{T})}^{-2n}[T_{\hat{a}}^n(|\sin(\cdot/2)|^\tau)](\xi) &\leqslant [T_{\widehat{a_j}}^n(|\sin(\cdot/2)|^\tau)](\xi)\\ &\leqslant \|\widehat{c_j}\|_{L_\infty(\mathbb{T})}^{2n}[T_{\hat{a}}^n(|\sin(\cdot/2)|^\tau)](\xi).\end{aligned}$$

Consequently, we have

$$\begin{aligned}\|1/\widehat{c_j}\|_{L_\infty(\mathbb{T})}^{-2}\|T_{\hat{a}}^n(|\sin(\cdot/2)|^\tau)\|_{L_p(\mathbb{T})}^{1/n} &\leqslant \|T_{\widehat{a_j}}^n(|\sin(\cdot/2)|^\tau)\|_{L_p(\mathbb{T})}^{1/n}\\ &\leqslant \|\widehat{c_j}\|_{L_\infty(\mathbb{T})}^2\|T_{\hat{a}}^n(|\sin(\cdot/2)|^\tau)\|_{L_p(\mathbb{T})}^{1/n}.\end{aligned}$$

Hence, we deduce that

$$\|1/\widehat{c_j}\|_{L_\infty(\mathbb{T})}^{-2}\rho_\tau(\hat{a},p) \leqslant \rho_\tau(\widehat{a_j},p) \leqslant \|\widehat{c_j}\|_{L_\infty(\mathbb{T})}^2\rho_\tau(\hat{a},p) \qquad \forall\, j \in \mathbb{N}, 1 \leqslant p \leqslant \infty. \tag{4.17}$$

By our assumption in (4.15), we have $\lim_{j\to\infty} \|1/\widehat{c_j}\|_{L_\infty(\mathbb{T})} = \lim_{j\to\infty} \|\widehat{c_j}\|_{L_\infty(\mathbb{T})} = 1$. Now by the definition of $\nu_2(\hat{a})$ and (4.17), we must have $\lim_{j\to\infty} \nu_2(\widehat{a_j},p) = \nu_2(\hat{a},p)$ for all $1 \leqslant p \leqslant \infty$.

If $\nu_2(\widehat{a_j},p) = \nu_2(\widehat{a_j})$, then $\nu_2(\hat{a},p) = \lim_{j\to\infty} \nu_2(\widehat{a_j},p) = \lim_{j\to\infty} \nu_2(\widehat{a_j}) = \nu_2(\hat{a})$.

If $\hat{a}$ has exponential decay, then we can write $\hat{a}(\xi) = \hat{c}(\xi)\hat{A}(\xi)$, where $\hat{c}$ is a 2π-periodic trigonometric polynomial and $\hat{A}$ has exponential decay satisfying $\hat{A}(\xi) \neq 0$

for all $\xi \in \mathbb{R}$. Now it is easy to see that there is a sequence $\{\widehat{A_j}\}_{j=1}^{\infty}$ of 2π-periodic trigonometric polynomials such that $\lim_{j\to\infty} \|\widehat{A_j}/\hat{A}\|_{L_\infty(\mathbb{T})} = \lim_{j\to\infty} \|\hat{A}/\widehat{A_j}\|_{L_\infty(\mathbb{T})} = 1$. Taking $\widehat{a_j}(\xi) := \hat{c}(\xi)\widehat{A_j}(\xi)$, then (4.15) holds, and by Lemma 4.4, $\nu_2(\widehat{a_j}, p) = \nu_2(\widehat{a_j})$ for all $1 \leqslant p \leqslant \infty$ and $j \in \mathbb{N}$. By what has been proved, we have $\nu_2(\hat{a}, p) = \lim_{j\to\infty} \nu_2(\widehat{a_j}, p) = \lim_{j\to\infty} \nu_2(\widehat{a_j}) = \nu_2(\hat{a})$. □

In passing, we mention that the quantity defined in [14, equation (2.16)] corresponds to $\nu_2(\hat{a}, 1)$ in (4.2) of this paper. For a mask with exponential decay, by Proposition 4.7, the quantity $\nu_2(\hat{a})$ in [14, equation (2.16)] agrees with the one in this paper. However, it is not clear whether $\nu_2(\hat{a}, p) = \nu_2(\hat{a})$ for all $1 \leqslant p \leqslant \infty$ if $\hat{a} \in C^\beta(\mathbb{T})$ for some $\beta > 0$.

5. Proof of Theorem 2.1. Since $\hat{a}(0) = 1$ and $\hat{a} \in C^\beta(\mathbb{T})$ with $\beta > 0$, we see that $\hat{\phi}$ in (1.2) is well defined and $\hat{\phi}$ is continuous. If $\{f_n\}_{n=1}^{\infty}$ is a Cauchy sequence in $L_{2,\infty,0}(\mathbb{R})$ and $\lim_{n\to\infty} \hat{f}(2^{-n}\xi) = 1$ for almost every $\xi \in \mathbb{R}$, by $\hat{\phi}(\xi) = \lim_{n\to\infty} \widehat{f_n}(\xi)$ for almost every $\xi \in \mathbb{R}$, then we must have $\phi \in L_{2,\infty,0}(\mathbb{R})$ and $\lim_{n\to\infty} \|f_n - \phi\|_{L_{2,\infty,0}(\mathbb{R})} = 0$.

We shall prove in the order that (i)⇒(ii)⇒(iii)⇒(iv) ⇒(v)⇒(ii) and (iii)⇒(i). Note that $\hat{a} \in C^\beta(\mathbb{T})$ implies $\hat{a} \in C^{\min(1,\beta)}(\mathbb{T})$. So, without loss of generality, we replace β by $\min(1, \beta)$. That is, we assume $0 < \beta \leqslant 1$, and the following proof depends only on the fact that $\hat{a} \in C^\alpha(\mathbb{T})$ for a small number $\alpha > 0$.

We show (i)⇒(ii) by constructing an admissible function η in $L_{2,\infty,0}(\mathbb{R})$ such that the shifts of η are stable in $L_2(\mathbb{R})$ and $\hat{\eta}$ is compactly supported; such an admissible initial function η will be used in several places in this proof. Since $\hat{\phi}$ is a continuous function with $\hat{\phi}(0) = 1$, there exists $0 < \varepsilon < \pi$ such that $1/2 \leqslant |\hat{\phi}(\xi)| \leqslant 3/2$ for all $|\xi| \leqslant \varepsilon$. Define a function $\hat{\eta}$ by

$$\hat{\eta}(\xi) := \begin{cases} \hat{\phi}(\xi) & \text{if } |\xi| \leqslant \varepsilon, \\ \hat{\phi}(-\varepsilon)(3\pi + 2\xi)/(3\pi - 2\varepsilon) & \text{if } -3\pi/2 \leqslant \xi < -\varepsilon, \\ \hat{\phi}(\varepsilon)(3\pi - 2\xi)/(3\pi - 2\varepsilon) & \text{if } \varepsilon < \xi \leqslant 3\pi/2, \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

Note that $\hat{\eta}(\xi) - \hat{a}(\xi/2)\hat{\eta}(\xi/2) = 0$ for all $|\xi| \leqslant \varepsilon$. By the assumption $\hat{a}(\pi) = 0$ in (i) and $\hat{a} \in C^\beta(\mathbb{T})$, we see that $|\hat{a}(\cdot + \pi)|^2/|\sin(\cdot/2)|^{2\beta} \in L_\infty(\mathbb{T})$. Since $\hat{\eta}$ is supported inside $[-3\pi/2, 3\pi/2]$, now we can easily verify that η is an admissible function in $L_{2,\infty,0}(\mathbb{R})$ with respect to $\hat{a}$, since the condition in (2.1) holds with $\tau = 2\beta > 0$. By the definition of $\hat{\eta}$, we have $1/32 \leqslant [\hat{\eta}, \hat{\eta}] \leqslant 9/2$. Therefore, the shifts of η are stable in $L_2(\mathbb{R})$. Taking $f = \eta$, it follows directly from (i) that (ii) holds.

Now we prove (ii)⇒(iii) without the condition that the initial function f is admissible. By the definition of f_n in (2.2) and induction, we deduce that

$$[\widehat{f_n}, \widehat{f_n}](\xi) = \sum_{k=0}^{2^n-1} \prod_{j=1}^{n} |\hat{a}(2^{-j}(\xi + 2\pi k))|^2 [\hat{f}, \hat{f}](2^{-n}(\xi + 2\pi k)) = (T_{\hat{a}}^n [\hat{f}, \hat{f}])(\xi). \tag{5.2}$$

Since the shifts of f are stable in $L_2(\mathbb{R})$ and $\{f_n\}_{n=1}^{\infty}$ is a Cauchy sequence in $L_{2,\infty,0}(\mathbb{R})$, there exists a positive constant C_1 such that

$$\|f_n\|^2_{L_{2,\infty,0}(\mathbb{R})} = \|[\widehat{f_n}, \widehat{f_n}]\|_{L_\infty(\mathbb{T})} \leqslant C_1$$

for all $n \in \mathbb{N}$ and $1/C_1 \leqslant [\hat{f}, \hat{f}](\xi) \leqslant C_1$ for almost every $\xi \in \mathbb{R}$. Now it follows from (5.2) that

$$0 \leqslant [T_{\hat{a}}^n 1](\xi) \leqslant C_1 (T_{\hat{a}}^n C_1^{-1})(\xi) \leqslant C_1 (T_{\hat{a}}^n [\hat{f}, \hat{f}])(\xi) = C_1 [\widehat{f_n}, \widehat{f_n}](\xi) \leqslant C_1^2.$$

That is, we have

$$\|T_{\hat{a}}^n 1\|_{L_\infty(\mathbb{T})} \leqslant C_1^2 \qquad \forall\, n \in \mathbb{N} \cup \{0\}. \tag{5.3}$$

Since $\hat{a} \in C^\beta(\mathbb{T})$, we have $|\hat{a}|^2 \in C^\beta(\mathbb{T})$, and there exists a positive constant C_2 such that

$$\big\| |\hat{a}|^2 - |\hat{a}(\cdot - h)|^2 \big\|_{L_\infty(\mathbb{T})} \leqslant C_2 h^\beta \qquad \forall\, h > 0. \tag{5.4}$$

Now we are going to extend an interesting idea in [23] to show that $T_{\hat{a}}^n g$, $n \in \mathbb{N}$, are equicontinuous if $g \in C^\tau(\mathbb{T})$ for some $\tau > 0$. For $g \in L_\infty(\mathbb{T})$, we denote

$$\omega_n(\xi, h) := \left| [T_{\hat{a}}^n g](\xi) - [T_{\hat{a}}^n g](\xi - h) \right|, \qquad \xi \in \mathbb{R}, h > 0, n \in \mathbb{N} \cup \{0\}. \tag{5.5}$$

Setting

$$C_3 := \max(C_1^2, 2^{1-\beta} C_1^4 C_2/(1 - 2^{-\beta})) < \infty,$$

we show that (5.3) and (5.4) imply that

$$\|h^{-\beta} \omega_{n+k}(\cdot, h)\|_{L_\infty(\mathbb{T})} \leqslant C_3 \Big(2^{-n\beta} \|(2^{-n}h)^{-\beta} \omega_k(\cdot, 2^{-n}h)\|_{L_\infty(\mathbb{T})} + \|T_{\hat{a}}^k g\|_{L_\infty(\mathbb{T})} \Big)$$

$$\forall\, h > 0,\ k, n \in \mathbb{N} \cup \{0\}, g \in L_\infty(\mathbb{T}). \tag{5.6}$$

By the definition of $T_{\hat{a}}$, we have

$$\begin{aligned}
&[T_{\hat{a}}^n g](\xi) - [T_{\hat{a}}^n g](\xi - h) \\
&\quad = |\hat{a}(\xi/2)|^2 \left([T_{\hat{a}}^{n-1} g](\xi/2) - [T_{\hat{a}}^{n-1} g](\xi/2 - h/2) \right) \\
&\qquad + |\hat{a}(\xi/2 + \pi)|^2 \left([T_{\hat{a}}^{n-1} g](\xi/2 + \pi) - [T_{\hat{a}}^{n-1} g](\xi/2 + \pi - h/2) \right) \\
&\qquad + \left(|\hat{a}(\xi/2)|^2 - |\hat{a}(\xi/2 - h/2)|^2 \right) [T_{\hat{a}}^{n-1} g](\xi/2 - h/2) \\
&\qquad + \left(|\hat{a}(\xi/2 + \pi)|^2 - |\hat{a}(\xi/2 + \pi - h/2)|^2 \right) [T_{\hat{a}}^{n-1} g](\xi/2 + \pi - h/2).
\end{aligned}$$

It follows from (5.4) that

$$\begin{aligned}
\omega_n(\xi, h) &\leqslant |\hat{a}(\xi/2)|^2 \omega_{n-1}(\xi/2, h/2) + |\hat{a}(\xi/2 + \pi)|^2 \omega_{n-1}(\xi/2 + \pi, h/2) \\
&\qquad + 2C_2 (h/2)^\beta \|T_{\hat{a}}^{n-1} g\|_{L_\infty(\mathbb{T})} \\
&= [T_{\hat{a}} \omega_{n-1}(\cdot, h/2)](\xi) + 2^{1-\beta} C_2 h^\beta \|T_{\hat{a}}^{n-1} g\|_{L_\infty(\mathbb{T})}.
\end{aligned}$$

Consequently, by induction on n, we deduce from the above inequality that for all $k, n \in \mathbb{N} \cup \{0\}$,

$$\begin{aligned}
\omega_{n+k}(\xi, h) &\leqslant [T_{\hat{a}}^n \omega_k(\cdot, 2^{-n}h)](\xi) \\
&\qquad + 2C_2 h^\beta \sum_{j=1}^{n} 2^{-j\beta} \|T_{\hat{a}}^{n+k-j} g\|_{L_\infty(\mathbb{T})} \|T_{\hat{a}}^{j-1} 1\|_{L_\infty(\mathbb{T})}.
\end{aligned} \tag{5.7}$$

In fact, (5.7) clearly holds for $n=0$ and all $k\in\mathbb{N}\cup\{0\}$. Suppose that (5.7) holds for n and all $k\in\mathbb{N}\cup\{0\}$. Then by induction hypothesis we have

$$\begin{aligned}\omega_{n+1+k}(\xi,h)&\leqslant[T_{\hat{a}}^n\omega_{k+1}(\cdot,2^{-n}h)](\xi)\\&\quad+2C_2h^\beta\sum_{j=1}^n2^{-j\beta}\|T_{\hat{a}}^{n+k+1-j}g\|_{L_\infty(\mathbb{T})}\|T_{\hat{a}}^{j-1}1\|_{L_\infty(\mathbb{T})}.\end{aligned}$$

Note that we proved in the inequality above (5.7) that

$$\omega_{k+1}(\xi,2^{-n}h)\leqslant[T_{\hat{a}}\omega_k(\cdot,2^{-1-n}h)](\xi)+2^{1-\beta}C_2(2^{-n}h)^\beta\|T_{\hat{a}}^kg\|_{L_\infty(\mathbb{T})}.$$

Applying the operator $T_{\hat{a}}^n$ on both sides of the above inequality, we deduce that

$$T_{\hat{a}}^n\omega_{k+1}(\xi,2^{-n}h)\leqslant[T_{\hat{a}}^{n+1}\omega_k(\cdot,2^{-1-n}h)](\xi)+2^{1-\beta}C_2(2^{-n}h)^\beta\|T_{\hat{a}}^kg\|_{L_\infty(\mathbb{T})}[T_{\hat{a}}^n1](\xi).$$

Combining all the above inequalities together, we see that (5.7) holds for $n+1$ and all $k\in\mathbb{N}\cup\{0\}$. So, by induction, (5.7) holds for all $k,n\in\mathbb{N}\cup\{0\}$.

By (5.3), we have

$$\|T_{\hat{a}}^{n+k-j}g\|_{L_\infty(\mathbb{T})}\leqslant\|T_{\hat{a}}^{n-j}1\|_{L_\infty(\mathbb{T})}\|T_{\hat{a}}^kg\|_{L_\infty(\mathbb{T})}\leqslant C_1^2\|T_{\hat{a}}^kg\|_{L_\infty(\mathbb{T})}$$

and

$$\|T_{\hat{a}}^n\omega_k(\cdot,2^{-n}h)\|_{L_\infty(\mathbb{T})}\leqslant\|T_{\hat{a}}^n1\|_{L_\infty(\mathbb{T})}\|\omega_k(\cdot,2^{-n}h)\|_{L_\infty(\mathbb{T})}\leqslant C_1^2\|\omega_k(\cdot,2^{-n}h)\|_{L_\infty(\mathbb{T})}.$$

Therefore, we deduce from (5.7) that

$$\begin{aligned}\|\omega_{n+k}(\cdot,h)\|_{L_\infty(\mathbb{T})}&\leqslant C_1^2\|\omega_k(\cdot,2^{-n}h)\|_{L_\infty(\mathbb{T})}+2C_1^4C_2h^\beta\|T_{\hat{a}}^kg\|_{L_\infty(\mathbb{T})}\sum_{j=1}^\infty2^{-j\beta}\\&\leqslant C_3\left(\|\omega_k(\cdot,2^{-n}h)\|_{L_\infty(\mathbb{T})}+h^\beta\|T_{\hat{a}}^kg\|_{L_\infty(\mathbb{T})}\right).\end{aligned}$$

That is, (5.6) has been proved. In particular, for any $\tau>0$, we take $g(\xi)=|\sin(\xi/2)|^\tau$ and $\nu:=\min(\beta,\tau)>0$. By Proposition 4.2, $g\in C^\nu(\mathbb{T})$. It is evident that $\|T_{\hat{a}}^0g\|_{L_\infty(\mathbb{T})}=\|g\|_{L_\infty(\mathbb{T})}=1$. By $g\in C^\nu(\mathbb{T})$, there exists a positive constant C_4 such that

$$(2^{-n}h)^{-\nu}\omega_0(\xi,2^{-n}h)=(2^{-n}h)^{-\nu}|g(\xi)-g(\xi-2^{-n}h)|\leqslant C_4. \tag{5.8}$$

Since $0<\nu\leqslant\beta$, we see that (5.4) holds with β being replaced by ν (now the constant C_2 in (5.4) may be different). That is, for all $k,n\in\mathbb{N}\cup\{0\}$ and $h>0$, (5.6) becomes

$$\begin{aligned}&\|h^{-\nu}\omega_{n+k}(\cdot,h)\|_{L_\infty(\mathbb{T})}\\&\quad\leqslant C_3\Big(2^{-n\nu}\|(2^{-n}h)^{-\nu}\omega_k(\cdot,2^{-n}h)\|_{L_\infty(\mathbb{T})}+\|T_{\hat{a}}^kg\|_{L_\infty(\mathbb{T})}\Big).\end{aligned} \tag{5.9}$$

Setting $k=0$ in (5.9), we deduce that

$$\|h^{-\nu}\omega_n(\cdot,h)\|_{L_\infty(\mathbb{T})}\leqslant C_3(2^{-n\nu}C_4+1)\leqslant C_3(C_4+1)<\infty\qquad\forall\,h>0,n\in\mathbb{N}. \tag{5.10}$$

By the definition of $\omega_n(\xi,h)$ in (5.5), this is equivalent to saying that

$$|[T_{\hat{a}}^ng](\xi_1)-[T_{\hat{a}}^ng](\xi_2)|\leqslant C_3(C_4+1)|\xi_1-\xi_2|^\nu\qquad\forall\,n\in\mathbb{N},\xi_1,\xi_2\in\mathbb{R}.$$

Note that $\hat{a}$ is continuous and $\|T_{\hat{a}}^n g\|_{L_\infty(\mathbb{T})} \leqslant \|T_{\hat{a}}^n 1\|_{L_\infty(\mathbb{T})}\|g\|_{L_\infty(\mathbb{T})} \leqslant C_1^2$. So, the sequence $\{T_{\hat{a}}^n g\}_{n=1}^\infty$ is bounded and equicontinuous in $C(\mathbb{T})$. By the Arzela–Ascoli theorem, there is a subsequence $\{T_{\hat{a}}^{n_k} g\}_{k=1}^\infty$ converging to $g_\infty \in C(\mathbb{T})$ as $k \to \infty$; that is,

$$\lim_{k\to\infty} \|T_{\hat{a}}^{n_k} g - g_\infty\|_{C(\mathbb{T})} = 0. \tag{5.11}$$

Since all $T_{\hat{a}}^{n_k} g \geqslant 0$, we must have $g_\infty \geqslant 0$. Now we show that if (ii) holds, then we must have $g_\infty \equiv 0$. Define

$$\begin{aligned} g_n(\xi) &:= g(2^{-n}\xi)|\widehat{f_n}(\xi)|^2 = |\sin(2^{-1-n}\xi)|^\tau |\widehat{f_n}(\xi)|^2 \\ &= |\sin(2^{-1-n}\xi)|^\tau |\hat{f}(2^{-n}\xi)|^2 \prod_{j=1}^n |\hat{a}(2^{-j}\xi)|^2. \end{aligned}$$

By induction, we observe that $[g_n, g_n] = T_{\hat{a}}^n([\hat{f}, \hat{f}]g)$ for all $n \in \mathbb{N}$. Since $[\hat{f}, \hat{f}] \geqslant 1/C_1$, we conclude that

$$[T_{\hat{a}}^n g](\xi) \leqslant C_1[T_{\hat{a}}^n([\hat{f}, \hat{f}]g)](\xi) = C_1[g_n, g_n](\xi).$$

Therefore,

$$\int_0^{2\pi} [T_{\hat{a}}^n g](\xi)\, d\xi \leqslant C_1 \int_0^{2\pi} [g_n, g_n](\xi)\, d\xi = C_1 \int_{\mathbb{R}} g(2^{-n}\xi)|\widehat{f_n}(\xi)|^2\, d\xi. \tag{5.12}$$

By $g(\xi) = |\sin(\xi/2)|^\tau$ and $\tau > 0$, we have $\lim_{n\to\infty} g(2^{-n}\xi) = g(0) = 0$. Since $\hat{a} \in C^\beta(\mathbb{T})$ with $\hat{a}(0) = 1$, we have $\lim_{n\to\infty} \prod_{j=1}^n |\hat{a}(2^{-j}\xi)|^2 = |\hat{\phi}(\xi)|^2$. Observing that

$$\begin{aligned} 0 \leqslant g(2^{-n}\xi)|\widehat{f_n}(\xi)|^2 &= g(2^{-n}\xi)|\hat{f}(2^{-n}\xi)|^2 \prod_{j=1}^n |\hat{a}(2^{-j}\xi)|^2 \\ &\leqslant \|f\|_{L_{2,\infty,0}}^2(\mathbb{R}) g(2^{-n}\xi) \prod_{j=1}^n |\hat{a}(2^{-j}\xi)|^2, \end{aligned}$$

we see that $\lim_{n\to\infty} g(2^{-n}\xi)|\widehat{f_n}(\xi)|^2 = 0$ for almost every $\xi \in \mathbb{R}$.

Since $\{f_n\}_{n=1}^\infty$ is a Cauchy sequence in $L_{2,\infty,0}(\mathbb{R})$, the sequence $\{f_n\}_{n=1}^\infty$ must also be a Cauchy sequence in $L_2(\mathbb{R})$ by $\|f\|_{L_2(\mathbb{R})} \leqslant \|f\|_{L_{2,\infty,0}(\mathbb{R})}$. Consequently,

$$\lim_{n\to\infty} \int_{\mathbb{R}} |\widehat{f_n}(\xi)|^2\, d\xi = \lim_{n\to\infty} \int_0^{2\pi} [\widehat{f_n}, \widehat{f_n}](\xi)\, d\xi$$

exists and is finite. Now by $0 \leqslant g(2^{-n}\xi)|\widehat{f_n}(\xi)|^2 \leqslant |\widehat{f_n}(\xi)|^2$ and the generalized Lebesgue dominated convergence theorem, we conclude that

$$\lim_{n\to\infty} \int_{\mathbb{R}} g(2^{-n}\xi)|\widehat{f_n}(\xi)|^2\, d\xi = \int_{\mathbb{R}} \lim_{n\to\infty} g(2^{-n}\xi)|\widehat{f_n}(\xi)|^2\, d\xi = \int_{\mathbb{R}} 0\, d\xi = 0.$$

Hence, by (5.12), we get

$$\lim_{n\to\infty} \int_0^{2\pi} [T_{\hat{a}}^n g](\xi)\, d\xi = 0. \tag{5.13}$$

Now it follows from (5.11) and (5.13) that

$$\int_0^{2\pi} g_\infty(\xi)\,d\xi = \lim_{k\to\infty}\int_0^{2\pi} [T_{\hat a}^{n_k} g](\xi)\,d\xi = 0.$$

Since $g_\infty \geqslant 0$, we must have $g_\infty = 0$. That is, (5.11) becomes

$$\lim_{k\to\infty} \|T_{\hat a}^{n_k} g\|_{L_\infty(\mathbb{T})} = 0. \tag{5.14}$$

Setting $k = n_k$ in (5.9), by (5.10), we deduce that

$$\|h^{-\nu}\omega_{n+n_k}(\cdot,h)\|_{L_\infty(\mathbb{T})} \leqslant C_3(2^{-n\nu}C_3(C_4+1) + \|T_{\hat a}^{n_k} g\|_{L_\infty(\mathbb{T})}).$$

So, when n and k are large enough, by (5.14) and $\nu > 0$, setting $N = n + n_k$, we must have

$$h^{-\nu}\omega_N(\xi,h) \leqslant \pi^{-\nu} \qquad \forall\, \xi \in \mathbb{R}, h > 0.$$

In particular, setting $\xi = 0$ or $-h$ in the above inequality, we have

$$\big|[T_{\hat a}^N g](h) - [T_{\hat a}^N g](0)\big| \leqslant \pi^{-\nu}|h|^\nu \qquad \forall\, h \in \mathbb{R}.$$

By (5.13) and Proposition 4.2, we see that $\hat a(\pi) = 0$. Since $g(0) = \hat a(\pi) = 0$, by induction, one can verify that $[T_{\hat a}^n g](0) = 0$ for all $n \in \mathbb{N}$. Therefore, it follows from the above inequality that

$$\frac{[T_{\hat a}^N g](h)}{|\sin(h/2)|^\nu} \leqslant (\pi/2)^\nu \frac{[T_{\hat a}^N g](h)}{|h|^\nu} \leqslant 2^{-\nu} < 1 \qquad \forall\, h \in [-\pi,\pi]\backslash\{0\}.$$

In order to show that (ii)⇒(iii), by Proposition 4.2, it suffices to show $\rho_\tau(\hat a,\infty) < 1$ for sufficiently small $\tau > 0$. Since $\hat a \in C^\beta(\mathbb{T})$ with $\beta > 0$, we must have $|\hat a(\cdot+\pi)|^2/|\sin(\cdot/2)|^{2\beta} \in L_\infty(\mathbb{T})$. So, for any $0 < \tau < \beta$, by $\nu = \min(\beta,\tau)$, we have $\nu = \tau$, and by Theorem 4.1, we must have

$$\rho_\tau(\hat a,\infty) = \inf_{n\in\mathbb{N}} \|[T_{\hat a}^n g]/g\|_{L_\infty(\mathbb{T})}^{1/n} \leqslant \big\|[T_{\hat a}^N g]/g\big\|_{L_\infty(\mathbb{T})}^{1/N} \leqslant 2^{-\nu/N} < 1.$$

So, (iii) holds for all $0 < \tau < \beta$, and therefore (iii) holds for all $\tau > 0$. Hence, (ii)⇒(iii).

By (iii) and Proposition 4.2, we have $\hat a(\pi) = 0$. So, it follows from $\hat a \in C^\beta(\mathbb{T})$ that $|\hat a(\cdot+\pi)|^2/|\sin(\cdot/2)|^{2\beta} \in L_\infty(\mathbb{T})$. Now it is straightforward to see that (iii)⇒(iv), since $\rho(\hat a) \leqslant \rho_{2\beta}(\hat a,\infty) < 1$.

By the definition of $\rho(\hat a)$, (iv) implies that $\rho_\tau(\hat a,\infty) < 1$ and $|\hat a(\cdot+\pi)|^2/|\sin(\cdot/2)|^\tau \in L_\infty(\mathbb{T})$ for some $\tau \geqslant 0$. On the other hand, since $\hat a(0) = 1$ and $\hat a$ is continuous, by $[T_{\hat a}^n 1](0) \geqslant |\hat a(0)|^n = 1$, we must have $\rho_0(\hat a,\infty) \geqslant 1$. Therefore, $\tau > 0$, and so (iv)⇒(v).

Now we show that (v)⇒(ii). By (v) and Proposition 4.2, $\hat a(\pi) = 0$ and $|\hat a(\cdot+\pi)|^2/|\sin(\cdot/2)|^\tau \in L_\infty(\mathbb{T})$. Let $\hat\eta$ be defined in (5.1). Then $\eta \in L_{2,\infty,0}(\mathbb{R})$ and the shifts of η are stable in $L_2(\mathbb{R})$. Consequently, it is easy to directly verify that

$$H := [\hat a(\cdot/2)\hat\eta(\cdot/2) - \hat\eta, \hat a(\cdot/2)\hat\eta(\cdot/2) - \hat\eta]/|\sin(\cdot/2)|^\tau \in L_\infty(\mathbb{T}). \tag{5.15}$$

So, η is an admissible function in $L_{2,\infty,0}(\mathbb{R})$ with respect to $\hat a$ such that the shifts of η are stable in $L_2(\mathbb{R})$. Taking $f = \eta$ and defining f_n as in (2.2), we show that $\{f_n\}_{n=1}^\infty$ is a Cauchy sequence in $L_{2,\infty,0}(\mathbb{R})$.

Note that

$$|\widehat{f_{n+1}}(\xi) - \widehat{f_n}(\xi)| = |\hat{a}(2^{-1-n}\xi)\hat{f}(2^{-1-n}\xi) - \hat{f}(2^{-n}\xi)| \prod_{j=1}^{n} |\hat{a}(2^{-j}\xi)|. \tag{5.16}$$

Consequently, we have

$$[\widehat{f_{n+1}} - \widehat{f_n}, \widehat{f_{n+1}} - \widehat{f_n}](\xi) = \left[T_{\hat{a}}^n([\hat{a}(\cdot/2)\hat{f}(\cdot/2) - \hat{f}, \hat{a}(\cdot/2)\hat{f}(\cdot/2) - \hat{f}])\right](\xi) \qquad \forall\, n \in \mathbb{N}.$$

By (5.15), we have

$$[\hat{a}(\cdot/2)\hat{f}(\cdot/2) - \hat{f}, \hat{a}(\cdot/2)\hat{f}(\cdot/2) - \hat{f}](\xi) = H(\xi)g(\xi)$$

and $H \in L_\infty(\mathbb{T})$, where $g(\xi) := |\sin(\xi/2)|^\tau$. Hence, we have

$$\|f_{n+1} - f_n\|^2_{L_{2,\infty,0}(\mathbb{R})} = \|[\widehat{f_{n+1}} - \widehat{f_n}, \widehat{f_{n+1}} - \widehat{f_n}]\|_{L_\infty(\mathbb{T})} \leqslant \|H\|_{L_\infty(\mathbb{T})} \|T_{\hat{a}}^n g\|_{L_\infty(\mathbb{T})}. \tag{5.17}$$

By our assumption in (v), we have $\rho_\tau(\hat{a}, \infty) < 1$, and therefore, for any ρ such that $\rho_\tau(\hat{a}, \infty) < \rho < 1$, there exists a positive constant C such that $\|T_{\hat{a}}^n g\|_{L_\infty(\mathbb{T})} \leqslant C\rho^n$ for all $n \in \mathbb{N}$. Now it follows from (5.17) that

$$\|f_{n+1} - f_n\|_{L_{2,\infty,0}(\mathbb{R})} \leqslant \|H\|^{1/2}_{L_\infty(\mathbb{T})} C^{1/2} \rho^{n/2} \qquad \forall\, n \in \mathbb{N}, \tag{5.18}$$

which yields, by $0 < \rho < 1$, that $\{f_n\}_{n=1}^\infty$ is a Cauchy sequence in $L_{2,\infty,0}(\mathbb{R})$. So, (v)⇒(ii).

Finally, we show that (iii)⇒(i). Let f be an admissible function in $L_{2,\infty,0}(\mathbb{R})$ with respect to $\hat{a}$ such that (2.1) holds for some $\tau > 0$. By Proposition 4.2 and (iii), we have $\hat{a}(\pi) = 0$. Now by $\hat{a} \in C^\beta(\mathbb{T})$, we must have $|\hat{a}(\cdot + \pi)|^2/|\sin(\cdot/2)|^{2\beta} \in L_\infty(\mathbb{T})$. Replacing τ by $\min(\tau, 2\beta)$, we see that (2.1) still holds and $|\hat{a}(\cdot + \pi)|^2/|\sin(\cdot/2)|^\tau \in L_\infty(\mathbb{T})$. By our assumption in (iii) and $\tau > 0$, we have $\rho_\tau(\hat{a}, \infty) < 1$. Now by the same proof for showing (v)⇒(ii), we see that $\{f_n\}_{n=1}^\infty$ is a Cauchy sequence in $L_{2,\infty,0}(\mathbb{R})$. Therefore, (iii)⇒(i).

Now we prove the rest of Theorem 2.1. Since $\nu_2(\hat{a}) > 0$, we must have $\hat{a}(\pi) = 0$, and (5.3) holds. Consequently,

$$\prod_{j=1}^{n} |\hat{a}(2^{-j}\xi)|^2 \leqslant \|T_{\hat{a}}^n 1\|_{L_\infty(\mathbb{T})} \leqslant C_1^2 \qquad \forall\, n \in \mathbb{N}, \xi \in \mathbb{R}. \tag{5.19}$$

Since $\hat{a} \in C^\beta(\mathbb{T})$ and since we assumed $0 < \beta \leqslant 1$, there exists a positive constant C such that $|\hat{a}(\xi_1) - \hat{a}(\xi_2)| \leqslant C|\xi_1 - \xi_2|^\beta$ for all $\xi_1, \xi_2 \in \mathbb{R}$. We deduce that

$$\begin{aligned}
|\hat{\phi}(\xi_1) - \hat{\phi}(\xi_2)| &= \left| \sum_{j=1}^{\infty} \left[\prod_{k=1}^{j-1} \hat{a}(2^{-k}\xi_1) \right] [\hat{a}(2^{-j}\xi_1) - \hat{a}(2^{-j}\xi_2)] \left[\prod_{\ell=j+1}^{\infty} \hat{a}(2^{-\ell}\xi_2) \right] \right| \\
&\leqslant \sum_{j=1}^{\infty} \left| \prod_{k=1}^{j-1} \hat{a}(2^{-k}\xi_1) \right| \times \left| \hat{a}(2^{-j}\xi_1) - \hat{a}(2^{-j}\xi_2) \right| \times \left| \prod_{\ell=j+1}^{\infty} \hat{a}(2^{-\ell}\xi_2) \right| \\
&\leqslant C_1^2 C \sum_{j=1}^{\infty} 2^{-j\beta} |\xi_1 - \xi_2|^\beta \leqslant |\xi_1 - \xi_2|^\beta C_1^2 C/(1 - 2^{-\beta}).
\end{aligned}$$

So, $\hat{\phi} \in \Lambda^\beta(\mathbb{R})$. Let $\hat{\eta}$ be defined in (5.1). Take $f = \eta$ and define f_n as in (2.2). Since $\nu_2(\hat{a}) > 0$, $\{f_n\}_{n=1}^\infty$ is a Cauchy sequence in $L_{2,\infty,0}(\mathbb{R})$. In particular, by $\lim_{n\to\infty} \widehat{f_n}(\xi) = \hat{\phi}(\xi)$, we have $\phi \in L_{2,\infty,0}(\mathbb{R})$ and $\lim_{n\to\infty} \|f_n - \phi\|_{L_{2,\infty,0}(\mathbb{R})} = 0$. By induction, we have $[\widehat{f_n}, \widehat{f_n}] = T_{\hat{a}}^n[\hat{f}, \hat{f}]$. Note that $[\widehat{f_n}, \widehat{f_n}]$ is continuous. Now define $g := [\hat{\eta}, \hat{\eta}]$. By the definition of $\hat{\eta}$ in (5.1) and $\hat{\phi} \in C^\beta(\mathbb{R})$, it is easy to check that $g \in C^\beta(\mathbb{T})$. Consequently, taking $k = 0$ in (5.6), wee see that a similar result as in (5.10) holds; that is, there exists a positive constant C_5 such that

$$h^{-\beta} \Big\| [\widehat{f_n}, \widehat{f_n}] - [\widehat{f_n}, \widehat{f_n}](\cdot - h) \Big\|_{L_\infty(\mathbb{T})} = h^{-\beta} \|\omega_n(\cdot, h)\|_{L_\infty(\mathbb{T})} \leqslant C_5 \qquad \forall\, h > 0, \tag{5.20}$$

where ω_n is defined in (5.5) and we used the fact $[\widehat{f_n}, \widehat{f_n}] = T_{\hat{a}}^n[\hat{f}, \hat{f}] = T_{\hat{a}}^n g$. Since $\lim_{n\to\infty} \|f_n - \phi\|_{L_{2,\infty,0}(\mathbb{R})} = 0$, we deduce that $\lim_{n\to\infty} \|[\widehat{f_n}, \widehat{f_n}] - [\hat{\phi}, \hat{\phi}]\|_{C(\mathbb{T})} = 0$. Now it follows from (5.20) that $\|[\hat{\phi}, \hat{\phi}] - [\hat{\phi}, \hat{\phi}](\cdot - h)\|_{C(\mathbb{T})} \leqslant C_5 h^\beta$ for all $h > 0$. So, $[\hat{\phi}, \hat{\phi}] \in \Lambda^\beta(\mathbb{T})$.

Now we prove (2.3). Since $0 \leqslant \nu < \nu_2(\hat{a})$, by the definition of $\nu_2(\hat{a})$, there exists $\tau > 0$ such that $\rho_\tau(\hat{a}, \infty) < 2^{-2\nu}$ and $|\hat{a}(\cdot + \pi)|^2/|\sin(\cdot/2)|^\tau \in L_\infty(\mathbb{T})$. Let $\hat{\eta}$ be defined in (5.1). Take $f = \eta$ and define f_n as in (2.2). As in the proof of (v)$\Rightarrow$(ii), we see that (5.18) holds for $\rho_\tau(\hat{a}, \infty) < \rho < 2^{-2\nu}$. Denote $g_n := \widehat{f_{n+1}} - \widehat{f_n}$ and

$$[g_n, g_n]_\nu(\xi) := \sum_{k\in\mathbb{Z}} |g_n(\xi + 2\pi k)|^2 (1 + |\xi + 2\pi k|^2)^\nu, \qquad \xi \in [-\pi, \pi].$$

Since $\hat{f} = \hat{\eta}$ is supported inside $[-3\pi/2, -\varepsilon] \cup [\varepsilon, 3\pi/2]$, it follows from (5.16) that g_n is supported inside $[-3\pi 2^n, -\varepsilon 2^n] \cup [\varepsilon 2^n, 3\pi 2^n]$. Therefore, for $\xi \in [-\pi, \pi]$, it follows from (5.18) that

$$\begin{aligned}
[g_n, g_n]_\nu(\xi) &= \sum_{|\xi+2\pi k| \leqslant 3\pi 2^n} |g_n(\xi + 2\pi k)|^2 (1 + |\xi + 2\pi k|^2)^\nu \\
&\leqslant (1 + (3\pi 2^n)^2)^\nu \sum_{k\in\mathbb{Z}} |g_n(\xi + 2\pi k)|^2 \\
&= (1 + (3\pi 2^n)^2)^\nu \|f_{n+1} - f_n\|^2_{L_{2,\infty,0}(\mathbb{R})} \leqslant C_6 (2^{2\nu}\rho)^n,
\end{aligned}$$

where $C_6 := 18^\nu \pi^{2\nu} C \|[h, h]\|_{L_\infty(\mathbb{T})} < \infty$. Since $[g_n, g_n]$ is continuous, we conclude that

$$[g_n, g_n]_\nu(\xi) \leqslant C_6 (2^{2\nu}\rho)^n \qquad \forall\, n \in \mathbb{N},\ \xi \in [-\pi, \pi]. \tag{5.21}$$

Since $\nu_2(\hat{a}) > 0$, we have $\phi = \eta + \sum_{n=0}^\infty (f_{n+1} - f_n)$ in $L_{2,\infty,0}(\mathbb{R})$ with $f_0 := \eta$. Since all $\hat{\phi}, \hat{\eta}$, and $g_n, n \in \mathbb{N}$, are continuous and g_n is supported inside $[-3\pi 2^n, -\varepsilon 2^n] \cup [\varepsilon 2^n, 3\pi 2^n]$, we must have $\hat{\phi}(\xi) = \hat{\eta}(\xi) + \sum_{n=0}^\infty g_n(\xi)$ for all $\xi \in \mathbb{R}$, where the series is in fact a finite sum for any ξ in any bounded set. Therefore, for all $\xi \in [-2\pi, 2\pi]$ and $N \geqslant 3$, we have

$$\begin{aligned}
\left(\sum_{|k|=N+1}^\infty (1 + |\xi + 2\pi k|^2)^\nu |\hat{\phi}(\xi + 2\pi k)|^2 \right)^{1/2} &\leqslant \sum_{n=\log_2(N/3)}^\infty [g_n, g_n]_\nu^{1/2}(\xi) \\
&\leqslant C_6^{1/2} \sum_{n=\log_2(N/3)}^\infty (2^\nu \rho^{1/2})^n \leqslant C_6^{1/2} (N/3)^{\log_2(2^\nu \rho^{1/2})} / (1 - 2^\nu \rho^{1/2}).
\end{aligned}$$

Since $2^\nu \rho^{1/2} < 1$, we have $\log_2(2^\nu \rho^{1/2}) < 0$, and therefore $\lim_{N\to\infty}(N/3)^{\log_2(2^\nu\rho^{1/2})} = 0$. Hence, the series $\sum_{k=-N}^{N}(1+|\xi+2\pi k|^2)^\nu |\hat\phi(\xi+2\pi k)|^2$ is uniformly convergent as $N\to\infty$ for all $\xi\in[-2\pi,2\pi]$. Since $\hat\phi$ is continuous, we conclude that $[\hat\phi,\hat\phi]_\nu \in C(\mathbb{T})$. That is, (2.3) holds. □

For a mask $\hat a \in C^\beta(\mathbb{T})$ with $\beta>0$, a more complicated argument can be used to show a stronger statement that $\hat\phi\in C^\beta(\mathbb{R})$ and $[\hat\phi,\hat\phi]\in C^\beta(\mathbb{T})$, instead of the result $\hat\phi\in\Lambda^{\min(1,\beta)}(\mathbb{R})$ and $[\hat\phi,\hat\phi]\in\Lambda^{\min(1,\beta)}(\mathbb{T})$ stated in Theorem 2.1. We shall address this technical issue elsewhere.

6. Proof of Theorem 2.3. Before we present the proof of Theorem 2.3, we need the two following auxiliary results.

PROPOSITION 6.1. *Let $f\in L_{2,1,\gamma_2}(\mathbb{R})$ with $\gamma_2>0$. Then for $0\leqslant\gamma_1<\gamma_2$,*

$$[\hat f,\hat f](\xi+i\zeta) := \sum_{k\in\mathbb{Z}} |\hat f(\xi+i\zeta)|^2 \leqslant C_{\gamma_2-\gamma_1}\|f\|^2_{L_{2,1,\gamma_2}(\mathbb{R})} \tag{6.1}$$
$$\forall\,\xi\in\mathbb{R},\ \zeta\in[-\gamma_1,\gamma_1],$$

where

$$C_\gamma := \left\| \sum_{k\in\mathbb{Z}} e^{-2\gamma|\cdot-k|} \right\|_{L_\infty(\mathbb{R})} < \infty$$

for $\gamma>0$. If in addition $\hat f(2\pi k)=0$ for all $k\in\mathbb{Z}$, then $f=h-h(\cdot-1)$ and $h\in L_{2,1,\gamma_1}(\mathbb{R})$ for all $0\leqslant\gamma_1<\gamma_2$, where $h:=\sum_{k=0}^\infty f(\cdot-k)$.

Proof. Since $f\in L_{2,1,\gamma_2}(\mathbb{R})$, by the Cauchy–Schwarz inequality, for $0\leqslant\gamma_1<\gamma_2$, we have

$$\int_0^1 \left(\sum_{k\in\mathbb{Z}} |f(x-k)|e^{\gamma_1|x-k|}\right)^2 dx \leqslant C_\gamma \int_0^1 \sum_{k\in\mathbb{Z}} |f(x-k)|^2 e^{2\gamma_2|x-k|}\,dx \tag{6.2}$$
$$= C_\gamma \|f\|^2_{L_{2,1,\gamma_2}(\mathbb{R})},$$

where $\gamma:=\gamma_2-\gamma_1>0$. Using the Fourier series of $[\hat f,\hat f]$ and $\hat f(\xi+i\zeta)=\widehat{e^{\zeta\cdot}f}(\xi)$, for any fixed $\zeta\in[-\gamma_1,\gamma_1]$ and almost every $\xi\in\mathbb{R}$, we deduce from (6.2) that

$$[\hat f,\hat f](\xi+i\zeta) = \left|\sum_{k\in\mathbb{Z}} e^{ik\xi}\int_{\mathbb{R}} e^{\zeta x} f(x)\overline{e^{\zeta(x+k)}f(x+k)}\,dx\right|$$
$$\leqslant \sum_{k\in\mathbb{Z}}\int_{\mathbb{R}} e^{|\zeta x|}|f(x)|e^{|\zeta(x+k)|}|f(x+k)|\,dx$$
$$= \int_0^1 \left(\sum_{k\in\mathbb{Z}} |f(x+k)|e^{|\zeta|\times|x+k|}\right)^2 dx \leqslant C_\gamma\|f\|^2_{L_{2,1,\gamma_2}(\mathbb{R})}.$$

Since $\hat f$ is continuous, it follows from the above inequality that (6.1) holds for all $\xi\in\mathbb{R}$.

If $\hat f(2\pi k)=0$ for all $k\in\mathbb{Z}$, then we have $\sum_{k\in\mathbb{Z}} f(\cdot-k)=0$. From the definition of h, we see that $f=h-h(\cdot-1)$ and $h=-\sum_{k=-\infty}^{-1} f(\cdot-k)$. We now verify

that $h \in L_{2,1,\gamma_3}$ for all $0 < \gamma_3 < \gamma_2$. Take γ_1 such that $\gamma_3 < \gamma_1 < \gamma_2$. Then by $h = \sum_{k=0}^{\infty} f(\cdot - k)$, we have

$$\sum_{j=0}^{\infty} |h(x-j)|e^{\gamma_3|x-j|} \leqslant \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} |f(x-j-k)|e^{\gamma_3|x-j|}$$

$$= \sum_{k=0}^{\infty} |f(x-k)|e^{\gamma_1|x-k|}e^{-\gamma_1|x-k|}\sum_{j=0}^{k} e^{\gamma_3|x-j|}.$$

For $x \in [0,1]$, by $\gamma_3 < \gamma_1 < \gamma_2$ and $k \geqslant 0$, we have

$$e^{-\gamma_1|x-k|}\sum_{j=0}^{k} e^{\gamma_3|x-j|} \leqslant e^{-\gamma_1(k-1)}\sum_{j=0}^{k} e^{\gamma_3(j+1)} \leqslant \frac{e^{\gamma_2+2\gamma_3}}{e^{\gamma_3}-1} := C.$$

Therefore, we deduce that

$$\sum_{j=0}^{\infty} |h(x-j)|e^{\gamma_3|x-j|} \leqslant C\sum_{k=0}^{\infty} |f(x-k)|e^{\gamma_1|x-k|}, \qquad x \in [0,1].$$

Similarly, using $h = -\sum_{k=-\infty}^{-1} f(\cdot - k)$, we have

$$\sum_{j=-\infty}^{-1} |h(x-j)|e^{\gamma_3|x-j|} \leqslant C\sum_{k=-\infty}^{-1} |f(x-k)|e^{\gamma_1|x-k|}, \qquad x \in [0,1].$$

Hence,

$$\sum_{j\in\mathbb{Z}} |h(x-j)|e^{\gamma_3|x-j|} \leqslant C\sum_{k\in\mathbb{Z}} |f(x-k)|e^{\gamma_1|x-k|}.$$

By (6.2), we conclude that

$$\|h\|^2_{L_{2,1,\gamma_3}(\mathbb{R})} = \int_0^1 \sum_{j\in\mathbb{Z}} |h(x-j)e^{\gamma_3|x-k|}|^2\,dx \leqslant \int_0^1 \left(\sum_{j\in\mathbb{Z}} |h(x-j)|e^{\gamma_3|x-j|}\right)^2 dx$$

$$\leqslant C\int_0^1 \left(\sum_{k\in\mathbb{Z}} |f(x-k)|e^{\gamma_1|x-k|}\right)^2 dx \leqslant CC_{\gamma_2-\gamma_1}\|f\|^2_{L_{2,1,\gamma_2}(\mathbb{R})}.$$

So, $h \in L_{2,1,\gamma_3}(\mathbb{R})$ for all $0 < \gamma_3 < \gamma_2$. This completes the proof. □

It follows from (6.1) and $\hat{f}(\xi + i\zeta) = \widehat{e^{\zeta\cdot}f}(\xi)$ that for any $1 \leqslant p, q \leqslant \infty$,

$$\|f\|^2_{L_{2,p,\gamma_1}(\mathbb{R})} \leqslant \|f\|^2_{L_{2,\infty,\gamma_1}(\mathbb{R})} \leqslant C_{\gamma_2-\gamma_1}\|f\|^2_{L_{2,1,\gamma_2}(\mathbb{R})} \leqslant C_{\gamma_2-\gamma_1}\|f\|^2_{L_{2,q,\gamma_2}(\mathbb{R})}$$
$$\forall\ 0 \leqslant \gamma_1 < \gamma_2. \tag{6.3}$$

Consequently, $L_{2,q,\gamma_2}(\mathbb{R}) \subseteq L_{2,p,\gamma_1}(\mathbb{R})$ for all $1 \leqslant p, q \leqslant \infty$ and $0 \leqslant \gamma_1 < \gamma_2$.

Now we have the following result on admissible functions in a cascade algorithm.

PROPOSITION 6.2. *Let $f \in L_{2,1,\gamma}(\mathbb{R})$ for some $\gamma > 0$ such that f satisfies*

$$\hat{f}(2\pi k) = 0 \qquad \forall\ k \in \mathbb{Z}\backslash\{0\}. \tag{6.4}$$

Then f is admissible with respect to $\hat{a}$ for any $\hat{a} \in C^\beta(\mathbb{T})$ with $\hat{a}(0) = 1$, $\hat{a}(\pi) = 0$, and $\beta > 0$.

Proof. Let $\widehat{f_1} := \hat{f}(2\cdot + 2\pi)$ and $\widehat{f_2} := \hat{f} - \hat{f}(2\cdot)$. Then $f_1, f_2 \in L_{2,1,\gamma/2}(\mathbb{R})$ and $\widehat{f_1}(2\pi k) = \widehat{f_2}(2\pi k) = 0$ for all $k \in \mathbb{Z}$. By Proposition 6.1, $f_1 = h_1 - h_1(\cdot - 1)$ and $f_2 = h_2 - h_2(\cdot - 1)$ for some $h_1, h_2 \in L_{2,1,\gamma_1}(\mathbb{R})$ with $0 < \gamma_1 < \gamma/2$. Therefore, we have

$$[\widehat{f_1}, \widehat{f_1}](\xi) = |1 - e^{-i\xi}|^2 [\widehat{h_1}, \widehat{h_1}](\xi) \quad \text{and} \quad [\widehat{f_2}, \widehat{f_2}](\xi) = |1 - e^{-i\xi}|^2 [\widehat{h_2}, \widehat{h_2}](\xi). \tag{6.5}$$

Since $\hat{a} \in C^\beta(\mathbb{T})$ with $\hat{a}(0) = 1$ and $\hat{a}(\pi) = 0$, there exists a positive constant C such that

$$|\hat{a}(\xi) - 1| \leqslant C|1 - e^{-i\xi}|^\beta \quad \text{and} \quad |\hat{a}(\xi + \pi)| \leqslant C|1 - e^{-i\xi}|^\beta \qquad \forall\, \xi \in \mathbb{R}. \tag{6.6}$$

Now it follows from (6.5) and (6.6) that

$$\begin{aligned}
&\sum_{k\in\mathbb{Z}} \big|\hat{a}(\xi/2 + 2\pi k)\hat{f}(\xi/2 + 2\pi k) - \hat{f}(\xi + 4\pi k)\big|^2 \\
&\qquad = \sum_{k\in\mathbb{Z}} \big|[\hat{a}(\xi/2) - 1]\hat{f}(\xi/2 + 2\pi k) + \widehat{f_2}(\xi/2 + 2\pi k)\big|^2 \\
&\qquad \leqslant 2|\hat{a}(\xi/2) - 1|^2 [\hat{f}, \hat{f}](\xi/2) + 2[\widehat{f_2}, \widehat{f_2}](\xi/2) \\
&\qquad \leqslant 2C^2|1 - e^{-i\xi/2}|^{2\beta} \big(\|[\hat{f}, \hat{f}]\|_{L_\infty(\mathbb{T})} + 2^{2-2\beta}\|[\widehat{h_2}, \widehat{h_2}]\|_{L_\infty(\mathbb{T})}\big).
\end{aligned}$$

Similarly, by $\hat{f}(\xi + 2\pi + 4\pi k) = \widehat{f_1}(\xi/2 + 2\pi k)$, we have

$$\begin{aligned}
&\sum_{k\in\mathbb{Z}} \big|\hat{a}(\xi/2 + \pi + 2\pi k)\hat{f}(\xi/2 + \pi + 2\pi k) - \hat{f}(\xi + 2\pi + 4\pi k)\big|^2 \\
&\qquad \leqslant 2|\hat{a}(\xi/2 + \pi)|^2 [\hat{f}, \hat{f}](\xi/2 + \pi) + 2[\widehat{f_1}, \widehat{f_1}](\xi/2) \\
&\qquad \leqslant 2C^2|1 - e^{-i\xi/2}|^{2\beta} \big(\|[\hat{f}, \hat{f}]\|_{L_\infty(\mathbb{T})} + 2^{2-2\beta}\|[\widehat{h_1}, \widehat{h_1}]\|_{L_\infty(\mathbb{T})}\big).
\end{aligned}$$

Letting $C_1 := 8C^2(\|[\hat{f}, \hat{f}]\|_{L_\infty(\mathbb{T})} + \|[\widehat{h_1}, \widehat{h_1}]\|_{L_\infty(\mathbb{T})} + \|[\widehat{h_2}, \widehat{h_2}]\|_{L_\infty(\mathbb{T})}) < \infty$, combining the above two inequalities, for almost every $\xi \in [-\pi, \pi]$, we have

$$[\hat{a}(\cdot/2)\hat{f}(\cdot/2) - \hat{f}, \hat{a}(\cdot/2)\hat{f}(\cdot/2) - \hat{f}](\xi) \leqslant C_1|1 - e^{-i\xi/2}|^{2\beta} \leqslant 2^\beta C_1 |\sin(\xi/2)|^{2\beta}.$$

Therefore, (2.1) holds with $\tau = 2\beta > 0$. ∎

Proof of Theorem 2.3. By (6.3), it is easy to see that (i)⇒(ii). (ii)⇒(iii) is obvious. To show that (iii)⇒(iv), we take $f = \max(0, 1 - |\cdot|)$. It is easy to check that f satisfies the condition in (6.4) and the shifts of f are stable. By $\hat{a}(\pi) = 0$ and Proposition 6.2, $f \in L_{2,p,\gamma}(\mathbb{R})$ is admissible with respect to $\hat{a}$. Therefore, (iii)⇒(iv).

If (iv) holds, by Proposition 6.1 or (6.3), $f \in L_{2,\infty,0}(\mathbb{R})$ and f is admissible with respect to $\hat{a}$. Now it follows from (iv) and (6.3) that $\{f_n\}_{n=1}^\infty$ is a Cauchy sequence in $L_{2,\infty,0}(\mathbb{R})$. Therefore, (ii) of Theorem 2.1 holds, and consequently $\nu_2(\hat{a}) > 0$. So, (iv)⇒(v).

To complete the proof, we have to show that (v)⇒(i), which is the major part of this proof. By Proposition 4.2, $\nu_2(\hat{a}) > 0$ implies $\hat{a}(\pi) = 0$. So, we can write $\hat{a}(\xi) = (1 + e^{-i\xi})\hat{A}(\xi)$, where $\hat{A}$ also has exponential decay of order r. By Theorems 2.1 and 4.1, it follows from $\nu_2(\hat{a}) > 0$ and $\hat{a}(\xi) = (1 + e^{-i\xi})\hat{A}(\xi)$ that $\inf_{n\in\mathbb{N}} \|T_{\hat{A}}^n 1\|_{L_\infty(\mathbb{T})}^{1/n} =$

$\rho_0(\hat{A},\infty)=\rho_2(\hat{a},\infty)<1$. Therefore, there exist $0<\rho<1$ and $N\in\mathbb{N}$ such that $\|T_{\hat{A}}^N 1\|_{L_\infty(\mathbb{T})}^{1/N}<\rho<1$. Since $\hat{A}$ has exponential decay of order r, the operator $[T_{\hat{A}}^n 1](\xi)$ in (1.5) is well defined for $\xi\in\Gamma_{2r}:=\{z\in\mathbb{C}\,:\,|\mathrm{Im}(z)|<2r\}$. Since $\hat{A}$ is a 2π-periodic continuous function on the strip Γ_r, there exists $\gamma_1>0$ such that

$$\|[T_{\hat{A}}^N 1](\cdot+i\zeta)\|_{L_\infty(\mathbb{T})}\leqslant\rho^N<1\qquad\forall\,\zeta\in[-\gamma_1,\gamma_1]. \tag{6.7}$$

Take m to be the smallest nonnegative integer such that $2^{1-m}r<\gamma_1$. For each $n\geqslant N+m$, we can uniquely write $n=Nk+j$, where $j\in\{m,m+1,\ldots,m+N-1\}$ and $k\in\mathbb{N}$. So, for every $\zeta\in[-\gamma,\gamma]$ and $0<\gamma<2r$, by (6.7) and the extended definition of $T_{\hat{A}}$ in (1.5), we have $|2^{-j}\zeta|\leqslant 2^{-m}\gamma<2^{1-m}r<\gamma_1$ and

$$\begin{aligned}\|[T_{\hat{A}}^n 1](\cdot+i\zeta)\|_{L_\infty(\mathbb{T})}&=\|[T_{\hat{A}}^j T_{\hat{A}}^{Nk}1](\cdot+i\zeta)\|_{L_\infty(\mathbb{T})}\\&\leqslant\|[T_{\hat{A}}^j 1](\cdot+i\zeta)\|_{L_\infty(\mathbb{T})}\|[T_{\hat{A}}^{Nk}1](\cdot+i2^{-j}\zeta)\|_{L_\infty(\mathbb{T})}\\&\leqslant\|[T_{\hat{A}}^j 1](\cdot+i\zeta)\|_{L_\infty(\mathbb{T})}\rho^{Nk}\leqslant C_1\rho^n,\end{aligned}$$

where

$$C_1:=\rho^{1-m-N}\sup\big\{\|[T_{\hat{A}}^j 1](\cdot+i\zeta)\|_{L_\infty(\mathbb{T})}\,:\,j=m,\ldots,m+N-1,\zeta\in[-\gamma,\gamma]\big\}<\infty.$$

That is, we have

$$\|[T_{\hat{A}}^n 1](\cdot+i\zeta)\|_{L_\infty(\mathbb{T})}\leqslant C_1\rho^n\qquad\forall\,n\geqslant N+m,\zeta\in[-\gamma,\gamma]. \tag{6.8}$$

Denote $\hat{F}(\xi):=\hat{a}(\xi/2)\hat{f}(\xi/2)-\hat{f}(\xi)$. Since f is admissible with respect to $\hat{a}$ and $F\in L_{2,1,\gamma}(\mathbb{R})$, we must have $\hat{F}(2\pi k)=0$ for all $k\in\mathbb{Z}$. By Proposition 6.1, $\hat{F}(\xi)=(1-e^{-i\xi})\hat{h}(\xi)$ for some $h\in L_{2,1,\gamma/2}(\mathbb{R})$. Denote $g_n:=\widehat{f_{n+1}}-\widehat{f_n}$ with $g:=g_0:=\widehat{f_1}-\hat{f}$. Then $g(\xi)=\hat{F}(\xi)$. Therefore, by Proposition 6.1, we conclude that

$$\begin{aligned}[g,g](\xi)=[\hat{F},\hat{F}](\xi)=|1-e^{-i\xi}|^2[\hat{h},\hat{h}](\xi)\leqslant C_2|1-e^{-i\xi}|^2\\ \forall\,\xi\in\Gamma_{\gamma_2}:=\{z\in\mathbb{C}\,:\,|\mathrm{Im}(z)|<\gamma_2\},\end{aligned} \tag{6.9}$$

where $\gamma_2:=\gamma/4>0$, $C_2:=C_{\gamma/4}\|h\|_{L_{2,1,\gamma/2}(\mathbb{R})}<\infty$, and $C_{\gamma/4}$ is defined in Proposition 6.1.

By the definition of f_n and by induction on n, for $n\geqslant n_0:=1-\log_2(\gamma_2/r)$, we have that $g_n(\xi)=g(2^{-n}\xi)\prod_{j=1}^n\hat{a}(2^{-j}\xi)$ for $\xi\in\Gamma_{2r}$ and g_n is holomorphic on Γ_{2r}, since $2^{-n}\Gamma_{2r}\subseteq\Gamma_{\gamma_2}$ for all $n\geqslant n_0$. Now by induction, it follows from (6.9) that

$$[g_n,g_n](\xi)=(T_{\hat{a}}^n[g,g])(\xi)\leqslant C_2(T_{\hat{a}}^n(|1-e^{-i\cdot}|^2))(\xi),\qquad\xi\in\Gamma_{2r}. \tag{6.10}$$

By $\hat{a}(\xi)=(1+e^{-i\xi})\hat{A}(\xi)$, we have $|\hat{a}(\xi)|^2=2^2|\cos(\xi/2)|^2|\hat{A}(\xi)|^2$. So, (4.8) holds with $\tau=1$. So, we deduce that

$$[T_{\hat{a}}^n(|1-e^{-i\cdot}|^2)](\xi)=|1-e^{-i\xi}|^2[T_{\hat{A}}^n 1](\xi)\leqslant(1+e^{2r})^2[T_{\hat{A}}^n 1](\xi),\qquad\xi\in\Gamma_{2r}.$$

Consequently, since $0<\gamma<2r$, it follows from (6.8) and (6.10) that

$$\begin{aligned}[g_n,g_n](\xi)\leqslant C_2(1+e^{2r})^2|[T_{\hat{A}}^n 1](\xi)|\leqslant C_1C_2(1+e^{2r})^2\rho^n\leqslant C_3\rho^n\\ \forall\,n\geqslant n_1,|\mathrm{Im}(\xi)|\leqslant\gamma,\end{aligned} \tag{6.11}$$

where $C_3 := C_1C_2(1+e^{2r})^2 < \infty$ and $n_1 = \max(n_0, N+m)$. Note that $g_n(\cdot + i\zeta)$ is the Fourier transform of $e^{\zeta\cdot}(f_{n+1} - f_n)$. For $\zeta \in [-\gamma, \gamma]$, we have

$$2\pi\|(f_{n+1}-f_n)e^{\zeta\cdot}\|^2_{L_2(\mathbb{R})} = \|g_n(\cdot+i\zeta)\|^2_{L_2(\mathbb{R})} = 2\pi\|[g_n, g_n](\cdot+i\zeta)\|_{L_1(\mathbb{T})}$$
$$\leqslant 2\pi\|[g_n, g_n](\cdot+i\zeta)\|_{L_\infty(\mathbb{T})}.$$

Now by the definition of the space $L_{2,1,\gamma}(\mathbb{R})$, it follows from the above inequality and (6.11) that

$$\|f_{n+1}-f_n\|^2_{L_{2,1,\gamma}(\mathbb{R})} = \|(f_{n+1}-f_n)e^{\gamma|\cdot|}\|^2_{L_2(\mathbb{R})} \leqslant 2C_3\rho^n \qquad \forall\, n \geqslant n_0.$$

Consequently, $\{f_n\}_{n=1}^\infty$ is a Cauchy sequence in $L_{2,1,\gamma}(\mathbb{R})$. So, (v)⇒(i).

Now we show that ϕ has exponential decay of order $2r$. Since $\hat\phi(\xi) := \prod_{j=1}^\infty \hat a(2^{-j}\xi)$ and $\hat a$ has exponential decay of order r with $\hat a(0) = 1$, we see that $\hat\phi$ can be extended into a holomorphic function in the strip Γ_{2r}. If $\nu_2(\hat a) > 0$, for every $0 < \gamma < 2r$, by (i) and $\lim_{n\to\infty} \widehat{f_n}(\xi) = \hat\phi(\xi)$ for all $\xi \in \mathbb{R}$ (here we additionally assumed that $\lim_{\xi\to 0}\hat f(\xi) = 1$, which is satisfied by many initial admissible functions), we see that $\lim_{n\to\infty}\|f_n - \phi\|_{L_{2,1,\gamma}(\mathbb{R})} = 0$ and hence $\phi \in L_{2,1,\gamma}(\mathbb{R})$. Therefore, $\phi \in L_{2,1,\gamma}(\mathbb{R})$ for all $0 \leqslant \gamma < 2r$. That is, (1.9) holds. □

7. Proof of Theorem 4.1. Before we present a proof of Theorem 4.1, we need the following result.

PROPOSITION 7.1. *Let $\hat a$ and f be 2π-periodic measurable functions. Assume that $f(\xi) > 0$ for almost every $\xi \in \mathbb{R}$ and $[T^n_{\hat a} f]/f \in L_\infty(\mathbb{T})$ for all $n \geqslant n_0$, where $n_0 \in \mathbb{N}$. Then*

$$\limsup_{n\to\infty} \|[T^n_{\hat a} f]/f\|^{1/n}_{L_\infty(\mathbb{T})} = \lim_{n\to\infty} \|[T^n_{\hat a} f]/f\|^{1/n}_{L_\infty(\mathbb{T})} = \inf_{n\geqslant n_0} \|[T^n_{\hat a} f]/f\|^{1/n}_{L_\infty(\mathbb{T})}. \tag{7.1}$$

Proof. Denote $\rho := \inf_{n\geqslant n_0} \|[T^n_{\hat a} f]/f\|^{1/n}_{L_\infty(\mathbb{T})}$. Then $\rho < \infty$ by $[T^{n_0}_{\hat a} f]/f \in L_\infty(\mathbb{T})$. Since for all $n \geqslant n_0$, $\|[T^n_{\hat a} f]/f\|^{1/n}_{L_\infty(\mathbb{T})} \geqslant \rho$, we have

$$\liminf_{n\to\infty} \|[T^n_{\hat a} f]/f\|^{1/n}_{L_\infty(\mathbb{T})} \geqslant \rho. \tag{7.2}$$

For any $\varepsilon > 0$, there exists $m \in \mathbb{N}$ such that $m \geqslant n_0$ and $\|[T^m_{\hat a} f]/f\|^{1/m}_{L_\infty(\mathbb{T})} \leqslant \rho + \varepsilon$. Since $f(\xi) > 0$ for almost every $\xi \in \mathbb{R}$, we deduce that

$$[T^m_{\hat a} f](\xi) \leqslant \|[T^m_{\hat a} f]/f\|_{L_\infty(\mathbb{T})} f(\xi) \leqslant f(\xi)(\rho+\varepsilon)^m \qquad \text{a.e. } \xi \in \mathbb{R}. \tag{7.3}$$

For each $n \geqslant 2m$, we can uniquely write $n = mN + j$, where $N \in \mathbb{N}$ and $j \in \{m, m+1, \dots, 2m-1\}$. Therefore, by (7.3),

$$[T^n_{\hat a} f](\xi) = [T^j_{\hat a}(T^{mN}_{\hat a} f)](\xi) \leqslant (\rho+\varepsilon)^{mN}[T^j_{\hat a} f](\xi).$$

Since $f(\xi) > 0$ for almost every $\xi \in \mathbb{R}$, the above inequality yields that

$$\frac{[T^n_{\hat a} f](\xi)}{f(\xi)} \leqslant (\rho+\varepsilon)^{mN}\frac{[T^j_{\hat a} f](\xi)}{f(\xi)} \leqslant C(\rho+\varepsilon)^n \qquad \text{a.e. } \xi \in \mathbb{R},$$

where

$$C := \max\{(\rho+\varepsilon)^{-j}\|[T^j_{\hat a} f]/f\|_{L_\infty(\mathbb{T})} \,:\, j = m, \dots, 2m-1\} < \infty.$$

Thus, $\|[T_{\hat a}^n f]/f\|_{L_\infty(\mathbb{T})}^{1/n} \leqslant C^{1/n}(\rho+\varepsilon)$ for all $n \geqslant 2m$. Consequently,

$$\limsup_{n\to\infty} \|[T_{\hat a}^n f]/f\|_{L_\infty(\mathbb{T})}^{1/n} \leqslant (\rho+\varepsilon).$$

Taking $\varepsilon \to 0$, we have $\limsup_{n\to\infty} \|[T_{\hat a}^n f]/f\|_{L_\infty(\mathbb{T})}^{1/n} \leqslant \rho$. By (7.2), (7.1) holds. ∎

Proof of Theorem 4.1. Without loss of generality, we can assume that $|\hat a(0)|^2 = 1$; otherwise, we consider $\hat a/|\hat a(0)|$. By $|\hat a(\xi)|^2 = 2^{2\tau}|\cos(\xi/2)|^{2\tau}|\hat A(\xi)|^2$, (4.8) holds. Setting $f = 1$ in (4.8), we have

$$\left[T_{\hat a}^n(|\sin(\cdot/2)|^{2\tau})\right](\xi) = |\sin(\xi/2)|^{2\tau}[T_{\hat A}^n 1](\xi) \qquad \forall\, n \in \mathbb{N} \text{ a.e. } \xi \in \mathbb{R}. \tag{7.4}$$

Since $|\sin(\xi/2)|^{2\tau} \leqslant 1$, it follows from (7.4) that

$$\|T_{\hat a}^n(|\sin(\cdot/2)|^{2\tau})\|_{L_\infty(\mathbb{T})} \leqslant \|T_{\hat A}^n 1\|_{L_\infty(\mathbb{T})}$$

for all $n \in \mathbb{N}$. Consequently, by $\hat A \in L_\infty(\mathbb{T})$ and Proposition 7.1, we conclude that

$$\rho := \rho_{2\tau}(\hat a, \infty) := \limsup_{n\to\infty} \left\|T_{\hat a}^n(|\sin(\cdot/2)|^{2\tau})\right\|_{L_\infty(\mathbb{T})}^{1/n} \leqslant \limsup_{n\to\infty} \|T_{\hat A}^n 1\|_{L_\infty(\mathbb{T})}^{1/n}. \tag{7.5}$$

On the other hand, by (7.4), we have

$$[T_{\hat A}^n 1](\xi) = \frac{[T_{\hat a}^n(|\sin(\cdot/2)|^{2\tau})](\xi)}{|\sin(\xi/2)|^{2\tau}} =: g_n(\xi) \qquad \forall\, n \in \mathbb{N}. \tag{7.6}$$

By the definition of $T_{\hat a}$ and g_n, we have

$$g_n(\xi) = |\hat A(\xi/2)|^2 g_{n-1}(\xi/2) + \frac{|\hat a(\xi/2+\pi)|^2}{|\sin(\xi/2)|^{2\tau}}[T_{\hat a}^{n-1}(|\sin(\cdot/2)|^{2\tau})](\xi/2+\pi).$$

Since $\tau \geqslant 0$, for almost every $\xi \in [-\pi,\pi]$, we have

$$|\hat a(\xi/2+\pi)|^2/|\sin(\xi/2)|^{2\tau} = |\hat A(\xi/2+\pi)|^2/|\cos(\xi/4)|^{2\tau} \leqslant 2^\tau \|\hat A\|_{L_\infty(\mathbb{T})},$$

from which we see that

$$g_n(\xi) \leqslant |\hat A(\xi/2)|^2 g_{n-1}(\xi/2) + 2^\tau \|\hat A\|_{L_\infty(\mathbb{T})} \|T_{\hat a}^{n-1}(|\sin(\cdot/2)|^{2\tau})\|_{L_\infty(\mathbb{T})} \quad \text{a.e. } \xi \in [-\pi,\pi].$$

By induction on n and $g_0 = 1$, we deduce from the above inequality that for almost every $\xi \in [-\pi,\pi]$ and all $n \in \mathbb{N}$,

$$\begin{aligned} g_n(\xi) &\leqslant \prod_{j=1}^n |\hat A(2^{-j}\xi)|^2 \\ &\quad + 2^\tau \|\hat A\|_{L_\infty(\mathbb{T})} \sum_{j=1}^{n-1} \|T_{\hat a}^j(|\sin(\cdot/2)|^{2\tau})\|_{L_\infty(\mathbb{T})} \prod_{k=1}^{n-j-1} |\hat A(2^{-k}\xi)|^2. \end{aligned} \tag{7.7}$$

Since $|\hat a|^2 \in C^\beta(\mathbb{T})$ and $|\hat a(0)|^2 = 1$, we define

$$\Phi(\xi) := \prod_{j=1}^\infty |\hat a(2^{-j}\xi)|^2, \qquad \xi \in \mathbb{R}.$$

Then Φ is continuous and $\Phi(0) = 1$. Thus, there exists $0 < \varepsilon_0 < \pi$ such that $1/2 \leqslant \Phi(\xi) \leqslant 3/2$ for all $\xi \in [-\varepsilon_0, \varepsilon_0]$. By

$$|\hat{a}(\xi)|^2 = 2^{2\tau} |\cos(\xi/2)|^{2\tau} |\hat{A}(\xi)|^2,$$

for $n \geqslant \log_2(\pi/\varepsilon_0)$ and $\xi \in [-\pi, \pi]$, we have

$$\begin{aligned}\prod_{j=1}^{n} |\hat{A}(2^{-j}\xi)|^2 &= \Big|\frac{\sin(2^{-1-n}\xi)}{\sin(\xi/2)}\Big|^{2\tau} \prod_{j=1}^{n} |\hat{a}(2^{-j}\xi)|^2 \\ &= \Big|\frac{\sin(2^{-1-n}\xi)}{\sin(\xi/2)}\Big|^{2\tau} \frac{\Phi(\xi)}{\Phi(2^{-n}\xi)} \leqslant \Big|\frac{2^{-1-n}\xi}{\xi/\pi}\Big|^{2\tau} \frac{\Phi(\xi)}{1/2} \\ &\leqslant 2^{1-2\tau}\pi^{2\tau} \|\Phi\|_{L_\infty([-\pi,\pi])} 2^{-2\tau n}.\end{aligned}$$

Therefore, we have

$$\prod_{j=1}^{n} |\hat{A}(2^{-j}\xi)|^2 \leqslant C_1 2^{-2\tau n} \qquad \forall\, \xi \in [-\pi, \pi] \tag{7.8}$$

with $C_1 := 2^{1-2\tau}\pi^{2\tau}\|\Phi\|_{L_\infty([-\pi,\pi])} < \infty$. Since $\prod_{j=1}^{n} |\hat{a}(2^{-j}\xi)|^2 = \Phi(\xi)/\Phi(2^{-n}\xi)$ and $1/2 \leqslant \Phi(\xi) \leqslant 3/2$ for $\xi \in [-\varepsilon_0, \varepsilon_0]$, we have

$$\begin{aligned}\|T_{\hat{a}}^n(|\sin(\cdot/2)|^{2\tau})\|_{L_\infty(\mathbb{T})} &\geqslant \Big\| |\sin(2^{-1-n}\cdot)|^{2\tau} \prod_{j=1}^{n} |\hat{a}(2^{-j}\cdot)|^2 \Big\|_{L_\infty([-\pi,\pi])} \\ &= \big\| |\sin(2^{-1-n}\cdot)|^{2\tau} \Phi(\cdot)/\Phi(2^{-n}\cdot) \big\|_{L_\infty([-\pi,\pi])} \\ &\geqslant |\sin(2^{-1-n}\varepsilon_0)|^{2\tau} \Phi(\varepsilon_0)/\Phi(2^{-n}\varepsilon_0) \\ &\geqslant 3^{-1}\pi^{-2\tau}\varepsilon_0^{2\tau} 2^{-2\tau n}.\end{aligned}$$

By the definition of ρ, it follows from the above inequality that $\rho \geqslant 2^{-2\tau}$. By (7.8) and the definition of ρ in (7.5), for any $\varepsilon > 0$, there exists a positive constant C with $C \geqslant C_1$ such that

$$\Big\| \prod_{j=1}^{n} |\hat{A}(2^{-j}\cdot)|^2 \Big\|_{L_\infty([-\pi,\pi])} \leqslant C 2^{-2\tau n} \leqslant C(\rho + \varepsilon)^n$$

and

$$\|T_{\hat{a}}^n(|\sin(\cdot/2)|^{2\tau})\|_{L_\infty(\mathbb{T})} \leqslant C(\rho + \varepsilon)^n \quad \forall\, n \in \mathbb{N}.$$

Now it follows from (7.7) and the above inequalities that

$$\|g_n\|_{L_\infty(\mathbb{T})} = \|g_n\|_{L_\infty([-\pi,\pi])} \leqslant C(\rho + \varepsilon)^n + 2^\tau \|\hat{A}\|_{L_\infty(\mathbb{T})} C^2 n (\rho + \varepsilon)^{n-1},$$

from which we deduce that $\limsup_{n\to\infty} \|g_n\|_{L_\infty(\mathbb{T})}^{1/n} \leqslant \rho + \varepsilon$. Taking $\varepsilon \to 0$, by the definition of g_n in (7.6), we conclude that

$$\limsup_{n\to\infty} \|T_{\hat{A}}^n 1\|_{L_\infty(\mathbb{T})}^{1/n} = \limsup_{n\to\infty} \|g_n\|_{L_\infty(\mathbb{T})}^{1/n} \leqslant \rho.$$

So, by Proposition 7.1, the proof is completed by the above inequality and (7.5). □

REFERENCES

[1] A. S. Cavaretta, W. Dahmen, and C. A. Micchelli, *Stationary Subdivision*, Mem. Amer. Math. Soc. 453, AMS, Providence, RI, 1991.

[2] C. K. Chui and J. Z. Wang, *On compactly supported spline wavelets and a duality principle*, Trans. Amer. Math. Soc., 330 (1992), pp. 903–915.

[3] A. Cohen, *Wavelets and Multiscale Signal Processing*, Chapman and Hall, New York, 1995.

[4] A. Cohen and I. Daubechies, *A stability criterion for biorthogonal wavelet bases and their related subband coding scheme*, Duke Math. J., 68 (1992), pp. 313–335.

[5] A. Cohen and I. Daubechies, *A new technique to estimate the regularity of refinable functions*, Rev. Mat. Iberoamericana, 12 (1996), pp. 527–591.

[6] A. Cohen, I. Daubechies, and J. C. Feauveau, *Biorthogonal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 45 (1992), pp. 485–560.

[7] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 61, SIAM, Philadelphia, 1992.

[8] I. Daubechies and J. C. Lagarias, *Two-scale difference equations* I. *Existence and global regularity of solutions*, SIAM J. Math. Anal., 22 (1991), pp. 1388–1410.

[9] I. Daubechies and Y. Huang, *A decay theorem for refinable functions*, Appl. Math. Lett., 7 (1994), pp. 1–4.

[10] N. Dyn and D. Levin, *Subdivision schemes in geometric modeling*, Acta Numer., 11 (2002), pp. 73–144.

[11] A.-H. Fan and Q. Y. Sun, *Regularity of Butterworth refinable functions*, Asia J. Math., 5 (2001), pp. 433–440.

[12] B. Han, *Analysis and construction of optimal multivariate biorthogonal wavelets with compact support*, SIAM J. Math. Anal., 31 (1999), pp. 274–304.

[13] B. Han, *Vector cascade algorithms and refinable function vectors in Sobolev spaces*, J. Approx. Theory, 124 (2003), pp. 44–88.

[14] B. Han, *On a conjecture about MRA Riesz wavelet bases*, Proc. Amer. Math. Soc., 134 (2006), pp. 1973–1983.

[15] B. Han and R.-Q. Jia, *Multivariate refinement equations and convergence of subdivision schemes*, SIAM J. Math. Anal., 29 (1998), pp. 1177–1199.

[16] B. Han and R. Q. Jia, *Characterization of Riesz bases of wavelets generated from multiresolution analysis*, Appl. Comput. Harmon. Anal., 23 (2007), pp. 321–345.

[17] B. Han and Z. Shen, *Wavelets with short support*, SIAM J. Math. Anal., 38 (2006), pp. 530–556.

[18] C. Herley and M. Vetterli, *Wavelets and recursive filter banks*, IEEE Trans. Signal Process., 41 (1993), pp. 2536–2556.

[19] R. Q. Jia, *Subdivision schemes in L_p spaces*, Adv. Comput. Math., 3 (1995), pp. 309–341.

[20] R. Q. Jia and C. A. Micchelli, *Using the refinement equation for the construction of pre-wavelets* II: *Power of two*, in Curves and Surfaces, P. J. Laurent, A. Le Méhauté, and L. L. Schumaker, eds., Academic Press, New York, 1991, pp. 209–246.

[21] R. Q. Jia, J. Z. Wang, and D. X. Zhou, *Compactly supported wavelet bases for Sobolev spaces*, Appl. Comput. Harmon. Anal., 15 (2003), pp. 224–241.

[22] W. Lawton, S. L. Lee, and Z. W. Shen, *Convergence of multidimensional cascade algorithms*, Numer. Math., 78 (1998), pp. 427–438.

[23] P. G. Lemarié-Rieusset, *Fonctions d'échelle interpolantes, polynômes de Bernstein et ondelettes non stationnaires*, Rev. Mat. Iberoamericana, 13 (1997), pp. 91–188.

[24] R. Lorentz and P. Oswald, *Criteria for hierarchical bases in Sobolev spaces*, Appl. Comput. Harmon. Anal., 8 (2000), pp. 32–85.

[25] A. Oppenheim and R. Schafer, *Digital Signal Processing*, Prentice–Hall, New York, 1975.

[26] Q. Sun, *Convergence of cascade algorithms and smoothness of refinable distributions*, Chinese Ann. Math. Ser. B, 24 (2003), pp. 367–386.

[27] M. Unser and T. Blu, *Fractional splines and wavelets*, SIAM Rev., 42 (2000), pp. 43–67.

[28] D. X. Zhou, *Norms concerning subdivision sequences and their applications in wavelets*, Appl. Comput. Harmon. Anal., 11 (2001), pp. 329–346.

SIAM J. MATH. ANAL.
Vol. 40, No. 1, pp. 103–133

PROPAGATION THROUGH GENERIC LEVEL CROSSINGS: A SURFACE HOPPING SEMIGROUP*

CLOTILDE FERMANIAN KAMMERER† AND CAROLINE LASSER‡

Abstract. We construct a surface hopping semigroup, which asymptotically describes nuclear propagation through crossings of electron energy levels. The underlying time-dependent Schrödinger equation has a matrix-valued potential, whose eigenvalue surfaces have a generic intersection of codimension two, three, or five in Hagedorn's classification. Using microlocal normal forms reminiscent of the Landau–Zener problem, we prove convergence to the true solution with an error of the order $\varepsilon^{1/8}$, where ε is the semiclassical parameter. We present numerical experiments for an algorithmic realization of the semigroup illustrating the convergence of the algorithm.

Key words. time-dependent Schrödinger system, eigenvalue crossing, microlocal normal form, surface hopping

AMS subject classifications. 35Q40, 41A60, 81V55

DOI. 10.1137/070686810

1. Introduction.

In the framework of time-dependent Born–Oppenheimer approximation, the dynamics of molecules can approximately be reduced to matrix-valued Schrödinger equations on the nucleonic configuration space,

$$\begin{cases} i\varepsilon\partial_t\psi^\varepsilon(q,t) = \left(-\frac{\varepsilon^2}{2}\Delta_q + V(q)\right)\psi^\varepsilon(q,t), \quad (q,t)\in\mathbb{R}^d\times\mathbb{R}, \\ \psi^\varepsilon(q,0) = \psi_0^\varepsilon(q); \end{cases} \tag{1.1}$$

see, for example, [12, 19]. The linear Schrödinger equation (1.1) has a unique global solution $\psi^\varepsilon \in C(\mathbb{R}, L^2(\mathbb{R}^d,\mathbb{C}^N))$ for all square-integrable initial data ψ_0^ε. The parameter $\varepsilon > 0$ is small and causes a highly oscillatory behavior of the solution in space and time. It can be thought of as the square root of the ratio of electronic mass and the average mass of the nuclei. Moreover, the solution itself does not have any direct physical interpretation. It is the position density $|\psi^\varepsilon(q,t)|^2$ which gives the probability of finding the nuclei in the configuration $q\in\mathbb{R}^d$ at time t. We are interested in an asymptotic description for the time evolution of quadratic quantities like the position density with the following properties. First, it shall be effective in the sense that it unfolds characteristic dynamical properties. Second, it shall be explicit enough, such that it allows an algorithmic realization. Third, the resulting algorithm shall be applicable on high-dimensional nucleonic configuration spaces $\mathbb{R}^d$, $d\gg 1$.

Hagedorn rigorously derived and classified Schrödinger systems for molecular propagation through electron energy level crossings of minimal multiplicity [13]. He obtained potentials of the form

$$V(q) = v(q)\,\mathrm{Id} + V_\ell\left(\phi(q)\right), \quad \ell\in\{2,3,3',5\},$$

*Received by the editors March 30, 2007; accepted for publication (in revised form) December 3, 2007; published electronically April 2, 2008.
http://www.siam.org/journals/sima/40-1/68681.html
†Université de Paris 12 Val de Marne, Laboratoire d' analyse et de mathématiques appliquées, 61 avenue du général de Gaulle, 94 010 Créteil cedex, France (clotilde.fermanian@univ-paris12.fr).
‡Freie Universität Berlin, Fachbereich Mathematik und Informatik, Arnimallee 6, 14195 Berlin, Germany (lasser@math.fu-berlin.de).

where $v(q) \in \mathcal{C}^\infty(\mathbb{R}^d, \mathbb{R})$ is a smooth real-valued function, Id is the identity matrix in $\mathbb{C}^{2\times 2}$ or $\mathbb{C}^{4\times 4}$, and $q \mapsto \phi(q)$ is a smooth vector-valued function with $\phi(q) \in \mathbb{R}^2$, $\mathbb{R}^3$, or $\mathbb{R}^5$. The matrices V_ℓ are given by

$$V_2(\phi) = \begin{pmatrix} \phi_1 & \phi_2 \\ \phi_2 & -\phi_1 \end{pmatrix}, \qquad V_3(\phi) = \begin{pmatrix} \phi_1 & \phi_2 + i\phi_3 \\ \phi_2 - i\phi_3 & -\phi_1 \end{pmatrix}, \tag{1.2}$$

$$V_{3'}(\phi) = \begin{pmatrix} \begin{pmatrix} \phi_1 & \phi_2 + i\phi_3 \\ \phi_2 - i\phi_3 & -\phi_1 \end{pmatrix} & \mathbf{0} \\ \mathbf{0} & \begin{pmatrix} \phi_1 & \phi_2 - i\phi_3 \\ \phi_2 + i\phi_3 & -\phi_1 \end{pmatrix} \end{pmatrix}, \tag{1.3}$$

$$V_5(\phi) = \begin{pmatrix} \phi_0 \ \mathrm{Id} & \begin{pmatrix} \phi_1 + i\phi_2 & \phi_3 + i\phi_4 \\ -\phi_3 + i\phi_4 & \phi_1 - i\phi_2 \end{pmatrix} \\ \begin{pmatrix} \phi_1 - i\phi_2 & -\phi_3 - i\phi_4 \\ \phi_3 - i\phi_4 & \phi_1 + i\phi_2 \end{pmatrix} & -\phi_0 \ \mathrm{Id} \end{pmatrix}. \tag{1.4}$$

For these four matrices, the eigenvalues are $\pm|\phi|$. Therefore, the eigenvalues of $V(q)$ are $v(q) \pm |\phi(q)|$, and $q^* \in \mathbb{R}^d$ is a point of crossing eigenvalues if and only if $\phi(q^*) = 0$. We shall say that this crossing is generic if

$$\mathrm{d}\phi \text{ is of maximal rank on } \{\phi = 0\},$$

i.e., of rank 2 for $\ell = 2$, of rank 3 for $\ell = 3, 3'$, and of rank 5 for $\ell = 5$. This explains why these crossings are usually referred to as codimension two, three, and five crossings and enlightens the choice of the index ℓ we have made. Hagedorn's codimension one crossings are not considered here, since they violate the above rank condition and also show a different dynamical behavior than systems with crossings of higher codimension. We set

$$N(2) = N(3) = 2, \quad N(3') = N(5) = 4,$$

so that the potential $V(q)$, the wave function $\psi^\varepsilon(q,t)$, and the differential $\mathrm{d}\phi(q)$ belong to $\mathbb{C}^{N(\ell)\times N(\ell)}$, $\mathbb{C}^{N(\ell)}$, and $\mathbb{R}^{\ell\times d}$, respectively. For $\ell = 3'$ we set $\mathbb{R}^{3'} = \mathbb{R}^3$. The orthogonal eigenprojectors

$$\Pi^\pm(q) = \tfrac{1}{2}\left(\mathrm{Id} \pm |\phi(q)|^{-1} V_\ell(\phi(q))\right)$$

of the matrix $V(q)$ have a conical singularity at points of crossing eigenvalues q^*; that is, $\nabla\Pi^\pm(q) = O(|q - q^*|^{-1})$ as $q \to q^*$. This motivates the notion of conical intersections, by which especially codimension two crossings are frequently referred to.

Eigenvalue crossings are ubiquitous in the quantum mechanical description of polyatomic molecules, that is, molecules with more than two nuclei. The collection [4] provides an exposition of this active area of research in theoretical chemistry. As for a prominent example of an ultrafast isomerization on the femtosecond time scale, a codimension two crossing of energy levels explains the effectiveness of the first step of vision, the cis-trans isomerization of retinal in rhodopsin; see also [14] and section 3 below for related numerical experiments.

The analysis of scalar Schrödinger equations teaches us that the direct study of the time evolution of quadratic quantities like the position density $|\psi^\varepsilon(q,t)|^2$ is

impossible. The oscillations of $\psi^\varepsilon(q,t)$ have to be taken into account, and one has to work in the space of positions and momenta (q,p), the phase space $\mathbb{R}^d_q \times \mathbb{R}^d_p$. Therefore, one studies the Wigner function of $\psi^\varepsilon(q,t)$ in a suitable ε-dependent scaling, which resolves the highly oscillatory features of the solution,

$$W^\varepsilon(\psi^\varepsilon(t))(q,p) = (2\pi)^{-d} \int_{\mathbb{R}^d} \psi^\varepsilon\big(q - \tfrac{\varepsilon}{2}v, t\big) \otimes \overline{\psi}^\varepsilon\big(q + \tfrac{\varepsilon}{2}v, t\big)\ \mathrm{e}^{i\,v\cdot p}\,\mathrm{d}v.$$

It plays the role of a generalized probability density on phase space. For square-integrable wave functions ψ, the Wigner function $W^\varepsilon(\psi)$ is a square-integrable function on phase space with values in the space of hermitian matrices. One recovers the position density by

$$|\psi(q)|^2 = \operatorname{tr} \int_{\mathbb{R}^d} W^\varepsilon(\psi)(q,p)\,\mathrm{d}p.$$

Besides, the action of the Wigner function against compactly supported smooth test functions $a \in \mathcal{C}^\infty_c(\mathbb{R}^{2d}, \mathbb{C}^{N(\ell)\times N(\ell)})$ is simply expressed in terms of the semiclassical pseudodifferential operator with symbol a, which is defined by

$$\mathrm{op}_\varepsilon(a)\psi(q) = (2\pi\varepsilon)^{-d} \int_{\mathbb{R}^{2d}} a\left(\tfrac{1}{2}(q+v), p\right) \mathrm{e}^{\frac{i}{\varepsilon}p\cdot(q-v)}\psi(v)\,\mathrm{d}v\,\mathrm{d}p$$

for $\psi \in L^2(\mathbb{R}^d, \mathbb{C}^{N(\ell)})$. Indeed, we have

$$\operatorname{tr} \int_{\mathbb{R}^{2d}} W^\varepsilon(\psi)(q,p)a(q,p)\,\mathrm{d}q\,\mathrm{d}p = \left(\mathrm{op}_\varepsilon(a)\psi\,,\,\psi\right)_{L^2(\mathbb{R}^d,\mathbb{C}^{N(\ell)})}.$$

It is our aim to construct an asymptotic semigroup, which approximately propagates the initial data's Wigner function for all generic level crossings of Hagedorn's classification. Our approximation relies on a microlocal normal form for operators with eigenvalue crossings, which has been derived in [1, 2, 6]. Roughly speaking, near a crossing point the Schrödinger operator

$$-i\varepsilon\partial_t - \tfrac{\varepsilon^2}{2}\Delta_q + V(q)$$

is equivalent to the normal form

$$-i\varepsilon\partial_t + V_\ell\Big(t, \mathrm{op}_\varepsilon(|\mathrm{d}\phi(q)p|^{-\frac{1}{2}}\pi_\ell(q,p)\phi(q))\Big),$$

where $\pi_\ell(q,p)$ denotes the orthogonal projection onto the hyperplane normal to the vector $\mathrm{d}\phi(q)p \in \mathbb{R}^\ell$. In the case $\ell = 2$, this resembles the Landau–Zener system

$$i\varepsilon\tfrac{\mathrm{d}}{\mathrm{d}t}\psi(t) = \begin{pmatrix} t & \gamma \\ \gamma & -t \end{pmatrix}\psi(t), \qquad \gamma > 0,$$

for which the probability that a solution starting at time $t = -\infty$ in the one eigenspace will have passed over to the other eigenspace at time $t = +\infty$, which has explicitly been computed by Landau [16] and Zener [21] in the 1930s. This famous Landau–Zener transition rate reads as

$$\exp\Big(-\frac{\pi}{\varepsilon}\gamma^2\Big),$$

and in [8] it is proven that the rate still gives the correct asymptotics if γ is replaced by a bounded operator. Our semigroup combines effective transitions between eigenspaces close to points of crossing eigenvalues on the one hand with classical transport in the adiabatic regime on the other hand. More precisely, the V-diagonal components $\Pi^{\pm}W^{\varepsilon}(\psi_0^{\varepsilon})\Pi^{\pm}$ of the initial Wigner function are transported along the Hamiltonian curves of the eigenvalues of the Schrödinger operator's symbol

$$\tfrac{1}{2}|p|^2 + v(q) \pm |\phi(q)|.$$

Whenever one of the trajectories $(q^{\pm}(t), p^{\pm}(t))$ attains a local minimal gap between the eigenvalues, there is an effective nonadiabatic transfer of weight according to the ε-dependent transition rate

$$\exp\left(-\frac{\pi}{\varepsilon}\frac{|\pi_\ell(q,p)\phi(q)|^2}{|\mathrm{d}\phi(q)p|}\right).$$

Since the rate is negligibly small, when the eigenvalue gap is larger than $\sqrt{\varepsilon}$, the nonadiabatic transfer of weight is effectively performed at times t^* with

$$t \mapsto |\phi(q^{\pm}(t))| \text{ has a local minimum in } t = t^* \text{ and } |\phi(q^{\pm}(t^*))| \leq R\sqrt{\varepsilon}$$

for some fixed $R > 0$. Our main result is that this dynamical description yields approximate solutions with an error of order $\varepsilon^{1/8}$ when choosing $R = \varepsilon^{-1/8}$. Moreover, it is explicit enough for an algorithmic realization, whose performance on a model for retinal in rhodopsin is studied here as well. The algorithm is a mathematical counterpart to the popular surface hopping algorithms of chemical physics introduced by Tully and Preston in [20].

Quantum dynamical descriptions in terms of classical transport as described above, that is, in the spirit of an Egorov theorem, are well established and have been given for Wigner functions, for example, in [10, 11]: for Schrödinger systems they hold to leading order in ε, until classical trajectories come close to a point of crossing eigenvalues. Then, as already mentioned, the adiabatic approximation is no longer valid, and there are leading order nonadiabatic transitions between the levels (the energy propagated on one level to the crossing may pass partially or completely on the other level). This phenomenon has been precisely analyzed in the case of Gaussian wave packet propagation by Hagedorn [13] for all generic electron level crossings. For initial data, which are less specific than Gaussian wave packets, the evolution of appropriate two-scale Wigner measures has been studied. These measures are weak limits of the Wigner function and incorporate information on concentration effects close to trajectories, which touch points of crossing eigenvalues, with respect to the second scale $\sqrt{\varepsilon}$. These Wigner measures have been analyzed for a linear codimension two crossing in [7], for general two-level systems in [8], and for all of Hagedorn's models in [5]. In [18], the results of [7] have been lifted to a leading order approximation of the Wigner function. Here, we aim at approximating the Wigner function for all generic crossings, while additionally proving a convergence rate.

We will proceed as follows. Section 2 constructs the surface hopping semigroup, states the main result, that is, the validity of our approximation with an error of order $\varepsilon^{1/8}$, and discusses the strategy of the proof. In section 3, numerical results are presented for an algorithmic realization of the semigroup applied to a retinal model. In section 4, the proof for propagation away from the crossing is carried out, while the microlocal normal form yields the correct nonadiabatic transition rates, as proven

in section 5. In section 6, the main result is extended to observables, which are more pertinent for the crossings with degenerate eigenvalues ($\ell = 3', 5$). Finally, the appendix summarizes basic facts of Weyl calculus.

2. Main result. Propagation through level crossings can be approximated by a proper combination of classical transport and nonadiabatic transitions. For this, we study the underlying classical flows and combine them with effective nonadiabatic transitions to an asymptotic semigroup.

2.1. Transport and transitions. We consider the classical flows

$$\Phi_\pm^{-t} : \mathbb{R}^{2d} \to \mathbb{R}^{2d}, \quad \Phi_\pm^{-t}(q_0, p_0) = \big(q^\pm(t), p^\pm(t)\big)$$

associated with the Hamiltonian curves of $\frac{1}{2}|p|^2 + v(q) \pm |\phi(q)|$. These curves are solutions to the Hamiltonian systems

$$\begin{cases} \dot q^\pm(t) = p^\pm(t), \;\; \dot p^\pm(t) = -\nabla v\,(q^\pm(t)) \mp {}^t\mathrm{d}\phi\,(q^\pm(t))\,\frac{\phi\left(q^\pm(t)\right)}{|\phi(q^\pm(t))|}, \\ q^\pm(0) = q_0, \;\; p^\pm(0) = p_0. \end{cases}$$

We consider only initial phase space points $(q_0, p_0) \in \mathbb{R}^{2d}$ such that for $t > 0$

$$\phi(q^\pm(t)) = 0 \;\Rightarrow\; \mathrm{d}\phi(q^\pm(t))p^\pm(t) \neq 0. \tag{2.1}$$

This condition guarantees that classical trajectories arrive transversally at the crossing set and have a unique smooth continuation through this singularity; see Proposition 1 in [5].

For a large class of test functions and under suitable restrictions on the time interval, the classical flows are enough for approximating the dynamics up to an error of order ε. Indeed, one considers observables $a \in \mathcal{C}_c^\infty(\mathbb{R}^{2d}, \mathbb{C}^{N(\ell)\times N(\ell)})$ such that

$$a = a^+\Pi^+ + a^-\Pi^-, \qquad a^\pm \in \mathcal{C}_c^\infty(\mathbb{R}^{2d} \setminus \{\phi = 0\}, \mathbb{C}). \tag{2.2}$$

For $\ell = 2, 3$, the eigenspaces are one-dimensional, and these observables focus on the V-diagonal elements of the Wigner matrix, where V-diagonal means diagonal with respect to the decomposition of $\mathbb{C}^{N(\ell)}$ by the eigenprojectors $\Pi^+(q)$ and $\Pi^-(q)$. For $\ell = 3', 5$, however, the eigenspaces are two-dimensional, and observables of the form (2.2) are not enough to completely resolve all dynamical features within the eigenspaces. We will address this issue in section 6.

For all times $t \in [0, T]$, such that the classical trajectories $\Phi_\pm^t$ arriving on the support of a have not passed the crossing set $\{\phi = 0\}$, the action of the Wigner function on $a = a^\pm\Pi^\pm$ obeys

$$\begin{aligned} \operatorname{tr} \int_{\mathbb{R}^{2d}} W^\varepsilon(\psi^\varepsilon(t))\,(q,p)\,\Pi^\pm(q) a^\pm(q,p)\,\mathrm{d}q\,\mathrm{d}p \\ - \operatorname{tr} \int_{\mathbb{R}^{2d}} \big(\Pi^\pm W^\varepsilon(\psi_0^\varepsilon)\Pi^\pm \circ \Phi_\pm^{-t}\big)(q,p)\,a^\pm(q,p)\,\mathrm{d}q\,\mathrm{d}p = O(\varepsilon) \end{aligned}$$

as $\varepsilon \to 0$. Such Egorov-type descriptions hold, until classical trajectories come close to a crossing point $(q, p) \in \{\phi = 0\}$ and leading order nonadiabatic transitions occur. These transitions depend on how the solution $\psi^\varepsilon(t)$ concentrates on the ingoing trajectories with respect to the scale $\sqrt{\varepsilon}$. For the linear codimension two crossing with $\phi(q) = q$, $q \in \mathbb{R}^2$, the two-scale Wigner measure's description of [7] is lifted

to an approximation of the Wigner function in [18]. This linear model has specific features; see also [9]. In particular, all classical trajectories which meet the crossing are included in the set $\{q \wedge p = 0\}$, where $q \wedge p := q_2 p_1 - q_1 p_2$ for $q, p \in \mathbb{R}^2$. The idea of [18] is to propagate the V-diagonal parts of the initial Wigner function along the classical trajectories and to apply the ε-dependent transition coefficient

$$T^{\varepsilon}_{lin}(q^*, p^*) = \exp\left(-\frac{\pi}{\varepsilon}\frac{|q^* \wedge p^*|^2}{|p^*|^3}\right),$$

as soon as the trajectories reach their minimal distance from the crossing set, which is easy to check since $q \cdot p = 0$ at such a point. Theorem 3.2 in [18] proves that under suitable conditions on the initial data this ε-dependent propagation is correct in the limit $\varepsilon \to 0$. We construct here an extension, which covers the general situation described above, and give a convergence proof including a convergence rate. Our approach draws from the understanding of the nonadiabatic mechanism as developed in [5].

2.2. A surface hopping semigroup. Let $R > 0$. In the general case, the crucial points in phase space are those where the classical trajectories attain a *local minimal gap between the two eigenvalues.* These points fulfill the condition

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\left|\phi\left(q^{\pm}(t)\right)\right|^2\right) = 2\,\mathrm{d}\phi\left(q^{\pm}(t)\right)p^{\pm}(t)\,\cdot\,\phi\left(q^{\pm}(t)\right) = 0,$$

and one performs an effective nonadiabatic transfer of weight, whenever a trajectory passes the set

$$S_{\varepsilon,R} = \left\{(q,p) \in \mathbb{R}^{2d} \mid |\phi(q)| \leq R\sqrt{\varepsilon},\, \mathrm{d}\phi(q)p \cdot \phi(q) = 0\right\}.$$

The microlocal normal form, which will be given later in Theorem 5.2, suggests the transition rate

$$T^{\varepsilon}(q^*, p^*) = \exp\left(-\frac{\pi}{\varepsilon}\frac{|\pi_\ell(q^*, p^*)\phi(q^*)|^2}{|\mathrm{d}\phi(q^*)p^*|}\right),$$

where $\pi_\ell(q^*, p^*)$ is the orthogonal projection from the Euclidean space $\mathbb{R}^\ell$ into the hyperplane normal to the vector $\mathrm{d}\phi(q^*)p^* \in \mathbb{R}^\ell$. Since $\mathrm{d}\phi(q^{\pm}(t))p^{\pm}(t)$ does not vanish when the considered trajectories arrive at the crossing set $\{\phi = 0\}$, it is also nonzero when arriving at the jump manifold $S_{\varepsilon,R}$ if $R\sqrt{\varepsilon}$ is small enough. Besides, for $\ell = 2$, one has

$$|\,\pi_\ell(q,p)\phi(q)\,| = |\,\phi(q) \wedge \tfrac{\mathrm{d}\phi(q)p}{|\mathrm{d}\phi(q)p|}\,|,$$

and we recover the transition coefficient $T^{\varepsilon}_{lin}(q^*, p^*)$ for $\phi(q) = q$, $q \in \mathbb{R}^2$.

We attach the labels -1 and $+1$ to phase space. For points $(q, p, j) \in \mathbb{R}^{2d}_{\pm} := \mathbb{R}^{2d} \times \{-1, +1\}$, we consider trajectories

$$\mathcal{T}^{(q,p,j)}_{\varepsilon,R} : [0, +\infty) \to \mathbb{R}^{2d}_{\pm},$$

which combine deterministic classical transport and random jumps between the levels at the manifold $S_{\varepsilon,R}$. More precisely, we set $\mathcal{T}^{(q,p,j)}_{\varepsilon,R}(t) = \left(\Phi^t_j(q,p), j\right)$ as long as $\Phi^t_j(q,p) \notin S_{\varepsilon,R}$. Whenever the deterministic flow $\Phi^t_j(q,p)$ hits the manifold $S_{\varepsilon,R}$ at

a point (q^*, p^*), a random jump from j to $-j$ occurs with probability $T^\varepsilon(q^*, p^*)$. For points (q, p, j) generating classical trajectories, which either violate the non-degeneracy condition (2.1) or do not leave the set $S_{\varepsilon,R}$, there is either no transport or no jump at all. Since the trajectories which touch the crossing set arrive there transversally, each path

$$(q, p, j) \to \mathcal{T}_{\varepsilon,R}^{(q,p,j)}(t)$$

has a finite number of jumps and remains in a bounded region of $\mathbb{R}^{2d}_{\pm}$ within a bounded time interval $[0, T]$. Away from the jump manifold $S_{\varepsilon,R} \times \{-1, +1\}$ each path is smooth. Hence, the random trajectories define a time-dependent Markov process with state space $\mathbb{R}^{2d}_{\pm}$. The associated transition function $P_{\varepsilon,R}(p, q, j; t, \Gamma)$ describes the probability of being at time t in the measurable set $\Gamma \subset \mathbb{R}^{2d}_{\pm}$ having started in (q, p, j). Its action on bounded measurable scalar functions $f : \mathbb{R}^{2d}_{\pm} \to \mathbb{C}$ defines a semigroup $(\mathcal{L}^t_{\varepsilon,R})_{t\geq 0}$ by

$$(\mathcal{L}^t_{\varepsilon,R} f)(q, p, j) := \int_{\mathbb{R}^{2d}\times\{-1,+1\}} f(x, \xi, k)\, P_{\varepsilon,R}(q, p, j; t, \mathrm{d}(x, \xi, k)).$$

For introducing the semigroup's action on Wigner functions, we use the following space of continuous V-diagonal test functions satisfying T^ε-dependent boundary conditions at the jump manifold.

DEFINITION 2.1. *A continuous function $a \in \mathcal{C}_c(\mathbb{R}^{2d} \setminus S_{\varepsilon,R}, \mathbb{C}^{N(\ell)\times N(\ell)})$ belongs to the space $\mathcal{C}_{\varepsilon,R}$ if it has the following properties:*

i. *$a = a^+\Pi^+ + a^-\Pi^-$ with $a^\pm \in \mathcal{C}_c(\mathbb{R}^{2d} \setminus S_{\varepsilon,R}, \mathbb{C})$.*

ii. *The function $f_a : (\mathbb{R}^{2d} \setminus S_{\varepsilon,R}) \times \{-1, +1\} \to \mathbb{C}$,*

$$f_a(q, p, +) = a^+(q, p), \quad f_a(q, p, -) = a^-(q, p),$$

satisfies for all $(q, p, j) \in S_{\varepsilon,R} \times \{-1, +1\}$

$$\begin{aligned}
&\lim_{\delta\to-0} f_a(q + \delta p, p + \delta(-\nabla v(q) - j\ {}^t\mathrm{d}\phi(q)\phi(q)/|\phi(q)|), j)\\
&= \lim_{\delta\to+0} (T^\varepsilon f_a)(q + \delta p, p + \delta(-\nabla v(q) + j\ {}^t\mathrm{d}\phi(q)\phi(q)/|\phi(q)|), -j)\\
&= \lim_{\delta\to+0} ((1 - T^\varepsilon) f_a)(q + \delta p, p + \delta(-\nabla v(q) - j\ {}^t\mathrm{d}\phi(q)\phi(q)/|\phi(q)|), j).
\end{aligned}$$

For test functions $a \in \mathcal{C}_{\varepsilon,R}$, the action of $(\mathcal{L}^t_{\varepsilon,R})_{t\geq 0}$ is naturally given by

$$(\mathcal{L}^t_{\varepsilon,R}\, a)(q, p) := \left(\mathcal{L}^t_{\varepsilon,R} f_a\right)(q, p, +1)\, \Pi^+(q) + \left(\mathcal{L}^t_{\varepsilon,R} f_a\right)(q, p, -1)\, \Pi^-(q).$$

By construction, the semigroup leaves $\mathcal{C}_{\varepsilon,R}$ invariant, and duality allows us to define its action on Wigner functions. More precisely, let $W^\varepsilon(\psi)$ be the Wigner function of some wave function $\psi \in L^2(\mathbb{R}^d, \mathbb{C}^{N(\ell)})$. Then, $\mathcal{L}^t_{\varepsilon,R} W^\varepsilon(\psi)$ acts on $a \in \mathcal{C}_{\varepsilon,R}$ by

$$\left(\mathcal{L}^t_{\varepsilon,R} W^\varepsilon(\psi), a\right) = \mathrm{tr} \int_{\mathbb{R}^{2d}} W^\varepsilon(\psi)(q, p)(\mathcal{L}^t_{\varepsilon,R}\, a)(q, p)\, \mathrm{d}q\, \mathrm{d}p,$$

defining a locally integrable function on phase space. We finally choose an ε-dependent hopping range $R(\varepsilon) = \varepsilon^{-1/8}$ and set

$$(\mathcal{L}^t_\varepsilon)_{t\geq 0} := (\mathcal{L}^t_{\varepsilon,R(\varepsilon)})_{t\geq 0}.$$

2.3. Assumptions and main result. Let $\psi^\varepsilon(t)$ be the solution of the Schrödinger equation (1.1) with initial datum ψ_0^ε. We now state the precise assumptions, under which the action of the semigroup $(\mathcal{L}_\varepsilon^t)_{t\geq 0}$ on the initial Wigner function $W^\varepsilon(\psi_0^\varepsilon)$ approximates the V-diagonal components of $W^\varepsilon(\psi^\varepsilon(t))$.

(A0) $V \in \mathcal{C}^\infty(\mathbb{R}^d, \mathbb{C}^{N(\ell)\times N(\ell)})$ is of subquadratic growth and of the form

$$V(q) = v(q)\,\mathrm{Id} + V_\ell\left(\phi(q)\right), \quad \ell \in \{2, 3, 3', 5\},$$

where the matrices $V_\ell(\phi(q))$ have been defined in (1.2), (1.3), and (1.4). We assume the eigenvalue crossings to be generic in the sense that $\mathrm{d}\phi$ is of maximal rank on the crossing set $\{\phi = 0\}$.

(A1) $(\psi_0^\varepsilon)_{\varepsilon>0}$ is a bounded family in $L^2\big(\mathbb{R}^d, \mathbb{C}^{N(\ell)}\big)$ associated with $\mathrm{Ran}\Pi^+$,

$$\|\Pi^-\psi_0^\varepsilon\|_{L^2(\mathbb{R}^d,\mathbb{C}^{N(\ell)})} = O(\varepsilon^{\beta_1}), \quad \beta_1 \geq 1/8.$$

We suppose that the initial data are localized away from the crossing $\{\phi = 0\}$; i.e., for all $b \in \mathcal{C}_c^\infty(\mathbb{R}^{2d}, \mathbb{C}^{N(\ell)\times N(\ell)})$ with $\mathrm{supp}(b) \subset \{|\phi| \leq R\sqrt{\varepsilon}\}$, $R = \varepsilon^{-1/8}$,

$$\int_{\mathbb{R}^{2d}} W^\varepsilon(\psi_0^\varepsilon)(q,p)\, b(q,p)\,\mathrm{d}q\,\mathrm{d}p = O(\varepsilon^{\beta_2}), \quad \beta_2 \geq 1/8.$$

We also assume localization away from the set

$$\left\{(q_0,p_0) \in \mathbb{R}^{2d} \mid \exists t > 0 : \phi(q^\pm(t)) = 0,\ \mathrm{d}\phi(q^\pm(t))p^\pm(t) = 0\right\},$$

which contains the points issuing classical trajectories, which arrive at the crossing without a unique continuation through it.

(A2) The test function $a \in \mathcal{C}_c^\infty\left(\mathbb{R}^{2d}, \mathbb{C}^{N(\ell)\times N(\ell)}\right)$ has its support at a distance larger than $R\sqrt{\varepsilon}$ with $R = \varepsilon^{-1/8}$ from the crossing; that is,

$$\mathrm{supp}(a) \cap \left\{(q,p) \in \mathbb{R}^{2d} \mid |\phi(q)| \leq R\sqrt{\varepsilon}\right\} = \emptyset, \qquad R = \varepsilon^{-1/8},$$

and

$$a(q,p) = a^+(q,p)\Pi^+(q) + a^-(q,p)\Pi^-(q), \qquad (q,p) \in \mathbb{R}^{2d},$$

with scalar-valued $a^\pm \in \mathcal{C}_c^\infty(\mathbb{R}^{2d}, \mathbb{C})$.

(A3) Within the time interval $[0,T]$, each of the plus-trajectories arriving at the support of a^+ at time T has performed at most one jump, generating minus-trajectories arriving at the support of a^-, which have not jumped at all.

Alternatively, assumptions (A1) and (A3) could also require that the initial data are associated with $\mathrm{Ran}\Pi^-$ and that each of the minus-trajectories arriving at the support of a^- at time T has performed at most one jump, generating plus-trajectories arriving at the support of a^+, which have not jumped at all.

THEOREM 2.2. *Let the potential V, the initial data $(\psi_0^\varepsilon)_{\varepsilon>0}$, the observable a, and the time interval $[0,T]$ fulfill assumptions* (A0), (A1), (A2), *and* (A3). *Let $\chi \in \mathcal{C}_c^\infty([0,T],\mathbb{R})$. Then, there exist positive constants $C, \varepsilon_0 > 0$ such that for all $0 < \varepsilon < \varepsilon_0$ the solution $\psi^\varepsilon(t)$ of the Schrödinger equation* (1.1) *satisfies*

$$\left|\mathrm{tr}\int_{\mathbb{R}^{2d+1}} \chi(t)\left(W^\varepsilon(\psi^\varepsilon(t)) - \mathcal{L}_\varepsilon^t W^\varepsilon(\psi_0)\right)(q,p)\, a(q,p)\,\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}t\right| \leq C\,\varepsilon^{1/8}. \tag{2.3}$$

Before entering the proof, we add some remarks. First, if one allows initial data in assumption (A1) with $\beta_1, \beta_2 > 0$, then the result holds with convergence

rate $\varepsilon^{\min(\beta_1,\beta_2,1/8)}$. Second, pointwise convergence holds on time intervals without nonadiabatic jumps, that is, when the solution has passed by the jump manifold; see also [18]. However, for pointwise convergence only the limit behavior without convergence rate can be deduced, since the constants C and ε_0 depend on the cut-off function χ and its derivatives in a way that possible oscillations in time are not controlled. Finally, in section 6 an extension of the approximation for the cases $\ell = 3', 5$ with degenerate eigenvalues is given. There, assumption (A2) is generalized to observables a which commute with V, that is, $a = \Pi^+ a \Pi^+ + \Pi^- a \Pi^-$.

2.4. Strategy of the proof. For notational convenience, we suppose $\beta_1 = \beta_2 = 1/2$. Otherwise, one has to add $O(\varepsilon^{1/8})$ in all the estimates. We work with the semigroup $(\mathcal{L}^t_{\varepsilon,R})_{t\geq 0}$ and prove convergence with an error of order

$$O(1/(R^5\sqrt{\varepsilon})) + O(R^3\sqrt{\varepsilon}) + O(1/R^2) + O(\sqrt{\varepsilon}\,|\ln\varepsilon|)$$

as $\varepsilon \to 0$ and $R \to +\infty$, which gives the claimed rate when choosing $R = \varepsilon^{-1/8}$. We distinguish between regions of large and small eigenvalue gap, that is, between sets $\{|\phi| > C\,R\sqrt{\varepsilon}\}$ with $C = \frac{1}{2}, 1$ on the one hand and $\{|\phi| \leq R\sqrt{\varepsilon}\}$ on the other hand. For a large gap we prove classical transport with an error of size $O(1/(R^5\sqrt{\varepsilon})) + O(1/R^2) + O(\sqrt{\varepsilon})$. Close to the crossing set, proving the relevance of nonadiabatic transitions, we use a microlocal normal form, which reduces the Schrödinger equation to a Landau–Zener-type problem with explicitly computable transition rates. There, we introduce an error of order $O(R^3\sqrt{\varepsilon}) + O(1/R^2) + O(\sqrt{\varepsilon}|\ln\varepsilon|) + O(1/(R^5\sqrt{\varepsilon}))$. The combination of both errors will then yield the final estimate of Theorem 2.2.

PROPOSITION 2.3. *Let $c \in \mathcal{C}^\infty_c(\mathbb{R}^{2d}, \mathbb{C})$, and let $b \in \mathcal{C}^\infty(\mathbb{R}^\ell, \mathbb{C})$ with ∇b compactly supported. If there exist $C > 0$ and $s_0 > 0$ such that*

$$\forall r \in [-s_0, s_0]: \quad \Phi^r_\pm(\mathrm{supp}(c)) \subset \{|\phi| > C\,R\sqrt{\varepsilon}\},$$

then for all $\chi \in \mathcal{C}^\infty_c(\mathbb{R}, \mathbb{R})$ and for all $s \in [t - s_0, t + s_0]$

$$\begin{aligned}
&\mathrm{tr} \int_{\mathbb{R}^{2d+1}} \chi(t)\, c(q,p)\, b\Big(\tfrac{\phi(q)}{R\sqrt{\varepsilon}}\Big)\, \Pi^\pm(q) W^\varepsilon(\psi^\varepsilon(t))(q,p)\, \mathrm{d}q\, \mathrm{d}p\, \mathrm{d}t \\
&= \mathrm{tr} \int_{\mathbb{R}^{2d+1}} \chi(t)\, c(q,p)\, b\Big(\tfrac{\phi(q)}{R\sqrt{\varepsilon}}\Big)\, \big(\Pi^\pm W^\varepsilon(\psi^\varepsilon(s))\, \Pi^\pm \circ \Phi^{-t+s}_\pm\big)(q,p)\, \mathrm{d}q\, \mathrm{d}p\, \mathrm{d}t \\
&\qquad + O\Big(\tfrac{1}{R^5\sqrt{\varepsilon}}\Big) + O\big(\tfrac{1}{R^2}\big) + O(\sqrt{\varepsilon}).
\end{aligned}$$

Proposition 2.3 will be proved in section 4. To use it for the main proof, we need to specify which points of $\mathrm{supp}(a^\pm)$ arrive close to the crossing. We denote the sets of trajectories arriving at (respectively, arising from) the crossing set $\{\phi = 0\}$ by

$$M^{\pm,in} = \big\{\Phi^t_\pm(q,p) \in \mathbb{R}^{2d} \mid \Phi^t_\pm(q,p) \notin \{\phi = 0\}, \exists t_0 < t : \Phi^{t_0}_\pm(q,p) \in \{\phi = 0\}\big\},$$

$$M^{\pm,out} = \big\{\Phi^t_\pm(q,p) \in \mathbb{R}^{2d} \mid \Phi^t_\pm(q,p) \notin \{\phi = 0\}, \exists t_0 > t : \Phi^{t_0}_\pm(q,p) \in \{\phi = 0\}\big\}.$$

The sets $M^{\pm,in/out}$ are smooth submanifolds of $\mathbb{R}^{2d}$. By construction of the semigroup, all phase space points generating backward trajectories passing through the zone of small gap $\{|\phi| \leq R\sqrt{\varepsilon}\}$ and performing a jump are contained in a neighborhood $\Omega^\pm$ of the intersection of the support of $a^\pm$ with $M^{\pm,out}$.

Some of the random trajectories reaching $\Omega^\pm$ touch the crossing set. We consider one of them, which arrives at time t_0 at (q_0, p_0) with $\phi(q_0) = 0$, $\mathrm{d}\phi(q_0)p_0 \neq 0$,

and choose the associated point (q_0, t_0, p_0, τ_0) in the phase space of space-time, where $\tau_0 = -\frac{1}{2}|p_0|^2 - v(q_0)$ is the energy coordinate. The normal form theorems [1, 2, 6] give neighborhoods of these points, on which the Schrödinger equation (1.1) microlocally reduces to a Landau–Zener-type problem

$$-i\varepsilon\partial_s v^\varepsilon = V_\ell(s, \tilde{z} + \gamma_\varepsilon(z,\zeta))v^\varepsilon + O(\varepsilon^\infty).$$

These model problems have explicitly computable transition rates in the scattering regime; see [7, 8]. The compact subset of $\{|\phi| < R\sqrt{\varepsilon}\}$, which is touched by the backward trajectories coming from $\Omega^\pm$, can be covered by finitely many of these neighborhoods projected to $\mathbb{R}^{2d}$, and without loss of generality we assume that one of them suffices.

Moreover, for each point in $\Omega^\pm$ being reached by a random trajectory at time t there are positive numbers $0 < t_{\mathrm{f}}^\pm < t_{\mathrm{i}}^\pm$, such that at time $t - t_{\mathrm{i}}^\pm$ and $t - t_{\mathrm{f}}^\pm$ the trajectories are contained in an annulus $\{C_1\sqrt{\varepsilon} < |\phi| < C_2\sqrt{\varepsilon}\}$ with $C_1, C_2 > 0$ and have performed their only jump within the interval $]t - t_{\mathrm{i}}^\pm, t - t_{\mathrm{f}}^\pm[$, whose length is denoted by $\delta_t^\pm = t_{\mathrm{i}}^\pm - t_{\mathrm{f}}^\pm$. These quantities are well defined, since the trajectories are transverse to the crossing set. Choosing $C_1 = \frac{R}{2}$ and $C_2 = R$, $\Omega^\pm$ can be covered by finitely many open sets, such that each of these sets can be associated with such points of time $t_{\mathrm{i}}^\pm$ and $t_{\mathrm{f}}^\pm$. Without loss of generality, we assume that one of them is enough. Then, we have by Proposition 2.3 with $C = \frac{1}{2}$

$$\begin{aligned}
&\operatorname{tr}\int_{\mathbb{R}^{2d+1}} \chi(t)\, a^\pm(q,p)\, \Pi^\pm(q) W^\varepsilon(\psi^\varepsilon(t))(q,p)\,\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}t \\
&= \operatorname{tr}\int_{\mathbb{R}^{2d+1}} \chi(t)\, a^\pm(q,p)\, \left(\Pi^\pm W^\varepsilon\big(\psi^\varepsilon(t - t_{\mathrm{f}}^\pm)\big)\, \Pi^\pm \circ \Phi_\pm^{-t_{\mathrm{f}}^\pm}\right)(q,p)\,\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}t \\
&\qquad\qquad + O(1/(R^5\sqrt{\varepsilon})) + O(1/R^2) + O(\sqrt{\varepsilon}).
\end{aligned}$$

Then, we perform a cut-off of the symbol $a^\pm \circ \Phi_\pm^{t_{\mathrm{f}}^\pm}$. We consider a smooth compactly supported function $\chi_0 \in \mathcal{C}_c^\infty(\mathbb{R}^\ell, \mathbb{R})$ such that $\chi_0(u) = 1$ on $\{|u| < 1\}$ and $\chi_0(u) = 0$ for $\{|u| > 2\}$. We write

$$\begin{aligned}
\left(a^\pm \circ \Phi_\pm^{t_{\mathrm{f}}^\pm}\right)(q,p) &= a_{BO}^\pm(q,p) + a_{LZ}^\pm(q,p), \\
a_{BO}^\pm(q,p) &= \left(a^\pm \circ \Phi_\pm^{t_{\mathrm{f}}^\pm}\right)(q,p)\left(1 - \chi_0\Big(\tfrac{\phi(q)}{R\sqrt{\varepsilon}}\Big)\right), \\
a_{LZ}^\pm(q,p) &= \left(a^\pm \circ \Phi_\pm^{t_{\mathrm{f}}^\pm}\right)(q,p)\,\chi_0\Big(\tfrac{\phi(q)}{R\sqrt{\varepsilon}}\Big).
\end{aligned}$$

Since the trajectories, which pass within the time interval $]t - t_{\mathrm{i}}^\pm, t - t_{\mathrm{f}}^\pm[$ through the support of $a_{BO}^\pm$, do not jump, Proposition 2.3 with $C = 1$ is enough to deal with the Born–Oppenheimer part. The analysis of the Landau–Zener part, however, involves nonadiabatic transitions. For points $(q,p) \in \operatorname{supp}(a_{LZ}^\pm)$ we have $|\phi(q)| \leq 2R\sqrt{\varepsilon}$. Hence, not all of the trajectories passing through the support of $a_{LZ}^\pm$ jump. Nevertheless, we argue as if all of them did. Indeed, the transition coefficients generated by these added jumps are exponentially small with respect to ε, since they occur for points (q,p) with $|\phi(q)| > R\sqrt{\varepsilon}$.

PROPOSITION 2.4. *Let $0 < t_{\mathrm{f}}^\pm < t_{\mathrm{i}}^\pm$ be such that at $t - t_{\mathrm{i}}^\pm$ and $t - t_{\mathrm{f}}^\pm$ all random trajectories arriving at time t in $\Omega^\pm$ are contained in $\left\{\frac{R}{2}\sqrt{\varepsilon} \leq |\phi| \leq R\sqrt{\varepsilon}\right\}$ and have*

performed their only jump within the interval $]t - t_{\rm i}^{\pm}, t - t_{\rm f}^{\pm}[$ *of length* $\delta_t^{\pm} = t_{\rm i}^{\pm} - t_{\rm f}^{\pm}$. *Then, for all* $\chi \in \mathcal{C}_c^{\infty}([0,T],\mathbb{R})$

$$\begin{aligned}
&\operatorname{tr} \int_{\mathbb{R}^{2d+1}} \chi(t)\, W^{\varepsilon}(\psi^{\varepsilon}(t))(q,p)\, a_{LZ}^{\pm}(q,p)\Pi^{\pm}(q)\,\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}t \\
&= \operatorname{tr} \int_{\mathbb{R}^{2d+1}} \chi(t)\, W^{\varepsilon}(\psi^{\varepsilon}(t-\delta_t^{\pm}))(q,p)(\mathcal{L}_{\varepsilon,R}^{\delta_t^{\pm}} a_{LZ}^{\pm})(q,p)\,\Pi^{+}(q)\,\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}t \\
&\qquad + O(1/R^2) + O(R^3\sqrt{\varepsilon}) + O(\sqrt{\varepsilon}|\ln\varepsilon|) + O(1/(R^5\sqrt{\varepsilon})). \qquad (2.4)
\end{aligned}$$

One observes that in the right-hand side of (2.4) only the plus-projector Π^+ appears. This comes from the fact that at time $t - t_{\rm i}^{\pm}$ the contribution on the minus-mode is of order $O(1/(R^5\sqrt{\varepsilon})) + O(1/R^2) + O(\sqrt{\varepsilon})$, which is due to Proposition 2.3 and the initial data being associated only with $\operatorname{Ran}\Pi^+$. Finally, again the classical transport result of Proposition 2.3 relates the right-hand side of (2.4) with the initial data, and the proof of our main result, Theorem 2.2, is complete.

3. Numerical experiments. Before giving the detailed proof, we present numerical experiments illustrating the theoretical convergence result and the effectiveness of the algorithm. We consider a two-level Schrödinger equation with codimension two crossing in two space dimensions, which models the photoisomerization of retinal in rhodopsin. This conformational change is considered as the first step of vision. In [14], computations with the model Hamiltonian

$$-\tfrac{\omega}{2}\partial_x^2 - \tfrac{1}{2m}\partial_\varphi^2 + \tfrac{1}{2}\omega x^2 + \begin{pmatrix} \frac{1}{2}W_0(1-\cos\varphi) & \lambda x \\ \lambda x & E_1 - \frac{1}{2}W_1(1-\cos\varphi) + \kappa x \end{pmatrix}$$

have qualitatively reproduced spectroscopic information of the molecule. The two effective coordinates $(\varphi, x) \in]-\frac{\pi}{2}, \frac{3\pi}{2}] \times \mathbb{R}$ consist of the reaction coordinate ϕ and a collective coordinate x. The parameters are $m^{-1} = 4.84 \cdot 10^{-4}$, $E_1 = 2.48$, $W_0 = 3.6$, $W_1 = 1.09$, $\omega = \lambda = 0.19$, and $\kappa = 0.1$ (all in eV, $\hbar = 1$); see note 18 in [14]. Setting

$$\varepsilon = m^{-1/2} = 0.022, \qquad q_1 = \varphi, \qquad q_2 = \frac{\varepsilon}{\sqrt{\omega}}\, x,$$

one obtains a rescaled Hamiltonian $-\frac{\varepsilon^2}{2}\Delta_q + V(q)$ with potential

$$V(q) = \tfrac{1}{2}(\beta q_2)^2 + \begin{pmatrix} \frac{1}{2}W_0(1-\cos q_1) & \alpha_1 q_2 \\ \alpha_1 q_2 & E_1 - \frac{1}{2}W_1(1-\cos q_2) + \alpha_2 q_2 \end{pmatrix},$$

whose parameters $(\alpha_1, \alpha_2) = \frac{\sqrt{\omega}}{\varepsilon}(\kappa, \lambda) \approx (2,\, 3.8)$ and $\beta = \omega/\varepsilon \approx 8.6$ are of order one with respect to ε. Fixing these values of (α_1, α_2) and β, we run a series of experiments for varying values of the semiclassical parameter ε and hopping ranges R,

$$\varepsilon \in \{0.0005, 0.001, 0.005, 0.01, 0.015, 0.02, 0.022, 0.03\}, \qquad R \in \{1, 2, 3\},$$

in the following set-up. We consider normalized Gaussian initial data associated with the plus-level, that is,

$$\psi_0^{\varepsilon}(q) = (\varepsilon\pi)^{-1/2} \exp\bigl(-\tfrac{1}{2\varepsilon}|q - q_0^{\varepsilon}|^2 + \tfrac{i}{\varepsilon}\, p_0 \cdot (q - q_0^{\varepsilon})\bigr)\, v^{+}(q),$$

TABLE 1

The table shows final level populations and particle numbers as well as the accuracy of the reference solver. The population of the upper level $\|\Pi^+\psi^\varepsilon(t)\|^2$ at the final time $T = 7\sqrt{\varepsilon}$ is computed by the reference solver, a pseudospectral Strang splitting scheme, and illustrates leading order nonadiabatic transitions for all test cases. Depending on the hopping range R, the final particle numbers of the surface hopping algorithm vary between 3000 and 30000. The reference accuracy is the difference in L^2-norm of the final wave function computed with full and half resolution. For all computations it is less than 10^{-4}.

ε	0.0005	0.001	0.005	0.01	0.015	0.02	0.022	0.03
$\|\Pi^+\psi^\varepsilon(t)\|^2$	0.499	0.576	0.792	0.378	0.263	0.276	0.276	0.239
♯ particles, $R=1$	5396	5229	3730	3548	3316	3274	3482	3292
♯ particles, $R=2$	6650	6353	4732	4766	4801	5569	6106	8281
♯ particles, $R=3$	7392	6881	5809	6155	7119	10 536	12 581	34 485
Ref. accuracy $\cdot 10^5$	3.44	2.89	7.57	2.56	2.47	2.45	2.47	2.63

where $v^+(q)$ is a normalized eigenvector of $V(q)$ for the eigenvalue $v(q)+|\phi(q)|$, which depends smoothly on q. The initial center in position space

$$q_0^\varepsilon = \big(1.63 - 4\sqrt{\varepsilon},\, 0.5\sqrt{\varepsilon}\big)$$

is chosen left of the two crossing points $(\gamma_\mathrm{l}, 0)$, $(\gamma_\mathrm{r}, 0)$, where $\gamma_{\mathrm{l,r}}$ are the two solutions of $\cos\varphi = 1 - 2E_1/(W_0 + W_1)$ for $\varphi \in]-\frac{\pi}{2}, \frac{3\pi}{2}]$, that is, $\gamma_\mathrm{l} \approx 1.63$ and $\gamma_\mathrm{r} \approx 4.65$. The initial momentum center and the time interval,

$$p_0 = (1, 0), \qquad [0, T] = [0, 7\sqrt{\varepsilon}],$$

are chosen such that the wave function passes only the left crossing point $(\gamma_\mathrm{l}, 0)$ once without returning to it again. The upper level populations $\|\Pi^+\psi^\varepsilon(t)\|^2$ at the final time $T = 7\sqrt{\varepsilon}$, which are given in Table 1, confirm that the experimental set-up produces leading order nonadiabatic transitions.

The numerical realization of the surface hopping semigroup $(\mathcal{L}^t_{\varepsilon,R})_{t\geq 0}$ is the same as for the simulations of models with a linear isotropic potential matrix presented in [17], up to adding the $R\sqrt{\varepsilon}$-dependent jump criterion

$$t \mapsto |\phi(q^\pm(t))| \text{ has a local minimum in } t = t^* \text{ and } |\phi(q^\pm(t^*))| \leq R\sqrt{\varepsilon}.$$

The initial sampling is performed on 16×16 grids in position and momentum space, and the classical transport is discretized by the explicit Runge–Kutta method of Dormand and Prince `DOPRI45`. For comparison, we have also solved the Schrödinger equation (1.1) by a numerically converged pseudospectral Strang splitting scheme. The solutions obtained with a 1024×512 space grid on the computational domain $[1.63 - 8\sqrt{\varepsilon}, 1.63 + 16\sqrt{\varepsilon}] \times [-6\sqrt{\varepsilon}, 6\sqrt{\varepsilon}]$ and with 10^4 time steps are regarded as a reference, since they differ in L^2-norm from the corresponding solution with a fourth of the grid points and half the time steps by less than 10^{-4}; see Table 1. We have computed the following quadratic quantities of the wave function at final time $T = 7\sqrt{\varepsilon}$: the level populations $\|\Pi^\pm\psi^\varepsilon(t)\|^2$ and the expectation values of position and momentum on each level,

$$\langle \Pi^\pm(q)\psi^\varepsilon(q,t), q\Pi^\pm(q)\psi^\varepsilon(q,t)\rangle, \qquad \langle \Pi^\pm(q)\psi^\varepsilon(q,t), -i\varepsilon\nabla_q\Pi^\pm(q)\psi^\varepsilon(q,t)\rangle.$$

We note that the reference solver restricts the length of the time interval to $7\sqrt{\varepsilon}$, since the dynamics can be resolved only as long as the solution stays well localized in the computational domain, while for the fixed number of 1024×512 grid points the size of the computational domain affects the numerical accuracy.

Comparing the outcome of the two algorithms, we find all errors within and below the corridor $[\frac{1}{5}\sqrt{\varepsilon}, 5\sqrt{\varepsilon}]$ (see Figure 1), which is better than the proven convergence rate $\varepsilon^{1/8}$. Moreover, the errors increase monotonically when increasing the semiclassical parameter; however, when entering the range $\varepsilon \geq 0.01$ we observe different dependencies. The level populations' error starts decreasing, while the error of the momentum expectation oscillates. We have no mathematical explanation for these observations and can make only an educated guess. The good convergence rate might be caused by the localization properties of the initial Gaussian wave packet combined with the short length of the time interval. Indeed, the error terms $O(1/R^2)$ and $O(1/(R^5\sqrt{\varepsilon}))$ in the approximation by classical transport in Proposition 2.3 might be negligible in this case as well as the contribution $O(1/R^2)$ in Proposition 2.4, which is due to localization in energy. The tendencies for a large semiclassical parameter, however, are clearly beyond the reach of our asymptotic analysis.

Table 1 shows that an increase of the hopping range R increases the number of final particles, that is, the number of jumps within the overall time interval. We obtain particle numbers around 3000 and 30000 for $R = 1$ and $R = 3$, respectively, resulting in half a minute and 5 minutes computing time for our implementation of the algorithm in MATLAB 7.0 on a 3GHz Pentium 4 computer. However, an enlarged hopping range need not improve the accuracy of the approximation. The plots in Figure 1 mostly display smaller errors for larger R, but the level populations in the physical relevant range of $\varepsilon = 0.022$ have the most accurate computation for $R = 2$.

Summarizing, the numerical experiments are consistent with the theoretical result, but also present a better convergence rate and tendencies in the range of a larger semiclassical parameter, which seem to be unexplainable by our asymptotic analysis. A systematic comparison with the well-established surface hopping algorithms of chemical physics is in progress.

4. Propagation outside the crossing zone. We now begin proving our main result. The first step is to establish the validity of the classical transport approximation in the zone of large eigenvalue gap $\{|\phi| \geq C\, R\sqrt{\varepsilon}\}$.

Proof of Proposition 2.3. Our aim is to prove

$$\operatorname{tr} \int \left\{\chi(t)c(q,p)b\Big(\tfrac{\phi(q)}{R\sqrt{\varepsilon}}\Big), \tau + \tfrac{1}{2}|p|^2 + v(q) \pm |\phi(q)|\right\} \Pi^{\pm}(q) W^{\varepsilon}(\psi^{\varepsilon}(t))(q,p)\, \mathrm{d}q\, \mathrm{d}p\, \mathrm{d}t$$
$$= O\big(\tfrac{1}{R^2}\big) + O\Big(\tfrac{1}{R^5\sqrt{\varepsilon}}\Big) + O(\sqrt{\varepsilon}),$$

since then classical transport follows immediately. The key argument is the estimation of the action of the commutator

$$K = \tfrac{1}{\varepsilon}\left[\chi(t)\mathrm{op}_{\varepsilon}\Big(c(q,p)b\Big(\tfrac{\phi(q)}{R\sqrt{\varepsilon}}\Big)\,\Pi^{\pm}(q)\Big)\ ,\ -i\varepsilon\partial_t - \tfrac{\varepsilon^2}{2}\Delta_q + V(q)\right]$$

on the solution of the Schrödinger equation (1.1). Indeed, observing that

$$(K\psi^{\varepsilon}, \psi^{\varepsilon})_{L^2(\mathbb{R}^{d+1})} = 0,$$

we are going to prove

$$(K\psi^{\varepsilon}, \psi^{\varepsilon})_{L^2(\mathbb{R}^{d+1})} = O\big(\tfrac{1}{R^2}\big) + O\Big(\tfrac{1}{R^5\sqrt{\varepsilon}}\Big) + O(\sqrt{\varepsilon})$$
$$+ \left(\mathrm{op}_{\varepsilon}\left(\left\{\chi(t)c(q,p)b\Big(\tfrac{\phi(q)}{R\sqrt{\varepsilon}}\Big), \tau + \tfrac{1}{2}|p|^2 + v(q) \pm |\phi(q)|\right\}\Pi^{\pm}(q)\right)\psi^{\varepsilon}, \psi^{\varepsilon}\right)_{L^2(\mathbb{R}^{d+1})},$$

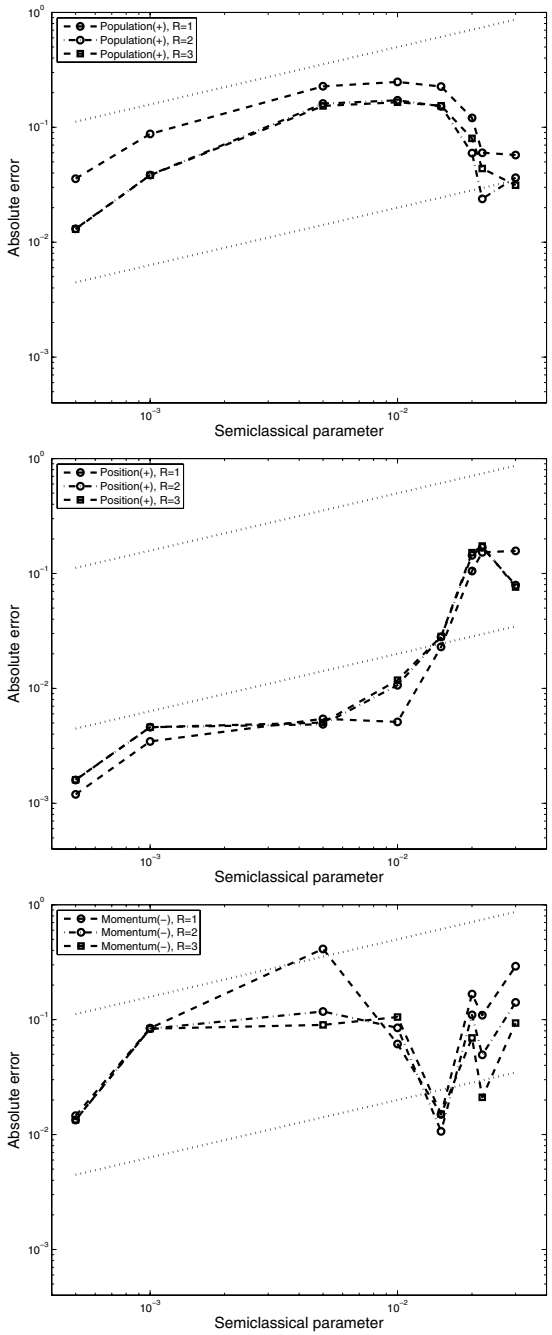


FIG. 1. *Double logarithmic plots of the absolute error, when comparing the outcome of the surface hopping algorithm with a numerically converged pseudospectral splitting scheme. The semiclassical parameter ε varies in the set $\{0.0005, 0.001, 0.005, 0.01, 0.015, 0.02, 0.022, 0.03\}$. The dashed, dashed-dotted, and solid lines refer to a hopping range $R = 1, 2, 3$, respectively. The three plots show the error of the level population $\|\Pi^+(q)\psi^\varepsilon(q,t)\|^2$, of the position expectation value $\langle \Pi^+(q)\psi^\varepsilon(q,t), q\Pi^+(q)\psi^\varepsilon(q,t)\rangle$, and of the momentum expectation $\langle \Pi^-(q)\psi^\varepsilon(q,t), -i\varepsilon\nabla_q\Pi^-(q)\psi^\varepsilon(q,t)\rangle$ at the final time $T = 7\sqrt{\varepsilon}$. All errors lie in and below the corridor defined by the two functions $\varepsilon \mapsto 5\sqrt{\varepsilon}$ and $\varepsilon \mapsto \frac{1}{5}\sqrt{\varepsilon}$, which are represented by dotted lines.*

where op_ε also denotes Weyl quantized operators acting on space-time variables. We use the scaling operator

$$T : L^2_{\mathrm{loc}}(\mathbb{R}^{d+1}) \to L^2_{\mathrm{loc}}(\mathbb{R}^{d+1}), \quad (T\psi)(t,q) = \varepsilon^{d/4}\psi(t,\sqrt{\varepsilon}q) \tag{4.1}$$

and write

$$T^*KT = \tfrac{1}{\varepsilon}\Big[\mathrm{op}_1\Big(\chi(t)c(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)b\Big(\tfrac{\phi(\sqrt{\varepsilon}q)}{R\sqrt{\varepsilon}}\Big)\,\Pi^\pm(\sqrt{\varepsilon}q)\Big)\ ,\ -i\varepsilon\partial_t - \tfrac{\varepsilon}{2}\Delta_q + V(\sqrt{\varepsilon}q)\Big].$$

We have to deal carefully with the ε, R dependence of the symbol in the left-hand side of the commutator. For all multi-indices $\alpha \in \mathbb{N}^d$,

$$D^\alpha\big(\Pi^\pm(q)\big) = O(|\phi(q)|^{-|\alpha|}).$$

Since $|\phi(q)| > C\,R\sqrt{\varepsilon}$ on the support of $b_{\varepsilon,R}(q,p) := c(q,p)b(\frac{\phi(q)}{R\sqrt{\varepsilon}})$, we have

$$D^\alpha_q\big((b_{\varepsilon,R}\Pi^\pm)(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\big) = O(1), \quad D^\alpha_p\big((b_{\varepsilon,R}\Pi^\pm)(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\big) = O(\varepsilon^{|\alpha|/2}) \tag{4.2}$$

for $\alpha \in \mathbb{N}^d$. By the symbolic calculus of Lemma A.2,

$$\begin{aligned}\tfrac{1}{\varepsilon}\big[\mathrm{op}_1\big((b_{\varepsilon,R}\Pi^\pm)\big(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p\big)\big),\mathrm{op}_1\big(\tfrac{\varepsilon}{2}|p|^2\big)\big] &= \tfrac{1}{i}\,\mathrm{op}_1(\{(b_{\varepsilon,R}\Pi^\pm)\big(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p\big),\tfrac{1}{2}|p|^2\}) \\ &= \tfrac{1}{i}\,\mathrm{op}_1\big(\{b_{\varepsilon,R}(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p),\tfrac{1}{2}|p|^2\}\Pi^\pm(\sqrt{\varepsilon}q)\big) - \tfrac{1}{i}\,\mathrm{op}_1\big(r_0(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\big),\end{aligned}$$

where

$$r_0(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p) = b_{\varepsilon,R}(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\sum_{j=1}^d \sqrt{\varepsilon}p_j(\partial_{q_j}\Pi^\pm)(\sqrt{\varepsilon}q). \tag{4.3}$$

Moreover, in view of $[\Pi^\pm, V] = 0$ and (4.2),

$$\begin{aligned}\tfrac{1}{\varepsilon}\big[\mathrm{op}_1\big((b_{\varepsilon,R}\Pi^\pm)\big(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p\big)\big)\,,\,V(\sqrt{\varepsilon}q)\big] &= \tfrac{1}{2i\varepsilon}\,\mathrm{op}_1(\{(b_{\varepsilon,R}\Pi^\pm)(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p),V(\sqrt{\varepsilon}q)\}) \\ &\quad - \tfrac{1}{2i\varepsilon}\,\mathrm{op}_1(\{V(\sqrt{\varepsilon}q),(b_{\varepsilon,R}\Pi^\pm)(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\}) + O(\varepsilon).\end{aligned}$$

Working on the Poisson brackets involving $V = v + V_\ell(\phi)$, we first observe that

$$\big\{b_{\varepsilon,R}\Pi^\pm, V_\ell(\phi)\big\} - \big\{V_\ell(\phi), b_{\varepsilon,R}\Pi^\pm\big\} = \Pi^\pm\,\{b_{\varepsilon,R}, V_\ell(\phi)\} - \{V_\ell(\phi), b_{\varepsilon,R}\}\,\Pi^\pm.$$

Using that $V_\ell(\phi) = |\phi|(\Pi^+ - \Pi^-)$ and $\partial_{q_j}\Pi^\pm = \Pi^\pm(\partial_{q_j}\Pi^\pm) + (\partial_{q_j}\Pi^\pm)\Pi^\pm$, we get

$$\big\{b_{\varepsilon,R}\Pi^\pm, V_\ell(\phi)\big\} - \big\{V_\ell(\phi), b_{\varepsilon,R}\Pi^\pm\big\} = \pm 2\{b_{\varepsilon,R},|\phi|\}\Pi^\pm \pm 2\,|\phi|\sum_{j=1}^d \big(\partial_{p_j}b_{\varepsilon,R}\big)\,\partial_{q_j}\Pi^\pm$$

and set

$$\begin{aligned} r_1(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p) &= \tfrac{1}{\varepsilon}|\phi(\sqrt{\varepsilon}q)|\sum_{j=1}^d \partial_{p_j}\big(b_{\varepsilon,R}(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\big)\,\partial_{q_j}\big(\Pi^\pm(\sqrt{\varepsilon}q)\big) \\ &= |\phi(\sqrt{\varepsilon}q)|\sum_{j=1}^d (\partial_{p_j}c)(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\,b\Big(\tfrac{\phi(\sqrt{\varepsilon}q)}{R\sqrt{\varepsilon}}\Big)\,(\partial_{q_j}\Pi^\pm)(\sqrt{\varepsilon}q).\end{aligned} \tag{4.4}$$

Now, collecting all the different pieces, we have

$$K = \mathrm{op}_\varepsilon\big(\{\chi(t)b_{\varepsilon,R}(q,p)\,,\,\tau+\tfrac12|p|^2+v(q)\pm|\phi(q)|\}\,\Pi^\pm(q)\big)$$
$$+\mathrm{op}_\varepsilon(\chi(t)r_0(q,p))+\mathrm{op}_\varepsilon(\chi(t)r_1(q,p))+O(\varepsilon).$$

Hence, our claim follows from the analysis of r_0 and r_1, which is carried out in Lemma 4.1. ∎

LEMMA 4.1. *Let ψ^ε solve the Schrödinger equation* (1.1). *For the matrix-valued functions r_0 and r_1 defined in* (4.3) *and* (4.4), *respectively, one has*

$$\left(\mathrm{op}_\varepsilon\left(\chi(t)r_0(q,p)\right)\psi^\varepsilon,\psi^\varepsilon\right)_{L^2(\mathbb{R}^{d+1})}=O(\sqrt{\varepsilon})+O(1/R^2)+O(1/(R^5\sqrt{\varepsilon})),$$

$$\left(\mathrm{op}_\varepsilon\left(\chi(t)r_1(q,p)\right)\psi^\varepsilon,\psi^\varepsilon\right)_{L^2(\mathbb{R}^{d+1})}=O(\sqrt{\varepsilon})+O(1/R^2).$$

Proof. Both functions have off-diagonal matrix structure; that is, $r_0(q,p)$ and $r_1(q,p)$ do not commute with $V(q)$. However, since r_1 contains an additional factor $|\phi|$, it is less singular than r_0, in the sense that for $\alpha\in\mathbb{N}^d$ and (q,p) with $|\phi(\sqrt{\varepsilon}q)|>CR\sqrt{\varepsilon}$

$$D_q^\alpha\big(r_0(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\big)=O(R^{-|\alpha|-1}\varepsilon^{-1/2}),\qquad D_q^\alpha\big(r_1(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\big)=O(R^{-|\alpha|}).$$

We begin with r_1. We write $r_1=\Pi^-r_1\Pi^++\Pi^+r_1\Pi^-$ and work successively with each part. Thus, without loss of generality, we suppose that $\Pi^-r_1\Pi^+=r_1$. The strategy is to reuse the Schrödinger equation, since

$$r_1=\Pi^-r_1\Pi^+=\tfrac{1}{2|\phi|}[r_1,V_\ell(\phi)]=\tfrac{1}{2|\phi|}[r_1,\tau+\tfrac12|p|^2+V].$$

With the scaling operator T defined in (4.1), we obtain

$$\mathrm{op}_\varepsilon(\chi(t)r_1(q,p))=T^*\mathrm{op}_1\Big(\Big[\tfrac{\chi(t)}{2|\phi(\sqrt{\varepsilon}q)|}\,r_1(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p),\varepsilon\tau+\tfrac{\varepsilon}{2}|p|^2+V(\sqrt{\varepsilon}q)\Big]\Big)\,T.$$

Then, by the symbolic calculus of Lemma A.2,

$$\mathrm{op}_1\Big(\Big[\tfrac{\chi(t)}{2|\phi(\sqrt{\varepsilon}q)|}\,r_1(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p),\varepsilon\tau+\tfrac{\varepsilon}{2}|p|^2+V(\sqrt{\varepsilon}q)\Big]\Big)$$
$$=\Big[\mathrm{op}_1\Big(\tfrac{\chi(t)}{2|\phi(\sqrt{\varepsilon}q)|}\,r_1(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\Big),\mathrm{op}_1\big(\varepsilon\tau+\tfrac{\varepsilon}{2}|p|^2+V(\sqrt{\varepsilon}q)\big)\Big]+\mathrm{op}_1(r_2(t,\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)),$$

where

$$r_2(t,\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)=\chi(t)\tilde r_1(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)+\varepsilon\tfrac{\chi'(t)}{2|\phi(\sqrt{\varepsilon}q)|}\,r_1(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)+\varepsilon\chi(t)\tilde r_2(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)$$

with

$$\tilde r_1(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)=\tfrac{1}{2i}\Big\{\tfrac{1}{2|\phi(\sqrt{\varepsilon}q)|}\,r_1(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\,,\,\tfrac{\varepsilon}{2}|p|^2+V(\sqrt{\varepsilon}q)\Big\}$$
$$-\tfrac{1}{2i}\Big\{\tfrac{\varepsilon}{2}|p|^2+V(\sqrt{\varepsilon}q)\,,\,\tfrac{1}{2|\phi(\sqrt{\varepsilon}q)|}\,r_1(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\Big\}.$$

The term $\tilde r_2(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)$ contains the commutator of second derivatives in p of the symbol $|\phi(\sqrt{\varepsilon}q)|^{-1}\,r_1(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)$ with second derivatives of $V(\sqrt{\varepsilon}q)$ and a linear combination of derivatives in (q,p) of order greater than or equal to three. Hence, by Lemma A.2

$$\mathrm{op}_\varepsilon\left(\chi(t)\tilde r_2(q,p)\right)=O(\varepsilon\sqrt{\varepsilon})+O(\varepsilon/R^4)\ \text{ in }\ \mathcal{L}(L^2(\mathbb{R}^{d+1})).$$

It remains to study $\tilde{r}_1$. Since derivatives in p of $r_1(\sqrt{\varepsilon}q, \sqrt{\varepsilon}p)$ generate powers of $\sqrt{\varepsilon}$, the bracket with $V(\sqrt{\varepsilon}q)$ gives

$$\mathrm{op}_1\Big(\chi(t)\Big\{\tfrac{1}{2|\phi(q\sqrt{\varepsilon})|}r_1(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p), V(\sqrt{\varepsilon}q)\Big\}\Big) = O\Big(\tfrac{\sqrt{\varepsilon}}{R}\Big) \text{ in } \mathcal{L}(L^2(\mathbb{R}^{d+1})).$$

For the bracket with $\frac{\varepsilon}{2}|p|^2$ one has

$$\Big\{\tfrac{1}{2|\phi(\sqrt{\varepsilon}q)|}r_1(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\,,\,\tfrac{\varepsilon}{2}|p|^2\Big\}$$
$$= \tfrac{\sqrt{\varepsilon}}{2|\phi(\sqrt{\varepsilon}q)|}\sqrt{\varepsilon}p\cdot\nabla_q\big(r_1(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\big) + \tfrac{\varepsilon}{2}\tfrac{\mathrm{d}\phi(\sqrt{\varepsilon}q)\sqrt{\varepsilon}p\cdot\phi(\sqrt{\varepsilon}q)}{|\phi(\sqrt{\varepsilon}q)|^3}\,r_1(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p).$$

Both terms give a contribution of order $O(1/R^2)$, but since they are not purely off-diagonal, the preceding commutator argument cannot be reiterated. Hence,

$$\mathrm{op}_\varepsilon(\chi(t)\tilde{r}_1(q,p)) = O\big(\tfrac{1}{R^2}\big) + O\Big(\tfrac{\sqrt{\varepsilon}}{R}\Big), \qquad \mathrm{op}_\varepsilon\Big(\varepsilon\tfrac{\chi'(t)}{|\phi(q)|}\,r_1(q,p)\Big) = O\Big(\tfrac{\sqrt{\varepsilon}}{R}\Big)$$

in $\mathcal{L}(L^2(\mathbb{R}^{d+1}))$, and we have proven one part of the lemma.

In the case of the more singular symbol r_0, the previous strategy results in an error of size

$$1/R\sqrt{\varepsilon}\,\big(O(\varepsilon\sqrt{\varepsilon}) + O(\varepsilon/R^4) + O(\sqrt{\varepsilon}/R) + O(1/R^2)\big)$$
$$= O(\sqrt{\varepsilon}) + O(1/R^2) + O(1/(R^3\sqrt{\varepsilon})).$$

However, the special form of r_0 allows us to ameliorate the term of order $O(1/(R^3\sqrt{\varepsilon}))$, which stems from the Poisson bracket with $\frac{\varepsilon}{2}|p|^2$. Indeed, we observe that

$$p\cdot\nabla\Pi^+(q) = \tfrac{1}{4|\phi(q)|^3}\left[V_\ell(\phi(q))\,,\,[V_\ell(\phi(q)), V_\ell(\mathrm{d}\phi(q)p)]\right].$$

Therefore, the bracket with $\frac{\varepsilon}{2}|p|^2$ in the $\tilde{r}_1$ term writes as

$$\Big\{\tfrac{1}{4|\phi(\sqrt{\varepsilon}q)|^3}\,b_{\varepsilon,R}(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\left[V_\ell(\phi(\sqrt{\varepsilon}q)), V_\ell(\mathrm{d}\phi(\sqrt{\varepsilon}q)\sqrt{\varepsilon}p)\right],\tfrac{\varepsilon}{2}|p|^2\Big\}$$
$$= \tfrac{\varepsilon}{|\phi(\sqrt{\varepsilon}q)|^4}\left[V_\ell(\phi(\sqrt{\varepsilon}q)), G_\varepsilon(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\right]$$

for some matrix-valued function G_ε with

$$D_q^\alpha\big(G_\varepsilon(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\big) = O(1), \qquad D_p^\alpha\big(G_\varepsilon(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\big) = O(\varepsilon^{|\alpha|/2})$$

for all $\alpha\in\mathbb{N}^d$. We then set

$$\tilde{r}_3(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p) := \tfrac{\varepsilon}{|\phi(\sqrt{\varepsilon}q)|^4}\left[V_\ell(\phi(\sqrt{\varepsilon}q)), G_\varepsilon(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\right]$$
$$= -\Big[\tfrac{\varepsilon}{|\phi(\sqrt{\varepsilon}q)|^4}\,G_\varepsilon(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\,,\,\tau+\tfrac{\varepsilon}{2}|p|^2+V(\sqrt{\varepsilon}q)\Big]$$

and obtain

$$\big(\mathrm{op}_\varepsilon(\chi(t)\tilde{r}_3(q,p))\,\psi^\varepsilon\,,\,\psi^\varepsilon\big)_{L^2(\mathbb{R}^{d+1})}$$
$$= \Big(T^*\mathrm{op}_1\Big(\Big[\tfrac{\varepsilon\chi(t)}{|\phi(\sqrt{\varepsilon}q)|^4}\,G_\varepsilon(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\,,\,\varepsilon\tau+\tfrac{\varepsilon}{2}|p|^2+V(\sqrt{\varepsilon}q)\Big]\Big)\,T\psi^\varepsilon\,,\,\psi^\varepsilon\Big)_{L^2(\mathbb{R}^{d+1})}$$

$$= \tfrac{1}{2i}\Big(T^*\mathrm{op}_1\Big(\Big\{\tfrac{\varepsilon\chi(t)}{|\phi(\sqrt{\varepsilon}q)|^4}\,G_\varepsilon(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\,,\,\varepsilon\tau+\tfrac{\varepsilon}{2}|p|^2+V(\sqrt{\varepsilon}q)\Big\}\Big)\,T\psi^\varepsilon\,,\,\psi^\varepsilon\Big)_{L^2(\mathbb{R}^{d+1})}$$

$$- \tfrac{1}{2i}\Big(T^*\mathrm{op}_1\Big(\Big\{\varepsilon\tau+\tfrac{\varepsilon}{2}|p|^2+V(\sqrt{\varepsilon}q)\,,\,\tfrac{\varepsilon\chi(t)}{|\phi(\sqrt{\varepsilon}q)|^4}\,G_\varepsilon(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)\Big\}\Big)\,T\psi^\varepsilon\,,\,\psi^\varepsilon\Big)_{L^2(\mathbb{R}^{d+1})}$$

$$+ \big(T^*\mathrm{op}_1(\tilde r_4(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p))T\psi^\varepsilon\,,\,\psi^\varepsilon\big)_{L^2(\mathbb{R}^{d+1})}\,,$$

where $\tilde r_4(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)$ contains the commutator of second order derivatives in p of $\varepsilon\,\chi(t)|\phi(\sqrt{\varepsilon}q)|^{-4}\,G_\varepsilon(\sqrt{\varepsilon}q,\sqrt{\varepsilon}p)$ with second derivatives of $V(\sqrt{\varepsilon}q)$ and a linear combination of higher order derivatives in (q,p). Hence, $\mathrm{op}_\varepsilon(\tilde r_4(q\sqrt{\varepsilon},p\sqrt{\varepsilon}))=O(\varepsilon/R^4)$ in $\mathcal{L}(L^2(\mathbb{R}^{d+1}))$. Since the Poisson brackets give a contribution of order $O(1/(R^5\sqrt{\varepsilon}))+O(1/R^4)$, the other part of the lemma is proven, too. ∎

5. Transitions near the crossing. The microlocal normal form used for proving Proposition 2.4 holds locally near some point $(q_0,t_0,p_0,\tau_0)\in\mathbb{R}^{2d+2}$ of the phase space of space-time, which is a crossing point in the sense that $\phi(q_0)=0$ and $\tau_0+v(q_0)+\frac12|p_0|^2=0$.

5.1. Localization in energy. For localization in energy, we consider a cut-off function $\theta\in\mathcal{C}_c^\infty(\mathbb{R})$, $0\le\theta\le1$, with $\theta(x)=1$ for $|x|\le\frac12$ and $\theta(x)=0$ for $|x|>1$. We set

$$\lambda^\pm(q,p,\tau)=\tau+v(q)+\tfrac12|p|^2\pm|\phi(q)|$$

and crucially use the following lemma for suitably reformulating Proposition 2.4.

LEMMA 5.1. *Let $c\in\mathcal{C}_c^\infty(\mathbb{R}^{2d+1+\ell},\mathbb{C})$. If $c_{\varepsilon,R}(t,q,p)=c(t,q,p,\phi(q)/(R\sqrt{\varepsilon}))$ is supported in $\{|\phi|\ge\frac{R}{2}\sqrt{\varepsilon}\}$, then*

$$\mathrm{tr}\int_{\mathbb{R}^{2d+1}} W^\varepsilon(\psi^\varepsilon(t))(q,p)c_{\varepsilon,R}(t,q,p)\Pi^\pm(q)\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}t=O\big(\tfrac{1}{R^2}\big)+O(\sqrt{\varepsilon})$$

$$+\Big(\mathrm{op}_\varepsilon\Big(c_{\varepsilon,R}(t,q,p)\theta\Big(\tfrac{\lambda^\pm(q,p,\tau)}{R\sqrt{\varepsilon}}\Big)\,\Pi^\pm(q)\Big)\,\psi^\varepsilon,\psi^\varepsilon\Big)_{L^2(\mathbb{R}^{d+1})}.$$

Proof. Writing $1-\theta(u)=u\,G(u)$ with $G\in\mathcal{C}^\infty(\mathbb{R})$, we have

$$1-\theta\Big(\tfrac{\lambda^\pm(q,p,\tau)}{R\sqrt{\varepsilon}}\Big)=\tfrac{1}{R\sqrt{\varepsilon}}\lambda^\pm(q,p,\tau)G\Big(\tfrac{\lambda^\pm(q,p,\tau)}{R\sqrt{\varepsilon}}\Big).$$

We now argue as in section 4, using the estimates on $\Pi^\pm$ and $\lambda^\pm(q,p,\tau)\Pi^\pm(q)=\Pi^\pm(q)\big(\tau+\frac12|p|^2+V(q)\big)$. The symbolic calculus of Lemma A.2 yields in $\mathcal{L}(L^2(\mathbb{R}^{d+1}))$

$$\mathrm{op}_\varepsilon\Big(c_{\varepsilon,R}(t,q,p)\,\Big(1-\theta\Big(\tfrac{\lambda^\pm(q,p,\tau)}{R\sqrt{\varepsilon}}\Big)\Big)\,\Pi^\pm(q)\Big)=O(\sqrt{\varepsilon})+O\big(\tfrac{1}{R^2}\big)$$

$$+\tfrac{1}{R\sqrt{\varepsilon}}\,\mathrm{op}_\varepsilon\Big(c_{\varepsilon,R}(t,q,p)G\Big(\tfrac{\lambda^\pm(q,p,\tau)}{R\sqrt{\varepsilon}}\Big)\,\Pi^\pm(q)\Big)\,\mathrm{op}_\varepsilon\big(\tau+\tfrac12|p|^2+V(q)\big).$$

Indeed, the derivatives of the projectors are less harmful than in section 4, since they are divided only by $\sqrt{\varepsilon}$ and not by ε. Since ψ^ε solves the equation, we obtain

$$\Big(\mathrm{op}_\varepsilon\Big(c_{\varepsilon,R}(t,q,p)\,\Big(1-\theta\Big(\tfrac{\lambda^\pm(q,p,\tau)}{R\sqrt{\varepsilon}}\Big)\Big)\,\Pi^\pm(q)\Big)\,\psi^\varepsilon,\psi^\varepsilon\Big)_{L^2(\mathbb{R}^{d+1})}=O\big(\tfrac{1}{R^2}\big)+O(\sqrt{\varepsilon}).$$ ∎

By Lemma 5.1, we introduce the energy cut-off on both sides of equality (2.4), adding an error of order $O(1/R^2) + O(\sqrt{\varepsilon})$. Then, the left-hand side reads

$$\operatorname{tr} \int_{\mathbb{R}^{2d+1}} \chi(t)\, W^\varepsilon(\psi^\varepsilon(t))(q,p) a^\pm_{LZ}(q,p)\, \Pi^\pm(q) \mathrm{d}q\, \mathrm{d}p\, \mathrm{d}t = O\big(\tfrac{1}{R^2}\big) + O(\sqrt{\varepsilon}) + \Big(\mathrm{op}_\varepsilon\Big(a^\pm_{LZ}(q,p)\chi(t)\,\theta\Big(\tfrac{\lambda^\pm(q,p,\tau)}{R\sqrt{\varepsilon}}\Big)\,\Pi^\pm(q)\Big)\,\psi^\varepsilon\,,\,\psi^\varepsilon\Big)_{L^2(\mathbb{R}^{d+1})}.$$

Using the notation

$$f^\pm_a(q,p,j) = \mathbf{1}_{\{j=\pm 1\}}(j)\, a^j(q,p), \qquad (q,p,j) \in \mathbb{R}^{2d}_\pm,$$

for a V-diagonal matrix-valued symbol a, the action of the semigroup on $a^\pm_{LZ}$ can be written as

$$(\mathcal{L}^{\delta^\pm_t}_{\varepsilon,R} a^\pm_{LZ})(q,p)\,\Pi^\pm(q) = (\mathcal{L}^{\delta^\pm_t}_{\varepsilon,R} f^\pm_{a_{LZ}})(q,p,+)\,\Pi^+(q).$$

Then, the right-hand side of (2.4) is

$$\operatorname{tr} \int_{\mathbb{R}^{2d+1}} \chi(t+\delta^\pm_t)\, W^\varepsilon(\psi^\varepsilon(t))(q,p)\,(\mathcal{L}^{\delta^\pm_t}_{\varepsilon,R} a^\pm_{LZ})(q,p)\,\Pi^+(q)\,\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}t = \Big(\mathrm{op}_\varepsilon\Big(\chi(t+\delta_t)\,\theta\Big(\tfrac{\lambda^+(q,p,\tau)}{R\sqrt{\varepsilon}}\Big)\,(\mathcal{L}^{\delta^\pm_t}_{\varepsilon,R} f^\pm_{a_{LZ}})(q,p,+)\,\Pi^+(q)\Big)\,\psi^\varepsilon\,,\,\psi^\varepsilon\Big)_{L^2(\mathbb{R}^{d+1})} + O\big(\tfrac{1}{R^2}\big) + O(\sqrt{\varepsilon}),$$

and the proof of Proposition 2.4 reduces to showing that

$$\Big(\mathrm{op}_\varepsilon\Big(a^\pm_{LZ}(q,p)\chi(t)\theta\Big(\tfrac{\lambda^\pm(q,p,\tau)}{R\sqrt{\varepsilon}}\Big)\,\Pi^\pm(q)\Big)\,\psi^\varepsilon\,,\,\psi^\varepsilon\Big)_{L^2(\mathbb{R}^{d+1})} = \Big(\mathrm{op}_\varepsilon\Big(\chi(t+\delta^\pm_t)\theta\Big(\tfrac{\lambda^+(q,p,\tau)}{R\sqrt{\varepsilon}}\Big)\,(\mathcal{L}^{\delta^\pm_t}_{\varepsilon,R} f^\pm_{a_{LZ}})(q,p,+)\,\Pi^+(q)\Big)\,\psi^\varepsilon\,,\,\psi^\varepsilon\Big)_{L^2(\mathbb{R}^{d+1})} + O\big(\tfrac{1}{R^2}\big) + O(R^3\sqrt{\varepsilon}) + O(\sqrt{\varepsilon}|\ln\varepsilon|) + O\Big(\tfrac{1}{R^5\sqrt{\varepsilon}}\Big). \tag{5.1}$$

For points $(q,t,p,\tau,j) \in \mathbb{R}^{2d+2} \times \{\pm 1\}$, we consider random trajectories

$$\mathcal{T}^{(q,t,p,\tau,j)}_{\varepsilon,R} : [0,+\infty) \to \big(\mathbb{R}^{2d+2} \times \{\pm 1\}\big)$$

with $\mathcal{T}^{(q,t,p,\tau,j)}_{\varepsilon,R}(r) = \big(q^j(r), r+t, p^j(r), \tau, j\big)$ as long as $(q^j(r),p^j(r)) \notin S_{\varepsilon,R}$ and a jump from j to $-j$ with probability $T^\varepsilon(q^*,p^*)$, whenever $(q^j(r),p^j(r))$ hits $S_{\varepsilon,R}$ at a point (q^*,p^*). We keep the notation $(\mathcal{L}^r_{\varepsilon,R})_{r\geq 0}$ for the associated semigroup and set

$$c^\pm_{\varepsilon,R}(q,t,p,\tau) = a^\pm_{LZ}(q,p)\,\chi(t)\,\theta\Big(\tfrac{\lambda^\pm(q,p,\tau)}{R\sqrt{\varepsilon}}\Big). \tag{5.2}$$

Since $r \mapsto \lambda^\pm(q^\pm(r),p^\pm(r),\tau)$ is a constant function, and since within $[t-t_{\mathrm{i}}, t-t_{\mathrm{f}}]$ all involved random trajectories perform a jump, we have

$$(\mathcal{L}^{\delta^\pm_t}_{\varepsilon,R} f^+_{c_{\varepsilon,R}})(q,t,p,\tau,+) = \chi(t+\delta_t)\,\theta\Big(\tfrac{\lambda^+(q,p,\tau)}{R\sqrt{\varepsilon}}\Big)\,(\mathcal{L}^{\delta_t}_{\varepsilon,R} f^+_{a_{LZ}})(q,p,+),$$

$$(\mathcal{L}^{\delta_t}_{\varepsilon,R} f^-_{c_{\varepsilon,R}})(q,t,p,\tau,+) = \chi(t+\delta_t)\,\theta\Big(\tfrac{\lambda^+(q,p,\tau)}{R\sqrt{\varepsilon}}\Big)\,(\mathcal{L}^{\delta_t}_{\varepsilon,R} f^-_{a_{LZ}})(q,p,+).$$

With this notation, (5.1) and, consequently, Proposition 2.4 are equivalent to

$$\begin{aligned}\Big(\mathrm{op}_\varepsilon\Big(c^\pm_{\varepsilon,R}(q,t,p,\tau)\Pi^\pm(q)\Big)\,\psi^\varepsilon\,,\,\psi^\varepsilon\Big)_{L^2(\mathbb{R}^{d+1})}\\ =\Big(\mathrm{op}_\varepsilon\Big((\mathcal{L}^{\delta^\pm_t}_{\varepsilon,R}f^\pm_{c_{\varepsilon,R}})(q,t,p,\tau,+)\,\Pi^+(q)\Big)\,\psi^\varepsilon\,,\,\psi^\varepsilon\Big)_{L^2(\mathbb{R}^{d+1})}\\ +O(1/R^2)+O(R^3\sqrt{\varepsilon})+O(\sqrt{\varepsilon}|\ln\varepsilon|)+O(1/(R^5\sqrt{\varepsilon})).\end{aligned}\tag{5.3}$$

We emphasize that the symbols $c^\pm_{\varepsilon,R}$ and $\mathcal{L}^{\delta^\pm_t}_{\varepsilon,R}f^\pm_{c_{\varepsilon,R}}$ are compactly supported inside the annulus $\{\frac{R}{2}\sqrt{\varepsilon}<|\phi|<R\sqrt{\varepsilon}\}$ at a distance of order $R\sqrt{\varepsilon}$ of $J^{\pm,out}$, where

$$J^{\pm,in/out}=\Big\{(q,t,p,\tau)\in\mathbb{R}^{2d+2}\mid(q,p)\in M^{\pm,in/out},\,\tau+v(q)+\tfrac12|p|^2\pm|\phi(q)|=0\Big\}$$

denote the submanifolds, which consist of all Hamiltonian trajectories entering (respectively, leaving) the crossing set.

5.2. The normal form. Let us first recall some basic facts about canonical transforms and Fourier integral operators. The phase space $T^*\mathbb{R}^{d+1}=\mathbb{R}^{d+1}\times\mathbb{R}^{d+1}$ is a symplectic space, once endowed with the symplectic form $\omega=\mathrm{d}\tau\wedge\mathrm{d}t+\mathrm{d}p\wedge\mathrm{d}q$. A canonical transform $\kappa:T^*\mathbb{R}^{d+1}\to T^*\mathbb{R}^{d+1}$ is a change of coordinates, which preserves the symplectic form. With a canonical transform κ, one associates a unitary operator $U:L^2(\mathbb{R}^{d+1})\to L^2(\mathbb{R}^{d+1})$ such that for all $a\in\mathcal{C}^\infty_c(\mathbb{R}^{2d+2},\mathbb{C}^{N(\ell)\times N(\ell)})$

$$U^*\mathrm{op}_\varepsilon(a)\,U=\mathrm{op}_\varepsilon(a\circ\kappa)+O(\varepsilon^2)$$

as bounded operators on $L^2(\mathbb{R}^{d+1})$; see, for example, section 2.2 in [7]. The operator U is a Fourier integral operator. The last equality extends to symbols of the form

$$b_{\varepsilon,R}(q,t,p,\tau)=b\Big(q,t,p,\tau,\tfrac{f(q,p)}{R\sqrt{\varepsilon}}\Big)$$

with $b\in\mathcal{C}^\infty_c(\mathbb{R}^{2d+3},\mathbb{C}^{N(\ell)\times N(\ell)})$ and $f\in\mathcal{C}^\infty(\mathbb{R}^{2d},\mathbb{R})$ according to

$$U^*\mathrm{op}_\varepsilon(b_{\varepsilon,R})\,U=\mathrm{op}_\varepsilon(b_{\varepsilon,R}\circ\kappa)+O(\sqrt{\varepsilon}).\tag{5.4}$$

The proof of this statement follows the proof of Lemma 2 in [7]: one uses symbolic calculus for the commutator of a usual semiclassical pseudodifferential operator and a two-scale one of the form $\mathrm{op}_\varepsilon(b_{\varepsilon,R})$, hence the gain of $\sqrt{\varepsilon}$.

We shall crucially use the following microlocal normal form result, which for codimension two and three crossings is proven in [1, 2], however, without the explicit equations (5.6)–(5.9). These equations, including the normal form for codimension five crossings, are provided in Theorem 1 and Proposition 4 of [6].

THEOREM 5.2 (see [6]). *We consider $\rho_0=(q_0,t_0,p_0,\tau_0=-v(q_0)-\frac12|p_0|^2)$ such that $\phi(q_0)=0$, $\mathrm{d}\phi(q_0)p_0\neq0$, and $\mathrm{d}\phi$ is of maximal rank near q_0. Then, there exists a local canonical transform κ from a neighborhood of ρ_0 into some neighborhood Ω of 0,*

$$\kappa:(q,t,p,\tau)\mapsto(z,s,\zeta,\sigma),\quad\kappa(\rho_0)=0.$$

There exist a Fourier integral operator U associated with κ^{-1} and an invertible matrix-valued symbol $A_\varepsilon=A_0+\varepsilon A_1+\varepsilon^2A_2+\cdots$ such that $v^\varepsilon=U^*\mathrm{op}_\varepsilon(A_\varepsilon)^{-1}\psi^\varepsilon$ satisfies for all $\varphi\in\mathcal{C}^\infty_c(\Omega,\mathbb{R})$

$$\mathrm{op}_\varepsilon(\varphi)\,\mathrm{op}_\varepsilon\left(-\sigma+V_\ell\left(s,\tilde z+\gamma_\varepsilon(z,\zeta)\right)\right)v^\varepsilon=O(\varepsilon^\infty)\tag{5.5}$$

in $L^2(\mathbb{R}^{d+1})$, *where* $z=(\tilde z,z')\in\mathbb{R}^d$ *with* $\tilde z\in\mathbb{R}^{\ell-1}$ *and* $\gamma_\varepsilon=\gamma_\varepsilon(z,\zeta)$ *is a vector-valued symbol* $\gamma_\varepsilon=\gamma_0+\varepsilon\gamma_1+\varepsilon^2\gamma_2+\cdots\in\mathbb{R}^{\ell-1}$ *with*

$$\gamma_\varepsilon=0 \ \text{ for } \ \ell=2, \qquad \gamma_\varepsilon=O(|\tilde z|^2) \ \text{ for } \ \ell>2.$$

$\tilde z\in\mathbb{R}^{\ell-1}$ *contains the coordinates of the vector* $|\mathrm{d}\phi(q)p|^{-1/2}\pi_\ell(q,p)\phi(q)$ *in an orthonormal basis of the hyperplane normal to* $\mathrm{d}\phi(q)p$ *up to* $O(|\phi(q)|^2)$, *while*

$$\begin{aligned} s&=-|\mathrm{d}\phi(q)p|^{-1/2}\tfrac{\mathrm{d}\phi(q)p}{|\mathrm{d}\phi(q)p|}\cdot\phi(q)+O(s^2+\sigma^2+|\tilde z|^2),\\ \sigma&=|\mathrm{d}\phi(q)p|^{-1/2}\left(\tau+\tfrac12|p|^2+v(q)\right)+O(s^2+\sigma^2+|\tilde z|^2). \end{aligned}\tag{5.6}$$

Moreover,

$$J^{\pm,in}=\{\sigma\mp s=0,\ \tilde z=0,\ s\le0\},\quad J^{\pm,out}=\{\sigma\pm s=0,\ \tilde z=0,\ s\ge0\},\tag{5.7}$$

and there exists $\gamma\in\{-1,+1\}$ *such that for all* $\rho=(q,t,p,\tau)$ *with* $\kappa(\rho)=(z,s,\zeta,\sigma)$

$$A_0^*(\rho)\left(\tau+\tfrac12|p|^2+V(q)\right)A_0(\rho)=\gamma\left(-\sigma+V_\ell(s,\tilde z+\gamma_0(z,\zeta))\right),\tag{5.8}$$

$$A_0^*(\rho)V_\ell\Big(\tfrac{\pi_\ell(q,p)\phi(q)}{|\pi_\ell(q,p)\phi(q)|}\Big)A_0(\rho)=\gamma\,V_\ell\Big(0,\tfrac{\tilde z}{|\tilde z|}\Big)+O(\sqrt{\sigma^2+s^2+|\tilde z|^2}).\tag{5.9}$$

We denote by

$$\widetilde\Pi^\pm(z,s,\zeta)=\tfrac12\left(\mathrm{Id}\mp\tfrac{1}{\sqrt{s^2+|\tilde z+\gamma_0(z,\zeta)|^2}}V_\ell(s,\tilde z+\gamma_0(z,\zeta))\right),$$

$$\tilde\lambda^\pm(z,s,\zeta,\sigma)=-\sigma\mp\sqrt{s^2+|\tilde z+\gamma_0(z,\zeta)|^2}$$

the spectral projectors and the eigenvalues of $-\sigma+V_\ell(s,\tilde z+\gamma_0(z,\zeta))$. Due to the relation $J^{\pm,in/out}\subseteq\{-\sigma\mp\sqrt{s^2+|\tilde z|^2}=0,\tilde z=0\}$, the labeling of $\widetilde\Pi^\pm$ coincides on $J^{\pm,in/out}$ with the one for $\Pi^\pm$.

Proposition 5.3. *There exist functions* $k^\pm$ *such that if* $\kappa(q,t,p,\tau)=(z,s,\zeta,\sigma)$, *the projectors* $\Pi^\pm$ *and* $\widetilde\Pi^\pm$ *are related by*

$$\widetilde\Pi^\pm(z,s,\zeta,\sigma)=\left(k^\pm A_0^*\Pi^\pm A_0\right)(q,t,p,\tau) \ \text{ on } \ \Sigma^\mp=\{\lambda^\mp=0\}.\tag{5.10}$$

If $S=\{\phi(q)=0,\ \tau+\frac12|p|^2+v(q)=0\}$, *then* $k^+_{|S}=k^-_{|S}=e\neq0$ *and* $(e\,A_0^*A_0)_{|S}=\mathrm{Id}_{|S}$. *Moreover, on* $\Sigma^\pm\cap\{0<|\phi|\le R\sqrt\varepsilon\}$,

$$\widetilde\Pi^\pm(z,s,\zeta,\sigma)=e(A_0^*\Pi^\pm A_0)(q,t,p,\tau)+O(R\sqrt\varepsilon).\tag{5.11}$$

Proof. For convenience, we set

$$P=\tau+\tfrac12|p|^2+V(q),\quad \widetilde P=-\sigma+V_\ell(s,\tilde z+\gamma_0(z,\zeta)).$$

By (5.8), the use of determinants gives $\Sigma^+\cup\Sigma^-=\kappa^{-1}(\{\tilde\lambda^+=0\}\cup\{\tilde\lambda^-=0\})$. Considering the equations of $J^{\pm,in/out}$, the only possibility is

$$\Sigma^\pm=\kappa^{-1}\left(\{\tilde\lambda^\pm=0\}\right).$$

Therefore, for $\rho \in \Sigma^+$ we have

$$\gamma \widetilde{P}(\kappa(\rho)) = \gamma(\tilde{\lambda}^- \widetilde{\Pi}^-)(\kappa(\rho)) = (A_0^* P A_0)(\rho) = (\lambda^- A_0^* \Pi^- A_0)(\rho).$$

The same argument for Σ^- gives (5.10).

The fact $k^+_{|S} = k^-_{|S}$ comes from the precise analysis of the Hamiltonian vector fields associated with the eigenvalues $\lambda^\pm$. Let $\rho \in S$. We find in [6, section 5] that if

$$H(\rho) = \lim_{\alpha \to 0^\mp} H_{\lambda^\pm}(\Phi^\alpha_\pm(q,p)), \quad H'(\rho) = \lim_{\alpha \to 0^\pm} H_{\lambda^\pm}(\Phi^\alpha_\pm(q,p)),$$

then there exists a nonzero function e such that

$$H = e(\partial_s + \partial_\sigma), \quad H' = e(\partial_s - \partial_\sigma) \text{ on } S.$$

Since $\lambda^\pm = 0$ and $\tilde{\lambda}^\pm = 0$ on S, we have $k^+_{|S} = k^-_{|S} = e$. Next, we consider the limit of the projectors $\Pi^\pm$ along outgoing trajectories,

$$\Pi^\mp_S(\rho) = \lim_{\alpha \to 0^-} \Pi^\mp(\Phi^\alpha_\pm(q,p)) = \tfrac{1}{2}\left(\mathrm{Id} \mp V_\ell\Big(\tfrac{\mathrm{d}\phi(q)p}{|\mathrm{d}\phi(q)p|}\Big)\right).$$

Then, (5.10) gives on S

$$eA_0^* A_0 = eA_0^*\left(\Pi^+_S + \Pi^-_S\right) A_0 = \left(\widetilde{\Pi}^+ + \widetilde{\Pi}^-\right) \circ \kappa = \mathrm{Id}.$$

Finally, let $\rho \in \Sigma^+$. By relation (5.8), we have for any vector $w \in \mathbb{C}^{N(\ell)}$ that $w \in \operatorname{Ker} \widetilde{P}(\kappa(\rho))$ if and only if $A_0(\rho) w \in \operatorname{Ker} P(\rho)$. Moreover, $\operatorname{Ker} P(\rho) = \operatorname{Ran} \Pi^+(\rho)$ and $\operatorname{Ker} \widetilde{P}(\kappa(\rho)) = \operatorname{Ran} \widetilde{\Pi}^+(\kappa(\rho))$. We therefore obtain

$$\operatorname{Ran} \widetilde{\Pi}^+(\kappa(\rho)) = \operatorname{Ran} (A_0^{-1} \Pi^+ A_0)(\rho).$$

Since $\sqrt{e} A_0$ is unitary on S, we have

$$(A_0^{-1}\Pi^+ A_0)^*(\rho) = (A_0^{-1}\Pi^+ A_0)(\rho) + O(R\sqrt{\varepsilon})$$

for $\rho \in \Sigma^+ \cap \{|\phi| \leq R\sqrt{\varepsilon}\}$. Since the two projectors $\widetilde{\Pi}^+(\kappa(\rho))$ and $(A_0^{-1}\Pi^+ A_0)(\rho)$ have the same range, while one of them is orthogonal and the other orthogonal up to $O(R\sqrt{\varepsilon})$, they coincide up to $O(R\sqrt{\varepsilon})$. The same argument holds for $\rho \in \Sigma^-$, and we have proven relation (5.11). ∎

As the next step towards proving the claimed identity (5.3), we perform the canonical change of coordinates for arriving at the microlocal normal form.

PROPOSITION 5.4. *Let $c^\pm_{\varepsilon,R} \in \mathcal{C}^\infty_c(\mathbb{R}^{2d+2}, \mathbb{C})$ be the functions defined in* (5.2) *and $v^\varepsilon = U^* \mathrm{op}_\varepsilon(A_\varepsilon)^{-1} \psi^\varepsilon$. Denote*

$$\widetilde{T}^\varepsilon(\tilde{z}) = \exp(-\tfrac{\pi}{\varepsilon}|\tilde{z}|^2). \tag{5.12}$$

Then, there exist functions $b^\pm \in \mathcal{C}^\infty_c(\mathbb{R}^{2d+2}, \mathbb{C})$ and $s^\pm_1 \in \mathbb{R}$, such that $b^\pm(z,s,\zeta,\eta)$ and $b^\pm(z, s^\pm_1 + s, \zeta, \eta)$ are compactly supported in $\{s > 0\}$ and $\{s < 0\}$, respectively, and satisfy

$$\left(\mathrm{op}_\varepsilon\Big(c^+_{\varepsilon,R}(q,t,p,\tau)\Pi^+(q)\Big)\psi^\varepsilon(q,t), \psi^\varepsilon(q,t)\right)_{L^2}$$

$$= \Big(\mathrm{op}_\varepsilon\Big(b^+\Big(z,s,\zeta,\tfrac{\tilde z}{R\sqrt\varepsilon}\Big)\Big)\, v_2^\varepsilon(z,s), v_2^\varepsilon(z,s)\Big)_{L^2} + O(R\sqrt\varepsilon),$$

$$\Big(\mathrm{op}_\varepsilon\Big((\mathcal{L}^{\delta_t}_{\varepsilon,R} f^+_{c_{\varepsilon,R}})(q,t,p,\tau,+)\Pi^+(q)\Big)\,\psi^\varepsilon(q,t), \psi^\varepsilon(q,t)\Big)_{L^2}$$

$$= \Big(\mathrm{op}_\varepsilon\Big((1-\widetilde T^\varepsilon(\tilde z))b^+\Big(z,s+s_1^+,\zeta,\tfrac{\tilde z}{R\sqrt\varepsilon}\Big)\Big)\, v_1^\varepsilon(z,s), v_1^\varepsilon(z,s)\Big)_{L^2} + O(R^3\sqrt\varepsilon),$$

$$\Big(\mathrm{op}_\varepsilon\Big(c^-_{\varepsilon,R}(q,t,p,\tau)\Pi^-(q)\Big)\,\psi^\varepsilon(q,t), \psi^\varepsilon(q,t)\Big)_{L^2}$$

$$= \Big(\mathrm{op}_\varepsilon\Big(b^-\Big(z,s,\zeta,\tfrac{\tilde z}{R\sqrt\varepsilon}\Big)\Big)\, v_1^\varepsilon(z,s), v_1^\varepsilon(z,s)\Big)_{L^2} + O(R\sqrt\varepsilon),$$

$$\Big(\mathrm{op}_\varepsilon\Big((\mathcal{L}^{\delta_t}_{\varepsilon,R} f^-_{c_{\varepsilon,R}})(q,t,p,\tau,+)\Pi^+(q)\Big)\,\psi^\varepsilon(q,t), \psi^\varepsilon(q,t)\Big)_{L^2}$$

$$= \Big(\mathrm{op}_\varepsilon\Big(\widetilde T^\varepsilon(\tilde z)\, b^-\Big(z,s+s_1^-,\zeta,\tfrac{\tilde z}{R\sqrt\varepsilon}\Big)\Big)\, v_1^\varepsilon(z,s), v_1^\varepsilon(z,s)\Big)_{L^2} + O(R^3\sqrt\varepsilon).$$

Proof. We prove only the first two equalities, since one deals similarly with the other two. By symbolic calculus and the transformation property (5.4), the canonical transform κ^{-1} of Theorem 5.2 acts as

$$\Big(\mathrm{op}_\varepsilon\Big(c^+_{\varepsilon,R}(q,t,p,\tau)\Pi^+(q)\Big)\,\psi^\varepsilon(q,t), \psi^\varepsilon(q,t)\Big)_{L^2}$$

$$= \Big(\mathrm{op}_\varepsilon\Big(\big((c^+_{\varepsilon,R}A_0^*\Pi^+A_0)\circ\kappa^{-1}\big)(z,s,\zeta,\sigma)\Big)\, v^\varepsilon(z,s)\ ,\ v^\varepsilon(z,s)\Big)_{L^2} + O(\sqrt\varepsilon).$$

The compactly supported function $c^+_{\varepsilon,R}\circ\kappa^{-1}$ is localized near $J^{+,out}$, that is, near $\{\sigma+s=0,\ \tilde z=0,\ s>0\}$. The relation (5.11) between the projectors gives a function $b\in\mathcal{C}^\infty_c(\mathbb{R}^{2d+2+\ell},\mathbb{C})$ compactly supported in $\{s>0\}$ such that

$$\big((c^+_{\varepsilon,R}A_0^*\Pi^+A_0)\circ\kappa^{-1}\big)(z,s,\zeta,\sigma)$$

$$= b\Big(z,s,\zeta,\sigma,\tfrac{\tilde z}{R\sqrt\varepsilon},\tfrac{\tilde\lambda^+(z,s,\zeta,\sigma)}{R\sqrt\varepsilon}\Big)\,\widetilde\Pi^+(z,s,\zeta)+O(R\sqrt\varepsilon)$$

$$=: b_{\varepsilon,R}(z,s,\zeta,\sigma)\widetilde\Pi^+(z,s,\zeta)+O(R\sqrt\varepsilon)$$

as functions in $\mathcal{C}^\infty_c(\mathbb{R}^{2d+2},\mathbb{C})$. Hence, we obtain

$$\Big(\mathrm{op}_\varepsilon\Big(c^+_{\varepsilon,R}(q,t,p,\tau)\Pi^+(q)\Big)\,\psi^\varepsilon(q,t)\ ,\ \psi^\varepsilon(q,t)\Big)_{L^2}$$

$$= \Big(\mathrm{op}_\varepsilon\Big(b_{\varepsilon,R}(z,s,\zeta,\sigma)\widetilde\Pi^+(z,s,\zeta)\Big)\, v^\varepsilon(z,s)\ ,\ v^\varepsilon(z,s)\Big)_{L^2} + O(R\sqrt\varepsilon).$$

For $|\tilde z|=O(R\sqrt\varepsilon)$ we have

$$\widetilde\Pi^+(z,s,\zeta)=\begin{pmatrix}0&0\\0&\mathrm{Id}\end{pmatrix}+O(R\sqrt\varepsilon)\ \text{ in }\ \{s>0\},$$

$$\widetilde\Pi^+(z,s,\zeta)=\begin{pmatrix}\mathrm{Id}&0\\0&0\end{pmatrix}+O(R\sqrt\varepsilon)\ \text{ in }\ \{s<0\}, \tag{5.13}$$

and therefore

$$\Big(\mathrm{op}_\varepsilon\Big(b_{\varepsilon,R}(z,s,\zeta,\sigma)\widetilde{\Pi}^+(z,s,\zeta)\Big)\,v^\varepsilon(z,s)\,,\,v^\varepsilon(z,s)\Big)_{L^2}$$
$$=\left(\mathrm{op}_\varepsilon\left(b_{\varepsilon,R}(z,s,\zeta,\sigma)\right)v_2^\varepsilon(z,s)\,,\,v_2^\varepsilon(z,s)\right)_{L^2}+O(R\sqrt{\varepsilon}).$$

We now remove the σ-dependence of the symbol. Taylor expanding around $\sigma=-\sqrt{s^2+|\tilde z+\gamma_0(z,\zeta)|^2}$, we write

$$b_{\varepsilon,R}(z,s,\zeta,\sigma)=b\Big(z,s,\zeta,-\sqrt{s^2+|\tilde z+\gamma_0(z,\zeta)|^2},\tfrac{\tilde z}{R\sqrt{\varepsilon}},0\Big)$$
$$+\tfrac{1}{R\sqrt{\varepsilon}}\tilde\lambda^+(s,z,\sigma,\zeta)\,G\Big(z,s,\zeta,\sigma,\tfrac{\tilde z}{R\sqrt{\varepsilon}},\tfrac{\tilde\lambda^+(z,s,\sigma,\zeta)}{R\sqrt{\varepsilon}}\Big)$$

with $G\in\mathcal{C}_c^\infty(\mathbb{R}^{2d+2+\ell},\mathbb{C})$. Since

$$\tilde\lambda^+(z,s,\zeta,\sigma)\tilde\Pi^+(z,s,\zeta)=\widetilde{\Pi}^+(z,s,\zeta)\,(\sigma-V_\ell(s,\tilde z+\gamma_0(z,\zeta))\,,$$

and since v^ε solves the Landau–Zener-type problem (5.5), an argument analogous to the proof of Lemma 5.1 yields

$$\left(\mathrm{op}_\varepsilon\left(b_{\varepsilon,R}(z,s,\zeta,\sigma)\right)v_2^\varepsilon(z,s)\,,\,v_2^\varepsilon(z,s)\right)_{L^2}=O\big(\tfrac{1}{R^2}\big)+O(\sqrt{\varepsilon})$$
$$+\Big(\mathrm{op}_\varepsilon\Big(b\big(z,s,\zeta,-\sqrt{s^2+|\tilde z+\gamma_0(z,\zeta)|^2},\tfrac{\tilde z}{R\sqrt{\varepsilon}},0\big)\Big)\,v_2^\varepsilon(z,s)\,,\,v_2^\varepsilon(z,s)\Big)_{L^2}.$$

Setting $b^+(z,s,\zeta,\eta)=b(z,s,\zeta,-s,\eta,0)$, we obtain

$$\Big(\mathrm{op}_\varepsilon\Big(c^+_{\varepsilon,R}(q,t,p,\tau)\Pi^+(q)\Big)\,\psi^\varepsilon(q,t)\,,\,\psi^\varepsilon(q,t)\Big)_{L^2}$$
$$=\Big(\mathrm{op}_\varepsilon\Big(b^+\Big(z,s,\zeta,\tfrac{\tilde z}{R\sqrt{\varepsilon}}\Big)\Big)\,v_2^\varepsilon(z,s)\,,\,v_2^\varepsilon(z,s)\Big)_{L^2}+O(R\sqrt{\varepsilon})+O\big(\tfrac{1}{R^2}\big).$$

Next, we focus on the second claimed identity, which contains nonadiabatic transitions. We have

$$\Big(\mathrm{op}_\varepsilon\Big((\mathcal{L}^{\delta_t}_{\varepsilon,R}f^+_{c_{\varepsilon,R}})(q,t,p,\tau,+)\Pi^+(q)\Big)\,\psi^\varepsilon(q,t)\,,\,\psi^\varepsilon(q,t)\Big)_{L^2}$$
$$=\Big(\mathrm{op}_\varepsilon\Big(\big((\mathcal{L}^{\delta_t}_{\varepsilon,R}f^+_{c_{\varepsilon,R}})A_0^*\Pi^+A_0\big)(\kappa^{-1}(z,s,\zeta,\sigma),+)\Big)\,v^\varepsilon(z,s)\,,\,v^\varepsilon(z,s)\Big)_{L^2}+O(\sqrt{\varepsilon}).$$

Let $r\mapsto(z^+(r),s^+(r),\zeta^+(r),\sigma^+(r))$ be the Hamiltonian trajectory of

$$\dot z=\partial_\zeta\tilde\lambda^+={}^t\mathrm{d}_\zeta\gamma_0(z,\zeta)(\tilde z+\gamma_0(z,\zeta))\,(s^2+|\tilde z+\gamma_0(z,\zeta)|^2)^{-1/2},$$
$$\dot s=\partial_\sigma\tilde\lambda^+=1,$$
$$\dot\zeta=-\partial_z\tilde\lambda^+=-{}^t\mathrm{d}_z(\tilde z+\gamma_0(z,\zeta))(\tilde z+\gamma_0(z,\zeta))(s^2+|\tilde z+\gamma_0(z,\zeta)|^2)^{-1/2},$$
$$\dot\sigma=-\partial_s\tilde\lambda^+=-s\,(s^2+|\tilde z+\gamma_0(z,\zeta)|^2)^{-1/2}$$

with $(z(0),s(0),\zeta(0),\sigma(0))=(z,s,\zeta,\sigma)=\kappa(q,t,p,\tau)$. A trajectory jumps for $r=r_*$ if $s^+(r_*)=O(|\phi(q)|^2)=O(R^2\varepsilon)$. Then, $\sigma^+(r_*)=O(R^2\varepsilon)$ as well. Since

$$\tfrac{\mathrm{d}}{\mathrm{d}r}\big(s^+(r)+\sigma^+(r)\big)=O(|\tilde z|^2)\quad\text{on}\quad J^{+,out}=\{\sigma+s=0,\,\tilde z=0,\,s>0\},$$

and since $|\tilde z| = O(R\sqrt{\varepsilon})$ on the support of our symbol, we have

$$z^+(r) = z + O(R^3\varepsilon^{3/2}), \quad s^+(r) = s + r,$$
$$\zeta^+(r) = \zeta + O(R\sqrt{\varepsilon}), \quad \sigma^+(r) = -s^+(r) + O(R^2\varepsilon).$$

These asymptotics together with conservation of energy along Hamiltonian flows yield

$$b\Big(z^+(r), s^+(r), \zeta^+(r), \sigma^+(r), \tfrac{\tilde z^+(r)}{R\sqrt{\varepsilon}}, \tfrac{\tilde\lambda^+(z^+(r),s^+(r),\zeta^+(r),\sigma^+(r))}{R\sqrt{\varepsilon}}\Big)$$
$$= b\Big(z, s+r, \zeta, -(s+r), \tfrac{\tilde z}{R\sqrt{\varepsilon}}, \tfrac{\tilde\lambda^+(z,s,\zeta,\sigma)}{R\sqrt{\varepsilon}}\Big) + O(R\sqrt{\varepsilon}).$$

Jumps occur for $|\phi(q)| \le R\sqrt{\varepsilon}$, that is, for

$$|\tilde z|^2 = |\mathrm{d}\phi(q)p|^{-1}|\pi_\ell(q,p)\phi(q)|^2 + O(|\phi(q)|^3)$$
$$= |\mathrm{d}\phi(q)p|^{-1}|\pi_\ell(q,p)\phi(q)|^2 + O(R^3\varepsilon^{3/2}).$$

Therefore, the transition rate $T^\varepsilon(q,p)$ reads in the new coordinates as

$$T^\varepsilon(q,p) = \exp\big(-\tfrac{\pi}{\varepsilon}|\mathrm{d}\phi(q)p|^{-1}|\pi_\ell(q,p)\phi(q)|^2\big)$$
$$= \exp\big(-\tfrac{\pi}{\varepsilon}|\tilde z|^2\big) + O(R^3\sqrt{\varepsilon}) = \widetilde T^\varepsilon(\tilde z) + O(R^3\sqrt{\varepsilon}),$$

and there exists $s_1^+ \in \mathbb{R}$ such that

$$\big((\mathcal{L}^{\delta_t}_{\varepsilon,R} f^+_{c_{\varepsilon,R}})A_0^*\Pi^+A_0\big)(\kappa^{-1}(z,s,\zeta,\sigma),+)$$
$$= \Big(1 - \widetilde T^\varepsilon(\tilde z)\Big)\, b\Big(z, s+s_1^+, \zeta, -(s+s_1^\pm), \tfrac{\tilde z}{R\sqrt{\varepsilon}}, \tfrac{\tilde\lambda^+(z,s,\zeta,\sigma)}{R\sqrt{\varepsilon}}\Big)\, \widetilde\Pi^+(z,s,\zeta) + O(R^3\sqrt{\varepsilon}).$$

By the asymptotics (5.13) of $\widetilde\Pi^+$ above $\{s<0\}$, we then get

$$\Big(\mathrm{op}_\varepsilon\Big(\big((\mathcal{L}^{\delta_t}_{\varepsilon,R} f^+_{c_{\varepsilon,R}})A_0^*\Pi^+A_0\big)(\kappa^{-1}(z,s,\zeta,\sigma),+)\Big)\, v^\varepsilon(z,s)\ ,\ v^\varepsilon(z,s)\Big)_{L^2} = O(R^3\sqrt{\varepsilon})$$
$$+ \Big(\mathrm{op}_\varepsilon\Big(\big(1-\widetilde T^\varepsilon(\tilde z)\big) b\Big(z, s+s_1^+, \zeta, -(s+s_1^\pm), \tfrac{\tilde z}{R\sqrt{\varepsilon}}, \tfrac{\tilde\lambda^+(z,s,\zeta,\sigma)}{R\sqrt{\varepsilon}}\Big)\Big)\, v_1^\varepsilon(z,s), v_1^\varepsilon(z,s)\Big)_{L^2}.$$

As before, we remove the σ-dependence of the symbol and obtain

$$\Big(\mathrm{op}_\varepsilon\Big((\mathcal{L}^{\delta_t}_{\varepsilon,R} f^+_{c_{\varepsilon,R}})(q,t,p,\tau,+)\Pi^+(q)\Big)\, \psi^\varepsilon(q,t)\ ,\ \psi^\varepsilon(q,t)\Big)_{L^2}$$
$$= \Big(\mathrm{op}_\varepsilon\Big(\big(1-\widetilde T^\varepsilon(\tilde z)\big) b^+\Big(z, s+s_1^+, \zeta, \tfrac{\tilde z}{R\sqrt{\varepsilon}}\Big)\Big)\, v_1^\varepsilon(z,s)\ ,\ v_1^\varepsilon(z,s)\Big)_{L^2} + O(R^3\sqrt{\varepsilon}). \qquad \square$$

5.3. The transitions. It remains to analyze $v^\varepsilon = U^*\mathrm{op}_\varepsilon(A_\varepsilon^{-1})\psi^\varepsilon$ for proving the equality of each pair in Proposition 5.4 up to an error of $O(1/R^2) + O(R^3\sqrt{\varepsilon}) + O(\sqrt{\varepsilon}|\ln\varepsilon|) + O(1/(R^5\sqrt{\varepsilon}))$, concluding the proof of our main result. We consider the solution u^ε of

$$-i\varepsilon\partial_s u^\varepsilon = \begin{pmatrix} s\,\mathrm{Id} & \sqrt{\varepsilon}\,G \\ \sqrt{\varepsilon}G^* & -s\,\mathrm{Id}\end{pmatrix} u^\varepsilon, \qquad u^\varepsilon_{|s=0} = v^\varepsilon_{|s=0}, \tag{5.14}$$

where G is one of the operators

$$G_2 = \tfrac{1}{\sqrt{\varepsilon}}\varphi\left(\tfrac{\tilde z}{R\sqrt{\varepsilon}}\right)\tilde z, \qquad G_3 = \tfrac{1}{\sqrt{\varepsilon}}\varphi\left(\tfrac{\tilde z}{R\sqrt{\varepsilon}}\right)(Z_1 + iZ_2),$$

$$G_5 = \tfrac{1}{\sqrt{\varepsilon}}\varphi\left(\tfrac{\tilde z}{R\sqrt{\varepsilon}}\right)\begin{pmatrix} Z_1+iZ_2 & Z_3+iZ_4 \\ -Z_3+iZ_4 & Z_1-iZ_2\end{pmatrix}$$

with $Z = \mathrm{op}_\varepsilon(\tilde z + \gamma_\varepsilon(z,\zeta))$ and $\varphi \in \mathcal{C}_c^\infty(\mathbb{R}^{\ell-1},\mathbb{R})$. In all three cases, G is a bounded operator on $L^2(\mathbb{R}^d)$ with $\|G\| = O(R)$, and we have

$$\|u^\varepsilon - v^\varepsilon\|_{L^2_{loc}(\mathbb{R}^{d+1})} = O(\varepsilon^\infty).$$

The following Landau–Zener-type formula is given in Proposition 7 of [8], up to the explicit error terms.

PROPOSITION 5.5. *Let u^ε be the solution of* (5.14). *There exist vector-valued functions $\alpha^\varepsilon = (\alpha_1^\varepsilon, \alpha_2^\varepsilon)$, $\omega^\varepsilon = (\omega_1^\varepsilon, \omega_2^\varepsilon) \in L^2(\mathbb{R}^d, \mathbb{C}^{N(\ell)})$ such that for any function $\chi \in \mathcal{C}_c^\infty(\{x \in \mathbb{R} \mid |x| \le R^2\},\mathbb{R})$ the families $(\chi(GG^*)\alpha_1^\varepsilon)_{\varepsilon>0}$, $(\chi(G^*G)\alpha_2^\varepsilon)_{\varepsilon>0}$, $(\chi(GG^*)\omega_1^\varepsilon)_{\varepsilon>0}$, $(\chi(G^*G)\omega_2^\varepsilon)_{\varepsilon>0}$ are bounded in $L^2(\mathbb{R}^d,\mathbb{C})$ and satisfy*

$$\chi(GG^*)u_1^\varepsilon(z,s) = \chi(GG^*)\mathrm{e}^{is^2/(2\varepsilon)}\left|\tfrac{s}{\sqrt{\varepsilon}}\right|^{i\frac{GG^*}{2}} k_1^\varepsilon(z) + O(R^2\sqrt{\varepsilon}),$$

$$\chi(G^*G)u_2^\varepsilon(z,s) = \chi(G^*G)\mathrm{e}^{-is^2/(2\varepsilon)}\left|\tfrac{s}{\sqrt{\varepsilon}}\right|^{-i\frac{G^*G}{2}} k_2^\varepsilon(z) + O(R^2\sqrt{\varepsilon})$$

in $L^2(\mathbb{R}^d,\mathbb{C})$, where $k_j^\varepsilon = \alpha_j^\varepsilon$ and $k_j^\varepsilon = \omega_j^\varepsilon$ for $s<0$ and $s>0$, respectively, $j \in \{1,2\}$. Moreover,

$$\begin{pmatrix}\omega_1^\varepsilon \\ \omega_2^\varepsilon\end{pmatrix} = \begin{pmatrix} a(GG^*) & -\overline{b}(GG^*)G \\ b(G^*G)G^* & a(G^*G)\end{pmatrix}\begin{pmatrix}\alpha_1^\varepsilon \\ \alpha_2^\varepsilon\end{pmatrix} \tag{5.15}$$

with

$$a(\lambda) = \mathrm{e}^{-\pi\lambda/2}, \qquad b(\lambda) = \tfrac{2i\mathrm{e}^{i\pi/4}}{\lambda\sqrt{\pi}}2^{-i\lambda/2}\mathrm{e}^{-\pi\lambda/4}\Gamma(1+i\tfrac{\lambda}{2})\sinh(\tfrac{\pi\lambda}{2}).$$

Proof. Lemma 7 in [8] is the crucial step in the proof of Proposition 7 for which we have to check that the leading order error estimate is indeed $O(R^2\sqrt{\varepsilon})$. For this, we turn to the explicit calculations in the proof of Lemma 11 in [7] and study the two integrals

$$A_0 = s^{-1+i\eta^2/2}\mathrm{e}^{-is^2/2}\int_{\mathbb{R}}\tilde\chi\left(\sqrt{2}-\sqrt{2-\tfrac{2z}{s^2}}\right)\left|\sqrt{2}-\sqrt{2-\tfrac{2z}{s^2}}\right|^{i\eta^2/2}\frac{\mathrm{e}^{iz}}{\sqrt{2-\frac{2z}{s^2}}}\,\mathrm{d}z,$$

$$B_0 = s^{1+i\eta^2/2}\int_{\mathbb{R}}(1-\tilde\chi(y))\,\mathrm{e}^{-\frac{i}{2}s^2(1+y^2-2\sqrt{2}y)}|y|^{i\eta^2/2}\mathrm{d}y,$$

where $\tilde\chi \in \mathcal{C}_c^\infty(\mathbb{R},\mathbb{R})$ is a function with $0 \le \tilde\chi \le 1$, $\tilde\chi(y) = 0$ for $|y| \ge \sqrt{2}/2$, and $\tilde\chi(y) = 1$ for $|y| \le \sqrt{2}/4$. The phase function $y \mapsto -\frac{1}{2}(1+y^2-2\sqrt{2}y)$ of B_0 has the stationary point $y = \sqrt{2}$, and Taylor expansion of $y \mapsto \ln|y|$ around $y = \sqrt{2}$ yields

$$B_0 = \sqrt{2\pi}\mathrm{e}^{-i\pi/4}2^{i\eta^2/4}s^{i\eta^2/2}\mathrm{e}^{is^2/2} + O(\eta^2 s^{-2}),$$

while integration by parts gives

$$A_0 = O(\eta^2 s^{-1})$$

as $\eta, s \to \infty$. Since the asymptotics of the other relevant integrals can be obtained analogously, the claimed error estimates follow by setting $s = O(\varepsilon^{-1/2})$ and $\eta = O(R)$. □

For implementing these Landau–Zener asymptotics, we need the following additional relations, which are literally contained in the proofs of Lemmas 8 and 9 in [8].

LEMMA 5.6. *For any $\chi \in \mathcal{C}_c^\infty(\mathbb{R}, \mathbb{R})$ and $b \in \mathcal{C}_c^\infty(\mathbb{R}^{2d+N(\ell)-1}, \mathbb{C})$, we have*

$$\chi\Big(\tfrac{|\tilde z|^2}{\varepsilon}\Big) = \chi(GG^*) + O(\sqrt{\varepsilon}) = \chi(G^*G) + O(\sqrt{\varepsilon}),$$

$$\Big|\tfrac{s}{\sqrt{\varepsilon}}\Big|^{\pm i\frac{G^*G}{2}} \operatorname{op}_\varepsilon\Big(b\Big(z, \zeta, \tfrac{\tilde z}{R\sqrt{\varepsilon}}\Big)\Big) \Big|\tfrac{s}{\sqrt{\varepsilon}}\Big|^{\mp i\frac{G^*G}{2}} = \operatorname{op}_\varepsilon\Big(b\Big(z, \zeta, \tfrac{\tilde z}{R\sqrt{\varepsilon}}\Big)\Big) + O(\sqrt{\varepsilon}|\ln\varepsilon|).$$

Now, we are ready to conclude the proof of Theorem 2.2. We discuss only the first pair of terms in Proposition 5.4, since the other pair can be dealt with analogously. We set

$$I^1_{\varepsilon,R} = \Big(\operatorname{op}_\varepsilon\Big(b^+\Big(z, s, \zeta, \tfrac{\tilde z}{R\sqrt{\varepsilon}}\Big)\Big)\, u_2^\varepsilon(z,s), u_2^\varepsilon(z,s)\Big)_{L^2},$$

$$I^2_{\varepsilon,R} = \Big(\operatorname{op}_\varepsilon\Big((1-\widetilde T^\varepsilon(\tilde z))b^+\Big(z, s+s_1^+, \zeta, \tfrac{\tilde z}{R\sqrt{\varepsilon}}\Big)\Big)\, u_1^\varepsilon(z,s), u_1^\varepsilon(z,s)\Big)_{L^2}.$$

By Lemma 5.6 and Proposition 5.5, we have

$$\begin{aligned} I^1_{\varepsilon,R} &= \Big(\operatorname{op}_\varepsilon\Big(b^+\Big(z, s, \zeta, \tfrac{\tilde z}{R\sqrt{\varepsilon}}\Big)\Big)\, \chi(G^*G)u_2^\varepsilon(z,s), \chi(G^*G)u_2^\varepsilon(z,s)\Big)_{L^2} + O(\sqrt{\varepsilon}) \\ &= \left(\operatorname{op}_\varepsilon\Big(b^+\Big(z, s, \zeta, \tfrac{\tilde z}{R\sqrt{\varepsilon}}\Big)\Big)\, \mathrm{e}^{-i\frac{s^2}{2\varepsilon}} \Big|\tfrac{s}{\sqrt{\varepsilon}}\Big|^{-i\frac{G^*G}{2}} \omega_2^\varepsilon(z), \mathrm{e}^{-i\frac{s^2}{2\varepsilon}} \Big|\tfrac{s}{\sqrt{\varepsilon}}\Big|^{-i\frac{G^*G}{2}} \omega_2^\varepsilon(z)\right)_{L^2} \\ &\qquad + O(R^2\sqrt{\varepsilon}) \\ &= \Big(\operatorname{op}_\varepsilon\Big(b^+\Big(z, s, \zeta, \tfrac{\tilde z}{R\sqrt{\varepsilon}}\Big)\Big)\, \omega_2^\varepsilon(z), \omega_2^\varepsilon(z)\Big)_{L^2} + O(R^2\sqrt{\varepsilon}) + O(\sqrt{\varepsilon}|\ln\varepsilon|) \\ &= \Big(\operatorname{op}_\varepsilon\Big(b^+\Big(z, s+s_1^+, \zeta, \tfrac{\tilde z}{R\sqrt{\varepsilon}}\Big)\Big)\, \omega_2^\varepsilon(z), \omega_2^\varepsilon(z)\Big)_{L^2} + O(R^2\sqrt{\varepsilon}) + O(\sqrt{\varepsilon}|\ln\varepsilon|) \end{aligned}$$

because $s_1^\pm = O(R\sqrt{\varepsilon})$ and analogously

$$\begin{aligned} I^2_{\varepsilon,R} = \Big(\operatorname{op}_\varepsilon\Big((1-\widetilde T^\varepsilon(\tilde z))\, b^+\Big(z, s+s_1^+, \zeta, \tfrac{\tilde z}{R\sqrt{\varepsilon}}\Big)\Big)\, \alpha_1^\varepsilon(z), \alpha_1^\varepsilon(z)\Big)_{L^2} \\ + O(R^2\sqrt{\varepsilon}) + O(\sqrt{\varepsilon}|\ln\varepsilon|). \end{aligned}$$

By the scattering identity (5.15), we have $\omega_2^\varepsilon = b(G^*G)G^*\alpha_1^\varepsilon + a(G^*G)\alpha_2^\varepsilon$ and

$$\begin{aligned} I^1_{\varepsilon,R} = \Big(\operatorname{op}_\varepsilon\Big(b^+\Big(z, s+s_1^+, \zeta, \tfrac{\tilde z}{R\sqrt{\varepsilon}}\Big)\Big)\, (b(G^*G)G^*\alpha_1^\varepsilon(z) + a(G^*G)\alpha_2^\varepsilon(z)), \\ b(G^*G)G^*\alpha_1^\varepsilon(z) + a(G^*G)\alpha_2^\varepsilon(z)\Big)_{L^2} + O(R^2\sqrt{\varepsilon}) + O(\sqrt{\varepsilon}|\ln\varepsilon|). \end{aligned}$$

Since the wave function $\psi^\varepsilon(q,t)$ is of order $O(1/R^2)+O(\sqrt{\varepsilon})+O(1/(R^5\sqrt{\varepsilon}))$ near the set $J^{-,in}=\{\sigma+s=0,\,\tilde z=0,\,s<0\}$, we have

$$\begin{pmatrix} 0 \\ v_2^\varepsilon(z,s)\end{pmatrix} = \mathrm{op}_\varepsilon\Big(\widetilde\Pi^-(z,s,\zeta)\Big)\, v^\varepsilon(z,s)+O(R\sqrt{\varepsilon})$$
$$= O(1/R^2)+O(R\sqrt{\varepsilon})+O(1/(R^5\sqrt{\varepsilon}))$$

as functions in $L^2_{loc}(\mathbb{R}^{d+1})$ localized near $J^{-,in}$. The preceding arguments expressing $I^1_{\varepsilon,R}$ and $I^2_{\varepsilon,R}$ in terms of α^ε and ω^ε then yield

$$a(G^*G)\alpha_2^\varepsilon(z)=O(1/R^2)+O(R^2\sqrt{\varepsilon})+O(\sqrt{\varepsilon}|\ln\varepsilon|)+O(1/(R^5\sqrt{\varepsilon}))$$

near $J^{-,in}$, and hence

$$I^1_{\varepsilon,R}=\Big(\mathrm{op}_\varepsilon\Big(b^+\Big(z,s+s_1^+,\zeta,\tfrac{\tilde z}{R\sqrt{\varepsilon}}\Big)\Big)\, b(G^*G)G^*\alpha_1^\varepsilon(z), b(G^*G)G^*\alpha_1^\varepsilon(z)\Big)_{L^2}$$
$$+O\big(\tfrac{1}{R^2}\big)+O(R^2\sqrt{\varepsilon})+O(\sqrt{\varepsilon}|\ln\varepsilon|)+O\Big(\tfrac{1}{R^5\sqrt{\varepsilon}}\Big).$$

Lemma 5.6 together with the relations $G^*b(GG^*)=b(G^*G)G^*$ and

$$\lambda|b(\lambda)|^2=1-\mathrm{e}^{-\pi\lambda}$$

implies

$$G\,\overline{b}(G^*G)\,\mathrm{op}_\varepsilon\Big(b^+\Big(z,s+s_1^+,\zeta,\tfrac{\tilde z}{R\sqrt{\varepsilon}}\Big)\Big)\, b(G^*G)G^*\alpha_1^\varepsilon(z)$$
$$= G\,\overline{b}(G^*G)b(G^*G)G^*\,\mathrm{op}_\varepsilon\Big(b^+\Big(z,s+s_1^+,\zeta,\tfrac{\tilde z}{R\sqrt{\varepsilon}}\Big)\Big)\,\alpha_1^\varepsilon(z)+O(\sqrt{\varepsilon})$$
$$= GG^*\overline{b}(GG^*)b(GG^*)\,\mathrm{op}_\varepsilon\Big(b^+\Big(z,s+s_1^+,\zeta,\tfrac{\tilde z}{R\sqrt{\varepsilon}}\Big)\Big)\,\alpha_1^\varepsilon(z)+O(\sqrt{\varepsilon})$$
$$= (1-\widetilde T^\varepsilon(\tilde z))\,\mathrm{op}_\varepsilon\Big(b^+\Big(z,s+s_1^+,\zeta,\tfrac{\tilde z}{R\sqrt{\varepsilon}}\Big)\Big)\,\alpha_1^\varepsilon(z)+O(\sqrt{\varepsilon}) \tag{5.16}$$

and finally

$$I^1_{\varepsilon,R}=I^2_{\varepsilon,R}+O(1/R^2)+O(R^2\sqrt{\varepsilon})+O(\sqrt{\varepsilon}|\ln\varepsilon|)+O(1/(R^5\sqrt{\varepsilon})).$$

6. Eigenvalues of multiplicity two. We have assumed that the matrix-valued observable a is V-diagonal in the sense that $a=a^+\Pi^++a^-\Pi^-$ with scalar-valued functions $a^\pm$. The more natural assumption, that a commutes with V,

$$a=\Pi^+a\Pi^++\Pi^-a\Pi^-,$$

does not change the situation in the case $\ell=2,3$ but enlarges the class of observables for $\ell=3',5$. For observables of this form, one has to modify the Markov process to account for a polarization effect. The state space requires an additional component $w\in\mathbb{C}^4$, and when the deterministic flow $\Phi_j^t(q,p)$ hits the jump manifold $S_{\varepsilon,R}$ in a point (q^*,p^*) a more general branching occurs. The state (q^*,p^*,j,w) changes with probability $T^\varepsilon(q^*,p^*)$ to $(q^*,p^*,-j,w)$ and with probability $1-T^\varepsilon(q^*,p^*)$ to $(q^*,p^*,j,\mathcal{R}(q^*,p^*)w)$, where

$$\mathcal{R}(q,p)=V_\ell\Big(\tfrac{\pi_\ell(q,p)\phi(q)}{|\pi_\ell(q,p)\phi(q)|}\Big).$$

This phenomenon is also described in Theorem 1 of [5] for two-scale Wigner measures. Our main result, Theorem 2.2, for the propagation of Wigner functions still applies for the semigroup, which incorporates polarization.

Proof of Theorem 2.2 *for* $a = \Pi^+ a \Pi^+ + \Pi^- a \Pi^-$ *if* $\ell = 3', 5$. Let us first prove classical transport. We set $A^+ = \Pi^+ a \Pi^+$ and focus on the + mode. We extensively use $\Pi^+ A^+ = A^+ \Pi^+ = A^+$. The strategy is similar to the one of section 4, and we have to focus on the Poisson brackets

$$\tfrac{1}{2}\{A^+(q,p), \tau + \tfrac{1}{2}|p|^2 + v(q) + V_\ell(\phi(q))\} - \tfrac{1}{2}\{\tau + \tfrac{1}{2}|p|^2 + v(q) + V_\ell(\phi(q)), A^+(q,p)\}.$$

We set $\mu(q,p,\tau) = \tau + \frac{1}{2}|p|^2 + v(q)$ and write

$$\{A^+, \mu\} = \Pi^+\{A^+, \mu\}\Pi^+ + A^+\{\Pi^+, \mu\} + \{\Pi^+, \mu\}A^+.$$

We observe that

$$r_0 = A^+\{\Pi^+, \mu\} + \{\Pi^+, \mu\}A^+ = -A^+(\nabla_q \Pi^+ \cdot p) - (\nabla_q \Pi^+ \cdot p)A^+$$

can be treated as in section 4. Indeed, since A^+ commutes with $V_\ell(\phi)$, one has for any matrix G that $A^+[V_\ell(\phi), G] = [V_\ell(\phi), A^+ G]$, $[V_\ell(\phi), G]A^+ = [V_\ell(\phi), GA^+]$, and consequently

$$\begin{aligned} r_0 &= -\tfrac{1}{4|\phi|^3}\left[V_\ell(\phi), \left[V_\ell(\phi), A^+ V_\ell(\mathrm{d}\phi p)\right]\right] - \tfrac{1}{4|\phi|^3}\left[V_\ell(\phi), \left[V_\ell(\phi), V_\ell(\mathrm{d}\phi p)A^+\right]\right] \\ &= -\left[\mu + V_\ell(\phi), \tfrac{1}{4|\phi|^3}\left[V_\ell(\phi), A^+ V_\ell(\mathrm{d}\phi p)\right]\right] - \left[\mu + V_\ell(\phi), \tfrac{1}{4|\phi|^3}\left[V_\ell(\phi), V_\ell(\mathrm{d}\phi p)A^+\right]\right]. \end{aligned}$$

The most harmful of the arising terms contains the brackets with $\frac{1}{2}|p|^2$, that is,

$$\tilde{r}_0 = -\left\{\tfrac{1}{2}|p|^2, \tfrac{1}{4|\phi|^3}\left[V_\ell(\phi), A^+ V_\ell(\mathrm{d}\phi p)\right]\right\} - \left\{\tfrac{1}{2}|p|^2, \tfrac{1}{4|\phi|^3}\left[V_\ell(\phi), V_\ell(\mathrm{d}\phi p)A^+\right]\right\}.$$

Since the term containing the derivatives of $V_\ell(\phi)$ vanishes,

$$-\tfrac{1}{4|\phi|^3}\left[V_\ell(\mathrm{d}\phi p), A^+ V_\ell(\mathrm{d}\phi p)\right] - \tfrac{1}{4|\phi|^3}\left[V_\ell(\mathrm{d}\phi p), V_\ell(\mathrm{d}\phi p)A^+\right] = 0,$$

there is a matrix-valued function G_ε with suitable bounds on its derivatives, such that $\tilde{r}_0 = |\phi|^{-4}[V_\ell(\phi), G_\varepsilon]$, and hence the other arguments of Lemma 4.1 apply for the analysis of r_0.

For the brackets with the matrix part, we write

$$\tfrac{1}{2}\{A^+, V_\ell(\phi)\} - \tfrac{1}{2}\{V_\ell(\phi), A^+\} = \Pi^+\{A^+, |\phi|\}\Pi^+ + |\phi|\left(\{A^+, \Pi^+\} - \{\Pi^+, A^+\}\right).$$

The second part,

$$r_1 = |\phi|\left(\{A^+, \Pi^+\} - \{\Pi^+, A^+\}\right) = |\phi|\left(\nabla_p A^+ \cdot \nabla_q \Pi^+ + \nabla_q \Pi^+ \nabla_p A^+\right),$$

is off-diagonal with respect to V, since $\nabla_p A^+ = \Pi^+ \nabla_p A^+ = \nabla_p A^+ \Pi^+$, $\Pi^\pm \Pi^\mp = 0$, and $\Pi^\pm \nabla_q \Pi^+ \Pi^\pm = 0$ imply

$$\Pi^\pm\left(\nabla_p A^+ \cdot \nabla_q \Pi^+ + \nabla_q \Pi^+ \nabla_p A^+\right)\Pi^\pm = 0.$$

Hence, Lemma 4.1 applies.

The importance of $\mathcal{R}(q,p)$ for the nonadiabatic transitions becomes clear, when recasting (5.16) in the previous section as

$$\begin{pmatrix} 0 & G^* \\ G & 0 \end{pmatrix} \overline{b}(G^*G)\,\mathrm{op}_\varepsilon\Big(b^+\Big(z, s+s_1^+, \zeta, \tfrac{\tilde z}{R\sqrt{\varepsilon}}\Big)\Big)\, b(G^*G) \begin{pmatrix} 0 & G \\ G^* & 0 \end{pmatrix} \begin{pmatrix} \alpha_1^\varepsilon(z) \\ 0 \end{pmatrix}$$

$$= V_\ell^*(0, \tfrac{\tilde z}{|\tilde z|})(G^*G)^{\frac12}\, \overline{b}(G^*G)\,\mathrm{op}_\varepsilon\Big(b^+\Big(z, s+s_1^+, \zeta, \tfrac{\tilde z}{R\sqrt{\varepsilon}}\Big)\Big)$$

$$\times\, b(G^*G)(G^*G)^{\frac12}\, V_\ell^*\left(0, \tfrac{\tilde z}{|\tilde z|}\right) \begin{pmatrix} \alpha_1^\varepsilon(z) \\ 0 \end{pmatrix} + O(\sqrt{\varepsilon})$$

$$= (1-\widetilde{T}^\varepsilon(\tilde z))\, V_\ell^*\left(0, \tfrac{\tilde z}{|\tilde z|}\right)\, \mathrm{op}_\varepsilon\Big(b^+\Big(z, s+s_1^+, \zeta, \tfrac{\tilde z}{R\sqrt{\varepsilon}}\Big)\Big)\, V_\ell\left(0, \tfrac{\tilde z}{|\tilde z|}\right) \begin{pmatrix} \alpha_1^\varepsilon(z) \\ 0 \end{pmatrix} + O(\sqrt{\varepsilon})$$

and observing that the normal form transformation relates $V_\ell(0, \tilde z/|\tilde z|)$ and $\mathcal{R}(q,p)$ by identity (5.9) of Theorem 5.2. ∎

Appendix A. Weyl calculus. For the convenience of the reader, we formulate the key technical lemma of the calculus of Weyl quantized pseudodifferential operators.

DEFINITION A.1. *A smooth matrix-valued function $a \in \mathcal{C}^\infty(\mathbb{R}^{2d}, \mathbb{C}^{N\times N})$ is of subquadratic growth if for all $|\alpha|+|\beta| \geq 2$ there exists $C_{\alpha,\beta} > 0$ such that*

$$\|\partial_q^\alpha \partial_p^\beta a\|_\infty \leq C_\alpha.$$

LEMMA A.2. *Let $a \in \mathcal{C}_c^\infty(\mathbb{R}^{2d}, \mathbb{C}^{N\times N})$, and let $b \in \mathcal{C}^\infty(\mathbb{R}^{2d}, \mathbb{C}^{N\times N})$ be of subquadratic growth. Then, for all $\psi \in \mathcal{C}_c^\infty(\mathbb{R}^d, \mathbb{C}^N)$*

$$\mathrm{op}_1(a)\mathrm{op}_1(b)\psi = \big(\mathrm{op}_1(ab) + \tfrac{1}{2i}\mathrm{op}_1\left(\{a,b\}\right) + \mathrm{op}_1(c) + \mathrm{op}_1(r)\big)\,\psi,$$

with $\{a,b\} = \partial_p a \partial_q b - \partial_q a \partial_p b$ the Poisson bracket, c a linear combination of

$$\partial_{q_j,p_j} a\, \partial_{q_j,p_j} b,\ \partial_{q_j}^2 a\, \partial_{p_j}^2 b,\ \partial_{p_j}^2 a\, \partial_{q_j}^2 b,$$

and $r \in \mathcal{C}^\infty(\mathbb{R}^{2d}, \mathbb{C}^{N\times N})$ such that

$$N_k(r) := \sup_{|\alpha|+|\beta|\leq k} \|\partial_q^\alpha \partial_p^\beta r\|_\infty \leq C_k \sum_{m+m'=k} \big(N_m(D^3 a)\, N_{m'}(D^3 b)\big)$$

for all $k \in \mathbb{N}$.

For a proof of this classical lemma, the reader can refer to [15] or to [3]. The theorem of Calderon and Vaillancourt implies that $\mathrm{op}_1(r)$ is a bounded operator on $L^2(\mathbb{R}^d, \mathbb{C}^N)$.

REFERENCES

[1] Y. COLIN DE VERDIÈRE, *The level crossing problem in semi-classical analysis* I. *The symmetric case*, Ann. Inst. Fourier (Grenoble), 53 (2003), pp. 1023–1054.

[2] Y. COLIN DE VERDIÈRE, *The level crossing problem in semi-classical analysis* II. *The Hermitian case*, Ann. Inst. Fourier (Grenoble), 54 (2004), pp. 1423–1441.

[3] M. DIMASSI AND J. SJÖSTRAND, *Spectral Asymptotics in the Semi-classical Limit*, London Math. Soc. Lecture Note Ser. 268, Cambridge University Press, Cambridge, UK, 1999.

[4] W. DOMCKE, D. YARKONY, AND H. KÖPPEL, EDS., *Conical Intersections*, World Scientific, Singapore, 2004.

[5] C. Fermanian Kammerer, *Wigner measures and molecular propagation through generic energy level crossings*, Rev. Math. Phys., 15 (2003), pp. 1285–1317.
[6] C. Fermanian Kammerer, *Normal forms for conical intersections in quantum chemistry*, Math. Phys. Electron. J., 13 (2007), paper 4.
[7] C. Fermanian Kammerer and P. Gérard, *Mesures semi-classiques et croisements de modes*, Bull. Soc. Math. France, 130 (2002), pp. 123–168.
[8] C. Fermanian Kammerer and P. Gérard, *A Landau-Zener formula for non-degenerated involutive codimension* 3 *crossings*, Ann. Henri Poincaré, 4 (2003), pp. 513–552.
[9] C. Fermanian Kammerer and C. Lasser, *Wigner measures and codimension two crossings*, J. Math. Phys., 44 (2003), pp. 507–527.
[10] P. Gérard, P. Markowich, N. Mauser, and F. Poupaud, *Homogenization limits and Wigner transforms*, Comm. Pure Appl. Math., 50 (1997), pp. 323–379.
[11] P. Gérard, P. Markowich, N. Mauser, and F. Poupaud, *Erratum: Homogenization limits and Wigner transforms*, Comm. Pure Appl. Math., 53 (2000), pp. 280–281.
[12] G. Hagedorn, *A time dependent Born-Oppenheimer approximation*, Comm. Math. Phys., 77 (1980), pp. 1–19.
[13] G. Hagedorn, *Molecular Propagation through Electron Energy Level Crossings*, Mem. Amer. Math. Soc. 111, AMS, Providence, RI, 1994.
[14] S. Hahn and G. Stock, *Quantum-mechanical modeling of the femtosecond isomerization in rhodopsin*, J. Phys. Chem. B, 104 (2000), pp. 1146–1149.
[15] L. Hörmander, *The Analysis of Linear Partial Differential Operators* III. *Pseudo-differential Operators*, Classics Math., Springer, Berlin, 1985.
[16] L. Landau, *Collected Papers of L. Landau*, Pergamon Press, Oxford, UK, 1965.
[17] C. Lasser, T. Swart, and S. Teufel, *Construction and validation of a rigorous surface hopping algorithm for conical crossings*, Commun. Math. Sci., 5 (2007), pp. 789–814.
[18] C. Lasser and S. Teufel, *Propagation through conical crossings: An asymptotic semigroup*, Comm. Pure Appl. Math., 58 (2005), pp. 1188–1230.
[19] H. Spohn and S. Teufel, *Adiabatic decoupling and time-dependent Born-Oppenheimer theory*, Comm. Math. Phys., 224 (2001), pp. 113–132.
[20] J. Tully and R. Preston, *Trajectory surface hopping approach to nonadiabatic molecular collisions: The reaction of H^+ with D_2*, J. Chem. Phys., 55 (1971), pp. 562–572.
[21] C. Zener, *Non-adiabatic crossing of energy levels*, Proc. Roy. Soc. Lond., 137 (1932), pp. 696–702.

SIAM J. MATH. ANAL.
Vol. 40, No. 1, pp. 134–166

ON A MODEL OF MULTIPHASE FLOW*

DEBORA AMADORI† AND ANDREA CORLI‡

Abstract. We consider a hyperbolic system of three conservation laws in one space variable. The system is a model for fluid flow allowing phase transitions; in this case the state variables are the specific volume, the velocity, and the mass density fraction of the vapor in the fluid. For a class of initial data having large total variation we prove the global existence of solutions to the Cauchy problem.

Key words. hyperbolic systems of conservation laws, phase transitions

AMS subject classifications. 35L65, 35L60, 35L67, 76T30

DOI. 10.1137/07069211X

1. Introduction.

We consider a model for the one-dimensional flow of an inviscid fluid capable of undergoing phase transitions. Both liquid and vapor phases are possible, as well as mixtures of them. In Lagrangian coordinates the model is

$$\begin{cases} v_t - u_x & = 0, \\ u_t + p(v,\lambda)_x & = 0, \\ \lambda_t & = 0\,. \end{cases} \tag{1.1}$$

Here $t > 0$ and $x \in \mathbb{R}$; moreover, $v > 0$ is the specific volume, u the velocity, and λ the mass density fraction of vapor in the fluid. Then $\lambda \in [0,1]$, with $\lambda = 0$ characterizing the liquid and $\lambda = 1$ the vapor phase; the intermediate values of λ model the mixtures of the two pure phases. The pressure is denoted by $p = p(v,\lambda)$; under natural assumptions the system is strictly hyperbolic.

This model is a simplified version of a model proposed by Fan [13], where also viscous and relaxation terms were taken into account. The model is isothermal (see (2.1) below); in the presence of phase transitions this physical assumption is meaningful for retrograde fluids. A study of the Riemann problem for a 2×2 relaxation approximation of (1.1) has been done in [10]. We focus here on the global existence of solutions to the Cauchy problem for (1.1), namely, for initial data

$$(v,u,\lambda)(0,x) = \big(v_o(x), u_o(x), \lambda_o(x)\big)$$

having finite total variation. This problem is motivated by the study of more complete models, where (1.1) is supplemented by source terms.

The problem of the global existence of solutions to strictly hyperbolic system of conservation laws has been studied for a long time; see [9, 11, 24, 25, 26] for general information. If the initial data have *small* total variation, then the Glimm theorem [14] applies; we refer the reader again to [9] for the analogous results obtained by a

*Received by the editors May 16, 2007; accepted for publication (in revised form) December 17, 2007; published electronically April 2, 2008. This work was supported by PROGETTO GNAMPA 2005 *Analisi Asintotica Per Sistemi Iperbolici Non Lineari*.
http://www.siam.org/journals/sima/40-1/69211.html
†Dipartimento di Matematica Pura e Applicata, Università degli Studi dell'Aquila, Via Vetoio, 67010 Coppito (AQ), Italy (amadori@univaq.it).
‡Dipartimento di Matematica, Università di Ferrara, Via Machiavelli 35, 44100 Ferrara, Italy (andrea.corli@unife.it).

wave-front tracking algorithm as well as for uniqueness and continuous dependence of the solutions on the initial data.

Some special systems allow, however, initial data with *large* total variation. For the system of isothermal gasdynamics Nishida [19] proved that it is sufficient that the variation $\mathrm{TV}(v_o, u_o)$ of the initial data is finite in order to have globally defined solutions. This result was extended by Nishida and Smoller [20] to any pressure law $p = k/v^\gamma$, $\gamma > 1$, provided that $(\gamma - 1)\mathrm{TV}(v_o, u_o)$ is small; related results are in [12]. For the full nonisentropic system of 3×3 gasdynamics, where $p = k \exp(\frac{\gamma-1}{R} s)/v^\gamma$ and s denotes the entropy, Liu [17, 16] proved the global existence of solutions if $(\gamma - 1)\mathrm{TV}(v_o, p_o)$ is small and $\mathrm{TV}(s_o)$ bounded. Temple [27] and Peng [21] obtained similar results. All these papers use the Glimm scheme. Analogous results making use of a wave-front tracking scheme have been given recently by Asakura [4, 5]; we point out that the use of wave-front tracking schemes in case of data with large variation is far from being trivial, and a deep analysis of the wave interactions is required. Very general results can be proved for systems with coinciding shock and rarefaction curves [8]; however, system (1.1) is not of this type.

In comparison with the above systems of gasdynamics, in (1.1) we keep a γ-law for the pressure with $\gamma = 1$ but add a dependence of p on λ: we then take $p = a(\lambda)/v$ for a suitable function a. System (1.1) has close connections to a system introduced by Benzoni-Gavage [6] and studied by Peng [22]; it seems, however, that the proof in [22] is not complete. A comparison of these models is done in subsection 3.1. We mention that the method of compensated compactness has also been applied to (1.1) (see [15, 7] and [18, sections 12.3 and 16]) but for different pressure laws.

In this paper we prove by a wave-front tracking scheme the global existence of solutions to (1.1) for a wide class of initial data with large total variation. We first introduce a *weighted total variation* (WTV) of $a(\lambda_o)$; this quantity arises in a natural way in the problem and also has an analytical meaning, being the logarithmic variation in the case of continuous functions. We prescribe a bound, on $\mathrm{WTV}\,(a(\lambda_o))$; for the variation $\mathrm{TV}(v_o, u_o)$ there is not such a bound, but, roughly speaking, the larger $\mathrm{TV}(v_o, u_o)$ is, the smaller $\mathrm{WTV}\,(a(\lambda_o))$ must be. An important point is that we give explicit expressions for these bounds; then our results are qualitatively different from some of those quoted above, where a generic smallness is required.

The plan of the paper is the following. The main result is stated in section 2, Theorem 2.2. The Riemann problem is reviewed in section 3 together with related results; proofs have been given in [2]. The definition of the algorithm is in section 4. The core of the proof is section 5—where interactions are studied in detail—and section 6—where we prove the convergence and consistence of the scheme. A careful analysis is needed due to the presence of large waves.

The paper is completed by two appendices. In the first one we prove the main result on the WTV. In the second we study the interaction of two shock waves to the light of section 5; namely, we look for precise bounds of the damping coefficient that controls the reflected wave produced in the interaction; we think that this analysis is interesting on its own. Good reading!

2. Main results. We consider the system of conservation laws (1.1). The pressure is given by

$$p(v, \lambda) = \frac{a^2(\lambda)}{v}, \tag{2.1}$$

where a is a smooth ($\mathbf{C^1}$) function defined on $[0, 1]$ satisfying for every $\lambda \in [0, 1]$

$$a(\lambda) > 0, \qquad a'(\lambda) > 0\,; \tag{2.2}$$

see Figure 2.1. For instance $a^2(\lambda) = k_0 + \lambda(k_1 - k_0)$ for $0 < k_0 < k_1$. As a consequence of (2.1) and (2.2) we have, for every $(v, \lambda) \in (0, +\infty) \times [0, 1]$,

$$p > 0, \qquad p_v < 0, \qquad p_{vv} > 0, \tag{2.3}$$

$$p_\lambda > 0, \qquad p_{v\lambda} < 0\,. \tag{2.4}$$

Remark that assumptions (2.3) and (2.4) are analogous to those usually made on the pressure in the full nonisentropic case [17], the entropy replacing λ.

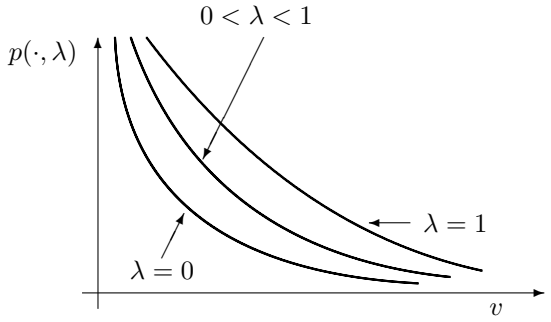


FIG. 2.1. *Pressure curves as functions of v.*

We denote $U = (v, u, \lambda) \in \Omega = (0, +\infty) \times \mathbb{R} \times [0, 1]$. Under assumptions (2.1) and (2.2) the system (1.1) is strictly hyperbolic in the whole Ω with eigenvalues $e_1 = -\sqrt{-p_v(v, \lambda)}$, $e_2 = 0$, $e_3 = \sqrt{-p_v(v, \lambda)}$. We write $c = \sqrt{-p_v} = a(\lambda)/v$. The eigenvectors associated with the eigenvalues e_i, $i = 1, 2, 3$, are $r_1 = (1, c, 0)$, $r_2 = (-p_\lambda, 0, p_v)$, $r_3 = (-1, c, 0)$. Because of the third inequality in (2.3) the eigenvalues e_1, e_3 are genuinely nonlinear with $\nabla e_i \cdot r_i = p_{vv}/(2c) > 0$, $i = 1, 3$, while e_2 is linearly degenerate. Pairs of Riemann invariants are $R_1 = \{u - a(\lambda) \log v, \lambda\}$, $R_2 = \{u, p\}$, $R_3 = \{u + a(\lambda) \log v, \lambda\}$.

We denote by $\mathrm{TV}(f)$ the total variation of a function f. In the case $f : \mathbb{R} \to (0, +\infty)$ we define the *weighted total variation* of f by

$$\mathrm{WTV}(f) = 2 \sup \sum_{j=1}^{n} \frac{|f(x_j) - f(x_{j-1})|}{f(x_j) + f(x_{j-1})},$$

where the supremum is taken over all $n \geq 1$ and $(n + 1)$-tuples of points x_j with $x_o < x_1 < \cdots < x_n$. This variation is motivated by the definition (3.6) of strength for the waves of the second family. If f is bounded and bounded away from zero, then

$$\frac{1}{\sup f} \mathrm{TV}(f) \leq \mathrm{WTV}(f) \leq \frac{1}{\inf f} \mathrm{TV}(f)\,.$$

PROPOSITION 2.1. *Consider $f : \mathbb{R} \to (0, +\infty)$; then*

$$\frac{\inf f}{\sup f} \mathrm{TV}\left(\log(f)\right) \leq \mathrm{WTV}(f) \leq \mathrm{TV}\left(\log(f)\right)\,. \tag{2.5}$$

Moreover, if $f \in C(\mathbb{R})$, then $\mathrm{WTV}(f) = \mathrm{TV}\left(\log(f)\right)$.

The proof is deferred to Appendix A. In (2.5), in the inequality on the right, the strict sign may occur if f is discontinuous; see Remark A.1.

We provide system (1.1) with initial data

$$U(x,0) = U_o(x) = (v_o(x), u_o(x), \lambda_o(x)) \tag{2.6}$$

for $x \in \mathbb{R}$. Denote $a_o(x) \doteq a(\lambda_o(x))$, $p_o(x) \doteq p(v_o(x), \lambda_o(x))$; remark that $\inf a_o(x) \geq a(0) > 0$. The main result of this paper now follows.

THEOREM 2.2. *Assume* (2.1), (2.2). *Consider initial data* (2.6) *with* $v_o(x) \geq \underline{v} > 0$ *for some constant* $\underline{v}$ *and* $0 \leq \lambda_o(x) \leq 1$. *For every* $m > 0$ *and a suitable function* $k(m) \in (0, 1/2)$ *the following holds. If*

$$\mathrm{TV}\left(\log(p_o)\right) + \frac{1}{\inf a_o}\mathrm{TV}(u_o) < 2\Big(1 - 2\mathrm{WTV}(a_o)\Big)m, \tag{2.7}$$

$$\mathrm{WTV}(a_o) < k(m), \tag{2.8}$$

then the Cauchy problem (1.1), (2.6) *has a weak entropic solution* (v, u, λ) *defined for* $t \in [0, +\infty)$. *Moreover, the solution is valued in a compact set of* Ω, *and there is a constant* $C(m)$ *such that for every* $t \in [0, +\infty)$

$$\mathrm{TV}\left(v(t,\cdot), u(t,\cdot)\right) \leq C(m)\,. \tag{2.9}$$

The function $k(m)$, whose expression is given in (6.22), deserves some comments. The interaction of two waves α, α' of the same family $i = 1, 3$ produces a wave β of the same family i and a "reflected" wave δ of the other family j ($j = 1, 3$, $j \neq i$). For a suitable definition of the strengths of the waves we prove that $|\delta| \leq d \cdot \min\{|\alpha|, |\alpha'|\}$ for a damping coefficient $d < 1$ depending on α and α'; see Lemma 5.6. The function k above depends essentially on the supremum of such coefficients d; we prove that $k(0) = 1/2$ and that $k(m)$ decreases to 0 as $m \to +\infty$. In particular then $\mathrm{WTV}(a_o) < 1/2$. The assumptions (2.7), (2.8) read as analogous to those in [20]: the larger m is, the smaller $k(m)$ is, and vice versa. The occurrence of a possible blow-up when the bound on $\mathrm{WTV}(a_o)$ does not hold is an interesting open problem.

The variation of λ_o appears both in condition (2.7), because of p_o, and in (2.8). Using the definition of the pressure, we can replace (2.7) by the slightly stronger condition $\mathrm{TV}\log(v_o) + 2\mathrm{TV}\log(a_o) + \frac{1}{\inf a_o}\mathrm{TV}(u_o) \leq 2\big(1 - 2\mathrm{WTV}(a_o)\big)m$ or even

$$\mathrm{TV}\log(v_o) + \frac{1}{a(0)}\mathrm{TV}(u_o) \leq 2m\left(1 - \frac{2m+1}{m}\mathrm{TV}(\log(a_o))\right)$$

by making use of (2.5). In particular if λ_o is constant, we recover the famous result by Nishida [19].

Clearly $\lambda(t,x) = \lambda_o(x)$ for any t because of the third equation in (1.1); this is why only v and u appear in the estimate (2.9). In other words system (1.1) can be rewritten as a p-system of two conservation laws with flux depending on x, namely, for the pressure law $p = p(v, \lambda_o(x)) = a^2(\lambda_o(x))/v$.

The proof of Theorem 2.2 makes use of a wave-front tracking scheme where we exploit the special structure of system (1.1) by differentiating the treatment of 1- and 3-waves from that of 2-waves. Our algorithm is a natural extension of that in [3], where the system for λ_o constant is studied, in the presence of a relaxation term.

Here we consider a linear functional as in [3] that accounts for the strengths of all 1- and 3-waves, with a weight $\xi > 1$ assigned to shock waves; a crucial point in

the proof is the choice of ξ as a function of m. This functional differs from that in [19, 4], where ξ is missing and only the variation of shocks is taken into account. Moreover, motivated again by [3], we do not introduce a simplified Riemann solver for interactions between 1- and 3-waves but only for interactions involving the 2-contact discontinuities. The interaction potential then uniquely considers interactions of 2-waves with 1- or 3-waves approaching it.

System (1.1) can be written in Eulerian coordinates. Denoting $\rho = 1/v$ the density, the pressure law becomes $p = a^2(\lambda)\rho$, and (1.1) turns into

$$\begin{cases} \rho_t + (\rho u)_x & = 0, \\ (\rho u)_t + \big(\rho u^2 + p(\rho, \lambda)\big)_x & = 0, \\ (\rho\lambda)_t + (\rho\lambda u)_x & = 0\,. \end{cases} \tag{2.10}$$

A global existence result of weak solutions for (2.10) holds by Theorem 2.2 because of [28].

3. Preliminaries.

3.1. Comparison with other models. In [6] many models for diphasic flows are proposed and studied. In a simple case (no source terms, the fluid in either a dispersed or separated configuration) and keeping notation as in [6, page 35], they can be written as

$$\begin{cases} (\rho_l R_l)_t + (\rho_l R_l u_l)_x & = 0, \\ (\rho_g R_g)_t + (\rho_g R_g u_g)_x & = 0, \\ (\rho_l R_l u_l + \rho_g R_g u_g)_t + (\rho_l R_l u_l^2 + \rho_g R_g u_g^2 + p)_x & = 0\,. \end{cases} \tag{3.1}$$

Here the indexes l and g stand for *liquid* and *gas*. Therefore ρ_l, R_l, u_l are the liquid density, phase fraction, and velocity, and analogously for the gas; clearly $R_l + R_g = 1$. The pressure law is $p = a^2\rho_g$ for $a > 0$ a constant. Equations (3.1) state the conservation of mass of either phase and the total momentum.

A case studied in [6, page 44] is when $u_l = u_g$ and ρ_l is constant, say, equal to 1. The unknown variables are then R_l, u, ρ_g, and it is assumed $0 < R_l < 1$ and $\rho_g > 0$; as a consequence $0 < R_g < 1$. Under these conditions, and still writing ρ_l instead of 1 for clarity, we define the concentration $c = \frac{\rho_g R_g}{\rho_l R_l} > 0$ and deduce the pressure law $p = a^2 c \frac{R_l}{1-R_l}$. We obtain exactly the model of [22]:

$$\begin{cases} (R_l)_t + (R_l u)_x & = 0, \\ (R_l c)_t + (R_l c u)_x & = 0, \\ (R_l(1+c)u)_t + \big(R_l(1+c)u^2 + p\big)_x & = 0\,. \end{cases} \tag{3.2}$$

This system is strictly hyperbolic for $c > 0$. Remark that the three eigenvalues of (3.2) coincide with u at $c = 0$, and if c vanishes identically, then (3.2) reduces to the pressureless gasdynamics system. System (3.2) is analogous to (2.10), but the pressure laws are different. In fact the variables ρ and λ of (2.10) write $\rho = \rho_l R_l + \rho_g R_g$ and $\lambda = \frac{\rho_g R_g}{\rho_l R_l + \rho_g R_g} = \frac{c}{1+c}$, and then $R_l = (1-\lambda)\rho$, $\rho_g = \frac{\lambda}{\rho^{-1} - (1-\lambda)}$. If we sum up the first two equations in (3.2), we find the first equation in (2.10); the third (resp., second) equation in (3.2) becomes the second (resp., third) equation in (2.10). The choice $p = a^2\rho_g$ for the pressure in (3.1) gives $p = a^2 \frac{\lambda}{\frac{1}{\rho} - (1-\lambda)}$.

Notice that the pressure vanishes in the presence of a pure liquid phase, and this is the main difference with (2.1), (2.2).

We now compare (2.10) and (3.2) in Lagrangian coordinates. Consider for (3.2) the change of coordinates $y = R_l dx - R_l u dt$ based on the streamlines of the liquid particles (because $R_l = \rho_l R_l$) [22]. Denote $w = \frac{1}{R_l} - 1 = \frac{R_g}{R_l} = \frac{c}{\rho_g}$. Then for $p = \frac{a^2 c}{w}$ system (3.2) turns into

$$\begin{cases} w_t - u_x & = 0, \\ ((1+c)u)_t + p_x & = 0, \\ c_t & = 0\,. \end{cases} \tag{3.3}$$

It is more interesting, however, to consider for system (3.2) the change $y = (1+c)R_l dx - (1+c)R_l u dt = \rho dx - \rho u dt$ into Lagrangian coordinates based on the streamlines of the *full* density ρ. Let w be as above and $v = \frac{w}{1+c} = \frac{R_g}{\rho}$. Then system (3.2) becomes system (1.1) with $a^2(\lambda) = a^2 \frac{c}{1+c} = a^2\lambda$. As a consequence the pressure law $p(v,\lambda) = a^2(\lambda)/v$ does not satisfy (2.2). This difficulty can be overcome as follows. Fix any $0 < a_1 < a_2 < a$ and consider for $c_i = \frac{a_i^2}{a^2 - a_i^2}$ the invariant domain $\{(R_l, u, c)\colon 0 < c_1 \le c \le c_2\}$ [22]. In this domain $0 < b_1 \le \lambda \le b_2 < 1$ for $b_i = a_i^2/a^2$. If we denote $\mu = \frac{\lambda - b_1}{b_2 - b_1}$, then the function $b(\mu) = a(\lambda) = a(b_1 + (b_2 - b_1)\mu)$ makes the pressure law $p(v,\lambda) = b^2(\mu)/v$, with $\mu \in [0,1]$, satisfy both conditions in (2.2).

3.2. Wave curves and the Riemann problem. In this section we recall some results about the wave curves for system (1.1) and the solution to the Riemann problem; see [2] for more details.

The shock-rarefaction curves through the point $U_o = (v_o, u_o, \lambda_o)$ for (1.1) are

$$\Phi_i(v, U_o) = (v, \phi_i(v, U_o), \lambda_o)\,, \qquad i = 1, 3, \tag{3.4}$$

$$\phi_1(v, U_o) = \begin{cases} u_o + a(\lambda_o)\cdot(v - v_o)/\sqrt{v v_o}, & v < v_o, \quad \text{shock}, \\ u_o + a(\lambda_o)\log(v/v_o), & v > v_o\,, \quad \text{rarefaction}, \end{cases}$$

$$\phi_3(v, U_o) = \begin{cases} u_o - a(\lambda_o)\log(v/v_o), & v < v_o, \quad \text{rarefaction}, \\ u_o - a(\lambda_o)\cdot(v - v_o)/\sqrt{v v_o}, & v > v_o\,, \quad \text{shock}, \end{cases}$$

$$\Phi_2(\lambda, U_o) = \left(v_o \frac{a^2(\lambda)}{a^2(\lambda_o)}, u_o, \lambda\right)\,, \qquad \lambda \in [0,1], \quad \text{contact discontinuity}. \tag{3.5}$$

The curves Φ_1, Φ_2, and Φ_3 are *plane* curves: Φ_1 and Φ_3 lie on the plane $\lambda = \lambda_o$, while Φ_2 on $u = u_o$.

DEFINITION 3.1 (wave strengths). *Under the notation* (3.4), (3.5) *we define the strength* ε_i *of an* i*-wave as*

$$\varepsilon_1 = \frac{1}{2}\log\left(\frac{v}{v_o}\right), \quad \varepsilon_2 = 2\,\frac{a(\lambda) - a(\lambda_o)}{a(\lambda) + a(\lambda_o)}, \quad \varepsilon_3 = \frac{1}{2}\log\left(\frac{v_o}{v}\right)\,. \tag{3.6}$$

According to this definition, rarefaction waves have positive strengths and shock waves have negative strengths. Given the initial datum $\lambda_o = \lambda_o(x)$, denote

$$a^* \doteq \sup_{x\in\mathbb{R}} a\big(\lambda_o(x)\big)\,, \quad a_* \doteq \inf_{x\in\mathbb{R}} a\big(\lambda_o(x)\big)\,, \quad [a]_* \doteq \frac{a^* - a_*}{a^* + a_*}\,. \tag{3.7}$$

Then $[a]_* \le \frac{a(1)-a(0)}{a(1)+a(0)} < 1$ and $|\varepsilon_2| \le 2[a]_* < 2$. It is useful to also define the function (see [22])

$$h(\varepsilon) = \begin{cases} \varepsilon & \text{if } \varepsilon \ge 0\,, \\ \sinh \varepsilon & \text{if } \varepsilon < 0\,. \end{cases} \tag{3.8}$$

Then we have for $i = 1, 3$

$$\phi_i(v, U_o) = u_o + a(\lambda_o) \cdot 2h(\varepsilon_i)\,. \tag{3.9}$$

At last we consider the Riemann problem. This is the initial value problem for (1.1) under the piecewise constant initial condition

$$(v, u, \lambda)(0, x) = \begin{cases} (v_\ell, u_\ell, \lambda_\ell) = U_\ell & \text{if } x < 0, \\ (v_r, u_r, \lambda_r) = U_r & \text{if } x > 0 \end{cases} \tag{3.10}$$

for U_ℓ and U_r in Ω. We denote $a_r = a(\lambda_r)$, $p_r = a_r^2/v_r$, and similarly for a_ℓ, p_ℓ.

PROPOSITION 3.2. *Fix any pair of states U_ℓ, U_r in Ω; then the Riemann problem* (1.1), (3.10) *has a unique Ω-valued solution in the class of solutions consisting of simple Lax waves. If ε_i is the strength of the i-wave, $i = 1, 2, 3$, then*

$$\varepsilon_3 - \varepsilon_1 = \frac{1}{2} \log\left(\frac{p_r}{p_\ell}\right), \qquad 2\big(a_\ell h(\varepsilon_1) + a_r h(\varepsilon_3)\big) = u_r - u_\ell\,.$$

Moreover, let $\underline{v} > 0$ be a fixed number. There exists a constant $C_1 > 0$ depending on $\underline{v}$ and $a(\lambda)$ such that if v_l, $v_r \ge \underline{v}$, then

$$|\varepsilon_1| + |\varepsilon_2| + |\varepsilon_3| \le C_1 |U_\ell - U_r|\,. \tag{3.11}$$

For the proof, see [2]. One can easily find that

$$\begin{aligned} |\varepsilon_1| + |\varepsilon_3| &\le \frac{1}{2} |\log(p_r) - \log(p_\ell)| + \frac{1}{2\min\{a_\ell, a_r\}} |u_r - u_\ell| \\ &\le \frac{1}{2} |\log(v_r) - \log(v_\ell)| + |\log(a_r) - \log(a_\ell)| + \frac{1}{2\min\{a_\ell, a_r\}} |u_r - u_\ell|\,. \end{aligned} \tag{3.12}$$

We remark that for any Riemann data $(v_\ell, u_\ell, \lambda_\ell)$, (v_r, u_r, λ_r), the λ component of the solution takes value λ_ℓ for $x < 0$ and λ_r for $x > 0$. The fact that the interfaces between different phases are connected by a stationary wave can then be interpreted as a "kinetic condition" [1], analogous to Maxwell's rule.

4. The approximate solution. In this section we define a wave-front tracking scheme [9] to build up piecewise constant approximate solutions to (1.1). More precisely, we follow the algorithm introduced in [3].

First, we approximate the initial data. For any $\nu \in \mathbb{N}$ we take a sequence $(v_o^\nu, u_o^\nu, \lambda_o^\nu)$ of piecewise constant functions with a finite number of jumps such that

(i) $\mathrm{TV} p_o^\nu \le \mathrm{TV} p_o$, $\mathrm{TV} u_o^\nu \le \mathrm{TV} u_o$, $\mathrm{WTV} a(\lambda_o^\nu) \le \mathrm{WTV} a(\lambda_o)$, $\inf a_o^\nu \ge \inf a_o$;
(ii) $\lim_{x \to -\infty} (v_o^\nu, u_o^\nu, \lambda_o^\nu)(x) = \lim_{x \to -\infty} (v_o, u_o, \lambda_o)(x)$;
(iii) $\|(v_o^\nu, u_o^\nu, \lambda_o^\nu) - (v_o, u_o, \lambda_o)\|_{\mathbf{L^1}} \le \frac{1}{\nu}$,

where $p_o^\nu = a^2(\lambda_o^\nu)/v_o^\nu$. Second, we define the approximate Riemann solver. We introduce positive parameters $\eta = \eta_\nu$, $\rho = \rho_\nu$; they control, respectively, the size of rarefactions and the threshold when a simplified Riemann solver is used. Define also a parameter $\hat{s} > 0$ strictly larger than all possible speeds of wave fronts of both families 1 and 3. These parameters will be determined at the end of section 6.

- At time $t = 0$ we solve the Riemann problems at each point of jump of $(v_o^\nu, u_o^\nu, \lambda_o^\nu)(0+, \cdot)$ as follows: shocks are not modified while rarefactions are approximated by fans of waves, each of them having size less than η. More precisely, a rarefaction of size ε is approximated by $N = [\varepsilon/\eta]+1$ waves whose size is $\varepsilon/N < \eta$; we set their speeds to be equal to the characteristic speed of the state at the right.
  Then $(v, u, \lambda)(t, \cdot)$ is defined until some wave fronts interact; by slightly changing the speed of some waves [9] we can assume that only *two* fronts interact at a time.
- When two wave fronts of families either 1 or 3 interact we solve the Riemann problem at the interaction point. If one of the incoming waves is a rarefaction, after the interaction it is prolonged (if it still exists) as a single discontinuity with speed equal to the characteristic speed of the state at the right. If a new rarefaction is generated, we employ the Riemann solver described before and divide it into a fan of waves having size less than η.
- When a wave front of family either 1 or 3 interacts with a 2-wave we proceed as follows. Let δ_2 be the size of the 2-wave and δ the size of the other wave.
  - If $|\delta_2\delta| \geq \rho$, we solve the Riemann problem as above, that is, with the *accurate Riemann solver*.
  - If $|\delta_2\delta| < \rho$, we prolong the 1- or 3- wave with a wave of the same family and size. Since the two waves do not commute, a *nonphysical* front is introduced [9], with fixed speed $\hat{s} > 0$. The size of a nonphysical wave is set to be $|u_r - u_\ell|$, where u_ℓ, u_r are the u components of the left and right states of the wave. We call this solver the *simplified Riemann solver*.
- When a nonphysical front interacts with a front of family 1, 2, or 3 ("physical"), we prolong the solution with a physical wave of the same size and a nonphysical one, consequently computing the intermediate value.

We refer for the last two items to Proposition 5.12 below. Remark that two nonphysical fronts cannot interact since they have the same constant speed $\hat{s}$. We denote by $\mathcal{NP}$ the set of nonphysical waves.

5. Interactions. Fix the index ν introduced in the previous section. We shall prove in subsection 6.1 that the algorithm described above is defined for any $t > 0$ and provides for any initial data $(v_o^\nu, u_o^\nu, \lambda_o^\nu)$ a piecewise constant approximate solution $(v^\nu, u^\nu, \lambda^\nu) = (v, u, \lambda)$, where we dropped for simplicity the index ν. Here we study the interaction of waves.

For $K_{np} > 0$ and $t > 0$ we define the functional L and the interaction potential Q, both referred to $(v, u, \lambda)(t, \cdot)$, by

$$L(t) = \sum_{i=1,3} |\gamma_i| + K_{np} L_{np}\,, \qquad L_{np} = \sum_{\gamma \in \mathcal{NP}} |\gamma|,$$

$$Q(t) = \sum_{\gamma_3 \text{ at the left of } \delta_2} |\gamma_3||\delta_2| + \sum_{\gamma_1 \text{ at the right of } \delta_2} |\delta_2||\gamma_1|\,. \tag{5.1}$$

Remark that L takes into account only the strengths of both 1- and 3-waves and that of nonphysical waves. For contact discontinuities we define

$$L_{cd} = \sum |\gamma_2| = \mathrm{WTV}a(\lambda_o^\nu)\,.$$

Finally, for $\xi \geq 1$ and $K \geq 0$ we introduce

$$L_\xi = \sum_{\substack{i=1,3\\ \gamma_i>0}} |\gamma_i| + \xi \sum_{\substack{i=1,3\\ \gamma_i<0}} |\gamma_i| + K_{np} L_{np}\,, \tag{5.2}$$

$$F = L_\xi + KQ\,. \tag{5.3}$$

For simplicity we omitted noting the dependence on K_{np} in the functional L_ξ and on K_{np}, ξ, K in F; the choice of K_{np} shall depend on that of K; see Proposition 5.12.

Observe that if λ_o is constant, then $Q = 0$ and $F = L_\xi$, whose variation was analyzed in Lemma 3.2 of [3]. Hence we will assume from now on that

$$A_o \doteq \mathrm{WTV}(a_o) > 0\,. \tag{5.4}$$

By assumption (i) in section 4, one has $L_{cd} \leq A_o$.

In the following sections we analyze in detail the different types of interactions. Recalling the definition of h, (3.8), and with the notation of Figure 5.1, we introduce the following identities (see (3.1), (3.2) in [22]):

$$\varepsilon_3 - \varepsilon_1 = \alpha_3 + \beta_3 - \alpha_1 - \beta_1, \tag{5.5}$$

$$a_\ell h(\varepsilon_1) + a_r h(\varepsilon_3) = a_\ell h(\alpha_1) + a_m h(\alpha_3) + a_m h(\beta_1) + a_r h(\beta_3)\,. \tag{5.6}$$

Formula (5.5) does not depend on λ and follows easily by equating the specific volumes v before and after the interaction time. By equating the velocities u we obtain (5.6). These properties are a consequence of the definition (3.6) of the strengths for 1- and 3-waves and of (3.9).

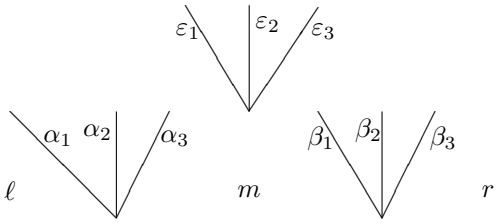


FIG. 5.1. *A general interaction pattern.*

5.1. Interactions with a 2-wave. We first consider the interactions of 1- or 3-waves with a 2-wave; see Figure 5.2.

PROPOSITION 5.1 (see [2]). *Denote by λ_ℓ, λ_r the side states of a 2-wave. The interactions of 1- or 3-waves with the 2-wave give rise to the following pattern of solutions:*

Interaction	*Outcome*	
	$\lambda_\ell < \lambda_r$	$\lambda_\ell > \lambda_r$
$2 \times 1R$	$1R + 2 + 3R$	$1R + 2 + 3S$
$2 \times 1S$	$1S + 2 + 3S$	$1S + 2 + 3R$
$3R \times 2$	$1S + 2 + 3R$	$1R + 2 + 3R$
$3S \times 2$	$1R + 2 + 3S$	$1S + 2 + 3S$

The next lemma is concerned instead with the *strengths* of waves involved in the interaction above. The inequalities (5.8) improve the inequality (3.3) in [22] in the special case of two interacting wave fronts, one of them being of the second family. More precisely, under the notation of [22] we find a term $1/(a_r + a_\ell)$ instead of $1/\min\{a_r, a_\ell\}$. The proof differs from Peng's. Our estimates are sharp: in some cases (5.8) reduces to an identity.

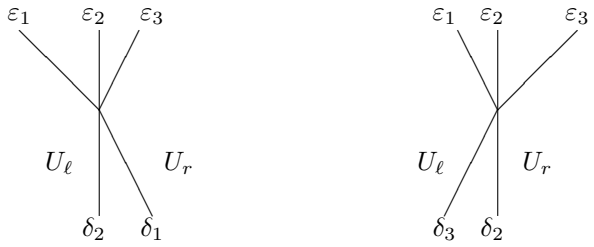


FIG. 5.2. *Interactions.* (a) *from the right;* (b) *from the left.*

LEMMA 5.2 (see [2]). *Assume that a* 1*-wave of strength* δ_1 *or a* 3*-wave of strength* δ_3 *interacts with a* 2*-wave of strength* $\delta_2 = 2(a_r - a_\ell)/(a_r + a_\ell)$. *Then the strengths* ε_i *of the outgoing waves satisfy* $\varepsilon_2 = \delta_2$ *and*

$$|\varepsilon_i - \delta_i| = |\varepsilon_j| \le \frac{1}{2}\,|\delta_2| \cdot |\delta_i| \le [a]_* |\delta_i| \tag{5.7}$$

for $i, j = 1, 3$, $i \neq j$. *Moreover,*

$$|\varepsilon_1| + |\varepsilon_3| \le \begin{cases} |\delta_1| + |\delta_1|[\delta_2]_+ & \text{if } 1 \text{ interacts}, \\ |\delta_3| + |\delta_3|[\delta_2]_- & \text{if } 3 \text{ interacts}. \end{cases} \tag{5.8}$$

Here $[x]_+ = \max\{x, 0\}$, $[x]_- = \max\{-x, 0\}$, $x \in \mathbb{R}$. Remark that the colliding 1- or 3-wave does not change sign across the interaction. Moreover, the functional L increases iff the incoming and the reflected waves are of the same type; this happens when the colliding wave is moving toward a more liquid phase.

Now we prove that F is decreasing for suitable K when an interaction with a 2-wave occurs. The potential Q is needed to balance the possible increase of L_ξ.

PROPOSITION 5.3. *Assume* $A_o < 2$ *and consider an interaction of a* 1*- or* 3*-wave with a* 2*-wave, with the notation of Lemma* 5.2. *Then* $\Delta Q < 0$. *If, moreover,*

$$\xi \ge 1 \qquad \text{and} \qquad K > \frac{2\xi}{2 - A_o}, \tag{5.9}$$

then

$$\xi|\varepsilon_j| = \xi\big||\varepsilon_i| - |\delta_i|\big| < \frac{K}{2}|\Delta Q|, \tag{5.10}$$

and hence $\Delta F < 0$.

Proof. We consider the interaction of a 1-wave with a 2-wave; the symmetric case follows in an analogous way. We use the notation as in Figure 5.2(a). We define $L^*_{cd} = L^-_{cd} + L^+_{cd}$, $L^\pm_{cd}$, meaning *right or left* of the 2-wave under consideration.

By assumption, one has

$$L_{cd} = L_{cd}^- + L_{cd}^+ + |\delta_2| = L_{cd}^* + |\delta_2| \le A_o < 2\,. \tag{5.11}$$

Recall that $\varepsilon_1 - \delta_1 = \varepsilon_3$ and by Proposition 5.1

$$|\varepsilon_1| - |\delta_1| = |\varepsilon_3| \qquad \text{if } \delta_2 > 0\,, \tag{5.12}$$

$$|\varepsilon_1| - |\delta_1| = -|\varepsilon_3| \qquad \text{if } \delta_2 < 0\,, \tag{5.13}$$

so that in particular $|\varepsilon_3| = \big||\varepsilon_1| - |\delta_1|\big|$. An estimate for ΔQ follows at once because of (5.11):

$$\Delta Q = -|\delta_2\delta_1| + (|\varepsilon_1| - |\delta_1|)\, L_{cd}^- + |\varepsilon_3| L_{cd}^+ \le \frac{1}{2}|\delta_2\delta_1|(L_{cd}^* - 2)$$

$$\le \frac{1}{2}|\delta_2\delta_1|(A_o - 2) < 0\,. \tag{5.14}$$

Hence, using (5.7), we get

$$\xi|\varepsilon_3| + \frac{K}{2}\Delta Q \le \frac{1}{2}|\delta_2\delta_1|\left\{\xi + \frac{K}{2}(A_o - 2)\right\} < 0 \tag{5.15}$$

because of (5.9); this proves (5.10). Finally, by using (5.15) we get

$$\Delta F = \Delta L_\xi + K\Delta Q \le \xi|\varepsilon_3| + \xi|\varepsilon_1 - \delta_1| + K\Delta Q < 0\,. \tag{5.16}$$

□

5.2. Interactions between 1- and 3-waves. Here we analyze the possible interactions between 1- and 3-waves. Two situations may occur; see Figure 5.3: either the waves belong to different families or they both belong to the same family. In this last case, at least one of the waves must be a shock.

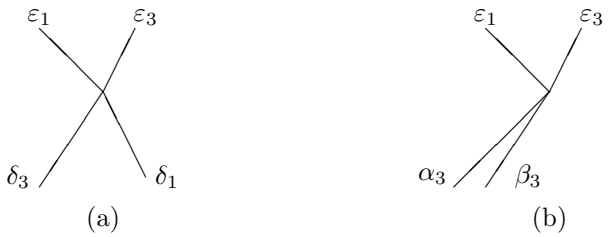


FIG. 5.3. *Interactions of* 1- *and* 3-*waves.*

LEMMA 5.4 (different families interacting). *If a wave of the third family interacts with a wave of the first family, they cross each other without changing their strength.*

Proof. See also Lemma 3.1 in [3]. Using notation as in Figure 5.3(a) we have $\varepsilon_3 - \varepsilon_1 = \delta_3 - \delta_1$ and $h(\varepsilon_1) + h(\varepsilon_3) = h(\delta_1) + h(\delta_3)$. The uniqueness of solutions to the Riemann problem implies $\varepsilon_1 = \delta_1$, $\varepsilon_3 = \delta_3$.

Remark that here $\Delta L_\xi = 0 = \Delta Q$ and then $\Delta F = 0$ for all $\xi \ge 1$ and K. □

LEMMA 5.5 (same family interacting: outcome). *Assume that a wave α_3 of the third family interacts with a wave β_3 of the third family, giving rise to waves ε_1, ε_3. Then*

(i) $\alpha_3 < 0$, $\beta_3 < 0 \Rightarrow \varepsilon_1 > 0$, $\varepsilon_3 < 0$,

(ii) $\alpha_3\beta_3 < 0 \Rightarrow \varepsilon_1 < 0$.

An analogous result holds for interacting waves of the first family.

Proof. The proof can be done in a geometric way by observing the mutual positions of the curves [19, 26]. A simple alternative proof by analytical arguments now follows. We have

$$\varepsilon_3 - \varepsilon_1 = \alpha_3 + \beta_3, \tag{5.17}$$

$$h(\varepsilon_1) + h(\varepsilon_3) = h(\alpha_3) + h(\beta_3)\,. \tag{5.18}$$

In case (i) these formulas read $\varepsilon_1 - \varepsilon_3 = |\alpha_3| + |\beta_3| > 0$ and $-h(\varepsilon_1) - h(\varepsilon_3) = \sinh(|\alpha_3|) + \sinh(|\beta_3|) > 0$. If it were $\varepsilon_3 > 0$, then $\varepsilon_1 > 0$ from the first equality and $\varepsilon_1 < 0$ from the second, a contradiction. Therefore $\varepsilon_3 < 0$ so that $\varepsilon_1 + |\varepsilon_3| = |\alpha_3| + |\beta_3|$ and $-h(\varepsilon_1) + \sinh(|\varepsilon_3|) = \sinh(|\alpha_3|) + \sinh(|\beta_3|)$. Analogously, if it were $\varepsilon_1 < 0$, using elementary inequalities we get $0 = \sinh(|\varepsilon_1|) + \sinh(|\alpha_3| + |\beta_3| + |\varepsilon_1|) - \sinh(|\alpha_3|) - \sinh(|\beta_3|) \geq 2\sinh(|\varepsilon_1|)$, a contradiction again. Hence $\varepsilon_1 > 0$.

In case (ii) assume $\alpha_3 < 0$, $\beta_3 > 0$; the other case is dealt with analogously since (5.17), (5.18) are symmetric in α_3, β_3. We have $\varepsilon_3 - \varepsilon_1 = -|\alpha_3| + |\beta_3|$ and $h(\varepsilon_1) + h(\varepsilon_3) = -\sinh(|\alpha_3|) + |\beta_3|$; then $[h(\varepsilon_1) + \varepsilon_1] + [h(\varepsilon_3) - \varepsilon_3] = |\alpha_3| - \sinh(|\alpha_3|) < 0$. If $\varepsilon_3 > 0$, this last equality becomes $h(\varepsilon_1) + \varepsilon_1 = |\alpha_3| - \sinh(|\alpha_3|) < 0$, which implies $\varepsilon_1 < 0$. If $\varepsilon_3 < 0$, then $h(\varepsilon_1) + \varepsilon_1 = [|\alpha_3| - \sinh(|\alpha_3|)] - [|\varepsilon_3| - \sinh(|\varepsilon_3|)]$. If it were $\varepsilon_1 > 0$, it would be $|\alpha_3| < |\varepsilon_3|$, since the map $x \mapsto x - \sinh x$ is decreasing; but from $|\varepsilon_3| + |\varepsilon_1| = |\alpha_3| - |\beta_3|$ we would get that $|\varepsilon_3| < |\alpha_3|$, a contradiction. Hence in all cases one has $\varepsilon_1 < 0$. ∎

Now we give sharper estimates for the interaction of waves of the same family: we prove that the strength of the reflected wave is bounded by the size of each incoming wave, multiplied by a damping factor smaller than 1. This property will be crucial in the next section, and it holds also for interactions with a 2-wave, with damping factor $[a]_*$; see (5.7). In the case below, however, the coefficient depends on the strengths of the incoming waves; this happens also when nonphysical waves are generated; see Proposition 5.12. We assume that

$$\begin{aligned} &\textit{the strength of any interacting } i\textit{-wave is less than } m \\ &\textit{for some } m > 0 \textit{ and } i = 1, 3. \end{aligned} \tag{5.19}$$

In the special case of interaction of waves of the same family producing two outgoing shocks we give a more precise result in Appendix B.

LEMMA 5.6 (same family interacting). *Consider the interaction of two waves of the same family, of sizes α_i and β_i, $i = 1, 3$, producing two outgoing waves ε_1, ε_3; assume* (5.19). *Then the following hold.*

(i) *There exists a damping coefficient $d = d(m)$, with $0 < d < 1$, such that*

$$|\varepsilon_j| \leq d(m) \cdot \min\{|\alpha_i|, |\beta_i|\}\,, \qquad j \neq i\,. \tag{5.20}$$

(ii) *If the incoming waves are both shocks, the resulting shock satisfies $|\varepsilon_i| > \max\{|\alpha_i|, |\beta_i|\}$. If the incoming waves have different signs, both the amount of shocks and the amount of rarefactions of the ith family decrease across the interaction.*

In any case

$$|\varepsilon_i| \leq |\alpha_i| + |\beta_i|\,. \tag{5.21}$$

Proof. (i) To fix the ideas, assume $i = 3$. We have

$$\varepsilon_3 - \varepsilon_1 = \alpha_3 + \beta_3, \tag{5.22}$$

$$h(\varepsilon_1) + h(\varepsilon_3) = h(\alpha_3) + h(\beta_3), \tag{5.23}$$

and then

$$h(\varepsilon_1) + h(\varepsilon_1 + \alpha_3 + \beta_3) = h(\alpha_3) + h(\beta_3)\,. \tag{5.24}$$

Remark that (5.24) is symmetric in α_3, β_3; from the implicit function theorem we find that $\varepsilon_1 = \varepsilon_1(\alpha_3, \beta_3)$ is $\mathbf{C^1}$. Using the notation $\varepsilon_1 = \tau$, $\alpha_3 = a$, $\beta_3 = b$, the identity (5.24) rewrites as

$$h(\tau) + h(\tau + a + b) - h(a) - h(b) = 0\,, \tag{5.25}$$

with $\tau = \tau(a, b)$. One verifies that $\tau(a, 0) = \tau(0, b) = 0$ and that

$$\tau_a = \frac{h'(a) - h'(\tau + a + b)}{h'(\tau) + h'(\tau + a + b)}, \qquad \tau_b = \frac{h'(b) - h'(\tau + a + b)}{h'(\tau) + h'(\tau + a + b)}\,.$$

As $a \to 0$, one has $\tau_a \to (1 - h'(b))/(1 + h'(b))$; then $|\tau_a(0, b)| < 1$, and it can be bounded by a positive constant less than 1 that depends on m. The same argument works for τ_b.

To complete the proof, we show that $|\tau| < \min\{|a|, |b|\}$ in the nontrivial case $a \neq 0 \neq b$. We argue by contradiction, using an argument of [23]. Suppose that $|\tau| \geq |a|$; we can assume $\tau > 0$, since the case $\tau < 0$ can be proved by using the equality (5.25) written in terms of $G(t) = -h(-t)$. Since the function h is increasing we have, for $\tau \geq |a|$,

$$h(\tau) \geq h(a), \qquad h(\tau + a + b) \geq h(b).$$

Moreover, one of the two inequalities is strict: if $a < 0$, the first; if $a > 0$, the second. Hence we contradict (5.25).

(ii) From [3] we already know that $\Delta L = |\varepsilon_1| + |\varepsilon_3| - |\alpha_3| - |\beta_3| \leq 0$, and hence (5.21).

If the incoming waves are both shocks, then (5.22) becomes

$$|\varepsilon_3| + |\varepsilon_1| = |\alpha_3| + |\beta_3|\,. \tag{5.26}$$

From (i) we have $|\varepsilon_1| < |\alpha_3|$, $|\beta_3|$ and hence the first part of (ii). On the other hand, if $\alpha_3\beta_3 < 0$, we have $\varepsilon_1 < 0$. We can assume $\alpha_3 < 0 < \beta_3$; hence (5.22) becomes $\varepsilon_3 = |\beta_3| - |\alpha_3| - |\varepsilon_1|$. If $\varepsilon_3 > 0$, then $|\varepsilon_3| < |\beta_3|$; if $\varepsilon_3 < 0$, using (i) again one finds that $|\varepsilon_3| < |\alpha_3|$. ∎

Remark 5.7. The damping coefficient $d(m)$ (see Figure 5.4) is given by

$$d(m) = \max_{\substack{|a| \leq m \\ |b| \leq m}} \frac{|\varepsilon(a, b)|}{\min\{|a|, |b|\}}\,,$$

where the function $\varepsilon(a, b)$ satisfies $h(\varepsilon) + h(\varepsilon + a + b) - h(a) - h(b) = 0$; see (5.24). Hence $d(m)$ increases with m and vanishes as $m \to 0$ because quadratic interaction estimates hold for m small.

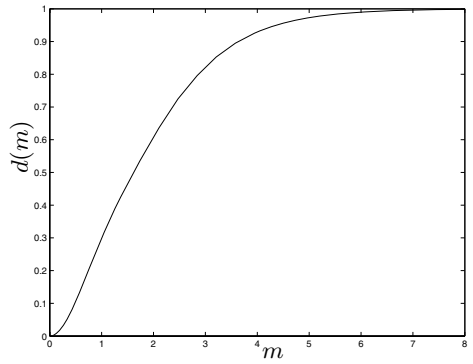


FIG. 5.4. *The coefficient* $d(m)$.

Moreover, it is asymptotic to 1 for m large. Indeed, from the proof of Lemma 5.6 we have $\tau_a(0,b) = \frac{1-h'(b)}{1+h'(b)}$; then $\tau_a(0,b) = 0$ if $b > 0$ and $|\tau_a(0,b)| = \frac{\cosh(b)-1}{\cosh(b)+1} \leq \frac{\cosh m-1}{\cosh m+1}$ if $b < 0$. Therefore $|\tau_a(0,b)| \leq \frac{\cosh m-1}{\cosh m+1}$ for every b, and an analogous estimate holds for $|\tau_b(0,a)|$. Hence $d(m) \geq \frac{\cosh m-1}{\cosh m+1} \doteq c(m)$; we refer to Lemma B.1 for the role of this quantity.

Remark 5.8. When a rarefaction interacts with a 1- or 3-wave, its size does not increase. Indeed, the size does not change upon interactions with waves of the other family by Lemma 5.4; if the rarefaction interacts with a shock of the same family, we apply Lemma 5.6(ii). Remark, moreover, that by Lemma 5.5, when a rarefaction and a shock of the same family interact, the reflected wave is never a rarefaction.

Remark 5.9. If two waves of the same family interact, the wave belonging to that family can be missing, while the "reflected" wave is always present. This follows easily from (5.22), (5.23).

PROPOSITION 5.10 (variation of F). *Consider the interactions of any two wave fronts of the same family,* 1 *or* 3*, and assume* (5.4), (5.19). *If*

$$1 < \xi < \frac{1}{d} \qquad \text{and} \qquad K < \frac{\xi - 1}{A_o}, \tag{5.27}$$

then $\Delta L_\xi < 0$ *and* $\Delta F < 0$.

Proof. Let two waves α_i, β_i interact, $i = 1, 3$, giving rise to waves ε_1, ε_3. We consider $i = 3$, the other case being analogous. Using (5.21), we get

$$\Delta Q = (|\varepsilon_3| - |\alpha_3| - |\beta_3|)\, L^+_{cd} + |\varepsilon_1| L^-_{cd} \leq |\varepsilon_1| L^-_{cd} \leq |\varepsilon_1| A_o\,. \tag{5.28}$$

Now we claim that

$$\Delta L_\xi + |\varepsilon_1|(\xi - 1) \leq 0\,. \tag{5.29}$$

From this estimate it follows that $\Delta F = \Delta L_\xi + K\Delta Q \leq |\varepsilon_1|\,(1 - \xi + KA_o) < 0$ because of (5.27). To prove our claim we consider the possible cases; we make use of (5.22).

$\boxed{SS \to RS}$. Since $\Delta L = 0$, then

$$\Delta L_\xi + (\xi - 1)|\varepsilon_1| = \xi(|\varepsilon_1| + |\varepsilon_3| - |\alpha_3| - |\beta_3|) \leq 0\,. \tag{5.30}$$

$\boxed{SR,\, RS \to SR}$. Assume $\alpha_3 < 0 < \beta_3$; then (5.29) reads $(2\xi - 1)|\varepsilon_1| + |\varepsilon_3| - |\beta_3| - \xi|\alpha_3| \le 0$. For later use we prove the stronger inequality

$$\xi^2|\varepsilon_1| + |\varepsilon_3| - |\beta_3| - \xi|\alpha_3| \le 0\,. \tag{5.31}$$

Indeed, from Lemma 5.6(ii) we have $|\varepsilon_3| < |\beta_3|$, while $\xi|\varepsilon_1| \le |\alpha_3|$ from (5.20), $(5.27)_1$.

$\boxed{SR,\, RS \to SS}$. Assume $\alpha_3 < 0 < \beta_3$; then (5.29) is $(2\xi - 1)|\varepsilon_1| + \xi(|\varepsilon_3| - |\alpha_3|) - |\beta_3| \le 0$. We prove also in this case the stronger inequality

$$\xi^2|\varepsilon_1| + \xi(|\varepsilon_3| - |\alpha_3|) - |\beta_3| \le 0. \tag{5.32}$$

Indeed, by (5.17) and again because of (5.20), $(5.27)_1$, one has

$$\begin{aligned}\xi^2|\varepsilon_1| + \xi(|\varepsilon_3| - |\alpha_3|) - |\beta_3| &= \xi^2|\varepsilon_1| + \xi(|\varepsilon_1| - |\beta_3|) - |\beta_3| \\ &= (\xi + 1)(\xi|\varepsilon_1| - |\beta_3|) \le 0\,.\end{aligned}$$

This proves the claim and concludes the proof. □

Remark 5.11. From the above proof we see that $\Delta L_\xi \le 0$ for $\xi = 1$. This was a key point in [19], where, however, a different choice of strengths was made. In [3] the inequality $\Delta L_\xi \le 0$ was proved to hold also for $1 < \xi \le \xi_o$ for some $\xi_o > 1$; the condition $(5.27)_1$ gives an estimate of such a threshold.

More precisely, in the first two cases of Proposition 5.10 we have $\Delta L_\xi \le 0$ for every $\xi \ge 1$. The third case is analyzed in detail in Lemma B.1; we prove there that $\Delta L_\xi \le 0$ for any $\xi > 1$ if $c(m) \le 1/2$, while we need $1 < \xi \le \frac{1}{2c(m)-1}$ if $c(m) > 1/2$.

5.3. Nonphysical waves. In this subsection we compute the strength of a nonphysical wave generated by an interaction and prove that it does not change in subsequent interactions. We introduce the following notation: given $U_\ell = (v_\ell, u_\ell, \lambda_\ell)$ and λ_r we define by

$$U^*_{\ell r} = \Phi_2(\lambda_r, U_\ell) = (A_{r\ell} v_\ell, u_\ell, \lambda_r)$$

the state on the right of a 2-wave with left state $U_\ell = (v_\ell, u_\ell, \lambda_\ell)$ and $\lambda = \lambda_r$ on the right, where $A_{r\ell} = a^2(\lambda_r)/a^2(\lambda_\ell)$. See (3.5) and [2].

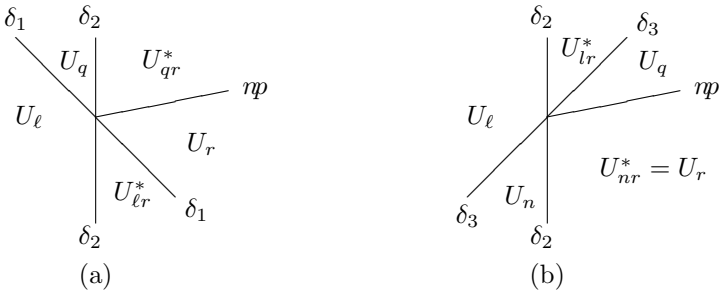


FIG. 5.5. *Simplified Riemann solver.*

PROPOSITION 5.12 (nonphysical waves). *Consider $U_\ell = (v_\ell, u_\ell, \lambda_\ell)$. Let $U_r = (v_r, u_r, \lambda_r)$ be connected to $U^*_{\ell r}$ by a 1-wave of size δ_1 and $U_q = (v_q, u_q, \lambda_\ell)$ be connected to U_ℓ by a 1-wave of size δ_1; see Figure* 5.5(a). *Assume* (5.19).

Then U^*_{qr} and U_r differ only in the u component; if δ_2 denotes the size of the 2-wave, there exists a constant $C_o = C_o(m)$ such that

$$\|U^*_{qr} - U_r\| = |u_q - u_r| \le C_o|\delta_2\delta_1|\,. \tag{5.33}$$

A similar result holds for the interaction of a 3-wave (see Figure 5.5(b)), again under (5.19).

Moreover, the size of a nonphysical wave does not change in subsequent interactions. For any $K > 0$ and $K_{np} < K/C_o$ at any interaction involving a nonphysical wave we have

$$\Delta F \le 0\,, \tag{5.34}$$

with $\Delta F < 0$ when a nonphysical wave is generated.

Proof. Recalling [2, Lemma 2], only the u component will be different after commutation of the 1- and the 2-wave. We find that

$$u_q - u_\ell = 2a_\ell h(\delta_1), \qquad u_r - u_\ell = 2a_r h(\delta_1),$$

and hence

$$|u_q - u_r| = 2|a_\ell - a_r| \cdot |h(\delta_1)| \le |\delta_2\delta_1| \cdot 2a(1) \max_{0<\eta\le m} \frac{\sinh\eta}{\eta}\,.$$

Then (5.33) follows with $C_o(m) \doteq 2a(1) \cdot \frac{\sinh m}{m}$.

Next, assume that a nonphysical wave interacts with a 2-wave. Since the values of u do not change across a 2-wave, the left and right values of u of the nonphysical wave do not change across the interaction; hence the size does not change.

Assume then that a nonphysical wave interacts with a 1- or 3-wave of size δ. Since λ is constant, we refer only to the components v, u. Let (v_ℓ, u_ℓ) and (v_ℓ, u_q) be the side states of the nonphysical wave before the interaction and (v_ℓ, u_q), (v_r, u_r) be the side states of the physical wave. After the interaction, let $(\tilde v_\ell, \tilde u_\ell)$ be the intermediate state. One has

$$u_r - u_q = 2a(\lambda)h(\delta) = \tilde u_\ell - u_\ell\,,$$

and hence $|u_q - u_\ell| = |u_r - \tilde u_\ell|$.

At last, we consider the functional F. The potential Q is unaltered when nonphysical waves interact with other waves. The only cases in which L_{np} changes are when a nonphysical wave arises. Assume that a 1- or a 3-wave of size δ interacts with a 2-wave of size δ_2, producing a wave of the same size and a nonphysical wave. Then $\Delta Q = -|\delta\delta_2|$ and $\Delta L_\xi = K_{np}\Delta L_{np} \le K_{np}C_o|\delta\delta_2|$; hence $\Delta F = \Delta L_\xi + K\Delta Q \le |\delta\delta_2|(K_{np}C_o - K)$, and then (5.34). ∎

5.4. Decreasing of the functional F and control of the variations. We first collect the previous results into a single proposition.

PROPOSITION 5.13 (local decreasing). *Consider the interaction of any two waves either of families 1, 2, 3 or nonphysical. Assume (5.19) for some $m > 0$; let $C_o = C_o(m)$ as in Proposition 5.12. Finally, let A_o satisfy*

$$0 < A_o < 2\frac{1-d}{3-d}\,. \tag{5.35}$$

If ξ, K, K_{np} satisfy

$$\frac{2-A_o}{2-3A_o} < \xi < \frac{1}{d}\,, \qquad \frac{2\xi}{2-A_o} < K < \frac{\xi-1}{A_o}\,, \qquad K_{np} < \frac{K}{C_o}, \tag{5.36}$$

then

$$\Delta F \le 0\,. \tag{5.37}$$

Proof. The condition on K comes from (5.9) and $(5.27)_2$. The interval where K lies is not empty if $A_o < \frac{2}{3}$ and $\xi > \frac{2-A_o}{2-3A_o}$; together with $(5.27)_1$ this gives the assumption required on ξ. In turn, it is possible to choose ξ in such an interval if (5.35) holds. Remark that $2\frac{1-d}{3-d} \le \frac{2}{3}$, so the previous condition on A_o holds. Therefore the assumptions of Propositions 5.3, 5.10, and 5.12 hold, and then (5.37) follows. ∎

PROPOSITION 5.14 (global decreasing). *Let $m > 0$; assume* (5.35), (5.36),

$$L(0+) < \frac{m}{2\xi - 1} \tag{5.38}$$

and that the approximate solution U is defined in $[0,T]$. Then $L(t) < m$ for any $t \in (0,T]$; as a consequence, condition (5.19) *holds for any* 1- *or* 3-*wave in U. Finally, $\Delta F(t) \le 0$ for all times $t \in [0,T]$.*

Proof. Since L may change value only at the times of interaction, we use an induction argument based on Proposition 5.13.

First, we have $L(0+) < m$ because $\xi \ge 1$ and (5.38). Now assume that $L(\tau) < m$ for all $0 < \tau < t$. By Proposition 5.13 one has both $\Delta F(\tau) \le 0$ and $\Delta F(t) \le 0$, so that

$$F(t) \le F(0+) \le \xi L(0+) + KQ(0+)\,.$$

Since $Q(0+) \le L(0+)L_{cd} \le L(0+)A_o$, and recalling $(5.36)_2$, we get

$$F(t) \;\le\; L(0+)\cdot(\xi + KA_o) \;\le\; L(0+)\cdot(2\xi - 1) \;<\; m\,, \tag{5.39}$$

again because of (5.38). Finally, since $L(t) \le L_\xi(t) \le F(t)$ we arrive at the conclusion. ∎

Remark 5.15. From (5.38) we see that, in order to have $L(t) < m$, the smaller ξ is, the larger the $L(0+)$ can be chosen.

Remark 5.16. From Proposition 5.14 we deduce that v remains bounded away from zero. Indeed, recalling (3.6), (3.5) and that v is constant across nonphysical waves, we have

$$\frac{1}{2}\mathrm{TV}(\log v(t,\cdot)) \le L(t) + \mathrm{TV}(\log a_o) \le m + \mathrm{TV}(\log a_o)\,.$$

6. The convergence and the consistence of the algorithm. In this section we prove Theorem 2.2. We first show that for fixed ν the algorithm introduced in section 4 gives an approximate solution defined for every $t > 0$; more precisely, we prove that at every time the number of interactions is bounded. Then we prove that the total amount of nonphysical waves in each approximate solution is very small. The convergence of a suitable subsequence is assured by Helly's theorem; then consistence follows.

6.1. Control of the number of interactions. We prove first that the size of the rarefactions in the scheme is small.

LEMMA 6.1. *Consider a rarefaction of size ε; then*

$$|\varepsilon| < \eta \mathrm{e}^{\frac{A_o}{2}} . \tag{6.1}$$

Proof. We analyze all possible situations. When the rarefaction is generated, one has $0 < \varepsilon < \eta$. When it interacts with a 1- or 3-wave, the size does not increase; see Remark 5.8. By Proposition 5.12 the size does not change when interactions with nonphysical waves occur.

The last case to be considered is when a rarefaction interacts with a 2-wave. In this case the size may increase; however, a rarefaction can meet a fixed 2-wave only once. Consider the case of a 1-rarefaction of size δ_1, as in Proposition 5.3, the other being analogous. If $\delta_2 < 0$, then the size decreases; see (5.13). If $\delta_2 > 0$, by (5.12) we have

$$|\varepsilon_1| = |\delta_1| + |\varepsilon_3| \leq |\delta_1| \left(1 + \frac{1}{2}|\delta_2|\right) < |\delta_1| \mathrm{e}^{\frac{|\delta_2|}{2}} .$$

Summarizing the three cases above, we get $|\varepsilon| < \eta \mathrm{e}^{L_{cd}^{+}/2}$ (or $|\varepsilon| < \eta \mathrm{e}^{L_{cd}^{-}/2}$) for a 1-rarefaction (resp., 3-rarefaction), where $L_{cd}^{\pm}$ is the sum of the 2-waves at the right or left of the rarefaction. Then (6.1) follows. ∎

Next, we prove that the number of interactions remains bounded in finite time, so that the approximate solution is well defined for all $t > 0$. We first give a lemma.

LEMMA 6.2. *Consider the wave-front tracking algorithm described in section* 4, *under the assumptions of Proposition* 5.14. *Then*

(i) *the number of interactions involving a* 2*-wave and solved by the accurate Riemann solver is finite;*

(ii) *the number of interactions where a new rarefaction of size $\varepsilon \geq \eta$ arises is finite.*

Proof. First consider (i) and refer to Proposition 5.3. Then, using (5.16), we have $\Delta F \leq \rho\,(\xi + K(A_o - 2)/2) < 0$, and hence F decreases by a uniform positive quantity; since it is nonincreasing, this can happen only a finite number of times.

Then consider (ii). After (i), it remains only to consider the case of two shocks of the same family interacting.

Under the notation of the corresponding case in the proof of Proposition 5.10 we have $\varepsilon = \varepsilon_1 \geq \eta$ and

$$\Delta F \leq |\varepsilon_1|\,(1 - \xi + KA_o) \leq \eta\,(1 - \xi + KA_o) < 0$$

because of (5.36). Arguing similarly as in (i), this can happen only a finite number of times. ∎

Regarding (ii) in Lemma 6.2, recall that if $\varepsilon \geq \eta$, then the new rarefaction must be split into more than one wave. Therefore Lemma 6.2 can be rephrased by saying that, except for finite interactions, the number of waves emitted in an interaction is at most three, and this case occurs precisely when a nonphysical wave is generated; moreover, in every interaction at most one wave per family is emitted.

In a schematic way, apart from a finite number of interactions, in our algorithm the following hold (we will consider the set of nonphysical waves as a fourth family of waves):

(a) the interaction of an i-wave, $i = 1, 3$, with a 2-wave is solved by a single i-wave, a 2-wave, and a 4-wave;
(b) in the interaction of just 1- and/or 3-waves, there is at most one outgoing wave of each family 1 and 3;
(c) the interaction of a i-wave, $i = 1, 2, 3$, with a 4-wave is solved by an i-wave and a 4-wave.

The next proposition extends the result of Lemma 2.5 in [3].

PROPOSITION 6.3. *Consider the wave-front algorithm described in section* 4 *and assume in the strip* $[0, T) \times \mathbb{R}$ *the following:*

for some $a_1 < a_2 < 0 < b_1 < b_2$ *the waves of the first (resp., third) family have speeds in the interval* $[a_1, a_2]$ *(resp.,* $[b_1, b_2]$*).*

Then the number of interactions in the region $[0, T) \times \mathbb{R}$ *is finite.*

Proof. Assume by contradiction that in the region $[0, T) \times \mathbb{R}$ there exists an infinite number of interactions. Unless we take a smaller T we can assume that the number of interactions is finite in every strip $[0, t] \times \mathbb{R}$, $0 < t < T$, and that T is an accumulation point for the times of interaction. Because of the finite propagation speed, the interaction points are also bounded in space.

Then there exists a bounded sequence (t_j, x_j), $j = 1, 2, \dots$, of interaction points such that

$$0 < t_j < t_{j+1} < T \quad \text{for all } j \quad \text{and } (t_j, x_j) \to (T, \bar{x})$$

for some $\bar{x}$. Denote $\mathcal{J} = \{(t_j, x_j) : j = 1, 2, \dots\}$ the set of all interaction points.

The situation described in items (a)–(c) above holds except in a finite number of interactions; let $\tau < T$ be the maximum time of these "exceptional" interactions. It is not restrictive to assume that $\tau < t_j < T$ for all $j = 1, 2, \dots$.

Starting from a point of $\mathcal{J}$, we "trace back" all the segments up to $t = 0$; we repeat the procedure for all the points of $\mathcal{J}$ and call $\mathcal{F}$ the set of the "traced segments" obtained in this way. In other words, a segment belongs to the set $\mathcal{F}$ iff it can be joined forward in time to some point of $\mathcal{J}$ by a continuous path along the wave fronts. The set $\mathcal{F}$ is not empty: for instance, two segments interacting at the point (t_j, x_j) belong to $\mathcal{F}$ for any $j = 1, 2, \dots,$. Observe the following dichotomy property of $\mathcal{F}$ which is used just below:

two interacting waves either both belong to $\mathcal{F}$ or none of them does; moreover, if at least one of the outgoing waves belong to $\mathcal{F}$, then both the incoming waves must belong to $\mathcal{F}$.

We now partition all the interaction points of the algorithm that occur for times $t > \tau$ into the following sets:

- $\mathcal{I}_0$: the interaction points where no ingoing wave belongs to $\mathcal{F}$;
- $\mathcal{I}_1$: the interaction points where both incoming waves belong to $\mathcal{F}$ and at most one outgoing segment belongs to $\mathcal{F}$;
- $\mathcal{I}_2$: the interaction points where exactly two outgoing segments belong to $\mathcal{F}$;
- $\mathcal{I}_3$: the interaction points where three outgoing waves all belong to $\mathcal{F}$.

Because of the dichotomy property quoted above no outgoing wave in case $\mathcal{I}_0$ can belong to $\mathcal{F}$. On the contrary, both incoming waves in $\mathcal{I}_1$, $\mathcal{I}_2$, $\mathcal{I}_3$ must belong to $\mathcal{F}$.

Recall that we are considering times $t > \tau$. Therefore the maximum number of emitted waves in an interaction is three, and this happens only in the situation considered in $\mathcal{I}_3$, that is, for interactions as in (a) above. The case of more than one emitted wave per family cannot occur, and so the outgoing waves in $\mathcal{I}_2$ belong to different families. By definition we have $\mathcal{J} \cap \mathcal{I}_0 = \emptyset$, and so $\mathcal{J} \subset \mathcal{I}_1 \cup \mathcal{I}_2 \cup \mathcal{I}_3$.

Let $\mathcal{V}(t)$ be the total number of wave fronts of the families 1, 2, and 3 that belong to $\mathcal{F}$ at time t. The functional $\mathcal{V}(t)$ is nonincreasing, and it decreases at least by 1 across $\mathcal{I}_1$. Then $\mathcal{I}_1$ is finite. As a consequence, all the interaction points of $\mathcal{J}$ belong to $\mathcal{I}_2 \cup \mathcal{I}_3$, except at most a finite number. Let $\tau_1 \in [\tau, T)$ be a time such that all points in $\mathcal{I}_1$ lie in $t < \tau_1$.

Let $P = \{x_1, \dots, x_{N_1}\}$ be the set of points of the x-axis where a 2-wave is located. We consider two cases and make use of [3, Lemma 2.5].

$\boxed{\bar{x} \notin P}$. In this case we can choose a time $\tau_1 < T$ such that after that time no segment belonging to $\mathcal{F}$ crosses a 2-wave. Then all the points in $\mathcal{J}$ with $t_j > \tau_1$ belong to $\mathcal{I}_2$.

Take a point $(t^*, x^*) \in \mathcal{J}$ with $t^* > \tau_1$, so that $(t^*, x^*) \in \mathcal{I}_2$. At (t^*, x^*) there are two outgoing segments, both of them belonging to $\mathcal{F}$; for $t > t^*$ and t close to t^* we define $\gamma_\ell(t)$ to be the segment on the left and $\gamma_r(t)$ that on the right. When $\gamma_\ell(t)$ (resp., $\gamma_r(t)$) reaches an interaction point, we prolong it by the segment of $\mathcal{F}$ outgoing from that point and located on the left (resp., on the right).

In this way we define recursively for any $t < T$ two paths: $\gamma_\ell(t)$ is made by segments of the families 1 or 3, while $\gamma_r(t)$ is made by segments of the families 3 or 4.

We claim that the speeds of the two paths are strictly separated. Indeed, if γ_r starts following a 3-segment, then γ_ℓ starts with a 1-segment and so will always follow 1-segments; therefore $\dot\gamma_\ell(t) \le a_2 < 0 < b_1 \le \dot\gamma_r(t)$. If γ_r starts following a 4-segment, then it will always follow 4-segments; in this case $\dot\gamma_\ell(t) \le b_2 < \hat{s} = \dot\gamma_r(t)$.

Thus for $c = \min\{\hat{s} - b_2, b_1 - a_2\} > 0$ we have $\gamma_r(t) - \gamma_\ell(t) \ge c(t - t^*)$. Now set $\rho = c(T - t^*)$ and choose $t_n \in (t^*, T)$, with $(t_n, x_n) \in \mathcal{J}$, such that

$$c(t_n - t^*) > \frac{\rho}{2}, \qquad (T - t_n)(|a_1| + \hat{s}) < \frac{\rho}{4}. \tag{6.2}$$

By definition of $\mathcal{F}$, the points $(t_n, \gamma_\ell(t_n))$ and $(t_n, \gamma_r(t_n))$ can be joined forward in time to points (t_h, x_h), resp., (t_k, x_k), of $\mathcal{J}$, with $h, k > n$. Then

$$x_h \le \gamma_\ell(t_n) + \hat{s}(t_h - t_n), \qquad x_k \ge \gamma_r(t_n) - |a_1|(t_k - t_n).$$

Thanks to (6.2), we get

$$x_k - x_h \ge c(t_n - t^*) - (|a_1| + \hat{s})(T - t_n) \; > \; \frac{\rho}{4}.$$

Since n can be taken arbitrarily large, the last inequality contradicts the convergence of the sequence x_j.

$\boxed{\bar{x} \in P}$. Consider case (a) above. The possibility that the outgoing waves belonging to $\mathcal{F}$ are precisely one physical and one nonphysical wave may happen only a finite number of times, since the functional $\mathcal{V}(t)$ is nonincreasing. Therefore we can assume that, for $t > \tau_1$, in case (a) the two outgoing physical waves belong to $\mathcal{F}$.

Let $(t^*, x^*) \in \mathcal{J}$ with $t^* > \tau_1$. As before we define for $t \in [t^*, T)$ two continuous paths $\gamma_\ell(t)$, $\gamma_r(t)$ starting at (t^*, x^*) in the following way.

At (t^*, x^*) there are either two or three outgoing segments belonging to $\mathcal{F}$; for times $t > t^*$ and sufficiently close to t^* we define $\gamma_\ell(t)$ to be the segment on the left and $\gamma_r(t)$ the one on the right. When $\gamma_\ell(t)$ ($\gamma_r(t)$) reaches an interaction point, it is prolonged by the segment of $\mathcal{F}$ on the left (resp., on the right). Then the path $\gamma_\ell(t)$ is made by segment of the families 1, 2, or 3, while $\gamma_r(t)$ is made by segments of families 3 or 4; in fact the interaction of a 1-wave with a 2-wave always produces a 4-wave (except in a finite number of cases).

Now we prove that the speeds of the paths $\gamma_\ell(t)$, $\gamma_r(t)$ are strictly separated. Indeed, if $\gamma_r(t)$ starts following a 3-segment, then γ_ℓ starts with either a 2- or a 1-segment and, by the remark at the beginning of this subcase, $\dot\gamma_\ell(t) \le 0 < b_1 \le \dot\gamma_r(t)$. If $\gamma_r(t)$ starts following a 4-segment, then $\dot\gamma_\ell(t) \le b_2 < \hat s = \dot\gamma_r(t)$.

Finally, for $c = \min\{\hat s - b_2, b_1\}$ the same argument exploited in the other case leads to a contradiction. □

6.2. Control of the total size of nonphysical fronts. Assume as above that the assumptions of Proposition 5.14 hold. We assign inductively to each wave α a *generation order* k_α as in [9, page 140]. This is done according to the following procedure. First, at time $t = 0$ each wave has order 1. Second, assume that two waves α and β interact at time t; if α and β belong to different families, the outgoing waves of those families keep the order of the incoming waves, and the other waves assume order $\max\{k_\alpha, k_\beta\} + 1$; if α and β belong to the same family, and the outgoing wave of that family takes the order $\min\{k_\alpha, k_\beta\}$, and the other waves are assigned order $\max\{k_\alpha, k_\beta\} + 1$.

When specialized to the current setting this has the following consequences:

- every 2-wave has order 1; when an i-wave, $i = 1, 3$, of order k interacts with a 2-wave, the outgoing i-wave has order k, and the other outgoing wave (of the family j, $j = 1, 3$, $j \neq i$, or a nonphysical wave) has order $k + 1$;
- in the interaction of a 1- with a 3-wave the waves cross without changing order; in the interaction of two waves α, β of the same family $i = 1, 3$, the outgoing wave of the family i takes order $\min\{k_\alpha, k_\beta\}$, and the wave of the family $j = 1, 3$, $j \neq i$, has order $\max\{k_\alpha, k_\beta\} + 1$;
- when a nonphysical wave interacts with any other wave, both waves cross without changing order; in particular a nonphysical wave keeps the order it has been assigned when generated.

For $t \ge 0$ not an interaction time and any $k = 1, 2, \dots$ define (see (5.2), (5.1))

$$V_k(t) = \sum_{\substack{\gamma>0\\ k_\gamma=k}} |\gamma| + \xi \sum_{\substack{\gamma<0\\ k_\gamma=k}} |\gamma| + K_{np} \sum_{\substack{\gamma\in\mathcal{NP}\\ k_\gamma=k}} |\gamma|,$$

$$Q_k(t) = \sum_{\substack{\gamma_3 \text{ at the left of } \delta_2\\ k_{\gamma_3}=k}} |\gamma_3||\delta_2| + \sum_{\substack{\gamma_1 \text{ at the right of } \delta_2\\ k_{\gamma_1}=k}} |\delta_2||\gamma_1|,$$

$$F_k(t) = V_k(t) + KQ_k(t),$$

and

$$\tilde V_k(t) = \sum_{\ell\ge k} V_\ell(t)\,, \qquad \tilde Q_k(t) = \sum_{\ell\ge k} Q_\ell(t)\,, \qquad \tilde F_k(t) = \tilde V_k(t) + K\tilde Q_k(t)\,.$$

We remark that

$$\tilde F_1(0+) = L_\xi(0+) + KQ(0+), \qquad \tilde F_k(0+) = 0 \qquad \text{for } k \ge 2.$$

Observe that if a nonphysical front interacts with another wave, the functionals above do not change; the same holds for interactions between 3- and 1-waves. Then we focus on interactions of waves of the same i family, $i = 1, 3$ (as usual denote $j = 1, 3$, $j \neq i$), and on interactions between 1- or 3-waves with a 2-wave.

For $h \in \mathbb{N}$, denote by I_h the set of times t when an interaction occurs between two waves α and β of families 1 or 3 with $\max\{k_\alpha, k_\beta\} = h$; denote by J_h the set of interaction times t of a 1- or 3-wave of order h with a 2-wave. Finally, denote $\mathcal{T}_h = I_h \cup J_h$ and $I = \bigcup_{h\geq 1} I_h$, $J = \bigcup_{h\geq 1} J_h$, $\mathcal{T} = I \cup J$.

In order to control the total size of nonphysical fronts we must strengthen the assumptions (5.35), (5.36) required in Proposition 5.14. First, for any fixed $m > 0$, instead of (5.35) we require the stronger condition

$$0 < A_o < \frac{1-\sqrt{d}}{2-\sqrt{d}}\,. \tag{6.3}$$

Then denote

$$\lambda \doteq \frac{1+KA_o}{\xi}\,, \qquad \lambda_2 \doteq \frac{\xi + KA_o}{K(2-A_o)-\xi}\,, \qquad \mu \doteq \max\left\{\lambda\,, \lambda_2, \frac{K_{np}C_o}{K}\right\}\,. \tag{6.4}$$

We need $0 < \mu < 1$. From $(5.36)_3$ we have $K_{np}C_o < K$; moreover, we have $0 < \lambda < 1$ and $\lambda_2 > 0$ because of $(5.36)_2$. At last $\lambda_2 < 1$ holds iff $\frac{\xi}{1-A_o} < K$; this condition is stronger than the left inequality in $(5.36)_2$. Hence, instead of (5.36) we assume

$$\frac{1-A_o}{1-2A_o} < \xi < \frac{1}{\sqrt{d}}\,, \qquad \frac{\xi}{1-A_o} < K < \frac{\xi-1}{A_o}\,, \qquad K_{np} < \frac{K}{C_o}\,. \tag{6.5}$$

Remark the new upper bound required on ξ. As in the proof of Proposition 5.13, the interval where K varies is not empty if $A_o < \frac{1}{2}$ and $\xi > \frac{1-A_o}{1-2A_o}$. In turn, we can find ξ satisfying $(6.5)_1$ if (6.3) holds; remark that $\frac{1-\sqrt{d}}{2-\sqrt{d}} \leq \frac{1}{2}$. The last condition in (6.5) coincides with that in (5.36).

Remark 6.4. Proposition 5.14 still holds under the stronger assumptions (6.3), (6.5) under the same condition (5.38), because the inequality on the right-hand side in $(6.5)_2$ has not changed; see (5.39).

PROPOSITION 6.5. *Fix $m > 0$ and assume* (6.3), (6.5). *We have the following:*

1. *$\tau \in \mathcal{T}_h$, $h \leq k-2$: then $\Delta\tilde{F}_k = \Delta F_k = 0$.*
2. *$\tau \in \mathcal{T}_{k-1}$: then $\Delta F_{k-1} < 0$, $\Delta\tilde{F}_k = \Delta F_k > 0$, and*

$$[\Delta\tilde{F}_k]_+ \leq \mu\left([\Delta F_{k-1}]_- - \sum_{\ell=1}^{k-2}[\Delta F_\ell]_+\right)\,. \tag{6.6}$$

3. *$\tau \in \mathcal{T}_h$, $h \geq k$: if $h = k$, then $\Delta F_k < 0$; in any case $\Delta\tilde{F}_k < 0$ and*

$$\sum_{\ell=1}^{k-1}[\Delta F_\ell]_+ < [\Delta\tilde{F}_k]_-\,. \tag{6.7}$$

Proof. As we pointed out above, for $\tau \in I$ only interactions of waves of the same family are taken into account; see Figure 6.1. Remark that by Proposition 5.14 we have

$$\sum_{\ell=1}^{k-1}\Delta F_\ell \;+\; \Delta\tilde{F}_k < 0\,. \tag{6.8}$$

1. If $h \leq k-2$, no waves with order $\geq k$ are involved, and then $\Delta\tilde{F}_k = \Delta F_k = 0$.
2. Let $h = k-1$. First, consider $\tau \in I_{k-1}$; then $\Delta\tilde{V}_k = \Delta V_k > 0$. We prove that

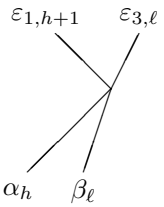


FIG. 6.1. *Interactions of 3-waves; $h \geq \ell$ denote generation orders.*

$$[\Delta \tilde{V}_k]_+ \leq \frac{1}{\xi} \left([\Delta V_{k-1}]_- - \sum_{\ell=1}^{k-2} [\Delta V_\ell]_+ \right) . \tag{6.9}$$

Indeed, from (5.30)–(5.32), we deduce that

$$\xi[\Delta \tilde{V}_k]_+ + \Delta V_{k-1} + \sum_{\ell=1}^{k-2} \Delta V_\ell \leq 0 . \tag{6.10}$$

If $\min\{k_\alpha, k_\beta\} = k-1$, then $\Delta V_\ell = 0$ for $\ell = 1, \dots, k-2$ and (6.10) becomes $\xi[\Delta \tilde{V}_k]_+ + \Delta V_{k-1} < 0$; this implies $\Delta V_{k-1} < 0$ and hence $[\Delta \tilde{V}_k]_+ < (1/\xi)[\Delta V_{k-1}]_-$, that is, (6.9). Moreover, in this case, one easily finds that $\Delta Q_{k-1} \leq 0$, because of (5.21).

If $\min\{k_\alpha, k_\beta\} = \ell \leq k-2$, then $\Delta V_{k-1} < 0$ and $\Delta Q_{k-1} < 0$, since no waves of order $k-1$ are present after the interaction. Therefore the estimate (6.10) becomes

$$\xi[\Delta \tilde{V}_k]_+ - [\Delta V_{k-1}]_- + \Delta V_\ell \leq 0 .$$

If $\Delta V_\ell \geq 0$, we get (6.9). If $\Delta V_\ell < 0$, we check directly that $\xi[\Delta \tilde{V}_k]_+ - [\Delta V_{k-1}]_- \leq \xi^2 |\varepsilon_1| - |\alpha_3| < 0$ because of (5.20) and $(6.5)_1$. This completes the proof of (6.9).

From this proof we see also that $\Delta F_{k-1} = \Delta V_{k-1} + K \Delta Q_{k-1} < 0$, since both terms in the sum are negative. Then we use (6.9) and $0 \leq \Delta \tilde{Q}_k \leq A_o \Delta \tilde{V}_k$ to get

$$0 < \Delta \tilde{F}_k \leq (1 + KA_o)[\Delta \tilde{V}_k]_+ \leq \lambda \left([\Delta V_{k-1}]_- - \sum_{\ell=1}^{k-2} [\Delta V_\ell]_+ \right) . \tag{6.11}$$

We now prove that

$$[\Delta Q_{k-1}]_- - \sum_{\ell=1}^{k-2} [\Delta Q_\ell]_+ \geq 0 . \tag{6.12}$$

We have only to consider the case in which $[\Delta Q_\ell]_+ > 0$ for some $\ell \leq k-2$; but in this case $[\Delta Q_{k-1}]_- - \sum_{\ell=1}^{k-2} [\Delta Q_\ell]_+ = L_{cd}^+ (|\alpha_3| + |\beta_3| - |\varepsilon_3|) \geq 0$ because of (5.21).

Therefore (6.6) for $\tau \in I_{k-1}$ follows from (6.11) and (6.12).

Second, assume $\tau \in J_{k-1}$; we prove that

$$[\Delta \tilde{F}_k]_+ \leq \mu [\Delta F_{k-1}]_- . \tag{6.13}$$

Indeed, if the reflected wave is a physical wave, then, under the notation of Proposition 5.3,

$$\Delta V_{k-1} \leq \xi \frac{|\delta_1 \delta_2|}{2} , \qquad \Delta Q_{k-1} \leq -|\delta_1 \delta_2| + \frac{|\delta_1 \delta_2|}{2} A_o = -\frac{|\delta_1 \delta_2|}{2} (2 - A_o)$$

so that $\Delta F_{k-1} \le [\xi - K(2-A_o)]\,|\delta_1\delta_2|/2 < 0$ because of (6.5). Then (6.13) follows since $\Delta\tilde F_k = \Delta F_k > 0$ and

$$[\Delta\tilde F_k]_+ \le \frac{|\delta_1\delta_2|}{2}(\xi + KA_o) = \frac{|\delta_1\delta_2|}{2}\,[K(2-A_o)-\xi]\cdot\lambda_2 \le \lambda_2[\Delta F_{k-1}]_- \le \mu[\Delta F_{k-1}]_-\,.$$

If the reflected wave is a nonphysical wave, we have, under the notation of Proposition 5.12,

$$0 < \Delta F_k = \Delta V_k \le K_{np}C_o|\delta\delta_2|\,, \qquad \Delta V_{k-1} = 0\,, \qquad \Delta Q_{k-1} = -|\delta\delta_2|,$$

and then

$$[\Delta F_{k-1}]_- = K|\delta\delta_2|\,, \qquad [\Delta F_k]_+ \le \frac{K_{np}C_o}{K}[\Delta F_{k-1}]_- \le \mu[\Delta F_{k-1}]_-\,.$$

The estimate (6.13) is then completely proved. From (6.13) we get (6.6) since no 1-, 3-, or nonphysical waves of order $\le k-2$ are involved in the interaction.

3. Finally, let $h \ge k$. We first consider $\tau \in I_h$. If $\min\{k_\alpha, k_\beta\} \ge k$, then $\Delta\tilde V_k = \Delta L_\xi < 0$ and $\Delta\tilde F_k = \Delta F < 0$. If $\min\{k_\alpha, k_\beta\} \le k-1$, assume $k_\alpha \ge k$ and $k_\beta \le k-1$; then $\Delta\tilde V_k \le \xi|\varepsilon_1| - |\alpha_3| \le (\xi d - 1)|\alpha_3| < 0$ by (5.20) and $(5.36)_1$. Moreover, $\Delta\tilde Q_k = |\varepsilon_1|L^-_{cd} - |\alpha_3|L^+_{cd} \le |\varepsilon_1|A_o \le dA_o|\alpha_3|$ again by (5.20). Then

$$\Delta\tilde F_k \;\le\; [(\xi + KA_o)d - 1]\cdot|\alpha_3|,$$

and since, because of (6.5),

$$(\xi + KA_o)d - 1 < (2\xi - 1)d - 1 < \left(\frac{2}{\sqrt{d}} - 1\right)d - 1 = -(1-\sqrt{d})^2 < 0\,,$$

we proved that $\Delta\tilde F_k < 0$. Now from (6.8) we deduce that $\sum_{\ell=1}^{k-1}\Delta F_\ell < [\Delta\tilde F_k]_-$. Since at most one nonzero term is present in the first sum, (6.7) follows. If $h = k$, we are in the same situation as in case 2, so we have $\Delta F_k < 0$.

Now assume $\tau \in J_h$. Then $\Delta\tilde F_k = \Delta F < 0$ and no 1-, 3-, or nonphysical waves of order $< k$ are present, so (6.7) holds. If $h = k$ and the reflected wave is physical, from the proof of Proposition 5.3 and (6.5) we find that

$$\Delta F_k = \Delta V_k + K\Delta Q_k \le \frac{|\delta_1\delta_2|}{2}\,(\xi - 2K + KA_o) < 0\,.$$

If $h = k$ and the reflected wave is nonphysical, then $\Delta V_k = 0$, $\Delta Q_k < 0$, and $\Delta F_k < 0$. □

Summarizing, for $\tau \in \mathcal{T}$ we have the following table:

$\mathcal{T}_h$; h	$\le k-2$	$k-1$	k	$\ge k+1$
ΔF_k	0	+	−	±
$\Delta\tilde F_k$	0	+	−	−

We write $\tilde F_k^{\pm}(t) = \sum_{\tau\le t}[\Delta\tilde F_k(\tau)]_\pm$ for $k \ge 2$. For simplicity the time τ in such sums is omitted.

LEMMA 6.6. *Under the assumptions of Proposition* 6.5 *we have*

$$\tilde F_2^+(t) \le \mu\left(L_\xi(0) + KQ(0)\right) + \sum_{\mathcal{T}_h,\, h\ge 2} \left[\Delta F_1\right]_+, \tag{6.14}$$

$$\tilde F_k^+(t) \le \mu\left(\tilde F_{k-1}^+(t) + \sum_{\mathcal{T}_h,\, h\ge k} \left[\Delta F_{k-1}\right]_+ - \sum_{\mathcal{T}_{k-1}} \sum_{\ell=1}^{k-2} \left[\Delta F_\ell\right]_+\right), \qquad k \ge 3\,. \tag{6.15}$$

Proof. By Proposition 6.5 (see also the table above), the functional F_k increases at times $\tau \in \mathcal{T}_{k-1}$ and decreases at times $\tau \in \mathcal{T}_k$, while it does not have a given sign at times $\tau \in \mathcal{T}_h$, with $h \ge k+1$.

First, by summing up (6.6) we obtain

$$\tilde F_k^+(t) \le \mu \sum_{\mathcal{T}_{k-1}} \left(\left[\Delta F_{k-1}\right]_- - \sum_{\ell=1}^{k-2} \left[\Delta F_\ell\right]_+\right) \tag{6.16}$$

for $k \ge 2$, where the last term in (6.16) is missing if $k = 2$.

Now recall that $F_1(0) = L_\xi(0) + KQ(0)$; therefore

$$F_1(t) \le L_\xi(0) + KQ(0) - \sum_{\mathcal{T}_1} \left[\Delta F_1\right]_- + \sum_{\mathcal{T}_h,\, h\ge 2} \left[\Delta F_1\right]_+$$

and then

$$\sum_{\mathcal{T}_1} \left[\Delta F_1\right]_- \le L_\xi(0) + KQ(0) + \sum_{\mathcal{T}_h,\, h\ge 2} \left[\Delta F_1\right]_+ . \tag{6.17}$$

On the other hand, $F_k(0) = 0$ for $k \ge 2$; from Proposition 6.5 we have

$$F_k(t) \le \sum_{\mathcal{T}_{k-1}} \left[\Delta F_k\right]_+ - \sum_{\mathcal{T}_k} \left[\Delta F_k\right]_- + \sum_{\mathcal{T}_h,\, h\ge k+1} \left[\Delta F_k\right]_+ .$$

Moreover,

$$\sum_{\mathcal{T}_{k-1}} \left[\Delta F_k\right]_+ = \sum_{\mathcal{T}_{k-1}} \left[\Delta \tilde F_k\right]_+ = \tilde F_k^+(t)$$

and then

$$\sum_{\mathcal{T}_k} \left[\Delta F_k\right]_- \le \tilde F_k^+(t) + \sum_{\mathcal{T}_h,\, h\ge k+1} \left[\Delta F_k\right]_+ . \tag{6.18}$$

From (6.16), (6.17), (6.18) we get (6.14), (6.15). □

Proposition 6.7 (a contraction property). *Under the assumptions of Proposition* 6.5, *for any* $t \ge 0$ *and* $k \ge 1$ *we have*

$$\tilde V_k(t) \le \tilde F_k(t) \le \mu^{k-1} \cdot \left(L_\xi(0) + KQ(0)\right) . \tag{6.19}$$

Proof. The estimate (6.19) holds for $k = 1$ because $\tilde F_1(t) = F(t) \le L_\xi(0) + KQ(0)$. Next, we prove by induction on $k \ge 2$ that for any t

$$\tilde F_k^+(t) \le \mu^{k-1}\left(L_\xi(0) + KQ(0)\right) + \sum_{\mathcal{T}_h,\, h\ge k} \sum_{\ell=1}^{k-1} \left[\Delta F_\ell\right]_+ . \tag{6.20}$$

Since by summing up (6.7) we obtain

$$\tilde{F}_k^-(t) \geq \sum_{\mathcal{T}_h,\, h\geq k} \sum_{\ell=1}^{k-1} [\Delta F_\ell]_+, \tag{6.21}$$

then (6.19) will follow from (6.20) for any $k \geq 2$ because of (6.21).

Formula (6.20) for $k = 2$ reduces to (6.14). Next, assume that (6.20) holds for some $k \geq 2$. By (6.15) and the induction assumption

$$\begin{aligned}
\tilde{F}_{k+1}^+(t) \;&\leq\; \mu \left(\tilde{F}_k^+(t) + \sum_{\mathcal{T}_h,\, h\geq k+1} [\Delta F_k]_+ - \sum_{\mathcal{T}_k} \sum_{\ell=1}^{k-1} [\Delta F_\ell]_+ \right) \\
&\leq \mu^k \left(L_\xi(0) + KQ(0) \right) + \mu \left(\sum_{\mathcal{T}_h,\, h\geq k} \sum_{\ell=1}^{k-1} [\Delta F_\ell]_+ + \sum_{\mathcal{T}_h,\, h\geq k+1} [\Delta F_k]_+ - \sum_{\mathcal{T}_k} \sum_{\ell=1}^{k-1} [\Delta F_\ell]_+ \right) \\
&\leq \mu^k \left(L_\xi(0) + KQ(0) \right) + \mu \sum_{\mathcal{T}_h,\, h\geq k+1} \sum_{\ell=1}^{k} [\Delta F_\ell]_+ .
\end{aligned}$$

Since $\mu < 1$, we get (6.20) for $k+1$. □

Remark 6.8. We now comment on the case $A_o = 0$. In this case system (1.1) reduces to the p-system with pressure law given by (2.1) for fixed λ. According to our front-tracking algorithm, stationary and nonphysical waves do not appear, and so $L_{cd} = Q = 0$; the algorithm reduces to the one introduced in [3]. Then Proposition 6.5 holds with $\tilde{V}_k$ and $1/\xi$ replacing $\tilde{F}_k$ and μ, respectively, and at last (6.19) reads

$$\tilde{V}_k(t) \leq \frac{1}{\xi^{k-1}} \cdot L_\xi(0)\,.$$

Next, we conclude that the total strength of all nonphysical waves is small by proceeding as in [9, page 142]. Recall the notation in section 4. First, by Remark 5.16, the sequence $\{v^\nu\}$ is uniformly bounded from above and away from 0; then the eigenvalues e_1 and e_3 are bounded, and this makes possible the choice of a suitable $\hat{s}$. We have two more parameters η, ρ to be chosen. Fix $\eta > 0$ with the condition $\eta = \eta_\nu \to 0$ as $\nu \to \infty$ and estimate the total number of waves of order $< k$. We have

$$\begin{aligned}
\sum_{\gamma\in\mathcal{NP}} |\gamma|(t) &\leq \tilde{V}_k(t) \;+ \sum_{\gamma\in\mathcal{NP},\, k_\gamma<k} |\gamma|(t) \\
&\leq \mu^{k-1} \cdot \left(L_\xi(0) + KQ(0) \right) \;+\; C_o\rho \cdot [\text{number of fronts of order} < k] \leq \frac{1}{\nu}
\end{aligned}$$

by choosing k sufficiently large to have the first term $\leq 1/(2\nu)$ and then choose ρ small enough to have the second term $\leq 1/(2\nu)$.

We now accomplish the proof of Theorem 2.2. First define

$$k(m) = \frac{1 - \sqrt{d(m)}}{2 - \sqrt{d(m)}}\,. \tag{6.22}$$

From the properties of the function $d(m)$ stated in Remark 5.7 we see that $k(0) = 1/2$ and that $k(m)$ is decreasing, tending to 0 for $m \to +\infty$. The assumption (2.8) implies that (6.3) holds.

Now, by hypotheses (2.7) it follows that we can choose ξ such that

$$\frac{1}{2}\mathrm{TV}\log(p_o) + \frac{1}{2\inf a_o}\mathrm{TV}(u_o) < \frac{m}{2\xi - 1} < \big(1 - 2A_o\big)m$$

and that $(6.5)_1$ holds. Hence, using (3.12) and (i) in section 4, we have

$$L(0+) \le \frac{1}{2}\mathrm{TV}\log(p_o) + \frac{1}{2\inf a_o}\mathrm{TV}(u_o) < \frac{m}{2\xi - 1}$$

so that the hypotheses (5.38) of Proposition 5.14 hold. Theorem 2.2 now follows along the lines of [9, section 7.4].

Appendix A. The weighted total variation. In this appendix we prove Proposition 2.1. Remark that the map $d(a,b) \doteq \frac{|a-b|}{a+b}$ is a distance on $\mathbb{R}_+$, as one can easily prove.

We start with the proof of the inequality on the right in (2.5). It is enough to prove that

$$2\sum_{j=1}^{n} \frac{|f(x_j) - f(x_{j-1})|}{f(x_j) + f(x_{j-1})} \le \sum_{j=1}^{n} |\log f(x_j) - \log f(x_{j-1})| \,. \tag{A.1}$$

We claim that

$$\log t \ge \frac{2(t-1)}{t+1} \quad \text{for } t \ge 1, \tag{A.2}$$

where the inequality is strict if $t > 1$. To prove the claim it is sufficient to notice that the function $\phi(t) = \log t - 2\frac{t-1}{t+1}$ vanishes in 1 and $\phi'(t) = \frac{(t-1)^2}{t(t+1)^2} > 0$ if $t > 1$.

We apply (A.2) to $t = x/y$ for $0 < y \le x$ and arguing by symmetry deduce that

$$|\log x - \log y| \ge \frac{2|x-y|}{x+y} \quad \text{for every } x, y > 0\,.$$

Then (A.1) follows. The proof of the inequality on the left in (2.5) is analogous, starting from the inequality

$$\frac{1}{t}\log t \le \frac{2(t-1)}{t+1} \quad \text{for } t \ge 1 \tag{A.3}$$

with strict inequality if $t > 1$.

Now assume that $f \in C(\mathbb{R})$. To show that $\mathrm{WTV}(f) = \mathrm{TV}\,(\log(f))$, we have to prove that the inequality

$$\mathrm{TV}\,(\log(f)) \le 2\sup \sum_{j=1}^{n} \frac{|f(x_j) - f(x_{j-1})|}{f(x_j) + f(x_{j-1})} \tag{A.4}$$

holds for any $[a,b] \subset \mathbb{R}$, the supremum being taken on the set of all partitions $a = x_o < x_1 < \cdots < x_n = b$, $n \in \mathbb{N}$. Consider any such partition; by the mean value theorem and by the intermediate value theorem applied to f, we get

$$\begin{aligned} |\log(f(x_j)) - \log(f(x_{j-1}))| &= \frac{|f(x_j) - f(x_{j-1})|}{\zeta_j} \\ &= \frac{f(x_j) + f(x_{j-1})}{2f(\eta_j)} \cdot 2\frac{|f(x_j) - f(x_{j-1})|}{f(x_j) + f(x_{j-1})} \end{aligned}$$

for some ζ_j between $f(x_j)$ and $f(x_{j-1})$ and $\eta_j \in [x_{j-1}, x_j]$. We exploit again the continuity of f in $[a,b]$. On one hand, its image is compact; then $\min_{[a,b]} f = m > 0$. On the other hand, f is uniformly continuous in $[a,b]$, so that for any $\varepsilon > 0$ there exists $\delta_\varepsilon > 0$ such that $|f(x) - f(y)| < \varepsilon$ if $|x - y| < \delta_\varepsilon$ for $x, y \in [a,b]$.

Now fix any $\varepsilon > 0$; without loss of generality we can consider partitions of the interval $[a,b]$ of mesh less than δ_ε. Assume for instance $f(x_{j-1}) \le f(x_j)$; then from the inequalities

$$f(x_j) - f(x_{j-1}) \ < \ \varepsilon\,, \qquad f(x_{j-1}) \le f(\eta_j) \le f(x_j)$$

it follows that

$$\frac{f(x_j) + f(x_{j-1})}{2f(\eta_j)} \le \frac{2f(x_{j-1}) + \varepsilon}{2f(x_{j-1})} \le 1 + \frac{\varepsilon}{m}\,.$$

The inequality (A.4) follows by remarking that then for any partition of mesh less than δ_ε

$$\sum_{j=1}^{n} |\log(f(x_j)) - \log(f(x_{j-1}))| \le \Big(1 + \frac{\varepsilon}{m}\Big)\, 2 \sum_{j=1}^{n} \frac{|f(x_j) - f(x_{j-1})|}{f(x_j) + f(x_{j-1})}\,.$$

The proof of Proposition 2.1 is complete.

Remark A.1. Observe that if $\mathrm{TV}(\log(f)) < \infty$ and f is discontinuous, then the inequality on the right in (2.5) is strict, because of the strict inequality in (A.2). For example, if f has a single jump and assumes the values $c > 0$ and $d > 0$, then $\mathrm{WTV}(f) = 2\frac{|c-d|}{c+d} < |\log c - \log d| = \mathrm{TV}(\log(f))$.

Remark, moreover, that WTV and TV are not equivalent, in the sense that there does not exist a positive constant C such that $C \cdot \mathrm{TV}(\log(f)) \le \mathrm{WTV}(f)$. This follows from the fact that clearly the inequality $C \log t \le \frac{2(t-1)}{t+1}$ does not hold for every $t \ge 1$.

Appendix B. Shock-rarefaction interactions. In this appendix we consider a particular case of Lemma 5.6. Actually, this is the only case needed in order to define a decreasing functional (see section 5); however, we needed further analysis for the control and treatment of the nonphysical waves.

LEMMA B.1 (the case SR, $RS \to SS$). *Consider the interaction of a shock α_i and a rarefaction β_i of the same family, $i = 1, 3$, producing two outgoing shocks ε_1, ε_3. Then there exists a smooth function B satisfying $|\alpha_i| \le B(\alpha_i) \le \min\{\sinh(|\alpha_i|), 2|\alpha_i|\}$ such that*

$$0 < \beta_i \le B(\alpha_i)\,. \tag{B.1}$$

Moreover, assume

$$|\alpha_i| \le m \tag{B.2}$$

for some $m > 0$ and denote $c = c(m) = \frac{\cosh(m)-1}{\cosh(m)+1}$. Then both the variation of shock waves and the reflected wave ε_j, $j \ne i$, $j = 1, 3$, are estimated by the interacting rarefaction as

$$|\varepsilon_1| + |\varepsilon_3| - |\alpha_i| \le (2c - 1) \cdot |\beta_i|\,, \tag{B.3}$$

$$|\varepsilon_j| \le c \cdot |\beta_i|\,. \tag{B.4}$$

Proof. We focus on the case $i = 3$, $j = 1$; see Figure 5.3(b). Therefore we consider $\alpha_3 < 0$, $\beta_3 > 0$ and $\varepsilon_1 < 0$, $\varepsilon_3 < 0$. Then (5.5), (5.6) become

$$|\varepsilon_1| - |\varepsilon_3| = |\beta_3| - |\alpha_3|, \tag{B.5}$$

$$\sinh(|\varepsilon_1|) + \sinh(|\varepsilon_3|) = \sinh(|\alpha_3|) - |\beta_3|\,. \tag{B.6}$$

From the second equation $|\varepsilon_1| < |\alpha_3|$, $|\varepsilon_3| < |\alpha_3|$, and $|\beta_3| < \sinh(|\alpha_3|)$; using the first equation and $|\varepsilon_3| < |\alpha_3|$, we get $|\varepsilon_1| < |\beta_3|$. Therefore in conclusion

$$|\varepsilon_1| < \min\{|\alpha_3|, |\beta_3|\}, \qquad |\varepsilon_3| < |\alpha_3|, \qquad |\beta_3| < \sinh(|\alpha_3|)\,. \tag{B.7}$$

Step 1: *notation.* We set $x = |\beta_3|$, $y = |\varepsilon_1|$, $z = |\alpha_3|$, so that

$$|\varepsilon_3| = y - x + z\,; \tag{B.8}$$

see Figure B.1. Under this notation, (B.6) writes as

$$F(x, y; z) = \sinh y + \sinh(y - x + z) - \sinh z + x = 0 \tag{B.9}$$

for $x \geq 0$, $y \geq 0$, $z \geq 0$, $y - x + z \geq 0$. By (B.7), any solution of (B.9) satisfies

$$y < z\,, \qquad y < x\,, \qquad x < \sinh(z)\,. \tag{B.10}$$

Step 2: *the threshold.* Observe that, despite the last inequality in (B.7), we may well have $|\beta_3| > |\alpha_3|$, that is, $x > z$. Consider in fact the limit case of $\varepsilon_3 = 0$: we have $y = x - z > 0$ and $\sinh(y) = \sinh(z) - x$ that give

$$\sinh(x - z) = \sinh(z) - x, \qquad x > z\,.$$

The last equality is the relation needed for β_3, α_3 in order to have that the shock and rarefaction cancel out exactly, giving rise only to a wave of the opposite family. Observe that the size of the rarefaction must be larger than the one of the shock. Under the notation above, the threshold curve separating the case of the outgoing waves S_1S_3 from the case S_1R_3 is given by

$$f(x, z) = \sinh(x - z) - \sinh(z) + x = 0\,. \tag{B.11}$$

Since $f_x = \cosh(x - z) + 1 > 0$ and $f(z, z) < 0$, the implicit equation $f(x, z) = 0$ is solved by $x = x_o(z) \geq z$ with $x_o'(z) = \frac{\cosh(x-z)+\cosh(z)}{\cosh(x-z)+1} > 0$ for every $z \geq 0$; the curve has for tangent at $(0, 0)$ the line $z = x$. Observe that $x_o(z) \leq 2z$ because $f(2z, z) = z \geq 0$. In conclusion

$$z \leq x_o(z) \leq 2z\,; \tag{B.12}$$

see Figure B.2(a). This estimate and the third inequality in (B.7) prove (B.1) for $B(\alpha_i) = x_o(|\alpha_i|)$. We can prove more than (B.12), that is,

$$\lim_{z \to +\infty} (x_o(z) - 2z) = 0\,. \tag{B.13}$$

Indeed, we show that the inequality $x_o(z) > 2z - q$ holds for large z and $q > 0$. This follows from $f(2z - q, z) = \sinh(z - q) - \sinh z + 2z - q \sim e^z\,(e^{-q} - 1)\,/2 \to -\infty$ for $z \to \infty$.

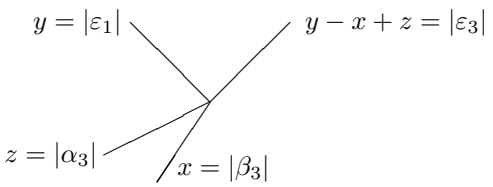


FIG. B.1. *Interactions.*

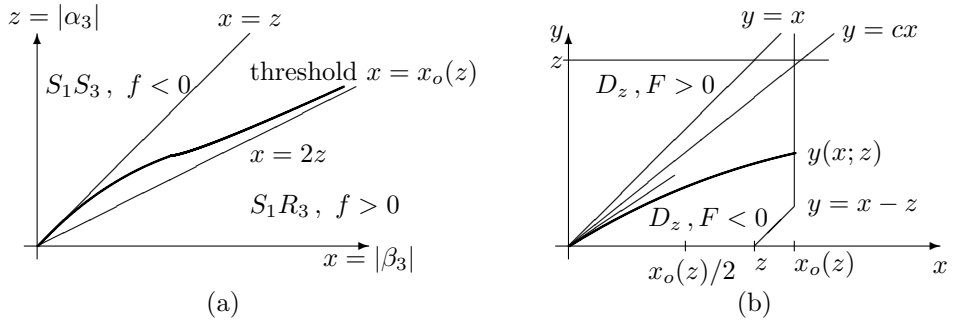


FIG. B.2. (a) *the threshold curve* $f(x, z) = \sinh(x - z) - \sinh(z) + x = 0$*;* (b) *the domain* D_z *and the function* $y = y(x; z)$*.*

Remark that the fact that ε_3 is a shock implies that

$$f(x, z) = \sinh(x - z) - \sinh(z) + x < 0\,. \tag{B.14}$$

Step 3: *the amount of shocks can increase.* From (B.5) we have

$$|\varepsilon_1| + |\varepsilon_3| - |\alpha_3| = 2|\varepsilon_1| - |\beta_3|\,. \tag{B.15}$$

We now prove that the inequality $|\varepsilon_1| + |\varepsilon_3| - |\alpha_3| < 0$, or equivalently $|\varepsilon_1| < \frac{1}{2}|\beta_3|$, does not hold if m is large.

The equation giving $|\varepsilon_1| = y$ in terms of $|\beta_3| = x$, for a given parameter $|\alpha_3| = z$, is (B.9), to be considered in the domain

$$D_z = \{(x, y):\ 0 \le x \le x_o(z)\,,\ y \ge \max\{0, x - z\}\} \qquad \text{for } z \ge 0,$$

where $x_o(z)$ satisfies (B.11); see Figure B.2(b). Since $F_y = \cosh y + \cosh(y - x + z) > 0$, the implicit equation (B.9) defines a function $y = y(x; z)$ with $0 \le y(x; z) \le x$ and $y(x; z) \le z$; see (B.10). Remark that $F(x, x; z) = \sinh x + x > 0$. Moreover, $F(x, 0; z) = \sinh(z - x) - \sinh z + x$ so that $F(0, 0; z) = 0$; by $F_x = 1 - \cosh(z - x) < 0$, we deduce that $F(x, 0; z) < 0$ if $x > 0$. By the implicit function theorem we have $y'(x; z) \ge 0$,

$$y'(0; z) = \frac{\cosh(z) - 1}{\cosh(z) + 1} \in [0, 1), \tag{B.16}$$

and $y''(0; z) = -\frac{4\sinh(z)}{(1+\cosh(z))^2} \le 0$. The function $\frac{\cosh(z)-1}{\cosh(z)+1}$ is increasing, and then for $z \in [0, m]$ its maximum is $\frac{\cosh(m)-1}{\cosh(m)+1}$; this quantity is strictly larger than $1/2$ if $m > \log(3 + 2\sqrt{2})$. Thus in general the estimate $y(x; z) < x/2$ cannot hold.

Step 4: *proof of the estimate.* From (B.15) we see that $|\varepsilon_1| + |\varepsilon_3| - |\alpha_3| \le (2c - 1) \cdot |\beta_3| \iff |\varepsilon_1| \le c \cdot |\beta_3|$, that is, that (B.3) and (B.4) are equivalent; we shall prove (B.4). To bypass the study of the function $y(x; z)$ we define

$$\Phi(x;z,c) = F\left(x, cx; z\right) = \sinh(cx) + \sinh\left(z - (1-c)x\right) - \sinh z + x\,.$$

If $1/2 \le c < 1$, then $z > (1-c)x$ and $\Phi(0;z,c) = 0$,

$$\Phi_x(x;z,c) = 1 + c\cosh(cx) - (1-c)\cosh\left(z-(1-c)x\right)\,,$$

$$\Phi_{xx}(x;z,c) = c^2\sinh(cx) + (1-c)^2\sinh\left(z-(1-c)x\right) > 0\,.$$

Therefore the function $x \to \Phi_x(x;z,c)$ is increasing, and then $\Phi(x;z,c) \ge 0$ if

$$\Phi_x(0;z,c) = (1+c) - (1-c)\cosh(z) > 0\,,$$

that is, if $\cosh(z) \le \frac{1+c}{1-c}$; this is just (B.2). Then $y(x;z) \le cx$ for all $x \in (0, x_o(z))$, and so (B.4) is proved. □

Remark B.2. From (B.16) we see that condition (B.2) is equivalent to the geometric condition $y'(0;z) = (\cosh z - 1)/(\cosh z + 1) < c$. Moreover, as we noticed in the above proof, condition (B.2) is equivalent to

$$\cosh(|\alpha_i|) \le \frac{1+c}{1-c}\,, \tag{B.17}$$

which, in turn, is equivalent to

$$|\alpha_i| \le \log\frac{(1+\sqrt{c})^2}{1-c}\,. \tag{B.18}$$

From the definition of the strength, one has that $|\alpha_i| = (1/2)\log(v_{max}/v_{min})$, where $v_{max} = \max\{v_\ell, v_r\}$, $v_{min} = \min\{v_\ell, v_r\}$, v_ℓ, v_r being, respectively, the left and right values of v for the wave of size α_i. Hence (B.18) is equivalent to

$$\sqrt{\frac{v_{max}}{v_{min}}} \le \frac{\left(1+\sqrt{c}\right)^2}{1-c}\,.$$

Remark B.3. In the proof above we showed that $\Delta L_{\text{shocks}} = |\varepsilon_1| + |\varepsilon_3| - |\alpha_3| = 2|\varepsilon_1| - |\beta_3|$ may be positive, differently from Nishida's paper, where it is always decreasing. This depends on the definition of the wave strengths, which was imposed to us in order to have good estimates when dealing with interactions with the 2-waves. In any case $\Delta L = \Delta L_{\text{shocks}} + \Delta L_{\text{rarefactions}} \le 0$.

Remark B.4. Under the notation and assumptions of Lemma B.1 we verify that

$$|\varepsilon_j| \le c \cdot |\alpha_i|\,. \tag{B.19}$$

This estimate, together with (B.4), allows us to obtain (in a special case) the analogue of (5.20) with c in place of d.

The proof makes use of a numerical computation. As in that lemma we consider the case $j = 1$, $i = 3$. Let $z > 0$ be fixed and $0 \le x \le x_o(z)$. From the proof of Lemma B.1 we deduce that $y = y(x,z)$; for z fixed the function $x \to y(x,z)$ is increasing. Then define

$$Y(z) = y(x_o(z), z) = x_o(z) - z\,.$$

In order to prove (B.19) it is sufficient to prove that

$$Y(z) \le cz \quad \text{if} \quad \cosh(z) \le \frac{1+c}{1-c}\,.$$

Remark that from (B.13) we know that $Y(z)/z \to 1$ for $z \to +\infty$; however, the constraint (B.17) implies that z is bounded. The inequality $Y(z) \le cz$ is equivalent to $x_o(z) - z \le cz$, i.e., $x_o(z) \le (1+c)z$. Therefore we need to prove that

$$\text{(B.20)} \quad \phi(z,c) \doteq \sinh(cz) - \sinh(z) + (1+c)z \ge 0 \quad \text{if} \quad \cosh(z) \le \frac{1+c}{1-c}\,.$$

Notice that $\cosh(z) \le \frac{1+c}{1-c}$ means $0 \le z \le z_c$ for $z_c = \log(\frac{1+c}{1-c} + \sqrt{(\frac{1+c}{1-c})^2 - 1}\,)$. Formula (B.20) is shown to hold true by numerical computations. Remark, however, that

$$\phi'(z,c) = c\cosh(cz) - \cosh(z) + 1 + c,$$

$$\phi''(z,c) = c^2\sinh(cz) - \sinh(z) \le 0,$$

so $\phi'(0,c) = 2c$, $\phi(\cdot,c)$ is concave, $\lim_{z\to+\infty}\phi(z,c) = -\infty$, and $\phi(\cdot,c)$ has a single point of maximum; at last $\phi'(z_c) = c\,(\cosh(cz_c) - \cosh(z_c)) < 0$. This concludes the proof of (B.19).

Remark that if $c = 1/2$, then $\phi(z,c) = \sinh(z/2) - \sinh z + \frac{3}{2}z$. If z is such that $\cosh z = 3$, then $\sinh z = 2\sqrt{2}$, $\sinh(z/2) = 1$, $z = \log(3+\sqrt{8})$, and (B.20) holds with strict inequality.

Acknowledgment. The authors thank Graziano Guerra for stimulating remarks and Umberto Massari for hints on BV functions.

REFERENCES

[1] R. Abeyaratne and J. K. Knowles, *Kinetic relations and the propagation of phase boundaries in solids*, Arch. Rational Mech. Anal., 114 (1991), pp. 119–154.

[2] D. Amadori and A. Corli, *A hyperbolic model of multiphase flow*, in Hyperbolic Problems: Theory, Numerics, Applications, S. Benzoni-Gavage and D. Serre, eds., Springer, Berlin, Heidelberg, 2008, pp. 407–414.

[3] D. Amadori and G. Guerra, *Global BV solutions and relaxation limit for a system of conservation laws*, Proc. Roy. Soc. Edinburgh Sect. A, 131 (2001), pp. 1–26.

[4] F. Asakura, *Wave-front tracking for the equations of isentropic gas dynamics*, Quart. Appl. Math., 63 (2005), pp. 20–33.

[5] F. Asakura, *Wave-front tracking for the equations of non-isentropic gas dynamics*, to appear.

[6] S. Benzoni-Gavage, *Analyse numérique des modèles hydrodynamiques d'écoulements diphasiques instationnaires dans les réseaux de production pétrolière*, Ph.D. thesis, Ecole Normale Supérieure de Lyon, Lyon, France, 1991.

[7] F. Bereux, E. Bonnetier, and P. G. LeFloch, *Gas dynamics system: Two special cases*, SIAM J. Math. Anal., 28 (1997), pp. 499–515.

[8] S. Bianchini, *The semigroup generated by a Temple class system with non-convex flux function*, Differential Integral Equations, 13 (2000), pp. 1529–1550.

[9] A. Bressan, *Hyperbolic Systems of Conservation Laws. The One-Dimensional Cauchy Problem*, Oxford University Press, Oxford, UK, 2000.

[10] A. Corli and H. Fan, *The Riemann problem for reversible reactive flows with metastability*, SIAM J. Appl. Math., 65 (2004), pp. 426–457.

[11] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, 2nd ed., Grundlehren Math. Wiss., 325, Springer-Verlag, Berlin, 2005.

[12] R. J. DiPerna, *Existence in the large for quasilinear hyperbolic conservation laws*, Arch. Rational Mech. Anal., 52 (1973), pp. 244–257.

[13] H. Fan, *On a model of the dynamics of liquid/vapor phase transitions*, SIAM J. Appl. Math., 60 (2000), pp. 1270–1301.

[14] J. Glimm, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math., 18 (1965), pp. 697–715.

[15] L. Gosse, *Existence of L^∞ entropy solutions for a reacting Euler system*, Port. Math. (N.S.), 58 (2001), pp. 473–484.
[16] T.-P. Liu, *Initial-boundary value problems for gas dynamics*, Arch. Rational Mech. Anal., 64 (1977), pp. 137–168.
[17] T.-P. Liu, *Solutions in the large for the equations of nonisentropic gas dynamics*, Indiana Univ. Math. J., 26 (1977), pp. 147–177.
[18] Y. Lu, *Hyperbolic Conservation Laws and the Compensated Compactness Method*, Chapman Hall/CRC Monogr. Surv. Pure Appl. Math., 128, Chapman and Hall/CRC, Boca Raton, FL, 2003.
[19] T. Nishida, *Global solution for an initial boundary value problem of a quasilinear hyperbolic system*, Proc. Japan Acad., 44 (1968), pp. 642–646.
[20] T. Nishida and J. A. Smoller, *Solutions in the large for some nonlinear hyperbolic conservation laws*, Comm. Pure Appl. Math., 26 (1973), pp. 183–200.
[21] Y.-J. Peng, *Solutions faibles globales pour l'équation d'Euler d'un fluide compressible avec de grandes données initiales*, Comm. Partial Differential Equations, 17 (1992), pp. 161–187.
[22] Y.-J. Peng, *Solutions faibles globales pour un modèle d'écoulements diphasiques*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 21 (1994), pp. 523–540.
[23] F. Poupaud, M. Rascle, and J.-P. Vila, *Global solutions to the isothermal Euler-Poisson system with arbitrarily large data*, J. Differential Equations, 123 (1995), pp. 93–121.
[24] D. Serre, *Systems of Conservation Laws. Vol.* 1, Cambridge University Press, Cambridge, UK, 1999.
[25] D. Serre, *Systems of Conservation Laws. Vol.* 2, Cambridge University Press, Cambridge, UK, 2000.
[26] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, 2nd ed., Springer-Verlag, New York, 1994.
[27] J. B. Temple, *Solutions in the large for the nonlinear hyperbolic conservation laws of gas dynamics*, J. Differential Equations, 41 (1981), pp. 96–161.
[28] D. H. Wagner, *Equivalence of the Euler and Lagrangian equations of gas dynamics for weak solutions*, J. Differential Equations, 68 (1987), pp. 118–136.

SIAM J. MATH. ANAL.
Vol. 40, No. 1, pp. 167–185

ON THE GLOBAL WELL-POSEDNESS OF THE CRITICAL QUASI-GEOSTROPHIC EQUATION*

HAMMADI ABIDI† AND TAOUFIK HMIDI†

Abstract. We prove the global well-posedness of the critical dissipative quasi-geostrophic equation for large initial data belonging to the critical Besov space $\dot{B}^0_{\infty,1}(\mathbb{R}^2)$.

Key words. quasi-geostrophic equation, global existence, Besov spaces

AMS subject classifications. 76D03, 35B33, 35Q35, 76D05

DOI. 10.1137/070682319

1. Introduction.

In this paper we are concerned with the initial value problem of the 2D (two-dimensional) dissipative quasi-geostrophic equation

$$(\mathrm{QG})_\alpha \begin{cases} \partial_t\theta + v\cdot\nabla\theta + |\mathrm{D}|^{2\alpha}\theta = 0, \\ \theta_{|t=0} = \theta^0, \end{cases}$$

where the scalar function θ represents the potential temperature and the parameter $\alpha \in]0,1]$. The velocity $v = (v^1, v^2)$ is determined by Riesz transforms of θ,

$$v = (-\partial_2|\mathrm{D}|^{-1}\theta, \partial_1|\mathrm{D}|^{-1}\theta) := (-R_2\theta, R_1\theta), \quad |\mathrm{D}| = \sqrt{-\Delta}.$$

In addition to its intrinsic mathematical importance, this equation serves as a 2D model in geophysical fluid dynamics; for more details about the subject, see [6, 19].

This equation has been intensively investigated, and much attention is devoted to the problem of global well-posedness. For the subcritical case ($\alpha > \frac{1}{2}$) the theory seems to be in a satisfactory state. Indeed, global existence and uniqueness for arbitrary initial data are established in various function spaces (see, for example, [8, 20]). However, the critical and supercritical cases, corresponding respectively to $\alpha = \frac{1}{2}$ and $\alpha < \frac{1}{2}$, are harder to deal with. In the supercritical case, we have until now only global results for small initial data; see, for instance, [3, 5, 12, 13, 23, 24]. For the critical case, Constantin, Córdoba, and Wu showed in [7] the global existence in Sobolev space H^1 under a smallness assumption of the L^∞ norm of θ^0. Many other relevant results can be found in [9, 14, 15, 17]. Very recently, Kiselev, Nazarov, and Volberg proved in [16] global well-posedness for arbitrary periodic smooth initial data by using an elegant argument of the modulus of continuity. In [2], Caffarelli and Vasseur established the global regularity of weak solutions associated with L^2 initial data.

The main goal of this work is to establish global well-posedness in the critical case when initial data belong to the homogeneous critical Besov space $\dot{B}^0_{\infty,1}(\mathbb{R}^2)$: we remove the periodic condition, and we weaken the initial regularity. Before giving our main result let us first specify our notion of critical spaces. Let θ be a solution of

*Received by the editors February 9, 2007; accepted for publication (in revised form) December 3, 2007; published electronically April 4, 2008.
http://www.siam.org/journals/sima/40-1/68231.html
†IRMAR, Université de Rennes 1, Campus de Beaulieu, 35042 Rennes cedex, France (hamadi.abidi@univ-rennes1.fr, thmidi@univ-rennes1.fr).

$(\mathrm{QG})_\alpha$ and $\lambda > 0$; then $\theta_\lambda(t,x) = \theta(\lambda t, \lambda x)$ is also a solution. One class of scaling invariant spaces is the homogeneous Besov spaces $(\dot{B}^{2/p}_{p,r})$, with p, $r \in [1,\infty]$.

Our first main result reads as follows.

THEOREM 1.1. *Let $\theta^0 \in \dot{B}^0_{\infty,1}$; then there exists a unique global solution θ to $(\mathrm{QG})_\alpha$ such that*

$$\theta \in \mathcal{C}(\mathbb{R}_+;\, \dot{B}^0_{\infty,1}) \cap L^1_{loc}(\mathbb{R}_+;\, \dot{B}^1_{\infty,1}).$$

The proof relies essentially on two facts: the first is the establishment of local existence, which is the major part of this paper, and the derivation of some smoothing effects of the solution, described in the next theorem. The second is the use of the modulus of continuity, as used in [16]. We mention that the property allowing us to remove the periodicity is the spatial decay of the solution. The key to our local existence result is some new estimates for the following transport-diffusion equation:

$$(\mathrm{TD}) \begin{cases} \partial_t \theta + v \cdot \nabla \theta + |\mathrm{D}|\theta = f, \\ \theta_{|t=0} = \theta^0, \end{cases}$$

where v and f are given and θ is the unknown scalar function. We now state our second main result.

THEOREM 1.2. *Let $s \in \,]-1,1[$, $r, \bar{r} \in [1,+\infty]$ with $r \geq \bar{r}$, $f \in \widetilde{L}^{\bar{r}}_{\mathrm{loc}}(\mathbb{R}_+; \dot{B}^{s+\frac{1}{\bar{r}}-1}_{\infty,1})$, and v be a divergence-free vector field belonging to $L^1_{\mathrm{loc}}(\mathbb{R}_+; \mathrm{Lip}(\mathbb{R}^2))$. We consider a smooth solution θ of the transport-diffusion equation* (TD)*; then there exists a constant C depending only on s such that for every $t \in \mathbb{R}_+$*

$$\|\theta\|_{\widetilde{L}^r_t \dot{B}^{s+\frac{1}{r}}_{\infty,1}} \leq C e^{C\int_0^t \|\nabla v(\tau)\|_{L^\infty} d\tau} \big(\|\theta^0\|_{\dot{B}^s_{\infty,1}} + \|f\|_{\widetilde{L}^{\bar{r}}_t \dot{B}^{s+\frac{1}{\bar{r}}-1}_{\infty,1}} \big).$$

Additionally, if $v = \nabla^\perp |\mathrm{D}|^{-1}\theta$, then we have for all $s \geq 1$

$$\|\theta\|_{\widetilde{L}^r_t \dot{B}^{s+\frac{1}{r}}_{\infty,1}} \leq C e^{C\int_0^t (\|\nabla\theta(\tau)\|_{L^\infty} + \|\nabla v(\tau)\|_{L^\infty}) d\tau} \big(\|\theta^0\|_{\dot{B}^s_{\infty,1}} + \|f\|_{\widetilde{L}^{\bar{r}}_t \dot{B}^{s+\frac{1}{\bar{r}}-1}_{\infty,1}} \big).$$

We use for the proof a new approach based on Lagrangian coordinates combined with paradifferential calculus and a new commutator estimate. This idea has been recently used by the second author in [11, 12].

The rest of the paper is structured as follows. In section 2 we review some basic results of Littlewood–Paley theory, and we give some useful lemmas. Section 3 deals with a new commutator estimate which is needed for the proof of Theorem 1.2, done in section 4. Theorem 1.1 is proved in section 5.

2. Preliminaries. Throughout the paper, C stands for a constant which may be different in each occurrence. We shall sometimes use the notation $A \lesssim B$ instead of $A \leq CB$, and $A \approx B$ means that $A \lesssim B$ and $B \lesssim A$. We denote by $\mathcal{F}f$ the Fourier transform of f and by $[\eta]$ the whole part of η.

One starts with recalling a traditional result that will be frequently used.

LEMMA 2.1 (Bernstein). *Let $f \in L^p(\mathbb{R}^2)$, with $1 \leq p \leq \infty$ and $0 < r < R$. Then there exists a constant $C > 0$ such that $\forall\, k \in \mathbb{N}$ and $\lambda > 0$ we have*

$$\mathrm{supp}\, \mathcal{F}f \subset B(0, \lambda r) \Longrightarrow \sup_{|\beta|=k} \|\partial^\beta f\|_{L^q} \leq C^k \lambda^{k+2(\frac{1}{p}-\frac{1}{q})} \|f\|_{L^p},$$

$$\mathrm{supp}\, \mathcal{F}f \subset \mathcal{C}(0, \lambda r, \lambda R) \Longrightarrow C^{-k}\lambda^k \|f\|_{L^p} \leq \sup_{|\beta|=k} \|\partial^\beta f\|_{L^p} \leq C^k \lambda^k \|f\|_{L^p}.$$

These estimates hold true if we replace the derivation ∂^β by $|\mathrm{D}|^{|\beta|}$.

To define Besov spaces we need to recall the homogeneous Littlewood–Paley decomposition based on a dyadic unity partition. Let φ be a smooth function supported in the ring $\mathcal{C} := \{\xi \in \mathbb{R}^2, \frac{3}{4} \leq |\xi| \leq \frac{8}{3}\}$ and such that

$$\sum_{q\in\mathbb{Z}} \varphi(2^{-q}\xi) = 1 \quad \text{for} \quad \xi \neq 0.$$

Now, for $u \in \mathcal{S}'$ we set

$$\forall q \in \mathbb{Z}, \quad \Delta_q u = \varphi(2^{-q}\mathrm{D})u \qquad \text{and} \qquad S_q u = \sum_{j\leq q-1} \Delta_j u.$$

We have the formal decomposition

$$u = \sum_{q\in\mathbb{Z}} \Delta_q\, u \quad \forall\, u \in \mathcal{S}'(\mathbb{R}^2)/\mathcal{P}[\mathbb{R}^2],$$

where $\mathcal{P}[\mathbb{R}^2]$ is the set of polynomials (see [18]). Moreover, the Littlewood–Paley decomposition satisfies the property of almost orthogonality:

$$\Delta_k\Delta_q u \equiv 0 \quad \text{if} \quad |k-q| \geq 2, \quad \text{and} \quad \Delta_k(S_{q-1}u\Delta_q u) \equiv 0 \quad \text{if} \quad |k-q| \geq 5. \tag{1}$$

We recall now the definition of Besov spaces. Letting $(p,m) \in [1,+\infty]^2$, $s \in \mathbb{R}$, and $u \in \mathcal{S}'$, we set

$$\|u\|_{\dot{B}^s_{p,m}} := \Big(2^{qs}\|\Delta_q u\|_{L^p}\Big)_{\ell^m}, \quad \dot{\mathcal{B}}^s_{p,m} := \Big\{u \in \mathcal{S} \;\Big|\; \|u\|_{\dot{B}^s_{p,m}} < \infty\Big\}.$$

- For $s < \frac{2}{p}$ (or $s \leq \frac{2}{p}$ if $m = 1$), we then define $\dot{B}^s_{p,m}$ as the completion of $\dot{\mathcal{B}}^s_{p,m}$ for $\|\cdot\|_{\dot{B}^s_{p,m}}$.
- If $k \in \mathbb{N}$ and $\frac{2}{p} + k - 1 \leq s < \frac{2}{p} + k$ (or $s = \frac{2}{p} + k$ if $m = 1$), then $\dot{B}^s_{p,m}$ is defined as the subset of distributions $u \in \mathcal{S}'$ such that $\partial^\beta u \in \dot{B}^{s-k}_{p,m}$ whenever $|\beta| = k$.

Another characterization of the homogeneous Besov spaces that will be needed later is the following; see, for instance, [21]. For $s \in]0,1[$, $p,m \in [1,\infty]$

$$\left(\int_{\mathbb{R}^2} \frac{\|u(\cdot - x) - u(\cdot)\|^m_{L^p}}{|x|^{sm}} \frac{dx}{|x|^2}\right)^{\frac{1}{m}} \approx \|u\|_{\dot{B}^s_{p,m}}, \tag{2}$$

with the usual modification if $m = \infty$.

In our next study we require two kinds of coupled space-time Besov spaces. The first one is defined in the following manner: for $T > 0$ and $m \geq 1$, we denote by $L^r_T\dot{B}^s_{p,m}$ the set of all tempered distributions u satisfying

$$\|u\|_{L^r_T\dot{B}^s_{p,m}} := \Big\|\Big(2^{qs}\|\Delta_q u\|_{L^p}\Big)_{\ell^m}\Big\|_{L^r_T} < \infty.$$

The second mixed space is $\widetilde{L}^r_T\dot{B}^s_{p,m}$, which is the set of tempered distributions u satisfying

$$\|u\|_{\widetilde{L}^r_T\dot{B}^s_{p,m}} := \Big(2^{qs}\|\Delta_q u\|_{L^r_T L^p}\Big)_{\ell^m} < \infty.$$

We can define in the same way the spaces $L^r_T B^s_{p,m}$ and $\widetilde{L}^r_T B^s_{p,m}$. The following embeddings are a direct consequence of Minkowski's inequality.

Let $s \in \mathbb{R}$, $r \geq 1$, and $\big(p, m\big) \in [1, \infty]^2$; then we have

$$L^r_T \dot{B}^s_{p,m} \hookrightarrow \widetilde{L}^r_T \dot{B}^s_{p,m} \quad \text{if} \quad m \geq r \quad \text{and} \tag{3}$$

$$\widetilde{L}^r_T \dot{B}^s_{p,m} \hookrightarrow L^r_T \dot{B}^s_{p,m} \quad \text{if} \quad r \geq m.$$

The next lemma will be useful.

PROPOSITION 2.2. *The following results hold true:*

- $$\dot{B}^s_{p,m} \hookrightarrow \dot{B}^{s-2(\frac{1}{p}-\frac{1}{p_1})}_{p_1,m_1} \quad \textit{for} \quad p \leq p_1 \quad \textit{and} \quad m \leq m_1.$$
- *Let* $|\mathrm{D}| := \sqrt{-\Delta}$ *and* $\sigma \in \mathbb{R}$*; then the operator* $|\mathrm{D}|^\sigma$ *is an isomorphism from* $\dot{B}^s_{p,m}$ *to* $\dot{B}^{s-\sigma}_{p,m}$.
- *Let* $\gamma \in]0,1[$, s_1, $s_2 \in \mathbb{R}$ *such that* $s_1 < s_2$ *and* $u \in \dot{B}^{s_1}_{p,\infty} \cap \dot{B}^{s_2}_{p,\infty}$*; then*

$$\|u\|_{\dot{B}^{\gamma s_1 + (1-\gamma)s_2}_{p,1}} \lesssim \|u\|^{\gamma}_{\dot{B}^{s_1}_{p,\infty}} \|u\|^{1-\gamma}_{\dot{B}^{s_2}_{p,\infty}}.$$

- *For* $s > 0$, $\dot{B}^s_{p,m} \cap L^\infty$ *is an algebra.*

We now recall some commutator estimates (see [4, 10] and the references therein).

LEMMA 2.3. *Let* $p, r \in [1, \infty]$, $1 = \frac{1}{r} + \frac{1}{r'}$, $\rho_1 < 1$, $\rho_2 < 1$*, and* v *be a divergence-free vector field of* $\mathbb{R}^2$*. Assume in addition that*

$$\rho_1 + \rho_2 + 2\min\{1, 2/p\} > 0 \quad \textit{and} \quad \rho_1 + 2/p > 0.$$

Then we have

$$\sum_{q \in \mathbb{Z}} 2^{q(\frac{2}{p}+\rho_1+\rho_2-1)} \big\| [\Delta_q, v \cdot \nabla] u \big\|_{L^1_t L^p} \lesssim \|v\|_{\widetilde{L}^r_t \dot{B}^{\frac{2}{p}+\rho_1}_{p,1}} \|u\|_{\widetilde{L}^{r'}_t \dot{B}^{\frac{2}{p}+\rho_2}_{p,1}}.$$

Moreover, we have for $s \in]-1, 1[$

$$\sum_{q \in \mathbb{Z}} 2^{qs} \big\| [\Delta_q, v \cdot \nabla] u \big\|_{L^p} \lesssim \|\nabla v\|_{L^\infty} \|u\|_{\dot{B}^s_{p,1}}.$$

If $v = \nabla^\perp |\mathrm{D}|^{-1} \theta$*, then the above estimate holds true for* $s \geq 1$ *if we replace* $\|\nabla v\|_{L^\infty}$ *by* $\|\nabla v\|_{L^\infty} + \|\nabla u\|_{L^\infty}$.

The following result is due to Vishik [22].

LEMMA 2.4. *Let* f *be a function in the Schwartz class and* ψ *a diffeomorphism preserving Lebesgue measure; then* $\forall\ p \in [1, +\infty]$ *and* $\forall\ j, q \in \mathbb{Z}$,

$$\|\Delta_j (\Delta_q f \circ \psi)\|_{L^p} \leq C 2^{-|j-q|} \|\nabla \psi^{\epsilon(j,q)}\|_{L^\infty} \|\Delta_q f\|_{L^p},$$

with

$$\epsilon(j, q) = \mathit{sign}(j - q).$$

The following result is proved in [9].

PROPOSITION 2.5. *Let v be a smooth divergence-free vector field and f be a smooth function. We assume that θ is a smooth solution of the equation*

$$\partial_t\theta + v\cdot\nabla\theta + \kappa|\mathrm{D}|^{2\alpha}\theta = f, \quad \text{with} \quad \kappa \geq 0 \quad \text{and} \quad \alpha \in [0,1].$$

Then for $p \in [1,+\infty]$ we have

$$\|\theta(t)\|_{L^p} \leq \|\theta(0)\|_{L^p} + \int_0^t \|f(\tau)\|_{L^p} d\tau.$$

We can find a proof of the next proposition in [12].

PROPOSITION 2.6. *Let $\mathcal{C}$ be a ring and $\alpha \in \mathbb{R}_+$. There exists a positive constant C such that, for any $p \in [1;+\infty]$, for any pair (t,λ) of positive real numbers, we have*

$$\operatorname{supp}\mathcal{F}u \subset \lambda\mathcal{C} \Rightarrow \|e^{-t|\mathrm{D}|^\alpha}u\|_{L^p} \leq Ce^{-C^{-1}t\lambda^\alpha}\|u\|_{L^p}.$$

3. Commutator estimate. The main result of this section is the following estimate that will play a crucial role for the proof of Theorem 1.2.

PROPOSITION 3.1. *Let v be a divergence-free vector field belonging to $L^1_{\mathrm{loc}}(\mathbb{R}_+;\mathrm{Lip}(\mathbb{R}^2))$. For $q \in \mathbb{Z}$ we denote by ψ_q the flow of the regularized vector field $S_{q-1}v$. Then for $f \in \dot{B}^1_{\infty,\infty}$ and for $q \in \mathbb{Z}$ we have*

$$\big\||\mathrm{D}|(\Delta_q f\circ\psi_q) - (|\mathrm{D}|\Delta_q f)\circ\psi_q\big\|_{L^\infty} \leq Ce^{CV(t)}V^{\frac12}(t)2^q\|\Delta_q f\|_{L^\infty},$$

where $V(t) = \|\nabla v\|_{L^1_t L^\infty(\mathbb{R}^2)}$ and C is an absolute constant.

Proof. We set $f_q := \Delta_q f$, and then it is obvious that

$$\begin{aligned}
|\mathrm{D}|(f_q\circ\psi_q) - (|\mathrm{D}|f_q)\circ\psi_q &= |\mathrm{D}|^{\frac12}\{(|\mathrm{D}|^{\frac12}f_q)\circ\psi_q\} - \{|\mathrm{D}|^{\frac12}(|\mathrm{D}|^{\frac12}f_q)\}\circ\psi_q \\
&\quad + |\mathrm{D}|^{\frac12}\big\{|\mathrm{D}|^{\frac12}(f_q\circ\psi_q) - (|\mathrm{D}|^{\frac12}f_q)\circ\psi_q\big\} \\
&:= \mathrm{I} + \mathrm{II}.
\end{aligned}$$

For the first term we apply Proposition 3.1 from [12], with $\alpha = \frac12$ and $F_q = |\mathrm{D}|^{\frac12}f_q$, yielding

$$\begin{aligned}
\|\mathrm{I}\|_{L^\infty} &\leq Ce^{CV(t)}(e^{CV(t)}-1)\|F_q\|_{\dot{B}^{\frac12}_{\infty,1}} \\
&\leq Ce^{CV(t)}(e^{CV(t)}-1)2^q\|f_q\|_{L^\infty} \\
&\lesssim e^{CV(t)}V^{\frac12}(t)2^q\|f_q\|_{L^\infty}.
\end{aligned}$$

For the second term we use the following formula for the fractional Laplacian:

$$|\mathrm{D}|^{\frac12}f(x) = C\int_{\mathbb{R}^2}\frac{f(x)-f(y)}{|x-y|^{\frac52}}dy.$$

Since the flow ψ_q preserves Lebesgue measure, we easily get

$$\begin{aligned}
|\mathrm{D}|^{\frac12}(f_q\circ\psi_q)(x) - (|\mathrm{D}|^{\frac12}f_q)\circ\psi_q(x) &= C\int_{\mathbb{R}^2}\frac{f_q(\psi_q(x)) - f_q(\psi_q(y))}{|x-y|^{\frac52}} \\
&\quad\times\left(1 - \frac{|x-y|^{\frac52}}{|\psi_q(x)-\psi_q(y)|^{\frac52}}\right)dy.
\end{aligned}$$

We denote $g_q(x)=f_q(\psi_q(x))$ and we set $h=x-y$:

$$|\mathrm{D}|^{\frac12}(f_q\circ\psi_q)(x)-(|\mathrm{D}|^{\frac12}f_q)\circ\psi_q(x)=C\int_{\mathbb{R}^2}\frac{g_q(x)-g_q(x-h)}{|h|^{\frac52}}\bar{\psi}_q(x,h)dh,$$

with

$$\bar{\psi}_q(x,h)=1-\frac{|h|^{\frac52}}{|\psi_q(x)-\psi_q(x-h)|^{\frac52}}.$$

It follows from law products and the embedding $\dot{B}^0_{\infty,1}\hookrightarrow L^\infty$ that

$$\begin{aligned}\|\mathrm{II}\|_{L^\infty}&\leq\big\||\mathrm{D}|^{\frac12}(f_q\circ\psi_q)-(|\mathrm{D}|^{\frac12}f_q)\circ\psi_q\big\|_{\dot{B}^{\frac12}_{\infty,1}}\\&\leq C\|\bar{\psi}_q\|_{L^\infty(\mathbb{R}^4)}\int_{\mathbb{R}^2}|h|^{-\frac52}\|g_q(\cdot)-g_q(\cdot-h)\|_{\dot{B}^{\frac12}_{\infty,1}}dh\\&+C\sup_{h\in\mathbb{R}^2}\|\bar{\psi}_q(\cdot,h)\|_{\dot{B}^{\frac12}_{\infty,1}}\int_{\mathbb{R}^2}|h|^{-\frac52}\|g_q(\cdot)-g_q(\cdot-h)\|_{L^\infty}dh\\&=J^1_q+J^2_q.\end{aligned}$$

We intend to estimate J^1_q. It is plain from the mean value theorem that

$$\frac{1}{\|\nabla\psi\|^{\frac52}_{L^\infty}}\leq\frac{|h|^{\frac52}}{|\psi(x)-\psi(x-h)|^{\frac52}}\leq\|\nabla\psi^{-1}\|^{\frac52}_{L^\infty},$$

which easily gives the inequality

$$\|\bar{\psi}_q\|_{L^\infty(\mathbb{R}^4)}\leq\max\Big(|1-\|\nabla\psi_q^{-1}\|^{\frac52}_{L^\infty}|;|1-\|\nabla\psi_q\|^{-\frac52}_{L^\infty}|\Big).$$

On the other hand, we have the classical estimates

$$\begin{aligned}e^{-C\|S_{q-1}\nabla v\|_{L^1_tL^\infty}}\leq\|\nabla\psi_q^{\mp1}\|_{L^\infty}\leq e^{C\|S_{q-1}\nabla v\|_{L^1_tL^\infty}}\\\text{and}\quad\|S_{q-1}\nabla v\|_{L^1_tL^\infty}\leq CV(t).\end{aligned}\tag{4}$$

We thus get

$$\|\bar{\psi}_q\|_{L^\infty(\mathbb{R}^4)}\leq Ce^{CV(t)}(e^{CV(t)}-1).\tag{5}$$

Using the definition of Besov spaces and the commutation of Δ_j with translation operators, one finds

$$\begin{aligned}&\int_{\mathbb{R}^2}|h|^{-\frac52}\|g_q(\cdot)-g_q(\cdot-h)\|_{\dot{B}^{\frac12}_{\infty,1}}dh\\&\qquad\leq\sum_j2^{\frac12j}\int_{\mathbb{R}^2}|h|^{-\frac12}\|\Delta_jg_q(\cdot)-(\Delta_jg_q)(\cdot-h)\|_{L^\infty}\frac{dh}{|h|^2}.\end{aligned}$$

Applying the characterization of Besov spaces (2) yields

$$\begin{aligned}\int_{\mathbb{R}^2} |h|^{-\frac{5}{2}} \|g_q(\cdot) - g_q(\cdot - h)\|_{\dot{B}^{\frac{1}{2}}_{\infty,1}} dh &\leq C \sum_j 2^{\frac{1}{2}j} \|\Delta_j g_q\|_{\dot{B}^{\frac{1}{2}}_{\infty,1}} \\ &\leq C \sum_{|j-k|\leq 1} 2^{j\frac{1}{2}} 2^{\frac{1}{2}k} \|\Delta_j \Delta_k g_q\|_{L^\infty} \\ &\leq C \|g_q\|_{\dot{B}^1_{\infty,1}}.\end{aligned}$$

Now we use the following interpolation estimate:

$$\begin{aligned}\|g_q\|_{\dot{B}^1_{\infty,1}} &\lesssim \|g_q\|^{\frac{1}{2}}_{L^\infty} \|\Delta g_q\|^{\frac{1}{2}}_{L^\infty} \\ &\lesssim \|f_q\|^{\frac{1}{2}}_{L^\infty} \|\Delta g_q\|^{\frac{1}{2}}_{L^\infty}.\end{aligned}$$

It is easy to check from Leibniz rule that

$$\Delta g_q = \Delta(f_q \circ \psi_q) = \sum_{i=1}^{d} \langle (\nabla^2 f_q) \circ \psi_q \cdot \partial_i \psi_q, \partial_i \psi_q \rangle + (\nabla f_q) \circ \psi_q \cdot \Delta \psi_q.$$

Applying the Bernstein inequality, we get

$$\|\Delta g_q\|_{L^\infty} \lesssim e^{CV(t)} 2^{2q} \|f_q\|_{L^\infty} + 2^q \|f_q\|_{L^\infty} \|\Delta \psi_q\|_{L^\infty}.$$

The derivative of the flow equation with respect to x and the use of the Gronwall and Bernstein inequalities gives

$$\begin{aligned}\|\nabla^2 \psi_q(t)\|_{L^\infty} &\lesssim e^{CV(t)} \int_0^t \|\nabla^2 S_{q-1} v(\tau)\|_{L^\infty} d\tau \\ &\lesssim e^{CV(t)} 2^q. \qquad (6)\end{aligned}$$

Combining both last estimates, we obtain

$$\|\Delta g_q\|_{L^\infty} \lesssim e^{CV(t)} 2^{2q} \|f_q\|_{L^\infty}. \qquad (7)$$

Putting together (5) and (7), we conclude that

$$\|J^1_q(t)\|_{L^\infty} \lesssim e^{CV(t)} (e^{CV(t)} - 1) 2^q \|f_q\|_{L^\infty}.$$

Let us now turn to the second term J^2_q. The integral term can be estimated from (2) as follows:

$$\int_{\mathbb{R}^2} |h|^{-\frac{5}{2}} \|g_q(\cdot) - g_q(\cdot - h)\|_{L^\infty} dh \lesssim \|g_q\|_{\dot{B}^{\frac{1}{2}}_{\infty,1}}.$$

According to the classical composition result we write

$$\begin{aligned}\|g_q(t)\|_{\dot{B}^{\frac{1}{2}}_{\infty,1}} &\lesssim \|\nabla \psi_q\|^{\frac{1}{2}}_{L^\infty} \|f_q\|_{\dot{B}^{\frac{1}{2}}_{\infty,1}} \\ &\lesssim e^{CV(t)} 2^{q\frac{1}{2}} \|f_q\|_{L^\infty}. \qquad (8)\end{aligned}$$

In order to estimate $\bar{\psi}_q$ we use the interpolation inequality

$$\|\bar{\psi}_q(\cdot,h)\|_{\dot{B}^{\frac{1}{2}}_{\infty,1}} \lesssim \|\bar{\psi}_q(\cdot,h)\|^{\frac{1}{2}}_{L^\infty}\|\nabla_x\bar{\psi}_q(\cdot,h)\|^{\frac{1}{2}}_{L^\infty}.$$

This leads in view of (5) to

$$\|\bar{\psi}_q(\cdot,h)\|_{\dot{B}^{\frac{1}{2}}_{\infty,1}} \leq Ce^{CV(t)}(e^{CV(t)}-1)^{\frac{1}{2}}\|\nabla_x\bar{\psi}_q(\cdot,h)\|^{\frac{1}{2}}_{L^\infty}. \tag{9}$$

The derivative of $\bar{\psi}_q$ with respect to x yields

$$\begin{aligned}|\nabla_x\bar{\psi}_q(x,h)| &\lesssim \frac{|h|^{\frac{7}{2}}}{|\psi_q(x)-\psi_q(x-h)|^{\frac{7}{2}}}\frac{|\nabla_x\psi_q(x)-\nabla_x\psi_q(x-h)|}{|h|}\\ &\lesssim \|\nabla\psi_q^{-1}\|^{\frac{7}{2}}_{L^\infty}\|\nabla^2\psi_q\|_{L^\infty}.\end{aligned}$$

Combining (4) and (6), we obtain

$$\|\nabla_x\bar{\psi}_q(t)\|_{L^\infty(\mathbb{R}^4)} \lesssim e^{CV(t)}2^q. \tag{10}$$

Plugging (10) into (9), we find

$$\|\bar{\psi}_q(\cdot,h)\|_{\dot{B}^{\frac{1}{2}}_{\infty,1}} \lesssim e^{CV(t)}V^{\frac{1}{2}}(t)2^{\frac{q}{2}}. \tag{11}$$

Thus we deduce from (11) and (8) that

$$\|J^2_q(t)\|_{L^\infty} \leq Ce^{CV(t)}V^{\frac{1}{2}}(t)2^q\|f_q(t)\|_{L^\infty}.$$

This achieves the proof. □

4. Proof of Theorem 1.2. The Fourier localized function $\theta_q := \Delta_q\theta$ satisfies

$$\partial_t\theta_q + S_{q-1}v\cdot\nabla\theta_q + |\mathrm{D}|\theta_q = -[\Delta_q, v\cdot\nabla]\theta + (S_{q-1}v-v)\cdot\nabla\theta_q + f_q := \mathcal{R}_q. \tag{12}$$

Let ψ_q denote the flow of the velocity $S_{q-1}v$, and set

$$\bar{\theta}_q(t,x) = \theta_q(t,\psi_q(t,x)) \quad \text{and} \quad \bar{\mathcal{R}}_q(t,x) = \mathcal{R}_q(t,\psi_q(t,x)).$$

Since ψ_q is an homeomorphism, then

$$\|\bar{\mathcal{R}}_q\|_{L^\infty} \leq \|[\Delta_q, v\cdot\nabla]\theta\|_{L^\infty} + \|(S_{q-1}v-v)\cdot\nabla\theta_q\|_{L^\infty} + \|f_q\|_{L^\infty}. \tag{13}$$

It is not hard to check that the function $\bar{\theta}_q$ satisfies

$$\partial_t\bar{\theta}_q + |\mathrm{D}|\bar{\theta}_q = |\mathrm{D}|(\theta_q\circ\psi_q) - (|\mathrm{D}|\theta_q)\circ\psi_q + \bar{\mathcal{R}}_q := \bar{\mathcal{R}}^1_q. \tag{14}$$

From Proposition 3.1 we find that for $q\in\mathbb{Z}$

$$\||\mathrm{D}|(\theta_q\circ\psi_q) - (|\mathrm{D}|\theta_q)\circ\psi_q\|_{L^\infty} \lesssim e^{CV(t)}V^{\frac{1}{2}}(t)2^q\|\theta_q(t)\|_{L^\infty}, \tag{15}$$

where $V(t) := \|\nabla v\|_{L^1_tL^\infty}$. Putting together (13) and (15) yields

$$\begin{aligned}\|\bar{\mathcal{R}}^1_q(t)\|_{L^\infty} \lesssim \|f_q(t)\|_{L^\infty} &+ \|(S_{q-1}v-v)\cdot\nabla\theta_q\|_{L^\infty} + \|[\Delta_q, v\cdot\nabla]\theta\|_{L^\infty}\\ &+ e^{CV(t)}V^{\frac{1}{2}}(t)2^q\|\theta_q(t)\|_{L^\infty}.\end{aligned}$$

Applying the operator Δ_j to (14) and using Proposition 2.6, we obtain

$$\|\Delta_j\bar{\theta}_q(t)\|_{L^\infty} \lesssim e^{-ct2^j}\|\Delta_j\theta_q^0\|_{L^\infty} + \int_0^t e^{-c(t-\tau)2^j}\|f_q(\tau)\|_{L^\infty}d\tau \tag{16}$$

$$+ e^{CV(t)}V^{\frac{1}{2}}(t)2^q\int_0^t e^{-c(t-\tau)2^j}\|\theta_q(\tau)\|_{L^\infty}d\tau$$

$$+ \int_0^t e^{-c(t-\tau)2^j}\|[\Delta_q, v\cdot\nabla]\theta(\tau)\|_{L^\infty}d\tau$$

$$+ \int_0^t e^{-c(t-\tau)2^j}\|(S_{q-1}v - v)\cdot\nabla\theta_q(\tau)\|_{L^\infty}d\tau. \tag{17}$$

Integrating this estimate with respect to the time and using the Young inequality, we get

$$\|\Delta_j\bar{\theta}_q\|_{L^r_tL^\infty} \lesssim 2^{-\frac{j}{r}}(1-e^{-crt2^j})^{\frac{1}{r}}\|\Delta_j\theta_q^0\|_{L^\infty} + 2^{-j(1+\frac{1}{r}-\frac{1}{\overline{r}})}\|f_q\|_{L^{\overline{r}}_tL^\infty}$$

$$+ e^{CV(t)}V^{\frac{1}{2}}(t)2^{(q-j)}\|\theta_q\|_{L^r_tL^\infty}$$

$$+ 2^{-\frac{j}{r}}\int_0^t \|[\Delta_q, v\cdot\nabla]\theta(\tau)\|_{L^\infty}d\tau$$

$$+ 2^{-\frac{j}{r}}\int_0^t \|(S_{q-1}v - v)\cdot\nabla\theta_q(\tau)\|_{L^\infty}d\tau. \tag{18}$$

Since the flow ψ is a homeomorphism, one writes

$$\begin{aligned}2^{q(s+\frac{1}{r})}\|\theta_q\|_{L^r_tL^\infty} &= 2^{q(s+\frac{1}{r})}\|\bar{\theta}_q\|_{L^r_tL^\infty}\\ &\le 2^{q(s+\frac{1}{r})}\left(\sum_{|j-q|>N}\|\Delta_j\bar{\theta}_q\|_{L^r_tL^\infty} + \sum_{|j-q|\le N}\|\Delta_j\bar{\theta}_q\|_{L^r_tL^\infty}\right)\\ &:= \mathrm{I} + \mathrm{II}.\end{aligned}$$

To estimate the term I we appeal to Lemma 2.4,

$$\begin{aligned}\|\Delta_j\bar{\theta}_q\|_{L^r_tL^\infty} &\lesssim 2^{-|q-j|}e^{C\int_0^t\|\nabla v(\tau)\|_{L^\infty}d\tau}\|\theta_q\|_{L^r_tL^\infty}\\ &\le C2^{-|q-j|}e^{CV(t)}\|\theta_q\|_{L^r_tL^\infty}.\end{aligned}$$

Therefore we get

$$\mathrm{I} \le C2^{-N}e^{CV(t)}2^{q(s+\frac{1}{r})}\|\theta_q\|_{L^r_tL^\infty}. \tag{19}$$

In order to bound the second term II we use (18):

$$\mathrm{II} \lesssim (1-e^{-crt2^q})^{\frac{1}{r}}2^{qs}\|\theta_q^0\|_{L^\infty} + 2^{N(\frac{1}{r}+1-\frac{1}{\overline{r}})}2^{q(s+\frac{1}{\overline{r}}-1)}\|f_q\|_{L^{\overline{r}}_tL^\infty}$$

$$+ 2^Ne^{CV(t)}V^{\frac{1}{2}}(t)2^{q(s+\frac{1}{r})}\|\theta_q\|_{L^r_tL^\infty}$$

$$+ 2^{\frac{N}{r}}2^{qs}\int_0^t\|[\Delta_q, v\cdot\nabla]\theta(\tau)\|_{L^\infty}d\tau$$

$$+ 2^{\frac{N}{r}}2^{qs}\int_0^t\|(S_{q-1}v - v)\cdot\nabla\theta_q(\tau)\|_{L^\infty}d\tau. \tag{20}$$

Denoting $Z_q^r(t) := 2^{q(s+\frac{1}{r})}\|\theta_q\|_{L^r_t L^\infty}$, we then obtain, in view of (19) and (20),

$$\begin{aligned}
Z_q^r(t) &\leq C(1-e^{-crt2^q})^{\frac{1}{r}}2^{qs}\|\theta_q^0\|_{L^\infty} + C2^{N(\frac{1}{r}+1-\frac{1}{\overline{r}})}2^{q(s+\frac{1}{\overline{r}}-1)}\|f_q\|_{L^{\overline{r}}_t L^\infty}\\
&\quad + C\big\{2^N e^{CV(t)}V^{\frac{1}{2}}(t) + 2^{-N}e^{CV(t)}\big\}Z_q^r(t)\\
&\quad + C2^{\frac{N}{r}}2^{qs}\int_0^t \|[\Delta_q, v\cdot\nabla]\theta(\tau)\|_{L^\infty}d\tau\\
&\quad + C2^{\frac{N}{r}}2^{qs}\int_0^t \|(S_{q-1}v - v)\cdot\nabla\theta_q(\tau)\|_{L^\infty}d\tau.
\end{aligned}$$

It is easy to check the existence of two absolute constants N and C_0 such that

$$V(t)\leq C_0 \Rightarrow C2^{-N}e^{CV(t)} + C2^N e^{CV(t)}V^{\frac{1}{2}}(t)\leq \frac{1}{2}\cdot$$

Thus we obtain under this condition

$$\begin{aligned}
Z_q^r(t) &\lesssim (1-e^{-crt2^q})^{\frac{1}{r}}2^{qs}\|\theta_q^0\|_{L^\infty} + 2^{q(s+\frac{1}{\overline{r}}-1)}\|f_q\|_{L^{\overline{r}}_t L^\infty}\\
&\quad + 2^{qs}\int_0^t \Big(\|[\Delta_q, v\cdot\nabla]\theta(\tau)\|_{L^\infty} + \|(S_{q-1}v - v)\cdot\nabla\theta_q\|_{L^\infty}\Big)d\tau. \qquad (21)
\end{aligned}$$

Summing over q and using Lemma 2.3 leads, for $V(t)\leq C_0$, to

$$\begin{aligned}
\|\theta\|_{\widetilde{L}^r_t \dot{B}^{s+\frac{1}{r}}_{\infty,1}} &\lesssim \|\theta^0\|_{\dot{B}^s_{\infty,1}} + \|f\|_{\widetilde{L}^{\overline{r}}_t \dot{B}^{s+\frac{1}{\overline{r}}-1}_{\infty,1}} + \int_0^t \|\nabla v(\tau)\|_{L^\infty}\|\theta(\tau)\|_{\dot{B}^s_{\infty,1}}d\tau\\
&\lesssim \|\theta^0\|_{\dot{B}^s_{\infty,1}} + \|f\|_{\widetilde{L}^{\overline{r}}_t \dot{B}^{s+\frac{1}{\overline{r}}-1}_{\infty,1}} + C_0\|\theta\|_{L^\infty_t \dot{B}^s_{\infty,1}}. \qquad (22)
\end{aligned}$$

Let us show how to conclude the proof in the case of $r=\infty$. If C_0 is sufficiently small, then we obtain from (22) the desired estimate:

$$\|\theta\|_{\widetilde{L}^\infty_t \dot{B}^s_{\infty,1}} \lesssim \|\theta^0\|_{\dot{B}^s_{\infty,1}} + \|f\|_{\widetilde{L}^{\overline{r}}_t \dot{B}^{s+\frac{1}{\overline{r}}-1}_{\infty,1}}. \qquad (23)$$

Now for an arbitrary positive time T we take a partition $(T_i)_{i=0}^M$ of $[0,T]$ such that $\int_{T_i}^{T_{i+1}}\|\nabla v(\tau)\|_{L^\infty}d\tau \approx C_0$. We can proceed analogously to the above calculus and obtain

$$\|\theta\|_{\widetilde{L}^\infty_{[T_i,T_{i+1}]}\dot{B}^s_{\infty,1}} \lesssim \|\theta(T_i)\|_{\dot{B}^s_{\infty,1}} + \|f\|_{\widetilde{L}^{\overline{r}}_{[T_i,T_{i+1}]}\dot{B}^{s+\frac{1}{\overline{r}}-1}_{\infty,1}}.$$

An iteration argument leads to

$$\|\theta\|_{\widetilde{L}^\infty_{[T_i,T_{i+1}]}\dot{B}^s_{\infty,1}} \leq C^{i+1}\Big(\|\theta^0\|_{\dot{B}^s_{\infty,1}} + \|f\|_{\widetilde{L}^{\overline{r}}_{[0,T_{i+1}]}\dot{B}^{s+\frac{1}{\overline{r}}-1}_{\infty,1}}\Big).$$

The triangle inequality and the fact that $C_0 M \simeq 1 + V(t)$ give

$$\|\theta\|_{\widetilde{L}^\infty_T \dot{B}^s_{\infty,1}} \leq Ce^{C\int_0^T\|\nabla v(\tau)\|_{L^\infty}}\Big(\|\theta^0\|_{\dot{B}^s_{\infty,1}} + \|f\|_{\widetilde{L}^{\overline{r}}_T \dot{B}^{s+\frac{1}{\overline{r}}-1}_{\infty,1}}\Big). \qquad (24)$$

Let us now turn to the case of finite r. Combining (22) and (23), we obtain under the assumption $V(t) \leq C_0$

$$\|\theta\|_{\widetilde{L}^r_t \dot{B}^{s+\frac{1}{r}}_{\infty,1}} \lesssim \|\theta^0\|_{\dot{B}^s_{\infty,1}} + \|f\|_{\widetilde{L}^{\overline{r}}_t \dot{B}^{s+\frac{1}{\overline{r}}-1}_{\infty,1}}. \tag{25}$$

This gives the result for a short time, and as for the case $r = \infty$, we obtain the required global estimate.

Concerning the last estimate of Theorem 1.2, we use in the commutator term of (21) the last part of Lemma 2.3. □

5. Proof of Theorem 1.1. The proof is divided into two parts: in the first one we construct a local unique solution and give a criterion of global existence. However, in the second part we discuss the global existence by reproducing an idea of [16].

5.1. Local existence. We aim to prove the following result.

PROPOSITION 5.1. *Given any $\theta^0 \in \dot{B}^0_{\infty,1}$, there is $T > 0$ such that the $(\mathrm{QG})_\alpha$ equation has a unique solution θ with*

$$\theta \in \widetilde{L}^\infty_T \dot{B}^0_{\infty,1} \cap L^1_T \dot{B}^1_{\infty,1}.$$

Moreover, for all $\beta \in \mathbb{R}_+$ we have $t^\beta \theta \in \widetilde{L}^\infty_T \dot{B}^\beta_{\infty,1}$.

Proof. The existence is based on Theorem 1.2 and an iterative method. We denote $\theta_0(t,x) := e^{-t|\mathrm{D}|}\theta^0(x)$, $v_0 := (-R_2\theta_0, R_1\theta_0)$, and θ_{n+1} the solution of the linear system

$$\begin{cases} \partial_t \theta_{n+1} + v_n \cdot \nabla \theta_{n+1} + |\mathrm{D}|\theta_{n+1} = 0, \\ v_n = (-R_2\theta_n, R_1\theta_n), \\ \theta_{n+1|t=0} = \theta^0. \end{cases}$$

Since $\theta_0 \in L^1(\mathbb{R}_+; \dot{B}^1_{\infty,1})$ and from the continuity of Riesz transforms in the homogeneous Besov spaces we find $v_0 \in L^1(\mathbb{R}_+; \dot{B}^1_{\infty,1})$. Thus by iteration and thanks to Theorem 1.2, one deduces that $\forall n \in \mathbb{N}$,

$$\theta_n \in \widetilde{L}^\infty(\mathbb{R}_+; \dot{B}^0_{\infty,1}) \cap L^1(\mathbb{R}_+; \dot{B}^1_{\infty,1}).$$

Step 1: *Uniform bounds.* Now we intend to obtain uniform bounds, with respect to the parameter n, for some $T > 0$ independent of n.

By (21), we have for all $T \geq 0$ such that

$$\int_0^T \|\theta_n(\tau)\|_{\dot{B}^1_{\infty,1}} d\tau \leq C_1 (:= CC_0) \tag{26}$$

the following estimate:

$$\begin{aligned} \|\theta_{n+1}\|_{\widetilde{L}^2_T \dot{B}^{\frac{1}{2}}_{\infty,1}} + \|\theta_{n+1}\|_{L^1_T \dot{B}^1_{\infty,1}} &\lesssim \sum_{q\in\mathbb{Z}} (1 - e^{-cT2^q})^{\frac{1}{2}} \|\Delta_q \theta^0\|_{L^\infty} \\ &+ \sum_{q\in\mathbb{Z}} \int_0^T \|[\Delta_q, v_n \cdot \nabla]\theta_{n+1}(\tau)\|_{L^\infty} d\tau \\ &+ \sum_{q\in\mathbb{Z}} \int_0^T \|(S_{q-1}v_n - v_n) \cdot \nabla \Delta_q \theta_{n+1}(\tau)\|_{L^\infty} d\tau. \end{aligned}$$

Since $\operatorname{div} v_n = 0$, then Lemma 2.3 combined with the continuity of Riesz transforms gives

$$\sum_{q\in\mathbb{Z}}\int_0^T \|[\Delta_q, v_n\cdot\nabla]\theta_{n+1}(\tau)\|_{L^\infty}d\tau \lesssim \|v_n\|_{\widetilde{L}^2_T\dot{B}^{\frac{1}{2}}_{\infty,\infty}}\|\theta_{n+1}\|_{\widetilde{L}^2_T\dot{B}^{\frac{1}{2}}_{\infty,1}}$$
$$\lesssim \|\theta_n\|_{\widetilde{L}^2_T\dot{B}^{\frac{1}{2}}_{\infty,\infty}}\|\theta_{n+1}\|_{\widetilde{L}^2_T\dot{B}^{\frac{1}{2}}_{\infty,1}}.$$

We deduce from the Hölder and Young inequalities that

$$\sum_{q\in\mathbb{Z}}\int_0^t \|(S_qv_n - v_n)\cdot\nabla\Delta_q\theta_{n+1}\|_{L^\infty}d\tau \lesssim \sum_{q\in\mathbb{Z}} 2^q\|\Delta_q\theta_{n+1}\|_{L^2_tL^\infty}\|S_qv_n - v_n\|_{L^2_tL^\infty}$$
$$\lesssim \sum_{q\in\mathbb{Z}} 2^{\frac{1}{2}q}\|\Delta_q\theta_{n+1}\|_{L^2_tL^\infty}\sum_{k\geq q}2^{\frac{1}{2}(q-k)}2^{\frac{1}{2}k}\|\Delta_kv_n\|_{L^2_tL^\infty}$$
$$\lesssim \|\theta_{n+1}\|_{\widetilde{L}^2_t\dot{B}^{\frac{1}{2}}_{\infty,1}}\|\theta_n\|_{\widetilde{L}^2_t\dot{B}^{\frac{1}{2}}_{\infty,\infty}}.$$

Therefore we obtain from the above inequalities

$$\|\theta_{n+1}\|_{\widetilde{L}^2_t\dot{B}^{\frac{1}{2}}_{\infty,1}} + \|\theta_{n+1}\|_{L^1_t\dot{B}^1_{\infty,1}} \lesssim \sum_{q\in\mathbb{Z}}(1-e^{-ct2^q})^{\frac{1}{2}}\|\Delta_q\theta^0\|_{L^\infty}$$
$$+ \|\theta_{n+1}\|_{\widetilde{L}^2_t\dot{B}^{\frac{1}{2}}_{\infty,1}}\|\theta_n\|_{\widetilde{L}^2_t\dot{B}^{\frac{1}{2}}_{\infty,1}}.$$

Thus there exists an absolute constant $\varepsilon_0 > 0$ such that, if

$$\sum_{q\in\mathbb{Z}}(1-e^{-cT2^q})^{\frac{1}{2}}\|\Delta_q\theta^0\|_{L^\infty} \leq \varepsilon_0, \tag{27}$$

then

$$\|\theta_{n+1}\|_{\widetilde{L}^2_T\dot{B}^{\frac{1}{2}}_{\infty,1}} + \|\theta_{n+1}\|_{L^1_T\dot{B}^1_{\infty,1}} \leq 2\varepsilon_0. \tag{28}$$

The existence of $T > 0$ is due to the Lebesgue theorem.

Hence, by using the estimate

$$\int_0^T \|\nabla v(\tau)\|_{L^\infty}d\tau \lesssim \int_0^T \|\theta(\tau)\|_{\dot{B}^1_{\infty,1}}d\tau$$

and Theorem 1.2, we obtain

$$\|\theta_{n+1}\|_{\widetilde{L}^\infty_T\dot{B}^0_{\infty,1}} \lesssim \|\theta^0\|_{\dot{B}^0_{\infty,1}}.$$

Thus we prove that the sequence $(v_n,\theta_n)_{n\in\mathbb{N}}$ is uniformly bounded in the space $\widetilde{L}^\infty_T\dot{B}^0_{\infty,1}\cap L^1_T\dot{B}^1_{\infty,1}$.

Step 2: *Strong convergence.* We will prove that (v_n,θ_n) is a Cauchy sequence in $\widetilde{L}^\infty_T\dot{B}^0_{\infty,1}$. Let $(n,m)\in\mathbb{N}^2$, $\theta_{n,m} =: \theta_{n+1}-\theta_{m+1}$, and $v_{n,m} := v_n - v_m$; then

$$\begin{cases} \partial_t\theta_{n,m} + v_n\cdot\nabla\theta_{n,m} + |\mathrm{D}|\theta_{n,m} = -v_{n,m}\cdot\nabla\theta_{m+1}, \\ \theta_{n,m_{|t=0}} = 0. \end{cases}$$

Applying Theorem 1.2 to this equation gives

$$\|\theta_{n,m}\|_{\widetilde{L}^\infty_t \dot{B}^0_{\infty,1}} \leq Ce^{C\|\theta_n\|_{L^1_t \dot{B}^1_{\infty,1}}} \int_0^t \|v_{n,m}\cdot\nabla\theta_{m+1}(\tau)\|_{\dot{B}^0_{\infty,1}} d\tau. \tag{29}$$

Thanks to Bony's decomposition [1], the embedding $\dot{B}^0_{\infty,1} \hookrightarrow L^\infty$, and the fact that $\operatorname{div} v_{n,m} = 0$,

$$\|v_{n,m}\cdot\nabla\theta_{m+1}\|_{\dot{B}^0_{\infty,1}} \lesssim \|v_{n,m}\|_{\dot{B}^0_{\infty,1}} \|\theta_{m+1}\|_{\dot{B}^1_{\infty,1}}. \tag{30}$$

Since Riesz transforms continuously map $\dot{B}^0_{\infty,1}$ into itself, we get

$$\|v_{n,m}\cdot\nabla\theta_{m+1}\|_{\dot{B}^0_{\infty,1}} \lesssim \|\theta_{n-1,m-1}\|_{\dot{B}^0_{\infty,1}} \|\theta_{m+1}\|_{\dot{B}^1_{\infty,1}}. \tag{31}$$

Thus we infer

$$\|\theta_{n,m}\|_{\widetilde{L}^\infty_t \dot{B}^0_{\infty,1}} \leq C\|\theta_{n-1,m-1}\|_{\widetilde{L}^\infty_t \dot{B}^0_{\infty,1}} e^{C\|\theta_n\|_{L^1_t \dot{B}^1_{\infty,1}}} \int_0^t \|\theta_{m+1}(\tau)\|_{\dot{B}^1_{\infty,1}} d\tau.$$

According to the inequality (28) one can choose ε_0 small such that

$$\|\theta_{n,m}\|_{\widetilde{L}^\infty_T \dot{B}^0_{\infty,1}} \leq \eta\|\theta_{n-1,m-1}\|_{\widetilde{L}^\infty_T \dot{B}^0_{\infty,1}}$$

with $\eta < 1$. Let us suppose that $n \geq m$; then by induction one finds

$$\|\theta_{n,m}\|_{\widetilde{L}^\infty_T \dot{B}^0_{\infty,1}} \lesssim \eta^m \|\theta^0\|_{\dot{B}^0_{\infty,1}}.$$

Thus $(\theta_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $\widetilde{L}^\infty_T \dot{B}^0_{\infty,1}$. Then there exists $\theta \in \widetilde{L}^\infty_t \dot{B}^0_{\infty,1}$ such that θ_n converges strongly to θ in $\widetilde{L}^\infty_t \dot{B}^0_{\infty,1}$. Moreover, the Fatou lemma and inequality (28) imply that $\theta \in L^1_t \dot{B}^1_{\infty,1}$. These pieces of information allow us to pass to the limit into the equation.

Step 3: *Uniqueness.* Let $X_T := L^\infty_T \dot{B}^0_{\infty,1} \cap L^1_T \dot{B}^1_{\infty,1}$, and θ_i, $i = 1, 2$ (v_i the corresponding velocity), be two solutions of the $(\mathrm{QG})_\alpha$ equation with the same initial data and belonging to the space X_T. We set $\theta_{1,2} = \theta_1 - \theta_2$ and $v_{1,2} = v_1 - v_2$; then it is plain that

$$\partial_t\theta_{1,2} + v^1\cdot\nabla\theta_{1,2} + |\mathrm{D}|\theta_{1,2} = -v_{1,2}\cdot\nabla\theta^2, \quad \theta_{1,2}{}_{|t=0} = 0.$$

Thanks to the inequalities (29) and (31), we have

$$\|\theta_{1,2}\|_{\widetilde{L}^\infty_t \dot{B}^0_{\infty,1}} \leq Ce^{C\|\theta_1\|_{L^1_t \dot{B}^1_{\infty,1}}} \int_0^t \|\theta_{1,2}\|_{\widetilde{L}^\infty_\tau \dot{B}^0_{\infty,1}} \|\theta_2(\tau)\|_{\dot{B}^1_{\infty,1}} d\tau.$$

Thus Gronwall's inequality gives the desired result.

Step 4: *Smoothing effect.* We will show the precise estimate: $\forall\, \beta \in \mathbb{R}_+$ we have

$$\|t^\beta\theta(t)\|_{\widetilde{L}^\infty_T \dot{B}^\beta_{\infty,1}} \leq C_\beta e^{C(\beta+1)\|\theta\|_{L^1_T \dot{B}^1_{\infty,1}}} \|\theta\|_{\widetilde{L}^\infty_T \dot{B}^0_{\infty,1}}. \tag{32}$$

It is clear that

$$\begin{cases} \partial_t(t^\beta\theta) + v\cdot\nabla(t^\beta\theta) + |\mathrm{D}|(t^\beta\theta) = \beta t^{\beta-1}\theta, \\ (t^\beta\theta)_{|t=0} = 0. \end{cases}$$

We will proceed by induction and start the proof with the case $\beta \in \mathbb{N}$.

For $\beta = 1$, we apply Theorem 1.2 with $\bar{r} = +\infty$,

$$\|t\theta(t)\|_{\widetilde{L}^\infty_T \dot{B}^1_{\infty,1}} \lesssim e^{C\|\theta\|_{L^1_T \dot{B}^1_{\infty,1}}} \|\theta\|_{\widetilde{L}^\infty_T \dot{B}^0_{\infty,1}}.$$

Assume that (32) holds for degree n; we will prove it for $n+1$.

Applying Theorem 1.2 to the equation of $t^{n+1}\theta$, we get

$$\begin{aligned}\|t^{n+1}\theta(t)\|_{\widetilde{L}^\infty_T \dot{B}^{n+1}_{\infty,1}} &\le C(n+1) e^{C\|\theta\|_{L^1_T \dot{B}^1_{\infty,1}}} \|t^n\theta\|_{\widetilde{L}^\infty_T \dot{B}^n_{\infty,1}} \\ &\le C_n e^{C(n+2)\|\theta\|_{L^1_T \dot{B}^1_{\infty,1}}} \|\theta\|_{\widetilde{L}^\infty_T \dot{B}^0_{\infty,1}}.\end{aligned}$$

For $\beta \in \mathbb{R}_+$, we have $[\beta] \le \beta < [\beta]+1$, and by interpolation, one has

$$\|t^\beta\theta\|_{\widetilde{L}^\infty_T(\dot{B}^\beta_{\infty,1})} \lesssim \Big\|t^{[\beta]}\theta\Big\|^{1+[\beta]-\beta}_{\widetilde{L}^\infty_T(\dot{B}^{[\beta]}_{\infty,1})} \Big\|t^{[\beta]+1}\theta\Big\|^{\beta-[\beta]}_{\widetilde{L}^\infty_T(\dot{B}^{[\beta]+1}_{\infty,1})}.$$

This completes the proof. □

5.2. Blowup criteria. The main result of this section is the following.

PROPOSITION 5.2. *Let T^* be the maximum local existence time of θ in $\widetilde{L}^\infty_T \dot{B}^0_{\infty,1} \cap L^1_T \dot{B}^1_{\infty,1}$. There exists an absolute constant $\varepsilon_0 > 0$ such that if $T^* < \infty$, then*

$$\liminf_{t \to T^*} (T^* - t)\|\nabla\theta(t)\|_{L^\infty} \ge \varepsilon_0.$$

Proof. From local existence theory and especially (27) we see that if $T^* < \infty$, then necessarily

$$\liminf_{t \to T^*} \sum_{q \in \mathbb{Z}} (1 - e^{-c(T^*-t)2^q})^{\frac{1}{2}} \|\theta_q(t)\|_{L^\infty} \ge \varepsilon_0;$$

otherwise we can continue the solution over T^*. It follows that

$$\liminf_{t \to T^*} \sum_{q \in \mathbb{Z}} (1 - e^{-c(T^*-t)2^q})^{\frac{1}{2}} \sup_{t \le T^*} \|\theta_q(t)\|_{L^\infty} \ge \varepsilon_0.$$

Consequently from the Lebesgue theorem we obtain

$$\|\theta\|_{\widetilde{L}^\infty_{T^*}(\dot{B}^0_{\infty,1})} = \infty.$$

Using the Bernstein inequality and the fact that $\|\theta_q\|_{L^\infty} \lesssim \|\theta^0\|_{L^\infty}$, we have

$$\begin{aligned}\varepsilon_0 &\le \liminf_{t \to T^*} \Bigg\{ \sum_{q \le N} (1 - e^{-c(T^*-t)2^q})^{\frac{1}{2}} \|\theta_q(t)\|_{L^\infty} \\ &\qquad + \sum_{q \ge N} (1 - e^{-c(T^*-t)2^q})^{\frac{1}{2}} \|\theta_q(t)\|_{L^\infty} \Bigg\} \\ &\lesssim \liminf_{t \to T^*} \Bigg\{ (T^*-t)^{\frac{1}{2}} \|\theta^0\|_{L^\infty} \sum_{q \le N} 2^{q/2} + \|\nabla\theta(t)\|_{L^\infty} \sum_{q \ge N} 2^{-q} \Bigg\} \\ &\lesssim \liminf_{t \to T^*} \Big\{ (T^*-t)\|\theta^0\|_{L^\infty} 2^N + \|\nabla\theta(t)\|_{L^\infty} 2^{-N} \Big\}.\end{aligned}$$

Choosing judiciously N, we obtain the desired result. □

5.3. Global existence. We will use the idea of [16]. Let T^* be the maximal time existence of the solution in the space $\widetilde{L}^{\infty}_{\rm loc}([0,T^*[,\dot{B}^0_{\infty,1})\cap L^1_{\rm loc}([0,T^*[,\dot{B}^1_{\infty,1})$. From the local existence, there exists $T_0>0$ such that

$$\forall t\in[0,T_0],\quad t\|\nabla\theta(t)\|_{L^\infty}\leq C\|\theta^0\|_{\dot{B}^0_{\infty,1}}.$$

Let λ be a real positive number that will be fixed later and $T_1\in\,]0,T_0[$. We define the set

$$I:=\Big\{T\in[T_1,T^*[;\forall t\in[T_1,T],\forall x\neq y\in\mathbb{R}^2,|\theta(t,x)-\theta(t,y)|<\omega_\lambda(|x-y|)\Big\},$$

where

$$\omega:\ \ \mathbb{R}_+\longrightarrow\mathbb{R}_+$$

is strictly nondecreasing, concave, $\omega(0)=0$, $\omega'(0)<+\infty$, $\lim_{\xi\longrightarrow 0^+}\omega''(\xi)=-\infty$, and

$$\omega_\lambda(|x-y|)=\omega(\lambda|x-y|).$$

The function ω is a modulus of continuity chosen as in [16]. We shall first check that I is nonempty. It suffices for this purpose to prove that T_1 belongs to I under suitable conditions over λ. Let C_0 be a large positive number such that

$$\omega(C_0)>2\|\theta^0\|_{L^\infty}. \tag{33}$$

Since ω is a strictly nondecreasing function, then we get from the maximum principle that

$$\forall x,y,\quad \lambda|x-y|\geq C_0\Rightarrow|\theta(T_1,x)-\theta(T_1,y)|\leq 2\|\theta^0\|_{L^\infty}<\omega_\lambda(|x-y|).$$

On the other hand, we have from the mean value theorem

$$|\theta(T_1,x)-\theta(T_1,y)|\leq|x-y|\|\nabla\theta(T_1)\|_{L^\infty}.$$

Let $0<\delta_0<C_0$. Then using the concavity of ω, one obtains

$$\lambda|x-y|\leq\delta_0\Rightarrow\omega_\lambda(|x-y|)\geq\frac{\omega(\delta_0)}{\delta_0}\lambda|x-y|.$$

If we choose λ so that

$$\lambda>\frac{\delta_0}{\omega(\delta_0)}\|\nabla\theta(T_1)\|_{L^\infty},$$

then we get

$$0<\lambda|x-y|\leq\delta_0\Rightarrow|\theta(T_1,x)-\theta(T_1,y)|<\omega_\lambda(|x-y|).$$

Let us now move to the case $\delta_0\leq\lambda|x-y|\leq C_0$. By an obvious computation we find

$$|\theta(T_1,x)-\theta(T_1,y)|\leq\frac{C_0}{\lambda}\|\nabla\theta(T_1)\|_{L^\infty}\quad\text{and}$$

$$\omega(\delta_0)\leq\omega(\lambda|x-y|).$$

Choosing λ such that

$$\lambda > \frac{C_0}{\omega(\delta_0)}\|\nabla\theta(T_1)\|_{L^\infty},$$

then we obtain

$$\delta_0 \leq \lambda|x-y| \leq C_0 \Rightarrow |\theta(T_1,x)-\theta(T_1,y)| < \omega_\lambda(|x-y|).$$

All the preceding conditions over λ can be obtained if we take

$$\lambda = \frac{\omega^{-1}(3\|\theta^0\|_{L^\infty})}{2\|\theta^0\|_{L^\infty}}\|\nabla\theta(T_1)\|_{L^\infty}. \tag{34}$$

From the construction, the set I is an interval of the form $[T_1, T_*)$. We have three possibilities. The first one is $T_* = T^*$, and in this case we must have $T^* = +\infty$ because the Lipschitz norm of θ does not blow up. The second one is $T_* \in I$, and we will show that is not possible. Indeed, let C_0 be as in (33); then $\forall\ t \in [T_1, T^*)$

$$\lambda|x-y| \geq C_0 \Rightarrow |\theta(t,x)-\theta(t,y)| < \omega_\lambda(|x-y|).$$

Since $\nabla\theta(t)$ belongs to $C(]0,T^*); \dot{B}^0_{\infty,1})$, then for $\epsilon > 0$ there exist $\eta_0, R > 0$ such that

$$\forall t \in [T_*, T_*+\eta_0] \Rightarrow \|\nabla\theta(t)\|_{L^\infty} \leq \|\nabla\theta(T_*)\|_{L^\infty} + \epsilon/2 \quad \text{and}$$

$$\|\nabla\theta(T_*)\|_{L^\infty(B^c_{(0,R)})} \leq \epsilon/2,$$

where $B_{(0,R)}$ is the ball of radius R and with center the origin.

Hence for $\lambda|x-y| \leq C_0$ and x or $y \in B^c_{(0,R+\frac{C_0}{\lambda})}$ we have for $t \in [T_*, T_*+\eta_0]$

$$\begin{aligned}|\theta(t,x)-\theta(t,y)| &\leq |x-y|\|\nabla\theta(t)\|_{L^\infty(B^c_{(0,R)})}\\ &\leq \epsilon|x-y|.\end{aligned}$$

On the other hand, from the concavity of ω we have

$$\lambda|x-y| \leq C_0 \Rightarrow \frac{\omega(C_0)}{C_0}\lambda|x-y| \leq \omega_\lambda(|x-y|).$$

Thus if we take ϵ sufficiently small such that

$$\epsilon < \frac{\omega(C_0)}{C_0}\lambda,$$

then we find that

$$\lambda|x-y| \leq C_0,\ x \text{ or } y \in B^c_{(0,R+\frac{C_0}{\lambda})} \Rightarrow |\theta(t,x)-\theta(t,y)| < \omega_\lambda(|x-y|).$$

It remains to study the case where $x, y \in B_{(0,R+\frac{C_0}{\lambda})}$. Since $\|\nabla^2\theta(T_*)\|_{L^\infty}$ is finite (see Proposition 5.1), we get for each $x \in \mathbb{R}^2$

$$|\nabla\theta(T_*,x)| < \lambda\omega'(0).$$

For the proof, see [16, p. 3]. From the continuity of $x \longmapsto |\nabla\theta(T_*, x)|$ we obtain

$$\|\nabla\theta(T_*)\|_{L^\infty(B_{(0,R+\frac{C_0}{\lambda})})} < \lambda\omega'(0).$$

Let $\delta_0 \ll 1$; then using the continuity in time of the quantity $\|\nabla\theta(t)\|_{L^\infty}$, one can find $\eta_1 > 0$ such that $\forall t \in [T_*, T_* + \eta_1]$

$$\|\nabla\theta(t)\|_{L^\infty(B_{(0,R+\frac{C_0}{\lambda})})} < \lambda\frac{\omega(\delta_0)}{\delta_0}.$$

For $\lambda|x-y| \le \delta_0$ and $x \neq y$ belonging together to $B_{(0,R+\frac{C_0}{\lambda})}$ we have

$$\begin{aligned}|\theta(t,x) - \theta(t,y)| &\le |x-y|\|\nabla\theta(t)\|_{L^\infty(B_{(0,R+\frac{C_0}{\lambda})})} \\ &< \lambda|x-y|\frac{\omega(\delta_0)}{\delta_0} \le \omega_\lambda(|x-y|).\end{aligned}$$

Now for the other case we have

$$\forall x, y \in B_{(0,R+\frac{C_0}{\lambda})}, \quad \delta_0 \le \lambda|x-y|; \quad |\theta(T_*,x) - \theta(T_*,y)| < \omega_\lambda(|x-y|);$$

then we get from a standard compact argument the existence of $\eta_2 > 0$ such that $\forall\ t \in [T_*, T_* + \eta_2]$

$$\forall x, y \in B_{(0,R+\frac{C_0}{\lambda})}, \quad \delta_0 \le \lambda|x-y|; \quad |\theta(t,x) - \theta(t,y)| < \omega_\lambda(|x-y|).$$

Taking $\eta = \min\{\eta_0, \eta_1, \eta_2\}$, we obtain that $T_* + \eta \in I$, which contradicts the fact that T_* is maximal.

The last case that we have to treat is that T_* does not belong to I. Thus we have by the time continuity of θ the existence of $x \neq y$ such that

$$\theta(T_*, x) - \theta(T_*, y) = \omega_\lambda(\xi), \quad \text{with} \quad \xi = |x-y|.$$

We will show that this scenario cannot occur, and more precisely

$$f'(T_*) < 0, \quad \text{where} \quad f(t) = \theta(t,x) - \theta(t,y).$$

This is impossible since $f(t) \le f(T_*)\ \forall t \in [0, T_*]$. The proof is the same as in [16], but for the convenience of the reader we will outline the proof. From the regularity of the solution we see that $(\mathrm{QG})_\alpha$ can be defined in the classical manner and

$$f'(T_*) = (u\cdot\nabla\theta)(T_*,x) - (u\cdot\nabla\theta)(T_*,y) + |\mathrm{D}|\theta(T_*,x) - |\mathrm{D}|\theta(T_*,y).$$

From [16] we have

$$(u\cdot\nabla\theta)(T_*,x) - (u\cdot\nabla\theta)(T_*,y) \le \Omega_\lambda(\xi)\omega'_\lambda(\xi),$$

where

$$\Omega_\lambda(\xi) = C\left(\int_0^\xi \frac{\omega_\lambda(\eta)}{\eta}d\eta + \xi\int_\xi^\infty \frac{\omega_\lambda(\eta)}{\eta^2}d\eta\right) = \Omega(\lambda\xi).$$

Again from [16],

$$|\mathrm{D}|\theta(T_*,x)-|\mathrm{D}|\theta(T_*,y)\le \frac{1}{\pi}\int_0^{\frac{\xi}{2}}\frac{\omega_\lambda(\xi+2\eta)+\omega_\lambda(\xi-2\eta)-2\omega_\lambda(\xi)}{\eta^2}d\eta$$
$$+\frac{1}{\pi}\int_{\frac{\xi}{2}}^{\infty}\frac{\omega_\lambda(2\eta+\xi)-\omega_\lambda(2\eta-\xi)-2\omega_\lambda(\xi)}{\eta^2}d\eta$$
$$\le \lambda\,\mathcal{I}(\lambda\xi),$$

where

$$\mathcal{I}(\xi)=\frac{1}{\pi}\int_0^{\frac{\xi}{2}}\frac{\omega(\xi+2\eta)+\omega(\xi-2\eta)-2\omega(\xi)}{\eta^2}d\eta$$
$$+\frac{1}{\pi}\int_{\frac{\xi}{2}}^{\infty}\frac{\omega(2\eta+\xi)-\omega(2\eta-\xi)-2\omega(\xi)}{\eta^2}d\eta.$$

Thus we get

$$f'(T_*)=\lambda(\Omega\omega'+\mathcal{I})(\lambda\xi).$$

Now, we choose the same function as [16] (see p. 5):

$$\omega(\xi)=\xi-\xi^{\frac{3}{2}}\quad \text{if } \xi\in[0,\delta],$$

and

$$\omega'(\xi)=\frac{\gamma}{\xi(4+\log(\xi/\delta))}\quad \text{if } \xi>\delta,$$

where δ and γ are small numbers and satisfy $0<\gamma<\delta$. It is shown in [16] that

$$\Omega(\xi)\omega'(\xi)+\mathcal{I}(\xi)<0\qquad \forall\xi\neq 0.$$

This yields $f'(T_*)<0$.

Finally we have $T^*=+\infty$ and

$$\forall t\in[T_1,+\infty),\quad \|\nabla\theta(t)\|_{L^\infty}\le\lambda.$$

The value of λ is given by (34).

REFERENCES

[1] J.-M. Bony, *Calcul symbolique et propagation des singularités pour les équations aux dérivées partielles non linéaires*, Ann. Ecole Norm. Sup., 14 (1981), pp. 209–246.

[2] L. Caffarelli and V. Vasseur, *Drift Diffusion Equations with Fractional Diffusion and the Quasi-Geostrophic Equations*, preprint, http://arXiv.org/abs/math/0608447.

[3] D. Chae and J. Lee, *Global well-posedness in the supercritical dissipative quasi-geostrophic equations*, Asymptot. Anal., 38 (2004), pp. 339–358.

[4] J.-Y. Chemin, *Perfect Incompressible Fluids*, Clarendon Press, Oxford University Press, New York, 1998.

[5] Q. Chen, C. Miao, and Z. Zhang, *A new Bernstein's inequality and the 2D dissipative quasi-geostrophic equation*, Comm. Math. Phys., 271 (2007), pp. 821–838.

[6] P. Constantin, A. Majda, and E. Tabak, *Formation of strong fronts in the* 2*D quasi-geostrophic thermal active scalar*, Nonlinearity, 7 (1994), pp. 1495–1533.
[7] P. Constantin, D. Córdoba, and J. Wu, *On the critical dissipative quasi-geostrophic equation*, Indiana Univ. Math. J., 50 (2001), pp. 97–107.
[8] P. Constantin and J. Wu, *Behavior of solutions of* 2*D quasi-geostrophic equations*, SIAM J. Math. Anal., 30 (1999), pp. 937–948.
[9] A. Córdoba and D. Córdoba, *A maximum principle applied to quasi-geostrophic equations*, Comm. Math. Phys., 249 (2004), pp. 511–528.
[10] R. Danchin, *Density-dependent incompressible viscous fluids in critical spaces*, Proc. Roy. Soc. Ed., 133 (2003), pp. 1311–1334.
[11] T. Hmidi, *Régularité höldérienne des poches de tourbillon visqueuses*, J. Math. Pures Appl. (9), 84 (2005), pp. 1455–1495.
[12] T. Hmidi and S. Keraani, *Global solutions of the super-critical* 2*D quasi-geostrophic equation in Besov spaces*, Adv. Math., 214 (2007), pp. 618–638.
[13] N. Ju, *Existence and uniqueness of the solution to the dissipative* 2*D quasi-geostrophic equations in the Sobolev space*, Comm. Math. Phys., 251 (2004), pp. 365–376.
[14] N. Ju, *On the two dimensional quasi-geostrophic equations*, Indiana Univ. Math. J., 54 (2005), pp. 897–926.
[15] N. Ju, *Global solutions to the two dimensional quasi-geostrophic equation with critical or super-critical dissipation*, Math. Ann., 334 (2006), pp. 627–642.
[16] A. Kiselev, F. Nazarov, and A. Volberg, *Global well-posedness for the critical* 2*D dissipative quasi-geostrophic equation*, Invent. Math., 167 (2007), pp. 445–453.
[17] F. Marchand and P.-G. Lemarié-Rieusset, *Solutions auto-similaires non radiales pour l'quation quasi-gostrophique dissipative critique*, C. R. Math. Acad. Sci. Paris, 341 (2005), pp. 535–538.
[18] J. Peetre, *New Thoughts on Besov Spaces*, Duke University Mathematical Series 1, Duke University Press, Durham, NC, 1976.
[19] J. Pedlosky, *Geophysical Fluid Dynamics*, Springer-Verlag, New York, 1987.
[20] S. Resnick, *Dynamical Problem in Nonlinear Advective Partial Differential Equations*, Ph.D. thesis, University of Chicago, 1995.
[21] H. Triebel, *Theory of Function Spaces*, Leipzig, Germany, 1983.
[22] M. Vishik, *Hydrodynamics in Besov spaces*, Arch. Ration. Mech. Anal., 145 (1998), pp. 197–214.
[23] J. Wu, *Solutions to the* 2*D quasi-geostrophic equations in Hölder spaces*, Nonlinear Anal., 62 (2005), pp. 579–594.
[24] J. Wu, *Global solutions of the* 2*D dissipative quasi-geostrophic equation in Besov spaces*, SIAM J. Math. Anal., 36 (2005), pp. 1014–1030.

SIAM J. MATH. ANAL.
Vol. 40, No. 1, pp. 186–214

A Γ-CONVERGENCE RESULT FOR THIN MARTENSITIC FILMS IN LINEARIZED ELASTICITY*

PETER HORNUNG†

Abstract. The elastic energy of a thin film Ω_h of thickness h with displacement $u : \Omega_h \to \mathbb{R}^3$ is given by the functional $E^h(u) = \int_{\Omega_h} W(\nabla u)$. We consider materials whose energy density W is linearly frame indifferent and vanishes on two linearized wells which are compatible in the plane but incompatible in the thickness direction. We prove compactness of displacement sequences $u^{(h)} : \Omega_h \to \mathbb{R}^3$ satisfying $E^h(u^{(h)}) \leq Ch^2$, and we derive the Γ-limit of the functionals $\frac{1}{h^2}E^h$ as $h \to 0$.

Key words. singular variational problems, thin films, martensitic phase transitions, Γ-convergence

AMS subject classifications. 74G65, 49J45, 74N99

DOI. 10.1137/070683167

1. Introduction.

The study of solid-solid phase transitions in thin elastic films leads to functionals of the form

$$E^h(v) = \int_{\Omega_h} W(\nabla v(x))\, dx, \tag{1}$$

where $\Omega_h = S \times (-\frac{h}{2}, \frac{h}{2})$ is a cylindrical domain of thickness h, $S \subset \mathbb{R}^2$ is a bounded Lipschitz domain, $v : \Omega_h \to \mathbb{R}^3$ is the elastic deformation (in the nonlinear setting) or the displacement (in the linearized setting), and W is a free energy density with n energy minima F_i, i.e., $W(F_i) = 0$ for $i = 1, \ldots, n$. In the context of nonlinear elasticity W is invariant under proper rotations, and in the context of linearized elasticity it is invariant under addition of skew symmetric matrices. In [10] it was observed (in the context of nonlinear elasticity) that for many materials which undergo austenite-martensite phase transitions, the low-energy states of very thin samples of material display a much richer variety of structures than bulk samples made of the same material. The reason is that three dimensional compatibility requires a plane on which two juxtaposed affine deformations coincide, i.e., that their gradients be rank-one connected. In contrast, two dimensional compatibility is already satisfied if there exists one in-plane vector on which the two deformations agree, so a rank-two connection between the gradients suffices. This fact leads to the existence of many nontrivial low-energy states, including laminates, tunnels, and tents; see, e.g., [11] for experimental results. This rich structure makes thin martensitic films particularly interesting for applications.

In this article we study the asymptotic behavior of thin martensitic films in the context of linearized elasticity. In our model the zero set of the energy density W consists of two linearized wells which are incompatible in bulk but compatible in the plane (see section 2 for details). We study the asymptotic behavior of the functionals (1) in the thin film-limit $h \to 0$. We prove compactness of displacement sequences

*Received by the editors February 20, 2007; accepted for publication (in revised form) December 3, 2007; published electronically April 4, 2008. This work was supported by the EU research and training network MULTIMAT MRTN-CT-2004-505226.

http://www.siam.org/journals/sima/40-1/68316.html

†Fachbereich Mathematik, Universität Duisburg-Essen, Lotharstrasse 65, 47057 Duisburg, Germany (peter.hornung@uni-due.de).

whose energy scales like h^2, and we derive the Γ-limit of the functionals $\frac{1}{h^2}E^h$ as the film thickness h converges to zero. To our knowledge this is the first Γ-convergence result for thin martensitic films in which both the domain and the image space are three dimensional and no interfacial energy term is added to the elastic energy. In our model, the formation of interfaces is penalized in a natural way by the interplay of nonzero film thickness with incompatibility of the energy wells in the thickness direction.

Thin films of single-well materials have been studied, e.g., in [1, 3, 12, 17] (in a linearized setting) and in [34, 13, 28, 29] (in a nonlinear setting). Thin films of multiwell materials have been studied, e.g., in [19] (in a linearized setting) and in [37, 10, 15, 6] (in a nonlinear setting).

To state our main result let us introduce the functionals

$$I^h(u;S) = \begin{cases} \frac{1}{h^2}\int_{S\times(-\frac{h}{2},\frac{h}{2})} W(\nabla u(x))\,dx & \text{if } u \in W^{1,2}(S\times(-\frac{h}{2},\frac{h}{2});\mathbb{R}^3), \\ +\infty & \text{otherwise} \end{cases}$$

and

$$I^0(w;S) = \begin{cases} \int_J k(\nu(x))d\mathcal{H}^1(x) & \text{if } w \in \mathcal{A}(S), \\ +\infty & \text{otherwise,} \end{cases}$$

where the class $\mathcal{A}(S)$ of admissible limiting displacements is given in (30) below, J denotes the jump set of $\operatorname{sym}\nabla' w \in BV$, and ν denotes the normal to it, which can assume only two values (up to a sign). The function k is a "surface tension" which depends on the normal and which we define in (29) below. Let us write v' to denote the first two entries of $v \in \mathbb{R}^3$, and let us call a domain $S \subset \mathbb{R}^2$ strictly star-shaped if there is $z \in S$ such that for all $z' \in \bar{S}$ the open segment (z, z') is contained in S. Our main result is the following theorem.

THEOREM 1. *Let $A, B \in \mathbb{R}^{3\times 3}$ satisfy* (i)–(iv) *from section* 2, *let W satisfy the conditions* (4)–(6) *below, and let $S \subset \mathbb{R}^2$ be a bounded strictly star-shaped Lipschitz domain. Then a Γ-type convergence $I^h(\cdot;S) \xrightarrow{\Gamma} I^0(\cdot;S)$ holds in the following sense:*

(i) Ansatz-free lower bound. *Let $w \in L^2(S;\mathbb{R}^2)$, $h_n \to 0$, let $v_n \in W^{1,2}(S\times(-\frac{h_n}{2},\frac{h_n}{2});\mathbb{R}^3)$, and set $w_n(x') = \frac{1}{h_n}\int_{-\frac{h_n}{2}}^{\frac{h_n}{2}} v_n'(x',x_3)\,dx_3$. If $w_n \to w$ in $L^2(S;\mathbb{R}^2)$, then $\liminf_{n\to\infty} I^{h_n}(v_n;S) \geq I^0(w;S)$.*

(ii) Existence of recovery sequences. *Let $w \in L^2(S;\mathbb{R}^2)$ and $h_n \to 0$. Then there is a sequence $v_n \in W^{1,2}(S\times(-\frac{h_n}{2},\frac{h_n}{2});\mathbb{R}^3)$ such that, setting $w_n(x') = \frac{1}{h_n}\int_{-\frac{h_n}{2}}^{\frac{h_n}{2}} v_n'(x',x_3)\,dx_3$, we have $w_n \to w$ strongly in $W^{1,2}(S;\mathbb{R}^2)$ and $\lim_{n\to\infty} I^{h_n}(v_n;S) = I^0(w;S)$.*

Remarks. (i) Theorem 1 is complemented by a compactness result for sequences v_n whose energy $E^{h_n}(v_n)$ scales like h_n^2 (Theorem 6 below).

(ii) Notice that in Theorem 1(ii) we state the existence of recovery sequences for any given sequence $h_n \to 0$.

(iii) The vertical average w_n can be interpreted as the in-plane displacement of the midplane S of the thin film; compare, e.g., [3].

(iv) The lower bound (i) is true for general (also non–star-shaped) Lipschitz domains S; see Theorem 12. Star-shapedness is used in the proof of the upper bound (ii) to show that limiting displacements with finitely many well-separated interfaces are energy dense and to avoid the necessity of a lateral matching of two local recovery sequences. The same technical difficulty concerning non–star-shaped domains occurs

in [19, 20] and in [18]. A figure depicting the problematic situation can be found in [18, Figure 4].

(v) The Γ-limit obtained in Theorem 1 has the same structure as that derived in [19]. The reason is that the functionals I^h are related to singularly perturbed functionals of the form

$$J^{(\varepsilon)}(u;S) = \int_S \frac{1}{\varepsilon} W_{2D}(\nabla u(x')) + \varepsilon |\nabla^2 u(x')|^2 \, dx'. \tag{2}$$

The asymptotic behavior of singularly perturbed functionals has been extensively studied in the literature (see [35, 27, 26, 5, 38] and, for gradient phase transitions, see [25, 18, 19, 20]). Recently, Conti and Schweizer derived the Γ-limit of the functionals (2) both under the assumption of linearized frame indifference [19] and under nonlinear frame indifference [20] of W_{2D}. Their results are restricted to two dimensions.

There are two crucial differences between functionals of the form (2) and the model studied in this article: In the former the domain and the image space are two dimensional and an extra term $\int |\nabla^2 u|^2$, weighted with some small parameter ε^2, is added to the elastic energy. The role of this term is to penalize the formation of phase interfaces. In our model (1), the domain and the image space are genuinely three dimensional and the energy functional does not involve higher derivatives: It is a key property of our model that no interfacial energy contribution is added to the elastic energy (this also contrasts with other thin film models [10, 37, 6]). The formation of phase interfaces is naturally penalized by the interplay of nonzero film thickness with the fact that the zero energy displacements $A + \text{Skew}$ and $B + \text{Skew}$ are incompatible in the thickness direction.

Lemma 14 makes the relation between (1) and (2) more precise. This requires a subtle mollification argument since (2) requires control on the second derivatives. Lemma 14 suggests that the small parameter ε should be interpreted as the film thickness h.

(vi) No nonlinear version of Theorem 1 has yet been proven. The only result in this direction is [15], where it is shown that the energy of thin-film deformations consisting of two phases scales like h^2. Notice that, in contrast to the linearized setting, the model considered in [20] would not be appropriate to describe thin martensitic films since it is too rigid. Nonlinearly elastic rods of multiphase materials were studied in [36].

This article is organized as follows. In section 2 we introduce some definitions and reduce the problem to a canonical form. Then we prove a two-well analogue of Korn's inequality, Theorem 3, which applies to incompatible linear wells. Then we apply this result to deduce the compactness result Theorem 6. In section 3 we obtain the lower bound, Theorem 12, by an abstract scaling argument. Finally, in section 4 we derive the upper bound by constructing three dimensional recovery sequences. The proof of Theorem 1 closes section 4.

Notation. We use the letter C to denote constants depending only on the domain and on W. Within an expression the explicit value of C may change from line to line. A bar above a given 3×3 matrix denotes its upper left 2×2 submatrix, and in general we use barred letters to denote 2×2 matrices. Primes on 3-vectors will denote the 2-vector consisting of the first two entries, so in particular $x = (x', x_3)$. For a matrix A we write $\operatorname{sym} A = \frac{1}{2}(A + A^T)$, $\operatorname{skew} A = \frac{1}{2}(A - A^T)$, and $|A|^2 = \operatorname{Tr}(A^T A)$, where Tr denotes the trace. By a subscript $_i$ we will denote the partial derivative with respect to the x_i-variable. By ∇' we denote the in-plane gradient, that is, $\nabla' w = (w_{,1}|w_{,2})$. For $h > 0$ we set $I_h = (-\frac{h}{2}, \frac{h}{2})$. All intervals in this article are implicitly assumed to

be nonempty and bounded. We use a dashed integral sign $\fint$ to denote the average. Often we will simply write $\{f = a\}$ instead of $\{x \in S : f(x) = a\}$. For $\rho > 0$ we set $[\rho] = \max\{n \in \mathbb{N} : n \le \rho\}$. If $E \subset \mathbb{R}^n$, then $|E|$ denotes its n dimensional Lebesgue measure and $\mathcal{H}^k(E)$ its k dimensional Hausdorff measure [24]. For $j \in \{1, 2\}$ we denote by e_j the jth unit vector, by $\pi_j : \mathbb{R}^2 \to \operatorname{span}\{e_j\}$ the orthogonal projection onto $\operatorname{span}\{e_j\}$, and by $\pi_j^\perp : \mathbb{R}^2 \to \{e_j\}^\perp$ the orthogonal projection onto $\{e_j\}^\perp$. If $E \subset \mathbb{R}^n$, then $B_\varepsilon(E) = \{x \in \mathbb{R}^n : \operatorname{dist}_E(x) < \varepsilon\}$.

2. Preliminaries and compactness. We consider the functional, defined for any Lipschitz domain $U \subset \mathbb{R}^2$,

$$I^h(u; U) = \begin{cases} \frac{1}{h^2} \int_{U \times I_h} W(\nabla u(x))\, dx & \text{if } u \in W^{1,2}(U \times I_h; \mathbb{R}^3), \\ +\infty & \text{otherwise.} \end{cases} \tag{3}$$

Throughout this article, W is assumed to satisfy the following conditions:

$$W : \mathbb{R}^{3\times 3} \to \mathbb{R} \text{ is continuous,} \tag{4}$$

$$\text{linearized frame indifference: } W(F) = W(\operatorname{sym} F) \text{ for all } F \in \mathbb{R}^{3\times 3}, \tag{5}$$

$$\text{quadratic growth and coercivity: } c_0 W_0(F) \le W(F) \le C_0 W_0(F), \tag{6}$$

where c_0, C_0 are positive constants. Here we have introduced the standard energy density $W_0(F) = \operatorname{dist}^2(\operatorname{sym} F, \{A, B\})$, where A and B are symmetric 3×3 matrices to be specified below. We define the reduced functional

$$I^h_{2D}(w; U) = \begin{cases} \int_U \frac{1}{h} W_{2D}(\nabla' w) + h|\nabla'^2 w|^2 & \text{if } w \in W^{2,2}(U; \mathbb{R}^2), \\ +\infty & \text{otherwise,} \end{cases}$$

where $W_{2D}(\bar F) = \operatorname{dist}^2(\operatorname{sym} \bar F, \{\bar A, \bar B\})$. We make the following assumptions on the wells A and B:

(i) $A \in \mathbb{R}^{3\times 3}$ and $B \in \mathbb{R}^{3\times 3}$ are symmetric; i.e., $A = \operatorname{sym} A$ and $B = \operatorname{sym} B$.

(ii) *Incompatibility in bulk.* $\operatorname{rank}(A - B + T) \ge 2$ for all $T \in \mathbb{R}^{3\times 3}$ with $\operatorname{sym} T = 0$.

(iii) *Compatibility in the plane.* There exists $\bar T \in \mathbb{R}^{2\times 2}$ with $\operatorname{sym} \bar T = 0$ such that $\operatorname{rank}(\bar A - \bar B + \bar T) \le 1$.

(iv) *Nondegeneracy.* $\det(\bar A - \bar B) \ne 0$.

Item (iii) is satisfied if and only if there exists $t \in \mathbb{R}$ such that

$$0 = \det\left(\bar A - \bar B + \begin{pmatrix} 0 & t \\ -t & 0 \end{pmatrix}\right) = \det(\bar A - \bar B) + t^2,$$

whence (iii) is equivalent to $\det(\bar A - \bar B) \le 0$ with equality if and only if $\bar A$ and $\bar B$ are rank-one connected. Thus (iii) and (iv) together are equivalent to $\det(\bar A - \bar B) < 0$. Table 11.1 in [7] shows that conditions (i)–(iv) are generically satisfied by real materials (in a linearized framework).

Let us now reduce the set of all matrices satisfying (i)–(iv) to a canonical form. Let $\tilde A$, $\tilde B$ satisfy conditions (i)–(iv) but be arbitrary otherwise. Then there is an orthogonal matrix $R \in O(3)$ with $Re_3 = e_3$ such that

$$R^T(\tilde B - \tilde A)R = \lambda_1 e_1 \otimes e_1 + \lambda_2 e_2 \otimes e_2 + \sum_{i=1}^{3} \tilde\mu_i \frac{e_i \otimes e_3 + e_3 \otimes e_i}{2},$$

where λ_i are the eigenvalues of the matrix $\bar{B} - \bar{A}$ and $\tilde{\mu}_i$ are some real numbers. By possibly choosing R differently (by interchanging the first two columns), we may assume that $\lambda_1 \geq \lambda_2$, so since $\det(\bar{A} - \bar{B}) < 0$, we must in fact have $\lambda_1 > 0 > \lambda_2$. Let $Q = \operatorname{diag}(|\lambda_1|^{-\frac{1}{2}}, |\lambda_2|^{-\frac{1}{2}}, 1)$ and set $\hat{B} = QR^T(\tilde{B} - \tilde{A})RQ$. This gives $\hat{B} = e_1 \otimes e_1 - e_2 \otimes e_2 + \sum_{i=1}^{3} \hat{\mu}_i \frac{e_i \otimes e_3 + e_3 \otimes e_i}{2}$, where $\hat{\mu}_i$ are related to $\tilde{\mu}_i$ and λ_i. Now we can find a proper rotation $\hat{Q} \in SO(3)$ with eigenvector e_3 such that

$$B = \hat{Q}^T \hat{B} \hat{Q} = e_1 \otimes e_2 + e_2 \otimes e_1 + \sum_{i=1}^{3} \mu_i \frac{e_i \otimes e_3 + e_3 \otimes e_i}{2} \tag{7}$$

for some $\mu_i \in \mathbb{R}$. Since the structural assumptions on the energy density W and on the shape of the domain (i.e., strict star-shapedness with respect to the origin and a cylindrical form $S \times I_h$) are invariant under the transformations introduced above, we obtain the following lemma.

LEMMA 2. *If Theorem* 1 *is shown for the special pairs A, B given by $A = 0$ and B as in* (7)*, then it holds for all possible choices of A and B which satisfy conditions* (i)–(iv).

2.1. Korn's inequality for two incompatible strains. The following theorem provides a generalization of Korn's inequality to the case of two incompatible linearized wells. A nonquantitative version of this result can be found in [22]; compare also [39, 9, 8, 21]. In [19] an example is provided which shows that no Korn-type rigidity like the one derived here can be expected in the case of two compatible wells.

THEOREM 3. *Let $\Omega \subset \mathbb{R}^n$ be a bounded connected Lipschitz domain, $n \geq 2$, and $K = (A + \mathrm{Skew}) \cup (B + \mathrm{Skew})$, where A and B are incompatible strains, i.e., $(B - A) + \mathrm{Skew}$ does not contain rank-one matrices. Then there exists a positive constant $C(\Omega, A, B)$ with the following property: For every $u \in W^{1,2}(\Omega; \mathbb{R}^n)$, there exists an associated $R \in K$ such that $\|\nabla u - R\|_{L^2(\Omega;\mathbb{R}^{n\times n})} \leq C(\Omega, A, B) \|\operatorname{dist}(\nabla u, K)\|_{L^2(\Omega;\mathbb{R}^{n\times n})}$.*

This theorem will follow from the interior estimate provided by the following lemma.

LEMMA 4. *With assumptions as in Theorem* 3 *and $\bar{U} \subset \Omega$, where U is Lipschitz and connected, there is a constant $C(U, \Omega, A, B)$ such that the following holds: For every $u \in W^{1,2}(\Omega; \mathbb{R}^n)$, there exists an associated $R \in K$ such that*

$$\|\nabla u - R\|_{L^2(U;\mathbb{R}^{n\times n})} \leq C(U, \Omega, A, B) \|\operatorname{dist}(\nabla u, K)\|_{L^2(\Omega;\mathbb{R}^{n\times n})}. \tag{8}$$

Proof. From (15) on this proof follows [14] rather closely with some minor changes. Define $d(F) = \operatorname{dist}(F, \{A, B\})$ and set $\varepsilon^2 = \int_\Omega \operatorname{dist}^2(\nabla u, K)$. Notice that $\operatorname{dist}^2(F, K) = d^2(\operatorname{sym} F)$ for all $F \in \mathbb{R}^{n\times n}$. Since $|F - A| \leq d(F) + |A - B|$, by Korn's inequality there exists $C > 0$ with the property that (8) is satisfied whenever u is such that $\varepsilon \geq 1$. Hence we may assume without loss of generality that $\varepsilon < 1$. By setting $\tilde{B} = B - A$ and applying the lemma to $\tilde{u}(x) = u(x) - Ax$ we may also assume without loss of generality that $A = 0$.

Denote by $P : \mathbb{R}^{n\times n} \to \big((\operatorname{span}\{B\}) \oplus \mathrm{Skew}\big)^\perp$ the orthogonal projection onto the orthogonal complement of the subspace $(\operatorname{span}\{B\}) \oplus \mathrm{Skew}$. Since $B + \mathrm{Skew}$ does not contain rank-one matrices, the only rank-one matrix contained in $(\operatorname{span}\{B\}) \oplus \mathrm{Skew}$ is the zero matrix. Thus $|P(a \otimes b)|^2 > 0$ for all $a, b \neq 0$. Hence, by continuity and by compactness of the sphere, P satisfies the Legendre–Hadamard ellipticity condition $\Lambda |a|^2 |b|^2 \geq |P(a \otimes b)|^2 \geq \lambda |a|^2 |b|^2$ for some $\Lambda > \lambda > 0$. Now let $w \in W^{1,2}(\Omega; \mathbb{R}^n)$ be

a weak solution of the linear elliptic system with constant coefficients

$$\text{div } P(\nabla w) = 0 \text{ in } \Omega, \tag{9}$$
$$w = u \text{ on } \partial\Omega.$$

Set $z = u - w$. Then $z \in W_0^{1,2}(\Omega; \mathbb{R}^n)$ is a weak solution of div $P(\nabla z) =$ div $P(\nabla u)$. Testing this with z gives

$$\int_\Omega P(\nabla z) : \nabla z = \int_\Omega P(\nabla u) : \nabla z \le \left(\int_\Omega |P(\nabla u)|^2 \right)^{\frac{1}{2}} \left(\int_\Omega |\nabla z|^2 \right)^{\frac{1}{2}}. \tag{10}$$

Since by ellipticity the left-hand side of (10) is greater than $\int_\Omega \lambda |\nabla z|^2$, we conclude

$$\int_\Omega |\nabla z|^2 \le C \int_\Omega |P(\nabla u)|^2 = C \int_\Omega \text{dist}^2(\text{sym } \nabla u,\ \text{span}\{B\}) \le C\varepsilon^2. \tag{11}$$

Thus it remains to prove that there exists $R \in K$ such that $\int_\Omega |\nabla w(x) - R|^2 \, dx \le C\varepsilon^2$, where C is independent of w.

Set $e_w =$ sym ∇w and let $y \in \Omega$ be such that $B(y, 2r) \subset \Omega$. By Korn's inequality there is a $C = C(n)$ (which by scaling invariance is independent of r), and $T \in \mathbb{R}^{n \times n}$ with sym $T = 0$ such that

$$\int_{B(y,2r)} |\nabla w - T|^2 \le C \int_{B(y,2r)} |e_w|^2. \tag{12}$$

Since by $P(M) = P(\text{sym } M)$ we have $P(T) = 0$, the mapping $v(x) = w(x) - Tx$ is a weak solution of

$$\text{div } P(\nabla v) = 0 \text{ in } \Omega,$$
$$v = u - Tx \text{ on } \partial\Omega. \tag{13}$$

By standard elliptic regularity for linear systems with constant coefficients (see, e.g., [30]), we obtain the inequality

$$\int_{B(y,r)} |\nabla^2 v|^2 \le \frac{C}{r^2} \int_{B(y,2r)} |\nabla v|^2 = \frac{C}{r^2} \int_{B(y,2r)} |\nabla w - T|^2. \tag{14}$$

We have $|\nabla e_w|^2 = \frac{1}{4} \sum_{i,j,k} (w_{i,jk} + w_{j,ik})^2 \le |\nabla^2 w|^2$. Hence by the choice of T and since $|\nabla^2 w|^2 = |\nabla^2 v|^2$ on $B(y, 2r)$, we conclude from (12) and (14) that

$$\int_{B(y,r)} |\nabla e_w|^2 \le \frac{C}{r^2} \int_{B(y,2r)} |e_w|^2. \tag{15}$$

This inequality holds for all $y \in \Omega$ with $B(y, 2r) \subset \Omega$.

Fix $r_0 \in (0, \frac{\text{dist}(U, \partial\Omega)}{4})$ such that there exists $c_0 > 0$ with the property that $|B_r(x) \cap U| \ge c_0 |B_r|$ for all $x \in U$ and for all $r \le r_0$. (Here and in what follows we will sometimes omit the center of the ball when denoting its volume.) Existence of such an r_0 follows from the Lipschitz property of U, and c_0 will depend on U. Covering $\bar{U}$ with finitely many balls of radius $\frac{1}{3}$ dist$(U, \partial\Omega)$ and applying (15) shows that $\int_U |\nabla e_w|^2$ is bounded by a constant independent of u (since $|e_w| \le d(e_w) + C$

and $\int_\Omega d^2(e_w) \le C\varepsilon^2$ by (11)). Hence, by applying Lemma 5 below with $K_1 = \{0\}$, $K_2 = \{B\}$, and $F = e_w$, we obtain

$$\min\left\{\int_U |e_w - B|^2, \int_U |e_w|^2\right\} \le C\left(\int_U d^2(e_w)\int_U |\nabla e_w|^2\right)^{\frac{n}{2(n-1)}} + \int_U d^2(e_w) \le C(\varepsilon^{\frac{n}{n-1}} + \varepsilon^2). \tag{16}$$

Set $\rho = \frac{|B|}{2}$, and let us assume that B is the minority phase in U; i.e., the set $E = \{x \in U : |e_w(x) - B|^2 \le \rho^2\}$ satisfies $|E| \le |\{x \in U : |e_w(x)|^2 \le \rho^2\}|$. (The case when A is the minority phase is treated similarly.) In particular, this implies $|E| \le |U \setminus E|$ by the choice of ρ. We have $\rho^2|E| \le \int_U |e_w|^2$ because $|e_w| \ge \rho$ on E, and similarly $\rho^2|E| \le \rho^2|U \setminus E| \le \int_U |e_w - B|^2$. Thus by (16), whenever $\varepsilon < 1$,

$$|E| \le C_1\left(\varepsilon^{\frac{n}{n-1}} + \varepsilon^2\right) \tag{17}$$

for some constant C_1 independent of u. Now we fix $\varepsilon_0 \in (0,1)$ such that $C_1(\varepsilon_0^{\frac{n}{n-1}} + \varepsilon_0^2) < \frac{c_0}{2}|B_{r_0}|$. From now on we assume that $\varepsilon \le \varepsilon_0$; the other case is treated at the end of this proof. From (17) we deduce that $|E| < \frac{c_0}{2}|B_{r_0}|$. Our aim is to show that

$$|E| \le C\varepsilon^2 \tag{18}$$

for a constant C independent of u. Using Lemma 5 (notice that the constant in its conclusion is invariant under a rescaling of the domain) as in (16) with B_r replacing U, and then dividing through $|B_r|$ and applying (15), one obtains

$$\min\left\{\fint_{B_r(x)} |e_w|^2, \fint_{B_r(x)} |e_w - B|^2\right\} \le C\left[\left(\mathcal{M}(|e_w|^2)(x)\fint_{B_r(x)} d^2(e_w)\right)^{\frac{n}{2(n-1)}} + \fint_{B_r(x)} d^2(e_w)\right] \tag{19}$$

for all $x \in \Omega$ and for all $r > 0$ such that $B_{2r}(x) \subset \Omega$. Here $\mathcal{M}$ denotes the Hardy–Littlewood maximal function, $\mathcal{M}(f)(x) = \sup_{r>0} \fint_{B_r(x)} |f|$. Above and in what follows we extend e_w by zero outside Ω.

Claim 1. The set $A_\infty = \{x \in \Omega : \mathcal{M}(|e_w|^2)(x) \ge 10|B|^2\}$ satisfies $|A_\infty| \le C\varepsilon^2$.

In fact, $\mathcal{M}(|e_w|^2) \le \mathcal{M}(|e_w|^2 - 5|B|^2)_+ + 5|B|^2$, whence $x \in A_\infty$ implies $\mathcal{M}(|e_w|^2 - 5|B|^2)_+(x) \ge 5|B|^2$. Since $|e_w|^2 \le 2(d^2(e_w) + |B|^2)$, we have $d^2(e_w) \ge \frac{1}{2}(|e_w|^2 - 5|B|^2)_+$. We conclude that $A_\infty \subset \{\mathcal{M}(d^2(e_w)) \ge \frac{5}{2}|B|^2\}$. Thus, by the Hardy–Littlewood maximal theorem [31, Chapter 4], $|A_\infty| \le |\{\mathcal{M}(d^2(e_w)) \ge \frac{5}{2}|B|^2\}| \le C\int_\Omega d^2(e_w)$, which proves Claim 1.

For almost every $x \in E \setminus A_\infty$ there is an $r_x \le r_0$ such that

$$\frac{|E \cap B_{r_x}(x)|}{|B_{r_x}(x)|} = \frac{c_0}{2}. \tag{20}$$

In fact, by the Lebesgue point theorem, for almost all $x \in E \setminus A_\infty$, we have $\frac{|E\cap B_r(x)|}{|B_r|} \to 1$ as $r \to 0$. On the other hand, $\frac{|E\cap B_r(x)|}{|B_r(x)|} \le \frac{|E|}{|B_r|}$, which is strictly less than $c_0/2$ for $r > r_0$ by the choice of ε_0. In particular, $B_{2r_x}(x) \subset \Omega$ for every x as above, by the choice of r_0.

By Vitali's covering theorem [24, Theorem 1, section 1.5] we can choose countably many $x_i \in E \setminus A_\infty$ satisfying (20) and such that

$$|E \setminus A_\infty| \le C \sum |B_{r_{x_i}}(x_i)| \tag{21}$$

with pairwise disjoint balls on the right-hand side. By (20) and since $|e_w|^2 \ge \rho^2$ on E and $|e_w - B|^2 \ge \rho^2$ on $U \setminus E$, for every i we have

$$\begin{aligned} \frac{c_0\rho^2}{2} &\le \frac{1}{|B_{r_{x_i}}|} \min\left\{ \int_{B_{r_{x_i}}(x_i)\cap E} |e_w|^2, \int_{B_{r_{x_i}}(x_i)\cap U\setminus E} |e_w - B|^2 \right\} \\ &\le C\left[\left(\fint_{B_{r_{x_i}}(x_i)} d^2(e_w) \right)^{\frac{n}{2n-2}} + \fint_{B_{r_{x_i}}(x_i)} d^2(e_w) \right]. \end{aligned} \tag{22}$$

For the first inequality we have used that $x_i \in U$, whence $|B_{r_{x_i}}(x_i) \cap U| \ge c_0 |B_{r_{x_i}}|$ by definition of c_0, and so $|B_{r_{x_i}}(x_i) \cap U \setminus E| \ge \frac{c_0}{2}|B_{r_{x_i}}|$ by (20). For the second inequality we have used (19) and the definition of A_∞. From (22) we have $\frac{c_0\rho^2}{2} \le 2C \max\left\{ (\fint_{B_{r_{x_i}}(x_i)} d^2(e_w))^{\frac{n}{2n-2}}, \fint_{B_{r_{x_i}}(x_i)} d^2(e_w) \right\}$. We conclude that $|B_{r_{x_i}}(x_i)| \le \max\left\{ \frac{4C}{c_0\rho^2}, \left(\frac{4C}{c_0\rho^2}\right)^{\frac{2n-2}{n}} \right\} \int_{B_{r_{x_i}}(x_i)} d^2(e_w)$. Summing over i, from the disjointedness of the $B_{r_{x_i}}(x_i)$ and from (21) we conclude that $|E \setminus A_\infty| \le C\varepsilon^2$, since $\int_\Omega d^2(e_w) \le C\varepsilon^2$ by (11). By Claim 1 this implies (18) in the case $\varepsilon \le \varepsilon_0$. But (18) also holds when $\varepsilon > \varepsilon_0$ by (17) (e.g., by choosing $C = \frac{2C_1}{\varepsilon_0^2}$ in (18); recall that $\varepsilon < 1$). Using (18) we can finally estimate

$$\int_U |e_w|^2 = \int_{U\setminus E} |e_w|^2 + \int_E |e_w|^2 \le C\left[\int_{U\setminus E} d^2(e_w) + |E| + \int_E d^2(e_w) \right] \le C\varepsilon^2.$$

The desired estimate now follows from Korn's inequality. □

The proof of Theorem 3 is completed using a cube decomposition of Ω and applying a weighted Poincaré inequality exactly as in the proof of Theorem 2 in [14]. We have used the following lemma, the proof of which is the same as that of Lemma 2.4 in [14], where one can replace ∇w throughout by an arbitrary matrix-valued $W^{1,2}$-function F.

LEMMA 5. *Let $n \ge 2$, let $\Omega \subset \mathbb{R}^n$ be a bounded and connected Lipschitz domain, and let K_1, K_2 be compact disjoint subsets of $\mathbb{R}^{n\times n}$, $K = K_1 \cup K_2$. Then there is a constant $C = C(K, \Omega)$, such that for any $F \in W^{1,2}(\Omega; \mathbb{R}^{n\times n})$*

$$\begin{aligned} \min_{i=1,2} \int_\Omega \operatorname{dist}^2(F, K_i) &\le C(K,\Omega) \left(\int_\Omega \operatorname{dist}^2(F,K) \int_\Omega |\nabla F|^2 \right)^{\frac{n}{2(n-1)}} \\ &\quad + C(K,\Omega) \int_\Omega \operatorname{dist}^2(F,K). \end{aligned}$$

2.2. Compactness. The following theorem provides the compactness result which complements the Γ-convergence result of Theorem 1. Its proof uses some facts which were derived in [15] (in order to prove a lower scaling bound in a nonlinearly elastic setting) and which in spirit are close to [28]. It is different from the Young

measure arguments used in the literature of singularly perturbed functionals (e.g., [27, 18, 19, 20]).

THEOREM 6. *Let $S \subset \mathbb{R}^2$ be a bounded Lipschitz domain, let $A, B \in \mathbb{R}^{2\times 2}$ be symmetric and such that $(B-A)+Skew$ does not contain rank-one matrices, and set $K = (A+Skew)\cup(B+Skew)$. Let $h_n \to 0$, set $\Omega_{h_n} = S \times I_{h_n}$, and suppose that a sequence $u_n \in W^{1,2}(\Omega_{h_n};\mathbb{R}^3)$ satisfies*

$$\limsup_{n\to\infty} \frac{1}{h_n^2} \int_{\Omega_{h_n}} \operatorname{dist}^2(\nabla u_n, K) < \infty. \tag{23}$$

Set $w_n(x') = \fint_{I_{h_n}} (u_n(x',x_3))'\,dx_3$. Then there exist a subsequence (not relabeled) and affine mappings $f_n : \mathbb{R}^2 \to \mathbb{R}^2$ with sym $\nabla' f_n = 0$, and there is $w_0 \in W^{1,2}(S;\mathbb{R}^2)$ satisfying sym $\nabla' w_0 \in BV(S;\{\bar A, \bar B\})$ such that $w_n + f_n \to w_0$ strongly in $W^{1,2}(S;\mathbb{R}^2)$.

Proof. For $h>0$ we consider a lattice of squares $S_{a,h} = a + (-\frac{h}{2},\frac{h}{2})^2, a \in h\mathbb{Z}^2$, and we let $S_h' = \bigcup_{\{a\in h\mathbb{Z}^2 : S_{a,h}\subset S\}} S_{a,h}$. Now apply Theorem 3 to $u^{(h)}$ (here and in what follows we write $u^{(h)}$ instead of u_n and h instead of h_n to avoid cumbersome notation) restricted to each cube $a + (-\frac{h}{2},\frac{h}{2})^3$ with $a \in h\mathbb{Z}^2$. This yields a piecewise constant map $R^{(h)} : S_h' \to K$ such that

$$\int_{S_{a,h}\times I_h} |\nabla u^{(h)}(x) - R^{(h)}(x')|^2\,dx \le C \int_{S_{a,h}\times I_h} \operatorname{dist}^2(\nabla u^{(h)}(x), K)\,dx. \tag{24}$$

Define the piecewise constant map $L^{(h)} : S_h' \to \{A,B\}$ by setting $L^{(h)}(x) = \text{sym}\ R^{(h)}(x)$. Let $\varepsilon > 0$ be sufficiently small (to be fixed below). We divide the family of squares $\{S_{a,h} : a \in h\mathbb{Z}^2 \text{ and } S_{a,h}\subset S\}$ into three different groups:

$$a \in \mathcal{A}_0^h \text{ if and only if } \int_{S_{a,h}\times I_h} \operatorname{dist}^2(\nabla u^{(h)}(x), K)\,dx \ge \varepsilon h^3. \tag{25}$$

If $a \notin \mathcal{A}_0^h$, then the matrix $L^{(h)}(a) \in \{A,B\}$ is such that $\frac{1}{h^3}\int_{S_{a,h}\times I_h} |\text{sym}\ \nabla u^{(h)} - L^{(h)}(a)|^2 \le C\varepsilon$. This follows from (24) and (25) by the definition of $L^{(h)}(a)$. Now define

$$a \in \mathcal{A}_1^h \text{ if and only if } a \notin \mathcal{A}_0^h \text{ and } L^{(h)}(a) = A,$$

$$a \in \mathcal{A}_2^h \text{ if and only if } a \notin \mathcal{A}_0^h \text{ and } L^{(h)}(a) = B. \tag{26}$$

For ε small enough, each square $S_{a,h}$ belongs to exactly one of these three groups. Thus the sets

$$\Omega_i^h = \operatorname{int}\left(\bigcup_{a\in\mathcal{A}_i^h} \bar S_{a,h}\right), \tag{27}$$

$i = 0, 1, 2$, are disjoint and cover S'_h up to an $\mathcal{H}^2$ null set. As in [15] one can prove that, for ε small enough, the following implication holds:

$$a \in \mathcal{A}_i^h, a' \in \mathcal{A}_j^h, \text{ and } \mathcal{H}^1(\bar{S}_{a',h} \cap \bar{S}_{a,h}) > 0 \implies j \in \{0, i\}, \tag{28}$$

that is, a square of type $\mathcal{A}_1^h$ can only have neighboring squares of type $\mathcal{A}_1^h$ or $\mathcal{A}_0^h$ (never of type $\mathcal{A}_2^h$), and the analogous statement holds with $\mathcal{A}_1^h$ and $\mathcal{A}_2^h$ swapped. But from (23) and from (25) we deduce $\#\mathcal{A}_0^h \le \frac{C}{h}$. Since the side-length of each square $S_{a,h}$ is h, this leads to the estimate $\mathcal{H}^1(\partial\Omega_1^h \setminus \partial S) \le C$; compare [15]. This implies that the characteristic functions $\chi_{\Omega_1^h}$ are bounded in $BV(S)$, whence they have a subsequence converging strongly in $L^1(S)$ and hence (by interpolation) in all $L^p(S)$ with $p < \infty$. Since $\#\mathcal{A}_0^h \le \frac{C}{h}$, we have $|\Omega_0^h| \le Ch$, whence $\chi_{\Omega_0^h} \to 0$ in $L^1(\mathbb{R}^2)$. Hence we also have strong $L^1(S)$-convergence of $\chi_{\Omega_2^h}$. Note that the respective limit functions, which we denote χ_{Ω_1} and χ_{Ω_2}, both belong to $BV(S)$.

On the other hand, $L^{(h)} = A\chi_{\Omega_1^h} + B\chi_{\Omega_2^h} + L^{(h)}\chi_{\Omega_0^h}$. Let us extend $L^{(h)}$ by zero to all of S. By the convergence $\chi_{S'_h} \to 1$ in $L^1(S)$ we obtain that $L^{(h)} \to L$ strongly in $L^2(S; \mathbb{R}^{2\times 2})$, where $L = \chi_{\Omega_1} A + \chi_{\Omega_2} B \in BV(S; \{A, B\})$. By (24), (23), and Jensen's inequality we have $\int_{S'_h} |\operatorname{sym} \nabla' w^{(h)} - \bar{L}^{(h)}|^2 \le Ch$. Using $L^{(h)} = 0$ on $S \setminus S'_h$ and applying Jensen's inequality, we find $\int_{S\setminus S'_h} |\operatorname{sym} \nabla' w^{(h)} - \bar{L}^{(h)}|^2 \le C|S \setminus S'_h| + Ch$. We conclude that $\operatorname{sym} \nabla' w^{(h)} \to \bar{L}$ strongly in $L^2(S; \mathbb{R}^{2\times 2})$. Since the subspace of symmetrized gradients is strongly closed in $L^2(S; \mathbb{R}^{2\times 2})$, there is a $w_0 \in W^{1,2}(S; \mathbb{R}^2)$ such that $\operatorname{sym} \nabla' w_0 = \bar{L} \in BV(S; \{\bar{A}, \bar{B}\})$. An application of Korn's and of Poincaré's inequalities on S for each h yields affine mappings $f^{(h)} : \mathbb{R}^2 \to \mathbb{R}^2$ with $\operatorname{sym} \nabla' f^{(h)} = 0$ (explicitly, e.g., $\nabla' f^{(h)} = \operatorname{skew} \fint_S \nabla'(w^{(h)} - w_0)$) such that $\|w^{(h)} + f^{(h)} - w_0\|^2_{W^{1,2}(S;\mathbb{R}^2)} \le C \int_S |\operatorname{sym} \nabla' w^{(h)} - \operatorname{sym} \nabla' w_0|^2$. This converges to zero because $\operatorname{sym} \nabla' w^{(h)} \to \bar{L}$ in $L^2(S; \mathbb{R}^{2\times 2})$. □

Remark. Let $y^{(h)}(x', x_3) = u^{(h)}(x', hx_3)$ be the rescaled displacements defined on $\Omega = S \times (-\frac{1}{2}, \frac{1}{2})$ and set $\nabla_h y = (\nabla' y | \frac{1}{h} y_{,3})$. Then the proof of Theorem 6 shows that $\operatorname{sym} \nabla_h y^{(h)} \to L$ strongly in $L^2(\Omega; \mathbb{R}^{3\times 3})$. Thus $\nabla'(y^{(h)})' \to \bar{L}$ and $y_{,3} \to 0$ strongly in L^2. If, using a scaling analogous to that in [17, section 1.3], we set $(\bar{y}^{(h)})' = (y^{(h)})'$ and $(\bar{y}^{(h)})_3 = h(y^{(h)})_3$, then Korn's inequality on Ω implies that there are affine mappings $F^{(h)} : \mathbb{R}^3 \to \mathbb{R}^3$ with $\operatorname{sym} \nabla F^{(h)} = 0$ and such that $\bar{y}^{(h)} + F^{(h)} \to F_L$ strongly in $W^{1,2}(\Omega; \mathbb{R}^3)$, where $F_L(x) = \left(\begin{smallmatrix} \bar{L}x' \\ 0 \end{smallmatrix}\right)$.

In [19, Proposition 2.2], the following characterization is provided for functions whose symmetrized gradient has bounded variation and is supported on two incompatible matrices $\bar{A}, \bar{B}$. (For earlier results in this direction, compare [23].)

PROPOSITION 7. *Let $S \subset \mathbb{R}^2$ be a bounded Lipschitz domain. Let $\bar{A}, \bar{B}$ satisfy* (iii)–(iv) *from the beginning of section* 2, *let ν_1, ν_2 be linearly independent solutions to $\bar{A} - \bar{B} + t(e_1 \otimes e_2 - e_2 \otimes e_1) = a \otimes \nu_i$, where $a \in \mathbb{R}^3$ and $t \in \mathbb{R}$, and let $w \in W^{1,2}(S; \mathbb{R}^2)$ satisfy sym $\nabla' w \in BV(S; \{\bar{A}, \bar{B}\})$. Then the jump set J of sym $\nabla' w$ consists of countably many disjoint segments whose endpoints belong to ∂S and which have normal directions ν_1 or ν_2. In addition, $\nabla' w$ is constant on each connected component of $S \setminus J$.*

3. Lower bound. In this section we prove part (i) of Theorem 1. From now on we assume that $A = 0$ and B is as in (7). This choice allows exactly two different directions for the interface normal; the directions are orthogonal to each other: Setting $T_1 = e_2 \otimes e_1 - e_1 \otimes e_2$ and $T_2 = e_1 \otimes e_2 - e_2 \otimes e_1$, we have $\bar{B} + T_1 = 2e_2 \otimes e_1$, giving the normal $\nu_1 = e_1$, and $\bar{B} + T_2 = 2e_1 \otimes e_2$, giving the normal $\nu_2 = e_2$. Notice that,

if $\bar{T} \in \mathbb{R}^{2\times 2} \setminus \{T_1, T_2\}$ with $\operatorname{sym} \bar{T} = 0$, then $\det(\bar{B} + \bar{T}) \neq 0$. Define the piecewise constant mappings $F_i^{\pm} : \mathbb{R}^2 \to \mathbb{R}^{2\times 2}$ by

$$F_i^{\pm}(x') = \begin{cases} 0 & \text{for } \pm x' \cdot \nu_i < 0, \\ \bar{B} + T_i & \text{otherwise,} \end{cases}$$

and set $w_i^{\pm}(x') = F_i^{\pm}(x')x'$; note that the jump set of $\nabla' w_i^{\pm}$ agrees with $\{\nu_i\}^{\perp}$.

DEFINITION 8. *Let $\sigma \in \{-,+\}$, $i \in \{1,2\}$, let J_1, J_2 be open intervals, and set $S = J_1 \times J_2$.*

(i) *We set $w_{i,S}^{\sigma}(x') = w_i^{\sigma}(x' - \xi e_i)$, where $\xi = \frac{1}{2}(\sup J_i - \inf J_i)$. The set $\mathcal{J}_{i,S} = S \cap (\xi\nu_i + \{\nu_i\}^{\perp})$ is called the interface of $w_{i,S}^{\sigma}$.*

(ii) *A pair of sequences (u_n, h_n) is called (σ, i)-admissible on S, provided that $h_n \in \mathbb{R}_+$, $h_n \to 0$, $u_n \in W^{1,2}(S \times I_{h_n}; \mathbb{R}^3)$, and $\fint_{I_{h_n}} u_n' \, dx_3 \to w_{i,S}^{\sigma}$ in $L^2(S; \mathbb{R}^2)$.*

(iii) *If $J_1 = \emptyset$ or $J_2 = \emptyset$, then we set $\mathcal{F}_i^{\sigma}(S) = 0$. Otherwise, we define*

$$\mathcal{F}_i^{\sigma}(S) = \inf \Big\{ \liminf_{n\to\infty} I^{h_n}(u_n; S) : \text{There exist } (h_n) \text{ and } (u_n) \text{ such that } (u_n, h_n) \text{ is } (\sigma, i)\text{-admissible on } S \Big\}.$$

We define

$$k(\nu_i) = \mathcal{F}_i^{+}\bigg(\Big(-\frac{1}{2}, \frac{1}{2}\Big) \times \Big(-\frac{1}{2}, \frac{1}{2}\Big)\bigg). \tag{29}$$

(iii) *A pair of sequences (u_n, h_n) is called a (σ, i)-recovery sequence on S, provided that (u_n, h_n) is (σ, i)-admissible on S and $\lim_{n\to\infty} I^{h_n}(u_n; S) = \mathcal{F}_i^{\sigma}(S)$.*

Remarks. (i) Our definition of $\mathcal{F}_i^{\sigma}$ differs slightly from the usual one. If one sets $\tilde{\mathcal{F}}_1^{\sigma}(J; \varepsilon) = \mathcal{F}_1^{\sigma}((-\varepsilon, \varepsilon) \times J)$ and $\tilde{\mathcal{F}}_2^{\sigma}(J; \varepsilon) = \mathcal{F}_2^{\sigma}(J \times (-\varepsilon, \varepsilon))$, then the $\tilde{\mathcal{F}}_i^{\sigma}$ correspond to the $\mathcal{F}_i^{\sigma}$ as defined, e.g., in [18, formula (4.2)].

(ii) In Lemma 13 we will show that for any *fixed* sequence $h_n \to 0$ there exist u_n such that (u_n, h_n) is a (σ, i)-recovery sequence.

(iii) When it is clear from the context which sequence $h_n \to 0$ is meant, then we will often just say that u_n is a (σ, i)-recovery sequence. Also, we will drop the prefix (σ, i) when it is clear from the context.

LEMMA 9. *Let J_1, J_2 be open intervals, set $S = J_1 \times J_2$, and let $i \in \{1,2\}$. Then $\mathcal{F}_i^{+}(S) = \mathcal{F}_i^{-}(S) = k(\nu_i)\ |\pi_i^{\perp}(S)|$. (Recall that $\pi_i^{\perp}$ denotes the orthogonal projection onto $\{e_i\}^{\perp} = \{\nu_i\}^{\perp}$.)*

Proof. Similar to [18, Lemma 4.3] or [19, Lemma 3.2], one can prove the lemma by showing the following facts. Let $J_1' \subset J_1$, $J_2' \subset J_2$ be open intervals and set $S' = J_1' \times J_2'$. Then

(i) $\mathcal{F}_i^{+}(S) = \mathcal{F}_i^{-}(S) =: \mathcal{F}_i(S)$.

(ii) Behavior under homotheties. $\mathcal{F}_i(x' + \lambda S) = \lambda \mathcal{F}_i(S)$ for all $x' \in \mathbb{R}^2$ and all $\lambda > 0$.

(iii) Monotonicity. $\mathcal{F}_i(S') \leq \mathcal{F}_i(S)$.

(iv) Concentration. $\mathcal{F}_i(S') = \mathcal{F}_i(S)$ if $|\pi_i^{\perp}(S')| = |\pi_i^{\perp}(S)|$. ∎

LEMMA 10. *Let $i \in \{1,2\}$, $\sigma \in \{+,-\}$, let J_1, J_2 be open intervals, let $\lambda > 0$, $y \in \mathbb{R}^2 \times \{0\}$, and let (v_n, h_n) be a (σ, i)-recovery sequence on $S = J_1 \times J_2$. Set $\hat{v}_n(x) = \lambda v_n(\frac{x-y}{\lambda})$ for all $n \in \mathbb{N}$. Then $(\hat{v}_n, \lambda h_n)$ is a (σ, i)-recovery sequence on $y' + \lambda S$.*

Proof. Since from Lemma 9 we have $\mathcal{F}_i^{\sigma}(y' + S) = \mathcal{F}_i^{\sigma}(S)$, we may assume without loss of generality that $y = 0$ and that S is centered about the origin. Since (v_n, h_n)

is (σ,i)-admissible on S, $(\hat{v}_n,\lambda h_n)$ is also (σ,i)-admissible on λS. By Lemma 9 and a change of variables we have $\limsup_{n\to\infty} I^{\lambda h_n}(\hat{v}_n;\lambda S)=\limsup_{n\to\infty}\lambda I^{h_n}(v_n;S)=\lambda\mathcal{F}_i^\sigma(S)=\mathcal{F}_i^\sigma(\lambda S)$. □

LEMMA 11. *Let $i\in\{1,2\}$, $\sigma\in\{+,-\}$, let J_j, J_j' be open intervals with $J_j'\subset J_j$, $j=1,2$, and set $S=J_1\times J_2$, $S'=J_1'\times J_2'$. Let (v_n,h_n) be a (σ,i)-recovery sequence on S. Then the following hold:*

(i) *If $\overline{S'}\cap\mathcal{J}_{i,S}=\emptyset$, then $\lim_{n\to\infty} I^{h_n}(v_n;S')=0$.*

(ii) *If $S'\cap\mathcal{J}_{i,S}\neq\emptyset$, then $\lim_{n\to\infty} I^{h_n}(v_n;S')=\mathcal{F}_i^\sigma(S')$.*

Proof. By translation invariance we may assume without loss of generality that S is centered about the origin, so $\mathcal{J}_{i,S}=S\cap\{\nu_i\}^\perp$. Statement (i) is an immediate consequence of Lemma 9. In fact, let $\varepsilon>0$ be so small that $S_\varepsilon=\{x'\in S:|x'\cdot\nu_i|<\varepsilon\}$ satisfies $S_\varepsilon\cap S'=\emptyset$. Then $\mathcal{F}_i^\sigma(S_\varepsilon)=\mathcal{F}_i^\sigma(S)$ by Lemma 9 because $|\pi_i^\perp(S_\varepsilon)|=|\pi_i^\perp(S)|$. Hence $\lim_{n\to\infty} I^{h_n}(v_n;S\setminus S_\varepsilon)=0$.

To prove statement (ii) notice that by Lemma 11(i) we may assume without loss of generality that $\pi_i(S')=\pi_i(S)$ (recall that π_i denotes the orthogonal projection onto the subspace spanned by $e_i=\nu_i$). In other words, $J_i'=J_i$, so S' is a stripe of width $|\pi_i^\perp(S')|$ perpendicular to the interface $\mathcal{J}_{i,S}$. Assume that (ii) is false, and so, since v_n is admissible on S', we have $\limsup_{n\to\infty} I^{h_n}(v_n;S')>\mathcal{F}_i^\sigma(S')$. If $S'=S$, then this would contradict the fact that v_n is a recovery sequence on S. Otherwise, denote by S_1, S_2 the two connected components of $S\setminus S'$. (If $S\setminus S'$ consists of only one connected component, then we call it S_1 and set $S_2=\emptyset$.) After passing to subsequences (not relabeled) we may assume that $I^{h_n}(v_n;S')$ and $I^{h_n}(v_n;S_j)$, $j=1,2$, converge. Hence we obtain the contradiction

$$\begin{aligned}\mathcal{F}_i^\sigma(S)&=\mathcal{F}_i^\sigma(S')+\mathcal{F}_i^\sigma(S_1)+\mathcal{F}_i^\sigma(S_2)\\&<\lim_{n\to\infty} I^{h_n}(v_n;S')+\lim_{n\to\infty} I^{h_n}(v_n;S_1)+\lim_{n\to\infty} I^{h_n}(v_n;S_2)\\&=\lim_{n\to\infty} I^{h_n}(v_n;S)=\mathcal{F}_i^\sigma(S).\end{aligned}$$

The first equality follows from Lemma 9; the strict inequality holds because $(v_n|_{S_j},h_n)$ are admissible on S_j and because by assumption we have $\lim I^{h_n}(v_n;S')>\mathcal{F}_i(S')$. The last equality holds because (v_n,h_n) is a recovery sequence on S. □

Now we define the set of admissible limiting functions as

$$\mathcal{A}(S)=\Big\{w\in W^{1,2}(S;\mathbb{R}^2):\operatorname{sym}\nabla' w\in BV(S;\{0,\bar{B}\})\Big\}\tag{30}$$

and the limiting functional

$$I^0(w;S)=\begin{cases}\int_J k(\nu(x))d\mathcal{H}^1(x) & \text{if } w\in\mathcal{A}(S),\\ +\infty & \text{otherwise.}\end{cases}\tag{31}$$

Here J denotes the jump set of $\operatorname{sym}\nabla' w$, also called the phase interface, and ν denotes the normal (the sign does not matter), which up to a sign can assume only the values $\nu_1=e_1$ and $\nu_2=e_2$.

THEOREM 12. *Let $S\subset\mathbb{R}^2$ be a bounded Lipschitz domain and $w\in L^2(S;\mathbb{R}^2)$. Then, for all $h_n\to 0$ and all $u_n\in L^2(S\times I_{h_n};\mathbb{R}^3)$ satisfying $\fint_{I_{h_n}} u_n' dx_3\to w$ in $L^2(S;\mathbb{R}^2)$, one has $\liminf_{n\to\infty} I^{h_n}(u_n;S)\geq I^0(w;S)$.*

Proof. If $\liminf_{n\to\infty} I^{h_n}(u_n;S)=\infty$, then there is nothing to prove. Otherwise, by passing to a subsequence (not relabeled) we may assume that the sequence

$I^{h_n}(u_n;S)$ converges, so, in particular, $\limsup_{n\to\infty} I^{h_n}(u_n;S)<\infty$. After passing to a further subsequence, Theorem 6 implies that there is a sequence of affine mappings $f_n:\mathbb{R}^2\to\mathbb{R}^2$ with $\operatorname{sym}\nabla f_n=0$ such that $w_n+f_n\to w_0$ in $W^{1,2}(S;\mathbb{R}^2)$ for some $w_0\in\mathcal{A}(S)$, where we have set $w_n=\fint_{I_{h_n}} u_n'\,dx_3$. Since $w_n\to w$ in $L^2(S;\mathbb{R}^2)$, we deduce that f_n converges in $L^2(S;\mathbb{R}^2)$, whence there is $\bar T\in\mathbb{R}^{2\times 2}$ with $\operatorname{sym}\bar T=0$ and a vector $c\in\mathbb{R}^2$ such that $f_n(x')\to c+\bar T x'$ for all $x'\in S$. Hence $w=w_0-\bar T x'-c$, and, in particular, we have $w\in\mathcal{A}(S)$. By the strong $W^{1,2}$-convergence of both w_n+f_n and f_n we have $w_n\to w$ in $W^{1,2}(S;\mathbb{R}^2)$. By Proposition 7 the jump set of $\operatorname{sym}\nabla' w$ consists of a countable union of disjoint segments $\mathcal{J}_k$ with normal ν_1 or ν_2. The rest of the proof is standard: One covers each $\mathcal{J}_k$ with a box, applies Lemma 9 to each box separately, and uses the minimality of $\mathcal{F}_i^{\pm}$ (see, e.g., the proof of Proposition 3.1 in [19] for the details). □

4. Upper bound. In this section we will show that for any admissible limit function w and for any given sequence $h_n\to 0$ one can find a sequence (v_n,h_n) such that $\fint_{I_{h_n}} v_n'\,dx_3\to w$ strongly in $W^{1,2}(S;\mathbb{R}^2)$ and $I^{h_n}(v_n;S)\to I^0(w;S)$. We will first show that given (σ,i) and *any* sequence $h_n\to 0$ one can find a sequence (v_n) such that (v_n,h_n) is a (σ,i)-recovery sequence on S. A key difference from the proof of the analogous Proposition 5.5 in [19] is that we do not rely on the existence of special recovery sequences which are affine away from the interface but work directly with an arbitrary recovery sequence.

LEMMA 13. *Let J_1,J_2 be open intervals and set $S=J_1\times J_2$. Let $\sigma\in\{+,-\}$, $i\in\{1,2\}$, and $H_n\to 0$ be given. Then we have*

$$\mathcal{F}_i^{\sigma}(S)=\inf\Big\{\liminf_{n\to\infty} I^{H_n}(u_n;S) : \text{there is } (u_n) \text{ such that } (u_n,H_n) \text{ is } (\sigma,i)\text{-admissible on } S\Big\}.$$

Proof. Clearly we must prove only the "$\geq$"-inequality, and by Lemmas 10 and 11 we may assume without loss of generality that $S=(-\frac12,\frac12)^2$. In fact, suppose Lemma 13 is shown for this particular case. Now let $H_n\to 0$ and an arbitrary S be given and assume without loss of generality (by translation invariance) that S is centered about the origin. Then there is $\lambda>0$ such that $\lambda S\subset(-\frac12,\frac12)^2$. By the special case of Lemma 13 there is a recovery sequence $(v_n,\lambda H_n)$ on $(-\frac12,\frac12)^2$. By Lemma 11(ii) we have that $(v_n|_{\lambda S},\lambda H_n)$ is a recovery sequence on λS. By Lemma 10, setting $\hat v_n(x)=\frac{1}{\lambda}v_n(\lambda x)$, we conclude that $(\hat v_n,H_n)$ is a recovery sequence on S.

So suppose that $S=(-\frac12,\frac12)^2$ and let $H_n\to 0$ be given. Let us restrict our attention to the case $\sigma=+$ and $i=1$, so the interface normal is $\nu_1=e_1$, the phase "0" is used on the left, $\{\nabla' w_1^+=0\}=S\cap\{x_1<0\}$, and $\{\nabla' w_1^+=\bar B+T_1\}=S\cap\{x_1>0\}$ up to a null set; the other cases are similar. Note that the infimum in the definition of $\mathcal{F}_i^{\sigma}(S)=k(\nu_i)$ is attained; i.e., there is a sequence $h_n\to 0$ and $v_n\in W^{1,2}(S\times I_{h_n};\mathbb{R}^3)$ such that $\fint_{I_{h_n}} v_n'dx_3\to w_1^+$ in $L^2(S;\mathbb{R}^2)$ and $\lim_{n\to\infty} I^{h_n}(v_n;S)=k(\nu_1)$. Since after passing to subsequences this equality remains valid, we may assume without loss of generality that $h_n\ll H_n$, i.e., $\alpha_n=\frac{H_n}{h_n}\to\infty$.

Set $y_1^{(n)}=\frac{1}{2\alpha_n}-\frac12$ and $y_{m+1}^{(n)}=y_m^{(n)}+\frac{1}{\alpha_n}$, $m=1,\dots,[\alpha_n]-1$, and let $S_m^{(n)}=(-\frac12,\frac12)\times(y_m^{(n)}-\frac{1}{2\alpha_n},y_m^{(n)}+\frac{1}{2\alpha_n})$. We recall (25)–(28) from the proof of Theorem 6 and apply them to v_n instead of $u^{(h)}$. Define $S_{m,1}^{(n)}=S_m^{(n)}\cap\Omega_1^{h_n}\cap\{\nabla' w_1^+=0\}$ and $S_{m,2}^{(n)}=S_m^{(n)}\cap\Omega_2^{h_n}\cap\{\nabla' w_1^+=\bar B+T_1\}$. It follows from the proof of Theorem 6 applied to v_n that $\chi_{\Omega_1^{h_n}}\to\chi_{\{\nabla' w_1^+=0\}\cap S}$ and $\chi_{\Omega_2^{h_n}}\to\chi_{\{\nabla' w_1^+=\bar B+T_1\}\cap S}$ strongly in $L^1(\mathbb{R}^2)$

(notice that, by uniqueness of the limits, the full sequence indeed converges). Now denote by $G_n = \{m = 1, \dots, [\alpha_n] : |S^{(n)}_{m,1}| > \frac{1}{3\alpha_n}$ and $|S^{(n)}_{m,2}| > \frac{1}{3\alpha_n}\}$ the index set of "good" stripes, and for $i = 1, 2$ set $M_{n,i} = \{m = 1, \dots, [\alpha_n] : |S^{(n)}_{m,i}| \le \frac{1}{3\alpha_n}\}$. We claim that

$$\frac{\#G_n}{\alpha_n} \to 1 \text{ as } n \to \infty. \tag{32}$$

Indeed, if (32) were false, then by definition of G_n there would exist $i \in \{1, 2\}$, $\gamma > 0$, and a subsequence (not relabeled) such that $\frac{\#M_{n,i}}{\alpha_n} \in (\frac{5}{6}\gamma, \gamma)$ for all n. For definiteness suppose that $i = 1$; the case $i = 2$ is similar. But for all n we have $\Omega_1^{h_n} \cap \{\nabla' w_1^+ = 0\} \subset (S \setminus \bigcup_{m=1}^{[\alpha_n]} S^{(n)}_m) \cup \bigcup_{m \notin M_{n,1}} S^{(n)}_{m,1} \cup \bigcup_{m \in M_{n,1}} S^{(n)}_{m,1}$. Taking measures on both sides, we find (notice that by definition $|S^{(n)}_{m,i}| \le \frac{1}{2\alpha_n}$)

$$\begin{aligned} |\Omega_1^{h_n} \cap \{\nabla' w_1^+ = 0\}| &\le 1 - \frac{[\alpha_n]}{\alpha_n} + \frac{1}{2\alpha_n}([\alpha_n] - \#M_{n,1}) + \frac{1}{3\alpha_n}\#M_{n,1} \\ &\le 1 - \frac{[\alpha_n]}{\alpha_n} + \frac{1}{2}\left(1 - \frac{5}{6}\gamma\right) + \frac{\gamma}{3}. \end{aligned}$$

As $n \to \infty$, the left-hand side converges to $\frac{1}{2}$ and the right-hand side converges to $\frac{1}{2} - \frac{1}{12}\gamma$. This contradiction proves (32).

From (32) one deduces that

$$\limsup_{n\to\infty} \big(\alpha_n \min_{m \in G_n} I^{h_n}(v_n; S^{(n)}_m)\big) \le \limsup_{n\to\infty} I^{h_n}(v_n; S). \tag{33}$$

In fact, if (33) were false, then (after passing to an unrelabeled subsequence) there would exist $\gamma > 0$ such that $I^{h_n}(v_n; S^{(n)}_m) - \frac{1}{\alpha_n} I^{h_n}(v_n; S) \ge \frac{\gamma}{\alpha_n}$ for all n and for all $m \in G_n$. Summing over $m \in G_n$ we would find $(1 - \frac{\#G_n}{\alpha_n}) I^{h_n}(v_n; S) \ge I^{h_n}(v_n; \bigcup_{m \in G_n} S^{(n)}_m) - \frac{\#G_n}{\alpha_n} I^{h_n}(v_n; S) \ge \frac{\#G_n}{\alpha_n}\gamma$. By (32) the left-hand side converges to zero as $n \to \infty$, while the right-hand side converges to γ, which is a contradiction proving (33).

Now choose $m_n \in G_n$ such that $I^{h_n}(v_n; S^{(n)}_{m_n}) = \min_{m \in G_n} I^{h_n}(v_n; S^{(n)}_m)$, and set $\hat{y}_n = y^{(n)}_{m_n}$. Let $r^{(n)} = (-\frac{l}{2\alpha_n}, \frac{l}{2\alpha_n}) \times (\hat{y}_n - \frac{1}{2\alpha_n}, \hat{y}_n + \frac{1}{2\alpha_n})$, where $l = 1 + \frac{k(\nu_1)}{k(\nu_2)}$. Consider the functions $g_n(x_1) = |\Omega_1^{h_n} \cap \big((x_1, 0) + r^{(n)}\big)|$. Since $m_n \in G_n$, there is $x_1 \le 0$ such that $g_n(x_1) > \frac{|r^{(n)}|}{2}$. The existence of such x_1 can be seen, e.g., by the following argument (another argument uses Fubini's theorem): Set $x_1^k = \frac{l}{2\alpha_n} - \frac{1}{2} + \frac{kl}{\alpha_n}$, $k = 0, \dots, N_n$, where $N_n = [\frac{\alpha_n}{2l} - \frac{1}{2}]$. Then $x_1^k \le 0$ for all $k = 0, \dots, N_n$ and

$$S^{(n)}_{m_n} \cap \{\nabla' w_1^+ = 0\} \subset \hat{r}^{(n)} \cup \bigcup_{k=0}^{N_n} \big((x_1^k, 0) + r^{(n)}\big), \tag{34}$$

where $\hat{r}^{(n)}$ is a rectangle with $|\hat{r}^{(n)}| \le |r^{(n)}|$. If $g_n(x_1^k) \le \frac{|r^{(n)}|}{2}$ for all $k = 0, \dots, N_n$, then $|\Omega_1^{h_n} \cap \big((x_1^k, 0) + r^{(n)}\big)| \le \frac{|r^{(n)}|}{2}$ for all k. Intersecting (34) with $\Omega_1^{h_n}$, taking measures, and multiplying by α_n, we find $\alpha_n |S^{(n)}_{m_n,1}| \le \alpha_n |\hat{r}^{(n)}| + \alpha_n (N_n + 1)\frac{|r^{(n)}|}{2} \le \frac{C}{\alpha_n} + \frac{1}{4}$ since $|r^{(n)}| = \frac{l}{\alpha_n^2}$. As $n \to \infty$, the right-hand side converges to $\frac{1}{4}$, while the left-hand side is greater than $\frac{1}{3}$ because $m_n \in G_n$, which is a contradiction. Similarly, one proves existence of $x_1 \ge 0$ such that $g_n(x_1) < \frac{|r^{(n)}|}{2}$.

Since g_n is continuous, we conclude that there is some $\hat{x}_n$ with $g_n(\hat{x}_n) = \frac{|r^{(n)}|}{2}$. By (33) and the choice of m_n the rectangle $\hat{S}_n = (\hat{x}_n - \frac{l}{2\alpha_n}, \hat{x}_n + \frac{l}{2\alpha_n}) \times (\hat{y}_n - \frac{1}{2\alpha_n}, \hat{y}_n + \frac{1}{2\alpha_n})$ satisfies $\limsup \alpha_n I^{h_n}(v_n; \hat{S}_n) \leq k(\nu_1)$. Recalling (25) we set $\hat{\mathcal{A}}_0^{(n)} = \{a \in \mathcal{A}_0^{h_n} : a + (-\frac{h_n}{2}, \frac{h_n}{2})^2 \subset \hat{S}_n\}$. Then by (25) we have $\#\hat{\mathcal{A}}_0^{(n)} \leq \frac{C}{h_n^3} \int_{\hat{S}_n} W(\nabla v_n) \leq \frac{C}{h_n \alpha_n}$. Hence $|\Omega_0^{h_n} \cap \hat{S}_n| \leq |\bigcup_{a \in \hat{\mathcal{A}}_0^{(n)}} (a + (-\frac{h_n}{2}, \frac{h_n}{2})^2)| + h_n \cdot \mathcal{H}^1(\partial \hat{S}_n) \leq \frac{C h_n}{\alpha_n}$. But $\frac{C h_n}{\alpha_n} \ll \frac{l}{\alpha_n^2} = |\hat{S}_n|$ because $H_n \to 0$. Hence $\frac{|\Omega_0^{h_n} \cap \hat{S}_n|}{|\hat{S}_n|} \to 0$, so we conclude that

$$\frac{|\Omega_i^{h_n} \cap \hat{S}_n|}{|\hat{S}_n|} \to \frac{1}{2} \text{ for } i = 1, 2 \text{ as } n \to \infty. \tag{35}$$

In fact, we have $|\hat{S}_n \cap \Omega_1^{h_n}| = \frac{|\hat{S}_n|}{2}$ by the choice of $\hat{x}_n$. Now set $V_n(x) = \alpha_n v_n\big(\frac{x}{\alpha_n} + (\hat{x}_n, \hat{y}_n)\big)$ and $S' = (-l/2, l/2) \times (-\frac{1}{2}, \frac{1}{2})$. Then $V_n \in W^{1,2}\big(S' \times (-\frac{H_n}{2}, \frac{H_n}{2}); \mathbb{R}^3\big)$ and $\limsup I^{H_n}(V_n; S') \leq k(\nu_1)$. Set $W_n = \fint_{I_{H_n}} V_n' dx_3$. After possibly reflecting V_n about $\{\nu_1\}^\perp$ we may assume that

$$\begin{aligned} &\left|\left\{x \in S' : x_1 < 0, \operatorname{sym} \nabla' W_n(x) \leq \frac{|B|}{4}\right\}\right| \\ &\qquad \leq \left|\left\{x \in S' : x_1 > 0, \operatorname{sym} \nabla' W_n(x) \leq \frac{|B|}{4}\right\}\right| \end{aligned} \tag{36}$$

for all n. By Theorem 6, for every subsequence there is a further subsequence, labeled with an index m, and there are affine mappings $F_m : \mathbb{R}^2 \to \mathbb{R}^2$ with $\operatorname{sym} \nabla F_m = 0$ such that $W_m + F_m \to W_0$ strongly in $W^{1,2}(S'; \mathbb{R}^2)$, where $W_0 \in \mathcal{A}(S')$. Moreover, we may assume that $\lim_{m \to \infty} I^{H_m}(V_m; S') = k(\nu_1)$. Theorem 12 implies that $I^0(W_0; S') \leq k(\nu_1)$. Since S' is a rectangle with sides parallel to e_1 and e_2, Proposition 7 implies that W_0 has either only interfaces with normal $\nu_1 = e_1$ or only interfaces with normal $\nu_2 = e_2$. If it had an interface of the latter type, we would obtain the contradiction $I^0(W_0; S') \geq lk(\nu_2) > k(\nu_1)$, by the choice of l. Thus W_0 has only interfaces with normal ν_1, and since $\limsup I^{H_n}(V_n; S') \leq k(\nu_1)$, there can be at most one such interface. On the other hand, by (35) we have $|\{\operatorname{sym} \nabla' W_0 = 0\}| = |\{\operatorname{sym} \nabla' W_0 = \bar{B}\}| = \frac{|S'|}{2}$, so there must be at least one interface. Hence W_0 has exactly one interface, it has normal ν_1, and by (35) it lies in $\{\nu_1\}^\perp$. By (36) we have $(-\frac{l}{2}, 0) \times (-\frac{1}{2}, \frac{1}{2}) = \{\operatorname{sym} \nabla' W_0 = 0\}$ up to a null set, so the mapping $w_1^+ - W_0$ is affine with $\operatorname{sym} \nabla'(w_1^+ - W_0) = 0$. We conclude that $\operatorname{sym} \nabla' W_m \to \operatorname{sym} \nabla' w_1^+$ strongly in $L^2(S'; \mathbb{R}^{2\times 2})$. Since the same limit is obtained for all subsequences, we conclude that the full sequence satisfies $\operatorname{sym} \nabla' W_n \to \operatorname{sym} \nabla' w_1^+$. By Korn's inequality on S', there exist affine mappings $\tilde{F}_n : \mathbb{R}^2 \to \mathbb{R}^2$ with $\operatorname{sym} \nabla' \tilde{F}_n = 0$ such that $W_n + \tilde{F}_n \to w_1^+$ strongly in $W^{1,2}(S'; \mathbb{R}^2)$. Denote by u_n the restriction of $x \mapsto V_n(x) + \binom{\tilde{F}_n(x')}{0}$ to S. This sequence satisfies $\fint_{I_{H_n}} u_n' \, dx_3 \to w_1^+$ in $W^{1,2}(S; \mathbb{R}^2)$ and $\limsup I^{H_n}(u_n; S) = k(\nu_1)$. □

Notation. From now on we will drop the index n when dealing with sequences $h_n \to 0$ because, in view of Lemma 13, there exists a recovery sequence (v_n, h_n) for a particular sequence $h_n \to 0$ if and only if there exists one for *every* sequence $h_n \to 0$. We say that there exists a (σ, i)-recovery sequence $u^{(h)}$ on S if for all (h_n) there exist (u_n) such that (u_n, h_n) is a (σ, i)-recovery sequence on S.

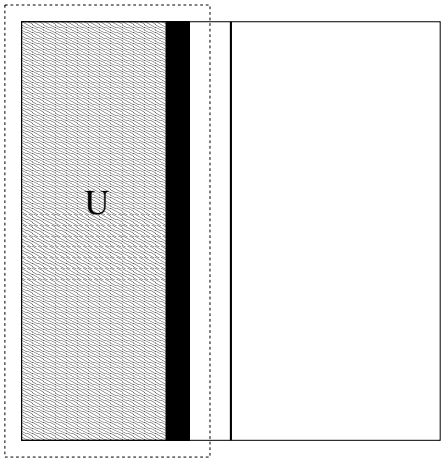


FIG. 1. *Proof of Lemma* 14*: The black region represents the interpolation layer. The dashed rectangle denotes the boundary of* V.

In a first modification step we will now change the recovery sequence furnished by Definition 8 and Lemma 13 in such a way that its vertical averages become smooth away from the interface. Lemma 14 provides a link between (1) and (2).

LEMMA 14. *Let* J_1, J_2 *be open intervals, set* $S = J_1 \times J_2$, *and let* $\sigma \in \{+,-\}$, $i \in \{1,2\}$. *Let* $\varepsilon > 0$ *and set* $Q = S \cap B_\varepsilon(\mathcal{J}_{i,S})$. *Then there exists a* (σ,i)*-recovery sequence* $u^{(h)} \in W^{1,2}(S \times I_h;\mathbb{R}^3)$ *on* S *such that* $w^{(h)}(x') = \fint_{I_h} (u^{(h)})'(x)\ dx_3$ *and* $\tau^{(h)}(x') = \fint_{I_h} u_3^{(h)}(x)\ dx_3$ *satisfy* $w^{(h)} \in C^\infty(S \setminus \bar{Q};\mathbb{R}^2)$ *and* $\tau^{(h)} \in C^\infty(S \setminus \bar{Q})$. *Moreover,*

$$\lim_{h\to 0} \left(I^h_{2D}(w^{(h)}; S \setminus Q) + h \int_{S\setminus Q} |\nabla' {}^2 \tau^{(h)}(x')|^2\ dx' \right) = 0. \tag{37}$$

Proof. Arguing as at the beginning of the proof of Lemma 13, we may assume without loss of generality that $S = (-\frac{1}{2},\frac{1}{2})^2$. Moreover, we will prove the statement for $i = 1$, $\sigma = +$ only, the other cases being analogous. Recall that $\{\nabla' w_1^+ = 0\} = \{x \in \mathbb{R}^2 : x_1 < 0\}$ up to a null set. Let $v^{(h)} \in W^{1,2}((2S) \times I_h;\mathbb{R}^3)$ be a $(+,1)$-recovery sequence on $2S$, so by Lemma 11(ii) we have $\lim_{h\to 0} I^h(v^{(h)}; S) = k(\nu_1)$. Fix $a > 0$ satisfying $(-9a, 9a) \times (-\frac{1}{2},\frac{1}{2}) \subset Q$, set $V = (-\frac{1+a}{2}, -\frac{a}{2}) \times (-\frac{1+a}{2}, \frac{1+a}{2})$, and let $U = (-\frac{1}{2}, -a) \times (-\frac{1}{2},\frac{1}{2})$, so $\bar{U} \subset V \subset \{\nabla' w_1^+ = 0\} \cap (2S)$. The situation is depicted in Figure 1. To obtain mappings defined on the full plate thickness, we mollify slicewise in horizontal planes: Let ψ be a standard mollifier supported on $(-\frac{1}{2},\frac{1}{2})^2$, and set $\psi_h(x') = \frac{1}{h^2}\psi(\frac{x'}{h})$. Set

$$\tilde{v}^{(h)}(x) = \big(\psi_h * v^{(h)}(\cdot, x_3)\big)(x') = \fint_{I_h^2} \psi\left(\frac{y'}{h}\right) v^{(h)}(x' - y', x_3)\ dy',$$

which for h small enough is well defined on $S \times I_h$ (recall that $v^{(h)}$ is defined on $(2S) \times I_h$). Recall the definition of S'_h and of $R^{(h)} : S'_h \to \mathrm{Skew} \cup (B + \mathrm{Skew})$ introduced before (24) in the proof of Theorem 6. Adopting the notation introduced there, we have $S \subset (2S)'_h$ for h small enough, so $R^{(h)}$ is defined everywhere on S.

Since $\nabla\tilde{v}^{(h)}(x) = \big(\psi_h * \nabla v^{(h)}(\cdot, x_3)\big)(x')$, we can estimate

$$\begin{aligned}
&\int_{U\times I_h} |\nabla\tilde{v}^{(h)}(x) - R^{(h)}(x')|^2 dx \\
&\quad\le \frac{C}{h^2}\int_{\text{spt }\psi_h} dy' \int_{U\times I_h} \Big(|\nabla v^{(h)}(x'-y', x_3) - R^{(h)}(x'-y')|^2 \\
&\qquad + |R^{(h)}(x'-y') - R^{(h)}(x')|^2\Big) dx \\
&\quad\le C\int_{V\times I_h} W(\nabla v^{(h)}(x))\, dx. \qquad (38)
\end{aligned}$$

In the first step we have applied Jensen's inequality and have added and subtracted $R^{(h)}(x'-y')$. In the last step we used that spt $\psi_h \subset (-\frac{h}{2}, \frac{h}{2})^2$ and applied the estimate

$$\int_U |R^{(h)}(x'+\zeta) - R^{(h)}(x')|^2\, dx' \le C\int_{V\times I_h} W(\nabla v^{(h)}(x))\, dx, \qquad (39)$$

which holds for all $\zeta \in \mathbb{R}^2$ with $|\zeta_1|, |\zeta_2| \le h$. (The estimate (39) can be derived by arguments similar to the first part of the proof of Theorem 4.1 in [28], with our Theorem 3 replacing their Theorem 3.1.) From (38) we deduce $I^h(\tilde{v}^{(h)}; U) \le CI^h(v^{(h)}; V)$ by (6) since sym $R^{(h)} \in \{A, B\}$. Let $w^{(h)} = \fint_{I_h}(v^{(h)})'\, dx_3$, $\tau^{(h)} = \fint_{I_h}(v^{(h)})_3\, dx_3$ and $\tilde{w}^{(h)} = \fint_{I_h}(\tilde{v}^{(h)})'\, dx_3$, $\tilde{\tau}^{(h)} = \fint_{I_h}(\tilde{v}^{(h)})_3\, dx_3$. Since sym $\bar{R}^{(h)} \in \{\bar{A}, \bar{B}\}$, by Jensen's inequality and by (38) we can estimate

$$\begin{aligned}
\int_{U\times I_h} W_{2D}(\nabla'\tilde{w}^{(h)}(x'))\, dx &\le \int_{U\times I_h} |\text{sym }\nabla'\tilde{w}^{(h)}(x') - \text{sym }\bar{R}^{(h)}(x')|^2\, dx \\
&\le \int_{U\times I_h} |\nabla\tilde{v}^{(h)}(x) - R^{(h)}(x')|^2\, dx \le C\int_{V\times I_h} W(\nabla v^{(h)}(x))\, dx.
\end{aligned}$$

Since $\nabla\tilde{v}^{(h)}(x) = \big(\psi_h * \nabla v^{(h)}(\cdot, x_3)\big)(x')$, for $\alpha \in \{1, 2\}$ we have $(\nabla\tilde{v}^{(h)})_{,\alpha}(x) = \int \psi_{,\alpha}(y')(\nabla v(x'-y', x_3) - \nabla v(x', x_3))dy'$ (where we have added a term which is zero by $\int \nabla'\psi_h = 0$). Using this together with Jensen's inequality and the fact that $|\nabla'\psi_h|^2 \le \frac{C}{h^2}$ while $|\text{spt }\psi_h| \le h^2$, an argument similar to (38) leads to $h\int_U |\nabla'^2\tilde{w}^{(h)}(x')|^2\, dx' + h\int_U |\nabla'^2\tilde{\tau}^{(h)}(x')|^2\, dx' \le CI^h(v^{(h)}; V)$. Summarizing, we have shown that

$$I^h_{2D}(\tilde{w}^{(h)}; U) + h\int_U |\nabla'^2\tilde{\tau}^{(h)}(x')|^2\, dx' \le CI^h(v^{(h)}; V). \qquad (40)$$

For $\kappa \in (0, a)$ let ϕ be a smooth cutoff function that decreases from one to zero within the transition layer $(-a-\kappa, -a)$ with $\|\phi'\|_\infty \le \frac{2}{\kappa}$. Consider the linear interpolation $u^{(h)}_\kappa(x) = v^{(h)}(x) + \phi(x_1)(\tilde{v}^{(h)}(x) - v^{(h)}(x))$. Since $\fint_{I_h}(\tilde{v}^{(h)})'\, dx_3 \to w_1^+$ and $\fint_{I_h}(v^{(h)})'\, dx_3 \to w_1^+$ in $L^2(U; \mathbb{R}^2)$, we also have $\fint_{I_h}(u^{(h)}_\kappa)'\, dx_3 \to w_1^+$ in $L^2(U; \mathbb{R}^2)$. Moreover, by (6) the energy $\int_{T_h} W(\nabla u^{(h)}_\kappa)$ on the transition layer

$T_h = (-a-\kappa, -a) \times (-\frac{1}{2}, \frac{1}{2}) \times I_h$ is bounded by

$$\begin{aligned} C \int_{T_h} W_0(\nabla u_\kappa^{(h)}(x))\, dx &\le C \int_{T_h} W_0(\nabla v^{(h)}(x)) + \frac{1}{\kappa^2} |\tilde{v}^{(h)}(x) - v^{(h)}(x)|^2 \\ &\quad + |\nabla \tilde{v}^{(h)}(x) - \nabla v^{(h)}(x)|^2\, dx \\ &\le \frac{C}{\kappa^2} \int_{T_h} W_0(\nabla v^{(h)}(x)) + |\nabla \tilde{v}^{(h)}(x) - \nabla v^{(h)}(x)|^2\, dx \\ &\le \frac{C}{\kappa^2} \int_{V \times I_h} W(\nabla v^{(h)}(x))\, dx. \end{aligned} \tag{41}$$

In passing to the second line we have assumed, by possibly adding a constant $c^{(h)}$ to $\tilde{v}^{(h)}$, that $\int_{T_h} \left(\tilde{v}^{(h)}(x) - v^{(h)}(x)\right) dx = 0$, so we could apply Poincaré's inequality to estimate the term involving $|\tilde{v}^{(h)} - v^{(h)}|^2$ (the varying domain causes no problem in the application of Poincaré's inequality; one could, e.g., apply it separately in the in-plane and in the x_3-directions). Note that $c^{(h)} \to 0$, since $w^{(h)}$ and $\tilde{w}^{(h)}$ converge to the same limit w_1^+ in $W^{1,2}(S;\mathbb{R}^2)$. In the last step in (41) we have used the fact that by (38) and since $\int_{U \times I_h} |\nabla v^{(h)} - R^{(h)}|^2 \le C \int_{V \times I_h} W(\nabla v^{(h)})$ (by the definition of $R^{(h)}$), we have $\int_{V \times I_h} |\nabla \tilde{v}^{(h)} - \nabla v^{(h)}|^2 \le C \int_{V \times I_h} W(\nabla v^{(h)})$. After applying the analogous construction to the right of the interface (adding a different constant $c_2^{(h)}$ to the corresponding $\tilde{v}^{(h)}$), the lemma follows because by Lemma 11(i) we have $\frac{1}{h^2} \int_{V \times I_h} W(\nabla v^{(h)}) \to 0$ as $h \to 0$. □

In the next lemma we further modify the recovery sequence such that the resulting functions are affine away from the interface. This is achieved via a two-step interpolation depicted in Figure 2. In the first step the recovery sequence is modified in such a way that it uses only one well away from the interface—namely, the one

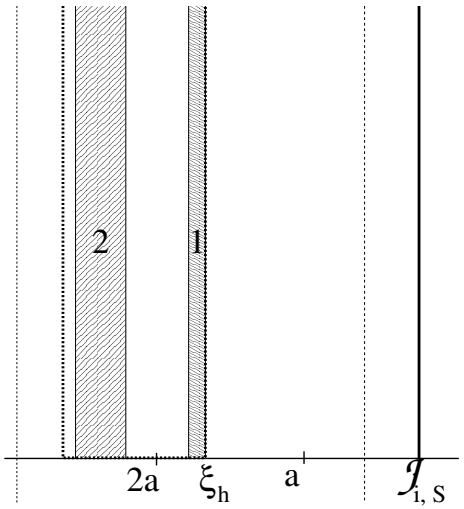


FIG. 2. *The shaded regions represent the interpolation layers whose numbers correspond to the steps in Lemma* 15. *The bold dashed lines belong to* ∂U_h, *the thin dashed lines to* ∂U, *and the solid horizontal line to* ∂S.

which is being used by the limiting mapping on that region. In a second step, it is further modified to become affine with gradient in the corresponding well.

LEMMA 15. *Let J_1, J_2 be open intervals, set $S = J_1 \times J_2$, and let $\sigma \in \{+,-\}$, $i \in \{1,2\}$. Then there is a (σ,i)-recovery sequence $v^{(h)}$ on S with the following property: For any $\varepsilon > 0$ there is $h_0 > 0$ such that, writing $Q = S \cap B_\varepsilon(\mathcal{J}_{i,S})$, the following holds: For $h \in (0,h_0)$ the mapping $v^{(h)}$ is affine on each connected component of $(S \setminus Q) \times I_h$, and $\nabla v^{(h)} \in Skew \cup (B + Skew)$ on $(S \setminus Q) \times I_h$.*

Proof. Arguing as at the beginning of the proof of Lemma 13, we may assume without loss of generality that $S = (-\frac{1}{2}, \frac{1}{2})^2$. We perform the construction only for $\sigma = +$ and $i = 1$ and only on the left side of the interface; the other cases are similar.

Set $\hat{S} = (-1,1)^2$ and fix $a \in (0, \frac{1}{100})$. Let $v^{(h)}$ be a $(+,1)$-recovery sequence on $\hat{S}$ as provided by Lemma 14, whose vertical averages $w^{(h)}(x') = \fint_{I_h} (v^{(h)}(x))' dx_3$ and $\tau^{(h)}(x') = \fint_{I_h} v_3^{(h)}(x) dx_3$ satisfy $w^{(h)} \in C^\infty(\hat{S} \setminus \{|x_1| < \frac{a}{10}\}; \mathbb{R}^2)$ and $\tau^{(h)} \in C^\infty(\hat{S} \setminus \{|x_1| < \frac{a}{10}\})$. Set $U = (-5a, -\frac{a}{2}) \times (-1,1)$. We may assume without loss of generality that $h < \frac{a}{100}$, and by Lemma 11(i) we have $I^h(v^{(h)}; U) \to 0$ as $h \to 0$. By the strong convergence $w^{(h)} \to w_1^+$ in $W^{1,2}(\hat{S}; \mathbb{R}^2)$ and since $U \subset \{\nabla' w_1^+ = 0\}$ we have $\int_U |\nabla' w^{(h)}|^2 \to 0$. Hence, using (37) in the conclusion of Lemma 14, we can write

$$I^h(v^{(h)}; U) + I^h_{2D}(w^{(h)}; U) + h \int_U |\nabla'^2 \tau^{(h)}(x')|^2 dx' + \int_U |\nabla' w^{(h)}(x')|^2 dx' = \eta_h, \tag{42}$$

where $\eta_h \to 0$ as $h \to 0$.

Step 1. *Interpolation to a displacement with low one-well energy.* As in the proof of Theorem 6 set $S_{z,h} = z + (-\frac{h}{2}, \frac{h}{2})^2$ and $\hat{S}'_h = \bigcup_{\{z \in h\mathbb{Z}^2 : S_{z,h} \subset \hat{S}\}} S_{z,h}$, and define the mapping $R^{(h)} : S'_h \to K$ to be constant on each $S_{z,h}$ with $z \in h\mathbb{Z}^2$ and such that $\int_{S_{z,h} \times I_h} |\nabla v^{(h)} - R^{(h)}|^2 \le C \int_{S_{z,h} \times I_h} \operatorname{dist}^2(\nabla v^{(h)}, K)$. (Here C is a universal constant given by applying Theorem 3 to a cube.) Let $G^{(h)} = h\mathbb{Z} \cap (-2a, -a)$ and set $N_h = \#G^{(h)}$. To every $\xi \in G^{(h)}$ define the column $Z_h(\xi) = \big((\xi - \frac{h}{2}, \xi + \frac{h}{2}) \times (-1,1)\big) \cap \hat{S}'_h$. By definition of U and since $h < \frac{a}{2}$, we have $Z_h(\xi) \subset U$ for all $\xi \in G^{(h)}$. Hence

$$\sum_{\xi \in G^{(h)}} \int_{Z_h(\xi)} |\nabla' w^{(h)}(x')|^2 dx' \le \int_U |\nabla' w^{(h)}(x')|^2 dx' \le \eta_h. \tag{43}$$

Now fix $\rho \in (0,1)$ such that $(1-4\rho)^2 > \frac{2}{3}$ and denote by $G_1^{(h)}$ the set of all $\xi \in G^{(h)}$ with the property that

$$\int_{Z_h(\xi)} |\nabla' w^{(h)}(x')|^2 dx' \le \frac{\eta_h}{[\rho N_h]}. \tag{44}$$

The estimate (43) implies that $\#G_1^{(h)} \ge (1-\rho)N_h$. Notice that by (44) and since $N_h \ge \frac{a}{2h}$, for $\xi \in G_1^{(h)}$ we have

$$\int_{Z_h(\xi)} |\nabla' w^{(h)}(x')|^2 dx' \le C\eta_h h. \tag{45}$$

Now set $L^{(h)} = \operatorname{sym} R^{(h)}$ and define $\mathcal{A}_i^h$ as in (25)–(26) and Ω_i^h as in (27), $i = 0,1,2$. By (42) and (25) we have $h \cdot \#(\mathcal{A}_0^h \cap U) \to 0$ as $h \to 0$. Hence the set $G_2^{(h)}$ of all $\xi \in G^{(h)}$ with the property that $Z_h(\xi) \cap \mathcal{A}_0^h = \emptyset$ satisfies $\frac{\#G_2^{(h)}}{N_h} \ge \frac{N_h - \#(\mathcal{A}_0^h \cap U)}{N_h} \to 1$

as $h \to 0$ because $hN_h \geq \frac{a}{2} > 0$. On the other hand, by the L^1-convergence $\chi_{\Omega_1^h} \to \chi_{\hat{S}\cap\{\nabla' w_1^+ = 0\}}$, the set $G_3^{(h)}$ of all $\xi \in G^{(h)}$ with the property that $Z_h(\xi) \cap \mathcal{A}_1^h \neq \emptyset$ satisfies $\frac{\#G_3^{(h)}}{N_h} \to 1$ as well. By these two cardinality estimates, for h small enough we have $\#(G_2^{(h)} \cap G_3^{(h)}) \geq (1-\rho)N_h$, and since also $\#G_1^{(h)} \geq (1-\rho)N_h$, we conclude that $\#(G_1^{(h)} \cap G_2^{(h)} \cap G_3^{(h)}) \geq (1-4\rho)N_h$. On the other hand, $N_h \geq \frac{a}{h} - 2 \geq (1-4\rho)\frac{a}{h}$ for small h. Hence by the choice of ρ we conclude

$$\#(G_1^{(h)} \cap G_2^{(h)} \cap G_3^{(h)}) > \frac{2a}{3h} \tag{46}$$

for small h. From (28) we deduce that if $\xi \in G_2^{(h)} \cap G_3^{(h)}$, then $Z_h(\xi) \subset \Omega_1^h$. Hence using that by definition sym $R^{(h)} = 0$ on Ω_1^h, from (24), (42), and the definition of $R^{(h)}$ we conclude that $\xi \in G_2^{(h)} \cap G_3^{(h)}$ implies

$$\begin{aligned}\int_{Z_h(\xi)\times I_h} |\text{sym } \nabla v^{(h)}(x)|^2 dx &= \int_{Z_h(\xi)\times I_h} |\text{sym } \nabla v^{(h)}(x) - \text{sym } R^{(h)}(\xi, x_2)|^2 dx \\ &\leq Ch^2 I^h(v^{(h)}; U) \leq C\eta_h h^2. \end{aligned} \tag{47}$$

Let $\tilde{J}_2^{(h)}$ be the set of all $\xi \in (-2a, -a)$ satisfying the property (P_h) (defined in the statement of Lemma 18 in the appendix) for $w^{(h)}$. Applying Lemma 18 on the domain $(-2a, -a) \times (-1, 1)$, for h small enough, we have $\mathcal{H}^1(\tilde{J}_2^{(h)}) \geq \frac{a}{2}$. For every $\xi \in \tilde{J}_2^{(h)}$ there is $\tilde{w}^{(h)} \in W^{1,2}((\xi - a, \xi) \times (-\frac{1}{2}, \frac{1}{2}); \mathbb{R}^2)$ satisfying $w^{(h)}(\xi, \cdot) = \tilde{w}^{(h)}(\xi, \cdot)$ in the trace sense and

$$\begin{aligned}\frac{1}{h}\int_{(\xi-a,\xi)\times(-\frac{1}{2},\frac{1}{2})} |\text{sym } \nabla' \tilde{w}^{(h)}(x')|^2 dx' &\leq C\left(I_{2D}^h(w^{(h)}; U) + \int_U |\nabla' w^{(h)}(x')|^2 dx'\right) \\ &\leq C\eta_h \end{aligned} \tag{48}$$

with C independent of ξ and h. We restrict the further construction to the domain of interest $S = (-\frac{1}{2}, \frac{1}{2})^2$. For $\xi \in (-2a, -a)$ define the columns

$$Z_h'(\xi) = \left(\xi - \frac{h}{8}, \xi\right) \times \left(-\frac{1}{2}, \frac{1}{2}\right) \tag{49}$$

and consider $\tilde{J}_1^{(h)} = \bigcup_{\xi \in \bigcap_{i=1}^3 G_i^{(h)}} (\xi - \frac{3h}{8}, \xi + \frac{3h}{8})$. By (46) we have $\mathcal{H}^1(\tilde{J}_1^{(h)} \cap (-2a, -a)) = \frac{3}{4}h \cdot (\#\bigcap_{i=1}^3 G_i^{(h)}) > \frac{a}{2}$. Since $\tilde{J}_2^{(h)} \subset (-2a, -a)$ and $\mathcal{H}^1(\tilde{J}_2^{(h)}) \geq \frac{a}{2}$ we conclude that there is $\xi_h \in \tilde{J}_1^{(h)} \cap \tilde{J}_2^{(h)}$ with $\xi_h \in (-2a, -a)$. Note that, in general, $\xi_h \notin G^{(h)}$ but that $Z_h'(\xi_h) \subset Z_h(\xi)$ for some $\xi \in \cap_{i=1}^3 G_i^{(h)}$ by the definitions of $\tilde{J}_1^{(h)}$ and $Z_h'(\xi)$. Let us introduce the set $U_h = (\xi_h - a, \xi_h) \times (-\frac{1}{2}, \frac{1}{2})$, which satisfies $U_h \subset (-3a, -a) \times (-\frac{1}{2}, \frac{1}{2}) \subset U$ for all h, since $\xi_h \in (-2a, -a)$. Setting

$$W_h = \text{skew} \fint_{U_h} \nabla' \tilde{w}^{(h)}(x')\, dx', \tag{50}$$

we apply Korn's inequality in the plane to deduce that there is an affine mapping $f^{(h)}$ with $\nabla f^{(h)} = W_h$ such that

$$\int_{U_h} |\tilde{w}^{(h)} - f^{(h)}|^2 + |\nabla' \tilde{w}^{(h)} - W_h|^2 dx' \leq C \int_{U_h} |\text{sym } \nabla' \tilde{w}^{(h)}|^2 dx' \leq Ch\eta_h, \tag{51}$$

where the last estimate holds by (48). Notice that C in (51) is independent of h because the constant appearing in Korn's inequality is invariant under translation of the domain. We claim that

$$W_h \to 0 \text{ in } \mathbb{R}^{2\times 2}. \tag{52}$$

Indeed, consider any subsequence. By the trace inequality and the fact that $\tilde{w}$ and w agree on $\{x \in S : x_1 = \xi_h\}$ we have

$$\begin{aligned}\int_{-\frac{1}{2}}^{\frac{1}{2}} |f^{(h)}(\xi_h, x_2)|^2 dx_2 &\le C \int_{-\frac{1}{2}}^{\frac{1}{2}} |f^{(h)}(\xi_h, x_2) - \tilde{w}^{(h)}(\xi_h, x_2)|^2 dx_2 \\ &\quad + C \int_{-\frac{1}{2}}^{\frac{1}{2}} |w^{(h)}(\xi_h, x_2)|^2 dx_2 \\ &\le C \int_{U_h} |f^{(h)}(x') - \tilde{w}^{(h)}(x')|^2 + |W_h - \nabla' \tilde{w}^{(h)}(x')|^2 \\ &\quad + |w^{(h)}(x')|^2 + |\nabla' w^{(h)}(x')|^2 dx',\end{aligned}$$

which tends to zero by (51) and since $w^{(h)} \to w_1^+$ in $W^{1,2}(S;\mathbb{R}^2)$. Hence (after passing to a subsequence) $f^{(h)}(\xi_h, x_2) \to 0$ for all $x_2 \in (-\frac{1}{2}, \frac{1}{2})$. From this and using that sym $W_h = 0$ we deduce (52).

Now we extend $\tilde{w}^{(h)}$ to a three dimensional displacement $\tilde{v}^{(h)}$ by defining

$$\tilde{v}^{(h)}(x) = \begin{pmatrix} \tilde{w}^{(h)}(x') \\ \tau^{(h)}(x') \end{pmatrix} + x_3 \begin{pmatrix} -\tau^{(h)}_{,1}(x') \\ -\tau^{(h)}_{,2}(x') \\ 0 \end{pmatrix}. \tag{53}$$

By (42) and (48) we have

$$\begin{aligned}\int_{U_h \times I_h} |\text{sym } \nabla \tilde{v}^{(h)}(x)|^2 \, dx &\le h^3 \int_{U_h} |\nabla'^2 \tau^{(h)}(x')|^2 \, dx' + h \int_{U_h} |\text{sym } \nabla' \tilde{w}^{(h)}(x')|^2 dx' \\ &\le C\eta_h h^2. \end{aligned} \tag{54}$$

(Later we will repeat this construction on the other side of the interface. Then one must replace the second summand in (53) by $x_3\big(\mu_2 - \tau^{(h)}_{,1}(x'),\ \mu_1 - \tau^{(h)}_{,2}(x'),\ \mu_3\big)^T$.) Now consider the interpolation

$$u^{(h)}(x) = v^{(h)}(x) + \phi^{(h)}(x_1)(\tilde{v}^{(h)}(x) - v^{(h)}(x)), \tag{55}$$

where $\phi^{(h)} : \mathbb{R} \to [0,1]$ denotes a smooth cutoff function that decreases from one to zero within the interval $(\xi_h - \frac{h}{8}, \xi_h)$, so spt $(\phi^{(h)})' \subset (\xi_h - \frac{h}{8}, \xi_h)$. We claim that

$$\int_{U_h \times I_h} |\text{sym } \nabla u^{(h)}(x)|^2 dx \le C\tilde{\eta}_h h^2, \tag{56}$$

where $\tilde{\eta}_h = \eta_h + |W_h|^2$ converges to zero as $h \to 0$.

To prove (56), recall that by the definition of U_h and (49) we have $Z_h'(\xi_h) \subset U_h$. Now notice that $u^{(h)} = \tilde{v}^{(h)}$ on $\big(U_h \times I_h\big) \setminus \big(Z_h'(\xi_h) \times I_h\big)$, whence by (54) we have

$\int_{(U_h\setminus Z_h'(\xi_h))\times I_h}|\text{sym }\nabla u^{(h)}|^2\le C\eta_h h^2$. It remains to prove (56) with the integration domain $Z_h'(\xi_h)\times I_h$ replacing $U_h\times I_h$. We make a standard calculation to obtain

$$\int_{Z_h'(\xi_h)\times I_h}|\text{sym }\nabla u^{(h)}(x)|^2dx$$

$$\le C\int_{Z_h'(\xi_h)\times I_h}|\text{sym }\nabla v^{(h)}(x)|^2+|\text{sym }\nabla\tilde v^{(h)}(x)|^2+\frac{1}{h^2}|\tilde v^{(h)}(x)-v^{(h)}(x)|^2dx. \tag{57}$$

Since $\xi_h\in G_2^{(h)}\cap G_3^{(h)}$, the first term on the right-hand side is estimated by (47). By (54) the second term in (57) satisfies $\int_{Z_h'(\xi_h)\times I_h}|\text{sym }\nabla\tilde v(x)|^2dx\le C\eta_h h^2$. Let us estimate the third term in (57). Since

$$\begin{pmatrix}w^{(h)}(x')\\ \tau^{(h)}(x')\end{pmatrix}=\fint_{I_h}v^{(h)}(x)dx_3=\fint_{I_h}v^{(h)}(x)+x_3(\nabla'\tau^{(h)}(x'))^Tdx_3$$

and since $w^{(h)}=\tilde w^{(h)}$ on the line $\{x\in S:x_1=\xi_h\}$, we can apply a Poincaré inequality (see, e.g., [16, Theorem 6.1-8]) to estimate the second term in the last step in (58) below. The first term in that step is estimated by the usual Poincaré inequality in the x_3-direction:

$$\int_{Z_h'(\xi_h)\times I_h}\frac{1}{h^2}|v^{(h)}(x)-\tilde v^{(h)}(x)|^2dx$$

$$\le\frac{C}{h^2}\int_{Z_h'(\xi_h)\times I_h}\left|v^{(h)}(x)+x_3\begin{pmatrix}\tau_{,1}^{(h)}(x')\\ \tau_{,2}^{(h)}(x')\\ 0\end{pmatrix}-\begin{pmatrix}w^{(h)}(x')\\ \tau^{(h)}(x')\end{pmatrix}\right|^2+|w^{(h)}(x')-\tilde w^{(h)}(x')|^2dx$$

$$\le C\int_{Z_h'(\xi_h)\times I_h}\left|v_{,3}^{(h)}(x)+\begin{pmatrix}\tau_{,1}^{(h)}(x')\\ \tau_{,2}^{(h)}(x')\\ 0\end{pmatrix}\right|^2+|\nabla'w^{(h)}(x')-\nabla'\tilde w^{(h)}(x')|^2dx. \tag{58}$$

To estimate the first term in (58), we observe that $(\tau_{,1}^{(h)}(x'),\tau_{,2}^{(h)}(x'),0)=\nabla\fint_{I_h}v_3^{(h)}(x)\,dx_3$, so we have

$$\int_{Z_h'(\xi_h)\times I_h}\left|v_{,3}^{(h)}(x)+\begin{pmatrix}\tau_{,1}^{(h)}(x')\\ \tau_{,2}^{(h)}(x')\\ 0\end{pmatrix}\right|^2dx\le\ C\int_{Z_h'(\xi_h)\times I_h}\left|v_{,3}^{(h)}(x)-\fint_{I_h}v_{,3}^{(h)}(x',z)dz\right|^2$$

$$+\left|\fint_{I_h}v_{,3}^{(h)}(x',z)dz+\left(\nabla\fint_{I_h}v_3^{(h)}(x',z)dz\right)^T\right|^2dx.$$

The second term is bounded by $\int_{Z_h'(\xi_h)\times I_h}|\text{sym }\nabla v^{(h)}|^2$ and can therefore be estimated by (47). By the x_3-independence of $R^{(h)}$, by Jensen's inequality, by definition of $R^{(h)}$,

and by (42), the first term can be estimated as follows:

$$\begin{aligned}\int_{Z_h'(\xi_h)\times I_h} \left| v_{,3}^{(h)}(x) - \fint_{I_h} v_{,3}^{(h)}(x',z)\,dz\right|^2\,dx \\ \leq C\int_{Z_h'(\xi_h)\times I_h} |v_{,3}^{(h)}(x) - R_3^{(h)}(x')|^2 + \left|\fint_{I_h} R_3^{(h)}(x') - v_{,3}^{(h)}(x',z)\,dz\right|^2\,dx \\ \leq C\eta_h h^2.\end{aligned}$$

Finally, the second term in (58) is bounded by

$$\begin{aligned}C\int_{Z_h'(\xi_h)\times I_h} |\nabla' w^{(h)}(x')|^2 + |\nabla'\tilde{w}^{(h)}(x') - W_h|^2 + |W_h|^2\,dx \\ \leq C\big(\eta_h h^2 + |Z_h'(\xi_h)\times I_h||W_h|^2\big).\end{aligned}$$

We have applied (52) and (45) multiplied by h (recall that $\xi_h \in G_1^{(h)}$), since here we are integrating over the thickness on the left-hand side. This proves (56) and finishes the first interpolation step.

Step 2. *Interpolation to an affine displacement.* We apply Lemma 16 to the mapping $u^{(h)}$ defined in (55) with $J_h = (\xi_h - a, \xi_h)$ instead of J_1, (so $J_h \times (-\frac{1}{2}, \frac{1}{2}) = U_h$) and with $t = \xi_h - \frac{a}{2}$ and $b = \frac{a}{4}$. Lemma 16 furnishes a mapping $\tilde{u}^{(h)}$ which agrees with $u^{(h)}$ on $\{x \in S\times I_h : x_1 > \xi_h - \frac{a}{2}\}$ and on $\{x \in S\times I_h : x_1 < \xi_h - \frac{3a}{4}\}$ agrees with an affine function $f^{(h)}$ with sym $\nabla f^{(h)} = 0$ (the mapping $\tilde{u}^{(h)}$ is at first not defined on $\{x_1 < \xi_h - a\}$, but since it is affine on $\{x \in S\times I_h : x_1 \in (\xi_h - a, \xi_h - \frac{3a}{4})\}$ we can extend it affinely). Moreover, $\tilde{u}^{(h)}$ satisfies $\int_{U_h\times I_h} |\text{sym }\nabla\tilde{u}^{(h)}|^2 \leq \frac{C}{a^4}\int_{U_h\times I_h} |\text{sym }\nabla u^{(h)}|^2$. Combining this with (56) and with the fact that $\tilde{u}^{(h)} = v^{(h)}$ on $\{x \in S\times I_h : x_1 > \xi_h\}$ and sym $\nabla\tilde{u}^{(h)} = 0$ on $\{x \in S\times I_h : x_1 < \xi_h - \frac{3a}{4}\}$, we conclude that

$$I^h(\tilde{u}^{(h)}; S) \to k(\nu_1). \tag{59}$$

Step 3. *Convergence.* Now we apply Steps 1 and 2 with obvious modifications also on the other side of the interface. Let us denote the resulting mappings by $\tilde{u}_a^{(h)}$. Hence $\tilde{u}_a^{(h)}$ is a $(+,1)$-recovery sequence on S which is affine on $\{x \in S\times I_h : |x_1| > 5a\}$ with sym $\nabla\tilde{u}_a^{(h)} = 0$ on $\{x \in S\times I_h : x_1 < -5a\}$ and sym $\nabla\tilde{u}_a^{(h)} = B$ on $\{x \in S\times I_h : x_1 > 5a\}$. By Proposition 19 there exists a sequence $a^{(h)} \to 0$ such that

$$\limsup_{h\to 0} I^h(\tilde{u}_{a^{(h)}}^{(h)}; S) = \limsup_{a\to 0}\limsup_{h\to 0} I^h(\tilde{u}_a^{(h)}; S) = k(\nu_1). \tag{60}$$

Theorem 6 implies that there exist affine mappings $f^{(h)}$ with sym $\nabla f^{(h)} = 0$ and $w \in \mathcal{A}(S)$ such that, after passing to an unrelabeled subsequence,

$$\bar{w}^{(h)} + f^{(h)} \to w \text{ strongly in } W^{1,2}(S;\mathbb{R}^2), \tag{61}$$

where we have set $\bar{w}^{(h)}(x') = \fint_{I_h} (\tilde{u}_{a^{(h)}}^{(h)}(x', x_3))'\,dx_3$. By Theorem 12 and (60) the limiting function w satisfies $I^0(w; S) \leq k(\nu_1)$. But by (61) and the properties of $\tilde{u}_{a^{(h)}}^{(h)}$ necessarily sym $\nabla' w = 0$ on $\{x \in S : x_1 < 0\}$ and sym $\nabla' w = \bar{B}$ on $\{x \in S : x_1 > 0\}$. Hence after possibly adding $w_1^+ - w$ to all $f^{(h)}$ (notice that $w_1^+ - w$ is affine with sym $\nabla'(w_1^+ - w) = 0$) we may assume that (61) holds with $w = w_1^+$ on the right-hand side. Since the same limit is obtained for every subsequence, (61) holds for the full

sequence with $w = w_1^+$, whence $x \mapsto \tilde{u}^{(h)}_{a^{(h)}}(x) + \binom{f^{(h)}(x')}{0}$ is the sought recovery sequence. ∎

LEMMA 16. *Let J_1, J_2 be open intervals, set $U = J_1 \times J_2$, let $i \in \{1,2\}$, and let $D \in \mathbb{R}^{3\times 3}$ be a symmetric matrix. Then there is a constant $C > 0$ such that the following holds: For every $h \in (0,1)$, for every $u^{(h)} \in W^{1,2}(U \times I_h; \mathbb{R}^3)$, for every $t \in J_i$, and for every $b \in (0,1)$ satisfying $t + b \in J_i$ (resp., $t - b \in J_i$), there exist $c^{(h)} \in \mathbb{R}^3$, $T^{(h)} \in \mathbb{R}^{3\times 3}$ with sym $T^{(h)} = 0$, and $\tilde{u}^{(h)} \in W^{1,2}(U \times I_h; \mathbb{R}^3)$ such that $\tilde{u}^{(h)} = u^{(h)}$ on $\{x \in U \times I_h : x_i < t\}$ (resp., $\{x \in U \times I_h : x_i > t\}$) and $\tilde{u}^{(h)} = Dx + T^{(h)}x + c^{(h)}$ on $\{x \in U \times I_h : x_i > t + b\}$ (resp., $\{x \in U \times I_h : x_i < t - b\}$) and such that $\int_{U\times I_h} |sym\ \nabla \tilde{u}^{(h)} - D|^2 \le \frac{C}{b^4} \int_{U\times I_h} |sym\ \nabla u^{(h)} - D|^2$. One can take*

$$T^{(h)} = skew \fint_{U\times I_h} \nabla u^{(h)}(x)\, dx \text{ and } c^{(h)} = \fint_{U\times I_h} u^{(h)}(x)\, dx. \tag{62}$$

Proof. We may assume without loss of generality that $D = 0$. (In fact, if $D \neq 0$, then apply the lemma with $D = 0$ to $\hat{u}^{(h)}(x) = u^{(h)}(x) - Dx$ instead of $u^{(h)}$ to obtain $\hat{u}^{(h)}$. Then define $\tilde{u}^{(h)}(x) = \hat{u}^{(h)}(x) + Dx$.) Moreover, we prove only the case when $i = 1$ and $t + b \in J_1$. By Proposition 17(ii) and Poincaré's inequality the mappings $f^{(h)}(x) = T^{(h)}x + c^{(h)}$ with $T^{(h)}$ and $c^{(h)}$ as in (62) satisfy

$$\int_{U\times I_h} |u^{(h)}(x) - f^{(h)}|^2 + |\nabla u^{(h)}(x) - T^{(h)}|^2 dx \le \frac{C}{h^2} \int_{U\times I_h} |\text{sym } \nabla u^{(h)}(x)|^2 dx. \tag{63}$$

Also, by Proposition 17(iii)

$$\int_{U\times I_h} |(u^{(h)})'(x) - (f^{(h)})'(x)|^2 dx \le C \int_{U\times I_h} |\text{sym } \nabla u^{(h)}(x)|^2 dx. \tag{64}$$

Fix a smooth cutoff function $\phi(x_1)$ which decreases from one to zero within the interval $(t + \frac{b}{4}, t + \frac{3b}{4})$. Set

$$\tilde{u}^{(h)}(x) = f^{(h)}(x) + \phi(x_1)(u^{(h)}(x) - f^{(h)}(x)) - x_3\phi'(x_1)(u_3^{(h)}(x) - f_3^{(h)}(x))e_1.$$

Then

$$\begin{aligned}\nabla \tilde{u}^{(h)}(x) &= T^{(h)} + \phi(x_1)(\nabla u^{(h)}(x) - T^{(h)}) + \begin{pmatrix} (u^{(h)} - f^{(h)})'(x) \\ 0 \end{pmatrix} \otimes e_1\phi'(x_1) \\ &\quad + (u^{(h)} - f^{(h)})_3(x)e_3 \otimes e_1\phi'(x_1) - (u^{(h)} - f^{(h)})_3(x)e_1 \otimes e_3\phi'(x_1) \\ &\quad - x_3 \left(\phi''(x_1)(u^{(h)} - f^{(h)})_3(x)e_1 \otimes e_1 + \phi'(x_1)e_1 \otimes (\nabla u_3^{(h)}(x) - (T^{(h)})^T e_3)\right).\end{aligned}$$

Upon taking the symmetric part of the above expression, the second line cancels, so we obtain

$$\begin{aligned}\int_{U\times I_h} |\text{sym } \nabla \tilde{u}^{(h)}|^2\, dx &\le C \int_{U\times I_h} |\text{sym } \nabla u^{(h)}|^2 + \frac{1}{b^2}|(u^{(h)} - f^{(h)})'|^2\, dx \\ &\quad + h^2 \int_{U\times I_h} \left(\frac{1}{b^4}|u^{(h)} - f^{(h)}|^2 + \frac{1}{b^2}|\nabla u^{(h)} - T^{(h)}|^2\right) dx,\end{aligned}$$

since $|\phi'| \le \frac{C}{b}$ and $|\phi''| \le \frac{C}{b^2}$. The last term is controlled by (63) and the $(u^{(h)} - f^{(h)})'$-term is controlled by (64). ∎

Proof of Theorem 1. By Lemma 2 we must prove the theorem only for the special case $A = 0$ and B as in (7). Statement (i) just rephrases the content of Theorem 12. The proof of (ii) is similar to that of Proposition 5.1 in [19]; compare also Step 2 in the proof of Theorem 5.6 in [18]. If $I^0(w) = \infty$, then the proof is trivial. Otherwise, $w \in \mathcal{A}(S)$, so by Proposition 7, w is affine on each connected component of $S \setminus \bigcup_{i=1}^{M} \mathcal{J}_i$, where the interfaces $\mathcal{J}_i \subset S$, $i = 1, \dots, M$ ($M \in \mathbb{N} \cup \{\infty\}$), are straight line segments parallel to e_1 or e_2 which intersect ∂S at their ends. By bounded variation we have $\sum_{i=1}^{M} \mathcal{H}^1(\mathcal{J}_i) < \infty$, so if $M = \infty$, then $\mathcal{H}^1(\mathcal{J}_i) \to 0$ as $i \to \infty$, so the $\mathcal{J}_i$ can accumulate only at ∂S. By translation invariance we may assume without loss of generality that S is strictly star-shaped with respect to the origin, so $\bar{S} \subset \eta S$ for any $\eta > 1$. Hence the restriction of $w_\eta(x) = \eta w(\frac{x}{\eta})$ to S is well defined and satisfies $w_\eta|_S \in \mathcal{A}(S)$. The mapping $w_\eta|_S$ has only finitely many interfaces $\mathcal{J}_i^\eta$, $i = 1, \dots, M_\eta$ (recall that the $\mathcal{J}_i$ accumulate only near ∂S), and they satisfy $\operatorname{dist}(\mathcal{J}_i^\eta, \mathcal{J}_k^\eta) > 0$ whenever $k \neq i$. By construction we have $\mathcal{J}_i^\eta = \tilde{\mathcal{J}}_i^\eta \cap S$, where $\tilde{\mathcal{J}}_i^\eta \subset \eta S$ are the interfaces of $w_\eta \in \mathcal{A}(\eta S)$. For $i = 1, \dots, M_\eta$, let $k_i \in \{1, 2\}$ be such that ν_{k_i} is normal to $\mathcal{J}_i^\eta$. Let $\varepsilon > 0$ be small (fixed below) and define the rectangles $S_i = \{x \in \mathbb{R}^2 : \pi_{k_i}(x) \in \pi_{k_i}(\tilde{\mathcal{J}}_i^\eta) + (-\varepsilon, \varepsilon) \text{ and } \pi_{k_i}^\perp(x) \in \pi_{k_i}^\perp(\tilde{\mathcal{J}}_i^\eta)\}$. Since the endpoints of $\tilde{\mathcal{J}}_i^\eta$ lie in $\partial(\eta S)$ and since the $\tilde{\mathcal{J}}_i^\eta$ do not intersect in ηS, for ε small enough the sides of ∂S_i which are parallel to ν_{k_i} do not intersect $\bar{S}$ and the S_i are pairwise disjoint. Define $\sigma_i \in \{-, +\}$ by the requirement that $w_\eta = w_{k_i, S_i}^{\sigma_i}$ on $S_i \cap \eta S$.

Now let $h_n \to 0$ be given and for all $i = 1, \dots, M_\eta$ let $(u_n^{(i)}, h_n)$ be a (σ_i, k_i)-recovery sequence on S_i. By Lemma 15 we may assume that $u_n^{(i)}$ is affine with $\operatorname{sym} \nabla u_n^{(i)} \in \{A, B\}$ in a neighborhood (relatively open in $S \times I_{h_n}$) of $(S \cap \partial S_i) \times I_{h_n}$. Hence we can extend each $u_n^{(i)}$ to all of $S \times I_{h_n}$ in such a way that $u_n^{(i)}$ is affine on each connected component of $(S \setminus S_i) \times I_{h_n}$. For $i, j = 1, \dots, M_\eta$ let us write $i \sim j$ whenever there exists a connected component S'_{ij} of $S \setminus \bigcup_{i=1}^{M_\eta} \mathcal{J}_i^\eta$ satisfying $S \cap \partial S'_{ij} = \mathcal{J}_i^\eta \cup \mathcal{J}_j^\eta$ (i.e., $\mathcal{J}_i^\eta$ and $\mathcal{J}_j^\eta$ are neighbors). By star-shapedness and the fact that $M_\eta < \infty$, and recalling that $u_n^{(i)}$ is affine on $(S \setminus S_i) \times I_{h_n}$, one inductively finds affine mappings $f_n^{(i)} : \mathbb{R}^3 \to \mathbb{R}^3$ with $\operatorname{sym} \nabla f_n^{(i)} = 0$, $i = 1, \dots, M_\eta$, such that $f_n^{(i)} + u_n^{(i)} = f_n^{(j)} + u_n^{(j)}$ on $S'_{ij} \setminus (S_i \cup S_j)$ whenever $i \sim j$. Now, for every $i = 1, \dots, M_\eta$, we set $v_n^\eta = u_n^{(i)} + f_n^{(i)}$ on S_i and extend it affinely to $S \times I_{h_n}$. Then $v_n^\eta \in W^{1,2}(S \times I_{h_n}; \mathbb{R}^3)$ is well defined by the choice of the $f_n^{(i)}$. Moreover, $\limsup_{n\to\infty} I^{h_n}(v_n^\eta; S) \leq I^0(w_\eta; S) + \rho_\eta$, where $\rho_\eta \downarrow 0$ as $\eta \downarrow 1$. By a diagonal sequence argument, this implies part (ii) of Theorem 1, since $w_\eta|_S \to w$ in $W^{1,2}(S; \mathbb{R}^2)$ and $I^0(w_\eta; S) \to I^0(w; S)$ as $\eta \downarrow 1$ (i.e., limiting displacements which arise as rescalings w_η of some $w \in \mathcal{A}(S)$ are energy dense; compare [2, p. 3]). □

Appendix. The following proposition was used in the proof of Lemma 16. Statement (ii) is Korn's inequality for thin films as presented in [3]; see also Problem 1.12 in [17].

PROPOSITION 17. *Let $S \subset \mathbb{R}^2$ be a bounded Lipschitz domain and let $A, B \in \mathbb{R}^{3\times 3}$ be such that rank$(A - B + F) \geq 2$ for all $F \in \mathbb{R}^{3\times 3}$ with sym $F = 0$. Then there is a constant $C(S)$ such that for all $h \in (0, 1)$ and for all $v^{(h)} \in W^{1,2}(S \times I_h; \mathbb{R}^3)$ the following hold:*

(i) *There exists a matrix sym $T^{(h)} \in \{A, B\}$ such that*

$$\int_{S\times I_h} |\nabla v^{(h)}(x) - T^{(h)}|^2 \, dx \leq \frac{C(S)}{h^2} \int_{S\times I_h} \operatorname{dist}^2(\operatorname{sym} \nabla v^{(h)}(x), \{A, B\}) \, dx.$$

(ii) *The estimate*

$$\int_{S\times I_h} |\nabla v^{(h)}(x) - T^{(h)}|^2 \, dx \le \frac{C(S)}{h^2} \int_{S\times I_h} |sym\ \nabla v^{(h)}(x)|^2 dx$$

holds for $T^{(h)} = skew \fint_{S\times I_h} \nabla v^{(h)}$.

(iii) *There exists* $c^{(h)} \in \mathbb{R}^2$ *such that*

$$\int_{S\times I_h} |(v^{(h)})'(x) - (T^{(h)}x)' - c^{(h)}|^2 dx \le C(S) \int_{S\times I_h} |sym\ \nabla v^{(h)}(x)|^2 \, dx$$

for the same $T^{(h)}$ *as in* (ii).

Proof. The proof of (i) is analogous to that of Theorem 10 in [29], with Theorem 3 replacing their geometric rigidity theorem. Statement (ii) can be proven in the same way, with Korn's inequality for one well replacing their geometric rigidity theorem. Another proof is given in [3] and [33]. Notice that if (ii) holds for some skew matrix, then it will also hold for the special choice $T^{(h)} = \text{skew} \fint_{S\times I_h} \nabla v^{(h)} dx$.

To prove statement (iii), set $w^{(h)}(x') = \fint_{I_h} (v^{(h)})'(x', x_3) dx_3$. From Korn's inequality in the plane and from Jensen's inequality we obtain

$$\begin{aligned} \int_S \left|\nabla' w^{(h)}(x') - \bar{T}^{(h)}\right|^2 \, dx' &\le C \int_S \left|\text{sym } \nabla' w^{(h)}(x')\right|^2 \, dx' \\ &\le \frac{C}{h} \int_{S\times I_h} |\text{sym } \nabla v^{(h)}(x)|^2 \, dx. \end{aligned} \tag{65}$$

With $c^{(h)} = \fint_S (w^{(h)}(x') - \bar{T}^{(h)}x') dx'$ we obtain

$$\begin{aligned} &\int_{S\times I_h} \left|(v^{(h)})'(x) - (T^{(h)}x)' - c^{(h)}\right|^2 \, dx \\ &\le C \int_{S\times I_h} \left|(v^{(h)})'(x) - (T_3^{(h)})' x_3 - w^{(h)}(x')\right|^2 + \left|w^{(h)}(x') - \bar{T}^{(h)}x' - c^{(h)}\right|^2 \, dx, \end{aligned}$$

where $T_3^{(h)}$ denotes the third column of $T^{(h)}$. The second term is estimated by applying Poincaré's inequality on S and then (65). To estimate the first term, notice that since the integration domain is symmetric, we have $w^{(h)}(x') = \fint_{I_h} (v'(x) - (T_3^{(h)})' x_3) \, dx_3$. Applying Poincaré's inequality in the x_3-direction for almost every x' and subsequently using (ii) shows that the first term is controlled by $\int_{S\times I_h} |\text{sym } \nabla v^{(h)}|^2$. □

The following lemma is a corollary of Proposition 4.1 in [19]. Notice that their ε corresponds to our h.

LEMMA 18. *Let* $a, l, d > 0$, *let* $U = (-l, l) \times (-d, d)$, *let* $\bar{A} = 0$ *and* $\bar{B} = e_1 \otimes e_2 + e_2 \otimes e_1$, *and let* $\bar{F} \in \{\bar{A}, \bar{B}\}$. *Then there are constants* $\eta_0, C_0 > 0$ *such that for every* $h \in (0, 1)$ *and* $w \in W^{2,2}(U; \mathbb{R}^2)$ *with*

$$I_{2D}^h(w; U) \le \eta_0 \text{ and } \int_U |sym\ \nabla' w(x') - \bar{F}|^2 \, dx' \le \eta_0$$

the set of $\xi \in (-l, l)$ *satisfying property* (P_h) *for* w *has measure not smaller than* l.

We say $\xi \in (-l,l)$ satisfies property (P_h) for $w \in W^{2,2}(U;\mathbb{R}^2)$ if, setting $U_1 = (\xi - a, \xi) \times (-\frac{d}{2}, \frac{d}{2})$ and $U_2 = (\xi, \xi + a) \times (-\frac{d}{2}, \frac{d}{2})$, for each $i = 1, 2$ there exist $\tilde{w}_i \in W^{1,2}\big(U_i;\mathbb{R}^2\big)$ with $\tilde{w}(\xi,\cdot) = w(\xi,\cdot)$ on $(-d/2, d/2)$ and

$$\frac{1}{h}\int_{U_i} |sym\ \nabla' \tilde{w}_i(x') - \bar{F}|^2 dx' \leq C_0 \left(I^h_{2D}\big(w;U\big) + \int_U |sym\ \nabla' w(x') - \bar{F}|^2 dx' \right).$$

An analogous result holds for lines of the form $\{x_2 = \xi\}$.

Proof. Set $\eta = I^h_{2D}(w;U) + \int_U |\mathrm{sym}\ \nabla' w - \bar{F}|^2$. By Proposition 4.1 in [19] there exists a Borel set $\Sigma \subset (-l,l)$ with $|\Sigma| \geq l$ and such that for all $\xi \in \Sigma$ there exists an affine mapping $w_\xi : \mathbb{R}^2 \to \mathbb{R}^2$ with $\mathrm{sym}\ \nabla' w_\xi = \bar{F}$ and

$$\|w(\xi,\cdot) - w_\xi(\xi,\cdot)\|^2_{H^{1/2}((-\frac{d}{2},\frac{d}{2});\mathbb{R}^2)} \leq Ch\eta. \tag{66}$$

By the properties of the $H^{1/2}$-norm (see, e.g., the appendix of [32] for a review), for $i = 1, 2$ there exist $v_i \in W^{1,2}(U_i;\mathbb{R}^2)$ such that

$$\int_{U_i} |\nabla' v_i(x')|^2\ dx' \leq \|w(\xi,\cdot) - w_\xi(\xi,\cdot)\|^2_{H^{1/2}((-\frac{d}{2},\frac{d}{2});\mathbb{R}^2)} \tag{67}$$

and $v_i(\xi,\cdot) = w(\xi,\cdot) - w_\xi(\xi,\cdot)$ on $(-\frac{d}{2}, \frac{d}{2})$ in the trace sense. Setting $\tilde{w}_i = v_i + w_\xi$, we find

$$\frac{1}{h}\int_{U_i} |\mathrm{sym}\ \nabla' \tilde{w}_i(x') - \bar{F}|^2\ dx' \leq \frac{C}{h}\int_{U_i} |\nabla' v_i(x')|^2 + |\mathrm{sym}\ \nabla' w_\xi(x') - \bar{F}|^2\ dx' \leq C\eta$$

and $\tilde{w}_i(\xi,\cdot) = w(\xi,\cdot)$ on $(-\frac{d}{2}, \frac{d}{2})$ in the trace sense. We have used (66)–(67) and the fact that $\mathrm{sym}\ \nabla' w_\xi = \bar{F}$. ∎

The following proposition is a standard diagonalization lemma (compare [4, Corollary 1.16] or [13, Lemma 7.2]).

PROPOSITION 19. *Let $a_{k,j}$ be a doubly indexed sequence of real numbers, $k, j \to \infty$. Then there exists a subsequence $k_j \to \infty$ such that*

$$\limsup_{j\to\infty} a_{k_j,j} = \limsup_{k\to\infty} \limsup_{j\to\infty} a_{k,j}.$$

Acknowledgments. This work is part of my Ph.D. thesis supervised by Prof. S. Müller (Max-Planck-Institute for Mathematics in the Sciences, Leipzig), whom I thank for his steady advice. I also wish to thank Prof. S. Conti for interesting discussions on the topic and for a lot of helpful comments on a preliminary version of this work.

REFERENCES

[1] E. ACERBI, G. BUTTAZZO, AND D. PERCIVALE, *A variational definition of an elastic string*, J. Elasticity, 25 (1991), pp. 137–148.

[2] G. ALBERTI, *Variational models for phase transitions: An approach via Γ-convergence*, in Calculus of Variations and Partial Differential Equations, Topics on Geometrical Evolution Problems and Degree Theory, G. Butazzo et al., eds., Springer-Verlag, Berlin, 2000, pp. 95–114.

[3] G. ANZELLOTTI, S. BALDO, AND D. PERCIVALE, *Dimension reduction in variational problems, asymptotic development in Γ-convergence and thin structures in elasticity*, Asymptot. Anal., 9 (1994), pp. 61–100.

[4] H. ATTOUCH, *Variational Convergence for Functions and Operators*, Pitman, Boston, 1984.

[5] A. C. Barroso and I. Fonseca, *Anisotropic singular perturbations—the vectorial case*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 527–571.
[6] P. Bělík and M. Luskin, *The Γ-convergence of a sharp interface thin film model with non-convex elastic energy*, SIAM J. Math. Anal., 38 (2006), pp. 414–433.
[7] K. Bhattacharya, *Microstructure of Martensite*, Oxford Series on Materials Modelling, Oxford University Press, Oxford, UK, 2003.
[8] K. Bhattacharya and G. Dolzmann, *Relaxation of some multi-well problems*, Proc. Roy. Soc. Edinburgh Sect. A, 131 (2001), pp. 279–320.
[9] K. Bhattacharya, N. B. Firoozye, R. D. James, and R. V. Kohn, *Restrictions on microstructure*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 843–878.
[10] K. Bhattacharya and R. D. James, *A theory of thin films of martensitic materials with applications to microactuators*, J. Mech. Phys. Solids, 47 (1999), pp. 531–576.
[11] K. Bhattacharya and R. D. James, *The material is the machine*, Science, 307 (2005), pp. 53–54.
[12] F. Bourquin, P. G. Ciarlet, G. Geymonat, and A. Raoult, *Γ-convergence et analyse asymptotique des plaques minces*, C. R. Acad. Sci. Paris Sér. I Math., 315 (1992), pp. 1017–1024.
[13] A. Braides, I. Fonseca, and G. Francfort, *3D-2D asymptotic analysis for inhomogeneous thin films*, Indiana Univ. Math. J., 49 (2000), pp. 1367–1403.
[14] N. Chaudhuri and S. Müller, *Rigidity estimate for two incompatible wells*, Calc. Var. Partial Differential Equations, 19 (2004), pp. 379–390.
[15] N. Chaudhuri and S. Müller, *Scaling of the energy for thin martensitic films*, SIAM J. Math. Anal., 38 (2006), pp. 468–477.
[16] P. G. Ciarlet, *Mathematical Elasticity Vol.* I, Studies in Mathematics and Its Applications 20, North-Holland, Amsterdam, 1988.
[17] P. G. Ciarlet, *Mathematical Elasticity Vol.* II, Studies in Mathematics and Its Applications 27, North-Holland, Amsterdam, 1997.
[18] S. Conti, I. Fonseca, and G. Leoni, *A Γ-convergence result for the two-gradient theory of phase transitions*, Comm. Pure Appl. Math., 55 (2002), pp. 857–936.
[19] S. Conti and B. Schweizer, *A sharp-interface limit for a two-well problem*, Arch. Ration. Mech. Anal., 179 (2006), pp. 413–452.
[20] S. Conti and B. Schweizer, *Rigidity and Gamma convergence for solid-solid phase transitions with SO(2)-invariance*, Comm. Pure Appl. Math., 59 (2006), pp. 830–868.
[21] C. DeLellis and L. Székelyhidi, Jr., *Simple proof of two-well rigidity*, C. R. Math. Acad. Sci. Paris, 343 (2006), pp. 367–370.
[22] A. De Simone and G. Friesecke, *On the problem of two linearized wells*, Calc. Var. Partial Differential Equations, 4 (1996), pp. 293–304.
[23] G. Dolzmann and S. Müller, *Microstructures with finite surface energy: The two-well problem*, Arch. Ration. Mech. Anal., 132 (1995), pp. 101–141.
[24] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.
[25] I. Fonseca, *Phase transitions of elastic solid materials*, Arch. Ration. Mech. Anal., 107 (1989), pp. 195–223.
[26] I. Fonseca and C. Mantegazza, *Second order singular perturbation models for phase transitions*, SIAM J. Math. Anal., 31 (2000), pp. 1121–1143
[27] I. Fonseca and L. Tartar, *The gradient theory of phase transitions for systems with two potential wells*, Proc. Roy. Soc. Edinburgh Sect. A, 111 (1989), pp. 89–102.
[28] G. Friesecke, R. D. James, and S. Müller, *A theorem on geometric rigidity and the derivation of nonlinear plate theory from three dimensional elasticity*, Comm. Pure Appl. Math., 55 (2002), pp. 1461–1506.
[29] G. Friesecke, R. D. James, and S. Müller, *A hierarchy of plate models derived from nonlinear elasticity by Gamma-convergence*, Arch. Ration. Mech. Anal., 180 (2006), pp. 183–236.
[30] M. Giaquinta, *Multiple Integrals in the Calculus of Variations and Nonlinear Elliptic Systems*, Princeton University Press, Princeton, NJ, 1983.
[31] M. Giaquinta, *Introduction to Regularity Theory for Nonlinear Elliptic Systems*, Birkhäuser, Basel, 1993.
[32] R. V. Kohn and S. Müller, *Surface energy and microstructure in coherent phase transitions*, Comm. Pure Appl. Math., 47 (1994), pp. 405–435.
[33] R. V. Kohn and M. Vogelius, *A new model for thin plates with rapidly varying thickness* II: *A convergence proof*, Quart. Appl. Math., 43 (1985), pp. 1–22.
[34] H. Le Dret and A. Raoult, *The nonlinear membrane model as variational limit of nonlinear three-dimensional elasticity*, J. Math. Pures Appl., 74 (1995), pp. 549–578.

[35] L. Modica and S. Mortola, *Un esempio die* Γ*-convergenza*, Boll. Un. Mat. Ital. B (5), 14 (1977), pp. 285–299.
[36] M. G. Mora and S. Müller, *Derivation of a rod theory for multiphase materials*, Calc. Var. Partial Differential Equations, 28 (2007), pp. 161–178.
[37] Y. C. Shu, *Heterogeneous thin films of martensitic materials*, Arch. Ration. Mech. Anal., 153 (2000), pp. 39–90.
[38] P. Sternberg, *The effect of a singular perturbation on nonconvex variational problems*, Arch. Ration. Mech. Anal., 101 (1988), pp. 209–260.
[39] K. Zhang, *Isolated microstructures on linear elastic strains*, Proc. R. Soc. Lond. A Ser. A math. Phys. Eng. Sci., 460 (2004), pp. 2993–3011.

SIAM J. MATH. ANAL.
Vol. 40, No. 1, pp. 215–237

DERIVATION OF A MACROSCOPIC RECEPTOR-BASED MODEL USING HOMOGENIZATION TECHNIQUES*

ANNA MARCINIAK-CZOCHRA† AND MARIYA PTASHNYK‡

Abstract. We study the problem of diffusive transport of biomolecules in the intercellular space, modeled as porous medium, and of their binding to the receptors located on the surface membranes of the cells. Cells are distributed periodically in a bounded domain. To describe this process we introduce a reaction-diffusion equation coupled with nonlinear ordinary differential equations on the boundary. We prove existence and uniqueness of the solution of this problem. We consider the limit, when the number of cells tends to infinity and at the same time their size tends to zero, while the volume fraction of the cells remains fixed. Using the homogenization technique of two-scale convergence, we show that the sequence of solutions of the original problem converges to the solution of the so-called macroscopic problem. To show the convergence of the nonlinear terms on the surfaces we use the unfolding method (periodic modulation). We discuss applicability of the result to mathematical description of membrane receptors of biological cells and compare the derived model with those previously considered.

Key words. homogenization, two-scale convergence, intercellular communication, receptor-ligand binding, reaction-diffusion equations, unfolding method (periodic modulation)

AMS subject classifications. 35B27, 74Q10, 74Q15, 35K57, 35K60

DOI. 10.1137/050645269

1. Introduction.

Regulatory and signaling molecules (ligands) act by binding and activating receptor molecules. Receptors are usually located in the cell membrane, with some exceptions such as lipophilic ligands, which are located in the cytoplasm [24, 33, 34]. Some receptors interact with surface-bound ligands, such as adhesion proteins and extracellular matrix components. Other receptors bind soluble ligands, such as growth factors and cytokines. There are also many ligands which are present in both forms. As an example, antibodies, which are secreted by B cells as soluble molecules, become surface-bound ligands for the Fc receptors upon binding to antigens deposited on the surface [25].

Soluble molecules which are secreted to the intercellular space and transported via diffusion provide cell-to-cell communication, which results in the activation of processes in cells at a distance from the original signal. This happens, for example, in the case of the bystander effect. There is strong evidence that unirradiated bystander cells respond to signals emitted by irradiated cells [28]. In another case, the interplay between the spatial transport of virons and interferons results in the formation of patterns of infected and resistant cells [10]. Intercellular signaling can also lead to the formation of spatially nonhomogeneous structures, which is especially evident in developmental processes [33, 34]. The effects of the spatial transport of the soluble molecules are even visible in experiments in which only spatial averages are measured in order to understand the time dynamics of a signaling pathway. There is evidence

*Received by the editors November 15, 2005; accepted for publication (in revised form) August 31, 2007; published electronically April 16, 2008.
http://www.siam.org/journals/sima/40-1/64526.html
†Center for Modeling and Simulation in the Biosciences (BIOMS), Institute of Applied Mathematics, University of Heidelberg, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany (anna.marciniak@iwr.uni-heidelberg.de).
‡Institute of Applied Mathematics, University of Heidelberg, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany (mariya.ptashnyk@iwr.uni-heidelberg.de).

that different mixing conditions strongly influence quantitative and also qualitative results of such experiments [17]. Therefore, there arises a need to explain how the intercellular transport of the molecules should be described on the macroscopic level.

Models proposed so far are mainly phenomenological and describe all the processes on the macroscale level represented by a two- or three-dimensional sheet of cells [29, 30, 31, 32, 42, 46]. However, the real geometry is much more complicated, and binding a soluble ligand to a cell surface receptor requires interaction of molecules diffusing in three-dimensional space with some molecules attached to a two-dimensional surface. Since the size of a cell is very small compared to the dimension of the whole tissue, systems that include cells have to be treated as multiscale systems.

The aim of the present work is to derive a macroscopic model of receptor-ligand binding, based on a microscopic description, using methods of asymptotic analysis. Such an approach is called homogenization and it hinges on demonstrating the convergence of solutions of a sequence of microscopic problems to the solution of the macroscopic problem in properly chosen function spaces. We use here the two-scale convergence, which was introduced in [2] and [36] for sequences of functions $\{u^\varepsilon\}$ bounded in L^2 or in H^1 on an ε-periodic domain. Then, in [37] and [3], the definition of two-scale convergence was extended to sequences of functions defined on ε-periodic hypersurfaces, with dependence on parameters. This extension was used to homogenize a diffusion-reaction process in a catalyst consisting of distributed bars [37]. A similar problem with convection was studied in [19] using the standard homogenization technique, the energy method. A model describing processes of diffusion, convection, and nonlinear reactions in a periodic array of cells was studied in [20]. In that paper, the convergence of the nonlinear terms was shown using their monotonicity. Homogenization of models of chemical reactive flows in domains with periodically distributed reactive solid grains was also recently studied by Conca et al. [9]. They considered a stationary reaction-diffusion model with nonlinear, fast growing but monotone kinetics on the the surface of reactive solid grains and a model of reaction-diffusion processes both inside and outside of grains. Homogenization of the reaction-diffusion-convection processes with linear reactions on the surface of microstructures was also considered by Hornung in [18].

The model presented in this paper includes the dynamics of molecule concentrations on the surface of microstructures described by nonlinear ordinary differential equations. Therefore, we apply the concept of two-scale convergence of functions from L^∞ on ε-periodic hypersurfaces. To show convergence of the nonlinear terms on the surface of microstructures we use the unfolding method (periodic modulation); see [5, 6, 7].

Our paper is organized as follows. First, we present a precise description of the considered ε-periodic geometry (section 2) and of the equations describing the microscopic nature of the receptor-ligand binding process (section 3). These equations are spatially scaled by ε. Then we show existence and uniqueness of solutions of the microscopic model (section 3.2) and a priori estimates (section 3.3). In section 4, after extension of the solutions from the porous domain to the whole domain, using a priori estimates, we show the convergence of solutions of the microscopic problem to the solutions of a macroscopic homogenized model. Effective macroscopic equations are derived in section 4.2 and formulated in Theorem 4.4. In section 5 we compare a derived macroscopic model of the receptor-ligand binding on cells surfaces with the phenomenological models previously discussed in the literature.

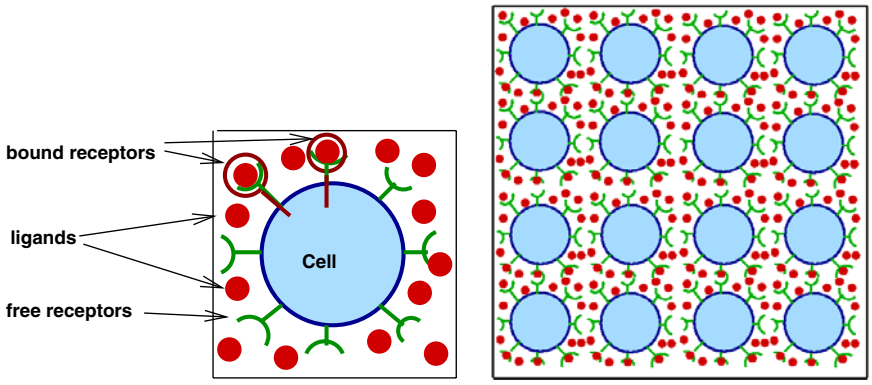


FIG. 1. *Geometry of the model. The array of the cells (on the right-hand side) consists of periodic repetition of the so-called standard cell, $Z = [0,1]^3$ (on the left-hand side), which corresponds to a single biological cell with the surrounding intercellular space.*

2. Problem formulation. We consider a model involving a system of cells, periodically distributed in a three-dimensional cube $\Omega = [a,b]^3$, $a, b \in \mathbb{R}$, $a < b$, with boundary Γ^N. For the mathematical formulation of the problem we consider the so-called standard cell, $Z = [0,1]^3$, periodically repeated over $\mathbb{R}^3$ with $Y_0 \subset Z$, an open subset with a smooth boundary Γ; $Y = Z \backslash \overline{Y}_0$; and ν, the outer normal of Y (see Figure 1).

Let $\varepsilon > 0$ be a given scale factor such that $\varepsilon = \frac{b-a}{n}$, $n \in \mathbb{N}$, denoting the ratio between the size of the cells and the size of the whole domain Ω. Then the geometric structure within the fixed domain Ω is obtained by intersecting the ε-multiple εZ with Ω. We define, for $k \in \mathbb{Z}^3$, a triple of integers; and e_i, unit vectors, $\Gamma^k = \Gamma + \sum_{i=1}^3 k_i e_i$, $Y_0^k = Y_0 + \sum_{i=1}^3 k_i e_i$, $Z^k = Z + \sum_{i=1}^3 k_i e_i$, $\Gamma^* = \cup\{\Gamma^k, \ k \in \mathbb{Z}^3\}$, $Z^* = \cup\{Z^k, \ k \in \mathbb{Z}^3\}$. We further define $\Omega_0^\varepsilon = \cup\{\varepsilon Y_0^k | \varepsilon Z^k \subset \Omega, \ k \in \mathbb{Z}^3\}$, $\Omega^\varepsilon = \Omega \backslash \Omega_0^\varepsilon$, $\Gamma^\varepsilon = \cup\{\varepsilon\Gamma^k | \varepsilon Z^k \subset \Omega, \ k \in \mathbb{Z}^3\}$.

Remark 2.1. The geometry defined above fulfills the assumptions that

1. cells (holes in the domain) do not touch the boundary $\partial\Omega$;
2. cells do not touch each other;
3. cells have smooth boundary.

These assumptions allow for the definition of the functions on the cell boundaries using periodic repetition, and the definition of extension as proposed in [8]. Therefore, these assumptions are important for the methods applied in this paper. Homogenization of the Neumann problem in domains with more complicated geometry was considered in [1] and [4].

We assume that new ligands and new free receptors are produced on the cell surface through a combination of recycling (dissociation of bound receptors) and *de novo* production within the cell. Free receptors exist only on the surfaces, while ligands are transported by diffusion within the intercellular space, which is a porous medium. A ligand reversibly binds to a free receptor, which results in a bound receptor that can be internalized into the cell. Bound receptors also dissociate. Both ligands and free receptors undergo natural decay. We denote the concentration of ligands by $l^\varepsilon : (0,T) \times \Omega^\varepsilon \to \mathbb{R}$. Bound and free receptor densities are denoted by $r_b^\varepsilon : (0,T) \times \Gamma^\varepsilon \to \mathbb{R}$ and $r_f^\varepsilon : (0,T) \times \Gamma^\varepsilon \to \mathbb{R}$, respectively. For simplicity we assume that all binding processes are governed by the law of mass action without saturation effects.

3. Microscopic model.

3.1. Model assumptions. The microscopic model consists of the following equations:

Diffusion equation for ligands in the intercellular space,

$$\frac{\partial}{\partial t}l^\varepsilon(t,x) = \nabla\cdot(D^\varepsilon(t,x)\nabla l^\varepsilon(t,x)) - \mu_l^\varepsilon(t,x)l^\varepsilon(t,x) + p_l^\varepsilon(t,x,l^\varepsilon(x,t)) \quad \text{in } (0,T)\times\Omega^\varepsilon,$$

$$\nu^\varepsilon\cdot\nabla_x l^\varepsilon(t,x) = 0 \quad \text{on } (0,T)\times\Gamma^N,$$

$$l^\varepsilon(x,t) = l_0(x), \quad t=0,\ x\in\Omega^\varepsilon. \tag{1}$$

Binding equation on the surfaces,

$$-D^\varepsilon(t,x)\nabla l^\varepsilon(t,x)\cdot\nu^\varepsilon = \varepsilon(b^\varepsilon(t,x)l^\varepsilon(t,x){r_f}^\varepsilon(t,x) - d^\varepsilon(t,x)r_b^\varepsilon(t,x)) \text{ on } (0,T)\times\Gamma^\varepsilon. \tag{2}$$

Reaction equations for receptors on the surfaces,

$$\frac{\partial}{\partial t}{r_f}^\varepsilon(x,t) = -\mu_f^\varepsilon(t,x){r_f}^\varepsilon(x,t) + p_r^\varepsilon(t,x,r_b^\varepsilon(x,t)) - b^\varepsilon(t,x){r_f}^\varepsilon(x,t)l^\varepsilon(x,t)$$

$$+ d^\varepsilon(t,x){r_b}^\varepsilon(x,t), \tag{3}$$

$$\frac{\partial}{\partial t}{r_b}^\varepsilon(x,t) = -\mu_b^\varepsilon(t,x){r_b}^\varepsilon(x,t) + b^\varepsilon(t,x){r_f}^\varepsilon(x,t)l^\varepsilon(x,t) - d^\varepsilon(t,x){r_b}^\varepsilon(x,t), \tag{4}$$

with initial conditions

$${r_f}^\varepsilon(x,t) = {r_f}_0(x), \quad t=0,\ x\in\Gamma^\varepsilon, \tag{5}$$

$${r_b}^\varepsilon(x,t) = r_{b0}(x), \quad t=0,\ x\in\Gamma^\varepsilon. \tag{6}$$

The following is a list of functional coefficients in these equations:

$\mu_l^\varepsilon : (0,T)\times\Omega\to\mathbb{R}$	rate of decay of ligands,
$p_l^\varepsilon : (0,T)\times\Omega\times\mathbb{R}\to\mathbb{R}$	production of ligands,
$D^\varepsilon : (0,T)\times\Omega\to\mathbb{R}^{3\times 3}$	diffusion coefficient for ligands,
$p_r^\varepsilon : (0,T)\times\Gamma^\varepsilon\times\mathbb{R}\to\mathbb{R}$	production of new free receptors,
$\mu_f^\varepsilon : (0,T)\times\Gamma^\varepsilon\to\mathbb{R}$	rate of decay of free receptors,
$\mu_b^\varepsilon : (0,T)\times\Gamma^\varepsilon\to\mathbb{R}$	rate of decay of bound receptors,
$d^\varepsilon : (0,T)\times\Gamma^\varepsilon\to\mathbb{R}$	rate of dissociation of bound receptors,
$b^\varepsilon : (0,T)\times\Gamma^\varepsilon\to\mathbb{R}$	rate of binding of ligands and free receptors,

where functions on Ω or Γ^ε are defined by Z-periodic function: $D_{i,j}^\varepsilon(t,x) = D_{i,j}(t,\frac{x}{\varepsilon})$, $p_l^\varepsilon(t,x,\xi) = p_l(t,\frac{x}{\varepsilon},\xi)$, $\mu_l^\varepsilon(t,x) = \mu_l(t,\frac{x}{\varepsilon})$, $\mu_f^\varepsilon(t,x) = \mu_f(t,\frac{x}{\varepsilon})$, $\mu_b^\varepsilon(t,x) = \mu_b(t,\frac{x}{\varepsilon})$, $b^\varepsilon(t,x) = b(t,\frac{x}{\varepsilon})$, $d^\varepsilon(t,x) = d(t,\frac{x}{\varepsilon})$, $p_r^\varepsilon(t,x,\xi) = p_r(t,\frac{x}{\varepsilon},\xi)$, defined on Z^* and Γ^*, respectively.

We assume that decay processes are linear and that binding is a product of the density of ligands and free receptors. The proposed functions are the simplest functions usually used to describe decay or binding processes (see the models described in [35]), modeled by the law of mass action. We assume that *de novo* production of free receptors, denoted by p_r, is regulated by bound receptors. We assume that p_r

is a bounded Lipschitz continuous function in r_b and is nonnegative for nonnegative values of r_b, for example a Michaelis–Menten function $p_r = \frac{m_1 r_b}{1+r_b}$. In addition, we assume that the production of ligands depends on their density. It could be regulated via some other receptors not considered in our model. Thus, we assume that p_l is a Lipschitz continuous function in l, nonnegative for nonnegative values of l.

Assumption 3.1.

1. $D \in L^\infty((0,T)\times Z)^{3\times 3}$, $\partial_t D \in L^\infty((0,T)\times Z)^{3\times 3}$, $(D(t,x)\xi,\xi) \ge d_0|\xi|^2$ for some $d_0 > 0$, for every $\xi \in \mathbb{R}^3$, a.a. $(t,x) \in (0,T)\times Z$.
2. $\mu_l \in L^\infty((0,T)\times Z)$ and $\mu_l \ge 0$ a.e. in $(0,T)\times Z$.
3. p_l is measurable in t and x, sublinear, i.e., $|p_l(t,x,\xi)| \le c_1 + c_2|\xi|$ for a.a. $(t,x) \in (0,T)\times Z$, Lipschitz continuous in ξ, and $p_l(t,x,\xi) \ge 0$ for $\xi \ge 0$.
4. $b \in C([0,T]; C^{0,\alpha}(\Gamma))$, $b \ge 0$, in $[0,T]\times\Gamma$, $\partial_t b \in L^\infty((0,T)\times\Gamma)$.
5. $d \in C([0,T]; C^{0,\alpha}(\Gamma))$, $d \ge 0$, in $[0,T]\times\Gamma$, $\partial_t d \in L^\infty((0,T)\times\Gamma)$.
6. $\mu_f, \mu_b \in C([0,T]; C^{0,\alpha}(\Gamma))$, $\mu_f \ge 0$, $\mu_b \ge 0$, in $[0,T]\times\Gamma$.
7. $p_r(\xi) \in C([0,T]; C^{0,\alpha}(\Gamma))$ for all $\xi \in \mathbb{R}$, $p_r(t,x,\xi) \ge 0$ for $\xi \ge 0$, p_r is bounded, i.e., $|p_r(t,x,\xi)| \le m_1$ for all $(t,x,\xi) \in (0,T)\times\Gamma\times\mathbb{R}$ and is Lipschitz continuous in ξ.

3.2. Existence of the solutions of the microscopic model. We start with a weak formulation of the microscopic model.

Definition 3.2. *The triple $(l^\varepsilon, r_f^\varepsilon, r_b^\varepsilon)$ is a solution of problem* (1)–(6) *if $l^\varepsilon \in L^2((0,T); H^1(\Omega^\varepsilon))$, $\partial_t l^\varepsilon \in L^2((0,T)\times\Omega^\varepsilon)$, $l^\varepsilon \in L^\infty((0,T)\times\Omega^\varepsilon)$, r_f^ε, $r_b^\varepsilon \in L^\infty((0,T)\times\Gamma^\varepsilon)$, $\partial_t r_f^\varepsilon$, $\partial_t r_b^\varepsilon \in L^\infty((0,T)\times\Gamma^\varepsilon)$ such that*

1.
$$\begin{aligned}(\partial_t l^\varepsilon, \phi)_{(0,T)\times\Omega^\varepsilon} &= -(D^\varepsilon \nabla l^\varepsilon, \nabla\phi)_{(0,T)\times\Omega^\varepsilon} - (\mu_l^\varepsilon l^\varepsilon, \phi)_{(0,T)\times\Omega^\varepsilon} \\ &\quad + (d^\varepsilon r_b^\varepsilon - b^\varepsilon r_f^\varepsilon l^\varepsilon, \phi)_{(0,T)\times\Gamma^\varepsilon} + (p_l^\varepsilon(l^\varepsilon), \phi)_{(0,T)\times\Omega^\varepsilon}\end{aligned} \tag{7}$$
for all $\phi \in L^2((0,T); H^1(\Omega^\varepsilon))$;
2. *l^ε satisfies the initial condition, i.e., $l^\varepsilon \to l_0$ in $L^2(\Omega^\varepsilon)$ as $t \to 0$;*
3.
$$\begin{cases} \dfrac{\partial}{\partial t} {r_f}^\varepsilon(x,t) = -\mu_f^\varepsilon {r_f}^\varepsilon(x,t) + p_r^\varepsilon(t,x,r_b^\varepsilon(x,t)) \\ \qquad\qquad - b^\varepsilon {r_f}^\varepsilon(x,t) l^\varepsilon(x,t) + d^\varepsilon {r_b}^\varepsilon(x,t), \\ \dfrac{\partial}{\partial t} {r_b}^\varepsilon(x,t) = -\mu_b^\varepsilon {r_b}^\varepsilon(x,t) + b^\varepsilon {r_f}^\varepsilon(x,t) l^\varepsilon(x,t) \\ \qquad\qquad - d^\varepsilon {r_b}^\varepsilon(x,t) \end{cases} \tag{8}$$
a.e. $(0,T)\times\Gamma^\varepsilon$;
4. *${r_f}^\varepsilon, {r_b}^\varepsilon$ satisfy the initial conditions* (5)–(6).

Here $(u,v)_{(0,T)\times\Omega^\varepsilon} = \int_0^T \int_{\Omega^\varepsilon} u\,v\,dx\,dt$ and $(u,v)_{(0,T)\times\Gamma^\varepsilon} = \varepsilon \int_0^T \int_{\Gamma^\varepsilon} u\,v\,d\gamma_x\,dt$.

Theorem 3.3. *Let Assumption* 3.1 *be satisfied and*

$$l_0 \in C^{0,\alpha}(\overline{\Omega}),\ \ l_0 \in H^1(\Omega),\ \ l_0 \ge 0,$$

$$r_{f0}, r_{b0} \in C^{0,\alpha}(\overline{\Omega}),\ \ r_{f0} \ge 0,\ \ r_{b0} \ge 0.$$

Then there exists a unique solution $(l^\varepsilon, r_f^\varepsilon, r_b^\varepsilon)$ of problem (1)–(6), *such that*

$$l^\varepsilon \in H^1(0,T; L^2(\Omega^\varepsilon)),\ l^\varepsilon \in L^2(0,T; H^1(\Omega^\varepsilon)),$$

$$l^\varepsilon \in C^{0,\beta/2}([0,T]; C^{0,\beta}(\overline{\Omega}^\varepsilon)),$$

$$r_f^\varepsilon, r_b^\varepsilon \in C^1([0,T]; C^{0,\beta}(\Gamma^\varepsilon)), \quad \textit{where} \quad \beta \in (0,\alpha],$$

$$\textit{and} \quad l^\varepsilon \ge 0, r_f^\varepsilon \ge 0, r_b^\varepsilon \ge 0.$$

Proof. Existence. The existence of a solution of the system (1), (3), (4) will be proved by showing the existence of a fix point of the operator K defined on $C([0,T]\times\overline{\Omega}^{\varepsilon})$ by $l^{n,\varepsilon}=K(l^{n-1,\varepsilon})$ with $l^{n,\varepsilon}$ given by

$$
(9)\quad \begin{cases} \partial_t\, l^{n,\varepsilon} = \nabla\cdot(D^{\varepsilon}\nabla l^{n,\varepsilon}) - \mu_l^{\varepsilon} l^{n,\varepsilon} + p_l^{\varepsilon}(l^{n-1,\varepsilon}), & t>0,\ x\in\Omega^{\varepsilon},\\ \nabla l^{n,\varepsilon}\cdot\nu^{\varepsilon} = 0, & t>0,\ x\in\Gamma^{N},\\ l^{n,\varepsilon} = l_0, & t=0,\ x\in\Omega^{\varepsilon},\\ -D^{\varepsilon}\nabla l^{n,\varepsilon}\cdot\nu^{\varepsilon} = \varepsilon(b^{\varepsilon} l^{n,\varepsilon} {r_f}^{n,\varepsilon} - d^{\varepsilon}{r_b}^{n,\varepsilon}), & t>0,\ x\in\Gamma^{\varepsilon}, \end{cases}
$$

$$
(10)\quad \begin{cases} \partial_t {r_f}^{n,\varepsilon} = -\mu_f^{\varepsilon}{r_f}^{n,\varepsilon} + p_r^{\varepsilon}({r_b}^{n,\varepsilon}) - b^{\varepsilon}{r_f}^{n,\varepsilon} l^{n-1,\varepsilon} + d^{\varepsilon}{r_b}^{n,\varepsilon}, & t>0,\ x\in\Gamma^{\varepsilon},\\ \partial_t {r_b}^{n,\varepsilon} = -\mu_b^{\varepsilon}{r_b}^{n,\varepsilon} + b^{\varepsilon}{r_f}^{n,\varepsilon} l^{n-1,\varepsilon} - d^{\varepsilon}{r_b}^{n,\varepsilon}, & t>0,\ x\in\Gamma^{\varepsilon},\\ {r_f}^{n,\varepsilon} = r_{f0}, & t=0,\ x\in\Gamma^{\varepsilon},\\ {r_b}^{n,\varepsilon} = r_{b0}, & t=0,\ x\in\Gamma^{\varepsilon}. \end{cases}
$$

For a given $l^{n-1,\varepsilon}\in C([0,T]\times\overline{\Omega}^{\varepsilon})$, $l^{n-1,\varepsilon}\geq 0$ on $[0,T]\times\overline{\Omega}^{\varepsilon}$, there exists a unique solution of system (10), $r_f^{n,\varepsilon}$, $r_b^{n,\varepsilon}\in C^1([0,T];C(\Gamma^{\varepsilon}))$, because the right-hand side of the system of ordinary differential equations (10) is Lipschitz continuous [45]. Since p_r is a nonnegative function for nonnegative values of r_b and $l^{n-1,\varepsilon}\geq 0$ on $[0,T]\times\Gamma^{\varepsilon}$ and $r_{f0}\geq 0$, $r_{b0}\geq 0$, we deduce that $r_f^{n,\varepsilon}\geq 0$, $r_b^{n,\varepsilon}\geq 0$ on $[0,T]\times\Gamma^{\varepsilon}$.

Using the Galerkin method and a priori estimates similar to the estimates in Lemma 3.4, we obtain the existence of a weak solution of (9), $l^{n,\varepsilon}\in L^2(0,T;H^1(\Omega^{\varepsilon}))$, $\partial_t l^{n,\varepsilon}\in L^2(0,T;L^2(\Omega^{\varepsilon}))$; see [23]. Since $l_0\in C^{0,\alpha}(\overline{\Omega})$, there exists $\max_{\overline{\Omega}^{\varepsilon}}|l_0|=M$. In addition, $r_f^{n,\varepsilon}\geq 0$ and $|r_b^{n,\varepsilon}|\leq C$. Thus, we may apply the result from [26] (Theorem 6.40) stating that for parabolic equations with uniformly elliptic operator, sublinear terms of lower order, bounded free terms, and bounded coefficients of Robin boundary conditions, the boundedness of the initial conditions implies the boundedness of the supremum of a solution. From this, we conclude that $\sup_{(0,T)\times\Omega^{\varepsilon}}|l^{n,\varepsilon}|\leq M_1$. Then, since $l_0\in C^{0,\alpha}(\overline{\Omega})$, $r_f^{n,\varepsilon}\geq 0$, and $r_b^{n,\varepsilon}\in C^1([0,T];C(\Gamma^{\varepsilon}))$, we conclude also that $l^{n,\varepsilon}\in C^{0,\beta/2}([0,T];C^{0,\beta}(\overline{\Omega}^{\varepsilon}))$ (see Theorem III.10.1 in [23], generalized for Robin boundary conditions, or [11] and [26]). Using the maximum principle and the continuity of $l^{n,\varepsilon}$, we obtain that $l^{n,\varepsilon}\geq 0$ in $[0,T]\times\overline{\Omega}^{\varepsilon}$ [12].

The space $C^{0,\beta/2}([0,T];C^{0,\beta}(\overline{\Omega}^{\varepsilon}))$ is compact embedded in $C([0,T]\times\overline{\Omega}^{\varepsilon})$. Then, by virtue of the Schauder theorem, there exists a fixed point of K, a solution of the microscopic problem l^{ε}, r_f^{ε}, and r_b^{ε}. In addition, we obtain that $l^{\varepsilon}\geq 0$, $r_f^{\varepsilon}\geq 0$, and $r_b^{\varepsilon}\geq 0$. Since $r_{f0}, r_{b0}\in C^{0,\alpha}(\Omega)$ and $l^{\varepsilon}\in C^{0,\beta/2}([0,T];C^{0,\beta}(\overline{\Omega}^{\varepsilon}))$, we conclude also that $r_f^{\varepsilon}, r_b^{\varepsilon}\in C^1([0,T];C^{0,\beta}(\Gamma^{\varepsilon}))$.

Uniqueness. Suppose there are two solutions of the problem $(l^{1,\varepsilon}, r_f^{1,\varepsilon}, r_b^{1,\varepsilon})$ and $(l^{2,\varepsilon}, r_f^{2,\varepsilon}, r_b^{2,\varepsilon})$. We denote $l^{\varepsilon}=l^{1,\varepsilon}-l^{2,\varepsilon}$ and choose $\phi=l^{\varepsilon}$. We calculate

$$
\frac{1}{2}\int_0^{\tau}\int_{\Omega^{\varepsilon}}\Big(\partial_t|l^{\varepsilon}|^2 + (D^{\varepsilon}\nabla l^{\varepsilon},\nabla l^{\varepsilon}) + \mu_l^{\varepsilon}|l^{\varepsilon}|^2\Big)\,dx\,dt = \int_0^{\tau}\int_{\Omega^{\varepsilon}}(p_l^{\varepsilon}(l^{1,\varepsilon}) - p_l^{\varepsilon}(l^{2,\varepsilon}))l^{\varepsilon}\,dx\,dt
$$

$$
+\varepsilon\int_0^{\tau}\int_{\Gamma^{\varepsilon}}((d^{\varepsilon}r_b^{1,\varepsilon} - b^{\varepsilon}r_f^{1,\varepsilon}l^{1,\varepsilon}) - (d^{\varepsilon}r_b^{2,\varepsilon} - b^{\varepsilon}r_f^{2,\varepsilon}l^{2,\varepsilon}))\,l^{\varepsilon}\,d\gamma\,dt
$$

for any $\tau \in [0,T]$. For r_f^ε and r_b^ε we obtain

$$\frac{\partial}{\partial t}(r_f^{1,\varepsilon} - r_f^{2,\varepsilon}) = -\mu_f^\varepsilon(r_f^{1,\varepsilon} - r_f^{2,\varepsilon}) + (p_r^\varepsilon(r_b{}^{1,\varepsilon}) - p_r^\varepsilon(r_b{}^{2,\varepsilon})) - b^\varepsilon(r_f^{1,\varepsilon} l^{1,\varepsilon} - r_f^{2,\varepsilon} l^{2,\varepsilon}) + d^\varepsilon(r_b^{1,\varepsilon} - r_b^{2,\varepsilon}),$$

$$\frac{\partial}{\partial t}(r_b^{1,\varepsilon} - r_b^{2,\varepsilon}) = -\mu_b^\varepsilon(r_b^{1,\varepsilon} - r_b^{2,\varepsilon}) + b^\varepsilon(r_f^{1,\varepsilon} l^{1,\varepsilon} - r_f^{2,\varepsilon} l^{2,\varepsilon}) - d^\varepsilon(r_b^{1,\varepsilon} - r_b^{2,\varepsilon}).$$

Integrating by parts with respect to time and summing up side by side the last two equations, we obtain

$$|r_f^{1,\varepsilon} - r_f^{2,\varepsilon}| + |r_b^{1,\varepsilon} - r_b^{2,\varepsilon}| \le \int_0^\tau \Big(\mu_f^1 |r_f^{1,\varepsilon} - r_f^{2,\varepsilon}| + \mu_b^1 |r_b^{1,\varepsilon} - r_b^{2,\varepsilon}| + c_r |r_b^{1,\varepsilon} - r_b^{2,\varepsilon}|\Big) dt$$

$$+ \int_0^\tau \Big(2b_1 \max_{[0,T]\times\Gamma^\varepsilon} |l^{1,\varepsilon}||r_f^{1,\varepsilon} - r_f^{2,\varepsilon}| + 2b_1 \max_{[0,T]\times\Gamma^\varepsilon} |r_f^{2,\varepsilon}||l^{1,\varepsilon} - l^{2,\varepsilon}| + 2d_1 |r_b^{1,\varepsilon} - r_b^{2,\varepsilon}|\Big)\, dt,$$

where c_r is the Lipschitz constant of p_r, $\mu_f^1 = \sup_{[0,T]\times\Gamma^\varepsilon} |\mu_f^\varepsilon|$, $\mu_b^1 = \sup_{[0,T]\times\Gamma^\varepsilon} |\mu_b^\varepsilon|$, $b_1 = \sup_{[0,T]\times\Gamma^\varepsilon} |b^\varepsilon|$, $d_1 = \sup_{[0,T]\times\Gamma^\varepsilon} |d^\varepsilon|$. The Gronwall lemma implies

$$|r_f^{1,\varepsilon} - r_f^{2,\varepsilon}| + |r_b^{1,\varepsilon} - r_b^{2,\varepsilon}| \le C \int_0^\tau |l^{1,\varepsilon} - l^{2,\varepsilon}|\, dt. \tag{11}$$

Using the above estimate and nonnegativity of b^ε and $r_f^{2,\varepsilon}$, we obtain

$$\frac{1}{2}\int_0^\tau \int_{\Omega^\varepsilon} \partial_t |l^\varepsilon|^2\, dx\, dt + d_0 \int_0^\tau \int_{\Omega^\varepsilon} |\nabla l^\varepsilon|^2\, dx\, dt + \int_0^\tau \int_{\Omega^\varepsilon} \mu_l^\varepsilon |l^\varepsilon|^2\, dx\, dt$$

$$\le C d_1^2 \varepsilon \frac{1}{2\delta} \int_0^\tau \int_{\Gamma^\varepsilon} \int_0^t |l^\varepsilon|^2\, ds\, d\gamma\, dt + \varepsilon \frac{\delta}{2} \int_0^\tau \int_{\Gamma^\varepsilon} |l^\varepsilon|^2\, d\gamma\, dt + c_l \int_0^\tau \int_{\Omega^\varepsilon} |l^\varepsilon|^2\, dx\, dt$$

$$+ C b_1 \varepsilon \max_{[0,T]\times\Gamma^\varepsilon} |l^{1,\varepsilon}| \int_0^\tau \int_{\Gamma^\varepsilon} \int_0^t |l^\varepsilon|^2\, ds\, d\gamma\, dt + C b_1 \varepsilon \max_{[0,T]\times\Gamma^\varepsilon} |l^{1,\varepsilon}| \int_0^\tau \int_{\Gamma^\varepsilon} |l^\varepsilon|^2\, d\gamma\, dt,$$

where c_l is the Lipschitz constant of p_l. Furthermore, using the estimate

$$\varepsilon \int_0^\tau \int_{\Gamma^\varepsilon} |l^\varepsilon|^2\, d\gamma\, dt \le c \int_0^\tau \int_{\Omega^\varepsilon} |l^\varepsilon|^2\, dx\, dt + c\varepsilon^2 \int_0^\tau \int_{\Omega^\varepsilon} |\nabla l^\varepsilon|^2\, dx\, dt, \tag{12}$$

we obtain

$$\frac{1}{2}\int_{\Omega^\varepsilon} |l^\varepsilon|^2\, dx + (d_0 - \varepsilon^2\delta) \int_0^\tau \int_{\Omega^\varepsilon} |\nabla l^\varepsilon|^2\, dx\, dt + \int_0^\tau \int_{\Omega^\varepsilon} \mu_l^\varepsilon |l^\varepsilon|^2\, dx\, dt$$

$$\le C \frac{1}{\delta} \int_0^\tau \int_0^t \int_{\Omega^\varepsilon} (|l^\varepsilon|^2 + |\nabla l^\varepsilon|^2)\, dx\, ds\, dt + c_l \int_0^\tau \int_{\Omega^\varepsilon} |l^\varepsilon|^2\, dx\, dt.$$

From the Gronwall lemma and $\mu_l^\varepsilon \ge 0$, taking the supremum over $\tau \in [0,T]$, we conclude that

$$\int_{\Omega^\varepsilon} |l^\varepsilon|^2\, dx + C \int_0^T \int_{\Omega^\varepsilon} |\nabla l^\varepsilon|^2\, dx\, dt \le 0$$

and, therefore, $l^{1,\varepsilon} = l^{2,\varepsilon}$ in $(0,T)\times\Omega^\varepsilon$. Due to (11), also $r_f^{1,\varepsilon} = r_f^{2,\varepsilon}$ and $r_b^{1,\varepsilon} = r_b^{2,\varepsilon}$ on $[0,T]\times\Gamma^\varepsilon$. □

3.3. A priori estimates for the microscopic solutions.

LEMMA 3.4. *For any solution of problem* (1)–(6) *from Theorem* 3.3 *the following estimates hold:*

$$\|l^\varepsilon\|_{L^2(0,T;H^1(\Omega^\varepsilon))} \le C, \quad \|\partial_t l^\varepsilon\|_{L^2(0,T;L^2(\Omega^\varepsilon))} \le C,$$

$$\|r_f^\varepsilon\|_{L^\infty((0,T)\times\Gamma^\varepsilon)} \le C, \quad \|r_b^\varepsilon\|_{L^\infty((0,T)\times\Gamma^\varepsilon)} \le C,$$

$$\|\partial_t r_f^\varepsilon\|_{L^2((0,T)\times\Gamma^\varepsilon)} \le C, \quad \|\partial_t r_b^\varepsilon\|_{L^2((0,T)\times\Gamma^\varepsilon)} \le C,$$

where C is a constant independent on ε.

Proof. To show the estimates for r_f^ε and r_b^ε we add (3) and (4) side by side and obtain

$$\partial_t(r_f^\varepsilon + r_b^\varepsilon) \le m_1.$$

Since r_f^ε and r_b^ε are nonnegative (see Theorem 3.3), we conclude that

$$||r_f^\varepsilon||_{L^\infty((0,T)\times\Gamma^\varepsilon)} \le C \quad \text{and} \quad ||r_b^\varepsilon||_{L^\infty((0,T)\times\Gamma^\varepsilon)} \le C.$$

Now we show the estimates for l^ε. We choose $\phi = l^\varepsilon$ as a test function in (7) and calculate

$$\frac{1}{2}\int_0^\tau \int_{\Omega^\varepsilon} \partial_t |l^\varepsilon|^2 \, dx\, dt + \int_0^\tau \int_{\Omega^\varepsilon} (D^\varepsilon \nabla l^\varepsilon, \nabla l^\varepsilon)\, dx\, dt + \int_0^\tau \int_{\Omega^\varepsilon} \mu_l^\varepsilon |l^\varepsilon|^2 \, dx\, dt$$
$$= \varepsilon \int_0^\tau \int_{\Gamma^\varepsilon} (d^\varepsilon r_b^\varepsilon - b^\varepsilon r_f^\varepsilon l^\varepsilon)\, l^\varepsilon \, d\gamma\, dt + \int_0^\tau \int_{\Omega^\varepsilon} p_l^\varepsilon(l^\varepsilon) l^\varepsilon \, dx\, dt$$

for any $\tau \in [0,T]$. Applying the Young inequality we obtain

$$\frac{1}{2}\int_{\Omega^\varepsilon} |l^\varepsilon|^2 \, dx + \int_0^\tau \int_{\Omega^\varepsilon} d_0 |\nabla l^\varepsilon|^2 \, dx\, dt + \int_0^\tau \int_{\Omega^\varepsilon} \mu_l^\varepsilon |l^\varepsilon|^2 \, dx\, dt$$
$$\le \frac{\varepsilon d_1}{2\delta} \int_0^\tau \int_{\Gamma^\varepsilon} |r_b^\varepsilon|^2 \, d\gamma\, dt + \varepsilon \frac{\delta}{2} \int_0^\tau \int_{\Gamma^\varepsilon} |l^\varepsilon|^2 \, d\gamma\, dt$$
$$- \varepsilon \int_0^\tau \int_{\Gamma^\varepsilon} b^\varepsilon r_f^\varepsilon |l^\varepsilon|^2 \, d\gamma\, dt + c_1 \int_0^\tau \int_{\Omega^\varepsilon} |l^\varepsilon|^2 \, dx\, dt + \frac{1}{2}\int_{\Omega^\varepsilon} |l_0^\varepsilon|^2 \, dx.$$

Now we use (12), $\mu_l^\varepsilon \ge 0$, $b^\varepsilon \ge 0$, and $r_f^\varepsilon \ge 0$ and obtain

$$\frac{1}{2}\int_{\Omega^\varepsilon} |l^\varepsilon|^2 \, dx + \int_0^\tau \int_{\Omega^\varepsilon} \left(d_0 - \frac{\delta\varepsilon^2}{2}\right) |\nabla l^\varepsilon|^2 \, dx\, dt$$
$$\le \frac{\varepsilon}{2\delta} \int_0^\tau \int_{\Gamma^\varepsilon} |r_b^\varepsilon|^2 \, d\gamma\, dt + c_1 \int_0^\tau \int_{\Omega^\varepsilon} |l^\varepsilon|^2 \, dx\, dt + \frac{1}{2}\int_{\Omega^\varepsilon} |l_0^\varepsilon|^2 \, dx.$$

Then, from the Gronwall lemma and the estimate for r_b^ε, it follows that

$$\int_\Omega |l^\varepsilon|^2 \, dx + \int_0^T \int_{\Omega^\varepsilon} |\nabla l^\varepsilon|^2 \, dx\, dt \le C.$$

Using the estimates for l^ε, r_f^ε, and r_b^ε, we conclude from (8) that

$$||\partial_t r_f^\varepsilon||_{L^2((0,T)\times\Gamma^\varepsilon)} \le C,$$

$$||\partial_t r_b^\varepsilon||_{L^2((0,T)\times\Gamma^\varepsilon)} \le C.$$

To obtain the estimates for $\partial_t l^\varepsilon$ we choose $\phi = \partial_t l^\varepsilon$ as a test function and calculate

$$\int_0^\tau \int_{\Omega^\varepsilon} |\partial_t l^\varepsilon|^2 \, dx \, dt + \frac{1}{2} \int_0^\tau \int_{\Omega^\varepsilon} \Big(\partial_t (D^\varepsilon \nabla l^\varepsilon, \nabla l^\varepsilon) - (\partial_t D^\varepsilon \nabla l^\varepsilon, \nabla l^\varepsilon) \Big) \, dx \, dt$$

$$= \int_0^\tau \int_{\Omega^\varepsilon} (p_l^\varepsilon(l^\varepsilon) - \mu_l^\varepsilon l^\varepsilon) \partial_t l^\varepsilon \, dx \, dt + \varepsilon \int_0^\tau \int_{\Gamma^\varepsilon} \Big(\partial_t (d^\varepsilon r_b^\varepsilon l^\varepsilon) - d^\varepsilon \partial_t r_b^\varepsilon l^\varepsilon - \partial_t d^\varepsilon r_b^\varepsilon l^\varepsilon \Big) d\gamma \, dt$$

$$+ \varepsilon \int_0^\tau \int_{\Gamma^\varepsilon} \Big(-\partial_t (b^\varepsilon r_f^\varepsilon |l^\varepsilon|^2) + b^\varepsilon \partial_t r_f^\varepsilon |l^\varepsilon|^2 + \partial_t b^\varepsilon r_f^\varepsilon |l^\varepsilon|^2 \Big) d\gamma \, dt.$$

Using the Young inequality we obtain

$$\begin{aligned}
(1-\delta) &\int_0^\tau \int_{\Omega^\varepsilon} |\partial_t l^\varepsilon|^2 \, dx \, dt + \frac{d_0}{2} \int_{\Omega^\varepsilon} |\nabla l^\varepsilon|^2 \, dx \\
&\le \frac{\varepsilon}{2\delta} \int_{\Gamma^\varepsilon} d_1^2 |r_b^\varepsilon|^2 d\gamma + D_2 \int_0^\tau \int_{\Omega^\varepsilon} |\nabla l^\varepsilon|^2 \, dx \, dt \\
&\quad + \varepsilon \frac{\delta}{2} \int_{\Gamma^\varepsilon} |l^\varepsilon|^2 \, d\gamma + \frac{\varepsilon}{2} \int_0^\tau \int_{\Gamma^\varepsilon} (d_1^2 |\partial_t r_b^\varepsilon|^2 + |\partial_t d|^2 |r_b^\varepsilon|^2) \, d\gamma \, dt \\
&\quad + \frac{\varepsilon}{2} \int_0^\tau \int_{\Gamma^\varepsilon} |l^\varepsilon|^2 \, d\gamma \, dt - \varepsilon \int_{\Gamma^\varepsilon} b^\varepsilon r_f^\varepsilon |l^\varepsilon|^2 \, d\gamma \\
&\quad + \frac{1}{2\delta} \int_0^\tau \int_{\Omega^\varepsilon} (|p_l^\varepsilon(l^\varepsilon)|^2 + \mu_l^\varepsilon |l^\varepsilon|^2) \, dx \, dt + \varepsilon \int_{\Gamma^\varepsilon} (d_1 r_{b0} l_0 + b_1 r_{f0} |l_0|^2) \, d\gamma \\
&\quad + D_1 \int_{\Omega^\varepsilon} |\nabla l_0|^2 \, dx + \varepsilon \int_0^\tau \int_{\Gamma^\varepsilon} (\partial_t b^\varepsilon r_f^\varepsilon + b^\varepsilon |\partial_t r_f^\varepsilon|) \, |l^\varepsilon|^2 \, d\gamma \, dt,
\end{aligned}$$

where $D_1 = \sup_{(0,T)\times\Omega} |D^\varepsilon|$, $D_2 = \sup_{(0,T)\times\Omega} |\partial_t D^\varepsilon|$. For the estimate of the last integral we use the embedding for a space of dimension $n = 3$, i.e., $L^\infty(0,T; H^1(\Omega^\varepsilon)) \subset L^4((0,T)\times\Gamma^\varepsilon)$,

$$\begin{aligned}
\varepsilon \int_0^\tau \int_{\Gamma^\varepsilon} b^\varepsilon |\partial_t r_f^\varepsilon| |l^\varepsilon|^2 \, d\gamma \, dt &\le \frac{b_1^2 \varepsilon}{2\delta} \int_0^\tau \int_{\Gamma^\varepsilon} |\partial_t r_f^\varepsilon|^2 \, d\gamma \, dt + \frac{\delta\varepsilon}{2} \int_0^\tau \int_{\Gamma^\varepsilon} |l^\varepsilon|^4 \, d\gamma \, dt \\
&\le \frac{b_1^2 \varepsilon}{2\delta} \int_0^\tau \int_{\Gamma^\varepsilon} |\partial_t r_f^\varepsilon|^2 \, d\gamma \, dt + \frac{\delta}{2} \sup_{[0,T]} \int_{\Omega^\varepsilon} |l^\varepsilon|^2 \, dx + \frac{\delta\varepsilon^2}{2} \sup_{[0,T]} \int_{\Omega^\varepsilon} |\nabla l^\varepsilon|^2 \, dx.
\end{aligned}$$

Using estimate (12) and the positivity of b^ε and r_f^ε we obtain

$$\int_0^T \int_{\Omega^\varepsilon} |\partial_t l^\varepsilon|^2 \, dx \, dt + \sup_{[0,T]} \int_{\Omega^\varepsilon} |\nabla l^\varepsilon|^2 \, dx \le C. \qquad \square$$

To obtain a priori estimates for functions defined in the domain independent of ε, we extend functions l^ε defined on Ω^ε to functions $\bar{l}^\varepsilon$ defined on the whole Ω.

3.4. Extension of l^ε. Since l^ε is defined only on Ω^ε, we extend it onto Ω; see [8] or [19] for the proof.

LEMMA 3.5. 1. *For $l \in H^1(Y)$ there exists an extension $\tilde{l}$ to Z, such that*

$$\|\tilde{l}\|_{L^2(Z)} \leq c\|l\|_{L^2(Y)} \quad \text{and} \quad \|\nabla \tilde{l}\|_{L^2(Z)} \leq c\|\nabla l\|_{L^2(Y)}.$$

2. *For $l^\varepsilon \in H^1(\Omega^\varepsilon)$ there exists an extension $\tilde{l}^\varepsilon$ to Ω, such that*

$$\|\tilde{l}^\varepsilon\|_{H^1(\Omega)} \leq c\|l^\varepsilon\|_{H^1(\Omega^\varepsilon)}.$$

Remark 3.1. For $l^\varepsilon \in L^2(0,T;H^1(\Omega^\varepsilon))$ we define $\bar{l}^\varepsilon(\cdot,t) := \tilde{l}^\varepsilon(\cdot,t)$ for a.a. t. Since the extension operator is linear, then $\bar{l}^\varepsilon \in L^2(0,T;H^1(\Omega))$.

We identify l^ε with the extension $\bar{l}^\varepsilon$. For the extended functions, we obtain a priori estimate of the supremum norm of l^ε.

LEMMA 3.6. *For any solution of problem* (1)–(6), *the following estimate holds:*

$$||l^\varepsilon||_{L^\infty((0,T)\times\Omega)} \leq C, \tag{13}$$

where C is a constant independent on ε.

Estimate (13) follows from the nonnegativity of l^ε, r_f^ε, r_b^ε, the boundedness of r_b^ε and l_0, and the estimate in Lemma 3.5; see Theorem 6.40 in [26] (for the sketch of proof see Appendix 6.1).

4. Convergence of solutions of microscopic problem.

4.1. Convergence of l^ε, r_f^ε, and r_b^ε. To show the convergence results we apply the method of two-scale convergence, introduced in [2] and [36], and extended further in [3, 37]. The definition and theorems concerning the two-scale convergence, used in this section are outlined in Appendix 6.2.

To show the compactness of l^ε we use the following Hilbert space.

DEFINITION 4.1 (see [47]). *Let $W^{\beta,2}(\Omega)$ with $\beta \in \mathbb{R}$, $\beta > 0$ be a Hilbert space defined as the completion of $C^\infty(\Omega)$ with respect to the norm*

$$\|u\|_{W^{\beta,2}(\Omega)} = \|u\|_{W^{k,2}(\Omega)} + \Big\{\int_\Omega \int_\Omega \frac{|u(x)-u(y)|^2}{|x-y|^{n+2(\beta-k)}} dxdy\Big\}^{\frac{1}{2}},$$

where $k = [\beta]$.

LEMMA 4.2. 1. *For a function $v^\varepsilon \in H^1(\Omega^\varepsilon)$ the following estimate holds:*

$$\varepsilon \int_{\Gamma^\varepsilon} |v^\varepsilon|^2 \, d\gamma_x \leq C \int_{\Omega^\varepsilon} |v^\varepsilon|^2 dx + C\varepsilon^2 \int_{\Omega^\varepsilon} |\nabla v^\varepsilon|^2 dx,$$

where C is a constant independent on ε.

2. *For a function $v^\varepsilon \in W^{\beta,2}(\Omega^\varepsilon)$, where $\frac{1}{2} < \beta < 1$, the following estimate holds:*

$$\varepsilon \int_{\Gamma^\varepsilon} |v^\varepsilon|^2 \, d\gamma_x \leq C \int_{\Omega^\varepsilon} |v^\varepsilon|^2 dx + C\varepsilon^{2\beta} \int_{\Omega^\varepsilon} \int_{\Omega^\varepsilon} \frac{|v^\varepsilon(x_1) - v^\varepsilon(x_2)|^2}{|x_1 - x_2|^{n+2\beta}} dx_1 dx_2,$$

where C is a constant independent on ε.

Proof. 1. For the proof see [19, Lemma 3].

2. For a function $v \in W^{\beta,2}(Y)$ the trace theorem implies

$$\int_\Gamma |v|^2 \, d\gamma_y \leq C \int_Y |v|^2 dy + C \int_Y \int_Y \frac{|v(y_1) - v(y_2)|^2}{|y_1 - y_2|^{n+2\beta}} dy_1 dy_2.$$

Changing variables, $y = x/\varepsilon$, we obtain

$$\int_{\varepsilon\Gamma_i} |v^\varepsilon|^2 \frac{d\gamma_x}{\varepsilon^{n-1}} \le C \int_{\varepsilon Y_i} |v^\varepsilon|^2 \frac{dx}{\varepsilon^n} + C \int_{\varepsilon Y_i}\int_{\varepsilon Y_i} \frac{|v^\varepsilon(x_1) - v^\varepsilon(x_2)|^2}{|x_1 - x_2|^{n+2\beta}} \varepsilon^{n+2\beta} \frac{dx_1}{\varepsilon^n}\frac{dx_2}{\varepsilon^n}.$$

Multiplying the inequality side by side with ε^{-n} and summing up over i from 1 to N implies the estimate of the lemma. ∎

Using a priori estimates derived in section 3.3 and the concept of the two-scale convergence, we obtain the following compactness result.

LEMMA 4.3. *There exist functions l, r_f, and r_b such that*

1. $l^\varepsilon \rightharpoonup l$ *in* $L^2(0,T;H^1(\Omega))$, $\partial_t l^\varepsilon \rightharpoonup \partial_t l$ *in* $L^2((0,T)\times\Omega)$, $l^\varepsilon \overset{*}{\rightharpoonup} l$ *in* $L^\infty((0,T)\times\Omega)$,
2. $l^\varepsilon \to l$ *in* $L^2(0,T;W^{\beta,2}(\Omega))$ *for* $\frac{1}{2} < \beta < 1$ *and* $\lim_{\varepsilon\to 0} ||l^\varepsilon - l||_{L^2((0,T)\times\Gamma^\varepsilon)} = 0$,
3. $l^\varepsilon \to l$ *two-scale,* $\nabla l^\varepsilon \to \nabla_x l + \nabla_y l_1$ *two-scale,* $l_1 \in L^2((0,T)\times\Omega; H^1_{per}(Z)/\mathbb{R})$,
4. $r_f^\varepsilon \to r_f$, $r_b^\varepsilon \to r_b$ *two-scale and* $r_f, r_b \in L^\infty((0,T)\times\Omega\times\Gamma)$,
5. $\partial_t r_f^\varepsilon \to \partial_t r_f$, $\partial_t r_b^\varepsilon \to \partial_t r_b$ *two-scale, and* $\partial_t r_f, \partial_t r_b \in L^2((0,T)\times\Omega\times\Gamma)$.

Proof. From the a priori estimates in Lemma 3.4, we obtain weak convergence $l^\varepsilon \rightharpoonup l$ in $L^2(0,T;H^1(\Omega))$, $\partial_t l^\varepsilon \rightharpoonup \partial_t l$ in $L^2((0,T)\times\Omega)$, and $l^\varepsilon \overset{*}{\rightharpoonup} l$ in $L^\infty((0,T)\times\Omega)$.

To obtain strong convergence of l^ε in $L^2((0,T),W^{\beta,2}(\Omega))$, $\frac{1}{2} < \beta < 1$, we use the compact embedding of $W^{\beta,2}(\Omega)$ in $H^1(\Omega)$ and apply the Lions–Aubin lemma [27] with $B = W^{\beta,2}(\Omega)$. Applying Lemma 4.2 we obtain the inequality

$$\|l^\varepsilon\|^2_{\Gamma^\varepsilon} \le c\|l^\varepsilon\|^2_{W^{\beta,2}(\Omega^\varepsilon)}.$$

It follows that

$$\|l^\varepsilon - l\|_{L^2((0,T)\times\Gamma^\varepsilon)} \le c\|l^\varepsilon - l\|^2_{L^2(0,T;W^{\beta,2}(\Omega^\varepsilon))} \le c\|l^\varepsilon - l\|^2_{L^2(0,T;W^{\beta,2}(\Omega))} \to 0 \text{ for } \varepsilon \to 0.$$

Since l^ε is bounded in $L^2(0,T;H^1(\Omega))$, the compactness theorem (see Theorem 6.3 in Appendix 6.2) implies the two-scale convergence of l^ε to the same function l and the existence of a function $l_1 \in L^2((0,T)\times\Omega; H^1_{per}(Z)/\mathbb{R})$ such that, up to a subsequence, ∇l^ε two-scale converges to $\nabla_x l(x) + \nabla_y l_1(x,y)$.

Invoking Theorem 6.5 (see Appendix 6.2) we obtain the two-scale convergence of r_f^ε and r_b^ε to functions in $L^\infty((0,T)\times\Omega\times\Gamma)$. Due to $||\partial_t r_f^\varepsilon||_{L^2((0,T)\times\Gamma^\varepsilon)} \le C$ and [37, Theorem 2.2], we conclude that $\partial_t r_f^\varepsilon \to v$ two-scale and $v \in L^2((0,T)\times\Omega\times\Gamma)$. Then

$$\int_0^T \int_{\Gamma\times\Omega} v\,\phi\,dx\,d\gamma_y\,dt = \lim_{\varepsilon\to 0} \int_0^T \int_{\Gamma^\varepsilon} \partial_t r_f^\varepsilon\,\phi\,d\gamma_x\,dt$$

$$= -\lim_{\varepsilon\to 0} \int_0^T \int_{\Gamma^\varepsilon} r_f^\varepsilon\,\partial_t\phi\,d\gamma_x\,dt = -\int_0^T \int_{\Gamma\times\Omega} r_f\,\partial_t\phi\,dx\,d\gamma_y\,dt.$$

Consequently, we conclude that $v = \partial_t r_f$. Analogously we obtain the two-scale convergence of $\partial_t r_b^\varepsilon$ to $\partial_t r_b$. ∎

4.2. Macroscopic equations.

THEOREM 4.4. *As $\varepsilon \to 0$, the sequence of solutions of the microscopic problem* (1)–(6) *converges to the weak solution* (l, r_f, r_b), $l \in H^1(0,T;L^2(\Omega))$, $l \in L^2(0,T;H^1(\Omega))$, $l \in L^\infty((0,T)\times\Omega)$, r_f, $r_b \in H^1(0,T;L^2(\Omega\times\Gamma))$, r_f, $r_b \in L^\infty((0,T)\times\Omega\times\Gamma)$,

of the following macroscopic problem:

(14)
$$\begin{cases} \partial_t l(t,x) = -\frac{1}{|Y|}\int_\Gamma (b(t,y) r_f(t,x,y) l(t,x) - d(t,y) r_b(t,x,y)) d\gamma_y \\ \qquad\qquad + (\nabla(S(t)\nabla l(t,x))) + \tilde{p}_l(t, l(t,x)) - \tilde{\mu}_l(t) l(t,x), & t>0, x\in\Omega, \\ \nabla l(t,x)\cdot\nu = 0, & t>0, x\in\Gamma^N, \\ l(t,x) = l_0(x), & t=0, x\in\Omega, \end{cases}$$

(15)
$$\begin{cases} \partial_t r_f(t,x,y) = p_r(t,y,r_b(t,x,y)) - b(t,y) r_f(t,x,y) l(t,x) \\ \qquad\qquad + d(t,y) r_b(t,x,y) - \mu_f(t,y) r_f(t,x,y), & y\in\Gamma, x\in\Omega, \\ \partial_t r_b(t,x,y) = b(t,y) r_f(t,x,y) l(t,x) - d(t,y) r_b(t,x,y) \\ \qquad\qquad - \mu_b(t,y) r_b(t,x,y), & y\in\Gamma, x\in\Omega, \\ r_f(t,x,y) = r_{f\,0}(x,y), & t=0, y\in\Gamma, x\in\Omega, \\ r_b(t,x,y) = r_{b0}(x,y), & t=0, y\in\Gamma, x\in\Omega, \end{cases}$$

where $\tilde{\mu}_l(t) = \frac{1}{|Y|}\int_Y \mu_l(t,y)\,dy$, $\tilde{p}(t,l) = \frac{1}{|Y|}\int_Y p(t,y,l)\,dy$, *and the matrix* S *is defined as* $s_{ij} = \frac{1}{|Y|}\sum_{k=1}^3 \int_Y (D_{ij}(t,y) + D_{ik}(t,y)\partial_{y_k} w_j)\,dy$ *with* w_i *being the solutions of the cell problem*

$$-\nabla_y(D(t,y)\nabla_y w_i) = \sum_{k=1}^3 \partial_{y_k} D_{ki}(t,y) \quad in\ Y, \qquad -D(t,y)\frac{\partial w_i}{\partial \nu} = \sum_{k=1}^3 D_{ki}(t,y)\nu_k \quad on\ \Gamma.$$

Proof. To derive a limit equation for l^ε we apply a standard two-scale convergence method and strong convergence of l^ε. Using in (7) a test function of the form $\phi(t,x) = \psi_0(t,x) + \varepsilon\psi_1(t,x,\frac{x}{\varepsilon})$, $\psi_0 \in C^\infty((0,T)\times\Omega)$, $\psi_1 \in C^\infty((0,T)\times\Omega; C^\infty_{per}(Z))$ and passing to the two-scale limit applying Lemma 4.3 yields

$$\begin{aligned} &|Y|\int_0^T\int_\Omega \partial_t l \psi_0(t,x)\,dx\,dt + |Y|\int_0^T\int_\Omega \tilde{\mu}_l(t)\, l(t,x)\,\psi_0(t,x)\,dx\,dt \\ &+ \int_0^T\int_\Omega\int_Y D(t,y)(\nabla_x l(t,x) + \nabla_y l_1(t,x,y))(\nabla_x\psi_0 + \nabla_y\psi_1)\,dy\,dx\,dt \\ &= -\int_0^T\int_\Omega\int_\Gamma [b(t,y) r_f(t,x,y) l(t,x) - d(t,y) r_b(t,x,y)]\psi_0(t,x)\,d\gamma_y\,dx\,dt \\ &+ |Y|\int_0^T\int_\Omega \tilde{p}_l(t,l)\,\psi_0\,dx\,dt. \end{aligned}$$

To show the convergence of the nonlinear term $b^\varepsilon r_f^\varepsilon l^\varepsilon$ of the boundary integral, we rewrite this integral as a sum of two integrals,

$$\begin{aligned} &\varepsilon\int_0^T\int_{\Gamma^\varepsilon} b^\varepsilon r_f^\varepsilon l^\varepsilon \left(\psi_0(t,x) + \varepsilon\psi_1\left(t,x,\frac{x}{\varepsilon}\right)\right) d\gamma_x\,dt \\ &\quad = \varepsilon\int_0^T\int_{\Gamma^\varepsilon} b^\varepsilon r_f^\varepsilon l \left(\psi_0(t,x) + \varepsilon\psi_1\left(t,x,\frac{x}{\varepsilon}\right)\right) d\gamma_x\,dt \\ &\qquad + \varepsilon\int_0^T\int_{\Gamma^\varepsilon} b^\varepsilon r_f^\varepsilon (l^\varepsilon - l) \left(\psi_0(t,x) + \varepsilon\psi_1\left(t,x,\frac{x}{\varepsilon}\right)\right) d\gamma_x\,dt. \end{aligned}$$

The first integral converges to $\int_0^T \int_\Omega \int_\Gamma b(t,y) r_f(t,x,y) l(t,x) \psi_0(t,x)\, d\gamma_y\, dx\, dt$ due to the two-scale convergence of r_f^ε. Since $||l^\varepsilon - l||_{L^2((0,T)\times\Gamma^\varepsilon)} \to 0$ as $\varepsilon \to 0$, we obtain for the second integral

$$\varepsilon \int_0^T \int_{\Gamma^\varepsilon} b^\varepsilon r_f^\varepsilon (l^\varepsilon - l) \left(\psi_0(t,x) + \varepsilon \psi_1\left(t,x,\frac{x}{\varepsilon}\right)\right) d\gamma_x\, dt$$

$$\leq \varepsilon \left(\int_0^T \int_{\Gamma^\varepsilon} |b^\varepsilon r_f^\varepsilon \psi_0|^2 d\gamma_x\, dt\right)^{1/2} \left(\int_0^T \int_{\Gamma^\varepsilon} |l^\varepsilon - l|^2 d\gamma_x\, dt\right)^{1/2}$$

$$+ \varepsilon^2 \left(\int_0^T \int_{\Gamma^\varepsilon} |b^\varepsilon r_f^\varepsilon \psi_1|^2 d\gamma_x\, dt\right)^{1/2} \left(\int_0^T \int_{\Gamma^\varepsilon} |l^\varepsilon - l|^2 d\gamma_x\, dt\right)^{1/2} \to 0 \quad \text{as} \quad \varepsilon \to 0.$$

To determinate the unknown function $l_1 \in L^2((0,T)\times\Omega; H^1_{\text{per}}(Y)/\mathbb{R})$, we set $\psi_0 = 0$ and obtain the equation

$$\int_0^T \int_{\Omega\times Y} D(t,y)(\nabla_x l(t,x) + \nabla_y l_1(t,x,y)) \nabla_y \psi_1(t,x,y)\, dt\, dx\, dy = 0$$

for all ψ_1. From this it follows that l_1 depends linearly on $\nabla_x l$, and it can be written in the form

$$l_1 = \sum_{i=1}^n \frac{\partial l}{\partial x_i} \cdot w_i,$$

where the functions w_i are defined as solutions of the cell problem

$$-\nabla(D(t,y)\nabla w_i) = \sum_{k=1}^3 \partial_{y_k} D_{ki}(t,y) \text{ in } Y, \quad -D(t,y)\frac{\partial w_i}{\partial \nu} = \sum_{k=1}^3 D_{ki}(t,y)\, \nu_k \text{ on } \Gamma.$$

Next, setting $\psi_1 = 0$, we obtain

$$\int_0^T \int_\Omega \int_Y \sum_{i,j=1}^n D_{ij}(t,y)(\partial_{x_i} l(t,x) + \sum_{k=1}^n \partial_{y_i} w_k \partial_{x_k} l(t,x)) \partial_{x_j} \psi_0(t,x)\, dy\, dx\, dt$$

$$= |Y| \int_0^T \int_\Omega \sum_{i,j=1}^n s_{ij} \partial_{x_i} \psi_0(t,x) \partial_{x_j} l(t,x)\, dy\, dx\, dt$$

with $s_{ij} = \frac{1}{|Y|} \sum_{k=1}^3 \int_Y (D_{ij}(t,y) + D_{ik}(t,y)\partial_{y_k} w_j)\, dy$.

The difficulty arises in passing to the limit in nonlinear terms in the ordinary differential equations on the surface of microstructures. We have to show that $p_r^\varepsilon(t,x,r_b^\varepsilon(t,x)) \to p_r(t,y,r_b(t,x,y))$ in the two-scale sense. To cope with this difficulty we apply the unfolding method (periodic modulation), developed in [7, 5, 6]. Following [5] and [6], we define a dilation operator.

DEFINITION 4.5. *For a given $\varepsilon > 0$, we define a dilation operator D^ε mapping measurable functions on $(0,T)\times\Gamma^\varepsilon$ to measurable functions on $(0,T)\times\Omega\times\Gamma$ by*

$$D^\varepsilon u(t,x,y) = u(t, c^\varepsilon(x) + \varepsilon y), \quad y \in \Gamma, \quad (t,x) \in (0,T)\times\Omega,$$

where $c^\varepsilon(x)$ denotes the lattice translation point of the ε-cell domain containing x, $c^\varepsilon(x) = \varepsilon[\frac{x}{\varepsilon}]$. We extend $D^\varepsilon u$ from Γ to $\bigcup_k(\Gamma + k)$ periodically.

Remark 4.1. The dilation operator D^ε is well defined for all $(t, x, y) \in (0, T) \times \Omega \times \Gamma$ under the assumption on the geometry of domain Ω^ε (cf. Remark 2.1).

To proceed, we have to establish the link between the two-scale convergence and the weak convergence of the dilated sequences. Following [6], we formulate the lemma on the convergence of $D^\varepsilon u^\varepsilon : (0, T) \times \Omega \times \Gamma \to \mathbb{R}$. We define $L^2_{\text{per}}(\Gamma)$ as the space of functions $f \in L^2(\Gamma)$ defined on Γ and periodically extended to $\Gamma^* = \bigcup_k(\Gamma + k)$.

LEMMA 4.6. *If $D^\varepsilon u^\varepsilon \rightharpoonup u^*$ weakly in $L^2((0, T) \times \Omega; L^2_{per}(\Gamma))$ and $u^\varepsilon \to u$ two-scale, then $u^* = u$ a.e. in $(0, T) \times \Omega \times \Gamma$.*

Proof. Let u^* be a weak limit of $D^\varepsilon u^\varepsilon$. Then, for a test function $\psi(t, x)h(y)$, where $\psi \in C_0^\infty((0, T) \times \Omega)$ and $h \in C^\infty_{\text{per}}(\Gamma)$, we obtain

$$\int_0^T \int_{\Omega\times\Gamma} D^\varepsilon u^\varepsilon(t, x, y)\psi(t, x)h(y)d\gamma_y dxdt$$

$$\to \int_0^T \int_{\Omega\times\Gamma} u^*(t, x, y)\psi(t, x)h(y)d\gamma_y dxdt \quad \text{as } \varepsilon \to 0.$$

On the other hand, we have

$$\int_0^T \int_{\Omega\times\Gamma} D^\varepsilon u^\varepsilon(t, x, y)\psi(t, x)h(y)d\gamma_y dxdt$$

$$= \int_0^T \int_{\Omega\times\Gamma} u^\varepsilon(t, \varepsilon y + c^\varepsilon(x))\psi(t, x)h(y)d\gamma_y dxdt$$

$$= \sum_{k=1}^N \int_0^T \int_{\varepsilon(Z+k)} \int_\Gamma u^\varepsilon(t, \varepsilon y + c^\varepsilon(x))\psi(t, x)h(y)d\gamma_y dxdt.$$

Changing variables $z = \varepsilon(y + k)$, where $c^\varepsilon(x) = \varepsilon[\frac{x}{\varepsilon}] = \varepsilon k$, and using the periodicity of h, we obtain

$$\int_0^T \sum_{k=1}^N \varepsilon^{-2} \int_{\varepsilon(\Gamma+k)} u^\varepsilon(t, z)h\left(\frac{z}{\varepsilon}\right) \int_{\varepsilon(Z+k)} \psi(t, x)\, dx\, d\gamma_z\, dt$$

$$= \varepsilon \int_0^T \sum_{k=1}^N \int_{\varepsilon(\Gamma+k)} u^\varepsilon(t, z)h\left(\frac{z}{\varepsilon}\right) \psi(t, z)d\gamma_z dt + c\varepsilon^2$$

$$\to \int_0^T \int_\Omega \int_\Gamma u(t, x, y)h(y)\psi(t, x)d\gamma_y dxdt,$$

since from the continuity of ψ we have the estimate

$$|\varepsilon^{-3} \int_{\varepsilon(Z+k)} \big(\psi(t, x) - \psi(t, z)\big)\, dx| \le c\varepsilon \text{ for } z \in \varepsilon(\Gamma + k).$$

Therefore, we conclude that $u^* = u$ a.e. in $(0, T) \times \Omega \times \Gamma$. ∎

In analogy to the above lemma and Lemma 2 in [5], we can prove the following properties of the dilation operator for oscillating surfaces.

LEMMA 4.7. *For* $u \in L^2((0,T)\times\Gamma^\varepsilon)$

$$||D^\varepsilon u||_{L^2(\Omega\times\Gamma)} = ||u||_{L^2(\Gamma^\varepsilon)}.$$

If $u \in L^2(\Omega\times\Gamma)$ *is constant in* y, *then* $D^\varepsilon u \to u$ *as* $\varepsilon \to 0$ *strongly in* $L^2(\Omega\times\Gamma)$.

Changing variables, $\Gamma^\varepsilon \ni x \to \varepsilon y + c^\varepsilon(x)$, $c^\varepsilon(x) = \varepsilon k$ for $x \in \Gamma^\varepsilon$, we obtain equations on the fixed domain $(0,T)\times\Omega\times\Gamma$,

$$\frac{\partial}{\partial t} D^\varepsilon {r_f}^\varepsilon(t,x,y) = -\mu_f(t,y) D^\varepsilon {r_f}^\varepsilon(t,x,y) + p_r(t,y,D^\varepsilon r_b^\varepsilon(t,x,y))$$

$$- b(t,y) D^\varepsilon {r_f}^\varepsilon(t,x,y) D^\varepsilon l^\varepsilon(t,x,y) + d(t,y) D^\varepsilon {r_b}^\varepsilon(t,x,y),$$

$$\frac{\partial}{\partial t} D^\varepsilon {r_b}^\varepsilon(t,x,y) = -\mu_b(t,y) D^\varepsilon {r_b}^\varepsilon(t,x,y)$$

$$+ b(t,y) D^\varepsilon {r_f}^\varepsilon(t,x,y) D^\varepsilon l^\varepsilon(t,x,y) - d(t,y) D^\varepsilon {r_b}^\varepsilon(t,x,y).$$

Applying the estimates for r_f^ε and r_b^ε, we obtain the estimates for $D^\varepsilon {r_f}^\varepsilon$ and $D^\varepsilon {r_b}^\varepsilon$ and the weak convergence of $D^\varepsilon {r_f}^\varepsilon$ to r_f and $D^\varepsilon {r_b}^\varepsilon$ to r_b in $L^2((0,T)\times\Omega; L^2_{\mathrm{per}}(\Gamma))$ (see Lemma 4.6). Since $\sup_{[0,T]\times\overline{\Omega}} |l^\varepsilon| \le C$, we conclude that $\sup_{[0,T]\times\Omega\times\Gamma} |D^\varepsilon l^\varepsilon| \le C$.

Now we prove the strong convergence of $D^\varepsilon {r_f}^\varepsilon$ and $D^\varepsilon {r_b}^\varepsilon$ in $L^2((0,T)\times\Omega; L^2_{\mathrm{per}}(\Gamma))$. For this we show that $D^\varepsilon {r_f}^\varepsilon$ and $D^\varepsilon {r_b}^\varepsilon$ are Cauchy sequences. We consider the equations for $D^{\varepsilon_n} {r_f}^{\varepsilon_n} - D^{\varepsilon_m} {r_f}^{\varepsilon_m}$ and $D^{\varepsilon_n} {r_b}^{\varepsilon_n} - D^{\varepsilon_m} {r_b}^{\varepsilon_m}$, with $n > m$, multiply them side by side with $D^{\varepsilon_n} {r_f}^{\varepsilon_n} - D^{\varepsilon_m} {r_f}^{\varepsilon_m}$ and $D^{\varepsilon_n} {r_b}^{\varepsilon_n} - D^{\varepsilon_m} {r_b}^{\varepsilon_m}$, respectively, and integrate over $\Omega\times\Gamma$.

$$\frac{\partial}{\partial t}\int_{\Omega\times\Gamma} |D^{\varepsilon_n} {r_f}^{\varepsilon_n} - D^{\varepsilon_m} {r_f}^{\varepsilon_m}|^2 dx d\gamma = -\int_{\Omega\times\Gamma} \mu_f(t,y) |D^{\varepsilon_n} {r_f}^{\varepsilon_n} - D^{\varepsilon_m} {r_f}^{\varepsilon_m}|^2 dx d\gamma$$

$$+ \int_{\Omega\times\Gamma} (p_r(t,y,D^{\varepsilon_n} {r_b}^{\varepsilon_n}) - p_r(t,y,D^{\varepsilon_m} r_b^{\varepsilon_m}))(D^{\varepsilon_n} {r_f}^{\varepsilon_n} - D^{\varepsilon_m} {r_f}^{\varepsilon_m}) dx d\gamma$$

$$- \int_{\Omega\times\Gamma} b(t,y)(D^{\varepsilon_n} {r_f}^{\varepsilon_n} D^{\varepsilon_n} l^{\varepsilon_n} - D^{\varepsilon_m} {r_f}^{\varepsilon_m} D^{\varepsilon_m} l^{\varepsilon_m})(D^{\varepsilon_n} {r_f}^{\varepsilon_n} - D^{\varepsilon_m} {r_f}^{\varepsilon_m}) dx d\gamma$$

$$- \int_{\Omega\times\Gamma} d(t,y)(D^{\varepsilon_n} {r_b}^{\varepsilon_n} - D^{\varepsilon_m} {r_b}^{\varepsilon_m})(D^{\varepsilon_n} {r_f}^{\varepsilon_n} - D^{\varepsilon_m} {r_f}^{\varepsilon_m}) dx d\gamma,$$

$$\frac{\partial}{\partial t}\int_{\Omega\times\Gamma} |D^{\varepsilon_n} {r_b}^{\varepsilon_n} - D^{\varepsilon_m} {r_b}^{\varepsilon_m}|^2 dx d\gamma = -\int_{\Omega\times\Gamma} \mu_b(t,y) |D^{\varepsilon_n} {r_b}^{\varepsilon_n} - D^{\varepsilon_m} {r_b}^{\varepsilon_m}|^2 dx d\gamma$$

$$+ \int_{\Omega\times\Gamma} b(t,y)(D^{\varepsilon_n} {r_f}^{\varepsilon_n} D^{\varepsilon_n} l^{\varepsilon_n} - D^{\varepsilon_m} {r_f}^{\varepsilon_m} D^{\varepsilon_m} l^{\varepsilon_m})(D^{\varepsilon_n} {r_b}^{\varepsilon_n} - D^{\varepsilon_m} {r_b}^{\varepsilon_m}) dx d\gamma$$

$$- \int_{\Omega\times\Gamma} d(t,y) |D^{\varepsilon_n} {r_b}^{\varepsilon_n} - D^{\varepsilon_m} {r_b}^{\varepsilon_m}|^2 \, dx d\gamma.$$

Using the Young inequality, we obtain

$$
\begin{aligned}
(16)\qquad & \frac{\partial}{\partial t}\int_{\Omega\times\Gamma}|D^{\varepsilon_n}{r_f}^{\varepsilon_n}-D^{\varepsilon_m}{r_f}^{\varepsilon_m}|^2dxd\gamma \\
&\le C_1\int_{\Omega\times\Gamma}|D^{\varepsilon_n}{r_f}^{\varepsilon_n}-D^{\varepsilon_m}{r_f}^{\varepsilon_m}|^2dxd\gamma \\
&\quad + C_2\int_{\Omega\times\Gamma}|D^{\varepsilon_n}{r_b}^{\varepsilon_n}-D^{\varepsilon_m}{r_b}^{\varepsilon_m}|^2dxd\gamma \\
&\quad + b_1\sup_{(0,T)\times\Omega\times\Gamma}|D^{\varepsilon_n}l^{\varepsilon_n}|\int_{\Omega\times\Gamma}|D^{\varepsilon_n}{r_f}^{\varepsilon_n}-D^{\varepsilon_m}{r_f}^{\varepsilon_m}|^2dxd\gamma \\
&\quad + C_3\int_{\Omega\times\Gamma}|D^{\varepsilon_n}l^{\varepsilon_n}-D^{\varepsilon_m}l^{\varepsilon_m}|^2dxd\gamma,
\end{aligned}
$$

$$
\begin{aligned}
(17)\qquad & \frac{\partial}{\partial t}\int_{\Omega\times\Gamma}|D^{\varepsilon_n}{r_b}^{\varepsilon_n}-D^{\varepsilon_m}{r_b}^{\varepsilon_m}|^2dxd\gamma \\
&\le C_4\int_{\Omega\times\Gamma}|D^{\varepsilon_n}{r_b}^{\varepsilon_n}-D^{\varepsilon_m}{r_b}^{\varepsilon_m}|^2dxd\gamma \\
&\quad + b_1\sup_{(0,T)\times\Omega\times\Gamma}|D^{\varepsilon_n}l^{\varepsilon_n}|\int_{\Omega\times\Gamma}|D^{\varepsilon_n}{r_f}^{\varepsilon_n}-D^{\varepsilon_m}{r_f}^{\varepsilon_m}|^2dxd\gamma \\
&\quad + C_5\int_{\Omega\times\Gamma}|D^{\varepsilon_n}l^{\varepsilon_n}-D^{\varepsilon_m}l^{\varepsilon_m}|^2dxd\gamma.
\end{aligned}
$$

Due to Lemma 4.7 and strong convergence of l^ε on Γ^ε, we obtain

$$
\int_0^T\int_{\Omega\times\Gamma}|D^\varepsilon l^\varepsilon-D^\varepsilon l|^2d\gamma dxdt=\varepsilon\int_0^T\int_{\Gamma^\varepsilon}|l^\varepsilon-l|^2d\gamma_x dt\le C\varepsilon.
$$

Therefore, since $D^{\varepsilon_n}l\to l$ strongly in $L^2((0,T)\times\Omega\times\Gamma)$ (see Lemma 4.7),

$$
\begin{aligned}
&\int_0^T\int_{\Omega\times\Gamma}|D^{\varepsilon_n}l^{\varepsilon_n}-D^{\varepsilon_m}l^{\varepsilon_m}|^2d\gamma dxdt \\
&\quad\le\int_0^T\int_{\Omega\times\Gamma}\Big(|D^{\varepsilon_n}l^{\varepsilon_n}-D^{\varepsilon_n}l|^2+|D^{\varepsilon_n}l-l|^2\Big)d\gamma dxdt \\
&\qquad+\int_0^T\int_{\Omega\times\Gamma}\Big(|D^{\varepsilon_m}l-l|^2+|D^{\varepsilon_m}l^{\varepsilon_m}-D^{\varepsilon_m}l|^2\Big)d\gamma dxdt \\
&\quad\le\varepsilon_n\int_0^T\int_{\Gamma^{\varepsilon_n}}|l^{\varepsilon_n}-l|^2d\gamma_x dt \\
&\qquad\varepsilon_m\int_0^T\int_{\Gamma^{\varepsilon_m}}|l^{\varepsilon_m}-l|^2d\gamma_x dt+\int_0^T\int_{\Omega\times\Gamma}\Big(|D^{\varepsilon_n}l-l|^2+|D^{\varepsilon_m}l-l|^2\Big)d\gamma dxdt \\
&\quad\le C(\varepsilon_n+\varepsilon_m).
\end{aligned}
$$

We add (16) and (17) side by side and integrate with respect to time. Using additionally the boundedness of $D^\varepsilon l^\varepsilon$ on $(0,T)\times\Omega\times\Gamma$, we obtain

$$||D^{\varepsilon_n} r_f{}^{\varepsilon_n} - D^{\varepsilon_m} r_f{}^{\varepsilon_m}||^2 + ||D^{\varepsilon_n} r_b{}^{\varepsilon_n} - D^{\varepsilon_m} r_b{}^{\varepsilon_m}||^2$$
$$\le C_1 \int_0^\tau \big(||D^{\varepsilon_n} r_f{}^{\varepsilon_n} - D^{\varepsilon_m} r_f{}^{\varepsilon_m}||^2 + ||D^{\varepsilon_n} r_b{}^{\varepsilon_n} - D^{\varepsilon_m} r_b{}^{\varepsilon_m}||^2\big) dt + C_2 \frac{1}{n},$$

where $C_1 = C_1(\sup_{(0,T)\times\Omega}|l^\varepsilon|, \sup_{(0,T)\times\Gamma}|\mu_f|, \sup_{(0,T)\times\Gamma}|\mu_b|, \sup_{(0,T)\times\Gamma}|b|, \sup_{(0,T)\times\Gamma}|d|, \sup_{(0,T)\times\Gamma^\varepsilon}|r_f^\varepsilon|)$. Then the Gronwall lemma yields

$$||D^{\varepsilon_n} r_f{}^{\varepsilon_n} - D^{\varepsilon_m} r_f{}^{\varepsilon_m}||_{L^2(\Omega\times\Gamma)} \le C\frac{1}{n},$$

$$||D^{\varepsilon_n} r_b{}^{\varepsilon_n} - D^{\varepsilon_m} r_b{}^{\varepsilon_m}||_{L^2(\Omega\times\Gamma)} \le C\frac{1}{n}.$$

Using strong convergence of $D^\varepsilon r_b{}^\varepsilon$, continuity of p_r, and weak convergence of $p_r(t,y,D^\varepsilon r_b^\varepsilon)$, which results from the boundedness of p_r, we obtain that $p_r(t,y,D^\varepsilon r_b^\varepsilon)$ weakly converges to $p_r(t,y,r_b(t,x,y))$ in $L^2((0,T)\times\Omega; L^2_{\mathrm{per}}(\Gamma))$.

Now we can take the two-scale limit in the equations on the boundary,

$$\varepsilon\int_0^T\int_{\Gamma^\varepsilon} \partial_t r_f^\varepsilon \psi_1\left(t,x,\frac{x}{\varepsilon}\right) d\gamma_x\, dt = \varepsilon\int_0^T\int_{\Gamma^\varepsilon} p_r^\varepsilon(t,x,r_b^\varepsilon(t,x))\psi_1\left(t,x,\frac{x}{\varepsilon}\right) d\gamma_x\, dt$$
$$+\varepsilon\int_0^T\int_{\Gamma^\varepsilon}\Big(-b^\varepsilon r_f^\varepsilon(t,x)l^\varepsilon(t,x) + d^\varepsilon(t,x)r_b^\varepsilon(t,x) - \mu_f^\varepsilon r_f^\varepsilon(t,x)\Big)\psi_1\left(t,x,\frac{x}{\varepsilon}\right) d\gamma_x\, dt,$$

$$\varepsilon\int_0^T\int_{\Gamma^\varepsilon} \partial_t r_b^\varepsilon(t,x)\psi_1\left(t,x,\frac{x}{\varepsilon}\right) d\gamma_x\, dt = \varepsilon\int_0^T\int_{\Gamma^\varepsilon} b^\varepsilon r_f^\varepsilon(t,x)l^\varepsilon(t,x)\psi_1\, d\gamma_x\, dt$$
$$+\varepsilon\int_0^T\int_{\Gamma^\varepsilon}\Big(-d^\varepsilon r_b^\varepsilon(t,x) - \mu_b^\varepsilon r_b^\varepsilon(t,x)\Big)\psi_1\left(t,x,\frac{x}{\varepsilon}\right) d\gamma_x\, dt.$$

The linear terms converge two-scale to their limit functions. The proof of convergence for the nonlinear term $b^\varepsilon r_f^\varepsilon(t,x)l^\varepsilon(t,x)$ is the same as in the equation for l^ε. Due to boundedness of p_r^ε and Lemma 4.6, $p_r^\varepsilon(t,x,r_b^\varepsilon)$ converges two-scale to $p_r(t,y,r_b(t,x,y))$. Therefore, we obtain the macroscopic equations for r_f and r_b. ∎

The uniqueness of the solution of the macroscopic problem can be proved in the same way as for the microscopic problem.

Remark 4.2. Properties of the macroscopic model: Using the framework of bounded invariant rectangles (see [44]) we can show that solutions of system (14)–(15) remain positive for positive initial conditions and that they are also uniformly bounded. This results from the assumption of the nonnegativity of the model parameters and their boundedness independent of time. Methods outlined in [44, Chapter 14] can be used without major modifications.

5. Discussion. In this work, using homogenization techniques, we studied the macroscopic limit of the microscopic model describing receptor-ligand dynamics on cell membranes and in the intercellular space. We tried to answer the question of how processes which take place in different "spaces," such as cells membranes, intercellular space, and also intracellular space, can be described by macroscopic models operating

in homogenized space. On one hand, this work provides a justification of previously proposed models, and on the other hand it is a starting point for further models.

Comparison of the macroscopic model (14)–(15) to the previously considered receptor-based model of the form

$$\begin{aligned}
\frac{\partial}{\partial t} r_f &= -\mu_f r_f + p_r(r_b) - b r_f l + d r_b, \\
\frac{\partial}{\partial t} r_b &= -\mu_b r_b + b r_f l - d r_b, \\
\frac{\partial}{\partial t} l &= \frac{1}{\gamma} \frac{\partial^2}{\partial x^2} l - \mu_l l - b r_f l + p_l(l) + d r_b,
\end{aligned} \tag{18}$$

defined on the macroscopic domain Ω, shows in which cases the "older" models can be derived from the microscopic description. Model (14)–(15) is equivalent to model (18) in the case when neither the model parameters nor the initial conditions for r_f and r_b depend on the surface variable y. It means that the processes described are homogeneous within each cell and that there is no heterogeneity in the dissociation or binding processes on the cell surfaces. For nonadherent cells one can consider receptor production, binding, dissociation, or decay to be uniformly distributed on the cell surface, which results in model coefficients being constant with respect to the surface variable y. Under such assumptions we obtain a macroscopic model, in which the integral in the equation for the ligands disappears and the only difference with respect to model (18) is that the kinetics are multiplied by a coefficient $\int_\Gamma d\gamma_y/|Y|$. However, there is now considerable evidence of the existence of lipid raft microdomains, called membrane rafts, which organize the membrane into specialized functional units [14, 15, 38, 39, 41]. Rafts were described mainly for T-cells and T-cell receptor [15, 39, 40], but now it is clear that they play an important role for many different receptor classes [13, 41]. There are observations that the structure of lipid rafts could control cellular processes such as signaling cascades [40, 43] and receptor synthesis and trafficking [21] as well as cell adhesion and migration [16]. Some membrane proteins are located preferentially on the raft domains, whereas others are excluded from them [14]. Such a situation corresponds to the nonhomogeneous initial distribution of receptors on the cell surface and also *de novo* production terms depending on the surface variable. Our studies show that in such a case the "older" type of receptor-based model is not relevant.

Another example of cells with nonhomogeneous membrane properties are adherent cells. In the case of adherent cells there are two types of polarity, top-bottom and front-back, and it is not easy for the ligand to get in contact with the bottom of the cell. One can imagine that receptors may be concentrated on the frontal end of the cell (this determines cell motility in the case of chemotaxis), and, therefore, all the receptor-ligand processes are nonhomogeneous within the membrane [22].

6. Appendix.

6.1. Supremum estimate for l^ε. We present here a sketch of the proof of Lemma 3.6 used in section 4.

LEMMA. *For any solution of problem* (1)–(6), *the following estimate holds:*

$$||l^\varepsilon||_{L^\infty((0,T)\times\Omega)} \leq C + 2k,$$

where C is a constant independent on ε and $k = \max\{1, \sup_\Omega |l_0|\}$.

Proof. To show the boundedness of l^ε we use the Moser iteration technique, described in the proof of [26, Theorem 6.15]. We choose as a test function $v = \psi(l^\varepsilon)(l^\varepsilon - k)_+$, where $\psi \geq 0$ is a bounded $C^1(\mathbb{R})$ function and which satisfies for $s > k$

$$0 \leq \frac{\psi'(s)(s-k)}{\psi(s)} \leq k_1.$$

Due to the fact that $l_0 \leq k$, we obtain

$$\begin{aligned}
(19) \qquad & \int_\Omega \int_0^{l^\varepsilon} \psi(s)(s-k)_+ \, ds\chi^\varepsilon \, dx + \int_0^\tau \int_\Omega (D^\varepsilon \nabla l^\varepsilon, \psi(l^\varepsilon)\nabla l^\varepsilon)\chi^\varepsilon \, dxdt \\
& \quad + \int_0^\tau \int_\Omega (D^\varepsilon \nabla l^\varepsilon, \psi'(l^\varepsilon)(l^\varepsilon - k)_+ \nabla l^\varepsilon)\chi^\varepsilon \, dxdt \\
& \quad + \int_0^\tau \int_\Omega \mu_l^\varepsilon l^\varepsilon \psi(l^\varepsilon)(l^\varepsilon - k)_+ \chi^\varepsilon dxdt \\
& = \int_0^\tau \int_\Omega p_l^\varepsilon(l^\varepsilon)\psi(l^\varepsilon)(l^\varepsilon - k)_+ \chi^\varepsilon \, dxdt \\
& \quad + \varepsilon \int_0^\tau \int_{\Gamma^\varepsilon} (d^\varepsilon r_b^\varepsilon - b^\varepsilon r_f^\varepsilon l^\varepsilon)\psi(l^\varepsilon)(l^\varepsilon - k)_+ d\gamma_x dt,
\end{aligned}$$

where χ is a characteristic function of Y periodically extended to Z^*, and $\chi^\varepsilon(x) = \chi(\frac{x}{\varepsilon})$. From the properties of ψ, we obtain

$$\int_k^s \psi(t)(t-k)dt \geq \frac{1}{2+k_1}\psi(s)(s-k) \quad \text{for} \quad s \geq k.$$

The third and fourth terms on the left-hand side of (19) are nonnegative; the third term on the right-hand side is nonpositive. Using $l^\varepsilon \geq k \geq 1$ and Lemma 4.2, we obtain the estimate

$$\begin{aligned}
& \varepsilon \int_0^\tau \int_{\Gamma^\varepsilon} \psi(l^\varepsilon)(l^\varepsilon - k)_+ d\gamma_x dt \\
& \quad \leq C \int_0^\tau \int_{\Omega^\varepsilon} \Big(\psi(l^\varepsilon)(l^\varepsilon - k)_+ + \varepsilon^2 \nabla(\psi(l^\varepsilon)(l^\varepsilon - k)_+) \Big) dxdt \\
& \quad \leq C(1 + \frac{1}{\delta})(1+k_1) \int_0^\tau \int_\Omega \psi(l^\varepsilon)|l^\varepsilon|^2 \chi^\varepsilon dxdt + C\varepsilon^2 \delta \int_0^\tau \int_\Omega \psi(l^\varepsilon)|\nabla l^\varepsilon|^2 \chi^\varepsilon dxdt.
\end{aligned}$$

Then boundedness of coefficients and sublinearity of p_l yields

$$\begin{aligned}
& \int_\Omega \psi(l^\varepsilon)(l^\varepsilon - k)^2 \chi^\varepsilon dx + d_0 \int_0^\tau \int_\Omega \psi(l^\varepsilon)|\nabla l^\varepsilon|^2 \chi^\varepsilon dxdt \\
& \qquad \leq C(1+k_1)^2 \int_0^\tau \int_\Omega \psi(l^\varepsilon)|l^\varepsilon|^2 \chi^\varepsilon dxdt.
\end{aligned}$$

Choosing $\psi(s) = (\min\{s, Z\}(1 - k/s)_+)^q$, where q, Z are positive constants, applying the Gronwall and Young inequalities, and taking $Z \to \infty$ leads to

$$\int_0^T \int_\Omega |l^\varepsilon|^{q+2} \chi^\varepsilon \, dxdt \leq C(q)|\Omega|Tk^{q+2}$$

for any positive q. Thus, for a fixed $q > 1$ we can choose $\psi(s) = s^{2q-2}(1-\frac{k}{s})_+^{(n+2)(q-1)}$ and conclude that

$$\int_\Omega (l^\varepsilon)^{2q}\left(1-\frac{k}{l^\varepsilon}\right)_+^{(n+2)q-n}\chi^\varepsilon dx + c(d_0)\int_0^\tau\int_\Omega (l^\varepsilon)^{2q-2}\left(1-\frac{k}{l^\varepsilon}\right)_+^{(n+2)(q-1)}|\nabla l^\varepsilon|^2\chi^\varepsilon dxdt$$

$$\leq Cq^2\int_0^\tau\int_\Omega (l^\varepsilon)^{2q}\left(1-\frac{k}{l^\varepsilon}\right)_+^{(n+2)(q-1)}\chi^\varepsilon dxdt.$$

Setting $h = (l^\varepsilon)^q((1-\frac{k}{l^\varepsilon})_+^{(n+2)q-n})^{1/2}$ gives $|\nabla h|^2 \leq c(n)q^2(l^\varepsilon)^{2q-2}(1-\frac{k}{l^\varepsilon})_+^{(n+2)(q-1)}$ $|\nabla l^\varepsilon|^2$. Using the property of extended function of l^ε, i.e., $||l^\varepsilon||_{H^1(\Omega)} \leq C||l^\varepsilon||_{H^1(\Omega^\varepsilon)}$, with a constant C independent of ε, yields

$$\sup_{(0,T)}\int_\Omega h^2 dx + \int_0^T\int_\Omega |\nabla h|^2 dxdt \leq Cq^4\int_0^T\int_\Omega (l^\varepsilon)^{2q}\left(1-\frac{k}{l^\varepsilon}\right)^{(n+2)(q-1)} dxdt.$$

Invoking the Sobolev embedding theorem on $(0,T)\times\Omega$, we obtain

$$\left(\int_0^T\int_\Omega h^{2\kappa}dxdt\right)^{1/\kappa} \leq Cq^4\int_0^T\int_\Omega (l^\varepsilon)^{2q}\left(1-\frac{k}{l^\varepsilon}\right)_+^{(n+2)q-n-2} dxdt,$$

where $\kappa = (n+2)/n$. Iterating the last inequality for $q = 1, \kappa, \kappa^2, \ldots$, as in [26], implies that

$$\sup_{(0,T)\times\Omega}|l^\varepsilon|^2\left(1-\frac{k}{l^\varepsilon}\right)_+^{n+2} \leq C\int_0^T\int_\Omega |l^\varepsilon|^2 dxdt.$$

Considering separately the cases $\sup_{(0,T)\times\Omega} l^\varepsilon \leq 2k$ and $\sup_{(0,T)\times\Omega} l^\varepsilon \geq 2k$ results in the estimate of the lemma. □

6.2. Two-scale convergence with parameters. We recall here the definition of two-scale convergence for functions dependent on parameters and several important results concerning this notion presented in [37]. The proofs are straightforward modifications of the proofs for the standard two-scale convergence method presented in [2].

DEFINITION 6.1. *Let (u_ε) be a sequence in $L^2(\Lambda\times\Omega)$, where ε is a sequence of strictly positive numbers, which tends to zero. (u_ε) is said to two-scale converge to a (unique) limit $u_0 \in L^2(\Lambda\times\Omega\times Z)$ iff for any $\phi \in \mathcal{D}(\Lambda\times\Omega, C^\infty_{per}(Z))$ we have*

$$\lim_{\varepsilon\to 0}\int_\Lambda\int_\Omega u_\varepsilon(\lambda,x)\phi\left(\lambda,x,\frac{x}{\varepsilon}\right)dxd\lambda = \int_\Lambda\int_\Omega\int_Z u_0(\lambda,x,y)\phi(\lambda,x,y)dxdyd\lambda.$$

THEOREM 6.2. *From each bounded sequence (u_ε) in $L^2(\Lambda\times\Omega)$ we can extract a subsequence which two-scale converges to $u_0 \in L^2(\Lambda\times\Omega\times Z)$.*

THEOREM 6.3. 1. *Let (u_ε) be a bounded sequence in $L^2(\Lambda, H^1(\Omega))$, which converges weakly to a limit function $u \in L^2(\Lambda, H^1(\Omega))$. Then there exists $u_1 \in L^2(\Lambda\times\Omega, H^1_{per}(Z))$ such that, up to a subsequence, u_ε two-scale converges to u and ∇u_ε two-scale converges to $\nabla u(\lambda,x) + \nabla_y u_1(\lambda,x,y)$.*

2. *Let (u_ε) and $(\varepsilon\nabla u_\varepsilon)$ be bounded sequences in $L^2(\Lambda\times\Omega))$. Then there exists $u_0 \in L^2(\Lambda\times\Omega, H^1_{per}(Z))$ such that, up to a subsequence, u_ε and $\varepsilon\nabla u_\varepsilon$ two-scale converge to $u_0(\lambda,x,y)$ and $\nabla_y u_0(\lambda,x,y)$, respectively.*

Now, we transfer the compactness results to the case of a sequence u_ε defined on an $(n-1)$-dimensional ε-periodic manifold $\Gamma^\varepsilon \in \Omega$. Let $\Gamma \in Z$ be a smooth $(n-1)$-dimensional manifold (in our application a sphere, $n=3$). Then Γ^ε is the union of all $\varepsilon\Gamma$. For each Γ^ε we consider the space $L^2(\Gamma^\varepsilon)$ equipped with the scalar product $(u,v)_{\Gamma^\varepsilon} := \varepsilon \int_{\Gamma^\varepsilon} u(x)v(x)dx$.

DEFINITION 6.4 (see [37]). *A sequence of functions $(w_\varepsilon) \in L^2(\Lambda \times \Gamma^\varepsilon)$ is said to two-scale converge to a limit $w \in L^2(\Lambda \times \Omega \times \Gamma)$ iff for any $\psi \in \mathcal{D}(\Lambda \times \Omega, C^\infty_{per}(\Gamma))$ we have*

$$\lim_{\varepsilon\to 0} \varepsilon \int_\Lambda \int_{\Gamma^\varepsilon} w^\varepsilon(\lambda,x)\psi\left(\lambda,x,\frac{x}{\varepsilon}\right) d\gamma_x d\lambda = \int_\Lambda \int_\Omega \int_\Gamma w(\lambda,x,y)\psi(\lambda,x,y)dx d\gamma_y d\lambda.$$

THEOREM 6.5. 1. *From each bounded sequence (w_ε) in $L^2(\Lambda\times\Gamma^\varepsilon)$ we can extract a subsequence which two-scale converges to $w \in L^2(\Lambda\times\Omega\times\Gamma)$.*

2. *If the sequence (w_ε) is bounded in $L^\infty(\Lambda\times\Gamma^\varepsilon)$, then the limit w belongs to $L^\infty(\Lambda\times\Omega\times\Gamma)$.*

Proof. For the proof of 1, see [37].

2. We know that if w^ε is bounded in $L^2((0,T)\times\Gamma^\varepsilon)$, then there exists $w \in L^2((0,T)\times\Omega\times\Gamma)$ such that $w^\varepsilon \to w$ two-scale [37]. Now we use the proof of that theorem and show that if w^ε is bounded in $L^\infty((0,T)\times\Gamma^\varepsilon)$, then $w^\varepsilon \to w$ two-scale and $w \in L^\infty((0,T)\times\Omega\times\Gamma)$. We define $\mu_\varepsilon(\phi) = \varepsilon\int_0^T \int_{\Gamma^\varepsilon} w^\varepsilon(t,x)\phi(t,x,\frac{x}{\varepsilon})\, d\gamma_x^\varepsilon\, dt$ and obtain

$$|\mu_\varepsilon(\phi)| \le ||w^\varepsilon||_{L^2((0,T)\times\Gamma^\varepsilon)} \Big(\int_0^T \int_{\Gamma^\varepsilon} \varepsilon \left|\phi\left(x,\frac{x}{\varepsilon}\right)\right|^2 d\gamma_x^\varepsilon\, dt\Big)^{\frac{1}{2}} \le c||\phi||^2_{C^0([0,T]\times\bar\Omega;C^0_{per}(\Gamma))}.$$

Therefore, $\{\mu_\varepsilon\}$ is a bounded sequence of functionals on $C^0([0,T]\times\bar\Omega; C^0_{per}(\Gamma))$. Since this space is a separable Banach space, there exists a subsequence of μ_ε that converges weakly* to μ. Using the boundedness of w^ε and a variant of the oscillation lemma [2], we obtain

$$|\mu(\phi)| = \lim_{\varepsilon\to 0} |\mu_\varepsilon(\phi)| \le C \lim_{\varepsilon\to 0} \Big(\int_0^T \int_{\Gamma^\varepsilon} \varepsilon \left|\phi\left(x,\frac{x}{\varepsilon}\right)\right|^2 d\gamma_x^\varepsilon\, dt\Big)^{\frac{1}{2}} = c||\phi||_{L^2((0,T)\times\Omega\times\Gamma)}.$$

Therefore, μ is a bounded functional on $L^2((0,T)\times\Omega\times\Gamma)$. The Riesz representation theorem implies the existence of a function $w \in L^2((0,T)\times\Omega\times\Gamma)$. Furthermore, $||w^\varepsilon||_{L^\infty((0,T)\times\Gamma^\varepsilon)} \le C$ yields

$$|\mu(\phi)| = \lim |\mu_\varepsilon(\phi)| \le C \lim_{\varepsilon\to 0} \int_0^T \int_{\Gamma^\varepsilon} \varepsilon \left|\phi\left(x,\frac{x}{\varepsilon}\right)\right| d\gamma_x^\varepsilon\, dt = c||\phi||_{L^1((0,T)\times\Omega\times\Gamma)}.$$

Finally, we conclude

$$||w||_{L^\infty((0,T)\times\Omega\times\Gamma)} = \frac{\langle w,\phi\rangle}{||\phi||_{L^1((0,T)\times\Omega\times\Gamma)}}$$

$$= \frac{|\mu(\phi)|}{||\phi||_{L^1((0,T)\times\Omega\times\Gamma)}} \le \frac{C||\phi||_{L^1((0,T)\times\Omega\times\Gamma)}}{||\phi||_{L^1((0,T)\times\Omega\times\Gamma)}} = C. \qquad \square$$

THEOREM 6.6 (see [37]). *Let (u_ε) and $(\varepsilon\nabla u_\varepsilon)$ be bounded sequences in $L^2(\Lambda\times\Gamma^\varepsilon))$. Then there exists $u_0 \in L^2(\Lambda\times\Omega, H^1_{per}(\Gamma))$ such that, up to a subsequence, u_ε and $\varepsilon\nabla u_\varepsilon$, two-scale converge to $u_0(\lambda,x,y)$ and $\nabla_y u_0(\lambda,x,y)$, respectively.*

REFERENCES

[1] E. Acerbi, V. Chiado Piat, G. Dal Maso, and D. Percivale, *An extension theorem from connected sets, and homogenization in general periodic domains*, Nonlinear Anal., 18 (1992), pp. 481–496.

[2] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[3] G. Allaire, A. Damlamian, and U. Hornung, *Two-scale convergence on periodic surfaces and applications*, in Proceedings of the International Conference on Mathematical Modelling of Flow through Porous Media, A. Bourgeat et al., eds., World Scientific, Singapore, 1996, pp. 15–25.

[4] G. Allaire and F. Murat, *Homogenization of the Neumann problem with nonisolated holes*, Asymptotic Anal., 7 (1993), pp. 81–95.

[5] T. Arbogast, J. Douglas, Jr., and U. Hornung, *Derivation of the double porosity model of single phase flow via homogenization theory*, SIAM J. Math. Anal., 21 (1990), pp. 823–836.

[6] A. Bourgeat, S. Luckhaus, and A. Mikelić, *Convergence of the homogenization process for a double-porosity model of immiscible two-phase flow*, SIAM J. Math. Anal., 27 (1996), pp. 1520–1543.

[7] D. Cioranescu, A. Damlamian, and G. Griso, *Periodic unfolding and homogenization*, C. R. Acad. Sci. Paris Sér. I Math., 335 (2002), pp. 99–104.

[8] D. Cioranescu and J. Saint Jean Paulin, *Homogenization in open sets with holes*, J. Math. Anal. Appl., 71 (1979), pp. 590–607.

[9] C. Conca, J. I. Diaz, A. Linan, and C. Timofte, *Homogenization in chemical reactive flows*, Electron. J. Differential Equations, 40 (2004), pp. 1–22.

[10] K. A. Duca, V. Lam, I. Keren, E. E. Endler, G. J. Letchworth, I. S. Novella, and J. Yin, *Quantifying viral propagation in vitro: Toward a method for characterization of complex phenotypes*, Biotechnology Progress, 17 (2001), pp. 1156–1165.

[11] L. Dung, *Remarks on Hölder continuity for parabolic equations and convergence to global attractors*, Nonlinear Anal., 41 (2000), pp. 921–941.

[12] L. Dung and H. Smith, *Strong positivity of solutions to parabolic and elliptic equations on nonsmooth domains*, J. Math. Anal. Appl., 275 (2002), pp. 208–221.

[13] J. I. Elliott, A. Surprenant, F. M. Marelli-Berg, J. C. Cooper, R. Cassady-Cain, C. Wooding, K. Linton, D. R. Alexander, and C. F. Higgins, *Membrane phospatidylserine distribution as a non-apoptotic signalling mechanism in lymphocytes*, Nature Cell Biol., 7 (2005), pp. 808–816.

[14] K. Gaus, M. Rodriguez, K. R. Ruberu, I. Gelissen, T. M. Sloane, L. Kritharides, and W. Jessup, *Domain-specific lipid distribution in macrophage plasma membranes*, J. Lipid Res., 46 (2005), pp. 1526–1538.

[15] O. O. Glebov and B. J. Nichols, *Lipid raft proteins have a random distribution during localized activation of the T-cell receptor*, Nature Cell Biol., 6 (2004), pp. 238–243.

[16] C. Gomez-Mounton, J. L. Abad, E. Mira, R. A. Lacalle, E. Gallaro, S. Jimenez-Baranda, I. Illa, A. Bernard, S. Manes, and A. C. Martinez, *Segregation of leading-edge and uropod components into specific lipid rafts during T cell polarization*, Proc. Natl. Acad. Sci. USA, 98 (2001), pp. 9642–9647.

[17] A. Gratchev, *Personal communication*, Medical Center in Mannheim, University of Heidelberg, 2005.

[18] U. Hornung, *Homogenization and Porous Media*, Springer-Verlag, New York, 1997.

[19] U. Hornung and W. Jäger, *Diffusion, convection, adsorption and reaction of chemicals in porous media*, J. Differential Equations, 92 (1991), pp. 199–225.

[20] U. Hornung, W. Jäger, and A. Mikelić, *Reactive transport through an array of cells with semi-permeable membranes*, RAIRO Modél. Math. Anal. Numér., 28 (1994), pp. 59–94.

[21] E. Ikonen, *Roles of lipid rafts in membrane transport*, Curr. Opin. Cell Biol., 13 (2001), pp. 470–477.

[22] B. Johnston and E. C. Butcher, *Chemokines in rapid leukocyte adhesion triggering and migration*, Immunology, 14 (2002), pp. 83–92.

[23] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Uralceva, *Linear and Quasi-Linear Equations of Parabolic Type*, AMS, Providence, RI, 1968.

[24] D. A. Lauffenburger and J. J. Linderman, *Receptors. Models for Binding, Trafficking, and Signaling*, Oxford University Press, New York, 1993.

[25] P. Li, P. Selvaraj, and C. Zhu, *Analysis of Competition binding between soluble and membrane-bound ligands for cell surface receptors*, Biophys. J., 77 (1999), pp. 3394–3408.

[26] G. M. Lieberman, *Second Order Parabolic Differential Equations*, World Scientific, Singapore, 1996.
[27] J. L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Paris, 1969.
[28] J. Little, *Genomic instability and bystander effects: A historical perspective*, Oncogene, 22 (2003), pp. 6978–6987.
[29] A. Marciniak-Czochra, *Receptor-based models with diffusion-driven instability for pattern formation in hydra*, J. Biol. Sys., 11 (2003), pp. 293–324.
[30] A. Marciniak-Czochra, *Receptor-based models with hysteresis for pattern formation in hydra*, Math. Biosci., 199 (2005), pp. 97–119.
[31] J. L. Martiel and A. Goldbeter, *A model based on receptor desensitization for cyclic AMP signaling in* Disctyostelium *cells*, Biophys. J., 57 (1987), pp. 807–828.
[32] P. B. Monk and H. G. Othmer, *Cyclic AMP oscillations in suspensions of* Disctyostelium Discoideum, Philos. Trans. Roy. Soc. London Ser. B, 323 (1989), pp. 185–224.
[33] W .A. Müller, *Pattern control in hydra: Basic experiments and concepts*, in Experimental and Theoretical Advances in Biological Pattern Formation, H. G. Othmer, P. K. Maini, and J. D. Murray, eds., Plenum Press, New York, 1993.
[34] W. A. Müller, *Developmental Biology*, Springer-Verlag, New York, 1997.
[35] J. Murray, *Mathematical Biology*, Springer-Verlag, Berlin, 2003.
[36] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–623.
[37] M. Neuss-Radu, *Some extensions of two-scale convergence*, C. R. Acad. Sci. Paris Sér. I Math., 332 (1996), pp. 899–904.
[38] P. O'Shea, *Physical landscapes in biological membranes: Physico-chemical terrains for spatio-temporal control of biomolecular interactions and behaviour*, Philos. Trans. Roy. Soc. London Ser. A, 363 (2005), pp. 575–588.
[39] I. Pecht and D. M. Gakamsky, *Spatial coordination of CD*8 *and TCR molecules controls antigen recognition by CD*8^+ *T-cells*, FEBS Lett., 579 (2005), pp. 3336–3341.
[40] T. M. Razzaq, P. Ozegbe, E. C. Jury, P. Sembi, N. M. Blackwell, and P. S. Kabouridis, *Regulation of T-cell receptor signalling by membrane microdomains*, Immunology, 113 (2004), pp. 413–426.
[41] P. B. Sehgal, *Plasma membrane rafts and chaperones in cytokine STAT signaling*, Acta Bioch. Pol., 50 (2003), pp. 583–594.
[42] J. A. Sherratt, P. K. Maini, W. Jäger, and W. Müller, *A receptor-based model for pattern formation in hydra*, Forma, 10 (1995), pp. 77–95.
[43] K. Simons and D. Toomre, *Lipid rafts and signal transduction*, Nat. Rev. Mol. Cell Biol., 1 (2000), pp. 31–39.
[44] J. Smoller, *Shock-Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1994.
[45] W. Walter, *Gewöhnliche Differentialgleichungen*, Springer-Verlag, Berlin, 1996.
[46] H. Wearing and J. A. Sherratt, *Keratinocyte growth factor signalling: A mathematical model of dermal-epidermal interaction in epidermal wound healing*, Math. Biosci., 165 (2000), pp. 41–62.
[47] J. Wloka, *Partielle Differentialgleichungen*, Teubner Verlag, Stuttgart, 1982.

SIAM J. MATH. ANAL.
Vol. 40, No. 1, pp. 238–265

INVERSE PROBLEM OF DETERMINING THE DENSITY AND TWO LAMÉ COEFFICIENTS BY BOUNDARY DATA*

M. BELLASSOUED†, O. IMANUVILOV‡, AND M. YAMAMOTO§

Abstract. In this paper we study an inverse problem of determining spatially varying density and two Lamé coefficients by a single measurement of solution in a subboundary over a time interval. By assuming that, in a neighborhood of the boundary of the spatial domain, the density and the Lamé coefficients are known, we prove a logarithmic stability estimate for the inverse problem with a single measurement of data on an arbitrarily given subboundary.

Key words. Carleman estimate, Lamé system, inverse problem, Fourier–Bros–Iagolnitzer transform

AMS subject classifications. 35B60, 35R25, 35R30, 74B05

DOI. 10.1137/070679971

1. Introduction.

This paper is concerned with the global stability in determining density and two Lamé coefficients in the classical isotropic elasticity system from data of the solution on a subboundary over a time interval. We will formulate our problem as follows: In a bounded domain $\Omega \subset \mathbb{R}^3$ with sufficiently smooth boundary $\Gamma = \partial\Omega$, we consider the isotropic elasticity system

$$\rho(x)\partial_t^2 \mathbf{u}(x,t) - \mathcal{L}_{\lambda,\mu}(x,\partial_x)\mathbf{u}(x,t) = 0, \quad (x,t) \in Q = \Omega \times (-T,T), \tag{1.1}$$

where

$$\mathcal{L}_{\lambda,\mu}(x,\partial_x)\mathbf{v}(x) \equiv \mu(x)\Delta\mathbf{v}(x) + (\mu(x)+\lambda(x))\,(\nabla \mathrm{div}\,\mathbf{v}(x)) + (\mathrm{div}\,\mathbf{v}(x))\,\nabla\lambda(x) + \left(\nabla\mathbf{v} + (\nabla\mathbf{v})^T\right)\nabla\mu(x), \qquad x \in \Omega.$$

Throughout this paper, t and $x = (x_1, x_2, x_3)$ denote the time variable and the spatial variable, respectively, and $\mathbf{u} = (u_1, u_2, u_3)^T$ denotes the displacement at the location x and the time t, where $\cdot^T$ denotes the transpose of matrices. We will assume that the density ρ and the Lamé parameters μ and λ satisfy

$$\rho, \lambda, \mu \in C^3(\overline{\Omega}), \quad \rho(x) > 0,\ \mu(x) > 0, \quad \lambda(x) + \mu(x) > 0 \quad \text{for } x \in \overline{\Omega}.$$

To system (1.1), we attach initial and boundary conditions:

$$\mathbf{u} = \Phi, \quad \partial_t \mathbf{u} = \Psi \quad \text{on} \quad \Omega \times \{0\} \tag{1.2}$$

*Received by the editors January 12, 2007; accepted for publication (in revised form) September 5, 2007; published electronically April 16, 2008.
http://www.siam.org/journals/sima/40-1/67997.html

†Department of Mathematics, Faculty of Sciences of Bizerte, 7021 Jarzouna Bizerte, Tunisia (mourad.bellassoued@fsb.rnu.tn).

‡Department of Mathematics, Colorado State University, 101 Weber Building, Fort Collins, CO 80523 (oleg@math.colostate.edu). This author was supported by NSF grant DMS 0205148.

§Department of Mathematical Sciences, The University of Tokyo, 3-8-1 Komaba, Meguro, Tokyo 153, Japan (myama@ms.u-tokyo.ac.jp). This author was supported partially by grant 15340027 from the Japan Society for the Promotion of Science and grant 17654019 from the Ministry of Education, Culture, Sports and Technology.

and

$$\mathbf{u} = \mathbf{g} \quad \text{on} \quad \Sigma \equiv \Gamma \times (-T, T). \tag{1.3}$$

There exists a unique weak solution $\mathbf{u} \in C([-T,T]; \mathbf{H}_0^1(\Omega))$ for suitable $\Phi, \Psi, \mathbf{g}$ under sufficient compatibility conditions, and by $\mathbf{u}(\lambda, \mu, \rho, \Phi, \Psi, \mathbf{g})$ we denote the solution to (1.1)–(1.3).

The main subject of this paper is the inverse problem of determining $\rho = \rho(x)$, $\lambda = \lambda(x)$, and $\mu = \mu(x)$ from observed data of the solution on a part of the boundary. It is an important problem, for example, in the geophysics to determine ρ, λ, and μ inside an elastic body from measurements on a subboundary. Thus our inverse problem is physically motivated.

1.1. Inverse problem. Let $\Gamma_1 \subset \Gamma$ be given arbitrarily, and let Φ_j, Ψ_j, $\mathbf{g}_j$, $1 \le j \le \mathcal{N}$, be appropriately given. Then we want to determine $\lambda(x)$, $\mu(x)$, $\rho(x)$, $x \in \Omega$, by measurements

$$\partial_\nu \mathbf{u}(\lambda, \mu, \rho, \Phi_j, \Psi_j, \mathbf{g}_j)(x,t), \quad (x,t) \in \Sigma_1 \equiv \Gamma_1 \times (-T,T),\ 1 \le j \le \mathcal{N}.$$

Here $\nu = \nu(x)$ denotes the unit outward normal vector and $\partial_\nu = \nabla \cdot \nu$.

Our formulation of the inverse problem requires only a finite number of observations (i.e., $\mathcal{N} < \infty$). As for inverse problems for a nonstationary Lamé system by infinitely many boundary observations (i.e., Dirichlet-to-Neumann map), we refer to Rachele [42], for example. Moreover see a monograph by Yakhno [48] for inverse problems for the Lamé system.

For the formulation with a finite number of observations, Bukhgeim and Klibanov [10] proposed a remarkable method based on a Carleman estimate and established the uniqueness for similar inverse problems for scalar partial differential equations. See also Baudouin and Puel [2], Bellassoued [4], [5], Bellassoued and Yamamoto [6], [7], Bukhgeim [8], Bukhgeim, Cheng, Isakov, and Yamamoto [9], Imanuvilov and Yamamoto [20], [21], [22], [23], [24], Isakov [25], [26], Isakov and Yamamoto [27], Khaidarov [29], Klibanov [30], [31], Klibanov and Timonov [33], Klibanov and Yamamoto [34], Kubo [35], Li [39], Puel and Yamamoto [40], [41], and Yamamoto [49].

A Carleman estimate is an inequality for a solution to a partial differential equation with weighted L^2-norm and is a strong tool also for proving the uniqueness in the Cauchy problem or the unique continuation for a partial differential equation with nonanalytic coefficients. Moreover Carleman estimates have been applied essentially for estimating the energy (e.g., Kazemi and Klibanov [28], Klibanov and Malinsky [32]), while we refer to [1] as another method for the energy estimate, which is, however, not applicable to our inverse problem.

As a pioneering work concerning a Carleman estimate, we refer to Carleman's paper [11], which proved what is now called a Carleman estimate and applied it for proving the uniqueness in the Cauchy problem for a two-dimensional elliptic equation. Since [11], the theory of Carleman estimates has been studied extensively. We refer to a general theory by Hörmander [14] in the case where the symbol of a partial differential equation is isotropic and functions under consideration have compact supports (that is, they and their derivatives of suitable orders vanish on the boundary of a domain). Later Carleman estimates for functions with compact supports have been obtained for partial differential operators with anisotropic symbols by Isakov [26]. For Carleman estimates for functions without compact supports, see Imanuvilov [17] and Tataru [46]. We further refer to Fursikov and Imanuvilov [13] and Imanuvilov [16]. As

for a direct derivation of pointwise Carleman estimates for hyperbolic equations which are applicable to functions without compact supports, see Klibanov and Timonov [33] and Lavrent'ev, Romanov, and Shishat·skiĭ [36].

The Carleman estimate for the nonstationary Lamé system was obtained for functions with compact supports, by Eller, Isakov, Nakamura, and Tataru [12], Ikehata, Nakamura, and Yamamoto [15], Imanuvilov, Isakov, and Yamamoto [18], and Isakov [25]. Lemma 2.3 is our Carleman estimate for the Lamé system whose solutions have not necessarily compact supports, and the Carleman estimate bounds also the second-order x-derivatives, which is possible because the right-hand side is estimated in the weighted H^1-norm in x, and we can use the identity $\Delta v = -\mathrm{rot}\,\mathrm{rot}\,v + \nabla(\mathrm{div}\,v)$ and the a priori estimate for the Dirichlet problem for the Laplace equation. By the methodology by [10] or [22] with such Carleman estimates, several uniqueness and stability results are available for the inverse problem for the Lamé system (1.1). That is, in [25] Isakov established the uniqueness in determining a single coefficient $\rho(x)$, using four measurements (i.e., $\mathcal{N} = 4$). Later [15] reduced the number of measurements to three. Recently [18] proved conditional stability and the uniqueness in the determination of all three functions λ, μ, and ρ, with two measurements, and Imanuvilov and Yamamoto [23], [24] proved conditional stability results with a single measurement, provided that initial data satisfy some nondegeneracy condition.

In all of the works [15], [18], [23], [24], [25], the authors assume some geometric condition of the observation subboundary. For such a kind of inverse problems, the uniqueness as well as the stability with boundary measurement on an arbitrary part of Γ are open problems. As for the corresponding unique continuation, we can refer to Bellassoued [3], Robbiano [43], and Tataru [46]. However, their methods are not applicable to the inverse problem. In our paper, by assuming that coefficients under consideration are given in a neighborhood of Γ, we will prove a stability result in the inverse problem. The coincidence of coefficients near the boundary is technically restrictive but acceptable from practical viewpoints, because one can directly know physical properties near the boundary.

Our main achievement of this paper is that we can take an arbitrary observation subboundary Γ_1 for the stability estimate. The key idea is a combination of the method (e.g., [10], [23]) by the Carleman estimates and the Fourier–Bros–Iagolnitzer (FBI) transformation which was used for sharp unique continuation by Robbiano [44]. More precisely, we apply the FBI transformation to change the problem near the boundary into a problem to which elliptic estimates can be applied.

1.2. Notation and statement of main results. In order to formulate our results, we need to introduce some notation. Let $x_0 \in \mathbb{R}^3 \setminus \overline{\Omega}$, $M_0 \geq 0$, $0 < \theta_0 \leq 1$, and $\theta_1 > 0$ be arbitrarily fixed, and let us introduce the conditions on a scalar function p:

$$\begin{cases} p(x) \geq \theta_1 > 0, \quad x \in \overline{\Omega}, \\ \\ \|p\|_{C^3(\overline{\Omega})} \leq M_0, \quad \dfrac{(\nabla p(x) \cdot (x - x_0))}{2p(x)} \leq 1 - \theta_0, \quad x \in \overline{\Omega \setminus \omega}. \end{cases} \tag{1.4}$$

Next we define an admissible set of unknown coefficients λ, μ, ρ. Let $\omega \subset \Omega$ be a given arbitrary neighborhood of the boundary Γ. For fixed functions ρ_0, λ_0, μ_0 on ω

and Φ, Ψ in Ω, and a given constant $M_1 > 0$, we set $\Lambda = \Lambda_{M_0,M_1,\theta_0,\theta_1}$

$$\Lambda = \Bigg\{ (\lambda,\mu,\rho) \in \left(C^3(\overline{\Omega})\right)^3 ; (\rho,\lambda,\mu) = (\rho_0,\lambda_0,\mu_0) \quad \text{in } \omega,$$

$$\left(\frac{\lambda+2\mu}{\rho}\right), \left(\frac{\mu}{\rho}\right) \text{ satisfy(1.4)}, \ \|\mathbf{u}(\lambda,\mu,\rho,\Phi,\Psi,\mathbf{g})\|_{W^{s,\infty}(Q)} \le M_1 \Bigg\}. \tag{1.5}$$

Throughout this paper, let $\mathbf{I}_3$ be the 3×3 identity matrix. We note that $\mathcal{L}_{\lambda,\mu}(x,\partial_x)\Phi(x)$ is the 3-column vector for a 3-column vector Φ. Moreover by $\{\mathbf{a}\}_j$ we denote the matrix (or vector) obtained from $\mathbf{a}$ after deleting the jth row, $\det_j A$ means $\det\{A\}_j$ for a square matrix A, and $\langle A\rangle_j$ is the matrix which is obtained from A by deleting the jth column of A. Furthermore we assume that the observation data are measured by the norm:

$$\epsilon(\Sigma_1) = \sum_{|\alpha|=1}^{2} \|\partial_x^\alpha \mathbf{u}(\lambda,\mu,\rho,\Phi,\Psi,\mathbf{g}) - \partial_x^\alpha \mathbf{u}(\widetilde{\lambda},\widetilde{\mu},\widetilde{\rho},\Phi,\Psi,\mathbf{g})\|^2_{L^2(\Sigma_1)} \tag{1.6}$$

for (λ,μ,ρ), $(\widetilde{\lambda},\widetilde{\mu},\widetilde{\rho}) \in \Lambda$.

1.3. Hypotheses $(\mathcal{H}_1)$–$(\mathcal{H}_2)$. Let (λ,μ,ρ) be an arbitrarily fixed element of Λ. For $\Phi = (\phi_1,\phi_2,\phi_3)^T$ and $\Psi = (\psi_1,\psi_2,\psi_3)^T$, we assume that there exist $j_1, j_2, \in \{1,\dots,6\}$ such that for all $x \in \overline{\Omega}$

$(\mathcal{H}_1)$

$$\det_{j_1} \begin{pmatrix} \mathcal{L}_{\lambda,\mu}(x,\partial_x)\Phi(x) & (\operatorname{div}\Phi(x))\mathbf{I}_3 & \left(\nabla\Phi(x) + (\nabla\Phi(x))^T\right)(x-x_0) \\ \mathcal{L}_{\lambda,\mu}(x,\partial_x)\Psi(x) & (\operatorname{div}\Psi(x))\mathbf{I}_3 & \left(\nabla\Psi(x) + (\nabla\Psi(x))^T\right)(x-x_0) \end{pmatrix} \neq 0,$$

$$(\mathcal{H}_2) \qquad \det_{j_2} \begin{pmatrix} \mathcal{L}_{\lambda,\mu}(x,\partial_x)\Phi(x) & \nabla\Phi(x) + (\nabla\Phi(x))^T & (\operatorname{div}\Phi)(x-x_0) \\ \mathcal{L}_{\lambda,\mu}(x,\partial_x)\Psi(x) & \nabla\Psi(x) + (\nabla\Psi(x))^T & (\operatorname{div}\Psi)(x-x_0) \end{pmatrix} \neq 0.$$

Now we are ready to state the main result, which proves that only one observation with a suitable initial value yields the logarithmic conditional stability for our inverse problem.

THEOREM 1.0. *Let $T > 0$ be sufficiently large for Ω, ω, and let $\Lambda = \Lambda_{M_0,M_1,\theta_0,\theta_1}$ be defined by* (1.5). *Moreover let (Φ,Ψ) satisfy the conditions* $(\mathcal{H}_1)$–$(\mathcal{H}_2)$. *Then there exist constants $C > 0$ and $\kappa \in (0,1)$ such that the following estimate holds:*

$$\|\widetilde{\lambda} - \lambda\|_{H^2(\Omega)} + \|\widetilde{\mu} - \mu\|_{H^2(\Omega)} + \|\widetilde{\rho} - \rho\|_{H^1(\Omega)} \le C \left[\log\left(2 + \frac{C}{\epsilon(\Sigma_1)}\right)\right]^{-\kappa}$$

for any $(\tilde{\lambda},\tilde{\mu},\tilde{\rho}) \in \Lambda$.

Here we note that $\epsilon(\Sigma_1)$ is given by (1.6) *and the constants C and $\kappa \in (0,1)$ are dependent on, Ω, ω, T, M_0, M_1 and independent of $(\lambda,\mu,\rho) \in \Lambda$.*

Our stability result requires only one measurement, $\mathcal{N} = 1$, and the stability result is of logarithmic rate and weaker than any Hölder stability. We notice that, with a suitable geometric condition of the observation subboundary, we can prove Hölder (or the Lipschitz) stability by means of the method in Imanuvilov and Yamamoto [23], [24]. By Theorem 1.0, we can readily derive the uniqueness in the inverse problem.

COROLLARY 1.1 (uniqueness). *Under the assumptions in Theorem* 1*, for all* $(\lambda, \mu, \rho), (\tilde{\lambda}, \tilde{\mu}, \tilde{\rho}) \in \Lambda$*, we have the uniqueness*

$$\partial_\nu \mathbf{u}(x,t) = \partial_\nu \widetilde{\mathbf{u}}(x,t), \quad (x,t) \in \Sigma_1 \quad \textit{implies} \quad (\lambda, \mu, \rho) = (\tilde{\lambda}, \tilde{\mu}, \tilde{\rho}) \quad \textit{in } \Omega.$$

For the determination of the three coefficients by a single measurement, we have to choose initial data satisfying conditions $(\mathcal{H}_1)$–$(\mathcal{H}_2)$. Thus conditions $(\mathcal{H}_1)$–$(\mathcal{H}_2)$ are not generic properties, and we should satisfy them artificially and a posteriori. Moreover, as the following example shows, we can take such Φ and Ψ.

Example of Ω*,* Φ*,* Ψ *meeting* $(\mathcal{H}_1)$–$(\mathcal{H}_2)$. For simplicity, we assume that $x_0 = (0,0,0)$, $\overline{\Omega}$ does not intersect any of $\{x_1 = 0\}$, $\{x_2 = 0\}$, $\{x_3 = 0\}$, and $\{x_1 + x_3 = 0\}$, and λ, μ are positive constants. For example, we take

$$\Phi(x) = \begin{pmatrix} 0 \\ x_1 x_2 \\ 0 \end{pmatrix}, \quad \Psi(x) = \begin{pmatrix} x_2^2 \\ 0 \\ x_2^2 \end{pmatrix}.$$

Then, by choosing $j_1 = j_2 = 6$, we can verify that $(\mathcal{H}_1)$–$(\mathcal{H}_2)$ are satisfied.

Thanks to the extra information $(\lambda, \mu, \rho) = (\tilde{\lambda}, \tilde{\mu}, \tilde{\rho})$ in a neighborhood ω of $\partial\Omega$, the sharp unique continuation by Bellassoued [3], implies $\mathbf{u}(\lambda, \mu, \rho, \Phi, \Psi, \mathbf{g})(x,t) = \mathbf{u}(\widetilde{\lambda}, \widetilde{\mu}, \widetilde{\rho}, \Phi, \Psi, \mathbf{g})(x,t)$ on $\partial(\Omega \setminus \overline{\omega}) \times (-T, T)$, provided that $T > 0$ is sufficiently large. Therefore the method in Imanuvilov, Isakov, and Yamamoto[18], and Imanuvilov and Yamamoto [24] directly yields the uniqueness in our inverse problem. However, our main result is concerned with the stability in the inverse problem, and the direct combination of the existing results in [18], [21], [22] does not work. For our purpose, we will use the FBI transformation according to Robbiano [44], [43].

The remainder of the paper is organized as follows. In section 2, we give key estimates. In section 3, we prove Theorem 1.0 on the basis of the weak observation estimate, that is, an estimate of $\mathbf{u}(\lambda, \mu, \rho, \Phi, \Psi, \mathbf{g}) - \mathbf{u}(\widetilde{\lambda}, \widetilde{\mu}, \widetilde{\rho}, \Phi, \Psi, \mathbf{g})$ by data on any small part of the boundary. Section 4 is devoted to the proof of the weak observation estimate.

2. Preliminaries and Carleman estimates. We set

$$\omega(\epsilon) = \{x \in \Omega;\ \operatorname{dist}(x, \partial\Omega) \leq \epsilon\}$$

and

$$\omega(\epsilon_1, \epsilon_2) = \{x \in \Omega;\ \epsilon_1 \leq \operatorname{dist}(x, \partial\Omega) \leq \epsilon_2\},$$

with $0 < \epsilon_1 < \epsilon_2$ and $\epsilon > 0$.

In this section we first derive several estimates. We choose $\epsilon_0, \epsilon_1, \epsilon_2 > 0$ such that

$$\omega(3\epsilon_0) \subset \omega \tag{2.1}$$

and

$$\omega(\epsilon_1, \epsilon_2) \subset \omega, \quad \epsilon_1 < \epsilon_2 < 8\epsilon_0. \tag{2.2}$$

We set

$$\omega_T(\epsilon) = \omega(\epsilon) \times [-T, T], \quad \omega_T(\epsilon_1, \epsilon_2) = \omega(\epsilon_1, \epsilon_2) \times [-T, T].$$

For δ_0 such that $0 < \delta_0 < T$, we set

$$Q_{\delta_0} = \Omega \times [-T + \delta_0, T - \delta_0], \quad Q_{\delta_0}(\epsilon) = (\Omega \setminus \omega(\epsilon)) \times [-T + \delta_0, T - \delta_0].$$

For formulating our Carleman estimate, we need some notation. Set

$$d = (\sup_{x \in \Omega} |x - x_0|^2 - \inf_{x \in \Omega} |x - x_0|^2)^{\frac{1}{2}}, \tag{2.3}$$

where $x_0 \notin \overline{\Omega}$ is arbitrarily fixed. We choose $\theta > 0$ such that

$$\theta + \frac{M_0 d}{\sqrt{\theta_1}} \sqrt{\theta} < \theta_0 \theta_1, \qquad \theta_1 \inf_{x \in \Omega} |x - x_0|^2 - \theta \sup_{x \in \Omega} |x - x_0|^2 > 0. \tag{2.4}$$

Here we note that, since $x_0 \notin \overline{\Omega}$, such $\theta > 0$ exists.

We introduce two functions $\psi, \varphi : \Omega \times \mathbb{R} \longrightarrow \mathbb{R}$ of class C^1 by setting

$$\psi(x,t) = |x - x_0|^2 - \theta |t|^2 \quad \text{for all } x \in \Omega, \quad -T \le t \le T,$$

$$\varphi(x,t) = e^{\beta \psi(x,t)}, \quad \beta > 0, \tag{2.5}$$

where $T > \frac{d}{\sqrt{\theta}}$. Therefore, by (2.4) and (2.5), we have

$$\varphi(x,0) \ge d_0, \quad \varphi(x, \pm T) < d_0, \tag{2.6}$$

with $d_0 = \exp(\beta \inf_{x \in \Omega} |x - x_0|)$. Thus, for a given sufficiently small $\eta > 0$, we can choose a sufficiently small $\delta_0 = \delta_0(\eta)$ such that

$$\varphi(x,t) \le d_0 - \eta \quad \text{for all } \ (x,t) \in Q \backslash \overline{Q_{2\delta_0}}. \tag{2.7}$$

We set $\nabla_{x,t} v(t,x) = (\frac{\partial v}{\partial x_1}, \frac{\partial v}{\partial x_2}, \frac{\partial v}{\partial x_3}, \frac{\partial v}{\partial t}) = (\nabla v, \partial_t v)$, and we shall use the weighted norm:

$$\|u\|^2_{H^{1,\tau}(Q)} = \tau^2 \|u\|^2_{L^2(Q)} + \|\nabla_{x,t} u\|^2_{L^2(Q)}.$$

Moreover $H_x^{k,\tau}(Q)$ is the Sobolev space equipped with the norm

$$\|u\|^2_{H_x^{k,\tau}(Q)} = \sum_{|\alpha| \le k} \tau^{2(k-|\alpha|)} \|\partial_x^\alpha u\|^2_{L^2(Q)}.$$

In what follows, $C > 0$ denote generic constants depending on $\widehat{\beta}$, τ_0, M_0, M_1, θ_0, θ_1, Ω, T, x_0, ω, χ and Φ, Ψ, ϵ, δ but independent of $\tau > \widehat{\tau}$.

2.1. Preliminary estimates. The first lemma is a classical Carleman estimate for a scalar hyperbolic equation (e.g., [14], [17], [25], [26]). See also Triggiani and Yao [47].

LEMMA 2.1. *Let φ be defined by* (2.5). *If $\frac{a}{\rho}$ satisfies* (1.4), *then there exists $\widehat{\beta} > 0$ such that for any $\beta > \widehat{\beta}$ we can choose $\widehat{\tau}(\beta) > 0$ such that the following estimate holds true:*

$$\tau \|e^{\tau\varphi} u\|^2_{H^{1,\tau}(Q)} \le C \left\| e^{\tau\varphi} \left(\rho \partial_t^2 - a\Delta \right) u \right\|^2_{L^2(Q)},$$

whenever a function $u \in H^2(Q)$ is supported in Q and $\tau > \widehat{\tau}$.

The next lemma is a weighted estimate for the elliptic equation which follows from the formula

$$|\Delta(\mathbf{v}e^{\tau\varphi})| = O(\tau^2)e^{\tau\varphi}|\mathbf{v}| + O(\tau)e^{\tau\varphi}|\nabla\mathbf{v}|$$

$$+e^{\tau\varphi}|\mathrm{rot}\,(\mathrm{rot}\,\mathbf{v}) - \nabla(\mathrm{div}\,\mathbf{v})|$$

in Q and the standard a priori estimate for the Dirichlet problem for the Poisson equation.

LEMMA 2.2. *There exists a constant $C > 0$ such that*

$$\frac{1}{\tau}\|e^{\tau\varphi}\mathbf{v}\|^2_{H^{2,\tau}_x(Q)} \le C(\tau\|e^{\tau\varphi}\mathbf{v}\|^2_{H^{1,\tau}_x(Q)} + \|e^{\tau\varphi}(\nabla\mathrm{div}\mathbf{v})\|^2_{L^2(Q)} + \|e^{\tau\varphi}(\nabla\mathrm{rot}\mathbf{v})\|^2_{L^2(Q)}),$$

whenever $\mathbf{v} \in H^2(Q)$ and $\mathbf{v}_{|\Sigma} = 0$.

2.2. Carleman estimate for the Lamé system. In order to prove a Carleman estimate, we have to assume a condition called the pseudoconvexity (e.g., [14]) where the coefficient of the principal term is involved. Since such a coefficient is unknown in our inverse problem, we need to establish a Carleman estimate with one possible explicit characterization (1.4) of the coefficients for the pseudoconvexity, and we will argue similarly to Bellassoued [4]. Moreover for our stability estimates, unlike [12], [26], we require a Carleman estimate for functions which do not have compact supports.

Now we will consider the three-dimensional isotropic nonstationary Lamé system

$$P(x,\partial_x,\partial_t)\mathbf{y}(x,t) = \rho(x)\partial_t^2\mathbf{y}(x,t) - \mathcal{L}_{\lambda,\mu}(x,\partial_x)\mathbf{y}(x,t) = \mathbf{f}(x,t), \quad (x,t) \in Q. \tag{2.8}$$

We have a Carleman estimate.

LEMMA 2.3. *There exists $\widehat{\beta} > 0$ such that for any $\beta > \widehat{\beta}$ we can choose $\tau_0 = \tau_0(\beta) > 0$ such that for any solution $\mathbf{y} \in H^2(Q)$ to problem (2.8) the following estimate holds true:*

$$\begin{aligned}\frac{1}{\tau}\|e^{\tau\varphi}\mathbf{y}\|^2_{H^{2,\tau}_x(Q_{\delta_0}(3\epsilon_0))} \le C(&\|e^{\tau\varphi}\mathbf{f}\|^2_{L^2(Q)} + \|e^{\tau\varphi}\nabla\mathbf{f}\|^2_{L^2(Q)} + e^{C\tau}\|\mathbf{y}\|^2_{H^2(\omega_T(\epsilon_0,3\epsilon_0))}\\ &+ e^{2\tau(d_0-\eta)}\|\mathbf{y}\|^2_{H^2(Q)})\end{aligned}$$

for any $\tau \ge \tau_0$, where the constant $C = C(\beta) > 0$ is independent of τ.

Proof. We introduce a cutoff function χ satisfying $0 \le \chi \le 1$, $\chi \in C_0^\infty(\mathbb{R}^3 \times \mathbb{R})$, $\chi_1 \in C_0^\infty(\mathbb{R}^3)$, $\chi_2 \in C_0^\infty(\mathbb{R})$, and

$$\chi(x,t) = \chi_1(x)\chi_2(t), \quad \chi_1(x) = \begin{cases} 0, & x \in \omega(\epsilon_0),\\ 1, & x \in \Omega \setminus \overline{\omega(3\epsilon_0)},\end{cases} \qquad \chi_2(t) = \begin{cases} 0, & |t| > T - \delta_0,\\ 1, & |t| < T - 2\delta_0.\end{cases}$$

Set $\mathbf{v}(x,t) = \chi(x,t)\mathbf{y}(x,t)$. Then we have

$$\partial_t^2\mathbf{v} - \frac{1}{\rho}\mathcal{L}_{\lambda,\mu}(x,\partial_x)\mathbf{v} = \frac{1}{\rho}\widetilde{\mathbf{f}} \quad \text{in } Q, \quad \mathbf{v} = 0 \quad \text{in } \omega(\epsilon_0), \tag{2.9}$$

where

$$\widetilde{\mathbf{f}}(x,t) = \chi(x,t)\mathbf{f} + [P,\chi_1]\chi_2\mathbf{y} + \chi_1[P,\chi_2]\mathbf{y}. \tag{2.10}$$

Let $v = \operatorname{div}\mathbf{v}$ and $\mathbf{w} = \operatorname{rot}\mathbf{v}$. Therefore, by [12], for example, we apply rot and div to (2.9) and obtain

$$\rho(x)\partial_t^2\mathbf{v} - \mu(x)\Delta\mathbf{v} + A_1(\mathbf{v}, v) = \widetilde{\mathbf{f}},$$

$$\rho(x)\partial_t^2 v - (2\mu(x) + \lambda(x))\,\Delta v + A_2(\mathbf{v}, v, \mathbf{w}) = R_1\widetilde{\mathbf{f}},$$
$$\rho(x)\partial_t^2\mathbf{w} - \mu(x)\Delta\mathbf{w} + A_3(\mathbf{v}, v, \mathbf{w}) = R_2\widetilde{\mathbf{f}},$$

where A_j and R_j are linear differential operators of the first order with bounded coefficients in Ω.

Since $\frac{\mu}{\rho}$ and $\frac{2\mu+\lambda}{\rho}$ satisfy (1.4), we have by Lemma 2.1

$$\begin{aligned}
&\tau(\|e^{\tau\varphi}\mathbf{v}\|^2_{H^{1,\tau}(Q)} + \|e^{\tau\varphi}v\|^2_{H^{1,\tau}(Q)} + \|e^{\tau\varphi}\mathbf{w}\|^2_{H^{1,\tau}(Q)})\\
\le& C(\|\tilde{\mathbf{f}}e^{\tau\varphi}\|^2_{L^2(Q)} + \|(\nabla\tilde{\mathbf{f}})e^{\tau\varphi}\|^2_{L^2(Q)} + \|\mathbf{v}e^{\tau\varphi}\|^2_{L^2(Q)} + \|(\nabla\mathbf{v})e^{\tau\varphi}\|^2_{L^2(Q)}\\
&+\|ve^{\tau\varphi}\|^2_{L^2(Q)} + \|(\nabla v)e^{\tau\varphi}\|^2_{L^2(Q)} + \|\mathbf{w}e^{\tau\varphi}\|^2_{L^2(Q)} + \|(\nabla\mathbf{w})e^{\tau\varphi}\|^2_{L^2(Q)}).
\end{aligned}$$

By taking $\tau > 0$ sufficiently large, we have

$$\tau(\|e^{\tau\varphi}\mathbf{v}\|^2_{H^{1,\tau}(Q)} + \|e^{\tau\varphi}v\|^2_{H^{1,\tau}(Q)} + \|e^{\tau\varphi}\mathbf{w}\|^2_{H^{1,\tau}(Q)})$$

$$\le C(\|\tilde{\mathbf{f}}e^{\tau\varphi}\|^2_{L^2(Q)} + \|(\nabla\tilde{\mathbf{f}})e^{\tau\varphi}\|^2_{L^2(Q)}). \tag{2.11}$$

On the other hand, since $[P, \chi_1]$ is a first-order differential operator which is supported in $\omega_T(\epsilon_0, 3\epsilon_0)$ and $[P, \chi_2]$ is a first-order differential operator which is supported in $Q\backslash Q_{2\delta_0}$, we obtain by (2.7) and (2.10)

$$\begin{aligned}
\|e^{\tau\varphi}\widetilde{\mathbf{f}}\|^2_{L^2(Q)} + \|e^{\tau\varphi}\nabla\widetilde{\mathbf{f}}\|^2_{L^2(Q)} \le C(&\|e^{\tau\varphi}\mathbf{f}\|^2_{L^2(Q)} + \|e^{\tau\varphi}\nabla\mathbf{f}\|^2_{L^2(Q)} + e^{C\tau}\|\mathbf{y}\|^2_{H^2(\omega_T(\epsilon_0,3\epsilon_0))}\\
&+e^{2\tau(d_0-\eta)}\|\mathbf{y}\|^2_{H^2(Q)}).
\end{aligned}$$

Thus, in terms of (2.11), we have

$$\begin{aligned}
&\tau(\|e^{\tau\varphi}\mathbf{v}\|^2_{H^{1,\tau}(Q)} + \|e^{\tau\varphi}\operatorname{div}\mathbf{v}\|^2_{H^{1,\tau}(Q)} + \|e^{\tau\varphi}\operatorname{rot}\mathbf{v}\|^2_{H^{1,\tau}(Q)})\\
\le& C(\|e^{\tau\varphi}\mathbf{f}\|^2_{L^2(Q)} + \|e^{\tau\varphi}\nabla\mathbf{f}\|^2_{L^2(Q)} + e^{C\tau}\|\mathbf{y}\|^2_{H^2(\omega_T(\epsilon_0,3\epsilon_0))} + e^{2\tau(d_0-\eta)}\|\mathbf{y}\|^2_{H^2(Q)}).
\end{aligned}$$

By applying Lemma 2.2, we obtain

$$\begin{aligned}
&\frac{1}{\tau}\|e^{\tau\varphi}\mathbf{v}\|^2_{H^{2,\tau}_x(Q)}\\
&\le C(\tau\|e^{\tau\varphi}\mathbf{v}\|^2_{H^{1,\tau}(Q)} + \|e^{\tau\varphi}\operatorname{div}\mathbf{v}\|^2_{H^{1,\tau}(Q)} + \|e^{\tau\varphi}\operatorname{rot}\mathbf{v}\|^2_{H^{1,\tau}(Q)})\\
&\le C(\|e^{\tau\varphi}\mathbf{f}\|^2_{L^2(Q)} + \|e^{\tau\varphi}\nabla\mathbf{f}\|^2_{L^2(Q)} + e^{C\tau}\|\mathbf{y}\|^2_{H^3(\omega_T(\epsilon_0,3\epsilon_0))} + e^{2\tau(d_0-\eta)}\|\mathbf{y}\|^2_{H^2(Q)}).
\end{aligned}$$

Since $\mathbf{v} = \chi\mathbf{y}$ and $\chi = 1$ in $Q_{\delta_0}(3\epsilon_0)$, we can replace $\mathbf{v}$ by $\mathbf{y}$ on the left-hand side, so that the proof of Lemma 2.3 is complete. □

2.3. Carleman estimate for a first-order partial differential operator. We consider a first-order partial differential equation

$$Av = \sum_{j=1}^{n} a_j(x)\partial_j v + a_0(x)v \equiv f(x), \quad x \in \Omega, \tag{2.12}$$

where

$$a_0 \in C(\overline{\Omega}), \quad a = (a_1, \dots, a_n) \in \left[C^1(\overline{\Omega})\right]^n, \tag{2.13}$$

and

$$|a(x) \cdot (x - x_0)| \geq c_0 > 0 \quad \text{on } \overline{\Omega}, \tag{2.14}$$

with a constant $c_0 > 0$. We set

$$\varphi_0(x) = \varphi(x, 0), \qquad x \in \Omega.$$

We have the following.

LEMMA 2.4. *In addition to* (2.14), *we assume that* $\|a_0\|_{C^2(\overline{\Omega})} \leq M$ *and* $\|a_i\|_{C^2(\overline{\Omega})} \leq M$, $1 \leq i \leq 3$. *Then there exists a constant* $\widehat{\beta} > 0$ *such that for all* $\beta > \widehat{\beta}$ *there exist* $\widehat{\tau} = \widehat{\tau}(\beta) > 0$ *and* $C = C(\widehat{\tau}, \widehat{\beta}, \Omega, \omega) > 0$ *such that*

$$\tau^2 \sum_{|\alpha| \leq 2} \int_{\Omega} |\partial_x^\alpha v|^2 e^{2\tau\varphi_0(x)} dx \leq C \sum_{|\alpha| \leq 2} \int_{\Omega} |\partial_x^\alpha f(x)|^2 e^{2\tau\varphi_0(x)} dx$$

for all $\tau > \widehat{\tau}$ *and* $v \in C_0^2(\Omega)$.

Proof. We will repeat the argument of Lemma 3.2 in [23], and, for completeness, we give the proof. We multiply both sides of (2.12) by $v(x)e^{2\tau\varphi_0(x)}$, and, by using the divergence theorem and $v \in C_0^2(\Omega)$, we obtain

$$\begin{aligned}
&\int_{\Omega} Av(x) \cdot v(x) e^{2\tau\varphi_0(x)} dx = \int_{\Omega} \nabla v(x) \cdot \left(e^{2\tau\varphi_0(x)} v(x) a(x)\right) dx + \int_{\Omega} a_0(x) |v(x)|^2 dx \\
&= -\int_{\Omega} v(x) \mathrm{div}\left(e^{2\tau\varphi_0(x)} v(x) a(x)\right) dx + \int_{\Omega} a_0(x) |v(x)|^2 e^{2\tau\varphi_0(x)} dx \\
&= -\int_{\Omega} |v|^2 e^{2\tau\varphi_0(x)} \mathrm{div}(a(x)) dx - 2\tau \int_{\Omega} |v(x)|^2 \nabla\varphi_0 \cdot a(x) e^{2\tau\varphi_0(x)} dx \\
&\quad - \int_{\Omega} e^{2\tau\varphi_0(x)} v(x) \nabla v(x) \cdot a(x) dx + \int_{\Omega} a_0(x) |v(x)|^2 e^{2\tau\varphi_0(x)} dx.
\end{aligned}$$

By (2.14), we obtain

$$|\nabla\varphi_0(x) \cdot a(x)| \geq 2\beta c_0, \quad \nabla v(x) \cdot a(x) = Av - a_0(x)v(x), \quad x \in \Omega,$$

so that the Cauchy–Schwarz inequality yields

$$\begin{aligned}
\tau \int_{\Omega} |v(x)|^2 e^{2\tau\varphi_0(x)} dx &\leq C \int_{\Omega} |(Av(x) \cdot v(x))| e^{2\tau\varphi_0(x)} dx + C \int_{\Omega} |v(x)|^2 e^{2\tau\varphi_0(x)} dx \\
&\leq \frac{C_\varepsilon}{\tau} \int_{\Omega} |Av(x)|^2 e^{2\tau\varphi_0(x)} dx + \varepsilon\tau \int_{\Omega} |v(x)|^2 e^{2\tau\varphi_0(x)} dx \\
&\quad + C \int_{\Omega} |v(x)|^2 e^{2\tau\varphi_0(x)} dx.
\end{aligned}$$

By choosing τ large and ε small, we obtain

$$\tau^2 \int_\Omega |v(x)|^2 e^{2\tau\varphi_0(x)} dx \le C \int_\Omega |Av(x)|^2 e^{2\tau\varphi_0(x)} dx. \tag{2.15}$$

Since $A(\partial_j v) = \partial_j f(x) - \sum_{k=1}^n (\partial_j a_k)\partial_k v - (\partial_j a_0)v$ and $\partial_j v|_{\partial\Omega} = 0$, we apply (2.15) to $\partial_j v$, so that

$$\begin{aligned}\tau^2 \int_\Omega |\partial_j v(x)|^2 e^{2\tau\varphi_0} dx &\le C \int_\Omega (|v(x)|^2 + |\nabla v(x)|^2) e^{2\tau\varphi_0} dx + C \int_\Omega |\partial_j f(x)|^2 e^{2\tau\varphi_0} dx \\ &\le C \int_\Omega (|f(x)|^2 + |\partial_j f(x)|^2) e^{2\tau\varphi_0} dx + C \int_\Omega |\nabla v(x)|^2 e^{2\tau\varphi_0} dx.\end{aligned}$$

Therefore

$$\tau^2 \int_\Omega |\nabla v(x)|^2 e^{2\tau\varphi_0} dx \le C \int_\Omega (|f(x)|^2 + |\nabla f(x)|^2) e^{2\tau\varphi_0} dx + C \int_\Omega |\nabla v(x)|^2 e^{2\tau\varphi_0} dx.$$

By taking $\tau_0 > 0$ sufficiently large, we have

$$\tau^2 \int_\Omega |\nabla v(x)|^2 e^{2\tau\varphi_0} dx \le C \int_\Omega \left(|f(x)|^2 + |\nabla f(x)|^2\right) e^{2\tau\varphi_0} dx.$$

Next we have

$$A(\partial_k \partial_\ell v) = \partial_k \partial_\ell f - \sum_{j=1}^3 (\partial_k a_j)(\partial_\ell \partial_j v) + (\partial_\ell a_j)(\partial_k \partial_j v) + K(v, \nabla v),$$

where K is a linear operator of v and ∇v with bounded coefficients in Ω. Noting that $\partial_k \partial_\ell v = 0$ on $\partial\Omega$, we apply (2.15) to $\partial_k \partial_\ell v$, and we can complete the proof of Lemma 2.4. □

2.4. Weak observation estimate. Let $\mathbf{v}$ satisfy

$$\rho(x)\partial_t^2 \mathbf{v} - \mathcal{L}_{\lambda,\mu}(x, \partial_x)\mathbf{v} = R(x,t) \quad \text{in } Q \equiv \Omega \times (-T, T)$$

and

$$\mathbf{v}(x,t) = 0 \quad \text{on } \Sigma \equiv \Gamma \times (-T, T),$$

where we assume that

$$R(x,t) = 0, \quad (x,t) \in \omega \times (-T, T).$$

The following proposition shows the stability in the unique continuation of solutions of the Lamé system from lateral boundary data on an arbitrarily small part Γ_1 of $\partial\Omega$.

LEMMA 2.5. *For sufficiently large $T > 0$, there exists a constant $C > 0$ such that*

$$\|\mathbf{v}\|^2_{H^2(\omega_T(\epsilon_0, 3\epsilon_0))} \le C \left[\log\left(2 + \frac{C}{\sum_{|\alpha|=1}^2 \|\partial_x^\alpha \mathbf{v}\|^2_{L^2(\Sigma_1)}}\right)\right]^{-1}.$$

In our case, the corresponding uniqueness is already proved in Bellassoued [3] and Eller, Isakov, Nakamura, and Tataru [12]. For similar stability results for a scalar hyperbolic equation, see Bellassoued and Yamamoto [6] and Robbiano [44]. The proof of the lemma is given in section 4.

3. Proof of the main result. This section is devoted to the proof of Theorem 1.0. The key is the combination of Lemma 2.5 and the existing method (e.g., [4]).

3.1. Notation and preliminary estimates. For simplicity, we set

$$\mathbf{u} = \mathbf{u}(\lambda, \mu, \rho, \Phi, \Psi, \mathbf{g}), \quad \widetilde{\mathbf{u}} = \mathbf{u}(\widetilde{\lambda}, \widetilde{\mu}, \widetilde{\rho}, \Phi, \Psi, \mathbf{g}),$$

and

$$\mathbf{u}_* = \mathbf{u} - \widetilde{\mathbf{u}}, \quad \rho_* = \rho - \widetilde{\rho}, \quad \lambda_* = \lambda - \widetilde{\lambda}, \quad \mu_* = \mu - \widetilde{\mu}.$$

Then we obtain

$$\widetilde{\rho}\partial_t^2 \mathbf{u}_*(x,t) = \mathcal{L}_{\widetilde{\lambda},\widetilde{\mu}}(x,\partial_x)\mathbf{u}_*(x,t) + \mathbf{f}(x,t) \quad \text{in} \quad Q, \tag{3.1}$$

$$\mathbf{u}_*(x,0) = \partial_t \mathbf{u}_*(x,0) = 0, \quad x \in \Omega, \tag{3.2}$$

and

$$\mathbf{u}_* = 0 \qquad \text{on} \quad \Sigma = \Gamma \times (-T,T). \tag{3.3}$$

Here we set

$$\begin{aligned}\mathbf{f}(x,t) &= -\rho_*(x)\partial_t^2 \mathbf{u}(x,t) + (\lambda_*(x) + \mu_*(x))\nabla(\operatorname{div}\mathbf{u})(x,t) + \mu_*(x)\Delta\mathbf{u}(x,t)\\ &\quad + (\operatorname{div}\mathbf{u})(x,t)\nabla\lambda_*(x) + \left(\nabla\mathbf{u}(x,t) + (\nabla\mathbf{u}(x,t))^T\right)\nabla\mu_*(x)\\ &= -\left(\rho_*(x)\partial_t^2\mathbf{u} - \mathcal{L}_{\lambda_*,\mu_*}(x,\partial_x)\mathbf{u}\right).\end{aligned}$$

Moreover we set

$$\mathbf{y}_1(x,t) = \partial_t^2 \mathbf{u}_*(x,t), \quad \mathbf{y}_2(x,t) = \partial_t^3 \mathbf{u}_*(x,t), \quad \mathbf{y}_3(x,t) = \partial_t^4 \mathbf{u}_*(x,t).$$

Then we have

$$\widetilde{\rho}\partial_t^2 \mathbf{y}_j - \mathcal{L}_{\widetilde{\lambda},\widetilde{\mu}}(x,\partial_x)\mathbf{y}_j = \partial_t^{j+2}\mathbf{f}. \tag{3.4}$$

For simplicity, it is convenient to use the following notation:

$$\mathcal{D} = \sum_{j=2}^{4} \|\partial_t^j \mathbf{u}_*\|^2_{H^2(\omega_T(\epsilon_0,3\epsilon_0))} = \sum_{j=1}^{3} \|\mathbf{y}_j\|^2_{H^2(\omega_T(\epsilon_0,3\epsilon_0))}$$

and

$$\mathcal{E} = \sum_{|\alpha|\le 1} \|e^{\tau\varphi}\partial_x^\alpha \rho_*\|^2_{L^2(Q)} + \sum_{|\alpha|\le 2} (\|e^{\tau\varphi}\partial_x^\alpha \lambda_*\|^2_{L^2(Q)} + \|e^{\tau\varphi}\partial_x^\alpha \mu_*\|^2_{L^2(Q)}),$$

$$\mathbf{z}_1(x) = \mathbf{y}_1(x,0), \quad \mathbf{z}_2(x) = \mathbf{y}_2(x,0).$$

We have the following.

LEMMA 3.1. *There exists a constant $\widehat{\beta} > 0$ such that for all $\beta \ge \widehat{\beta}$ there exist $\widehat{\tau}$ and $C > 0$ such that*

$$\sum_{|\alpha|\le 2} \int_{\Omega\setminus\omega(3\epsilon_0)} \tau^{4-2|\alpha|} \left(|\partial_x^\alpha \mathbf{z}_1(x)|^2 + |\partial_x^\alpha \mathbf{z}_2(x)|^2\right) e^{2\tau\varphi_0(x)} dx \le C(\tau^4 e^{2\tau(d_0-\eta)} + \tau^2\mathcal{E} + e^{C\tau}\mathcal{D})$$

for all large $\tau > \widehat{\tau}$.

Here we recall that $\varphi_0(x) = \varphi(x,0)$, $x \in \Omega$.

Proof. Noting that $\mathbf{u}_* \in W^{8,\infty}(Q)$, we can apply Lemma 2.3 to (3.4), so that we have

$$\frac{1}{\tau}\left\|e^{\tau\varphi}\mathbf{y}_j\right\|^2_{H^{2,\tau}_x(Q_{\delta_0}(3\epsilon_0))} \le C(\|e^{\tau\varphi}\partial_t^{j+2}\mathbf{f}\|^2_{L^2(Q)} + \|e^{\tau\varphi}\nabla(\partial_t^{j+2}\mathbf{f})\|^2_{L^2(Q)}$$
$$+e^{C\tau}\left\|\mathbf{y}_j\right\|^2_{H^2(\omega_T(\epsilon_0,3\epsilon_0))} + e^{2\tau(d_0-\eta)}\left\|\mathbf{y}_j\right\|^2_{H^2(Q)}), \quad j = 1,2,3. \tag{3.5}$$

By the definition of $\mathbf{f}$, we have

$$\left|\nabla(\partial_t^{j+2}\mathbf{f}(x,t))\right|^2 \le C\left(\sum_{|\alpha|\le 1}|\partial_x^\alpha \rho_*(x)|^2 + \sum_{|\alpha|\le 2}\left(|\partial_x^\alpha\lambda_*(x)|^2 + |\partial_x^\alpha\mu_*(x)|^2\right)\right)$$

and

$$\left|\partial_t^{j+2}\mathbf{f}(x,t)\right|^2 \le C\left(|\rho_*(x)|^2 + \sum_{|\alpha|\le 1}\left(|\partial_x^\alpha\lambda_*(x)|^2 + |\partial_x^\alpha\mu_*(x)|^2\right)\right) \quad \text{in } Q.$$

Therefore

$$\sum_{j=1}^{5}\left(\|e^{\tau\varphi}\nabla(\partial_t^j\mathbf{f})\|^2_{L^2(Q)} + \|e^{\tau\varphi}\partial_t^j\mathbf{f}\|^2_{L^2(Q)}\right) \le C\mathcal{E}.$$

Thus (3.5) implies that

$$\frac{1}{\tau}\sum_{j=1}^{3}\left\|e^{\tau\varphi}\mathbf{y}_j\right\|^2_{H^{2,\tau}_x(Q_{\delta_0}(3\epsilon_0))} \le C(e^{2\tau(d_0-\eta)} + \mathcal{E} + e^{C\tau}\mathcal{D}). \tag{3.6}$$

On the other hand, we introduce a cutoff function $\chi_2 \in C_0^\infty(\mathbb{R})$ satisfying $0 \le \chi_2 \le 1$ and

$$\chi_2(t) = \begin{cases} 0, & |t| > T - \delta_0, \\ 1, & |t| < T - 2\delta_0. \end{cases}$$

Then we have

$$\int_{\Omega\setminus\omega(3\epsilon_0)}|\partial_x^\alpha\mathbf{z}_j(x)|^2 e^{2\tau\varphi_0(x)}dx = \int_{-T}^{0}\frac{\partial}{\partial t}\left(\int_{\Omega\setminus\omega(3\epsilon_0)}|(\partial_x^\alpha\mathbf{y}_j)(x,t)|^2\chi_2(t)^2e^{2\tau\varphi}dx\right)dt$$
$$= 2\int_{-T}^{0}\int_{\Omega\setminus\omega(3\epsilon_0)}(\partial_x^\alpha\mathbf{y}_{j+1}\cdot\partial_x^\alpha\mathbf{y}_j)\chi_2^2(t)e^{2\tau\varphi}dxdt$$
$$+2\tau\int_{-T}^{0}\int_{\Omega\setminus\omega(3\epsilon_0)}|\partial_x^\alpha\mathbf{y}_j|^2\chi_2^2(\partial_t\varphi)e^{2\tau\varphi}dxdt$$
$$+2\int_{-T}^{0}\int_{\Omega\setminus\omega(3\epsilon_0)}|\partial_x^\alpha\mathbf{y}_j|^2\chi_2'(t)\chi_2(t)e^{2\tau\varphi}dxdt, \quad j = 1,2.$$

Since $\chi_2'(t)$ is supported in $[-T,-T+2\delta_0]\cup[T-2\delta_0,T]$, by the Cauchy–Schwarz

inequality we obtain

$$\sum_{j=1}^{2}\sum_{|\alpha|\le 2}\int_{\Omega\setminus\omega(3\epsilon_0)}\tau^{4-2|\alpha|}\left|\partial_x^\alpha \mathbf{z}_j(x)\right|^2 e^{2\tau\varphi_0(x)}dx$$

$$\le C\left(\tau\sum_{j=1}^{3}\|e^{\tau\varphi}\mathbf{y}_j\|^2_{H_x^{2,\tau}(Q_{\delta_0}(3\epsilon_0))}+\tau^4 e^{2\tau(d_0-\eta)}\right). \tag{3.7}$$

Therefore (3.6) completes the proof of the lemma. □

3.2. Estimation for the two Lamé coefficients. We will consider a first-order partial differential equations in λ_*, μ_*, and ρ_*. That is, by (3.1), (3.2), and $\mathbf{u},\mathbf{v}\in W^{8,\infty}(Q)$, we have

$$\widetilde{\rho}\partial_t^2\mathbf{u}_*(x,0)=\mathbf{f}(x,0),\quad \widetilde{\rho}\partial_t^3\mathbf{u}_*(x,0)=\partial_t\mathbf{f}(x,0). \tag{3.8}$$

For simplicity, for $x\in\overline{\Omega}$, we set

$$\mathbf{a}=\begin{pmatrix}-\frac{1}{\rho}\mathcal{L}_{\lambda,\mu}(x,\partial_x)\Phi\\ \\ -\frac{1}{\rho}\mathcal{L}_{\lambda,\mu}(x,\partial_x)\Psi\end{pmatrix},\quad \mathbf{b}_1=\begin{pmatrix}\operatorname{div}\Phi\\0\\0\\\operatorname{div}\Psi\\0\\0\end{pmatrix},$$

$$\mathbf{b}_2=\begin{pmatrix}0\\\operatorname{div}\Phi\\0\\0\\\operatorname{div}\Psi\\0\end{pmatrix},\quad \mathbf{b}_3=\begin{pmatrix}0\\0\\\operatorname{div}\Phi\\0\\0\\\operatorname{div}\Psi\end{pmatrix},$$

$$(\mathbf{d}_1,\mathbf{d}_2,\mathbf{d}_3)=\begin{pmatrix}\nabla\Phi+(\nabla\Phi)^T\\ \\ \nabla\Psi+(\nabla\Psi)^T\end{pmatrix},\quad \mathbf{G}=\begin{pmatrix}\widetilde{\rho}\mathbf{z}_1(x)-(\lambda_*+\mu_*)\nabla(\operatorname{div}\Phi)-\mu_*\Delta\Phi\\ \\ \widetilde{\rho}\mathbf{z}_2(x)-(\lambda_*+\mu_*)\nabla(\operatorname{div}\Psi)-\mu_*\Delta\Psi\end{pmatrix}.$$

Then we can rewrite (3.8) as

$$\mathbf{a}\rho_*+\sum_{k=1}^{3}\mathbf{b}_k\partial_k\lambda_*(x)=\mathbf{G}-\sum_{k=1}^{3}\mathbf{d}_k\partial_k\mu_*(x). \tag{3.9}$$

In view of (3.9), we show the following.

LEMMA 3.2. *Let assumptions* $(\mathcal{H}_1)$*–*$(\mathcal{H}_2)$ *hold true. Then there exists a constant* $C>0$ *such that the following estimate holds:*

$$\sum_{|\alpha|\le 2}\int_\Omega\left(\left|\partial_x^\alpha\mu_*(x)\right|^2+\left|\partial_x^\alpha\lambda_*(x)\right|^2\right)e^{2\tau\varphi_0(x)}dx\le C(\tau^2e^{2\tau(d_0-\eta)}+\mathcal{E}+e^{C\tau}\mathcal{D})$$

provided that τ *is large.*

Proof. By (3.9), for $j_1\in\{1,2,3,4,5,6\}$, we have

$$\{\mathbf{a}\}_{j_1}\rho_*+\sum_{k=1}^{3}\{\mathbf{b}_k\}_{j_1}\partial_k\lambda_*(x)=\{\mathbf{G}\}_{j_1}-\sum_{k=1}^{3}\{\mathbf{d}_k\}_{j_1}\partial_k\mu_*(x)\quad\text{on }\overline{\Omega}.$$

For any fixed $x \in \overline{\Omega}$, we regard this as a system of five linear equations with respect to four unknowns ρ_*, $\partial_1\lambda_*$, $\partial_2\lambda_*$, $\partial_3\lambda_*$, and so, for the existence of solutions, we need the consistency of the coefficients, that is,

$$\det_{j_1}\left(\mathbf{a}, \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \mathbf{G} - \sum_{k=1}^{3} \mathbf{d}_k \partial_k \mu_*\right) = 0 \quad \text{on } \overline{\Omega}.$$

Hence

$$\sum_{k=1}^{3} \det_{j_1}(\mathbf{a}, \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \mathbf{d}_k)\partial_k\mu_* = \det_{j_1}(\mathbf{a}, \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \mathbf{G}) := f \quad \text{on } \overline{\Omega} \tag{3.10}$$

by the linearity of the determinant with respect to the column vectors. Here we note that

$$\sum_{|\alpha|\le 2} |\partial_x^\alpha f(x)| \le C\left(\sum_{|\alpha|\le 2}(|\partial_x^\alpha \mathbf{z}_1(x)| + |\partial_x^\alpha \mathbf{z}_2(x)|) + \sum_{|\alpha|\le 2}(|\partial_x^\alpha \lambda_*(x)| + |\partial_x^\alpha \mu_*(x)|)\right). \tag{3.11}$$

By (3.10) the function $\mu_* \in C_0^2(\Omega)$ solves the following first-order partial differential equation:

$$\sum_{k=1}^{3} a_k(x)\partial_k\mu_*(x) = f(x), \quad x \in \Omega,$$

where

$$a_k(x) = \det_{j_1}(\mathbf{a}, \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \mathbf{d}_k).$$

In view of $(\mathcal{H}_1)$, we can apply Lemma 2.4 to $v = \mu_*$, and we obtain

$$\tau^2 \sum_{|\alpha|\le 2} \int_\Omega |\partial_x^\alpha \mu_*|^2 e^{2\tau\varphi_0(x)} dx \le C \sum_{|\alpha|\le 2} \int_{\Omega\setminus\omega(3\epsilon_0)} |\partial_x^\alpha f(x)|^2 e^{2\tau\varphi_0(x)} dx,$$

where we have used $\mu_* \equiv 0$ in $\omega \supset \omega(3\epsilon_0)$. Hence (3.11) and Lemma 3.1 yield

$$\begin{aligned}
&\tau^2 \sum_{|\alpha|\le 2} \int_\Omega |\partial_x^\alpha \mu_*(x)|^2 e^{2\tau\varphi_0(x)} dx \le C \sum_{j=1}^{2} \sum_{|\alpha|\le 2} \int_{\Omega\setminus\omega(3\epsilon_0)} |\partial_x^\alpha \mathbf{z}_j(x)|^2 e^{2\tau\varphi_0(x)} dx \\
&+C \sum_{|\alpha|\le 2} \int_\Omega \left(|\partial_x^\alpha \lambda_*(x)|^2 + |\partial_x^\alpha \mu_*(x)|^2\right) e^{2\tau\varphi_0(x)} dx \\
&\le C \sum_{|\alpha|\le 2} \int_\Omega \left(|\partial_x^\alpha \lambda_*(x)|^2 + |\partial_x^\alpha \mu_*(x)|^2\right) e^{2\tau\varphi_0(x)} dx + C(e^{C\tau}\mathcal{D} + \tau^4 e^{2\tau(d_0-\eta)} + \tau^2\mathcal{E})
\end{aligned}$$

for all large $\tau > 0$. Similarly, by assumption $(\mathcal{H}_2)$ we can argue for λ_* and obtain

$$\begin{aligned}
&\tau^2 \sum_{|\alpha|\le 2} \int_\Omega |\partial_x^\alpha \lambda_*(x)|^2 e^{2\tau\varphi_0(x)} dx \\
&\le C \sum_{|\alpha|\le 2} \int_\Omega \left(|\partial_x^\alpha \lambda_*(x)|^2 + |\partial_x^\alpha \mu_*(x)|^2\right) e^{2\tau\varphi_0(x)} dx + C(e^{C\tau}\mathcal{D} + \tau^4 e^{2\tau(d_0-\eta)} + \tau^2\mathcal{E}).
\end{aligned}$$

Hence by adding the two inequalities, we have

$$\begin{aligned}
&\tau^2 \sum_{|\alpha|\le 2} \int_\Omega \left(|\partial_x^\alpha \mu_*(x)|^2 + |\partial_x^\alpha \lambda_*(x)|^2 \right) e^{2\tau\varphi_0(x)} dx \\
&\le C \sum_{|\alpha|\le 2} \int_\Omega \left(|\partial_x^\alpha \mu_*(x)|^2 + |\partial_x^\alpha \lambda_*(x)|^2 \right) e^{2\tau\varphi_0(x)} dx \\
&\quad + C(\tau^4 e^{2\tau(d_0-\eta)} + \tau^2 \mathcal{E} + e^{C\tau}\mathcal{D})
\end{aligned}$$

for all large $\tau > 0$. By taking $\tau > 0$ large, we can absorb the first term on the right-hand side into the left-hand side, and the proof is complete. ∎

3.3. Estimation for the density.

LEMMA 3.3. *Let assumptions* $(\mathcal{H}_1)$–$(\mathcal{H}_2)$ *hold true. Then there exists a constant* $C > 0$ *such that*

$$\sum_{|\alpha|\le 1} \int_\Omega |\partial_x^\alpha \rho_*(x)|^2 e^{2\tau\varphi_0(x)} dx \le C(\tau^2 e^{2\tau(d_0-\eta)} + \mathcal{E} + e^{C\tau}\mathcal{D}),$$

provided that τ *is large.*

Proof. By (3.9), we have

$$\mathbf{a}\rho_*(x) = -\sum_{k=1}^3 \mathbf{b}_k \partial_k \lambda_*(x) + \mathbf{G} - \sum_{k=1}^3 \mathbf{d}_k \partial_k \mu_*(x), \quad x \in \Omega.$$

Moreover, by $(\mathcal{H}_1)$–$(\mathcal{H}_2)$, we see that $|\mathbf{a}(x)| > 0$ for $x \in \overline{\Omega}$, so that

$$|\rho_*(x)| \le C\,|\mathbf{G}(x)| + C \sum_{|\alpha|\le 1} \left(|\partial_x^\alpha \lambda_*(x)| + |\partial_x^\alpha \mu_*(x)| \right).$$

Similarly we have

$$|\partial_j \rho_*(x)| \le C \left(|\mathbf{G}(x)| + |\nabla \mathbf{G}(x)| + \sum_{|\alpha|\le 2} \left(|\partial_x^\alpha \lambda_*(x)| + |\partial_x^\alpha \mu_*(x)| \right) \right).$$

Hence

$$\begin{aligned}
\sum_{|\alpha|\le 1} \int_\Omega |\partial_x^\alpha \rho_*(x)|^2 e^{2\tau\varphi_0(x)} dx &\le C \sum_{|\alpha|\le 1} \int_{\Omega\setminus\omega(3\epsilon_0)} |\partial_x^\alpha \mathbf{G}(x)|^2 e^{2\tau\varphi_0(x)} dx \\
&\quad + C \sum_{|\alpha|\le 2} \int_\Omega \left(|\partial_x^\alpha \lambda_*(x)|^2 + |\partial_x^\alpha \mu_*(x)|^2 \right) e^{2\tau\varphi_0(x)} dx.
\end{aligned}$$

Furthermore we see that

$$\sum_{|\alpha|\le 1} |\partial_x^\alpha \mathbf{G}(x)|^2 \le C \sum_{|\alpha|\le 1} \left(|\partial_x^\alpha \mathbf{z}_1(x)|^2 + |\partial_x^\alpha \mathbf{z}_2(x)|^2 \right) + C \sum_{|\alpha|\le 1} \left(|\partial_x^\alpha \lambda_*(x)|^2 + |\partial_x^\alpha \mu_*(x)|^2 \right).$$

Therefore

$$\begin{aligned}
\sum_{|\alpha|\le 1} \int_\Omega |\partial_x^\alpha \rho_*(x)|^2 e^{2\tau\varphi_0(x)} dx &\le C \sum_{|\alpha|\le 1} \int_{\Omega\setminus\omega(3\epsilon_0)} \left(|\partial_x^\alpha \mathbf{z}_1(x)|^2 + |\partial_x^\alpha \mathbf{z}_1(x)|^2 \right) e^{2\tau\varphi_0(x)} dx \\
&\quad + C \sum_{|\alpha|\le 2} \int_\Omega \left(|\partial_x^\alpha \lambda_*(x)|^2 + |\partial_x^\alpha \mu_*(x)|^2 \right) e^{2\tau\varphi_0(x)} dx.
\end{aligned}$$

By Lemmas 3.1 and 3.2, we obtain the conclusion of the lemma. ∎

3.4. Completion of the proof of the main result. In terms of Lemmas 3.2 and 3.3, we will now complete the proof of Theorem 1.0.

Since $\varphi(x,0) > \varphi(x,t)$ for $t \neq 0$, by the Lebesgue theorem, we have

$$\sum_{|\alpha|\leq 1}\int_Q |\partial_x^\alpha \rho_*(x)|^2 e^{2\tau\varphi} dxdt = \sum_{|\alpha|\leq 1}\int_\Omega |\partial_x^\alpha \rho_*(x)|^2 e^{2\tau\varphi(x,0)} \left(\int_{-T}^T e^{2\tau(\varphi(x,t)-\varphi(x,0))}dt\right) dx$$
$$= o(1)\sum_{|\alpha|\leq 1}\int_\Omega |\partial_x^\alpha \rho_*(x)| e^{2\tau\varphi(x,0)} dx$$

as $\tau \to \infty$. Similarly we obtain

$$\sum_{|\alpha|\leq 2}\int_Q \left(|\partial_x^\alpha \lambda_*(x)|^2 + |\partial_x^\alpha \mu_*(x)|^2\right) e^{2\tau\varphi} dxdt$$
$$= o(1)\sum_{|\alpha|\leq 2}\int_\Omega \left(|\partial_x^\alpha \lambda_*(x)|^2 + |\partial_x^\alpha \mu_*(x)|^2\right) e^{2\tau\varphi(x,0)} dx,$$

as $\tau \to \infty$. We set

$$\mathcal{E}_0 = \sum_{|\alpha|\leq 1} \|e^{\tau\varphi_0}\partial_x^\alpha \rho_*\|^2_{L^2(\Omega)} + \sum_{|\alpha|\leq 2} (\|e^{\tau\varphi_0}\partial_x^\alpha \lambda_*\|^2_{L^2(\Omega)} + \|e^{\tau\varphi_0}\partial_x^\alpha \mu_*\|^2_{L^2(\Omega)}).$$

Therefore, by Lemmas 3.2 and 3.3, we obtain

$$\mathcal{E}_0 \leq C(\tau^2 e^{2\tau(d_0-\eta)} + e^{C\tau}\mathcal{D}) + o(1)\mathcal{E}_0$$

for all large $\tau > 0$.

By (2.6) and $\sup_{\tau>0}(\tau^2 e^{-\tau\eta}) < \infty$, we have

$$\sum_{|\alpha|\leq 1}\int_\Omega |\partial_x^\alpha \rho_*(x)|^2 dx + \sum_{|\alpha|\leq 2}\int_\Omega \left(|\partial_x^\alpha \lambda_*(x)|^2 + |\partial_x^\alpha \mu_*(x)|^2\right) dx \leq Ce^{-\tau\eta} + Ce^{C\tau}\mathcal{D} \tag{3.12}$$

for all large $\tau > \tau_0$. We replace C by $Ce^{C\tau_0}$, and (3.12) holds for any $\tau > 0$. We may assume that $\mathcal{D} < 1$.

Now we choose $\tau > 0$ such that

$$e^{C\tau}\mathcal{D} = e^{-\tau\eta},$$

that is,

$$\tau = -\frac{1}{\eta + C}\log \mathcal{D}.$$

Therefore (3.12) implies

$$\|\rho_*\|^2_{H^1(\Omega)} + \|\lambda_*\|^2_{H^2(\Omega)} + \|\mu_*\|^2_{H^2(\Omega)} \leq 2C\mathcal{D}^{\frac{\eta}{\eta+C}}$$

and

$$\|\rho_*\|^2_{H^1(\Omega)} + \|\lambda_*\|^2_{H^2(\Omega)} + \|\mu_*\|^2_{H^2(\Omega)} \leq C\|\mathbf{u}_*\|^{2\sigma}_{H^6(\omega_T(\epsilon_0,3\epsilon_0))},$$

with $\sigma = \frac{\eta}{\eta+C} \in (0,1)$. Then, noting that $\|\mathbf{u}_*\|_{W^{8,\infty}(Q)} \le M_1$ and the interpolation inequality

$$\|\mathbf{u}_*\|_{H^6(\omega_T(\epsilon_0,3\epsilon_0))} \le C\|\mathbf{u}_*\|^{\frac{1}{3}}_{H^2(\omega_T(\epsilon_0,3\epsilon_0))}\|\mathbf{u}_*\|^{\frac{2}{3}}_{H^8(\omega_T(\epsilon_0,3\epsilon_0))},$$

we apply Lemma 2.5 to $\mathbf{v} = \mathbf{u}_*$, so that

$$\|\rho_*\|^2_{H^1(\Omega)} + \|\lambda_*\|^2_{H^2(\Omega)} + \|\mu_*\|^2_{H^2(\Omega)} \le C\left[\log\left(2 + \frac{C}{\epsilon(\Sigma_1)}\right)\right]^{-2\kappa},$$

where $\kappa \in (0,1)$. Thus the proof of Theorem 1.0 is complete.

4. Proof of Lemma 2.5. We will now prove Lemma 2.5. This will be done in terms of the FBI transformation. For $(\lambda,\mu,\rho) \in \Lambda$, let us recall that $\mathbf{v}$ is a given solution to

$$\rho(x)\partial_t^2\mathbf{v} - \mathcal{L}_{\lambda,\mu}(x,\partial_x)\mathbf{v} = R(x,t) \quad \text{in } Q \equiv \Omega\times(-T,T), \tag{4.1}$$

with the Dirichlet boundary condition

$$\mathbf{v}(x,t) = 0 \quad \text{on } \Sigma \equiv \Gamma\times(-T,T). \tag{4.2}$$

Here and henceforth we assume that

$$R(x,t) = 0, \quad (x,t) \in \omega\times(-T,T). \tag{4.3}$$

4.1. Preliminary and elliptic estimation. Since we can choose $T > 0$ sufficiently large, we may assume that $T > 2$. Denote for $r > 0$

$$\Omega_r = \Omega\times(-r,r); \quad \omega_r(\epsilon_0,3\epsilon_0) = \omega(\epsilon_0,3\epsilon_0)\times(-r,r);$$
$$\Sigma_r = \Gamma\times(-r,r), \quad \Sigma_{1,r} = \Gamma_1\times(-r,r).$$

We introduce $\theta \in C_0^\infty(\mathbb{R})$ to be a cutoff function defined by

$$\theta(t) = \begin{cases} 1, & |t| \le (T-2), \\ 0, & |t| \ge (T-1). \end{cases}$$

Let $\gamma > 0$. We introduce the partial FBI transformation $\mathcal{F}_\gamma$. It is defined for $\mathbf{u} \in \{\mathcal{S}(\mathbb{R}^4)\}^3$, the space of rapidly decreasing functions, by

$$\mathbf{u}_{\gamma,t}(x,s) = \mathcal{F}_\gamma\mathbf{u}(x,z) = \sqrt{\frac{\gamma}{2\pi}}\int_{\mathbb{R}} e^{-\frac{\gamma}{2}(z-y)^2}\theta(y)\mathbf{u}(x,y)dy, \quad z = t+is.$$

Then

$$|D_x^\alpha\mathcal{F}_\gamma\mathbf{u}(x,z)| \le C\sqrt{\frac{\gamma}{2\pi}}e^{\gamma s^2}e^{-\frac{\gamma}{2}(\mathrm{dist}(t,supp(\theta\mathbf{u})))^2}\sup_{x\in\mathbb{R}^3}\|D_x^\alpha\mathbf{u}(x,.)\|^2_{L^2(\mathbb{R})}$$

for any $\mathbf{u} \in C_0^\infty(\mathbb{R}^3\times\mathbb{R})$ (see [45]).

Henceforth C_j, C denote generic constants which are independent of λ,γ,r,τ. Next we assume that T is sufficiently large, $s \in [-3r,3r]$, and $t \in \left[-\frac{T}{2},\frac{T}{2}\right]$. In particular we assume that $\frac{T}{2} > r$. We introduce a cutoff function χ_3 satisfying $0 \le \chi_3 \le 1$, $\chi_3 \in C_0^\infty(\mathbb{R}^3)$, and

$$\chi_3(x) = \begin{cases} 1, & \text{if } x \in \omega(6\epsilon_0), \\ 0, & \text{if } x \in \Omega\setminus\omega(7\epsilon_0). \end{cases}$$

Let $\mathbf{v}(x,t)$ satisfy (4.1). By setting $\mathbf{u}(x,t) = \chi_3(x)\mathbf{v}(x,t)$ and noting that $R(x,t)$ is zero in $\omega(7\epsilon_0)$, we obtain

$$\rho(x)\partial_t^2\mathbf{u} - \mathcal{L}_{\lambda,\mu}(x,\partial_x)\mathbf{u} = -\left[\mathcal{L}_{\lambda,\mu}(x,\partial_x),\chi_3\right]\mathbf{v} \quad \text{in } Q \tag{4.4}$$

and

$$\mathbf{u}(x,t) = 0 \quad \text{on } \Sigma. \tag{4.5}$$

In connection with the operator $\rho(x)\partial_t^2 - \mathcal{L}_{\lambda,\mu}(x,\partial_x)$, we define an elliptic operator by

$$Q_{\rho,\lambda,\mu} = \rho(x)\partial_s^2 + \mathcal{L}_{\lambda,\mu}(x,\partial_x).$$

Noting that

$$\partial_s \int_{\mathbb{R}} e^{-\frac{\gamma}{2}(is+t-y)^2}\theta(y)\mathbf{u}(x,y)dy = i\int_{\mathbb{R}} e^{-\frac{\gamma}{2}(z-y)^2}\partial_y\left[\theta(y)\mathbf{u}(x,y)\right]dy,$$

by integration by parts, we have

$$\begin{aligned} &Q_{\rho,\lambda,\mu}\mathbf{u}_{\gamma,t}(x,s) = F_{\gamma,t}(x,s) + G_{\gamma,t}(x,s) := \mathbf{f}(x,s), (x,s) \in \ \Omega_{3r}, \\ &\mathbf{u}_{\gamma,t}(x,s) = 0, \qquad\qquad (x,s) \in \ \Sigma_{3r}, \end{aligned} \tag{4.6}$$

where

$$F_{\gamma,t}(x,s) = -\sqrt{\frac{\gamma}{2\pi}}\int_{\mathbb{R}} e^{-\frac{\gamma}{2}(z-y)^2}\left(2\theta'(y)\partial_y\mathbf{u}(x,y) + \theta''(y)\mathbf{u}(x,y)\right)dy$$

and

$$G_{\gamma,t}(x,s) = \sqrt{\frac{\gamma}{2\pi}}\int_{\mathbb{R}} e^{-\frac{\gamma}{2}(z-t)^2}\theta(y)\left[\mathcal{L}_{\lambda,\mu}(x,\partial_x),\chi\right]\mathbf{v}(x,y)dy.$$

Since $\theta' = \frac{d\theta}{dy}$ and $\theta'' = \frac{d^2\theta}{dy^2}$ are supported in $T-2 \le |y| \le T-1$, there exists $\eta > 0$, independent of T, such that

$$\|F_{\gamma,t}\|_{H^1(\Omega_{3r})} \le Ce^{-\eta\gamma T}\|\mathbf{u}\|_{H^2(Q)} \quad \forall t \in \left[-\frac{T}{2},\frac{T}{2}\right]. \tag{4.7}$$

Moreover there exists $C_1 > 0$, independent of T, such that

$$\|\mathbf{u}_{\gamma,t}\|_{H^2(\Omega_{3r})} \le Ce^{C_1\gamma}\|\mathbf{u}\|_{H^2(Q)} \quad \forall t \in \left[-\frac{T}{2},\frac{T}{2}\right]. \tag{4.8}$$

By the definition, we easily obtain

$$G_{\gamma,t}(x,s) = 0 \quad \forall x \in \omega(6\epsilon_0). \tag{4.9}$$

Let K be a compact set in $\overline{\Omega}\times(-3r,3r)$ and let $\psi(x,s)$ be a C^∞ function satisfying $\nabla_{x,s}\psi(x,s) \neq 0$ on K. Let

$$\varphi(x,s) = e^{-\beta\psi(x,s)},$$

where $\beta > 0$ is sufficiently large.

Henceforth we set

$$\|u\|^2_{H^j_\tau(\Omega_{3r})} = \sum_{|\alpha|\le j} \tau^{2j-2|\alpha|} \|\partial^\alpha u\|^2_{L^2(\Omega_{3r})}$$

and

$$\|u\|^2_{H^j_\tau(\Sigma_{3r})} = \sum_{|\alpha|\le j} \tau^{2j-2|\alpha|} \|\partial^\alpha u\|^2_{L^2(\Sigma_{3r})}.$$

We note that the right-hand sides include not only tangential derivatives but also normal derivatives on Σ_{3r}.

Consider a scalar second-order elliptic operator

$$P(x,D) = \rho(x)\partial_s^2 + \sum_{j,k=1}^n a_{jk}(x)\partial_j\partial_k + \sum_{j=1}^n b_j(x)\partial_j + c.$$

If we assume that P has C^2 coefficients, then the following Carleman estimate holds true:

$$\begin{aligned} \frac{C}{\tau}\|e^{\tau\varphi}u\|^2_{H^2_\tau(\Omega_{3r})} &\le \|e^{\tau\varphi}Pu\|^2_{L^2(\Omega_{3r})} + \|e^{\tau\varphi}u\|^2_{H^2_\tau(\Sigma_{3r})}, \\ C\tau\|e^{\tau\varphi}u\|^2_{H^1_\tau(\Omega_{3r})} &\le \|e^{\tau\varphi}Pu\|^2_{L^2(\Omega_{3r})} + \tau\|e^{\tau\varphi}u\|^2_{H^1_\tau(\Sigma_{3r})}, \end{aligned} \tag{4.10}$$

whenever $u \in C_0^\infty(K)$ and $\tau > \tau_0$ (see, for example, [37] and [19], respectively, for the first and second Carleman estimates).

By (4.10) we can derive the following Carleman estimate for the elliptic system (4.6).

LEMMA 4.1. *There exist $\tau_0 > 0$ and $C > 0$ such that*

$$\frac{C}{\tau}\|e^{\tau\varphi}\mathbf{u}\|^2_{H^2_\tau(\Omega_{3r})} \le \|e^{\tau\varphi}Q_{\rho,\lambda,\mu}\mathbf{u}\|^2_{H^1(\Omega_{3r})} + \|e^{\tau\varphi}\mathbf{u}\|^2_{H^2_\tau(\Sigma_{3r})},$$

whenever $\mathbf{u} \in C_0^\infty(K)$ and all $\tau > \tau_0$.

Proof. In order to prove Lemma 4.1 we will extend system (4.6) for three unknown functions u_1, u_2, u_3 to a new one for four unknown functions by introducing $v = \mathrm{div}\mathbf{u}$. We refer to Eller, Isakov, Nakamura, and Tataru [12] and Ikehata, Nakamura, and Yamamoto [15]. If $\mathbf{u}$ solves

$$Q_{\rho,\lambda,\mu}\mathbf{u} = \rho(x)\partial_s^2\mathbf{u} + \mathcal{L}_{\lambda,\mu}(x,\partial_x)\mathbf{u} = \mathbf{f},$$

then

$$\rho(x)\partial_s^2 v + (\lambda(x) + 2\mu(x))\Delta v + A_{1,1}(v,\mathbf{u}) = \mathrm{div}\mathbf{f}$$

and

$$\rho(x)\partial_s^2\mathbf{u} + \mu(x)\Delta\mathbf{u} + A_{1,2}(v,\mathbf{u}) = \mathbf{f},$$

where $A_{1,1}$, $A_{1,2}$ are (matrix) linear partial differential operators with measurable and bounded coefficients. By using the scalar Carleman estimate (4.10), we obtain

$$\frac{C}{\tau}\|e^{\tau\varphi}\mathbf{u}\|^2_{H^2_\tau(\Omega_{3r})} \le \|e^{\tau\varphi}\mathbf{f}\|^2_{L^2(\Omega_{3r})} + \|e^{\tau\varphi}\mathbf{u}\|^2_{H^2_\tau(\Sigma_{3r})} + \|e^{\tau\varphi}v\|^2_{H^1_\tau(\Omega_{3r})}$$

and

$$C\tau \left\|e^{\tau\varphi}v\right\|^2_{H^1_\tau(\Omega_{3r})} \le \left\|e^{\tau\varphi}\mathrm{div}\mathbf{f}\right\|^2_{L^2(\Omega_{3r})} + \tau \left\|e^{\tau\varphi}v\right\|^2_{H^1_\tau(\Sigma_{3r})} + \left\|e^{\tau\varphi}\mathbf{u}\right\|^2_{H^1_\tau(\Omega_{3r})}.$$

Therefore, by choosing τ large, we obtain

$$\begin{aligned}\frac{C}{\tau}\left\|e^{\tau\varphi}\mathbf{u}\right\|^2_{H^2_\tau(\Omega_{3r})} \le \left\|e^{\tau\varphi}\mathbf{f}\right\|^2_{L^2(\Omega_{3r})} + \frac{1}{\tau}\left\|e^{\tau\varphi}\nabla\mathbf{f}\right\|^2_{L^2(\Omega_{3r})} \\ + \left\|e^{\tau\varphi}\mathbf{u}\right\|^2_{H^2_\tau(\Sigma_{3r})} + \left\|e^{\tau\varphi}v\right\|^2_{H^1_\tau(\Sigma_{3r})}.\end{aligned}$$

This completes the proof of Lemma 4.1. □

Now we can argue similarly to [6], with suitable modifications. We introduce a cutoff function χ_4 satisfying $0 \le \chi_4 \le 1$, $\chi_4 \in C_0^\infty(\mathbb{R})$, and

$$\chi_4(\eta) = \begin{cases} 0, & \text{if } \eta \le \frac{1}{2},\ \eta \ge 8, \\ 1, & \text{if } \eta \in \left[\frac{3}{4}, 7\right]. \end{cases}$$

Now we proceed to the estimation near Γ_1.

4.2. Estimation near the boundary part Γ_1. We shall estimate $\mathbf{u}_{\gamma,t}$ in a ball $B_1 = B(x^{(1)}, r) = \left\{x \in \mathbb{R}^3;\ |x - x^{(0)}| \le r\right\}$ over a small interval $(-r, r)$ by the velocity trace (in the normal direction) in the given part $\Sigma_{1,3r} = \Gamma_1 \times (-3r, 3r) \subset \Sigma_{3r}$.

LEMMA 4.2. *Let $\mathbf{u}_{\gamma,t}$ be a solution to* (4.6). *Then there exist $\widetilde{B}_1 \equiv B_1 \times (-r, r) \subset \Omega_r$ and $\nu_0 \in]0, 1[$ such that*

$$\left\|\mathbf{u}_{\gamma,t}\right\|_{H^2(\widetilde{B}_1)} \le C\left(\left\|F_{\gamma,t}\right\|_{H^1(\Omega_{3r})} + \left\|\mathbf{u}_{\gamma,t}\right\|_{H^2(\Sigma_{1,3r})}\right)^{\nu_0}\left(\left\|\mathbf{u}_{\gamma,t}\right\|_{H^2(\Omega_{3r})}\right)^{1-\nu_0} \tag{4.11}$$

for some positive constant C.

Proof. Let us choose $\delta > 0$ and $x^{(0)} \in \mathbb{R}^3 \backslash \overline{\Omega}$ such that

$$\delta < \frac{\epsilon_0}{4}, \quad \overline{B(x^{(0)}, \delta)} \cap \overline{\Omega} = \emptyset, \quad B(x^{(0)}, 2\delta) \cap \Omega \ne \emptyset, \quad B(x^{(0)}, 4\delta) \cap \Gamma \subset \Gamma_1. \tag{4.12}$$

That is, $x^{(0)}$ is an outer point of $\overline{\Omega}$ and is near Γ_1. We define the functions $\psi_0(x, s)$ and $\varphi_0(x, s)$ by

$$\psi_0(x, s) = \left|x - x^{(0)}\right|^2 + s^2, \quad \varphi_0(x, s) = e^{-\frac{\widetilde{\beta}}{\delta^2}\psi_0(x,s)}.$$

Here we choose $\widetilde{\beta} > 0$ sufficiently large. Denote

$$\mathbf{w}_{\gamma,t}(x, s) = \chi_4\left(\frac{\psi_0}{\delta^2}\right)\mathbf{u}_{\gamma,t}(x, s).$$

By applying Lemma 4.1, we obtain

$$\frac{C}{\tau}\left\|e^{\tau\varphi_0}\mathbf{w}_{\gamma,t}\right\|^2_{H^2_\tau(\Omega_r)} \le \left\|e^{\tau\varphi_0}Q_{\rho,\lambda,\mu}\mathbf{w}_{\gamma,t}\right\|^2_{H^1(\Omega_{3r})} + \left\|e^{\tau\varphi_0}\chi_4\left(\frac{\psi_0}{\delta^2}\right)\mathbf{u}_{\gamma,t}\right\|^2_{H^2_\tau(\Sigma_{3r})}$$

for $\tau > \tau_0$. Therefore by (4.9), we have

$$\begin{aligned} Q_{\rho,\lambda,\mu}\mathbf{w}_{\gamma,t}(x,s) &= \chi_4\left(\frac{\psi_0}{\delta^2}\right) Q_{\rho,\lambda,\mu}\mathbf{u}_{\gamma,t}(x,s) + \left[Q_{\rho,\lambda,\mu}, \chi_4\left(\frac{\psi_0}{\delta^2}\right)\right]\mathbf{u}_{\gamma,t}(x,s) \\ &= \chi_4\left(\frac{\psi_0}{\delta^2}\right)(F_{\gamma,t}(x,s) + G_{\gamma,t}(x,s)) + \left[Q_{\rho,\lambda,\mu}, \chi_4\left(\frac{\psi_0}{\delta^2}\right)\right]\mathbf{u}_{\gamma,t}(x,s) \\ &= \chi_4\left(\frac{\psi_0}{\delta^2}\right) F_{\gamma,t}(x,s) + \left[Q_{\rho,\lambda,\mu}, \chi_4\left(\frac{\psi_0}{\delta^2}\right)\right]\mathbf{u}_{\gamma,t}(x,s). \end{aligned}$$

Since $[Q_{\rho,\lambda,\mu}, \chi_4(\frac{\psi_0}{\delta^2})]$ is supported in

$$\left|x - x^{(0)}\right|^2 + s^2 \le \frac{3}{4}\delta^2, \quad 7\delta^2 \le \left|x - x^{(0)}\right|^2 + s^2 \le 8\delta^2,$$

by taking (4.12) into account, we see that $\left|x - x^{(0)}\right| \ge \delta$ for all $x \in \overline{\Omega}$ and $\Omega \cap \left\{x;\, \left|x - x^{(0)}\right| \le \frac{3}{4}\delta^2\right\} \ne \emptyset$, so that we obtain

$$\begin{aligned} \frac{C}{\tau}e^{2\tau e^{-4\tilde{\beta}}}\left\|\mathbf{u}_{\tau,t}\right\|^2_{H^2_\tau((\delta^2\le\psi_0\le 4\delta^2)\cap\Omega)} \le{}& \tau^2 e^{2\tau e^{-7\tilde{\beta}}}\left\|\mathbf{u}_{\gamma,t}\right\|^2_{H^2(\psi_0\le 8\delta^2)} \\ &+\tau^2 e^{2\tau e^{-\tilde{\beta}/2}}\left\|F_{\gamma,t}\right\|^2_{H^1(\Omega_{3r})} \\ &+e^{2\tau e^{-\tilde{\beta}/2}}\|\mathbf{u}_{\gamma,t}\|^2_{H^2(\Sigma_{1,3r})}. \end{aligned}$$

We can select $r > 0$ and $x^{(1)} \in \Omega$ such that

$$\operatorname{dist}(x^{(1)}, \Gamma) \ge 4r, \quad \widetilde{B}_1 = B(x^{(1)}, r) \times [-r, r] \subset \left\{\delta^2 \le \psi_0(x,s) \le 4\delta^2\right\}.$$

Then for $\tau > \tau_0$ we have

$$\|\mathbf{u}_{\gamma,t}\|^2_{H^2(\widetilde{B}_1)} \le Ce^{C_1\tau}\left[\|F_{\gamma,t}\|^2_{H^1(\Omega_{3r})} + \|\mathbf{u}_{\gamma,t}\|^2_{H^2(\Sigma_{1,3r})}\right] + e^{-C_2\tau}\left\|\mathbf{u}_{\gamma,t}\right\|^2_{H^2(\Omega_{3r})}.$$

The inequality holds for any $\tau > 0$ by replacing $C > 0$ by $Ce^{C_1\tau_0}$. Now minimize the right-hand side with respect to τ, and with $\nu_0 = \frac{C_2}{C_1+C_2}$ we have

$$\|\mathbf{u}_{\gamma,t}\|^2_{H^2(\widetilde{B}_1)} \le C\left(\|F_{\gamma,t}\|^2_{H^1(\Omega_{3r})} + \|\mathbf{u}_{\gamma,t}\|^2_{H^2(\Sigma_{1,3r})}\right)^{\nu_0}\left(\left\|\mathbf{u}_{\gamma,t}\right\|^2_{H^2(\Omega_{3r})}\right)^{1-\nu_0}.$$

This completes the proof of the lemma. ∎

4.3. Estimation near the boundary. In this subsection we extend the estimation from $\widetilde{B}_1$ to $\omega_r(\epsilon_0, 4\epsilon_0)$. In order to accomplish this, we use the techniques developed in [44]. This will be done by continuing estimates (4.11). Let $B(x^{(j)}, r)$, $2 \le j \le N$, be a covering of $\omega(\epsilon_0, 4\epsilon_0)$. We can assume that $x^{(j)}$ satisfies $\operatorname{dist}(x^{(j)}, \Gamma) \ge 4r$. In what follows, we assume without any restriction in generality that

$$B(x^{(j+1)}, r) \subset B(x^{(j)}, 2r),$$

and we set

$$\widetilde{B}_j = B(x^{(j)}, r) \times (-r, r); \quad 2 \le j \le N.$$

LEMMA 4.3. *Let $\mathbf{u}_{\gamma,t}$ be a solution to* (4.6). *Then there exist a constant $\nu \in (0,1)$ and $C > 0$ such that*

$$\|\mathbf{u}_{\gamma,t}\|_{H^2(\widetilde{B}_{k+1})} \leq C \left(\|F_{\gamma,t}\|_{H^1(\Omega_{3r})} + \|\mathbf{u}_{\gamma,t}\|_{H^2(\widetilde{B}_k)} \right)^{\nu} \left(\|\mathbf{u}_{\gamma,t}\|_{H^2(\Omega_{3r})} \right)^{1-\nu}, \quad k \geq 1. \tag{4.13}$$

Proof. We define the functions $\psi_k(x,s)$ and $\varphi_k(x,s)$ by

$$\psi_k(x,s) = \left| x - x^{(k)} \right|^2 + s^2, \quad \varphi_k(x,s) = e^{-\frac{\widetilde{\beta}}{r^2}\psi_k(x,s)}.$$

Moreover we set

$$\mathbf{w}_{\gamma,t}(x,s) = \chi_4\left(\frac{\psi_k}{r^2}\right)\mathbf{u}_{\gamma,t}(x,s).$$

By applying Lemma 4.1 in an interior domain, we obtain

$$\frac{C}{\tau}\left\|e^{\tau\varphi_k}\mathbf{w}_{\gamma,t}\right\|^2_{H^2_\tau(\Omega_r)} \leq \left\|e^{\tau\varphi_k}Q_{\rho,\lambda,\mu}\mathbf{w}_{\gamma,t}\right\|^2_{H^1(\Omega_{3r})}. \tag{4.14}$$

In the same way as the proof of Lemma 4.2, we have

$$Q_{\rho,\lambda,\mu}\mathbf{w}_{\lambda,t}(x,s) = \chi_4\left(\frac{\psi_k}{r^2}\right)F_{\lambda,t}(x,s) + \left[Q_{\rho,\lambda,\mu}, \chi_4\left(\frac{\psi_k}{r^2}\right)\right]\mathbf{u}_{\gamma,t}(x,s).$$

Since $[Q_{\rho,\lambda,\mu}, \chi_4(\frac{\psi_k}{r^2})]$ is supported in

$$\frac{r^2}{2} \leq \left|x - x^{(0)}\right|^2 + s^2 \leq r^2, \quad 7r^2 \leq \left|x - x^{(0)}\right|^2 + s^2 \leq 8r^2,$$

it follows from (4.14) that

$$\begin{aligned}\frac{C}{\tau}e^{2\tau e^{-5\widetilde{\beta}}}\left\|\mathbf{u}_{\gamma,t}\right\|^2_{H^2_\tau(r^2\leq\psi_k\leq 5r^2)} &\leq \tau^2 e^{2\tau e^{-\widetilde{\beta}/2}}\left\|\mathbf{u}_{\gamma,t}\right\|^2_{H^2(\psi_k\leq r^2)} + \tau^2 e^{2\tau e^{-7\widetilde{\beta}}}\left\|\mathbf{u}_{\gamma,t}\right\|^2_{H^2(\Omega_{3r})} \\ &\quad + \tau^2 e^{2\tau e^{-\widetilde{\beta}/2}}\left\|F_{\gamma,t}\right\|^2_{H^1(\Omega_{3r})},\end{aligned}$$

and hence for τ large we obtain

$$\begin{aligned}Ce^{2\tau e^{-5\widetilde{\beta}}}\left\|\mathbf{u}_{\gamma,t}\right\|^2_{H^2_\tau(\psi_k\leq 5r^2)} &\leq e^{2\tau e^{-\widetilde{\beta}/3}}\left\|\mathbf{u}_{\gamma,t}\right\|^2_{H^2(\psi_k\leq r^2)} + e^{2\tau e^{-6\widetilde{\beta}}}\left\|\mathbf{u}_{\gamma,t}\right\|^2_{H^2(\Omega_{3r})} \\ &\quad + e^{2\tau e^{-\widetilde{\beta}/3}}\left\|F_{\gamma,t}\right\|^2_{H^1(\Omega_{3r})}.\end{aligned}$$

Thus we obtain

$$\left\|\mathbf{u}_{\gamma,t}\right\|^2_{H^2(\psi_k\leq 5r^2)} \leq e^{C_1\tau}\left[\left\|\mathbf{u}_{\gamma,t}\right\|^2_{H^2(\psi_k\leq r^2)} + \left\|F_{\gamma,t}\right\|^2_{H^1(\Omega_{3r})}\right] + e^{-C_2\tau}\left\|\mathbf{u}_{\gamma,t}\right\|^2_{H^2(\Omega_{3r})}.$$

Now minimize the right-hand side with respect to τ, and with $\nu = \frac{C_2}{C_1+C_2}$ we obtain

$$\left\|\mathbf{u}_{\gamma,t}\right\|^2_{H^2(\psi_k\leq 5r^2)} \leq C\left(\left\|F_{\gamma,t}\right\|^2_{H^1(\Omega_{3r})} + \left\|\mathbf{u}_{\gamma,t}\right\|^2_{H^2(\psi_k\leq r^2)}\right)^{\nu}\left(\left\|\mathbf{u}_{\gamma,t}\right\|^2_{H^2(\Omega_{3r})}\right)^{1-\nu}.$$

Since

$$\widetilde{B}_{k+1} \subset \left\{\psi_k(s,x) \leq 5r^2\right\}, \quad \left\{\psi_k(x,s) \leq r^2\right\} \subset \widetilde{B}_k,$$

we obtain (4.13). This completes the proof of the lemma. □

LEMMA 4.4. *Let $\mathbf{u}_{\gamma,t}$ be a solution to* (4.6). *Then there exist a constant $C>0$ and $\mu=\nu^n$ such that*

$$\|\mathbf{u}_{\gamma,t}\|_{H^2(\widetilde{B}_n)} \le C\left(\|F_{\gamma,t}\|_{H^1(\Omega_{3r})}+\|\mathbf{u}_{\gamma,t}\|_{H^2(\widetilde{B}_1)}\right)^{\mu}\left(\|\mathbf{u}_{\gamma,t}\|_{H^2(\Omega_{3r})}\right)^{1-\mu}, \quad n\ge 1.$$

Here $\nu\in(0,1)$ is the constant given in Lemma 4.3.

Proof. Put

$$a_k=\|\mathbf{u}_{\gamma,t}\|_{H^2(\widetilde{B}_k)}, \quad A=\|F_{\gamma,t}\|_{H^1(\Omega_{3r})}, \quad B=\|\mathbf{u}_{\gamma,t}\|_{H^2(\Omega_{3r})}.$$

By (4.13) we have

$$a_{k+1}\le (C^{\frac{1}{1-\nu}}B)^{1-\nu}(a_k+A)^{\nu}.$$

By applying Lemma 4 in [38] (see also [37]), we obtain for all $\mu\in\,]0,\nu^n]$

$$a_n\le 2^{\frac{1}{1-\nu}}CB^{1-\mu}(a_1+A)^{\mu}.$$

This completes the proof of the lemma. □

LEMMA 4.5. *Let $\mathbf{u}_{\lambda,t}$ be a solution to* (4.6). *Then there exist $C>0$ and $C_1>0$ such that for all n there exist $C_n>0$ and $T_n>0$ such that*

$$C\|\mathbf{u}_{\gamma,t}\|^2_{H^2(\widetilde{B}_n)} \le e^{-C_1\gamma}\|\mathbf{u}\|^2_{H^2(Q)}+e^{C_n\gamma}\|\mathbf{u}\|^2_{H^2(\Sigma_1)} \tag{4.15}$$

for all $t\in\left[-\frac{T}{2},\frac{T}{2}\right]$, where $T>T_n$.

Proof. By Lemma 4.4 and the Young inequality, we easily obtain

$$\|\mathbf{u}_{\gamma,t}\|_{H^2(\widetilde{B}_n)} \le \epsilon^p\|\mathbf{u}_{\gamma,t}\|_{H^2(\Omega_{3r})}+\epsilon^{-p'}\left[\|F_{\gamma,t}\|_{H^1(\Omega_{3r})}+\|\mathbf{u}_{\gamma,t}\|_{H^2(\widetilde{B}_1)}\right]$$

for all $\epsilon>0$. Here

$$p=\frac{1}{1-\mu}, \quad p'=\frac{1}{\mu}, \quad \text{and} \quad \mu=\nu^n.$$

By using estimates (4.7) and (4.8), we have for all $t\in\left[-\frac{T}{2},\frac{T}{2}\right]$

$$\|\mathbf{u}_{\gamma,t}\|_{H^2(\widetilde{B}_n)} \le \epsilon^p e^{C_1\gamma}\|\mathbf{u}\|_{H^2(Q)}+\epsilon^{-p'}\left[e^{-\eta T\gamma}\|\mathbf{u}\|_{H^2(Q)}+\|\mathbf{u}_{\gamma,t}\|_{H^2(\widetilde{B}_1)}\right].$$

By selecting

$$\epsilon=e^{-\frac{2C_1}{p}\gamma},$$

we obtain

$$\|\mathbf{u}_{\gamma,t}\|_{H^2(\widetilde{B}_n)} \le e^{-C_1\gamma}\|\mathbf{u}\|_{H^2(Q)}+e^{-(\eta T-\frac{2C_1p'}{p})\gamma}\|\mathbf{u}\|_{H^2(Q)}+e^{\frac{2C_1p'}{p}\gamma}\|\mathbf{u}_{\gamma,t}\|_{H^2(\widetilde{B}_1)}$$

for all $t\in\left[-\frac{T}{2},\frac{T}{2}\right]$ and $\gamma>0$. Take T sufficiently large such that

$$\eta T-\frac{2C_1p'}{p}>C_1,$$

and we obtain

$$\|\mathbf{u}_{\gamma,t}\|_{H^2(\widetilde{B}_n)} \leq e^{-C_1\gamma}\|\mathbf{u}\|_{H^2(Q)} + e^{\kappa_1\gamma}\|\mathbf{u}_{\gamma,t}\|_{H^2(\widetilde{B}_1)}, \tag{4.16}$$

where we set

$$\kappa_1 = \frac{2C_1p'}{p}.$$

Similarly we obtain from Lemma 4.2 and the Young inequality

$$\|\mathbf{u}_{\gamma,t}\|_{H^2(\widetilde{B}_1)} \leq \epsilon^{p_0}e^{C_1\gamma}\|\mathbf{u}\|_{H^2(Q)} + \epsilon^{-p_0'}\left[e^{-\eta T\gamma}\|\mathbf{u}\|_{H^2(Q)} + \|\mathbf{u}_{\gamma,t}\|_{H^2(\Sigma_{1,3r})}\right],$$

where

$$p_0 = \frac{1}{1-\nu_0}, \quad p_0' = \frac{1}{\nu_0}.$$

Selecting $\epsilon = e^{-(\frac{2C_1+\kappa_1}{p_0})\gamma}$, we obtain for some positive constant κ_2

$$\|\mathbf{u}_{\gamma,t}\|_{H^2(\widetilde{B}_1)} \leq e^{-(C_1+\kappa_1)\gamma}\|\mathbf{u}\|_{H^2(Q)} + e^{-(\eta T - \frac{(2C_1+\kappa_1)p_0'}{p_0})\gamma}\|\mathbf{u}\|_{H^2(Q)} + e^{\kappa_2\gamma}\|\mathbf{u}_{\gamma,t}\|_{H^2(\Sigma_{1,3r})}.$$

Take T large such that

$$\left(\eta T - \frac{(2C_1+\kappa_1)p_0'}{p_0}\right) > C_1 + \kappa_1.$$

Then we obtain

$$\|\mathbf{u}_{\gamma,t}\|_{H^2(\widetilde{B}_1)} \leq e^{-(C_1+\kappa_1)\gamma}\|\mathbf{u}\|_{H^2(Q)} + e^{\kappa_2\gamma}\|\mathbf{u}_{\gamma,t}\|_{H^2(\Sigma_{1,3r})}.$$

By inserting this into (4.16), we have

$$\|\mathbf{u}_{\gamma,t}\|_{H^2(\widetilde{B}_n)} \leq e^{-C_1\gamma}\|\mathbf{u}\|_{H^2(Q)} + e^{C_n\gamma}\|\mathbf{u}_{\gamma,t}\|_{H^2(\Sigma_{1,3r})}$$

for some positive constant C_n. This completes the proof of (4.15). □

4.4. End of the proof of Lemma 2.5. We shall complete the proof of Lemma 2.5 in this subsection.

We fix $T > \max_{1\leq n\leq N} T_n$. Addition of inequalities (4.15) for $n \in \{1,\ldots,N\}$ yields

$$\|\mathbf{u}_{\gamma,t}\|_{H^2(\omega_r(\epsilon_0,3\epsilon_0))} \leq e^{-C_3\gamma}\|\mathbf{u}\|_{H^2(Q)} + e^{C_4\gamma}\|\mathbf{u}\|_{H^2(\Sigma_1)}, \quad t \in \left[-\frac{T}{2},\frac{T}{2}\right], \tag{4.17}$$

for some positive constants C_3 and C_4. We set $\mathbf{u}_\gamma(x,t) = \mathbf{u}_{\gamma,t}(x,0)$. Then we have

$$\mathbf{u}_\gamma(x,t) = \sqrt{\frac{\gamma}{2\pi}}\int_{\mathbb{R}} e^{-\frac{\gamma}{2}(t-y)^2}\theta(y)\mathbf{u}(x,y)dy = (K_\gamma * \theta\mathbf{u})(x,t),$$

where

$$K_\gamma(t) = \sqrt{\frac{\gamma}{2\pi}}e^{-\frac{\gamma}{2}t^2}.$$

LEMMA 4.6. *Let $T^1 = \frac{T}{2} - r$. Then we have*

$$\|\mathbf{u}_\gamma\|^2_{H^2(\omega_{T^1}(\epsilon_0,3\epsilon_0))} \le e^{-\mu\gamma}\|\mathbf{u}\|^2_{H^1(Q)} + e^{\mu'\gamma}\|\mathbf{u}\|^2_{H^2(\Sigma_1)}$$

for some positive constants μ and μ'.

Proof. By the Cauchy formula, for ϱ such that $0 < \varrho < r$, we obtain

$$\mathbf{u}_\gamma(x,a) = \frac{1}{2i\pi}\int_{|w-a|=\varrho} \frac{\mathbf{u}_\gamma(x,w)}{w-a}dw.$$

Thus, by using the polar coordinate, we obtain

$$|\mathbf{u}_\gamma(x,a)|^2 \le C_5 \int_0^{2\pi} \left|\mathbf{u}_\gamma(x,a+\varrho e^{i\theta})\right|^2 d\theta.$$

We integrate $\varrho \in (0,r)$ and obtain

$$|\mathbf{u}_\gamma(x,a)|^2 \le \frac{C_5}{r}\int_0^r\int_0^{2\pi} \left|\mathbf{u}_\gamma(x,a+\varrho e^{i\theta})\right|^2 d\theta d\varrho.$$

Therefore, for $x \in \omega(\epsilon_0,3\epsilon_0)$ and $a \in [-\frac{T}{2}+r, \frac{T}{2}-r]$, we have

$$|\mathbf{u}_\gamma(x,a)|^2 \le C_6\int_{|s|\le r,|t-a|\le r} |\mathbf{u}_\gamma(x,t+is)|^2\, dsdt = C_6\int_{|s|\le r,|t-a|\le r} |\mathbf{u}_{\gamma,t}(x,s)|^2\, dsdt$$

and

$$\begin{aligned}\|\mathbf{u}_\gamma(\cdot,a)\|^2_{L^2(\omega(\epsilon_0,3\epsilon_0))} &\le C_6\int_{-r}^r\int_{|t-a|\le r} \|\mathbf{u}_{\gamma,t}\|^2_{L^2(\omega_r(\epsilon_0,3\epsilon_0))}\, dtds \\ &\le C_7\int_{-r}^r\int_{-T/2}^{T/2} \|\mathbf{u}_{\gamma,t}\|^2_{L^2(\omega_r(\epsilon_0,3\epsilon_0))}\, dtds. \end{aligned} \tag{4.18}$$

We integrate $a \in [-T^1, T^1]$, and by using (4.17) we obtain

$$\|\mathbf{u}_\gamma\|^2_{L^2(\omega_{T^1}(\epsilon_0,3\epsilon_0))} \le C_T(e^{-C_3\gamma}\|\mathbf{u}\|^2_{H^2(Q)} + e^{C_4\gamma}\|\mathbf{u}\|^2_{H^2(\Sigma_1)}).$$

By using the same argument to $\partial^\alpha \mathbf{u}_\gamma$, $|\alpha| \le 2$, we complete the proof of Lemma 4.6. ∎

LEMMA 4.7. *Let $\mathbf{u}$ be a solution of (4.4). Then there exist $C_7 > 0$ and $C_8 > 0$ such that*

$$\|\mathbf{u}\|_{H^2(\omega_{T^1}(\epsilon_0,3\epsilon_0))} \le \frac{C_7}{\sqrt{\gamma}}\|\mathbf{u}\|_{H^3(Q)} + e^{C_7\gamma}\|\mathbf{u}\|_{H^2(\Sigma_1)}.$$

Proof. By $\widehat{\mathbf{u}}(x,\tau)$ we denote the Fourier transform of $\mathbf{u}(x,t)$ in t. We have

$$\widehat{\theta\mathbf{u}}(x,\tau) - \widehat{\mathbf{u}_\gamma}(x,\tau) = (1-\widehat{K_\gamma})\widehat{\theta\mathbf{u}}(x,\tau).$$

Furthermore we can directly verify that

$$\left|(1-\widehat{K_\gamma})(\tau)\right| \le \frac{\tau^2}{\gamma},$$

so that we obtain for $T^1 = T/2 - r$

$$\|\mathbf{u} - \mathbf{u}_\gamma\|_{L^2(\omega_{T^1}(\epsilon_0,3\epsilon_0))} \leq \frac{C_8}{\sqrt{\gamma}} \|\mathbf{u}\|_{H^1(Q)}.$$

Similarly we have

$$\|\mathbf{u} - \mathbf{u}_\gamma\|_{H^2(\omega_{T^1}(\epsilon_0,3\epsilon_0))} \leq \frac{C_9}{\sqrt{\gamma}} \|\mathbf{u}\|_{H^3(Q)}.$$

Hence

$$\begin{aligned}\|\mathbf{u}\|_{H^2(\omega_{T^1}(\epsilon_0,3\epsilon_0))} &\leq C_{10}\left[\|\mathbf{u} - \mathbf{u}_\gamma\|_{H^2(\omega_{T^1}(\epsilon_0,3\epsilon_0))} + \|\mathbf{u}_\gamma\|_{H^2(\omega_{T^1}(\epsilon_0,3\epsilon_0))}\right] \\ &\leq C_{11}\left[\frac{1}{\sqrt{\gamma}}\|\mathbf{u}\|_{H^3(Q)} + \|\mathbf{u}_\gamma\|_{H^2(\omega_{T^1}(\epsilon_0,3\epsilon_0))}\right].\end{aligned}$$

On the other hand, by Lemma 4.6, we obtain

$$\|\mathbf{u}_\gamma\|^2_{H^2(\omega_{T^1}(\epsilon_0,3\epsilon_0))} \leq e^{-\mu\gamma}\|\mathbf{u}\|^2_{H^2(Q)} + e^{\mu'\gamma}\|\mathbf{u}\|^2_{H^2(\Sigma_1)}$$

for some positive constants μ and μ'. This complete the proof of the lemma. □

We now turn to the proof of Lemma 2.6. By Lemma 4.7 and $\mathbf{u} = \chi_3\mathbf{v}$, we obtain

$$\|\mathbf{v}\|_{H^2(\omega_{T^1}(\epsilon_0,3\epsilon_0))} \leq \frac{C_7}{\sqrt{\gamma}}\|\mathbf{v}\|_{H^3(Q)} + e^{C_7\gamma}\|\mathbf{v}\|_{H^2(\Sigma_1)}.$$

By (1.6), we obtain

$$\|\mathbf{u}\|_{H^2(\omega_{T^1}(\epsilon_0,3\epsilon_0))} \leq \frac{M}{\sqrt{\gamma}} + e^{C_7\gamma}\epsilon(\Sigma_1).$$

By selecting

$$\gamma = \frac{1}{2C_7}\log\left(2 + \frac{M}{\epsilon(\Sigma_1)}\right),$$

the proof of Lemma 2.5 is complete. □

Acknowledgment. The authors thank the anonymous referee for very useful comments.

REFERENCES

[1] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[2] L. Baudouin and J.-P. Puel, *Uniqueness and stability in an inverse problem for the Schrödinger equation*, Inverse Problems, 18 (2002), pp. 1537–1554.

[3] M. Bellassoued, *Unicité et contrôle pour le système de Lamé*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 561–592.

[4] M. Bellassoued, *Global logarithmic stability in inverse hyperbolic problem by arbitrary boundary observation*, Inverse Problems, 20 (2004), pp. 1033–1052.

[5] M. Bellassoued, *Uniqueness and stability in determining the speed of propagation of second-order hyperbolic equation with variable coefficients*, Appl. Anal., 83 (2004), pp. 983–1014.

[6] M. Bellassoued and M. Yamamoto, *Logarithmic stability in determination of a coefficient in an acoustic equation by arbitrary boundary observation*, J. Math. Pures Appl. 85 (2006), pp. 193–224.
[7] M. Bellassoued and M. Yamamoto, *Determination of a Coefficient in the Wave Equation with a Single Measurement*, preprint.
[8] A. L. Bukhgeim, *Introduction to the Theory of Inverse Problems*, VSP, Utrecht, 2000.
[9] A. L. Bukhgeim, J. Cheng, V. Isakov, and M. Yamamoto, *Uniqueness in determining damping coefficients in hyperbolic equations*, in Analytic Extension Formulas and Their Applications, S. Saitoh et al., eds., Springer, New York, 2001, pp. 27–46.
[10] A. L. Bukhgeim and M. V. Klibanov, *Global uniqueness of class of multidimensional inverse problems*, Soviet Math. Dokl., 24 (1981), pp. 244–247.
[11] T. Carleman, *Sur un problème d'unicité pour les systeème d'équations aux dérivées partielles à deux variables indépendents*, Ark. Mat. Astr. Fys. 2B (1939), pp. 1–9.
[12] M. Eller, V. Isakov, G. Nakamura, and D. Tataru, *Uniqueness and stability in the Cauchy problem for Maxwell and elasticity systems*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, Vol. 14, North-Holland, Amsterdam, 2002, pp. 329–349.
[13] A. V. Fursikov and O. Yu. Imanuvilov, *Controllability of Evolution Equations*, Seoul National University, Seoul, 1996.
[14] L. Hörmander, *Linear Partial Differential Operators*, Springer, Berlin, 1963.
[15] M. Ikehata, G. Nakamura, and M. Yamamoto, *Uniqueness in inverse problems for the isotropic Lamé system*, J. Math. Sci. Univ. Tokyo, 5 (1998), pp. 627–692.
[16] O. Yu. Imanuvilov, *Controllability of parabolic equations*, Sb. Math., 186 (1995), pp. 879–900.
[17] O. Yu. Imanuvilov, *On Carleman estimates for hyperbolic equations*, Asymptot. Anal., 32 (2002), pp. 185–220.
[18] O. Yu. Imanuvilov, V. Isakov, and M. Yamamoto, *An inverse problem for the dynamical Lamé system with two sets of boundary data*, Comm. Pure Appl. Math., 56 (2003), pp. 1366–1382.
[19] O. Yu. Imanuvilov and J.-P. Puel, *Global Carleman estimates for weak solutions of elliptic nonhomogeneous Dirichlet problems*, Int. Math. Res. Not., 16 (2003), pp. 883–913.
[20] O. Yu. Imanuvilov and M. Yamamoto, *Lipshitz stability in inverse parabolic problems by Carleman estimate*, Inverse Problems, 14 (1998), pp. 1229–1249.
[21] O. Yu. Imanuvilov and M. Yamamoto, *Global Lipschitz stability in an inverse hyperbolic problem by interior observations*, Inverse Problems, 17 (2001), pp. 717–728.
[22] O. Yu. Imanuvilov and M. Yamamoto, *Determination of a coefficient in an acoustic equation with single measurement*, Inverse Problems, 19 (2003), pp. 157–171.
[23] O. Yu. Imanuvilov and M. Yamamoto, *Carleman estimates for the non-stationary Lamé system and the application to an inverse problem*, ESAIM Control Optim. Calc. Var., 11 (2005), pp. 1–56.
[24] O. Yu. Imanuvilov and M. Yamamoto, *Carleman estimates for the three-dimensional nonstationary Lamé system and applications to an inverse problem*, in Control Theory of Partial Differential Equations, Lect. Notes Pure Appl. Math. 242, O. Imanuvilov, G. Leugering, R. Triggiani, and B.-Y. Zhang, eds., Chapman & Hall/CRC, Boca Raton, FL, 2005, pp. 337–374.
[25] V. Isakov, *A nonhyperbolic Cauchy problem for $\Box_b\Box_c$ and its applications to elasticity theory*, Comm. Pure Appl. Math., 39 (1986), pp. 747–767.
[26] V. Isakov, *Inverse Problems for Partial Differential Equations*, Springer, Berlin, 2005.
[27] V. Isakov and M. Yamamoto, *Carleman estimates with the Neumann boundary condition and its applications to the observability inequality and inverse problems*, Contemp. Math., 268 (2000), pp. 191–225.
[28] M. A. Kazemi and M. V. Klibanov, *Stability estimates for ill-posed Cauchy problems involving hyperbolic equations and inequality*, Appl. Anal., 50 (1993), pp. 93–102.
[29] A. Khaĭdarov, *On stability estimates in multidimensional inverse problems for differential equation*, Soviet Math. Dokl., 38 (1989), pp. 614–617.
[30] M. V. Klibanov, *Inverse problems in the "large" and Carleman bounds*, Differ. Equ., 20 (1984), pp. 755–760.
[31] M. V. Klibanov, *Inverse problems and Carleman estimates*, Inverse Problems, 8 (1992), pp. 575–596.
[32] M. V. Klibanov and J. Malinsky, *Newton-Kantorovich method for 3-dimensional potential inverse scattering problem and stability of the hyperbolic Cauchy problem with time dependent data*, Inverse Problems, 7 (1991), pp. 577–595.
[33] M. V. Klibanov and A. Timonov, *Carleman Estimates for Coefficient Inverse Problems and*

Numerical Applications, VSP, Utrecht, 2004.
[34] M. V. Klibanov and M. Yamamoto, *Lipschitz stability of an inverse problem for an acoustic equation*, Appl. Anal., 85 (2006), pp. 515–538.
[35] M. Kubo, *Uniqueness in inverse hyperbolic problems: Carleman estimate for boundary value problems*, J. Math. Kyoto Univ., 40 (2000), pp. 451–473.
[36] M. M. Lavrent'ev, V. G. Romanov, and S. P. Shishat·skiĭ, *Ill-posed Problems of Mathematics Physics and Analysis*, AMS, Providence, RI, 1986.
[37] G. Lebeau and L. Robbiano, *Contrôle exact de l'équation de la chaleur*, Comm. Partial Differential Equations, 20 (1995), pp. 335–356.
[38] G. Lebeau and L. Robbiano, *Stabilisation de l'équation des ondes par le bord*, Duke Math. J., 86 (1997), pp. 465–491.
[39] S. Li, *An inverse problem for Maxwell's equations in bi-isotropic media*, SIAM J. Math. Anal., 37 (2005), pp. 1027–1043.
[40] J. P. Puel and M. Yamamoto, *On a global estimate in a linear inverse hyperbolic problem*, Inverse Problems, 12 (1996), pp. 995–1002.
[41] J. P. Puel and M. Yamamoto, *Generic well ill-posedness in a multidimensional hyperbolic inverse problem*, J. Inverse Ill-Posed Probl., 5 (1997), pp. 55–83.
[42] L. Rachele, *An inverse problem in elastodynamics: Uniqueness of the wave speeds in the interior*, J. Differential Equations, 162 (2000), pp. 300–325.
[43] L. Robbiano, *Théorème d'unicité adapté au contrôle des solutions des problèmes hyperboliques*, Comm. Partial Differential Equations, 16 (1991), pp. 789–800.
[44] L. Robbiano, *Fonction de coût et contrôle des solutions des équations hyperboliques*, Asymptot. Anal., 10 (1995), pp. 95–115.
[45] L. Robbiano and C. Zuily, *Uniqueness in the Cauchy problem for operators with partially holomorphic coefficients*, Invent. Math., 131 (1998), pp. 493–539.
[46] D. Tataru, *Carleman estimates and unique continuation for solutions to boundary value problems*, J. Math. Pures Appl., 75 (1996), pp. 367–408.
[47] R. Triggiani and P. F. Yao, *Carleman estimates with no lower-order terms for general Riemann wave equations. Global uniqueness and observability in one shot*, Appl. Math. Optim., 46 (2002), pp. 331–375.
[48] V. G. Yakhno, *Inverse Problems for Differential Equations of Elasticity*, Nauka, Novosibirsk, 1990.
[49] M. Yamamoto, *Uniqueness and stability in multidimensional hyperbolic inverse problems*, J. Math. Pures Appl., 78 (1999), pp. 65–98.

SIAM J. MATH. ANAL.
Vol. 40, No. 1, pp. 266–277

THE HUNTER–SAXTON EQUATION: A GEOMETRIC APPROACH*

JONATAN LENELLS†

Abstract. We provide a rigorous foundation for the geometric interpretation of the Hunter–Saxton equation as the equation describing the geodesic flow of the $\dot{H}^1$ right-invariant metric on the quotient space $Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$ of the infinite-dimensional Banach manifold $\mathcal{D}^k(\mathbb{S})$ of orientation-preserving H^k-diffeomorphisms of the unit circle $\mathbb{S}$ modulo the subgroup of rotations $Rot(\mathbb{S})$. Once the underlying Riemannian structure has been established, the method of characteristics is used to derive explicit formulas for the geodesics corresponding to the $\dot{H}^1$ right-invariant metric, yielding, in particular, new explicit expressions for the spatially periodic solutions of the initial-value problem for the Hunter–Saxton equation.

Key words. diffeomorphism group, Hunter–Saxton equation, geodesic flow

AMS subject classifications. 35Q53, 53C21, 58D30

DOI. 10.1137/050647451

1. Introduction.

The Hunter–Saxton equation

$$u_{txx} = -2u_x u_{xx} - u u_{xxx}, \quad t > 0,\ x \in \mathbb{R}, \tag{1.1}$$

models the propagation of weakly nonlinear orientation waves in a massive nematic liquid crystal director field, x being the space variable in a reference frame moving with the unperturbed wave speed and t being a slow time variable [6]. Equation (1.1) is a bivariational, completely integrable system with a bi-Hamiltonian structure [7], leading to the existence of an infinite family of commuting Hamiltonian flows together with associated conservation laws. Smooth solutions of (1.1) break down in finite time as the slope $u_x \to -\infty$. The existence of global weak solutions was considered in [8]. Weak dissipative solutions of (1.1) have been studied in [1], where also a discussion of different notions of weak solutions can be found.

It was first noticed in [9] that, for spatially periodic functions, (1.1) describes the geodesic flow on the homogeneous space $Rot(\mathbb{S})\backslash\mathcal{D}(\mathbb{S})$ of the infinite-dimensional Lie group $\mathcal{D}(\mathbb{S})$ of orientation-preserving diffeomorphisms of the unit circle $\mathbb{S}$ modulo the subgroup of rotations $Rot(\mathbb{S})$, endowed with the $\dot{H}^1$ right-invariant metric given at the identity by

$$\langle [u], [v] \rangle = \int_{\mathbb{S}} u_x v_x dx, \qquad [u], [v] \in T_{[id]}\big(Rot(\mathbb{S})\backslash\mathcal{D}(\mathbb{S})\big),$$

where $[id]$ denotes the equivalence class in $Rot(\mathbb{S})\backslash\mathcal{D}(\mathbb{S})$ of the identity map $id \in \mathcal{D}(\mathbb{S})$. This makes (1.1) one of the equations (others include the well-known Euler, Burgers, Korteweg–de Vries, and Camassa–Holm equations [4, 3, 15, 13]) that arises as the Euler equation for the geodesic flow corresponding to a right-invariant metric.

In this paper we provide a rigorous foundation for this geometric interpretation of (1.1) within the periodic setting (for the case on the line further technical complications arise due to the need of considering weighted Sobolev spaces—see the discussion

*Received by the editors December 13, 2005; accepted for publication (in revised form) September 14, 2007; published electronically April 16, 2008.
http://www.siam.org/journals/sima/40-1/64745.html
†Department of Mathematics, University of California, Santa Barbara, CA 93106 (jonatan@math.ucsb.edu).

in [2]). Two similar but distinct approaches are possible. We may choose to consider the geodesic flow of the $\dot{H}^1$ right-invariant metric either (1) on the homogeneous space $Rot(\mathbb{S})\backslash\mathcal{D}(\mathbb{S})$, where $\mathcal{D}(\mathbb{S})$ denotes the Fréchet Lie group of smooth diffeomorphisms of $\mathbb{S}$, or (2) on the Banach manifold $Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$ for $k > 3/2$, where $\mathcal{D}^k(\mathbb{S})$ incorporates all diffeomorphisms of $\mathbb{S}$ of Sobolev class H^k. We will pursue the latter approach. Note that $\mathcal{D}^k(\mathbb{S})$ is a topological group but not a Lie group as the group operation $(\psi, \varphi) \mapsto \psi \circ \varphi$ for $\psi, \varphi \in \mathcal{D}^k(\mathbb{S})$ is continuous but not smooth due to derivative loss (cf. [5]). The advantage of working on $Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$ is that the theory of Riemannian geometry on Banach manifolds is available—whereas nearly all results familiar from finite-dimensional Riemannian geometry immediately generalize to Banach manifolds (see [10]); a transition to Fréchet manifolds introduces several technical complications (see [5]). In particular, there are no general existence and uniqueness results for differential equations in Fréchet spaces, making it difficult to study geodesic flow and parallel translation. Also, on a Fréchet manifold, since the inverse mapping theorem does not hold, neither the Lie group exponential map nor the Riemannian exponential map is necessarily a local diffeomorphism at the identity. Another advantage of working with the wider class $Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$ is that when studying partial differential equations it is often preferable to work in Sobolev spaces rather than in the category of C^∞-maps.

We will construct a smooth affine connection on $Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$ compatible with the $\dot{H}^1$ right-invariant metric. Once this has been established, the general theory of Riemannian geometry on Banach manifolds immediately yields existence of a smooth curvature tensor, existence of normal neighborhoods, existence and uniqueness results for the geodesic flow and parallel translation, locally length-minimizing properties of the geodesics, etc. In this paper we focus on the geodesic flow and its relation to the Hunter–Saxton equation. In particular, we find that if $\varphi \in C^2([0,T);\mathcal{D}^k(\mathbb{S}))$ is a C^2-curve in $\mathcal{D}^k(\mathbb{S})$, then the quotient curve $t \mapsto [\varphi(t)]$ is a geodesic in $Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$ with respect to the $\dot{H}^1$ right-invariant metric if and only if $u = \varphi_t \circ \varphi^{-1} \in C([0,T);H^k(\mathbb{S})) \cap C^1([0,T);H^{k-1}(\mathbb{S}))$ satisfies (1.1). In a subsequent section the method of characteristics is used to derive explicit formulas for the geodesics corresponding to the $\dot{H}^1$ right-invariant metric. As a byproduct we also obtain explicit formulas for the spatially periodic solutions of the initial-value problem for (1.1). Note that this presents new solutions to (1.1) with respect to earlier presentations (cf. [6, 1])—the requirement that the solutions be spatially periodic introduces nontrivial constants of integration which alter the derivation and its outcome.

The quotient space $Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$, the $\dot{H}^1$ right-invariant metric $\langle\cdot,\cdot\rangle$, and a Christoffel map Γ are introduced in section 2. In section 3 it is proved that the covariant derivative induced by Γ is the unique covariant derivative compatible with the $\dot{H}^1$ right-invariant metric. In section 4 the connection between the geodesic flow and the Hunter–Saxton equation is explained, while in section 5 explicit formulas for the geodesics are obtained by means of the method of characteristics.

2. The quotient space $Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$. Let $\mathbb{S}$ be the circle of length one and let D_x denote differentiation with respect to x. For $X = [0,1]$ or $X = \mathbb{S}$ we let, for $n \geq 0$, $H^n(X)$ be the space of all functions on X of Sobolev class H^n. By restriction of a periodic function to the unit interval, $H^n(\mathbb{S})$ may be viewed as a closed linear subspace of $H^n[0,1]$.

For an integer $k \geq 3$,[1] let $\mathcal{D}^k(\mathbb{S})$ denote the Banach manifold of orientation-

[1]Even though $k > 3/2$ is sufficient for $\mathcal{D}^k(\mathbb{S})$ to be a topological group [4], we will assume $k \geq 3$ for simplicity; see [11] for an extension of the geometric approach which incorporates weak solutions.

preserving diffeomorphisms of $\mathbb{S}$ of class H^k (cf. [14]). By $Rot(\mathbb{S}) \subset \mathcal{D}^k(\mathbb{S})$ we denote the subgroup of rotations $x \mapsto x + d$, $d \in \mathbb{R}$. Let $Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$ be the space of right cosets $Rot(\mathbb{S}) \circ \varphi = \{\varphi(\cdot) + d \big| \, d \in \mathbb{R}\}$ for $\varphi \in \mathcal{D}^k(\mathbb{S})$. We will construct a global canonical chart on $Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$.

Put $M^k = \{\varphi \in \mathcal{D}^k(\mathbb{S}) \big| \, \varphi(0) = 0\}$. Then the map

$$\varphi \mapsto \big(\varphi(0), \varphi(\cdot) - \varphi(0)\big) : \mathcal{D}^k(\mathbb{S}) \to \mathbb{S} \times M^k \tag{2.1}$$

is a diffeomorphism. Note that M^k can be characterized as

$$M^k = \{\varphi \in H^k[0,1] \big| \, \varphi_x \in H^{k-1}(\mathbb{S}), \varphi_x > 0, \varphi(0) = 0, \varphi(1) = 1\},$$

or, equivalently,

$$M^k = \{u + id \big| \, u \in H^k(\mathbb{S}), u_x > -1, u(0) = 0\}, \tag{2.2}$$

where $id \in \mathcal{D}^k(\mathbb{S})$ is the identity map $id(x) = x$ for $x \in \mathbb{S}$.

From (2.1) we obtain a natural identification $Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S}) \simeq M^k$ given by

$$[\varphi] \mapsto \varphi - \varphi(0), \tag{2.3}$$

where $[\varphi]$ denotes the equivalence class of $\varphi \in \mathcal{D}^k(\mathbb{S})$. Let $\mathbf{E}^k \subset H^k(\mathbb{S})$ be the closed linear subspace

$$\mathbf{E}^k = \{u \in H^k(\mathbb{S}) \big| \, u(0) = 0\}$$

with topology induced from $H^k(\mathbb{S})$. The representation (2.2) shows that M^k is an open subset of the closed hyperplane $id + \mathbf{E}^k \subset H^k[0,1]$. Hence M^k provides a global chart for $Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$. Moreover, $T(Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})) \simeq TM^k \simeq M^k \times \mathbf{E}^k$, so that a vector field X on $Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$ can be viewed as a map $M^k \to \mathbf{E}^k$.

The $\dot{H}^1$ right-invariant metric $\langle \cdot, \cdot \rangle$ on $Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$ is most easily defined in the global chart M^k as follows (see also [9]). Let $A = -D_x^2$ and introduce a positive definite symmetric bilinear form $\langle \cdot, \cdot \rangle_{id}$ on $T_{id}M^k \simeq \mathbf{E}^k$ by

$$\langle u, v \rangle_{id} = \int_{\mathbb{S}} uAv dx = \int_{\mathbb{S}} u_x v_x dx, \qquad u, v \in \mathbf{E}^k.$$

and extend it to $T_\varphi M^k$ for any $\varphi \in M^k$ by right-invariance, so that, for $U, V \in T_\varphi M^k \simeq \mathbf{E}^k$,

$$\langle U, V \rangle_\varphi = \int_{\mathbb{S}} U \circ \varphi^{-1} A (V \circ \varphi^{-1}) dx.$$

We also introduce the closed linear subspace

$$\mathbf{F}^k = \left\{ f \in H^k(\mathbb{S}) \, \middle| \, \int_{\mathbb{S}} f dx = 0 \right\}.$$

It is straightforward to check that $A = -D_x^2$ is an isomorphism $\mathbf{E}^k \to \mathbf{F}^{k-2}$. Denoting its inverse by $A^{-1} : \mathbf{F}^{k-2} \to \mathbf{E}^k$, we infer that, for $u \in H^{k-1}(\mathbb{S})$,

$$-\big(A^{-1} D_x(u)\big)(x) = \int_0^x u(y) dy - x \int_{\mathbb{S}} u dx. \tag{2.4}$$

Now define $\Gamma : M^k \times \mathbf{E}^k \times \mathbf{E}^k \to \mathbf{E}^k$ by

$$\Gamma(\varphi, U, V) = -\frac{1}{2}\Big(A^{-1}D_x\big((U\circ\varphi^{-1})_x(V\circ\varphi^{-1})_x\big)\Big)\circ\varphi, \tag{2.5}$$

and notice that Γ is right-invariant in the sense that

$$\Gamma(\psi, U, V)\circ\varphi = \Gamma(\psi\circ\varphi, U\circ\varphi, V\circ\varphi), \qquad \varphi,\psi\in M^k, \quad U,V\in\mathbf{E}^k. \tag{2.6}$$

The motivation for the definition of Γ comes in section 3, where we will see that Γ is the Christoffel map (the infinite-dimensional analogue of the Christoffel symbols Γ^i_{jk} well known from finite-dimensional Riemannian geometry) for the affine connection on $Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$ compatible with the $\dot{H}^1$ right-invariant metric.

Although $\mathcal{D}^k(\mathbb{S})$ is a smooth Banach manifold it is not a Lie group. Indeed, the group operation $(\psi,\varphi)\mapsto\psi\circ\varphi : \mathcal{D}^k(\mathbb{S})\times\mathcal{D}^k(\mathbb{S})\to\mathcal{D}^k(\mathbb{S})$ is continuous but not C^1; right multiplication $R_\varphi : \psi\mapsto\psi\circ\varphi$ is smooth, whereas left multiplication $L_\psi : \varphi\mapsto\psi\circ\varphi$ is continuous but not C^1 due to derivative loss (see [5]). Similarly, M^k is a Banach manifold and a topological group but not a Lie group. Therefore, it is not a priori clear that the $\dot{H}^1$ metric and the Christoffel map Γ are smooth objects (that right multiplication is smooth is not enough). The following two propositions deal with this technicality (see [12] for detailed proofs of similar results in the case of the Camassa–Holm equation).

Proposition 2.1. *The map*

$$[\varphi]\mapsto\langle\cdot,\cdot\rangle_{[\varphi]} : Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})\to L^2_{sym}(T_{[\varphi]}\big(Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})\big);\mathbb{R})$$

is a smooth section of the bundle $L^2_{sym}(T\big(Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})\big);\mathbb{R})$.

Proposition 2.2. *Let* Γ *be defined by* (2.5). *The map*

$$\varphi\mapsto\Gamma(\varphi,\cdot,\cdot) : M^k\to L^2_{sym}(\mathbf{E}^k;\mathbf{E}^k)$$

is smooth.

3. Covariant derivative. In this section the covariant derivative induced by the Christoffel map Γ is shown to be compatible with the $\dot{H}^1$ right-invariant metric.

We first recall the general definition of a covariant derivative. Let $\mathcal{M}$ be a Banach manifold endowed with a Riemannian metric $\langle\cdot,\cdot\rangle$ and let $\mathfrak{X}(\mathcal{M})$ denote the space of smooth vector fields on $\mathcal{M}$. For $X,Y\in\mathfrak{X}(\mathcal{M})$ the Lie bracket $[X,Y]$ is defined locally by

$$[X,Y](m) = DY(m)\cdot X(m) - DX(m)\cdot Y(m).$$

Definition 3.1. *An* $\mathbb{R}$*-bilinear operator* $(X,Y)\mapsto\nabla_XY : \mathfrak{X}(\mathcal{M})\times\mathfrak{X}(\mathcal{M})\to\mathfrak{X}(\mathcal{M})$ *is a* Riemannian covariant derivative *if it satisfies*

(a) $X(m)=0$ *implies* $(\nabla_XY)(m)=0$ *for* $m\in\mathcal{M}$ *and* $X,Y\in\mathfrak{X}(\mathcal{M})$ *(punctual dependence on* X*),*
(b) $\nabla_XY-\nabla_YX=[X,Y]$ *for* $X,Y\in\mathfrak{X}(\mathcal{M})$ *(torsion-free),*
(c) $\nabla_X(fY)=(\mathcal{L}_Xf)Y+f\nabla_XY$ *for* $f\in C^\infty(\mathcal{M})$, $X,Y\in\mathfrak{X}(\mathcal{M})$ *(derivation in* Y*),*
(d) $\mathcal{L}_X\langle Y,Z\rangle=\langle\nabla_XY,Z\rangle+\langle Y,\nabla_XZ\rangle$ *for* $X,Y,Z\in\mathfrak{X}(\mathcal{M})$ *(compatible with the metric).*

Define the operator $\nabla : \mathfrak{X}(Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S}))\times\mathfrak{X}(Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})) \to \mathfrak{X}(Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S}))$ in the global chart M^k by

$$(\nabla_X Y)(\varphi) = DY(\varphi)\cdot X(\varphi) - \Gamma(\varphi, Y(\varphi), X(\varphi)), \tag{3.1}$$

where $X, Y : M^k \to \mathbf{E}^k$ are representatives in the chart M^k of two vector fields on $Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$.

In the finite-dimensional case, given a Riemannian metric $\langle\cdot,\cdot\rangle$ on a manifold $\mathcal{M}$ there automatically exists a Riemannian covariant derivative ∇ compatible with $\langle\cdot,\cdot\rangle$. For vector fields X, Y, Z on $\mathcal{M}$, $\nabla_X Y$ is defined as the unique vector field such that

$$2\langle\nabla_X Y, Z\rangle = -\langle[Y,X],Z\rangle - \langle X,[Y,Z]\rangle - \langle Y,[X,Z]\rangle + \mathcal{L}_X\langle Y,Z\rangle + \mathcal{L}_Y\langle Z,X\rangle - \mathcal{L}_Z\langle X,Y\rangle. \tag{3.2}$$

Indeed, the bracket $\langle\cdot,\cdot\rangle$ establishes an isomorphism $T_m\mathcal{M} \to T_m^*\mathcal{M}$ for each $m \in \mathcal{M}$, so since the right-hand side is a continuous linear functional of $Z(m)$, existence of $(\nabla_X Y)(m)$ follows immediately.

This approach does not apply to $Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$ endowed with the $\dot{H}^1$ right-invariant metric. The right-hand side of (3.2) is a continuous linear functional of $Z(\varphi)$ for each $\varphi \in M^k$. But the topology of $T_\varphi M^k \simeq \mathbf{E}^k$ induced by the H^k inner product is much stronger than the topology defined by the $\dot{H}^1$ right-invariant metric $\langle\cdot,\cdot\rangle_\varphi$—the $\dot{H}^1$ right-invariant metric is a *weak Riemannian metric* on $Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$. Therefore there are elements in $T_\varphi^* M^k$ that cannot be expressed as $\langle V,\cdot\rangle_\varphi$ for some $V \in T_\varphi M^k$; the spaces $T_\varphi M^k \simeq \mathbf{E}^k$ and $T_\varphi^* M^k \simeq \mathbf{E}^{k*}$ are in duality with respect to the H^k inner product, not with respect to $\langle\cdot,\cdot\rangle_\varphi$. The explicit formula for Γ will help us circumvent this difficulty.

However, even for weak Riemannian metrics *uniqueness* of the Riemannian covariant derivative can be deduced from (3.2). For if ∇ satisfies (a)–(d) of Definition 3.1, then, writing down property (d) for the cyclic permutations of $X, Y, Z \in \mathfrak{X}(M)$, we get

$$\mathcal{L}_X\langle Y,Z\rangle = \langle\nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle,$$

$$\mathcal{L}_Y\langle Z,X\rangle = \langle\nabla_Y Z, X\rangle + \langle Z, \nabla_Y X\rangle,$$

$$\mathcal{L}_Z\langle X,Y\rangle = \langle\nabla_Z X, Y\rangle + \langle X, \nabla_Z Y\rangle.$$

Adding the first two and subtracting the third of these relations, (3.2) drops out. Since $\langle\cdot,\cdot\rangle$ is nondegenerate, (3.2) shows the uniqueness of ∇.

In the proof of Theorem 3.2 the identity

$$\frac{d}{d\epsilon}\bigg|_{\epsilon=0} U\circ(\varphi+\epsilon V)^{-1} = -(U\circ\varphi^{-1})_x V\circ\varphi^{-1}, \qquad \varphi\in\mathcal{D}^k(\mathbb{S}), \quad U,V\in H^k(\mathbb{S}), \tag{3.3}$$

will be used. It is a consequence of

$$\frac{d}{d\epsilon}\bigg|_{\epsilon=0}(\varphi+\epsilon V)^{-1} = -\frac{V\circ\varphi^{-1}}{\varphi_x\circ\varphi^{-1}} \qquad \text{and} \qquad \frac{U_x\circ\varphi^{-1}}{\varphi_x\circ\varphi^{-1}} = (U\circ\varphi^{-1})_x.$$

THEOREM 3.2. *The bilinear map ∇ given by* (3.1) *defines a unique Riemannian covariant derivative on $Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$ compatible with the $\dot{H}^1$ right-invariant metric.*

Proof. Properties (a)–(c) are immediate from the local formula defining ∇. To establish (d) we compute, for vector fields $X, Y, Z : M^k \to \mathbf{E}^k$,

$$
\begin{aligned}
&(\mathcal{L}_X\langle Y,Z\rangle)(\varphi)\\
&\quad= \frac{d}{d\epsilon}\Big|_{\epsilon=0}\int_{\mathbb{S}} A\big(Y(\varphi+\epsilon X(\varphi))\circ(\varphi+\epsilon X(\varphi))^{-1}\big)\\
&\qquad\cdot Z(\varphi+\epsilon X(\varphi))\circ(\varphi+\epsilon X(\varphi))^{-1}dx\\
&\quad= \int_{\mathbb{S}} A\Big((DY(\varphi)\cdot X(\varphi))\circ\varphi^{-1} - (Y(\varphi)\circ\varphi^{-1})_x X(\varphi)\circ\varphi^{-1}\Big)Z(\varphi)\circ\varphi^{-1}dx\\
&\qquad+\int_{\mathbb{S}} A\Big((DZ(\varphi)\cdot X(\varphi))\circ\varphi^{-1} - (Z(\varphi)\circ\varphi^{-1})_x X(\varphi)\circ\varphi^{-1}\Big)Y(\varphi)\circ\varphi^{-1}dx,
\end{aligned}
$$

where formula (3.3) was used to carry out the differentiation. Define $u, v, w \in \mathbf{E}^k$ by $u = X(\varphi)\circ\varphi^{-1}$, $v = Y(\varphi)\circ\varphi^{-1}$, and $w = Z(\varphi)\circ\varphi^{-1}$. We get

$$
\begin{aligned}
(\mathcal{L}_X\langle Y,Z\rangle)(\varphi) &= \int_{\mathbb{S}} A\Big(\big(DY(\varphi)\cdot X(\varphi)\big)\circ\varphi^{-1}\Big)wdx\\
&+\int_{\mathbb{S}} A\Big(\big(DZ(\varphi)\cdot X(\varphi)\big)\circ\varphi^{-1}\Big)vdx - \int_{\mathbb{S}} A(v_x u)wdx - \int_{\mathbb{S}} A(w_x u)vdx.
\end{aligned}
\tag{3.4}
$$

On the other hand

$$
\langle\nabla_X Y, Z\rangle_\varphi = \int_{\mathbb{S}}\Big(DY(\varphi)\cdot X(\varphi) - \Gamma\big(\varphi, Y(\varphi), X(\varphi)\big)\Big)\circ\varphi^{-1}A(Z(\varphi)\circ\varphi^{-1})dx.
$$

Since $\Gamma(\varphi, Y(\varphi), X(\varphi)) = -\frac{1}{2}\big(A^{-1}D_x(v_x u_x)\big)\circ\varphi$, we get

$$
\langle\nabla_X Y, Z\rangle_\varphi = \int_{\mathbb{S}}\big(DY(\varphi)\cdot X(\varphi)\big)\circ\varphi^{-1}A(w)dx + \frac{1}{2}\int(v_x u_x)_x w dx.
\tag{3.5}
$$

Now, recalling that $A = -D_x^2$, it is easy to check that

$$
-\int_{\mathbb{S}} A(v_x u)wdx - \int_{\mathbb{S}} A(w_x u)vdx = \frac{1}{2}\int(v_x u_x)_x w dx + \frac{1}{2}\int(w_x u_x)_x v dx
$$

so by (3.4) and (3.5) we obtain

$$
(\mathcal{L}_X\langle Y,Z\rangle)(\varphi) = \langle\nabla_X Y, Z\rangle_\varphi + \langle Y, \nabla_X Z\rangle_\varphi.
$$

This proves that ∇ also satisfies (d). □

4. Geodesics and the Hunter–Saxton equation. Since the map Γ defined in (2.5) is a smooth Christoffel map (see Proposition 2.2), all the usual constructions for affine connections on Banach manifolds (cf. [10]) can easily be carried out on $Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$. Here we will be concerned with the geodesics on $Rot(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$ and their relation to the Hunter–Saxton equation.

4.1. Parallel translation. Let $J \subset \mathbb{R}$ be an open interval and let $\varphi : J \to M^k$ be a C^2-curve. A lift $V : J \to TM^k$ of φ is *φ-parallel* if

$$V_t = \Gamma(\varphi, V, \varphi_t), \qquad t \in J,$$

which is equivalent to $\nabla_{\varphi_t} V \equiv 0$. Applying the general theory for Banach manifolds, we get the following result.

PROPOSITION 4.1. *Let $t_0 \in J$. Given $V_0 \in T_{\varphi(0)}M^k$, there exists a unique φ-parallel lift $t \mapsto V(t; V_0) : J \to TM^k$ such that $V(t_0; V_0) = V_0$.*

Define two continuous maps $u, v : J \to T_{id}M^k \simeq \mathbf{E}^k$ by

$$u(t) = T_{\varphi(t)}R_{\varphi(t)^{-1}}(\varphi_t(t)) = \varphi_t(t) \circ \varphi(t)^{-1}$$

and

$$v(t) = T_{\varphi(t)}R_{\varphi(t)^{-1}}(V(t)) = V(t) \circ \varphi(t)^{-1}.$$

Note that u, v are not C^1-maps as

$$u_t = \varphi_{tt} \circ \varphi^{-1} + \frac{\varphi_{tx} \circ \varphi^{-1}}{\varphi_x \circ \varphi^{-1}}$$

is in general an element in $\mathbf{E}^{k-1}$ but not in $\mathbf{E}^k$. Nevertheless, it is clear that

$$u, v \in C(J; \mathbf{E}^k) \cap C^1(J; \mathbf{E}^{k-1}).$$

THEOREM 4.2. *Let $\varphi : J \to M^k$ be a C^2-curve and $V : J \to TM^k$ a lift of φ. Define $u, v : J \to \mathbf{E}^k$ by*

$$v(t) = V(t) \circ \varphi(t)^{-1}, \qquad u(t) = \varphi_t(t) \circ \varphi(t)^{-1},$$

so that $u, v \in C(J; \mathbf{E}^k) \cap C^1(J; \mathbf{E}^{k-1})$. The following statements are equivalent:

(a) *V is φ-parallel.*

(b) *u and v satisfy, for $t \in J$,*

$$v_t = \Gamma(id, v, u) - v_x u \quad \text{in} \quad \mathbf{E}^{k-1}. \tag{4.1}$$

(c) *u and v solve the equation*

$$v_{txx} = -\frac{3}{2}v_{xx}u_x - \frac{1}{2}v_x u_{xx} - v_{xxx}u, \qquad t \in J,\ x \in \mathbb{S}. \tag{4.2}$$

Proof. First note that

$$v_x u = (V \circ \varphi^{-1})_x \varphi_t \circ \varphi^{-1} = \frac{V_x \circ \varphi^{-1}}{\varphi_x \circ \varphi^{-1}} \varphi_t \circ \varphi^{-1} = -V_x \circ \varphi^{-1} \cdot (\varphi^{-1})_t. \tag{4.3}$$

Suppose V is φ-parallel. Using (4.3), we compute in $\mathbf{E}^{k-1}$

$$v_t = V_t \circ \varphi^{-1} + V_x \circ \varphi^{-1} \cdot (\varphi^{-1})_t = \Gamma(\varphi, V, \varphi_t) \circ \varphi^{-1} - v_x u = \Gamma(id, v, u) - v_x u,$$

where we used the right invariance (2.6) of Γ. Conversely, if (4.1) holds, then (4.3) yields

$$V_t \circ \varphi^{-1} = v_t - V_x \circ \varphi^{-1} \cdot (\varphi^{-1})_t = \Gamma(id, v, u) - v_x u - V_x \circ \varphi^{-1} \cdot (\varphi^{-1})_t = \Gamma(\varphi, V, \varphi_t) \circ \varphi^{-1},$$

showing that V is φ-parallel. This establishes the equivalence of (a) and (b).

Suppose (b) holds. By the definition (2.5) of Γ we rewrite (4.1) as

$$v_t = -\frac{1}{2}A^{-1}D_x(v_x u_x) - v_x u. \tag{4.4}$$

Applying $A = -D_x^2$ to both sides of (4.4) gives (4.2). Conversely, suppose (c) holds. Since at each fixed time both sides of (4.2) belong to $\mathbf{F}^{k-3}$, we may apply the isomorphism $A^{-1} : \mathbf{F}^{k-3} \to \mathbf{E}^{k-1}$ to obtain (4.4). Hence (b) and (c) are equivalent. □

4.2. Geodesics. A C^2-map $\varphi : J \to M^k$ is a geodesic if

$$\varphi_{tt} = \Gamma(\varphi, \varphi_t, \varphi_t).$$

Just like for parallel translation we can express the geodesic equation as an equation for $u = \varphi_t \circ \varphi^{-1}$.

THEOREM 4.3. *Let $\varphi : J \to M^k$ be a C^2-curve and define $u : J \to \mathbf{E}^k$ by $u(t) = \varphi_t(t) \circ \varphi(t)^{-1}$ so that $u \in C(J; \mathbf{E}^k) \cap C^1(J; \mathbf{E}^{k-1})$. Then φ is a geodesic if and only if u solves the Hunter–Saxton equation*

$$u_{txx} = -2u_x u_{xx} - u u_{xxx}, \qquad t \in J,\ x \in \mathbb{S}. \tag{4.5}$$

Proof. φ is a geodesic if and only if φ_t is φ-parallel. The equivalence of (a) and (c) of Theorem 4.2 shows that φ_t is φ-parallel if and only if u satisfies (4.5). □

The condition that $\varphi : J \to M^k \subset \mathcal{D}^k(\mathbb{S})$ imposes the condition $\varphi(t)\big|_{x=0} = 0$ for all t. We now remove this restriction.

Let $\rho : \mathcal{D}^k(\mathbb{S}) \to \mathit{Rot}(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$ be the quotient map. Assume $\varphi : J \to \mathcal{D}^k(\mathbb{S})$ is a C^1-curve and denote by $[\varphi] = \rho \circ \varphi : J \to \mathit{Rot}(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$ the induced curve in $\mathit{Rot}(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$. A lift $V : J \to T\mathcal{D}^k(\mathbb{S})$ of φ gives rise to a lift $[V] : J \to T\big(\mathit{Rot}(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})\big)$ of $[\varphi]$ given by $[V] = T\rho \circ V$. The next result, which is a fairly straightforward consequence of Theorem 4.2, characterizes the $[\varphi]$-parallel lifts $[V]$ of $[\varphi]$.

PROPOSITION 4.4. *Let $\varphi : J \to \mathcal{D}^k(\mathbb{S})$ be a C^2-curve and let V be a C^1-lift of φ. Define $u, v \in C(J; H^k(\mathbb{S})) \cap C^1(J; H^{k-1}(\mathbb{S}))$ by*

$$u(t) = \varphi_t(t) \circ \varphi(t)^{-1}, \quad v(t) = V(t) \circ \varphi(t)^{-1}.$$

Then $[V] = T\rho \circ V$ is $[\varphi]$-parallel if and only if u, v satisfy

$$v_{txx} = -\frac{3}{2}v_{xx}u_x - \frac{1}{2}v_x u_{xx} - v_{xxx}u, \qquad t \in J, \quad x \in \mathbb{S}. \tag{4.6}$$

The final theorem in this section says that $t \mapsto [\varphi(t)] : J \to \mathit{Rot}(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$ is a geodesic if and only if $u = \varphi_t \circ \varphi^{-1}$ solves the Hunter–Saxton equation.

THEOREM 4.5. *Let $\varphi : J \to \mathcal{D}^k(\mathbb{S})$ be a C^2-curve and define $u : J \to H^k(\mathbb{S})$ by $u(t) = \varphi_t(t) \circ \varphi(t)^{-1}$ so that $u \in C(J; H^k(\mathbb{S})) \cap C^1(J; H^{k-1}(\mathbb{S}))$. Then $[\varphi]$ is a geodesic in $\mathit{Rot}(\mathbb{S})\backslash\mathcal{D}^k(\mathbb{S})$ if and only if u is a solution of the Hunter–Saxton equation*

$$u_{txx} = -2u_x u_{xx} - u u_{xxx}, \qquad t \in J,\ x \in \mathbb{S}. \tag{4.7}$$

Proof. $[\varphi]$ is a geodesic if and only if $[\varphi]_t = [\varphi_t]$ is $[\varphi]$-parallel. By Proposition 4.4 this occurs if and only if (4.6) holds with $u = v = \varphi_t \circ \varphi^{-1}$. □

5. Explicit formulas. In this section the method of characteristics is adopted to obtain explicit formulas for the geodesics; as a byproduct new explicit solutions of the periodic Hunter–Saxton equation drop out.

DEFINITION 5.1. *By a solution of the Hunter–Saxton equation with initial data $u_0 \in H^k(\mathbb{S})$, $k \geq 3$, we mean a function $u(t,x)$ with $u \in C([0,T); H^3(\mathbb{S})) \cap C^1([0,T); H^2(\mathbb{S}))$ for some maximal time of existence $T > 0$ such that $u(0) = u_0$ and*

$$u_{txx} = -2u_x u_{xx} - uu_{xxx}, \qquad t \in [0,T),\ x \in \mathbb{S}. \tag{5.1}$$

The next lemma is straightforward to prove.

LEMMA 5.2. *Let $J \subset \mathbb{R}$ be an open interval and let $u \in C(J; H^3(\mathbb{S})) \cap C^1(J; H^2(\mathbb{S}))$. The map*

$$F : (t,\psi) \mapsto u(t) \circ \psi : J \times \mathcal{D}^2(\mathbb{S}) \to H^2(\mathbb{S})$$

is C^1.

Now let $u_0 \in H^k(\mathbb{S})$ and suppose $u(t,x)$ is a solution of the Hunter–Saxton equation with initial data u_0. Integration of (5.1) from 0 to x for each fixed time t gives

$$u_{tx} = -\frac{1}{2}u_x^2 - uu_{xx} + c(t), \qquad t \in [0,T),\ x \in \mathbb{S}, \tag{5.2}$$

for some function $c : [0,T) \to \mathbb{R}$. Since $\int_{\mathbb{S}} u_{tx}(t,x)dx = 0$ for each $t \in [0,T)$, we get

$$\int_{\mathbb{S}} \left(-\frac{1}{2}u_x^2 - uu_{xx} + c(t) \right) dx = 0.$$

An integration by parts yields

$$c(t) = -\frac{1}{2}\int_{\mathbb{S}} u_x^2(t,x)dx.$$

Moreover, by (5.2),

$$\frac{1}{2}\frac{d}{dt}\int_{\mathbb{S}} u_x^2 dx = \int_{\mathbb{S}} u_x u_{tx} dx = -\int_{\mathbb{S}} \left(\frac{1}{2}u_x^3 + uu_x u_{xx} \right) dx = 0,$$

showing that $c(t) = -\frac{1}{2}\int_{\mathbb{S}} u_x^2(t,x)dx$ is a constant function. Assuming that u_0 is nontrivial, we may rescale u_0 so that $\int_{\mathbb{S}} u_{0x}^2 dx = 4$. Then (5.2) becomes

$$u_{tx} = -\frac{1}{2}u_x^2 - uu_{xx} - 2, \qquad t \in [0,T),\ x \in \mathbb{S}.$$

By Lemma 5.2 and the local existence and uniqueness theorem for differential equations in Banach spaces, there exists a unique map $\varphi \in C^1([0,T_1); \mathcal{D}^2(\mathbb{S}))$ such that $\varphi(0) = id$ and

$$\varphi_t(t) = u(t) \circ \varphi(t), \qquad t \in [0,T_1), \tag{5.3}$$

for some maximal existence time $T_1 > 0$. Since

$$(u_x \circ \varphi)_t = u_{tx} \circ \varphi + u_{xx} \circ \varphi \cdot \varphi_t = \big(u_{tx} + uu_{xx}\big) \circ \varphi,$$

this gives

$$(u_x \circ \varphi)_t = \left(-\frac{1}{2}u_x^2 - 2\right) \circ \varphi = -\frac{1}{2}(u_x \circ \varphi)^2 - 2, \qquad t \in [0, T_1),\ x \in \mathbb{S}.$$

Therefore, for a fixed $x \in \mathbb{S}$, the function $t \mapsto (u_x \circ \varphi)(t, x)$ solves the differential equation

$$\dot{z}(t) = -\frac{1}{2}(z(t)^2 + 4). \tag{5.4}$$

The general solution of (5.4) is

$$z(t) = 2\tan\left(\arctan\left(\frac{z(0)}{2}\right) - t\right).$$

As $(u_x \circ \varphi)(0, x) = u_{0x}(x)$ for $x \in \mathbb{S}$, we infer that

$$(u_x \circ \varphi)(t, x) = 2\tan\left(\arctan\left(\frac{u_{0x}(x)}{2}\right) - t\right), \qquad t \in [0, T_1),\ x \in \mathbb{S}. \tag{5.5}$$

Differentiation of (5.3) yields $\varphi_{tx} = u_x \circ \varphi \cdot \varphi_x$, so that, by (5.5),

$$\varphi_{tx}(t, x) = 2\tan\left(\arctan\left(\frac{u_{0x}(x)}{2}\right) - t\right)\varphi_x(t, x).$$

Using the fact that $\varphi_x(0, x) = 1$ for $x \in \mathbb{S}$, we get

$$\varphi_x(t, x) = \exp\left(2\int_0^t \tan\left(\arctan\left(\frac{u_{0x}(x)}{2}\right) - s\right)ds\right), \qquad t \in [0, T_1),\ x \in \mathbb{S}. \tag{5.6}$$

Since

$$2\int_0^t \tan\left(\arctan\left(\frac{u_{0x}(x)}{2}\right) - s\right)ds = \ln\left(\cos^2\left(\arctan\left(\frac{u_{0x}(x)}{2}\right) - t\right)\right) - C(x),$$

where

$$C(x) = \ln\left(\cos^2\left(\arctan\left(\frac{u_{0x}(x)}{2}\right)\right)\right),$$

equation (5.6) yields

$$\varphi_x(t, x) = \frac{\cos^2\left(\arctan\left(\frac{u_{0x}(x)}{2}\right) - t\right)}{\cos^2\left(\arctan\left(\frac{u_{0x}(x)}{2}\right)\right)}. \tag{5.7}$$

We rewrite this as

$$\varphi_x(t, x) = \left(\cos t + \frac{u_{0x}(x)}{2}\sin t\right)^2, \quad t \in [0, T_1),\ x \in \mathbb{S}. \tag{5.8}$$

Hence

$$\varphi(t, x) - \varphi(t, 0) = x\frac{1 + \cos 2t}{2} + \frac{1}{4}\int_0^x u_{0x}^2(y)dy\frac{1 - \cos 2t}{2} + \frac{u_0(x) - u_0(0)}{2}\sin 2t.$$

Since, by (2.4),

$$-\big(A^{-1}D_x(u_{0x}^2)\big)(x) = \int_0^x u_{0x}^2(y)dy - x\int_{\mathbb{S}} u_{0x}^2 dx = \int_0^x u_{0x}^2(y)dy - 4x,$$

we infer that, for $t \in [0, T_1)$ and $x \in \mathbb{S}$,

$$\varphi(t,x) - \varphi(t,0) = x - \frac{1}{8}\big(A^{-1}D_x(u_{0x}^2)\big)(x)\big(1-\cos 2t\big) + \frac{u_0(x)-u_0(0)}{2}\sin 2t. \tag{5.9}$$

The right-hand side of this equation is well-defined for all times $t < T^*$, where

$$T^* = \frac{\pi}{2} + \arctan\bigg(\frac{1}{2}\min_{x\in\mathbb{S}} u_{0x}(x)\bigg).$$

Indeed, it follows from (5.8) that T^* is the first time for which the right-hand side of (5.9) ceases to be a bijective map $\mathbb{S} \to \mathbb{S}$.

The invariance of (5.9) under the transformation $\varphi(t,x) \mapsto \varphi(t,x) + c(t)$ for an arbitrary function $c(t)$ corresponds to the invariance of the Hunter–Saxton equation under the transformation $u(t) \mapsto u(t, \cdot - c(t)) + c'(t)$. It follows that the geodesic $\varphi(t,x)$ can be extended to all times $t < T^*$ unless $\lim_{t\uparrow T} |\varphi(t,0)| = \infty$ for some $T < T^*$. This artificial blow-up at some $T < T^*$ can be avoided by requiring that $u_0(0) = 0$ and that φ be a curve in M^k; this fixes $\varphi(t,0) = 0$ at all times. We summarize what we have obtained in a theorem.

THEOREM 5.3. *Suppose $u_0 \in H^k(\mathbb{S})$ satisfies $\langle u_0, u_0\rangle_{id} = \int_{\mathbb{S}} u_{0x}^2 dx = 4$ and $u_0(0) = 0$. Let $\varphi : [0, T^*) \to M^k$ be the unique geodesic with $\varphi(0) = id$ and $\varphi_t(0) = u_0$ existing for some maximal time $T^* > 0$.*

Then

$$\varphi(t) = id - \frac{1}{8}\big(A^{-1}D_x(u_{0x}^2)\big)\big(1-\cos 2t\big) + \frac{u_0}{2}\sin 2t, \tag{5.10}$$

and T^* is the first time for which $\varphi(t) : \mathbb{S} \to \mathbb{S}$ ceases to be bijective given by

$$T^* = \frac{\pi}{2} + \arctan\bigg(\frac{1}{2}\min_{x\in\mathbb{S}} u_{0x}(x)\bigg) < \frac{\pi}{2}.$$

Moreover, defining $u \in C([0,T^*); \mathbf{E}^k) \cap C^1([0,T^*); \mathbf{E}^{k-1})$ by

$$u(t) = \bigg(-\frac{1}{4}\big(A^{-1}D_x(u_{0x}^2)\big)\sin 2t + u_0\cos 2t\bigg)\circ\varphi(t)^{-1},$$

the set of solutions of the Hunter–Saxton equation with initial data u_0 (see Definition 5.1) is exactly the set of maps

$$\big\{t \mapsto u(t, \cdot - c(t)) + c'(t)\big\} \subset C([0,T); H^k(\mathbb{S})) \cap C^1([0,T); H^{k-1}(\mathbb{S})),$$

where $T \le T^*$ is the maximal time of existence, $c : [0,T) \to \mathbb{R}$ is an arbitrary C^1-function with $c(0) = c'(0) = 0$, and, if $T < T^*$, then $|c(t)| \to \infty$ as $t \uparrow T < T^*$.

Remark. For nonconstant initial data u_0 the assumption $\int_{\mathbb{S}} u_{0x}^2 dx = 4$ is just a matter of scaling; it sets the speed of the geodesic to be 4. The case of a constant u_0 is degenerate as it means that the initial velocity of the geodesic vanishes. For the sake of completeness we note that if u_0 is constant the general solution of (1.1) is $u(t,x) \equiv \bar{u}(t)$, where $\bar{u} : [0,T) \to \mathbb{R}$ is a C^1-function of time with $\bar{u}(0) = u_0$.

Acknowledgments. The research presented in this paper was carried out while the author was a visitor at the Mittag-Leffler Institute in Stockholm, Sweden. The author also thanks the two referees for helpful comments on a first version of the manuscript.

REFERENCES

[1] A. Bressan and A. Constantin, *Global solutions of the Hunter–Saxton equation*, SIAM J. Math. Anal., 37 (2005), pp. 996–1026.
[2] A. Constantin, *Existence of permanent and breaking waves for a shallow water equation: A geometric approach*, Ann. Inst. Fourier (Grenoble), 50 (2000), pp. 321–362.
[3] A. Constantin and B. Kolev, *On the geometric approach to the motion of inertial mechanical systems*, J. Phys. A, 35 (2002), pp. R51–R79.
[4] D. Ebin and J. E. Marsden, *Groups of diffeomorphisms and the motion of an incompressible fluid*, Ann. of Math. (2), 92 (1970), pp. 102–163.
[5] R. Hamilton, *The inverse function theorem of Nash and Moser*, Bull. Amer. Math. Soc., 7 (1982), pp. 65–222.
[6] J. K. Hunter and R. Saxton, *Dynamics of director fields*, SIAM J. Appl. Math., 51 (1991), pp. 1498–1521.
[7] J. K. Hunter and R. Saxton, *On a completely integrable nonlinear hyperbolic variational equation*, Phys. D, 79 (1994), pp. 361–386.
[8] J. K. Hunter and Y. Zheng, *On a nonlinear hyperbolic variational equation* I. *Global existence of weak solutions*, Arch. Rational Mech. Anal., 129 (1995), pp. 305–353.
[9] B. Khesin and G. Misiolek, *Euler equations on homogeneous spaces and Virasoro orbits*, Adv. Math., 176 (2003), pp. 116–144.
[10] S. Lang, *Differential and Riemannian Manifolds*, 3rd ed., Springer-Verlag, New York, 1995.
[11] J. Lenells, *Weak geodesic flow and global solutions of the Hunter-Saxton equation*, Discrete Contin. Dyn. Syst., 18 (2007), pp. 643–656.
[12] J. Lenells, *Riemannian geometry on the diffeomorphism group of the circle*, Ark. Mat., 45 (2007), pp. 297–325.
[13] G. Misiołek, *A shallow water equation as a geodesic flow on the Bott-Virasoro group*, J. Geom. Phys., 24 (1998), pp. 203–208.
[14] H. Omori, *Infinite-Dimensional Lie Groups*, Transl. Math. Monogr. 158, AMS, Providence, RI, 1997.
[15] V. Ovsienko and B. Khesin, *The (super) KdV equation as an Euler equation*, Funct. Anal. Appl., 21(1987), pp. 81–82.

SIAM J. MATH. ANAL.
Vol. 40, No. 1, pp. 278–291

ON THE SCHRÖDINGER EQUATION WITH DISSIPATIVE NONLINEARITIES OF DERIVATIVE TYPE*

NAKAO HAYASHI†, PAVEL I. NAUMKIN‡, AND HIDEAKI SUNAGAWA†

Abstract. We consider the cubic nonlinear Schrödinger equations of derivative type with small initial data. We present a structural condition on the nonlinear terms under which the corresponding Cauchy problem has a dissipative nature.

Key words. Schrödinger equation, dissipative nonlinearity, derivative coupling

AMS subject classifications. 35Q55, 35B40

DOI. 10.1137/070689103

1. Introduction.

We consider the initial value problem for the nonlinear Schrödinger equation of derivative type:

$$\begin{cases} i\partial_t u + \frac{1}{2}\partial_x^2 u = N(u, \partial_x u), & t > 0,\ x \in \mathbf{R}, \\ u(0,x) = u_0(x), & x \in \mathbf{R}, \end{cases} \tag{1.1}$$

where $i = \sqrt{-1}$, $\partial_t = \partial/\partial t$, $\partial_x = \partial/\partial x$, and u is a complex-valued unknown function. We will occasionally write u_x for $\partial_x u$, and $\overline{u}$ denotes the complex conjugate of u. The nonlinear term $N(u, u_x)$ is a cubic homogeneous polynomial in $(u, \overline{u}, u_x, \overline{u}_x)$ with complex coefficients, and it satisfies so-called gauge invariance, that is,

$$N(e^{i\theta}v, e^{i\theta}q) = e^{i\theta}N(v,q) \quad \text{for} \quad v, q \in \mathbf{C} \ \text{ and } \ \theta \in \mathbf{R}. \tag{1.2}$$

The aim of this paper is to present a structural condition on the nonlinear term N under which the corresponding forward Cauchy problem (1.1) has a dissipative nature. To explain the motivation, let us begin with the simplest case where N is independent of u_x, i.e., $N = \lambda|u|^2 u$ with $\lambda \in \mathbf{C}$. Then it is easy to see that

$$\|u(t)\|_{L^2}^2 - 2\,\mathrm{Im}\,\lambda \int_0^t \|u(\tau)\|_{L^4}^4\, d\tau = \|u_0\|_{L^2}^2, \tag{1.3}$$

which suggests a dissipative structure if $\mathrm{Im}\,\lambda < 0$. In fact, it is proved in [18] that the solution decays like $O((t\log t)^{-1/2})$ in L_x^∞ as $t \to +\infty$ if $\mathrm{Im}\,\lambda < 0$ and u_0 is small enough. Since the nontrivial free solution (i.e., the solution to (1.1) for $N \equiv 0$, $u_0 \neq 0$) only decays like $O(t^{-1/2})$, this gain of additional logarithmic time decay reflects a dissipative character. Now we turn our attention to the general gauge-invariant cubic nonlinear terms involving both u and u_x. Note that we cannot expect the conservation law like (1.3) anymore. However, as we shall show below, similar

*Received by the editors April 23, 2007; accepted for publication (in revised form) September 14, 2007; published electronically April 16, 2008.
http://www.siam.org/journals/sima/40-1/68910.html

†Department of Mathematics, Graduate School of Science, Osaka University, Toyonaka, Osaka 560-0043, Japan (nhayashi@math.wani.osaka-u.ac.jp, sunagawa@math.sci.osaka-u.ac.jp). The third author was partially supported by MEXT through a Grant-in-Aid for Young Scientists (B) (18740066).

‡Instituto de Matemáticas, Universidad Nacional Autónoma de México, Campus Morelia, AP 61-3 (Xangari), Morelia CP 58089, Michoacán, Mexico (pavelni@matmor.unam.mx). The research of this author is partially supported by CONACYT.

time decay is still valid if $\sup_{\xi\in\mathbf{R}} \operatorname{Im} N(1,i\xi) < 0$. It gives an answer to one of the questions left unsolved in the previous work [20].

Before stating our result, we introduce function spaces. For $s,\nu \ge 0$, let $H^{s,\nu}$ be the weighted Sobolev space given by $\{\phi \in L^2 : \|\phi\|_{H^{s,\nu}} = \|\langle x\rangle^\nu \langle i\partial_x\rangle^s \phi\|_{L^2} < \infty\}$, where $\langle x\rangle = (1+x^2)^{1/2}$. In particular we set $H^s = H^{s,0}$, which is typically an L^2-based Sobolev space of order s. The main result is as follows.

THEOREM 1.1. *Suppose that N satisfies*

$$\operatorname{Im} N(1,i\xi) \le 0 \quad \textit{for} \quad \xi \in \mathbf{R}. \tag{1.4}$$

Let $u_0 \in H^{2,1}\cap H^3$, and $\varepsilon = \|u_0\|_{H^{2,1}} + \|u_0\|_{H^3}$ is sufficiently small. Then (1.1) admits a unique global solution $u \in C([0,\infty); H^{2,1}\cap H^3)$. Moreover, the following asymptotic expression is valid as $t \to +\infty$ uniformly in $x \in \mathbf{R}$:

$$\begin{aligned} u(t,x) &= \frac{a(\frac{x}{t})\exp\Big\{ i\frac{x^2}{2t} - i|a(\frac{x}{t})|^2 \operatorname{Re} N(1,i\frac{x}{t}) \int_0^{\log t} \frac{d\sigma}{1-2\operatorname{Im} N(1,i\frac{x}{t})(|a(\frac{x}{t})|^2\sigma + b(\frac{x}{t}))} \Big\}}{\sqrt{t}\sqrt{1-2\operatorname{Im} N(1,i\frac{x}{t})\big(|a(\frac{x}{t})|^2\log t + b(\frac{x}{t})\big)}} \\ &\quad + O\big(t^{-3/4+\mu}\big), \end{aligned} \tag{1.5}$$

where $\mu > 0$ is an arbitrarily small constant, and $a(\xi)$, $b(\xi)$ are complex-valued continuous functions of $\xi \in \mathbf{R}$ which satisfy $|a(\xi)| \le C\varepsilon\langle\xi\rangle^{-2}$, $|b(\xi)| \le C\varepsilon^4\langle\xi\rangle^{-4}$ with some positive constant C.

Remark 1.1. As an immediate consequence of (1.5), we can see that

$$\|u(t)\|_{L^\infty} = O((t\log t)^{-1/2}) \quad \text{as} \quad t \to +\infty$$

if $\sup_{\xi\in\mathbf{R}} \operatorname{Im} N(1,i\xi) < 0$. Moreover, as we shall see in Remark 4.1 below, the L^2 norm of $u(t)$ also decays like $O((\log t)^{-1/2})$. Therefore, by interpolation, we obtain

$$\|u(t)\|_{L^p} = O\big(t^{-(1/2-1/p)}(\log t)^{-1/2}\big) \quad \text{as} \quad t \to +\infty$$

for all $p \in [2,\infty]$.

Remark 1.2. Our result can be also viewed as an extension of [9] (see also [3], [2], [4], [10], [13], [15], [16], [17], [18], and [21]) because (1.5) is reduced to

$$u(t,x) = \frac{1}{\sqrt{t}} a(x/t) e^{i\left(x^2/(2t) - |a(x/t)|^2 \operatorname{Re} N(1,ix/t)\log t\right)} + O\big(t^{-3/4+\mu}\big) \quad \text{as} \quad t \to +\infty$$

when $\operatorname{Im} N(1,i\xi) \equiv 0$. An example of the nonlinearity which was not considered previously but satisfies (1.4) is $N = \alpha\bar{u}u_x^2$ with $\operatorname{Im}\alpha > 0$.

Remark 1.3. When $\operatorname{Im} N(1,i\xi_0) > 0$ for some $\xi_0 \in \mathbf{R}$, the authors do not know any global existence or nonexistence results for (1.1). However, in view of the denominator of the leading term of (1.5), it is quite reasonable to expect that small amplitude solutions can blow up in finite time if condition (1.4) is violated. Concerning the lifespan T_ε of the solution for (1.1) with $u_0(x) = \varepsilon\varphi(x)$, the following explicit lower bound is obtained in [20]:

$$\liminf_{\varepsilon\downarrow 0} \varepsilon^2 \log T_\varepsilon \ge \frac{1}{\sup_{\xi\in\mathbf{R}} \big(2|\hat{\varphi}(\xi)|^2 \operatorname{Im} N(1,i\xi)\big)},$$

where $\hat{\varphi}$ denotes the Fourier transform of φ. It may be an interesting open problem to consider whether the corresponding upper estimate holds or not.

Remark 1.4. When we put

$$A(s,\xi)=\frac{a(\xi)\exp\Big\{-i|a(\xi)|^2\,\mathrm{Re}\,N(1,i\xi)\int_0^s\frac{d\sigma}{1-2\,\mathrm{Im}\,N(1,i\xi)(|a(\xi)|^2\sigma+b(\xi))}\Big\}}{\sqrt{1-2\,\mathrm{Im}\,N(1,i\xi)(|a(\xi)|^2s+b(\xi))}},$$

the asymptotic expression (1.5) can be interpreted as

$$u(t,x)\overset{t\to\infty}{\sim}\frac{e^{ix^2/(2t)}}{\sqrt{t}}A\Big(\log t,\frac{x}{t}\Big);\quad i\partial_sA=N(1,i\xi)|A|^2A.$$

This tells us that asymptotic behavior of the solution for (1.1) is characterized by that of the solution for the simpler ordinary differential equation. At the level of the reduced equation, the dissipative nature is transparent via the identity

$$\partial_s\big(|A(s,\xi)|^2\big)=2\,\mathrm{Im}\,N(1,i\xi)|A(s,\xi)|^4.$$

Note that analogous results have been obtained in [19] for nonlinear Klein–Gordon equations and in [14] for nonlinear wave equations.

The rest of this paper is organized as follows: In the next section, we recall several useful identities and inequalities. Section 3 is devoted to getting an a priori estimate, from which global existence follows immediately. After that, we will derive the large time asymptotics of $u(t)$ in section 4 by applying a lemma on ordinary differential equations.

2. Preliminaries. We collect here several identities and inequalities which are useful in the proof of our main theorem. In what follows, we will denote several positive constants by the same letter, C, which is possibly different line by line.

First we put $\mathcal{J}=\mathcal{J}(t)=x+it\partial_x$. It is well known that

$$[\partial_x,\mathcal{J}]=1\quad\text{and}\quad\Big[i\partial_t+\frac{1}{2}\partial_x^2,\mathcal{J}\Big]=0,\tag{2.1}$$

where $[\cdot,\cdot]$ denotes the commutator, i.e., $[\mathcal{A},\mathcal{B}]=\mathcal{A}\mathcal{B}-\mathcal{B}\mathcal{A}$ for linear operators $\mathcal{A}$ and $\mathcal{B}$. Also we have

$$\partial_x(\phi\overline{\psi})=\frac{1}{it}\Big\{(\mathcal{J}\phi)\overline{\psi}-\phi(\overline{\mathcal{J}\psi})\Big\},\tag{2.2}$$

$$\mathcal{J}\big(\theta\phi\overline{\psi}\big)=(\mathcal{J}\theta)\phi\overline{\psi}+\theta(\mathcal{J}\phi)\overline{\psi}-\theta\phi(\overline{\mathcal{J}\psi})\tag{2.3}$$

for smooth functions θ, ϕ, and ψ. Next we denote by $\mathcal{U}=\mathcal{U}(t)$ the free Schrödinger evolution group defined by

$$\big(\mathcal{U}\phi\big)(x)=\frac{e^{-i\pi/4}}{\sqrt{2\pi t}}\int_{\mathbf{R}}e^{i(x-y)^2/(2t)}\phi(y)\,dy.$$

From the relation $\mathcal{J}=\mathcal{U}x\mathcal{U}^{-1}$, we see that

$$\|\mathcal{F}\mathcal{U}^{-1}\phi\|_{H^1_\xi}\le C\|(1+|x|)\mathcal{U}^{-1}\phi\|_{L^2_x}\le C\big(\|\phi\|_{L^2_x}+\|\mathcal{J}\phi\|_{L^2_x}\big),\tag{2.4}$$

where

$$\big(\mathcal{F}\phi\big)(\xi)=\hat{\phi}(\xi)=\frac{1}{\sqrt{2\pi}}\int_{\mathbf{R}}e^{-iy\xi}\phi(y)\,dy.$$

It is also well known that $\mathcal{U}$ is decomposed into $\mathcal{MDFM}$, where

$$\big(\mathcal{M}\phi\big)(x) = e^{ix^2/(2t)}\phi(x),$$
$$\big(\mathcal{D}\phi\big)(x) = \frac{e^{-i\pi/4}}{\sqrt{t}}\phi\Big(\frac{x}{t}\Big).$$

With this notation, we set $\mathcal{W} = \mathcal{FMF}^{-1}$ so that $\mathcal{U} = \mathcal{MDWF}$. It follows from the inequalities $\|\phi\|_{L^\infty} \le C\|\phi\|_{L^2}^{1/2}\|\partial_x\phi\|_{L^2}^{1/2}$ and $|e^{ix^2/(2t)} - 1| \le Ct^{-1/2}|x|$ that

$$\|(\mathcal{W}-1)\phi\|_{L^\infty} \le Ct^{-1/4}\|\phi\|_{H^1}, \quad \|(\mathcal{W}^{-1}-1)\phi\|_{L^\infty} \le Ct^{-1/4}\|\phi\|_{H^1}. \tag{2.5}$$

Consequently we have

$$\begin{aligned}\|\phi - \mathcal{MDFU}^{-1}\phi\|_{L^\infty} &= \|\mathcal{MD}(\mathcal{W}-1)\mathcal{FU}^{-1}\phi\|_{L^\infty}\\ &= t^{-1/2}\|(\mathcal{W}-1)\mathcal{FU}^{-1}\phi\|_{L^\infty}\\ &\le Ct^{-3/4}\|\mathcal{FU}^{-1}\phi\|_{H^1}\\ &\le Ct^{-3/4}\big(\|\phi\|_{L^2} + \|\mathcal{J}\phi\|_{L^2}\big)\end{aligned} \tag{2.6}$$

and

$$\begin{aligned}\|\phi\|_{L^\infty} &\le \|\mathcal{MDFU}^{-1}\phi\|_{L^\infty} + \|\phi - \mathcal{MDFU}^{-1}\phi\|_{L^\infty}\\ &\le t^{-1/2}\|\mathcal{FU}^{-1}\phi\|_{L^\infty} + Ct^{-3/4}\big(\|\phi\|_{L^2} + \|\mathcal{J}\phi\|_{L^2}\big).\end{aligned} \tag{2.7}$$

Also it follows from the inequality $\|\mathcal{W}^{-1}\phi\|_{L^\infty} \le Ct^{1/2}\|\phi\|_{L^1}$ that

$$\begin{aligned}\|\mathcal{FU}^{-1}(\theta\phi\overline{\psi})\|_{L^\infty} &= \frac{1}{t}\Big\|\mathcal{W}^{-1}\Big\{(\mathcal{WFU}^{-1}\theta)(\mathcal{WFU}^{-1}\phi)\overline{(\mathcal{WFU}^{-1}\psi)}\Big\}\Big\|_{L^\infty}\\ &\le \frac{C}{t^{1/2}}\big\|(\mathcal{WFU}^{-1}\theta)(\mathcal{WFU}^{-1}\phi)(\mathcal{WFU}^{-1}\psi)\big\|_{L^1}\\ &\le \frac{C}{t^{1/2}}\|\theta\|_{L^2}\|\phi\|_{L^2}\big\|\mathcal{WFU}^{-1}\psi\big\|_{L^\infty}\\ &\le C\|\theta\|_{L^2}\|\phi\|_{L^2}\Big\{\|\psi\|_{L^\infty} + t^{-3/4}\big(\|\psi\|_{L^2} + \|\mathcal{J}\psi\|_{L^2}\big)\Big\}.\end{aligned} \tag{2.8}$$

3. A priori estimates. This section is devoted to getting suitable a priori estimate for the solution of (1.1). Since the local existence is well known (see, e.g., [1], [11], [13]), we can deduce global existence from this estimate. From now on, let $u(t)$ be the solution of (1.1) for $t \in [0,T]$ with some $T > 0$ and define

$$E_\delta(T) = \sup_{0\le t\le T}\left\{(1+t)^{-\delta}\sum_{j=0}^{1}\|\mathcal{J}^j u(t)\|_{H^{3-j}} + (1+t)^{1/2}\|u(t)\|_{W^{2,\infty}}\right\}$$

with $\delta \in (0, 1/8]$ fixed. Here and later on as well, $W^{k,p}$ denotes an L^p-based Sobolev space of order k. What we are going to show is the following.

LEMMA 3.1. *Let $K \ge 1$ and assume $E_\delta(T) \le K\varepsilon$, where $\varepsilon = \|u_0\|_{H^{2,1}} + \|u_0\|_{H^3}$. Then we have*

$$E_\delta(T) \le C_1 e^{C_2K^2\varepsilon^2}(1 + K^3\varepsilon^2 + K^5\varepsilon^4)\varepsilon,$$

where C_1, C_2 are positive constants independent of ε, K, and T, but possibly dependent on δ.

Once this lemma is established, global existence follows immediately. Indeed, when we put $\varepsilon_0 = K^{-2}$ and choose K so large that $C_1 e^{C_2 K^{-2}}(1+K^{-1}+K^{-3}) \le K/2$, it follows from the above lemma that $E_\delta(T) \le K\varepsilon$ implies $E_\delta(T) \le K\varepsilon/2$ for any $\varepsilon \in (0, \varepsilon_0]$. Then, by the continuity argument (see, e.g., [12]), we see that $E_\delta(T) \le K\varepsilon$ must hold as long as the solution exists. Therefore the local solution can be extended to the global one.

Now, we turn to the proof of Lemma 3.1. It will be divided into two parts, i.e., L^∞ and L^2 estimates. The argument below is a refinement of section 2 of [9].

3.1. L^∞ estimates. In this part, we consider the estimate of $(1+t)^{1/2}\|u(t)\|_{W^{2,\infty}}$. We set

$$\alpha(t,\xi) = \mathcal{F}\big(\mathcal{U}^{-1}u(t)\big)(\xi)$$

and $\alpha_k(t,\xi) = (i\xi)^k\alpha(t,\xi)$ so that $\partial_x^k u = \mathcal{M}\mathcal{D}\mathcal{W}\alpha_k$. Using the inequality (2.7) for $t \in [1,T]$ and the Sobolev imbedding for $t \in [0,1]$, we can see that the problem is reduced to getting the bound of $\|\alpha_k(t)\|_{L^\infty}$ for $t \in [1,T]$, $0 \le k \le 2$, under the assumption $E_\delta(T) \le K\varepsilon$. More precisely, our objective here is to obtain the following estimate:

$$\sup_{t\in[1,T]} \|\alpha_k(t)\|_{L^\infty} \le C(\varepsilon + K^3\varepsilon^3), \qquad k = 0,1,2.$$

Let us first consider the case of $k = 0$. Applying the operator $\mathcal{F}\mathcal{U}^{-1}$ to the equation, we have

$$\begin{aligned} i\partial_t\alpha &= \mathcal{F}\mathcal{U}^{-1}N(u,u_x) \\ &= \mathcal{W}^{-1}\mathcal{D}^{-1}\mathcal{M}^{-1}N(\mathcal{M}\mathcal{D}\mathcal{W}\alpha, \mathcal{M}\mathcal{D}\mathcal{W}\alpha_1) \\ &= \frac{1}{t}\mathcal{W}^{-1}N(\mathcal{W}\alpha, \mathcal{W}\alpha_1) \\ &= \frac{1}{t}N(1,i\xi)|\alpha|^2\alpha + \rho_0, \end{aligned}$$

where

$$\rho_0 = \frac{1}{t}\Big\{\mathcal{W}^{-1}N(\mathcal{W}\alpha, \mathcal{W}\alpha_1) - N(\alpha,\alpha_1)\Big\}.$$

Observe that (1.4) implies

$$\begin{aligned} \partial_t\big(|\alpha|^2\big) &= 2\,\mathrm{Im}\big(\overline{\alpha}\, i\partial_t\alpha\big) \\ &= \frac{2}{t}\,\mathrm{Im}N(1,i\xi)|\alpha|^4 + 2\,\mathrm{Im}\big(\overline{\alpha}\rho_0\big) \\ &\le 2|\alpha||\rho_0| \end{aligned}$$

and that (2.5) yields

$$\begin{aligned} \|\rho_0\|_{L^\infty} &\le t^{-1}\|(\mathcal{W}^{-1}-1)N(\mathcal{W}\alpha,\mathcal{W}\alpha_1)\|_{L^\infty} \\ &\quad + t^{-1}\|N(\mathcal{W}\alpha,\mathcal{W}\alpha_1) - N(\alpha,\alpha_1)\|_{L^\infty} \\ &\le CK^3\varepsilon^3 t^{-5/4}. \end{aligned} \tag{3.1}$$

So we deduce that

$$\|\alpha(t)\|_{L^\infty} \le \|\alpha(1)\|_{L^\infty} + \int_1^t \|\rho_0(\tau)\|_{L^\infty} d\tau \le C\varepsilon + CK^3\varepsilon^3.$$

Next we turn to the case of $k = 1, 2$. We split $N(u, u_x)$ into $g_0(u, u_x)u + g_1(u, u_x)u_x$, where $g_0(u, q) = \lambda_1|u|^2 + \lambda_2 u\overline{q} + \lambda_3|q|^2$, $g_1(u, q) = \lambda_4|u|^2 + \lambda_5 q\overline{u} + \lambda_6|q|^2$ with complex constants $\lambda_1, \dots, \lambda_6$, and we set

$$h_k = \partial_x^k\big(N(u, u_x)\big) - \sum_{j=0}^{1} g_j(u, u_x)\partial_x^{j+k}u$$

so that

$$i\partial_t\alpha_k = \mathcal{F}\mathcal{U}^{-1}\sum_{j=0}^{1} g_j(u, u_x)\partial_x^{j+k}u + \mathcal{F}\mathcal{U}^{-1}h_k.$$

To find out the leading term of the right-hand side, we analyze the action of $\mathcal{F}\mathcal{U}^{-1}$ to $g_j(u, u_x)\partial_x^{j+k}u$ carefully. For $j = 0$, we have

$$\begin{aligned}\mathcal{F}\mathcal{U}^{-1}\Big(g_0(u, u_x)\partial_x^k u\Big) &= \mathcal{W}^{-1}\mathcal{D}^{-1}\mathcal{M}^{-1}\Big(g_0(\mathcal{M}\mathcal{D}\mathcal{W}\alpha, \mathcal{M}\mathcal{D}\mathcal{W}\alpha_1)\mathcal{M}\mathcal{D}\mathcal{W}\alpha_k\Big)\\ &= \frac{1}{t}\mathcal{W}^{-1}\Big(g_0(\mathcal{W}\alpha, \mathcal{W}\alpha_1)\mathcal{W}\alpha_k\Big)\\ &= \frac{1}{t}g_0(\alpha, \alpha_1)\alpha_k + \sigma_k,\end{aligned}$$

where

$$\sigma_k = \frac{1}{t}\Big\{\mathcal{W}^{-1}\big(g_0(\mathcal{W}\alpha, \mathcal{W}\alpha_1)\mathcal{W}\alpha_k\big) - g_0(\alpha, \alpha_1)\alpha_k\Big\}.$$

For $j = 1$, let us introduce

$$\tilde{h}_k = g_1(u, u_x)\partial_x^{1+k}u + (\lambda_4 u\overline{u}_x + \lambda_5|u_x|^2 + \lambda_6 u_x\overline{u}_{xx})\partial_x^k u.$$

Then, noting the relation $\lambda_4\alpha\overline{\alpha_1} + \lambda_5|\alpha_1|^2 + \lambda_6\alpha_1\overline{\alpha_2} = -i\xi g_1(1, i\xi)|\alpha|^2$, we see that

$$\begin{aligned}&\mathcal{F}\mathcal{U}^{-1}\Big(g_1(u, u_x)\partial_x^{1+k}u\Big)\\ &= -\mathcal{F}\mathcal{U}^{-1}\big\{(\lambda_4 u\overline{u}_x + \lambda_5|u_x|^2 + \lambda_6 u_x\overline{u}_{xx})\partial_x^k u\big\} + \mathcal{F}\mathcal{U}^{-1}\tilde{h}_k\\ &= -\frac{1}{t}\mathcal{W}^{-1}\Big\{\big(\lambda_4(\mathcal{W}\alpha)(\overline{\mathcal{W}\alpha_1}) + \lambda_5|\mathcal{W}\alpha_1|^2 + \lambda_6(\mathcal{W}\alpha_1)(\overline{\mathcal{W}\alpha_2})\big)\mathcal{W}\alpha_k\Big\} + \mathcal{F}\mathcal{U}^{-1}\tilde{h}_k\\ &= -\frac{1}{t}(\lambda_4\alpha\overline{\alpha_1} + \lambda_5|\alpha_1|^2 + \lambda_6\alpha_1\overline{\alpha_2})\alpha_k + \tilde{\sigma}_k + \mathcal{F}\mathcal{U}^{-1}\tilde{h}_k\\ &= \frac{1}{t}i\xi g_1(1, i\xi)|\alpha|^2\alpha_k + \tilde{\sigma}_k + \mathcal{F}\mathcal{U}^{-1}\tilde{h}_k,\end{aligned}$$

where

$$\begin{aligned}\tilde{\sigma}_k = &-\frac{\lambda_4}{t}\Big\{\mathcal{W}^{-1}\big((\mathcal{W}\alpha)(\overline{\mathcal{W}\alpha_1})\mathcal{W}\alpha_k\big) - (\alpha\overline{\alpha_1})\alpha_k\Big\}\\ &-\frac{\lambda_5}{t}\Big\{\mathcal{W}^{-1}\Big(|\mathcal{W}\alpha_1|^2\mathcal{W}\alpha_k\Big) - |\alpha_1|^2\alpha_k\Big\}\\ &-\frac{\lambda_6}{t}\Big\{\mathcal{W}^{-1}\big((\mathcal{W}\alpha_1)(\overline{\mathcal{W}\alpha_2})\mathcal{W}\alpha_k\big) - \alpha_1\overline{\alpha_2}\alpha_k\Big\}.\end{aligned}$$

Summing up, we have

$$\begin{aligned} i\partial_t\alpha_k &= \big(g_0(1,i\xi)+i\xi g_1(1,i\xi)\big)|\alpha|^2\alpha_k+\sigma_k+\tilde\sigma_k+\mathcal{FU}^{-1}(h_k+\tilde h_k)\\ &= N(1,i\xi)|\alpha|^2\alpha_k+\rho_k, \end{aligned}$$

where $\rho_k=\sigma_k+\tilde\sigma_k+\mathcal{FU}^{-1}(h_k+\tilde h_k)$. Thus we deduce as in the previous case that

$$\|\alpha_k(t)\|_{L^\infty}\le\|\alpha_k(1)\|_{L^\infty}+\int_1^t\|\rho_k(\tau)\|_{L^\infty}\,d\tau\le C\varepsilon+CK^3\varepsilon^3,$$

provided that the estimate $\|\rho_k\|_{L^\infty}\le CK^3\varepsilon^3t^{-5/4}$ is verified. So the remaining task is to check this for $k=1,2$. In order to estimate $\|\mathcal{FU}^{-1}(h_k+\tilde h_k)\|_{L^\infty}$, we rewrite h_k as

$$\sum_{j=0}^1\sum_{l=1}^k\begin{pmatrix}k\\ l\end{pmatrix}\partial_x^{l-1}\Big(\partial_x\big(g_j(u,u_x)\big)\Big)\partial_x^{j+k-l}u$$

and apply (2.2) to $\partial_x\big(g_j(u,u_x)\big)$. Also we express $\tilde h_k$ as

$$\lambda_4u\partial_x\big(\overline{u}\partial_x^ku\big)+\lambda_5u_x\partial_x\big(\overline{u}\partial_x^ku\big)+\lambda_6u_x\partial_x\big(\overline{u}_x\partial_x^ku\big)$$

and use (2.2). Then we see that $h_k+\tilde h_k$ can be written as a linear combination of

$$\frac1t\big(\mathcal{J}\partial_x^{k_1}u\big)\big(\partial_x^{k_2}u\big)\overline{\big(\partial_x^{k_3}u\big)}\quad\text{or}\quad\frac1t\overline{\big(\mathcal{J}\partial_x^{k_1}u\big)}\big(\partial_x^{k_2}u\big)\big(\partial_x^{k_3}u\big)$$

with $k_1,k_2,k_3\le2$. Hence it follows from (2.8) that

$$\|\mathcal{FU}^{-1}(h_k+\tilde h_k)\|_{L^\infty}\le CK^3\varepsilon^3t^{-3/2+2\delta}. \tag{3.2}$$

On the other hand, we deduce as the derivation of (3.1) that

$$\|\sigma_k+\tilde\sigma_k\|_{L^\infty}\le CK^3\varepsilon^3t^{-5/4}. \tag{3.3}$$

From (3.2) and (3.3), it follows that

$$\|\rho_k\|_{L^\infty}\le CK^3\varepsilon^3t^{-5/4}+CK^3\varepsilon^3t^{-3/2+2\delta}\le CK^3\varepsilon^3t^{-5/4} \tag{3.4}$$

for $k=1,2$, as desired.

3.2. L^2 estimates. In the remainder of this section, we consider the bound of $(1+t)^{-\delta}\sum_{j=0}^1\|\mathcal{J}^ju(t)\|_{H^{3-j}}$. It is enough to show that

$$\sum_{j=0}^1\|\mathcal{J}^ju(t)\|_{L^2}\le C\varepsilon+CK^3\varepsilon^3(1+t)^\delta \tag{3.5}$$

and

$$\sum_{j=0}^1\|\partial_x^{3-j}\mathcal{J}^ju(t)\|_{L^2}\le Ce^{CK^2\varepsilon^2}(\varepsilon+K^3\varepsilon^3+K^5\varepsilon^5)(1+t)^\delta \tag{3.6}$$

for $t\in[0,T]$ under the assumption $E_\delta(T)\le K\varepsilon$.

First we consider the easier estimate (3.5). It follows from the standard energy method that

$$\frac{d}{dt}\|u(t)\|_{L^2} \le C\|u\|^2_{W^{1,\infty}}\|u\|_{H^1} \le CK^3\varepsilon^3(1+t)^{-1+\delta}.$$

Also, when we remember (2.1) and (2.3), we see that

$$\begin{aligned}&\left(i\partial_t + \frac{1}{2}\partial_x^2\right)\mathcal{J}u\\ &= \frac{\partial N}{\partial u}(u,u_x)\mathcal{J}u - \frac{\partial N}{\partial \overline{u}}(u,u_x)\overline{\mathcal{J}u} + \frac{\partial N}{\partial q}(u,u_x)\mathcal{J}\partial_x u - \frac{\partial N}{\partial \overline{q}}(u,u_x)\overline{\mathcal{J}\partial_x u},\end{aligned}$$

where the variable q is responsible for u_x, and $\partial/\partial u$, $\partial/\partial\overline{u}$, $\partial/\partial q$, $\partial/\partial\overline{q}$ are defined by

$$\begin{aligned}\frac{\partial}{\partial u} &= \frac{1}{2}\left(\frac{\partial}{\partial \operatorname{Re} u} - i\frac{\partial}{\partial \operatorname{Im} u}\right), \quad \frac{\partial}{\partial \overline{u}} = \frac{1}{2}\left(\frac{\partial}{\partial \operatorname{Re} u} + i\frac{\partial}{\partial \operatorname{Im} u}\right),\\ \frac{\partial}{\partial q} &= \frac{1}{2}\left(\frac{\partial}{\partial \operatorname{Re} q} - i\frac{\partial}{\partial \operatorname{Im} q}\right), \quad \frac{\partial}{\partial \overline{q}} = \frac{1}{2}\left(\frac{\partial}{\partial \operatorname{Re} q} + i\frac{\partial}{\partial \operatorname{Im} q}\right),\end{aligned}$$

respectively. Then the energy method again implies

$$\frac{d}{dt}\|\mathcal{J}u(t)\|_{L^2} \le C\|u\|^2_{W^{1,\infty}}(\|u\|_{H^1} + \|\mathcal{J}u\|_{H^1}) \le CK^3\varepsilon^3(1+t)^{-1+\delta},$$

which gives us (3.5).

Next we consider (3.6). To obtain this estimate, we use the gauge transformation technique since the standard energy estimate may cause a derivative loss, as we have just seen. The following energy inequality is established in section 2 of [13] (see also [1] and [11]).

LEMMA 3.2. *Let $\psi(t,x)$ be a smooth function satisfying*

$$i\partial_t\psi + \frac{1}{2}\partial_x^2\psi + b_1(t,x)\partial_x\psi + b_2(t,x)\partial_x\overline{\psi} = f(t,x)$$

for $(t,x) \in [0,T]\times\mathbf{R}$ with some smooth functions $b_1(t,x)$, $b_2(t,x)$, $f(t,x)$ and some $T>0$. Then we have

$$\frac{d}{dt}\|e^{P(t,\cdot)}\psi(t,\cdot)\|_{L^2} \le CB(t)\|e^{P(t,\cdot)}\psi(t,\cdot)\|_{L^2} + \|e^{P(t,\cdot)}f(t,\cdot)\|_{L^2}$$

for $t\in[0,T]$, where

$$P(t,x) = \int_{-\infty}^{x} \operatorname{Re} b_1(t,y)\,dy, \tag{3.7}$$

$$B(t) = \sum_{k=1}^{2}\left(\|\partial_x b_k(t,\cdot)\|_{L^\infty} + \|b_k(t,\cdot)\|^2_{L^\infty}\right) + \sup_{x\in\mathbf{R}}\left|\int_{-\infty}^{x}\partial_t b_1(t,y)\,dy\right|. \tag{3.8}$$

Since $\partial_x^{3-j}\mathcal{J}^j u$ satisfies

$$\begin{aligned}&\left(i\partial_t + \frac{1}{2}\partial_x^2\right)(\partial_x^{3-j}\mathcal{J}^j u)\\ &= \frac{\partial N}{\partial q}(u,u_x)\partial_x(\partial_x^{3-j}\mathcal{J}^j u) + (-1)^j\frac{\partial N}{\partial \overline{q}}(u,u_x)\partial_x(\overline{\partial_x^{3-j}\mathcal{J}^j u}) + R_j\end{aligned}$$

with

$$\|R_j\|_{L^2} \le CK^3\varepsilon^3(1+t)^{-1+\delta} \tag{3.9}$$

for $j = 0, 1$, we can apply Lemma 3.2 to obtain

$$\frac{d}{dt}\|e^{P(t)}\partial_x^{3-j}\mathcal{J}^j u(t)\|_{L^2} \le CB(t)\|e^{P(t)}\partial_x^{3-j}\mathcal{J}^j u(t)\|_{L^2} + \|e^{P(t)}R_j(t)\|_{L^2}, \tag{3.10}$$

where $P(t,x)$ and $B(t)$ are given by (3.7) and (3.8) with

$$b_1 = -\frac{\partial N}{\partial q}(u, u_x), \quad b_2 = (-1)^{j+1}\frac{\partial N}{\partial \overline{q}}(u, u_x).$$

Note that

$$\begin{aligned}
\|u(t)\|_{H^1_x} &= \|\langle\xi\rangle\mathcal{F}\mathcal{U}^{-1}u(t)\|_{L^2_\xi} \\
&= \Big\|\langle\xi\rangle^{-1}\big(\alpha(t) + \xi^2\alpha(t)\big)\Big\|_{L^2_\xi} \\
&\le \left(\int_{-\infty}^{\infty}\frac{d\xi}{1+\xi^2}\right)^{1/2}\big(\|\alpha(t)\|_{L^\infty_\xi} + \|\alpha_2(t)\|_{L^\infty_\xi}\big) \\
&\le CK\varepsilon,
\end{aligned}$$

whence

$$e^{\pm P(t,x)} \le e^{C\|u(t)\|^2_{H^1}} \le e^{CK^2\varepsilon^2}. \tag{3.11}$$

As for the estimate of $B(t)$, we observe that the identity

$$\partial_t\big(\phi\overline{\psi}\big) = -\frac{i}{2}\partial_x\big(\phi\overline{\psi_x} - \phi_x\overline{\psi}\big) + i\phi\overline{\left(i\partial_t + \frac{1}{2}\partial_x^2\right)\psi} - i\overline{\psi}\left(i\partial_t + \frac{1}{2}\partial_x^2\right)\phi$$

adapted to $b_1 = \sum_{j,k=0}^1 \lambda_{jk}(\partial_x^j u)\overline{(\partial_x^k u)}$. Then we have

$$\begin{aligned}
B(t) &\le C\|u\|^2_{W^{2,\infty}} + C\|u\|^4_{W^{1,\infty}} + C\sum_{j,k=0}^{1}\int_{\mathbf{R}}|\partial_x^j u|\,|\partial_x^k N(u,u_x)|\,dx \\
&\le C\|u\|^2_{W^{2,\infty}} + C\|u\|^4_{W^{1,\infty}} + C\|u\|^2_{W^{2,\infty}}\|u\|^2_{H^1} \\
&\le C(K^2\varepsilon^2 + K^4\varepsilon^4)(1+t)^{-1}.
\end{aligned} \tag{3.12}$$

By (3.9)–(3.12) we obtain

$$\begin{aligned}
\frac{d}{dt}\|e^{P(t)}\partial_x^{3-j}\mathcal{J}^j u(t)\|_{L^2} &\le C(K^2\varepsilon^2 + K^4\varepsilon^4)(1+t)^{-1}e^{CK^2\varepsilon^2}\|\partial_x^{3-j}\mathcal{J}^j u(t,\cdot)\|_{L^2} \\
&\quad + Ce^{CK^2\varepsilon^2}K^3\varepsilon^3(1+t)^{-1+\delta} \\
&\le Ce^{CK^2\varepsilon^2}(K^3\varepsilon^3 + K^5\varepsilon^5)(1+t)^{-1+\delta},
\end{aligned}$$

whence

$$\sum_{j=0}^{1}\|\partial_x^{3-j}\mathcal{J}^j u(t)\|_{L^2} \le e^{CK^2\varepsilon^2}\sum_{j=0}^{1}\|e^{P(t)}\partial_x^{3-j}\mathcal{J}^j u(t)\|_{L^2}$$
$$\le e^{CK^2\varepsilon^2}\left\{C\varepsilon + Ce^{CK^2\varepsilon^2}(K^3\varepsilon^3+K^5\varepsilon^5)\int_0^t(1+\tau)^{-1+\delta}\,d\tau\right\}$$
$$\le Ce^{CK^2\varepsilon^2}(\varepsilon+K^3\varepsilon^3+K^5\varepsilon^5)(1+t)^{\delta}.$$

The proof of Lemma 3.1 is complete. □

4. Large time asymptotics. Now we are in a position to show the asymptotic expression (1.5). In view of (2.6), it suffices to find the asymptotics of $\alpha(t,\xi)$. The following lemma is essentially due to [7], [8], [14], [19], though we shall state a slightly refined version here. For the convenience of the readers, the proof will be provided in the appendix.

LEMMA 4.1. *Let $\beta_0(\xi)$, $\kappa(\xi)$ be bounded continuous functions and suppose that $\operatorname{Im}\kappa(\xi)\le 0$ for all $\xi\in\mathbf{R}$. Let $r(t,\xi)$ satisfy*

$$\sup_{\xi\in\mathbf{R}}|r(t,\xi)| = O(t^{-1-\lambda}) \qquad (t\to+\infty)$$

with some constant $\lambda>0$. If $\beta(t,\xi)$ solves the differential equation

$$i\frac{\partial\beta}{\partial t} = \varepsilon^2\frac{\kappa(\xi)}{t}|\beta|^2\beta + \varepsilon^2 r(t,\xi), \quad \beta(1,\xi)=\beta_0(\xi) \tag{4.1}$$

for sufficiently small ε, then there exist continuous functions $\beta_\infty(\xi)$ and $\gamma(\xi)$, which satisfy $|\beta_\infty(\xi)|\le C$ and $|\gamma(\xi)|\le C\varepsilon^2$, such that

$$\beta(t,\xi) = \frac{\beta_\infty(\xi)\exp\left\{-i\varepsilon^2|\beta_\infty(\xi)|^2\operatorname{Re}\kappa(\xi)\int_0^{\log t}\frac{d\sigma}{1-2\varepsilon^2\operatorname{Im}\kappa(\xi)(|\beta_\infty(\xi)|^2\sigma+\gamma(\xi))}\right\}}{\sqrt{1-2\varepsilon^2\operatorname{Im}\kappa(\xi)(|\beta_\infty(\xi)|^2\log t+\gamma(\xi))}}$$
$$+O\big(t^{-\lambda+\mu}\big)$$

as $t\to+\infty$ uniformly in $\xi\in\mathbf{R}$, where μ is an arbitrarily small number.

Now, we set $\beta(t,\xi)=\varepsilon^{-1}\langle\xi\rangle^2\alpha(t,\xi)$, $\kappa(\xi)=\langle\xi\rangle^{-4}N(1,i\xi)$, and

$$r(t,\xi) = \frac{1}{\varepsilon^2}\left(i\frac{\partial\beta}{\partial t} - \varepsilon^2\frac{\kappa(\xi)}{t}|\beta|^2\beta\right).$$

Then, since

$$r(t,\xi) = \frac{1}{\varepsilon^3}\left\{\left(i\frac{\partial\alpha}{\partial t}-\frac{1}{t}N(1,i\xi)|\alpha|^2\alpha\right)+\left(i\frac{\partial\alpha_2}{\partial t}-\frac{1}{t}N(1,i\xi)|\alpha|^2\alpha_2\right)\right\}$$
$$= \frac{1}{\varepsilon^3}\Big(\rho_0(t,\xi)+\rho_2(t,\xi)\Big),$$

we deduce from (3.1) and (3.4) that

$$\sup_{\xi\in\mathbf{R}}|r(t,\xi)| \le \frac{1}{\varepsilon^3}(\|\rho_0(t)\|_{L^\infty}+\|\rho_2(t)\|_{L^\infty}) \le Ct^{-5/4}.$$

Therefore we can apply Lemma 4.1 with $\lambda = 1/4$ to get the asymptotics of β as $t \to +\infty$. Putting $a(\xi) = e^{-i\pi/4}\varepsilon\langle\xi\rangle^{-2}\beta_\infty(\xi)$ and $b(\xi) = \varepsilon^2\langle\xi\rangle^{-4}\gamma(\xi)$, we have

$$\alpha(t,\xi) = \frac{a(\xi)\exp\left\{i\pi/4 - i|a(\xi)|^2\,\mathrm{Re}N(1,i\xi)\int_0^{\log t}\frac{d\sigma}{1-2\,\mathrm{Im}N(1,i\xi)(|a(\xi)|^2\sigma+b(\xi))}\right\}}{\sqrt{1-2\,\mathrm{Im}N(1,i\xi)(|a(\xi)|^2\log t + b(\xi))}} + O\big(t^{-1/4+\mu}\big)$$

as $t \to +\infty$. Finally, using (2.6), we arrive at the desired asymptotic expression for $u(t,x)$. □

Remark 4.1. We can also deduce the L^2-decay of $u(t)$ because

$$\|u(t)\|_{L^2_x} = \varepsilon\|\langle\xi\rangle^{-2}\beta(t)\|_{L^2_\xi} \le C\varepsilon\|\beta(t)\|_{L^\infty_\xi}$$

and

$$\|\beta(t)\|_{L^\infty_\xi} = O((\log t)^{-1/2}) \quad \text{as} \quad t \to +\infty$$

if $\sup_{\xi\in\mathbf{R}} \mathrm{Im}\,N(1,i\xi) < 0$.

Remark 4.2. It is possible to weaken (1.2) to a certain extent, but impossible to remove it completely. In fact, when (1.2) is replaced by

$$N(e^{i\theta},0) = e^{i\theta}N(1,0) \quad \text{for} \quad \theta \in \mathbf{R}, \tag{4.2}$$

we can modify the above argument combining the idea of [5], [6] (see also the appendix of [19]) to show that Theorem 1.1 is still valid if $N(1,i\xi)$ in the statement is replaced by

$$\frac{1}{2\pi}\int_0^{2\pi} N(e^{i\theta}, i\xi e^{i\theta})e^{-i\theta}\,d\theta.$$

Note that (4.2) is just what excludes u^3, $\overline{u}^3$, $u\overline{u}^2$ from all possible cubic nonlinear terms, but it is not a technical assumption because for these three nonlinearities we can find a class of initial data for which the solution has another kind of asymptotic profile than (1.5) (see [7] and [8] for details).

Appendix. We give a proof of Lemma 4.1 following [19] and [14] with some modifications. Because of the uniqueness for (4.1), $\beta(t,\xi)$ admits the decomposition

$$\beta(t,\xi) = \frac{p(t,\xi)}{\sqrt{q(t,\xi)}},$$

where $p(t,\xi)$ and $q(t,\xi)$ satisfy

$$\begin{cases} \partial_t p(t,\xi) = -i\varepsilon^2\dfrac{\mathrm{Re}\,\kappa(\xi)}{t}\dfrac{\big|p(t,\xi)\big|^2}{q(t,\xi)}p(t,\xi) - i\varepsilon^2\sqrt{q(t,\xi)}r(t,\xi), \\ \partial_t q(t,\xi) = -2\varepsilon^2\dfrac{\mathrm{Im}\,\kappa(\xi)}{t}\big|p(t,\xi)\big|^2, \\ p(1,\xi) = \beta_0(\xi), \quad q(1,\xi) = 1. \end{cases} \tag{A.1}$$

Note that $p(t,\xi)$ is complex-valued, while $q(t,\xi)$ is real and strictly positive. In order to obtain the desired conclusion, it is sufficient to get the asymptotics of $p(t,\xi)$ and $q(t,\xi)$ as $t \to +\infty$.

We first show that there exists a positive constant A such that

$$\sup_{(t,\xi)\in[1,\infty)\times\mathbf{R}} |p(t,\xi)| < A. \tag{A.2}$$

We shall argue by contradiction: Suppose that for any $A > \sup_{\xi\in\mathbf{R}} |\beta_0(\xi)|$ there exists a finite time $T_A \in (1,\infty)$ such that

$$\sup_{(t,\xi)\in[1,T_A)\times\mathbf{R}} |p(t,\xi)| \le A \quad \text{and} \quad \sup_{\xi\in\mathbf{R}} |p(T_A,\xi)| = A.$$

Then, from the second equation of (A.1), we have

$$1 \le q(t,\xi) \le 1 - 2\varepsilon^2 \operatorname{Im}\kappa(\xi) A^2 \log t$$

for $t \in [1, T_A]$. On the other hand, it follows from the first equation of (A.1) that

$$\begin{aligned}
\partial_t\Big(|p(t,\xi)|^2\Big) &= 2\operatorname{Re}\Big(\overline{p(t,\xi)}\partial_t p(t,\xi)\Big) \\
&= 2\varepsilon^2 \operatorname{Re}\Big(\overline{p(t,\xi)}\sqrt{q(t,\xi)}r(t,\xi)\Big) \\
&\le 2\varepsilon^2 \big|p(t,\xi)\big|\,\big|\sqrt{q(t,\xi)}r(t,\xi)\big|.
\end{aligned}$$

So we have

$$\begin{aligned}
|p(T_A,\xi)| &\le |\beta_0(\xi)| + \int_1^{T_A} \varepsilon^2 \big|\sqrt{q(\tau,\xi)}r(\tau,\xi)\big|\,d\tau \\
&\le C + \int_1^{\infty} C\varepsilon^2 (1 + A^2\varepsilon^2 \log\tau)^{1/2} \tau^{-1-\lambda} d\tau \\
&\le C(1 + A\varepsilon).
\end{aligned}$$

When we choose $A = 4C$ and $\varepsilon_1 = 1/A$, we have

$$\sup_{\xi\in\mathbf{R}} |p(T_A,\xi)| \le \frac{A}{2} < A$$

for $\varepsilon \in (0, \varepsilon_1]$, which is the desired contradiction. Hence (A.2) must hold for some A. Also we have

$$1 \le q(t,\xi) \le 1 - 2\varepsilon^2 \operatorname{Im}\kappa(\xi) A^2 \log t$$

for any $t \ge 1$. Next we define

$$\Theta(t,\xi) = \int_1^t \frac{\operatorname{Re}\kappa(\xi)|p(\tau,\xi)|^2}{q(\tau,\xi)} \frac{d\tau}{\tau}$$

so that $\partial_t(p(t,\xi)e^{i\varepsilon^2\Theta(t,\xi)}) = -i\varepsilon^2\sqrt{q(t,\xi)}r(t,\xi)e^{i\varepsilon^2\Theta(t,\xi)}$. Since

$$\sup_{\xi\in\mathbf{R}} \Big|\sqrt{q(t,\xi)}r(t,\xi)\Big| \le C(1+\varepsilon^2\log t)^{1/2} t^{-1-\lambda} \le C t^{-1-\lambda+\mu},$$

we obtain

$$\sup_{\xi\in\mathbf{R}} |p(t,\xi) - p_\infty(\xi) e^{-i\varepsilon^2\Theta(t,\xi)}| \le C\varepsilon^2 t^{-\lambda+\mu},$$

where

$$p_\infty(\xi) = \beta_0(\xi) - i\varepsilon^2 \int_1^\infty \sqrt{q(\tau,\xi)} r(\tau,\xi) e^{i\varepsilon^2 \Theta(\tau,\xi)}\, d\tau.$$

We also set $q_\infty(t,\xi) = 1 - 2\varepsilon^2 \operatorname{Im}\kappa(\xi)\big(|p_\infty(\xi)|^2 \log t + \gamma(\xi)\big)$ with

$$\gamma(\xi) = \int_1^\infty \big(|p(\tau,\xi)|^2 - |p_\infty(\xi)|^2\big)\frac{d\tau}{\tau}.$$

Noting that

$$\begin{aligned}\big||p(t,\xi)|^2 - |p_\infty(\xi)|^2\big| &\le \big|p(t,\xi) - p_\infty(\xi) e^{-i\varepsilon^2\Theta(t,\xi)}\big|\big(|p(t,\xi)| + |p_\infty(\xi)|\big)\\ &\le C\varepsilon^2 t^{-\lambda+\mu},\end{aligned}$$

we can see that

$$|q(t,\xi) - q_\infty(t,\xi)| \le 2\varepsilon^2 |\operatorname{Im}\kappa(\xi)| \int_t^\infty \big||p(\tau,\xi)|^2 - |p_\infty(\xi)|^2\big|\frac{d\tau}{\tau} \le C\varepsilon^4 t^{-\lambda+\mu}.$$

Let us also introduce

$$\Phi(t,\xi) = \frac{\operatorname{Re}\kappa(\xi)}{t}\left(\frac{|p(t,\xi)|^2}{q(t,\xi)} - \frac{|p_\infty(\xi)|^2}{q_\infty(t,\xi)}\right).$$

Then we have

$$i\varepsilon^2\Theta(t,\xi) = i\varepsilon^2 |p_\infty(\xi)|^2 \operatorname{Re}\kappa(\xi) \int_1^t \frac{d\tau}{q_\infty(\tau,\xi)} + i\varepsilon^2 \int_1^\infty \Phi(\tau,\xi)\, d\tau - i\varepsilon^2 \int_t^\infty \Phi(\tau,\xi)\, d\tau$$

and

$$\begin{aligned}\int_t^\infty |\Phi(\tau,\xi)|\, d\tau &\le C\int_t^\infty \left(\frac{\big||p(\tau,\xi)|^2 - |p_\infty(\xi)|^2\big|}{q(\tau,\xi)} + \frac{|p_\infty(\xi)|^2\big|q(\tau,\xi) - q_\infty(\tau,\xi)\big|}{q(\tau,\xi) q_\infty(\tau,\xi)}\right)\frac{d\tau}{\tau}\\ &\le C\varepsilon^2 \int_t^\infty \frac{d\tau}{\tau^{1+\lambda-\mu}}\\ &\le C\varepsilon^2 t^{-\lambda+\mu}.\end{aligned}$$

Therefore, putting $\beta_\infty(\xi) = p_\infty(\xi)\exp\big(-i\varepsilon^2\int_1^\infty \Phi(\tau,\xi)\, d\tau\big)$, we obtain

$$\begin{aligned}p(t,\xi) &= p_\infty(\xi) e^{-i\varepsilon^2\Theta(t,\xi)} + O(t^{-\lambda+\mu})\\ &= \beta_\infty(\xi)\exp\left(-i\varepsilon^2 |p_\infty(\xi)|^2 \operatorname{Re}\kappa(\xi)\int_1^t \frac{d\tau}{q_\infty(\tau,\xi)}\right) + O(t^{-\lambda+\mu})\\ &= \beta_\infty(\xi)\exp\left(-i\varepsilon^2 |\beta_\infty(\xi)|^2 \operatorname{Re}\kappa(\xi)\int_0^{\log t} \frac{d\sigma}{1 - 2\varepsilon^2 \operatorname{Im}\kappa(\xi)\big(|\beta_\infty(\xi)|^2\sigma + \gamma(\xi)\big)}\right)\\ &\quad + O(t^{-\lambda+\mu})\end{aligned}$$

as well as

$$\begin{aligned}\frac{1}{\sqrt{q(t,\xi)}} &= \frac{1}{\sqrt{q_\infty(t,\xi)}} + \frac{q_\infty(t,\xi) - q(t,\xi)}{\sqrt{q(t,\xi)q_\infty(t,\xi)}\big(\sqrt{q(t,\xi)} + \sqrt{q_\infty(t,\xi)}\big)}\\ &= \frac{1}{\sqrt{1 - 2\varepsilon^2 \operatorname{Im}\kappa(\xi)\big(|\beta_\infty(\xi)|^2 \log t + \gamma(\xi)\big)}} + O(t^{-\lambda+\mu}),\end{aligned}$$

whence

$$\begin{aligned}\beta(t,\xi) &= \frac{p(t,\xi)}{\sqrt{q(t,\xi)}}\\ &= \frac{\beta_\infty(\xi)\exp\left(-i\varepsilon^2|\beta_\infty(\xi)|^2\operatorname{Re}\kappa(\xi)\int_0^{\log t}\frac{d\sigma}{1-2\varepsilon^2\operatorname{Im}\kappa(\xi)\left(|\beta_\infty(\xi)|^2\sigma+\gamma(\xi)\right)}\right)}{\sqrt{1-2\varepsilon^2\operatorname{Im}\kappa(\xi)\left(|\beta_\infty(\xi)|^2\log t+\gamma(\xi)\right)}}\\ &\quad + O(t^{-\lambda+\mu})\end{aligned}$$

as $t\to+\infty$. □

REFERENCES

[1] H. Chihara, *Local existence for the semilinear Schrödinger equations in one space dimension*, J. Math. Kyoto Univ., 34 (1994), pp. 353–367.

[2] N. Hayashi and P. I. Naumkin, *Asymptotic behavior in time of solutions to the derivative nonlinear Schrödinger equation revisited*, Discrete Contin. Dynam. Systems, 3 (1997), pp. 383–400.

[3] N. Hayashi and P. I. Naumkin, *Asymptotic behavior in time of solutions to the derivative nonlinear Schrödinger equation*, Ann. Inst. H. Poincaré Phys. Théor., 68 (1998), pp. 159–177.

[4] N. Hayashi and P. I. Naumkin, *Asymptotics for large time of solutions to the nonlinear Schrödinger and Hartree equations*, Amer. J. Math., 120 (1998), pp. 369–389.

[5] N. Hayashi and P. I. Naumkin, *Large time behavior of solutions for derivative cubic nonlinear Schrödinger equations without a self-conjugate property*, Funkcial. Ekvac., 42 (1999), pp. 311–324.

[6] N. Hayashi and P. I. Naumkin, *Asymptotics of small solutions to nonlinear Schrödinger equations with cubic nonlinearities*, Int. J. Pure Appl. Math., 3 (2002), pp. 255–273.

[7] N. Hayashi and P. I. Naumkin, *Large time behavior for the cubic nonlinear Schödinger equation*, Canad. J. Math., 54 (2002), pp. 1065–1085.

[8] N. Hayashi and P. I. Naumkin, *On the asymptotics for cubic nonlinear Schrödinger equations*, Complex Var. Theory Appl., 49 (2004), pp. 339–373.

[9] N. Hayashi, P. I. Naumkin, and H. Uchida, *Large time behavior of solutions for derivative cubic nonlinear Schrödinger equations*, Publ. Res. Inst. Math. Sci., 35 (1999), pp. 501–513.

[10] N. Hayashi and T. Ozawa, *Modified wave operators for the derivative nonlinear Schrödinger equation*, Math. Ann., 298 (1994), pp. 557–576.

[11] N. Hayashi and T. Ozawa, *Remarks on nonlinear Schrödinger equations in one space dimension*, Differential Integral Equations, 7 (1994), pp. 453–461.

[12] L. Hörmander, *Lectures on Nonlinear Hyperbolic Differential Equations*, Math. Appl. (Berlin) 26, Springer-Verlag, Berlin, 1997.

[13] S. Katayama and Y. Tsutsumi, *Global existence of solutions for nonlinear Schrödinger equations in one space dimension*, Comm. Partial Differential Equations, 19 (1994), pp. 1971–1997.

[14] H. Kubo, *Asymptotic behavior of solutions to semilinear wave equations with dissipative structure*, Discrete Contin. Dyn. Syst. (Suppl.), 2007 (2007), pp. 602–613.

[15] H. Lindblad and A. Soffer, *Scattering and small data completeness for the critical nonlinear Schrödinger equation*, Nonlinearity, 19 (2006), pp. 345–353.

[16] T. Ozawa, *Long range scattering for nonlinear Schrödinger equations in one space dimension*, Comm. Math. Phys., 139 (1991), pp. 479–493.

[17] T. Ozawa, *On the nonlinear Schrödinger equations of derivative type*, Indiana Univ. Math. J., 45 (1996), pp. 137–163.

[18] A. Shimomura, *Asymptotic behavior of solutions for Schrödinger equation with dissipative nonlinearities*, Comm. Partial Differential Equations, 31 (2006), pp. 1407–1423.

[19] H. Sunagawa, *Large time behavior of solutions to the Klein-Gordon equation with nonlinear dissipative terms*, J. Math. Soc. Japan, 58 (2006), pp. 379–400.

[20] H. Sunagawa, *Lower bounds of the lifespan of small data solutions to the nonlinear Schrödinger equations*, Osaka J. Math., 43 (2006), pp. 771–789.

[21] Y. Tsutsumi, *The null gauge condition and the one dimensional nonlinear Schrödinger equation with cubic nonlinearity*, Indiana Univ. Math. J., 43 (1994), pp. 241–254.

SIAM J. MATH. ANAL.
Vol. 40, No. 1, pp. 292–305

QUASILINEAR PARABOLIC SYSTEMS WITH MIXED BOUNDARY CONDITIONS ON NONSMOOTH DOMAINS*

MATTHIAS HIEBER† AND JOACHIM REHBERG‡

Abstract. In this paper we investigate quasilinear systems of reaction-diffusion equations with mixed Dirichlet–Neumann boundary conditions on nonsmooth domains. Using techniques from maximal regularity and heat-kernel estimates we prove the existence of a unique solution to systems of this type.

Key words. quasilinear parabolic system, mixed Dirichlet–Neumann conditions, L^∞-coefficients

AMS subject classifications. Primary, 35A05, 35B65; Secondary, 35K15/20

DOI. 10.1137/070683829

1. Introduction.

The theory of quasilinear parabolic systems has many applications to evolution problems in natural sciences, see, e.g., [2], [1], [5], [6], [20], [10], [16], [30], and [41]. In this paper we investigate in particular systems of reaction-diffusion equations with *mixed* Dirichlet–Neumann boundary conditions on nonsmooth domains $\Omega \subset \mathbb{R}^n$ for $n = 2, 3$ of the form

$$\begin{aligned} u_k' - \operatorname{div}(G_k(v)\mu_k \nabla v_k) &= R_k(t,v), \qquad t \in]T_0, T[, x \in \Omega, \\ u_k &= b_k\, F_k(v_k), \qquad t \in [T_0, T[, x \in \Omega, \\ \nu \cdot \mu_k \nabla v_k &= 0, \qquad t \in [T_0, T[, x \in \Gamma_N, \\ v_k &= \phi_k, \qquad t \in [T_0, T[, x \in \Gamma_D, \\ v_k(T_0) &= v_{0k}, \qquad x \in \Omega. \end{aligned} \tag{1.1}$$

Here $v = (v_1, \dots, v_m)$, $\mu_k \in L^\infty(\Omega, M_{n\times n})$ are diffusion coefficients, $b_k \in L^\infty(\Omega)$ are reference densities, and R_k, G_k, F_k denote the reaction, diffusion, and superposition terms for $k \in \{1, \dots, m\}$.

In many concrete problems which are described as a system of the form (1.1), the underlying domain is nonsmooth and the coefficient functions b_k and μ_k are discontinuous. We therefore aim for minimal smoothness assumptions on the boundary $\partial\Omega$ of Ω, the coefficient functions b_k and μ_k, as well as on the interface between the Neumann boundary part Γ_N of $\partial\Omega$ and the Dirichlet boundary part $\Gamma_D = \partial\Omega \setminus \Gamma_N$. More precisely, we generally assume that $\Omega \subset \mathbb{R}^n$ is a Lipschitz domain and $\Omega \cup \Gamma_N$ is regular in the sense of Gröger (see the definition below). Note that the situation where the boundary of Ω is smooth and consists of two *separated* parts, one with Dirichlet and the other with Neumann boundary conditions, has been studied before by Amann in [2].

Our setting includes even nonlocal diffusion terms as occurring in models describing the diffusion of bacteria (see [7], [8] and the references therein). In detail, the velocity at which the motion takes place is given by Fourier's law and the (relative)

*Received by the editors February 27, 2007; accepted for publication (in revised form) September 14, 2007; published electronically April 23, 2008.
http://www.siam.org/journals/sima/40-1/68382.html
†Technische Universität Darmstadt, Fachbereich Mathematik, Schlossgartenstr. 7, D-64298 Darmstadt, Germany (hieber@mathematik.tu-darmstadt.de).
‡Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, D-10117 Berlin, Germany (rehberg@wias-berlin.de).

diffusion coefficient $G = G_k$ depends on the solution v in the form

$$G(v) = \eta\left(\int_\Omega z(x)v(x)\,dx\right),$$

where η is a (strictly) positive, continuously differentiable function on $\mathbb{R}$.

Our approach includes reaction terms R_k depending discontinously on time t, which is important in many examples (see [41], [26], [30]), in particular in the control theory of parabolic equations. Alternatively, the reader should think, e.g., of a manufacturing process for semiconductors, where at a certain moment light is switched on/off and, of course, parameters in the chemical process change abruptly.

An interesting example for a reaction term stems from the thermistor problem (see [38], [3] and the references therein, see also [6]) which describes the combined processes of heat conduction and electrical conduction in a body. This is of importance in the industrially important process of electrical welding. To be precise, the reaction term is of the form $\sigma(v)|\nabla\varphi|^2$, where φ is the solution of an auxiliary elliptic problem, into which v enters as a parameter (see details in section 4.2). Observe that the quadratic gradient term is a critical one and in general not easy to handle (see [38]), while in our context it does not cause additional difficulties.

Note that the original formulation of the evolution equation in terms of balance laws takes the form (see [39, Ch. 21], see also [5])

$$\frac{\partial}{\partial t}\int_{\Omega'} u_k\,dx + \int_{\partial\Omega'} \nu\cdot j_k\,d\sigma = \int_{\Omega'} R_k\,dx; \quad j_k = j_k(v) = G_k(v)\mu_k\nabla v_k, \tag{1.2}$$

where Ω' stands for any (Lipschitzian) subdomain of Ω. Within the variational theory of weak solutions, however, the indicator functions of the subdomains are not admissible test functions. Therefore the integral formulation (1.2) is equivalent to the above evolution equation only if the weak solutions have some additional regularity. It is the main advantage of the present concept that the divergence of the corresponding current $j_k(v)$ indeed is a function, not only a distribution. In a strict sense, only this justifies the application of Gauss' theorem to calculate the normal components of the currents over boundaries of suitable subdomains. Moreover, the fact $\operatorname{div} j_k \in L^p$ is also of importance for the numerical treatment of (1.1), as the formulation (1.2) is the basis of finite volume methods (see [18])—namely in the sense of local balances. Global existence results for (1.1) cannot be expected within such a general approach (see, e.g., [17] or [6] and the references therein).

In contrast to many papers where existence and uniqueness results for quasilinear parabolic systems are based on the construction of an appropriate evolution operator (see, e.g., [1]), our approach relies heavily on maximal L^p-estimates for the linear part of (1.1). In fact, after rewriting (1.1) as an abstract evolution equation in $L^p(\Omega)^m$ of the form

$$\begin{aligned} w' - H(t,w)\big(\operatorname{div}(\mu\nabla w)\big) &= S(t,w),\\ w(T_0) &= v_0 - \phi(T_0), \end{aligned} \tag{1.3}$$

our strategy to solve (1.3) follows the approach of Clément and Li [10] and Prüss [34]. The advantage of the given situation (1.1) is that subtle techniques from harmonic analysis as well as heat-kernel methods can be used to prove the central L^p-estimates of the linear part. In order to apply these methods in our situation, one needs embedding properties of certain interpolation spaces between the domain of the L^p-realization of

the underlying elliptic operators and $L^p(\Omega)$ into $W^{1,2p}(\Omega)$. This embedding property rests on the assumption that the operators formally defined by

$$-\nabla \cdot \mu_k \nabla + 1 : W^{1,q}_{\Gamma_N}(\Omega) \to W^{-1,q}_{\Gamma_N}(\Omega)$$

provide topological isomorphisms for some $q > n$. Note that this assumption restricts the physical dimension of the problems to two and three, because one knows that the solution for a mixed boundary value problem generically has singularities, precisely, one cannot expect that its gradient lies in L^4; see Shamir's famous counterexample [37]. On the other hand, in two and three dimensions the assumption is fulfilled for many geometric constellations and (even discontinuous) coefficient functions; see section 4.

2. Preliminaries. Let $\Omega \subset \mathbb{R}^n$ be a bounded Lipschitz domain and assume that $n = 2$ or $n = 3$. (Concerning the notions "Lipschitz domain" and "domain with Lipschitz boundary" we follow [23, section 1.2.1].) Denote by $\Gamma_N \subset \partial\Omega$ an open subset of $\partial\Omega$. For $1 < q < \infty$ we define $W^{1,q}_{\Gamma_N}(\Omega)$ as the closure of

$$\{\psi|_\Omega : \psi \in C_c^\infty(\mathbb{R}^n),\ \operatorname{supp}\psi \cap (\partial\Omega \backslash \Gamma_N) = \emptyset\}$$

in the Sobolev space $W^{1,q}(\Omega)$. If $q = 2$, we write $H^1(\Omega)$ or $H^1_{\Gamma_N}(\Omega)$ instead of $W^{1,2}(\Omega)$ or $W^{1,2}_{\Gamma_N}(\Omega)$. Of course, if $\Gamma_N = \emptyset$, then $W^{1,q}_{\Gamma_N}(\Omega) = W^{1,q}_0(\Omega)$. Moreover, throughout this work we always suppose that $\Omega \cup \Gamma_N$ is regular in the sense of Gröger (see [24]), this means that, for all $x \in \partial\Omega$ there exist open sets $U_x, V_x \subset \mathbb{R}^n$ and a bi-Lipschitz transform Ψ_x from U_x onto V_x such that $x \in U_x, \Psi_x(x) = 0$ and $\Psi_x(U_x \cap (\Omega \cup \Gamma_N))$ coincides with one of the sets

$$\begin{aligned}
E_1 &:= \{x \in \mathbb{R}^n : \max_{l=1\ldots,n} |x_l| < 1, x_n < 0\},\\
E_2 &:= \{x \in \mathbb{R}^n : \max_{l=1,\ldots,n} |x_l| < 1, x_n \le 0\},\\
E_3 &:= \{x \in E_2 : x_n < 0 \text{ or } x_1 > 0\}.
\end{aligned}$$

It is not hard to see that every Lipschitz domain, and also its closure, is regular in the sense of Gröger, the corresponding model sets are then E_1 or E_2, respectively; see [23]. Moreover, if $\Omega \subset \mathbb{R}^2$ is a bounded Lipschitz domain and $\partial\Omega \setminus \Gamma_N$ is the finite union of (nondegenerate) closed arc pieces from the boundary, then $\Omega \cup \Gamma_N$ is regular in the sense of Gröger.

Finally, for $k \in \{1, \ldots, m\}$, let $\mu_k \in L^\infty(\Omega, M_{n\times n})$, where $M_{n\times n}$ denotes the set of all real, symmetric $n \times n$ matrices. Suppose that additionally

$$\inf_{x\in\Omega} \inf_{|\varsigma|=1} \mu_k(x)\varsigma \cdot \varsigma > 0. \tag{2.1}$$

For a closed subspace $V \subseteq H^1(\Omega)$ such that $H^1_0(\Omega) \subseteq V$ we define the form $a_k : V \times V \to \mathbb{R}$ by

$$a_k(u, v) := -\int_\Omega \mu_k \nabla u \cdot \nabla v \, \mathrm{d}x, \quad u, v \in V.$$

The form induces a continuous mapping $\mathcal{A}_k : V \to V'$ such that

$$a_k(u, v) = (\mathcal{A}_k u | v), \quad u, v \in V. \tag{2.2}$$

Here, for $v \in L^2(\Omega)$, $f_v(u) := (v|u)_{L^2}$ defines an element $f_v \in V'$ and $v \mapsto f_v : L^2(\Omega) \to V'$ defines a continuous injection. In the following, we identify v with f_v. We then define the operator A_k as

$$D(A_k) := \{u \in V : \exists f \in L^2(\Omega), a_k(u,\phi) = (f|\phi)\ \forall \phi \in V\}, \tag{2.3}$$

$$A_k u := f. \tag{2.4}$$

It is well known that A_k generates an analytic semigroup on $L^2(\Omega)$ which is positivity preserving. Furthermore, this semigroup extends to a C_0-semigroup of contractions on $L^p(\Omega)$ for all $1 < p < \infty$; see, [22, Thm. 4.9]. The realization of its generator in L^p is denoted by A_k^p.

3. Main result. We start this section by giving precise assumptions on the coefficients and functions being involved in problem (1.1). In order to do so, let $0 \le T_0 < T_1$ and set $J :=]T_0, T_1[$. For $k \in \{1, \dots, m\}$ let $\mu_k \in L^\infty(\Omega, M_{n\times n})$ and assume that (2.1) is satisfied.

Moreover, let for every $k \in \{1, \dots, m\}$ the functions $b_k \in L^\infty(\Omega; \mathbb{R})$ be bounded from below by some positive constant.

We assume the following for all $k \in \{1, \dots, m\}$.

(Op) There exists $p > \frac{n}{2}$ such that each $\mathcal{A}_k - Id$ is a topological isomorphism from $W^{1,2p}_{\Gamma_N}(\Omega)$ onto $W^{-1,2p}_{\Gamma_N}(\Omega)$. For all what follows we fix a number $r > \frac{4p}{2p-n}$.

(Su) There exists $f_k \in C^2(\mathbb{R})$, positive, with strictly positive derivative, such that F_k is the superposition operator induced by f_k; i.e., $F_k(v)(x) = (f_k \circ v)(x) = f_k(v(x))$, $x \in \Omega$.

(Ga) The mapping $G_k : \left(W^{1,2p}(\Omega)\right)^m \to W^{1,2p}(\Omega)$ is locally Lipschitz.

(Gb) For any ball in $\left(W^{1,2p}(\Omega)\right)^m$ there exists $\delta > 0$ such that $G_k(u) \ge \delta$ for all u from this ball.

(Ra) The function $R_k : J \times \left(W^{1,2p}(\Omega)\right)^m \to L^p(\Omega)$ is of Carathéodory type, i.e., $R_k(\cdot, u)$ is measurable for all $u \in \left(W^{1,2p}(\Omega)\right)^m$ and $R_k(t, \cdot)$ is continuous for a.a. $t \in J$.

(Rb) $R_k(\cdot, 0) \in L^r(J, L^p(\Omega))$ and for $\beta > 0$ there exists $g_\beta \in L^r(J)$ such that

$$\|R_k(t,u) - R_k(t,\tilde{u})\|_{L^p} \le g_\beta(t)\|u - \tilde{u}\|_{W^{1,2p}}, \quad t \in J,$$

provided $\max(\|u\|_{W^{1,2p}}, \|\tilde{u}\|_{W^{1,2p}}) \le \beta$.

(BC) $\phi_k \in C(\overline{J}; W^{1,2p}(\Omega)) \cap W^{1,r}(J; L^p(\Omega))$ and $A_k\phi_k(t) = 0$ for all $t \in J$.

(IC) $v_{0k} - \phi_k(T_0) \in (L^p(\Omega), D(A_k^p))_{1-\frac{1}{r},r}$.

The assumptions imply that the system (1.1) may be (formally) rewritten as a quasilinear system of the form

$$\begin{aligned} w_k' - H_k(t,w)A_k w_k &= T_k(t,w), \quad k = 1, \dots, m, \\ w(T_0) &= v_0 - \phi(T_0), \end{aligned} \tag{3.1}$$

where

(3.2)

$$T_k(t,w) := (b_k f_k'(w_k+\phi_k(t)))^{-1}[\nabla G_k(w+\phi(t))\cdot[\mu_k \nabla(w_k+\phi_k(t))]] + Q_k(t,w) - \frac{\partial \phi_k}{\partial t}(t)$$

with

$$H_k(t,z) := \frac{G_k(z+\phi(t))}{b_k\ f_k'(z_k+\phi_k(t))}, \quad t \in J,\ z \in \left(W^{1,2p}(\Omega)\right)^m, \tag{3.3}$$

$$Q_k(t,z) := \frac{R_k(t,z+\phi(t))}{b_k\ f_k'(z_k+\phi_k(t))}, \quad t \in J,\ z \in \left(W^{1,2p}(\Omega)\right)^m. \tag{3.4}$$

We are now in the position to state the main result of this paper.

THEOREM 3.1. *Let $1 < r, p < \infty$ such that $r > \frac{4p}{2p-n}$, where $n \in \{2, 3\}$. Assume that the assumptions* (Op), (Su), (Ga), (Gb), (Ra), (Rb), (BC), *and* (IC) *are satisfied. Then there exists a unique local solution $w = (w_1, \dots, w_m)$ for* (3.1) *on an interval $I =]T_0, T[$ satisfying*

$$w_k \in W^{1,r}(I; L^p(\Omega)) \cap L^r(I; D(A_k)), \quad k \in \{1, \dots, m\}. \tag{3.5}$$

COROLLARY 3.2. *Each w_k is Hölder-continuous simultaneously in space and time.*

Some remarks at this point are in order.

Remark 3.3.

(a) We refer the reader to section 4 for precise geometric and smoothness conditions implying the validity of assumption (Op).

(b) Besides the exponential, a typical example for a function f satisfying assumption (Su) is the Fermi–Dirac distribution function

$$f(t) := \frac{2}{\sqrt{\pi}} \int_0^\infty \frac{\sqrt{s}}{1 + e^{s-t}} \, \mathrm{d}s.$$

(c) Suppose that v_k coincides on Γ_D with a function $\phi \in C^1(J, W^{1,2p}(\Omega))$. Then there exists ϕ_k satisfying assumption (BC).

(d) Note that condition (BC) implies $\nu \cdot \mu_k \nabla \phi_k = 0$ on Γ_N. This, together with the property (3.5), yields the Neumann boundary condition for v_k on Γ_N; see [19, section II.2], [9, section 1.2].

4. Examples. Consider Ω and Γ_N, the subset of $\partial\Omega$ on which the Neumann boundary condition is prescribed. In this section we describe geometric configurations and coefficient functions for which Theorem 3.1 holds true. Furthermore, we present concrete examples of mappings G_k and reaction terms R_k fitting in our framework.

4.1. Geometric configurations and coefficients. In this subsection we will give a list of examples in order to show that (Op) is not an unjustified ad hoc assumption but fulfilled in many cases, even if the geometry is nonsmooth, the boundary conditions are mixed, and the coefficients may jump. Notice that for—realistic—nonsmooth situations a (strict) upper bound for the integrability index for the gradient of the solution is (at most) 4. That means in detail that, this limitation may be caused only by nonsmoothness of the domain (see [28, Thm. A]), only by the occurrence of mixed boundary conditions (despite smooth data, see [37]), or only by nonsmooth coefficients (see [33] or [14]).

We start with a result, due to Gröger [24], which completely covers the two-dimensional case.

PROPOSITION 4.1. *Assume that $\Omega \cup \Gamma_N$ is regular in the sense of Gröger and that the coefficient function satisfies the assumptions made in section* 2. *Then there exists $q > 2$ such that $\mathcal{A}_k - Id$ is a topological isomorphism from $W^{1,q}_{\Gamma_N}(\Omega)$ onto $W^{-1,q}_{\Gamma_N}(\Omega)$.*

In what follows we present three-dimensional settings, here always supposing that the domain Ω is a domain with Lipschitz boundary. We begin with the case where the coefficient function is continuous at least in a neighborhood of the boundary.

PROPOSITION 4.2. *Assume that $\Gamma_N = \emptyset$ or $\Gamma_N = \partial\Omega$ (pure Dirichlet or pure Neumann case). $\Omega_\circ \subset \Omega$ is another domain which is C^1 and which does not touch the boundary of Ω. $\mu_k|_{\Omega_\circ} \in BUC(\Omega_\circ)$ and $\mu_k|_{\Omega \setminus \bar{\Omega}_\circ} \in BUC(\Omega \setminus \bar{\Omega}_\circ)$. Then (Op) holds.*

Remark 4.3. The proposition is one of the main results from [15] and rests heavily on regularity results for the Dirichlet/Neumann Laplacian (see [28], [42]) and nontrivial estimates around points from $\partial\Omega_\circ$. In particular, the reader should carefully notice that the C^1 property of $\Omega_\circ$ is not dispensable without completely losing the result; see Elschner's counterexample in [15], see also [14].

The next proposition describes two cases with pure Dirichlet conditions ($\Gamma_N = \emptyset$), where the discontinuities of the coefficient function are located also near the boundary.

PROPOSITION 4.4. *There exists $q > 3$ such that $\mathcal{A}_k - Id$ is a topological isomorphism from $W^{1,q}_{\Gamma_N}(\Omega)$ onto $W^{-1,q}_{\Gamma_N}(\Omega)$ if one of the following conditions is satisfied.*

(i) *Ω is a polyhedron. There are hyperplanes $\mathcal{H}_1, \dots, \mathcal{H}_n$ in $\mathbb{R}^3$ which meet at most in a vertex of the polyhedron such that the coefficient function μ_k is constantly a real, symmetric, positive definite 3×3 matrix on each of the connected components of $\Omega \setminus \cup_{l=1}^n \mathcal{H}_l$. Moreover, for every edge on the boundary, induced by a hyperplane $\mathcal{H}_l$, the angles between the outer boundary plane and the hyperplane $\mathcal{H}_l$ do not exceed π.*

(ii) *$\Omega_\circ \subset \Omega$ is a Lipschitz domain such that $\partial\Omega_\circ \cap \Omega$ is a C^1 surface. Moreover, $\partial\Omega$ and $\partial\Omega_\circ$ meet suitably, this means that for any point x from the boundary of $\partial\Omega \cap \partial\Omega_\circ$ whithin $\partial\Omega$ there is an open neighborhood $\mathcal{U}_x$ of x in $\mathbb{R}^3$ and a C^1 diffeomorphism Φ_x from $\mathcal{U}_x$ onto an open subset of $\mathbb{R}^3$ such that*
- *$\Phi_x(\mathcal{U}_x \cap \Omega)$ equals a convex polyhedron $\mathcal{K}_x$,*
- *$\Phi_x(\mathcal{U}_x \cap \Omega \cap \partial\Omega_\circ) = \mathcal{K}_x \cap \mathcal{H}_x$, where $\mathcal{H}_x$ is a plane which contains $\Phi_x(x)$ and an inner point of $\mathcal{K}_x$. $\mu_k|_{\Omega_\circ} \in BUC(\Omega_\circ)$ and $\mu_k|_{\Omega\setminus\bar{\Omega}_\circ} \in BUC(\Omega \setminus \bar{\Omega}_\circ)$.*

The constellation (i) is treated in [14], the second in [15].

The last cases we present really affect the case of mixed boundary conditions.

PROPOSITION 4.5. *Assumption* (Op) *is also satisfied in any of the following two cases.*

(i) *Ω is a convex polyhedron, $\overline{\Gamma_N} \cap (\partial\Omega\setminus\Gamma_N)$ is a finite union of line segments, $\mu_k \equiv 1$.*

(ii) *Ω is a three-dimensional prismatic domain with triangular basis. Γ is half of one of its upright sides. Ξ is a plane that intersects the (relative) boundary of Γ within $\partial\Omega$ in only finitely many points. The coefficient function is constant on both components of $\Omega \setminus \Xi$.*

Remark 4.6. The assertion for (i) is shown in [11] (see Corollary 3.12), while the proof for (ii) is given in [25]. Notice that (ii) is by no means artificial: $\Omega \cup \Gamma$ may (alternatively) be taken as Gröger's third local model set in the description of geometric settings including mixed boundary conditions; see section 2 or [24] for further details. Additionally, it is the first three-dimensional constellation treated in the literature—in view of (Op)—which includes mixed boundary conditions and a discontinuous coefficient function, which is necessary, e.g., in the modeling of heterogeneous semiconductors [36]. (See [25] for other settings and further details.)

Remark 4.7. The operator $\mathcal{A}_k - Id$ also fulfills (Op) if the following condition is satisfied: there is a covering of $\bar{\Omega}$ by open sets $\mathcal{U}_1, \dots, \mathcal{U}_l$ such that for every $j \in \{1, \dots, l\}$ the setting $\Omega_j := \Omega \cap \mathcal{U}_j$, $\Gamma_j := \Gamma_N \cap \mathcal{U}_j$, and the restriction of μ_k to $\Omega \cap \mathcal{U}_j$ satisfy one of the conditions of the foregoing propositions; compare [24, Lemma 2]. Additionally, one can show by perturbation arguments (as, e.g., carried out in [15]) that in many cases admissible constellations are preserved under C^1 diffeomorphisms.

In the following we illustrate two admissible three-dimensional settings. On the left-hand side of Figure 4.1 one assumes Neumann conditions on the top of the upper cuboid, otherwise Dirichlet conditions. On the right-hand side of the figure, the

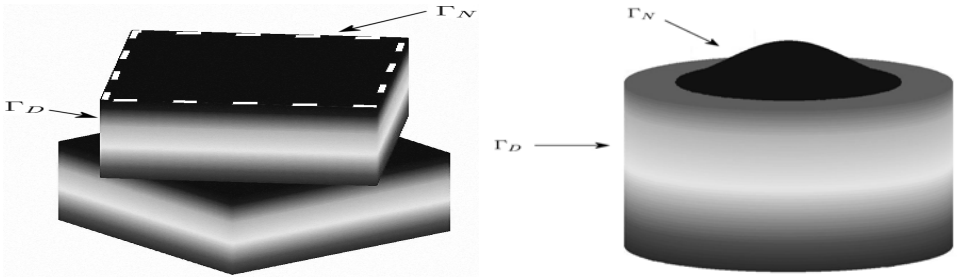


FIG. 4.1.

boundary of the cylinder is subject to Dirchlet conditions exept for the upper "hat," where Neumann conditions are prescribed.

4.2. Examples for the nonlinearities G_k and the reaction terms R_k. Next we give two examples for the operators G_k.

Example 4.8. Let $g_k : \mathbb{R}^m \mapsto]0, \infty[$ be a twice continuously differentiable function and define $G_k(z)(x) = g_k(z(x))$ if $z \in \left(W^{1,2p}\right)^m$ and $x \in \Omega$.

In many applications g_k depends only on one variable and is a multiple of the exponential function.

As the second example we present a nonlocal operator arising in the diffusion of bacteria; see [7], [8], and the references therein.

Example 4.9. Let η be a continuously differentiable function on $\mathbb{R}$ which is bounded from above and below by positive constants. Assume $\varphi \in L^2(\Omega)$ and define

$$G_k(z) := \eta\left(\int_\Omega z_k \varphi dx\right), \quad z = (z_1, \dots, z_m) \in \left(W^{1,2p}\right)^m.$$

Now we give two examples for mappings R_k.

Example 4.10. Assume that $[T_0, T_1[= \cup_{l=1}^{j} [t_l, t_{l+1}[$ is a (disjoint) decomposition of $[T_0, T_1[$ and let for $l \in \{1, \dots, j\}$

$$S_l : \mathbb{R}^m \times \mathbb{R}^{nm} \mapsto \mathbb{R}$$

be a function which satisfies the following condition: for any compact set $K \subset \mathbb{R}^m$ there is a constant L_K such that for any $a, \tilde{a} \in K$ and $b, \tilde{b} \in \mathbb{R}^{nm}$ the inequality

$$\begin{aligned} |S_l(a,b) - S_l(\tilde{a},\tilde{b})| \le\, & L_K |a - \tilde{a}|_{\mathbb{R}^m} \left(|b|^2_{\mathbb{R}^{nm}} + |\tilde{b}|^2_{\mathbb{R}^{nm}}\right) \\ & + L_K |b - \tilde{b}|_{\mathbb{R}^{nm}} \left(|b|_{\mathbb{R}^{nm}} + |\tilde{b}|_{\mathbb{R}^{nm}}\right) \end{aligned}$$

holds. We define a mapping $S : [T_0, T_1[\times \mathbb{R}^m \times \mathbb{R}^{nm} \mapsto \mathbb{R}$ by setting

$$S(t, a, b) := S_l(a, b) \quad \text{if} \quad t \in [t_l, t_{l+1}[.$$

The function S defines a mapping R in the following way: if z is the restriction of an $\mathbb{R}^m$-valued, continuously differentiable function on $\mathbb{R}^n$ to Ω, then we put

$$R(t, z, \nabla z)(x) = S(t, z(x), (\nabla z)(x)) \quad \text{for } x \in \Omega \tag{4.1}$$

and afterwards extend R by continuity to the whole set $[T_0, T_1[\times \left(W^{1,2p}(\Omega)\right)^m$.

Example 4.11 (electrical and heat conduction). Assume that $\sigma : \mathbb{R} \mapsto]0, \infty[$ is a continuously differentiable function. Further, let $\mathcal{S} : W^{1,2p} \mapsto W^{1,2p}$ be the mapping which assigns to $z \in W^{1,2p}$ the solution φ of the elliptic problem

$$-\nabla \cdot \sigma(z)\nabla\varphi = 0$$

(including boundary conditions, see [38], [3], [6]). If one defines

$$R(z) = \sigma(z)|\nabla(\mathcal{S}(z))|^2,$$

then, under a reasonable supposition on the domain and the boundary conditions, the mapping R satisfies assumption (Ra).

5. Tools for the proof of Theorem 3.1. Let $1 < s < \infty$ and let B be a densely defined sectorial operator in a Banach space X. Let again $J =]T_0, T_1[$ for some $T_0, T_1 > 0$. We say that the linear evolution equation

$$\begin{aligned} u' + Bu &= f, \\ u(T_0) &= 0 \end{aligned} \tag{5.1}$$

admits maximal L^s-regularity on J if for any $f \in L^s(J; X)$ there exists a unique function $u \in W^{1,s}(J; X) \cap L^s(J; D(B))$ satisfying (5.1) in the L^s-sense. In that case, we write $B \in MR(s, X)$. Observe that

$$W^{1,s}(J; X) \cap L^s(J; D(B)) \hookrightarrow C(\overline{J}; X_s), \tag{5.2}$$

where X_s is the real interpolation space $(X, D(B))_{1-\frac{1}{s},s}$. Consider now the quasilinear problem

$$\begin{aligned} u'(t) + \mathcal{B}(t, u(t))u(t) &= F(t, u(t)), \quad t \in J, \\ u(T_0) &= u_0. \end{aligned} \tag{5.3}$$

Here $u_0 \in X_s$, $B := \mathcal{B}(T_0, u_0)$, and $\mathcal{B} : J \times X_s \to \mathcal{L}(D(B); X)$ is continuous. $F : J \times X_s \to X$ is a Carathéodory map. We assume the following Lipschitz conditions on $\mathcal{B}$ and F.

(B) For each $R > 0$ there exists a constant $C_R > 0$, such that

$$\begin{aligned} &\|\mathcal{B}(t, u)v - \mathcal{B}(t, \tilde{u})v\|_X \\ &\quad \le C_R \, ||u - \tilde{u}||_{X_s} ||v||_{D(B)}, \ t \in J, u, \tilde{u} \in X_s, ||u||_s, \ ||\tilde{u}||_s \le R, \ v \in D(B). \end{aligned} \tag{5.4}$$

(F) $F(\cdot, 0) \in L^s(J; X)$ and for each $R > 0$ there is a function $\eta_R \in L^s(J)$ such that

$$\|F(t, u) - F(t, \tilde{u})\|_X \le \eta_R(t) \, \|u - \tilde{u}\|_s \text{ a.a. } t \in J, \ u, \tilde{u} \in X_s, \ ||u||_s, \|\tilde{u}\|_s \le R. \tag{5.5}$$

Then the following existence and uniqueness result due to Clément and Li [10] and Prüss [34] holds true.

THEOREM 5.1. *Assume that* (B) *and* (F) *are satisfied and that* $B := \mathcal{B}(T_0, u_0)$ *has the property of maximal* L^s*-regularity. Then there exists* $T \in]T_0, T_1[$ *such that* (5.3) *admits a unique solution* u *on* $I :=]T_0, T[$ *satisfying*

$$u \in W^{1,s}(I; X) \cap \ L^s(I; D(B)).$$

In order to verify the crucial condition that $B = \mathcal{B}(T_0, u_0)$ has maximal L^s-regularity in our situation, we need the following results on traces, heat kernels, their multiplicative perturbations, and maximal L^s-regularity.

Consider a closed subspace V of $H^1(\Omega)$ which includes $H^1_0(\Omega)$. Let $\varrho \in L^\infty(\Omega, M_{n\times n})$ and assume that it is elliptic in the sense of (2.1). Define a bilinear form $a : V \times V \to \mathbb{R}$ on V by

$$a(u,v) = -\int_\Omega \varrho \nabla u \cdot \nabla u \, \mathrm{d}x, \quad u, v \in V.$$

Let A be the operator associated with a in $L^2(\Omega)$ and let $(e^{tA})_{t\geq 0}$ be the semigroup on $L^2(\Omega)$ generated by A. The following result gives sufficient conditions on the subspace V such that $(e^{tA})_{t\geq 0}$ satisfies an upper Gaussian bound. More precisely, the following Proposition holds; see, [4, Ch. 4].

PROPOSITION 5.2. *Assume that V is a closed subspace of $H^1(\Omega)$ satisfying*
(a) $H^1_0(\Omega) \subseteq V$,
(b) *V has the L^1-H^1 extension property,*
(c) $u \in V$ *implies* $|u|,\ \inf(|u|, 1) \in V$,
(d) $u \in V,\ v \in H^1(\Omega),\ |v| \leq u$ *implies* $v \in V$.
Then e^{tA} satisfies an upper Gaussian estimate, i.e.,

$$(e^{tA} f)(x) = \int_\Omega K_t(x,y) f(y) \mathrm{d}y \quad \text{a.a. } x \in \Omega,\ f \in L^2(\Omega)$$

for some measurable function $K_t : \Omega \times \Omega \to \mathbb{R}_+$ and there exists constants $M, a > 0$ and $\omega \in \mathbb{R}$ such that

$$0 \leq K_t(x,y) \leq \frac{M}{t^{\frac{n}{2}}} \, e^{\frac{-a|x-y|^2}{t}} \, e^{\omega t}, \quad t > 0,\ \text{a.a. } x, y \in \Omega. \tag{5.6}$$

LEMMA 5.3. *Let $H^1_{\Gamma_N}(\Omega)$ be defined as above. Then $V := H^1_{\Gamma_N}(\Omega)$ satisfies the assumptions* (a)–(d) *of Proposition* 5.2.

Proof. We will not give a detailed proof—which is given in [4] in the case of domains with Lipschitz boundary. Let us only notice that assertion (a) is obvious and that (b) is assured by the fact that Lipschitz domains are extensions domains for $H^{1,2}$; see [21, Thm. 3.10] or [32, section 1.1.16]. Inspecting the corresponding proofs (which are given via localization, Lipschitz diffeomorphism, and reflection), one recognizes that the extension mapping simultaniously extends continuously to L^1. The assertions (c) and (d) may be proved essentially as in [4] (see Example 4.3). In fact, the continuity results for the mappings $u \to \inf(u, v)$ and $u \to |u|$ given in [31] are stated without any restriction on the regularity of the domain Ω. ∎

Consider the semigroup e^{tA_k} on $L^2(\Omega)$ generated by A_k associated with the form a_k defined in (2.2) with $V = H^1_{\Gamma_N}(\Omega)$. It follows from Proposition 5.2 and Lemma 5.3 that e^{tA_k} is a positive semigroup on $L^2(\Omega)$ satisfying an upper Gaussian bound. Hence, $(e^{tA_k})_{t\geq 0}$ extends to a positive C_0-semigroup of contractions on $L^q(\Omega)$ for all $1 \leq q < \infty$.

PROPOSITION 5.4. *Let $b \in L^\infty(\Omega, \mathbb{R})$ such that $\inf_{x\in\Omega} |b(x)| \geq \delta$ for some $\delta > 0$. Let $1 < s, q < \infty$. Then $bA_k \in MR(s, L^q(\Omega))$ for all $k \in \{1, \dots, m\}$.*

Proof. Let $k \in \{1, \dots, m\}$. By the above remark, e^{tA_k} is a positive contraction semigroup on $L^q(\Omega)$ satisfying an upper Gaussian bound. Hence, the kernel K_t of $e^{t(A_k - \alpha Id)})_{t\geq 0}$ satisfies (5.6) with $\omega = 0$ for suitable $\alpha \in \mathbb{R}$. Moreover, $A_k - \alpha Id$ is self-adjoint in $L^2(\Omega)$. By a result due to Duong and Ouhabaz [13], the semigroup on

$L^2(\Omega)$ generated by $b(A_k - \alpha Id)$ satisfies an upper Gaussian bound with $\omega = 0$ as well. Thus $b(A_k - \alpha Id) \in MR(s, L^q(\Omega))$ by a result of Hieber and Prüss (see [27] or [12, Thm. 4.8]). Finally, $bA_k \in MR(s, L^q(\Omega))$ due to the lower order perturbation result of maximal regularity; see [12, Prop. 4.3]. ∎

PROPOSITION 5.5. *Let $p > \frac{n}{2}$ be the number from Assumption* (Op) *and assume that $\theta \in (\frac{1}{2} + \frac{n}{4p}, 1]$. Then*

$$[L^p, D(A_k^p)]_\theta \hookrightarrow W^{1,2p}_{\Gamma_N}(\Omega).$$

A proof for the three-dimensional case is given in [35]; the two-dimensional case requires only obvious modifications. A complete, but technically more involved, proof for the two-dimensional case is contained in [29], see Thm. 5.2.

COROLLARY 5.6. *Let $r > \frac{4p}{2p-n}$. Then*

$$(L^p, D(A_k^p))_{1-\frac{1}{r},r} \hookrightarrow W^{1,2p}_{\Gamma_N}(\Omega).$$

Proof. Let θ be any number from the interval $]\frac{1}{2} + \frac{n}{4p}, 1 - \frac{1}{r}[$. By interpolation

$$(L^p, D(A_k^p))_{1-\frac{1}{r},r} \hookrightarrow (L^p, D(A_k^p))_{\theta,1} \hookrightarrow [L^p, D(A_k^p)]_\theta.$$

Then the assertion follows from the embedding property of the complex interpolation space into $W^{1,2p}_{\Gamma_N}(\Omega)$ established in Proposition 5.5. ∎

6. Proof of the main result. We first set $X := \big(L^p(\Omega)\big)^m$, $\mathcal{D} := \times_{k=1}^m D(A_k^p)$, and $X_r := (X, \mathcal{D})_{1-\frac{1}{r},r}$ for r as above. By assumption (IC), $w_0 \in X_r$. Further, for every pair $(t, z) \in [T_0, T_1) \times W^{1,2p}(\Omega)^m$ we define the mapping $H(t, z) : X \mapsto X$ via

$$\varphi := (\varphi_1, \dots, \varphi_m) \mapsto \big(H_1(t, z)\varphi_1, \dots, H_m(t, z)\varphi_m\big). \tag{6.1}$$

Since $H_k(t, z) \in L^\infty(\Omega)$ and since H_k possesses a strictly positive lower bound, it follows that

$$D(H_k(t, z)A_k^p) = D(A_k^p).$$

In particular, $D(H_k(T_0, w_0)A_k^p))$ is dense in $L^p(\Omega)$ (see [22, Thms. 4.5 and 4.7]).

Consider the mapping $\mathcal{B} : J \times X_r \to \mathcal{L}(\mathcal{D}; X)$ given by

$$\mathcal{B}(t, z)\varphi := H(t, z)(A_1^p\varphi_1, \dots, A_m^p\varphi_m), \quad \varphi = (\varphi_1, \dots, \varphi_m) \in \mathcal{D}.$$

By Corollary 5.6 and Morrey's theorem we have

$$X_r \hookrightarrow \big(W^{1,2p}_{\Gamma_N}(\Omega)\big)^m \hookrightarrow \big(C^\alpha(\Omega)\big)^m$$

for some $\alpha > 0$. Thus, the assumed properties on F_k, G_k, and ϕ_k imply that

$$\mathcal{B} : J \times X_r \to \mathcal{L}(\mathcal{D}; X)$$

is continuous. Moreover, for $\beta > 0$ there exists $C_\beta > 0$ such that

$$\|H(t, z) - H(t, \tilde{z})\|_\infty \le C_\beta \|z - \tilde{z}\|_{W^{1,2p}}$$

provided $t \in J$ and $\|z\|_{X_r}$ and $\|\tilde{z}\|_{X_r} \le \beta$. Hence, (5.4) from assertion (B) is fulfilled.

Furthermore, (5.5) from assertion (F) holds due to the assumed properties of F_k, G_k, ϕ, R_k and Corollary 5.6. It remains to verify the key condition of Theorem 5.1, namely that $B := \mathcal{B}(T_0, w_0)$ has the property of maximal regularity. To this end, recall that $H(T_0, w_0) \in \big(L^\infty(\Omega)\big)^m$ with a strictly positive lower bound in each component. Thus, $B \in MR(r, X)$ by Proposition 5.4. Finally, an application of Theorem 5.1 ends the proof of Theorem 3.1.

It remains to show that if w is a solution of (3.1), then $v := w + \phi$ provides a solution of (1.1). This will be done in the appendix. ∎

We now give a proof of Corollary 3.2; in fact we prove the following sharper result.

LEMMA 6.1. *There exists $\beta > 0$ such that each component w_k of the solution w of (3.1) belongs to the space $C^\beta(]T_0, T[; W^{1,2p}_\Gamma(\Omega)) \hookrightarrow C^\beta(]T_0, T[; C^\alpha(\Omega))$.*

Proof. We write for short $D_k = D(A_k)$ and $I =]T_0, T[$. Then

$$W^{1,r}(I; L^p) \cap L^r(I; D_k) \hookrightarrow C(\bar{I}; (L^p, D_k)_{1-\frac{1}{r},r}) \hookrightarrow C(\bar{I}; [L^p, D_k]_\theta)$$

if $\theta \in]0, 1 - \frac{1}{r}[$.

Moreover, we have the embedding

$$W^{1,r}(I; L^p) \hookrightarrow C^\delta(I; L^p) \quad \text{with} \quad \delta = 1 - \frac{1}{r}.$$

Fix $\theta \in]\frac{1}{2} + \frac{n}{4p}, 1 - \frac{1}{r}[$ and let $\lambda \in (0, 1)$ be given such that

$$\theta\lambda > \frac{1}{2} + \frac{n}{4p}.$$

In view of Proposition 5.5 and the reiteration theorem for complex interpolation (see [40, Ch. 1.9.3]) we obtain

$$\begin{aligned}
\frac{\|w_k(t) - w_k(s)\|_{W^{1,2p}}}{|t-s|^{\delta(1-\lambda)}} &\le c\, \frac{\|w_k(t) - w_k(s)\|_{[L^p, D_k]_{\theta\lambda}}}{|t-s|^{\delta(1-\lambda)}} \sim \frac{\|w_k(t) - w_k(s)\|_{\left[L^p, [L^p, D_k]_\theta\right]_\lambda}}{|t-s|^{\delta(1-\lambda)}} \\
&\le \hat{c}\, \frac{\|w_k(t) - w_k(s)\|^{1-\lambda}_{L^p}}{|t-s|^{\delta(1-\lambda)}} \|w_k(t) - w_k(s)\|^\lambda_{[L^p, D_k]_\theta} \\
&= \hat{c} \left(\frac{\|w_k(t) - w_k(s)\|_{L^p}}{|t-s|^\delta}\right)^{1-\lambda} \left(2 \sup_{s \in \bar{I}} \|w_k(s)\|_{[L^p, D_k]_\theta}\right)^\lambda. \qquad ∎
\end{aligned}$$

7. Appendix. It remains to show that if w is a solution of (3.1), then $v := w + \phi$ provides a solution of (1.1). One easily recognizes that all the manipulations which transfrom (1.1) into (3.1) are straightforward to justify within the distributional calculus, except one. Therefore, we will give a strict justification of this point in the following lemma. Throughout this appendix, $f : \mathbb{R} \mapsto \mathbb{R}$ is always assumed to be twice continuously differentiable.

LEMMA 7.1. *Assume $p, r \in]1, \infty[$ and $v \in W^{1,r}(]T_0, T[; L^p) \cap C([T_0, T]; C(\bar{\Omega}))$. Then the function $]T_0, T[\ni t \mapsto f(v(t))$ belongs to $W^{1,r}(]T_0, T[; L^p)$ and its distributional derivative is the function $]T_0, T[\ni t \mapsto f'(v(t))v'(t) \in L^r(]T_0, T[; L^p)$.*

Remark 7.2. We denote by $C^1(]T_0, T[; L^p)$ the space of all L^p-valued, continuously differentiable functions on $]T_0, T[$ with bounded derivatives on $]T_0, T[$.

In order to give a proof of Lemma 7.1 we use the following result.

LEMMA 7.3. *Let $[T_0, T] \ni t \mapsto \psi(t, \cdot)$ be a mapping belonging to $C([T_0, T]; C(\bar{\Omega})) \cap C^1(]T_0, T[; L^p)$. Then the mapping*

$$]T_0, T[\ni t \mapsto f(\psi(t, \cdot)) \tag{7.1}$$

takes its values in $C(\bar{\Omega}) \hookrightarrow L^p$. It is continuously differentiable when regarded as L^p-valued and its derivative in a point $s \in]T_0, T[$ is equal to the L^p-function $f'(\psi(s, \cdot))\psi'(s)$.

Proof. The first assertion is obvious. Concerning the second one, the set $\{\psi(t,x)/ x \in \Omega, t \in [T_0,T]\}$ is bounded. Since f is twice continuously differentiable, for $s,t \in]T_0,T[$ and $x \in \Omega$ one may apply Taylor's formulae:

$$\frac{f(\psi(t,x)) - f(\psi(s,x))}{t-s} = f'(\psi(s,x))\frac{[\psi(t,x)-\psi(s,x)]}{t-s} \tag{7.2}$$

$$+\int_0^1 (1-\tau)f''((1-\tau)\psi(t,x)+\tau\psi(s,x))\,d\tau \quad \frac{[\psi(t,x)-\psi(s,x)]^2}{t-s}. \tag{7.3}$$

The family $\{f'(\psi(s,\cdot))\frac{[\psi(t,\cdot)-\psi(s,\cdot)]}{t-s}\}_t$ converges by the supposition on the differentiablity of the mapping $t \mapsto \psi(t,\cdot)$ in L^p to $f'(\psi(s,\cdot))\psi'(s)$ if t approaches s . It remains to show that the expression in (7.3) approaches zero in L^p. This follows easily from the uniform boundedness of the values $f''((1-\tau)\psi(t,x)+\tau\psi(s,x))$, the boundedness of $\{\frac{[\psi(t,\cdot)-\psi(s,\cdot)]}{t-s}\}_t$ in L^p, and the convergence of $[\psi(t,\cdot)-\psi(s,\cdot)]$ to zero in $C(\bar\Omega)$ for t approaching s. The continuity of the derivative follows from the continuity of ψ' and the continuity of the function $t \mapsto f'(\psi(t,\cdot))$ in $C(\bar\Omega)$. ∎

LEMMA 7.4. *Let $v \in W^{1,r}(]T_0,T[;L^p) \cap C([T_0,T];C(\bar\Omega))$. Then there is a sequence $\{\psi_l\}_l$ in $C([T_0,T];C(\bar\Omega)) \cap C^1(]T_0,T[;L^p(\Omega))$ such that $\psi_l \mapsto v$ in $C([T_0,T];C(\bar\Omega))$ and $\psi_l' \mapsto v'$ in $L^r(]T_0,T[;L^p)$.*

Proof. Let us define a continuous extension $\tilde v$ to all of $\mathbb{R}$ which additionally has compact support as follows: we put

$$\hat v(t) := \begin{cases} v(T_0+(T_0-t)) & \text{if} \quad t \in]T_0-(T-T_0),T_0[, \\ v(t) & \text{if} \quad t \in [T_0,T], \\ v(T-(t-T) & \text{if} \quad t \in]T,T+(T-T_0)[\end{cases} \tag{7.4}$$

(reflection at T_0, T, respectively). Afterwards we multiply $\hat v$ by a real-valued, continuously differentiable function which is identical 1 on $[T_0,T]$ and which has its support in $]T_0-(T-T_0)/2, T+(T-T_0)/2[$. We define this product as $\tilde v$ and identify $\tilde v$ with its extension by zero to whole $\mathbb{R}$. Oviously, $\tilde v|_{[T_0,T]} = v$; further one verifies the property $\tilde v \in W^{1,r}(\mathbb{R};L^p) \cap C(\mathbb{R};C(\bar\Omega))$. Let ϑ be the usual mollifier function

$$\vartheta(s) = \begin{cases} \frac{1}{\int e^{-\frac{1}{1-s^2}}ds} e^{-\frac{1}{1-s^2}} & \text{if} \quad |s|<1, \\ 0 & \text{else on } \mathbb{R} \end{cases}$$

and $\vartheta_l(s) := l\vartheta(l\,s)$. Now we put

$$\psi_l(t) := \begin{cases} \int_{T_0}^t (\tilde v' * \vartheta_l)(s)\,ds + (\tilde v * \vartheta_l)(T_0) & \text{if} \quad t \ge T_0, \\ -\int_t^{T_0} (\tilde v' * \vartheta_l)(s)\,ds + (\tilde v * \vartheta_l)(T_0) & \text{if} \quad t < T_0. \end{cases} \tag{7.5}$$

Then ψ_l is nothing else but $\tilde v * \vartheta_l$. This yields $\psi_l \mapsto v$ in $C([T_0,T];C(\bar\Omega))$. On the other hand, (7.5) immediately gives $\psi_l' = \tilde v' * \vartheta_l$. This means that $\psi_l' \mapsto \tilde v'$ in $L^r(\mathbb{R};L^p)$, which implies $\psi_l'|_{]T_0,T[} \mapsto v'$ in $L^r(]T_0,T[;L^p)$. ∎

We now turn to the proof of Lemma 7.1. Let $\{\psi_l\}_l$ be the sequence from the previous lemma and $\varphi \in C_0^\infty(]T_0,T[)$. Then, considering the function $]T_0,T[\ni t \mapsto f(v(t))$ as an L^p-valued distribution, one gets by the definition of the weak derivative

$$\big(f(v)\big)'(\varphi) = -f(v)(\varphi') = -\int_{T_0}^T f(v(s))\varphi'(s)\,ds = -\int_{T_0}^T \lim_{l\mapsto\infty} f(\psi_l(s))\varphi'(s)\,ds$$

$$= \lim_{l\mapsto\infty} -\int_{T_0}^T f(\psi_l(s))\varphi'(s)\,ds.$$

By Lemma 7.3, each $f(\psi_l)$ even has a strong (time) derivative which equals $f'(\psi_l)\psi_l'$. From this and integrating by parts one gets

$$-\int_{T_0}^{T} f(\psi_l(s))\varphi'(s)\,ds = \int_{T_0}^{T} f'(\psi_l(s))\psi_l'(s)\varphi(s)ds.$$

By construction, $\psi_l \mapsto v$ in $C([T_0,T];C(\bar{\Omega}))$, $\psi_l' \mapsto v'$ in $L^r(]T_0,T[;L^p)$, which implies $f'(\psi_l(\cdot))\psi_l'\varphi \mapsto f'(v(\cdot))v'\varphi$ in $L^r(]T_0,T[;L^p)$. But the integral is a continuous mapping from $L^r(]T_0,T[;L^p)$ into L^p; this finally gives

$$\begin{aligned}\int_{T_0}^{T} f'(v(s))v'(s)\varphi(s)\,ds &= \int_{T_0}^{T} \lim_{l\mapsto\infty} f'(\psi_l(s))\psi_l'(s)\varphi(s)ds \\ &= \lim_{l\mapsto\infty}\int_{T_0}^{T} f'(\psi_l(s))\psi_l'(s)\varphi(s)ds \\ &= \lim_{l\mapsto\infty} -\int_{T_0}^{T} f(\psi_l(s))\varphi'(s)\,ds = \big(f(v)\big)'(\varphi).\end{aligned}$$

Thus, Lemma 7.1 is proved.

Acknowlegments. The second author would like to thank K. Gröger and J. Griepentrog for stimulating discussions on the subject of the paper.

REFERENCES

[1] H. Amann, *Linear and Quasilinear Parabolic Problems*, Birkhäuser, Basel, Boston, Berlin, 1995.

[2] H. Amann, *Nonhomogeneous linear and quasilinear elliptic and parabolic boundary value problems*, in Function Spaces, Differential Operators and Nonlinear Analysis, Teubner-Texte Math. 133, H.-J. Schmeisser et al., eds., Stuttgart, 1993, pp. 9–126.

[3] K. Andrews, P. Shi, M. Shillor, and S. Wright, *Thermoelastic contact with Barber's heat exchange condition*, Appl. Math. Optim., 28 (1993), pp. 11–48.

[4] W. Arendt and A. F. M. ter Elst, *Gaussian estimates for second order elliptic operators with boundary conditions*, J. Operator Theory, 38 (1997), pp. 87–130.

[5] M. Chaplain and G. Lolas, *Mathematical modelling of cancer cell invasion of tissue: The role of the urokinase plasminogen activation system*, Math. Models Methods Appl. Sci., 15 (2005), pp. 1685–1734.

[6] S. N. Antontsev and M. Chipot, *The thermistor problem: Existence, smoothness, uniqueness, blowup*, SIAM J. Math. Anal., 25 (1994), pp. 1128–1156.

[7] N. H. Chang and M. Chipot, *On some mixed boundary value problems with nonlocal diffusion*, Adv. Math. Sci. Appl., 14 (2004), pp. 1–24.

[8] M. Chipot and B. Lovat, *On the asymptotic behavior of some nonlocal problems*, Positivity, 3 (1999), pp. 65–81.

[9] P. G. Ciarlet, *The finite element method for elliptic problems*, Stud. Math. Appl., North–Holland, Amsterdam, New York, Oxford, 1979.

[10] P. Clément and S. Li, *Abstract parabolic quasilinear equations and application to a groundwater flow problem*, Adv. Math. Sci. Appl., 3 (1994), pp. 17–32.

[11] M. Dauge, *Neumann and mixed problems on curvilinear polyhedra*, Integral Equations Operator Theory, 15 (1992), pp. 227–261.

[12] R. Denk, M. Hieber, and J. Prüss, *$\mathcal{R}$-boundedness, Fourier multipliers and problems of elliptic and parabolic type*, Mem. Amer. Math. Soc., 166 (2003), pp. viii+114.

[13] X. T. Duong and E. M. Ouhabaz, *Complex multiplicative perturbations of elliptic operators: Heat kernel bounds and holomorphic functional calculus*, Differential Integral Equations, 12 (1999), pp. 395–418.

[14] J. Elschner, H. Kaiser, J. Rehberg, and G. Schmidt, *$W^{1,q}$ regularity results for elliptic transmission problems on heterogeneous polyhedra*, Math. Models Methods Appl. Sci., 17 (2007), pp. 593–615.

[15] J. Elschner, J. Rehberg, and G. Schmidt, *Optimal regularity for elliptic transmission problems including C^1 interfaces*, Interfaces Free Bound., 9 (2007), pp. 233–252.
[16] J. Escher, *On quasilinear fully parabolic boundary value problems*, Differential Integral Equations, 7 (1994), pp. 1325–1343.
[17] M. Fila and H. Matano, *Blow up in nonlinear heat equations from the dynamical systems point of view*, in Handbook of Dynamical Systems, Vol. 2, B. Fiedler, ed., North–Holland, Amsterdam, 2002.
[18] J. Fuhrmann and H. Langmach, *Stability and existence of solutions of time-implicit finite volume schemes for viscous nonlinear conservation laws*, Appl. Numer. Math., 37 (2001), pp. 201–230.
[19] H. Gajewski, K. Gröger, and K. Zacharias, *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen*, Akademie-Verlag, Berlin, 1974.
[20] H. Gajewski and K. Gröger, *Reaction-diffusion processes of electrical charged species*, Math. Nachr., 177 (1996), pp. 109–130.
[21] E. Giusti, *Metodi diretti nel calcolo delle variazioni*, Unione Matematica Italiana, Bologna, 1994.
[22] J. A. Griepentrog, H. C. Kaiser, and J. Rehberg, *Heat kernel and resolvent properties for second order elliptic differential operators with general boundary conditions on L^p*, Adv. Math. Sci. Appl., 11 (2001), pp. 87–112.
[23] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.
[24] K. Gröger, *A $W^{1,p}$-estimate for solutions to mixed boundary value problems for second order elliptic differential equations*, Math. Ann., 283 (1989), pp. 679–687.
[25] R. Haller-Dintelmann, H.-C. Kaiser, and J. Rehberg, *Elliptic Model Problems Including Mixed Boundary Conditions and Material Heterogeneities*, Preprint 1203, WIAS, Berlin 2007.
[26] M. Heinkenschloss and F. Troeltzsch, *Analysis of the Langrange-SQP-Newton method for the control of a phase field equation*, Control Cybernet., 28 (1999), pp. 177–211.
[27] M. Hieber and J. Prüss, *Heat kernels and maximal L^p-L^q estimates for parabolic evolution equations*, Comm. Partial Differential Equations, 22 (1997), pp. 1647–1669.
[28] D. Jerison and C. Kenig, *The inhomogeneous Dirichlet problem in Lipschitz domains*, J. Funct. Anal., 130 (1995), pp. 161–219.
[29] H. C. Kaiser, J. Rehberg, and H. Neidhardt, *Classical solutions of quasilinear parabolic systems on two dimensional domains*, NoDEA Nonlinear Differential Equation Appl., 13 (2006), pp. 287–310.
[30] P. Krejci, E. Rocca, and J. Sprekels, *Nonlocal Temperature-Dependent Phase-Field Models for Non-Isothermal Phase Transitions*, Preprint 1006, WIAS, Berlin, 2007.
[31] M. Marcus and V. Mizel, *Every superposition operator mapping one Sobolev space into another is continuous*, J. Funct. Anal., 33 (1979), pp. 217–229.
[32] V. Maz'ya, *Sobolev Spaces*, Springer, Berlin, Heidelberg, New York, Tokyo, 1985.
[33] N. G. Meyers, *An L^p-estimate for the gradient of solutions of second order elliptic divergence equations*, Ann. Sc. Norm. Super. Pisa Cl. Sci. (4), 17 (1963), pp. 189–206.
[34] J. Prüss, *Maximal regularity for evolution equations in L^p-spaces*, Conf. Semin. Math. Univ. Bari, 285 (2002), pp. 1–39.
[35] J. Rehberg, *Quasilinear parabolic equations in L^p*, in Nonlinear Elliptic and Parabolic Problems. A special tribute to the work of Herbert Amann, Prog. Nonlinear Differential Equations Appl., M. Chipot and J. Escher, eds., 2005, pp. 413–419.
[36] S. Selberherr, *Analysis and Simulation of Semiconductors*, Springer, Wien, 1984
[37] E. Shamir, *Regularization of mixed second-order elliptic problems*, Israel J. Math. 6, (1968), pp. 150–168.
[38] P. Shi, M. Shillor, and X. Xu, *Existence of a solution to the Stefan problem with Joule's heating*, J. Differential Equations, 105 (1993), pp. 239–263.
[39] A. Sommerfeld, *Thermodynamics and statistical mechanics*, Lectures on Theoretical Physics V, Academic Press, New York, 1956.
[40] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, North–Holland, Amsterdam, New York, Oxford, 1978.
[41] A. Unger and F. Troeltzsch, *Fast solutions of optimal control problems in the selective cooling of steel*, ZAMM Z. Angew. Math. Mech., 81 (2001), pp. 447–456.
[42] D. Zanger, *The inhomogeneous Neumann problem in Lipschitz domains*, Comm. Partial Differential Equations, 25 (2000), pp. 1771–1808.

SIAM J. MATH. ANAL.
Vol. 40, No. 1, pp. 306–326

A VARIATIONAL INEQUALITY ARISING FROM EUROPEAN INSTALLMENT CALL OPTIONS PRICING*

FAHUAI YI†, ZHOU YANG†, AND XIAOHUA WANG‡

Abstract. In this paper we consider a parabolic variational inequality arising from European continuous installment call options pricing and prove the existence and uniqueness of the solution to the problem. Moreover, we obtain C^∞ regularity and the bounds of the free boundary, as well as the limit of the free boundary as $\tau = T - t \to +\infty$. Eventually we show its numerical result by the binomial method.

Key words. free boundary, variational inequality, option pricing, European installment call options

AMS subject classification. 35R35

DOI. 10.1137/060670353

1. Introduction.

In this paper we consider a parabolic variational inequality arising from the model of European continuous installment call options pricing. More precisely, we will find $C(S,t)$ satisfying

$$
\begin{cases}
\partial_t C + \frac{\sigma^2}{2} S^2 \partial_{SS} C + (r-q) S \partial_S C - rC = L^* \\
\qquad\qquad \text{if } C > 0 \text{ and } (S,t) \in (0,+\infty) \times (0,T\,], \\
\partial_t C + \frac{\sigma^2}{2} S^2 \partial_{SS} C + (r-q) S \partial_S C - rC \le L^* \\
\qquad\qquad \text{if } C = 0 \text{ and } (S,t) \in (0,+\infty) \times (0,T\,], \\
C(S,T) \ = (S-K)^+, \qquad\qquad S \in [\,0,+\infty),
\end{cases}
\tag{1.1}
$$

where σ, r, L^*, and K are positive constants and q is a nonnegative constant.

In the appendix we present the financial and stochastic background of this problem.

If $L^* = 0$ in problem (1.1), it is standard European call option, which has an explicit analytic formula of solution and does not have free boundary at all (see [11]). We will find that the case $L^* > 0$ is more complicated than the case $L^* = 0$.

There are some papers in the field of install options, such as [7], [8], [1], [6], in which the authors developed the models and numerical analysis. Particularly, Alobaidi, Mallier, and Deakin showed the behavior of the free boundary $S_s(\tau)$ in (1.1) close to expire by Laplace transforms in [1], that is,

$$
S_s(\tau) \sim K \exp\{-\sigma(-\tau \ln \tau)^{1/2}(1 + o(1))\} \quad \text{as } \tau = T - t \to 0^+. \tag{1.2}
$$

*Received by the editors September 20, 2006; accepted for publication (in revised form) September 17, 2007; published electronically April 23, 2008. The project supported by National Natural Science Foundation of China (10671075), National Natural Science Foundation of GuangDong province (5005930), and University Special Research Fund for Ph.D. Program (20060574002).
http://www.siam.org/journals/sima/40-1/67035.html
†School of Mathematical Sciences, South China Normal University, Guangzhou 510631, China (fhyi@scnu.edu.cn, yangzhou1975@yahoo.com).
‡School of Mathematical Sciences, South China University of Technology, Guangzhou 510640, China.

Since (1.1) is a degenerate backward parabolic problem, we transform it into a familiar forward nondegenerate parabolic variational inequality problem. Thus, letting

$$V(x,\tau) = C(S,t)/K, \qquad \tau = T - t, \qquad x = \ln(S/K), \qquad L = L^*/K, \tag{1.3}$$

then we have

$$\begin{cases} \partial_\tau V - \mathcal{L}_x V = -L & \text{if } V > 0 \quad \text{and} \quad (x,\tau) \in \mathbb{R} \times (0,T\,], \\ \partial_\tau V - \mathcal{L}_x V \geq -L & \text{if } V = 0 \quad \text{and} \quad (x,\tau) \in \mathbb{R} \times (0,T\,], \\ V(x,0) \ = (e^x - 1)^+, & \qquad\qquad\quad\ \ x \in \mathbb{R}, \end{cases} \tag{1.4}$$

where

$$\mathcal{L}_x V = \frac{\sigma^2}{2}\partial_{xx} V + \left(r - q - \frac{\sigma^2}{2}\right)\partial_x V - rV. \tag{1.5}$$

We focus our attention on the monotonicity, regularity, and bounds of the free boundary as well as the limit of the free boundary as $\tau \to +\infty$ in problem (1.4).

As we know, the behavior $\partial_\tau V \geq 0$ is quite important for investigation of the regularity properties of the free boundary. Based on this behavior it can be deduced that $\partial_\tau V$ is continuous across the free boundary and C^∞ regularity of the free boundary follows (for the one-dimensional case, see [9], and for higher dimensions, see [4]). In the absence of condition $\partial_\tau V \geq 0$, the latest development is that $\partial_\tau V$ is continuous for almost all time (see [3]) and the free boundary possesses C^∞ regularity locally around some points which are energetically characterized (see [5]). Those results manifest that the analysis for regularity of the free boundary is not thoroughly solved.

In our case $V(x,0) = e^x - 1$ if $x \geq 0$; combining Lemma 4.4 in this paper, we see that

$$\partial_\tau V(x,0) = \mathcal{L}_x(e^x - 1) - L = r - qe^x - L \quad \text{if } \ x > 0.$$

It is clear that $\partial_\tau V(x,0) < 0$ if $q > 0$ and x is large enough.

In the next section, we will construct a transformation (2.1) and deduce a new unknown function $v(y,\tau)$ satisfying the variational inequality (2.2), which is important to the property $\partial_\tau v \geq 0$; then we prove the existence and uniqueness of the $W^{2.1}_{p,\,loc}$ solution to the new parabolic variational inequality (2.2). The main work is in sections 3 and 4. In section 3 we prove that the new free boundary is monotonic and C^∞-smooth based on the results in section 2. Moreover, we will show the starting point, the bounds of the free boundary, and the limit behavior of the free boundary as $\tau \to +\infty$. In section 4, we come back to consider the behavior of the free boundary of problem (1.4): it is C^∞-smooth on $(0,+\infty)$, and its monotonicity depends on the relationship of parameters σ, r, q, and L in the problem. In section 5, we focus our attention on the monotonicity of free boundary with respect to the parameters r, q, and L. In the last section, we provide numerical result applying the binomial method.

2. Existence and uniqueness of $W^{2,1}_{p,\,loc}$ solution of problem (1.4). As mentioned in the previous section, applying the transformation

$$y = x + (r - q - p^*)\tau, \quad p^* = \max\{r,\, L\}, \quad v(y,\tau) = V(x,\tau), \tag{2.1}$$

problem (1.4) then becomes

$$\begin{cases} \partial_\tau v - \mathcal{L}_y v = -L & \text{if } v > 0 \quad \text{and} \quad (y,\tau) \in \mathbb{R} \times (0,T\,], \\ \partial_\tau v - \mathcal{L}_y v \geq -L & \text{if } v = 0 \quad \text{and} \quad (y,\tau) \in \mathbb{R} \times (0,T\,], \\ v(y,0) \ = (e^y - 1)^+, & \qquad\qquad\quad\ \ y \in \mathbb{R}, \end{cases} \tag{2.2}$$

where

$$\mathcal{L}_y v = \frac{\sigma^2}{2}\partial_{yy} v + \left(p^* - \frac{\sigma^2}{2}\right)\partial_y v - rv. \tag{2.3}$$

Since the problem lies in the unbounded domain $\Omega_T = \mathbb{R}\times(0,T\,]$, we first consider the problem in the bounded domain $\Omega_T^n = (-n,n)\times(0,T\,],\ n \in \mathbb{Z}^+$:

$$\begin{cases} \partial_\tau v_n - \mathcal{L}_y v_n = -L & \text{if } v_n > 0 \quad \text{and} \quad (y,\tau) \in \Omega_T^n, \\ \partial_\tau v_n - \mathcal{L}_y v_n \geq -L & \text{if } v_n = 0 \quad \text{and} \quad (y,\tau) \in \Omega_T^n, \\ v_n(-n,\tau) = 0, \qquad \partial_y v_n(n,\tau) = e^{n+(p^*-r)\tau}, & \qquad\qquad\quad\ \ \tau \in [\,0,T\,], \\ v_n(y,0) \ = (e^y - 1)^+, & \qquad\qquad\quad\ \ y \in [\,-n,n\,]. \end{cases} \tag{2.4}$$

LEMMA 2.1. *For any fixed $n \in \mathbb{Z}^+$, there exists a unique solution $v_n \in C(\overline{\Omega_T^n}) \cap W_p^{2,1}(\Omega_T^n \backslash B_\delta(P_0))$ to problem (2.4), where $\forall\ 1 < p < +\infty$, and $\delta > 0$, $P_0 = (0,0)$, $B_\delta(P_0) = \{(y,\tau) : y^2 + \tau^2 \leq \delta^2\}$. Moreover, if n is large enough, we have*

$$(e^y - 1)^+ \leq v_n \leq e^{y+(p^*-r)\tau}, \tag{2.5}$$

$$\partial_\tau v_n \geq 0, \tag{2.6}$$

$$\partial_y v_n \geq 0. \tag{2.7}$$

Proof. As usual we define a penalty function $\beta_\varepsilon(t)$ (see Figure 1), which satisfies

$$\varepsilon > 0 \text{ and small enough}, \quad \beta_\varepsilon(t) \in C^\infty(-\infty,+\infty),$$
$$\beta_\varepsilon(t) \leq 0, \qquad 0 \leq \beta_\varepsilon'(t) \leq 2L/\varepsilon, \qquad \beta_\varepsilon'' \leq 0,$$

and

$$\beta_\varepsilon(t) = \begin{cases} 0, & t \geq 2\varepsilon, \\ \frac{2L}{\varepsilon}t - 3L - r\varepsilon, & t \leq 3\varepsilon/2. \end{cases} \tag{2.8}$$

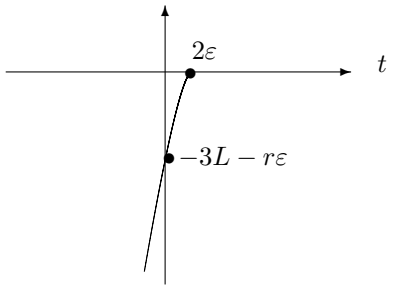


FIG. 1.

Then

$$\beta_\varepsilon(\varepsilon) = -L - r\varepsilon, \qquad \beta_\varepsilon(0) = -3L - r\varepsilon. \tag{2.9}$$

Since $(e^y-1)^+$ is not smooth enough, we need to smooth it. Define $\pi_\varepsilon(t)$ (see Figure 2):

$$\pi_\varepsilon(t) = \begin{cases} t, & t \geq \varepsilon, \\ 0, & t \leq -\varepsilon, \end{cases} \tag{2.10}$$

$\pi_\varepsilon(t) \in C^\infty$, $0 \leq \pi'_\varepsilon(t) \leq 1$, $\pi''_\varepsilon(t) \geq 0$, $\lim_{\varepsilon\to 0^+} \pi_\varepsilon(t) = t^+$.

Following the idea in [10], construct an approximation of problem (2.4):

$$\begin{cases} \partial_\tau v_{\varepsilon,n} - \mathcal{L}_y v_{\varepsilon,n} + \beta_\varepsilon(v_{\varepsilon,n}) = -L & \text{in } \Omega_T^n, \\ v_{\varepsilon,n}(-n,\tau) = 0, \qquad \partial_y v_{\varepsilon,n}(n,\tau) = e^{n+(p^*-r)\tau}, & \tau \in [\,0,T\,], \\ v_{\varepsilon,n}(y,0) = \pi_\varepsilon(e^y-1), & y \in [\,-n,n\,]. \end{cases} \tag{2.11}$$

Applying Schauder's fixed point theorem, it is not difficult to get the existence of the $W_p^{2,1}$ solution to problem (2.11). The proof of uniqueness is standard as well, so we omit the details.

If we can prove that, as ε is small enough,

$$\pi_\varepsilon(e^y-1) \leq v_{\varepsilon,n} \leq e^{y+(p^*-r)\tau}, \tag{2.12}$$

then

$$-3L - r\varepsilon \leq \beta_\varepsilon(v_{\varepsilon,n}) \leq 0.$$

This means that $\beta_\varepsilon(v_{\varepsilon,n})$ is a bounded function and its bound is independent of ε. It is deduced that, by $W_p^{2,1}$ and C^α $(0<\alpha<1)$ estimates of the parabolic problem,

$$|v_{\varepsilon,n}|_{W_p^{2,1}(\Omega_T^n \setminus B_\delta(P_0))} \leq C,$$
$$|v_{\varepsilon,n}|_{C^\alpha(\overline{\Omega_T^n})} \leq C,$$

where C is independent of ε. It is not difficult to derive that, as $\varepsilon \to 0^+$,

$$v_{\varepsilon,n} \to v_n \ \text{ in } \ W_p^{2,1}(\Omega_T^n \setminus B_\delta(P_0)) \text{ weakly} \qquad \text{and} \qquad v_{\varepsilon,n} \to v_n \ \text{ in } \ C(\overline{\Omega_T^n}),$$

where v_n is the solution of problem (2.4).

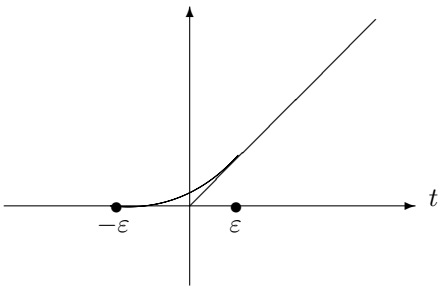


FIG. 2.

Next we prove (2.12). In the first, if ε is small enough, the properties of β_ε, π_ε, and (2.9) imply that

$$\begin{aligned}
&\partial_\tau \pi_\varepsilon(e^y-1) - \mathcal{L}_y\,\pi_\varepsilon(e^y-1) + \beta_\varepsilon(\pi_\varepsilon(e^y-1)) + L\\
&= -\frac{\sigma^2}{2}\pi_\varepsilon''(e^y-1)e^{2y} - \max\{r,\,L\}\pi_\varepsilon'(e^y-1)e^y + r\pi_\varepsilon(e^y-1) + \beta_\varepsilon(\pi_\varepsilon(e^y-1)) + L\\
&\le \begin{cases} r\pi_\varepsilon(e^y-1) + \beta_\varepsilon(\pi_\varepsilon(e^y-1)) + L \le r\varepsilon + \beta_\varepsilon(\varepsilon) + L \le 0, & e^y-1 \le \varepsilon,\\ -\max\{r,\,L\}e^y + r(e^y-1) + L \le 0, & e^y-1 > \varepsilon. \end{cases}
\end{aligned}$$

Furthermore, from the initial and boundary conditions in (2.11), we deduce that if ε is small enough,

$$\begin{cases} \pi_\varepsilon(e^y-1) = 0 = v_{\varepsilon,\,n}(y,\tau), & y=-n,\\ \partial_y\pi_\varepsilon(e^y-1) = \partial_y(e^y-1) = e^n \le e^{n+(p^*-r)\tau} = \partial_y v_{\varepsilon,\,n}(y,\tau), & y=n,\\ \pi_\varepsilon(e^y-1) = v_{\varepsilon,\,n}(y,0), & \tau=0. \end{cases}$$

From the comparison principle, we deduce $\pi_\varepsilon(e^y-1) \le v_{\varepsilon,\,n}$.

On the other hand, if $2\varepsilon < e^{-n}$, it is not difficult to check $\beta_\varepsilon(e^{y+(p^*-r)\tau}) \ge \beta_\varepsilon(e^{-n}) \ge \beta_\varepsilon(2\varepsilon) = 0$; hence, from (2.3),

$$\begin{aligned}
&\partial_\tau e^{y+(p^*-r)\tau} - \mathcal{L}_y e^{y+(p^*-r)\tau} + \beta_\varepsilon(e^{y+(p^*-r)\tau}) + L\\
&\ge e^{y+(p^*-r)\tau}\left(p^* - r - \frac{\sigma^2}{2} - \left(p^* - \frac{\sigma^2}{2}\right) + r\right) + 0 + L = L \ge 0 \quad \forall\,(y,\tau)\in\Omega_T^n.
\end{aligned}$$

Moreover, if $2\varepsilon < e^{-n}$, there holds

$$e^y = \pi_\varepsilon(e^y) \ge \pi_\varepsilon(e^y-1).$$

From the initial and boundary conditions in (2.11), we see that if $2\varepsilon < e^{-n}$,

$$\begin{cases} e^{y+(p^*-r)\tau} \ge 0 = v_{\varepsilon,\,n}(y,\tau), & y=-n,\\ \partial_y e^{y+(p^*-r)\tau} = e^{n+(p^*-r)\tau} = \partial_y v_{\varepsilon,\,n}(y,\tau), & y=n,\\ e^{y+(p^*-r)\tau} = e^y \ge \pi_\varepsilon(e^y-1) = v_{\varepsilon,\,n}(y,0), & \tau=0. \end{cases}$$

Then the comparison principle implies $e^{y+(p^*-r)\tau} \ge v_{\varepsilon,\,n}$; hence, we obtain (2.12), and (2.5) is a consequence of (2.12).

In the following, we prove (2.6). In fact, for any small $\delta > 0$, $v_{\varepsilon,\,n}(x,\tau+\delta)$ satisfies, by (2.11),

$$\begin{cases} \partial_\tau v_{\varepsilon,\,n}(y,\tau+\delta) - \mathcal{L}_y v_{\varepsilon,\,n}(y,\tau+\delta) + \beta_\varepsilon(v_{\varepsilon,\,n}(y,\tau+\delta)) = -L, & \\ \qquad\qquad (y,\tau)\in(-n,n)\times(0,T-\delta\,], & \\ v_{\varepsilon,\,n}(-n,\tau+\delta) = 0 = v_{\varepsilon,\,n}(-n,\tau), & \tau\in[\,0,T-\delta\,],\\ \partial_y v_{\varepsilon,\,n}(n,\tau+\delta) = e^{n+(p^*-r)(\tau+\delta)} \ge \partial_y v_{\varepsilon,\,n}(n,\tau), & \tau\in[\,0,T-\delta\,],\\ v_{\varepsilon,\,n}(y,0+\delta)\ \ge \pi_\varepsilon(e^y-1) = v_{\varepsilon,\,n}(y,0), & y\in[\,-n,n\,]. \end{cases}$$

Applying the comparison principle for solutions of PDEs with respect to initial and boundary values, we obtain

$$v_{\varepsilon,\,n}(y,\tau+\delta) \ge v_{\varepsilon,\,n}(y,\tau) \quad \forall\,(y,\tau)\in(-n,n)\times[\,0,T-\delta\,] \qquad \text{and} \qquad \partial_\tau v_{\varepsilon,\,n} \ge 0.$$

Take $\varepsilon \to 0^+$ in the above inequalities. Then (2.6) follows.

For proving (2.7), derive (2.11) with respect to y, and denote $W = \partial_y v_{\varepsilon,n}$; then

$$\begin{cases} \partial_\tau W - \mathcal{L}_y W + \beta'_\varepsilon(v_{\varepsilon,n})W = 0, & (y,\tau) \in \Omega_T^n, \\ W(-n,\tau) = \partial_y v_{\varepsilon,n}(-n,\tau), & \tau \in [\,0,T\,], \\ W(n,\tau) = e^{n+(p^*-r)\tau} \geq 0, & \tau \in [\,0,T\,], \\ W(y,0) = \pi'_\varepsilon(e^y-1)e^y \geq 0, & y \in [\,-n,n\,]. \end{cases} \tag{2.13}$$

Since $v_{\varepsilon,n}(x,\tau)$ achieves its minimum 0 at each point of $x=-n$, so $\partial_y v_{\varepsilon,n}(-n,\tau) \geq 0$, then the minimum principle implies

$$\partial_y v_{\varepsilon,n} = W \geq 0. \tag{2.14}$$

Take $\varepsilon \to 0^+$. Then we obtain (2.7).

At last, we prove uniqueness. Suppose v_1 and v_2 are two $W^{2.1}_{p,loc}(\Omega_T^n) \bigcap C(\overline{\Omega_T^n})$ solutions to problem (2.4), and denote

$$\mathcal{N} = \{(y,\tau) : v_1(y,\tau) < v_2(y,\tau),\ -n < y < n,\ 0 < \tau \leq T\}.$$

Suppose it is not empty; then if $(y,\tau) \in \mathcal{N}$,

$$v_2(y,\tau) > 0, \quad \partial_\tau v_2 - \mathcal{L}_y v_2 = -L.$$

Denote $W = v_2 - v_1$. Then W satisfies

$$\begin{cases} \partial_\tau W - \mathcal{L}_y W \leq 0, & (y,\tau) \in \mathcal{N}, \\ W(y,0) = 0 & \text{on } \partial_p\mathcal{N} \setminus (\{n\} \times [\,0,T\,]), \\ \partial_y W(y,0) = 0 & \text{on } \partial_p\mathcal{N} \cap (\{n\} \times [\,0,T\,]), \end{cases}$$

where $\partial_p\mathcal{N}$ is the parabolic boundary of the domain $\mathcal{N}$. Applying the A-B-P maximum principle (see [13]), we have $W \leq 0$ in $\mathcal{N}$, which contradicts the definition of $\mathcal{N}$. □

THEOREM 2.2. *There exists a unique solution $v \in C(\overline{\Omega_T}) \cap W^{2,1}_p(\Omega_T^R \backslash B_\delta(P_0))$ to problem* (2.2), *where $\forall\ R > 0,\ \delta > 0,\ 1 < p < +\infty$. And $\partial_y v \in C(\Omega_T)$,*

$$(e^y - 1)^+ \leq v \leq e^{y+(p^*-r)\tau}, \tag{2.15}$$

$$\partial_\tau v \geq 0, \tag{2.16}$$

$$\partial_y v \geq 0. \tag{2.17}$$

Proof. Applying

$$(\partial_\tau - \mathcal{L}_y)\, 0 = 0,$$

we rewrite problem (2.4) as

$$\begin{cases} \partial_\tau v_n - \mathcal{L}_y v_n = f(y,\tau) & \text{in } \Omega_T^n, \\ v_n(-n,\tau) = 0, \qquad \partial_y v_n(n,\tau) = e^{y+(p^*-r)\tau}, & \tau \in [\,0,T\,], \\ v_n(y,0) = (e^y-1)^+, & y \in [\,-n,n\,]. \end{cases} \tag{2.18}$$

$v_n \in W^{2.1}_{p,loc}(\,\Omega_T^n)$ implies $f(y,\tau) \in L^p_{loc}(\,\Omega_T^n)$ and

$$f(y,\tau) = I_{\{v_n>0\}}(-L) \qquad \text{a.e. in } \Omega_T^n, \tag{2.19}$$

where I_A denotes the indicator function of the set A.

Hence, for any fixed $R > \delta > 0$, if $n > R$, combining (2.5), we have the following $W_p^{2.1}$ uniform interior estimates in the domain $\Omega_T^R \backslash B_\delta(P_0)$:

$$\|v_n\|_{W_p^{2.1}(\Omega_T^R \backslash B_\delta(P_0))} \leq C(\|v_n\|_{L^\infty(\Omega_T^R)} + \|(e^y - 1)^+\|_{C^2([-R,-\delta]\cup[\delta,R])} + \|f(y,\tau)\|_{L^\infty(\Omega_T^R)}) \leq C,$$

where C depends on R but is independent of n. Let $n \to \infty$; then we have, possibly a subsequence,

$$v_n \rightharpoonup v_R \text{ in } W_p^{2.1}(\Omega_T^R \backslash B_\delta(P_0)) \text{ weakly} \qquad \text{as} \quad n \to +\infty.$$

Moreover, the Sobolev imbedding theorem implies

$$\partial_y v_n \to \partial_y v_R \text{ in } C(\Omega_T^R \backslash B_\delta(P_0)) \quad \text{as} \quad n \to +\infty.$$

Define $v = v_R$ if $y \in [-R, R]$. It is clear that v is reasonably defined and v is the solution of problem (2.2); moreover, $\partial_y v \in C(\Omega_T)$ and the C^α estimate implies $v \in C(\overline{\Omega_T})$.

Inequalities (2.15), (2.16), (2.17) are a consequence of (2.5), (2.6), (2.7), respectively. The proof of the uniqueness is the same as in the proof of Lemma 2.1. □

From the transformation (2.1), it is not difficult to see the following.

THEOREM 2.3. *There exists a unique solution* $V \in C(\overline{\Omega_T}) \cap W_p^{2,1}(\Omega_T^R \backslash B_\delta(P_0))$ *to problem* (1.4), *where* $\forall\ R > 0,\ \delta > 0,\ 1 < p < +\infty$. *And* $\partial_x V \in C(\Omega_T)$,

$$(e^{x+(r-q-p^*)\tau} - 1)^+ \leq V \leq e^{x-q\tau}, \tag{2.20}$$

$$\partial_x V \geq 0, \tag{2.21}$$

$$\partial_\tau V \geq (r - q - p^*)\, \partial_x V. \tag{2.22}$$

3. Characterizations of the free boundary of problem (2.2). In this section, we consider the behavior of the free boundary of problem (2.2). Denote

$$\mathbf{NR}_y = \{(y,\tau) :\ v(y,\tau) > 0\} \qquad \text{(nontransaction region)},$$
$$\mathbf{SR}_y = \{(y,\tau) :\ v(y,\tau) = 0\} \qquad \text{(stop region)}.$$

Applying (2.17), v is monotonic increasing with respect to y, so we can define the free boundary

$$y_s(\tau) = \sup\{y :\ v(y,\tau) = 0\}, \quad \tau > 0, \qquad \text{between } \mathbf{NR}_y \text{ and } \mathbf{SR}_y.$$

THEOREM 3.1. $y_s(\tau) \in C[0,T] \cap C^\infty(0,T]$ *and is strictly decreasing with* $y_s(0) = 0$ *(see Figure* 3*).*

Proof. We divide the proof into four steps.

Step 1. From (2.16) and (2.17), we see that $y_s(\tau)$ is decreasing in $[0,T]$.

Step 2. Prove $y_s(\tau)$ is continuous in $[0,T]$ and $y_s(0) = 0$.

In the first we prove $y_s(\tau)$ is continuous in $[0,T]$. Otherwise, there exists a domain $(y_1, y_2) \times (\tau_0, T)$ $(y_1 < y_2,\ 0 \leq \tau_0 < T)$ such that

$$\begin{cases} \partial_\tau v - \mathcal{L}_y v = -L, & (y,\tau) \in (y_1,y_2) \times (\tau_0, T), \\ v(y,\tau_0) = 0, & y \in (y_1, y_2). \end{cases}$$

Then we have $\partial_\tau v(y,\tau_0) = -L < 0$ for any $y_1 < y < y_2$, which contradicts (2.16).

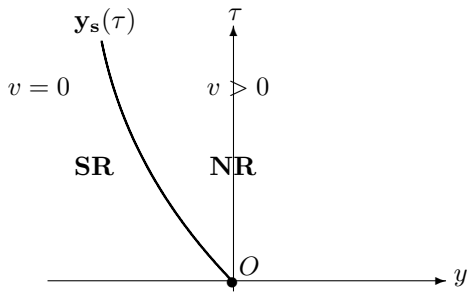


FIG. 3.

In the same way we can prove $y_s(0) \geq 0$; moreover, since $v(y,0) = (e^y - 1)^+ > 0$ if $y > 0$, we have $y_s(0) = 0$.

Step 3. Prove $y_s(\tau)$ is strictly decreasing in $[\,0, T\,]$.

Otherwise there exists a domain $(y_0, 1) \times (\tau_1, \tau_2)$ $(y_0 \leq 0,\ 0 < \tau_1 < \tau_2 \leq T)$ such that

$$\text{(3.1)} \qquad \begin{cases} \partial_\tau v - \mathcal{L}_y v = -L, & (x,\tau) \in (y_0, 1) \times (\tau_1, \tau_2), \\ v(y_0, \tau) = 0, & \tau \in (\tau_1, \tau_2). \end{cases}$$

Then $W = \partial_\tau v$ satisfies

$$\begin{cases} \partial_\tau W - \mathcal{L}_y W = 0, & (x,\tau) \in (y_0, 1) \times (\tau_1, \tau_2), \\ W(y_0, \tau) = 0, & \tau \in (\tau_1, \tau_2). \end{cases}$$

Since $W = \partial_\tau v \geq 0$, W achieves its nonpositive minimum at $y = y_0$. Applying the maximum principle, we have $\partial_y W(y_0, \tau) > 0$ for any $\tau \in (\tau_1, \tau_2)$. On the other hand, we can deduce that $\partial_y v(y_0, \tau) = 0$ for any $\tau \in (\tau_1, \tau_2)$ by $\partial_y v \in C(\Omega_T)$. So, $\partial_y W(y_0, \tau) = \partial_{\tau y} v(y_0, \tau) = 0$ for any $\tau \in (\tau_1, \tau_2)$; thus we get a contradiction. Hence $y_s(\tau)$ is strictly decreasing in $(0, T\,]$.

Step 4. Since $\partial_\tau v \geq 0$ and 0 is the lower obstacle, it can be proved that $y_s(\tau) \in C^{0,1}(0, T\,]$ by the method developed by Friedman in [9]. Moreover $y_s(\tau) \in C^\infty(0, T\,]$ by a bootstrap argument. □

THEOREM 3.2. *The free boundary $y_s(\tau)$ satisfies*

$$\text{(3.2)} \quad \ln\left[\frac{2L}{2r+\sigma^2}\left(1 - \frac{1}{1+r\tau}\right)\right] \leq y_s(\tau) + (p^* - r)\tau \leq \ln\left[\frac{2L}{2r+\sigma^2}\left(1 + \frac{1}{L\tau}\right)\right].$$

Proof. We divide the proof into four steps.

Step 1. In the first, we give a transformation

$$\text{(3.3)} \qquad z = y + (p^* - r)\tau, \qquad v^*(z,\tau) = v(y,\tau) = V(x,\tau).$$

Then we have

$$\text{(3.4)} \quad \begin{cases} \partial_\tau v^* - \mathcal{L}_z v^* = -L & \text{if } v^* > 0 \quad \text{and} \quad (z,\tau) \in \mathbb{R} \times (0, T\,], \\ \partial_\tau v^* - \mathcal{L}_z v^* \geq -L & \text{if } v^* = 0 \quad \text{and} \quad (z,\tau) \in \mathbb{R} \times (0, T\,], \\ v^*(z,0) = (e^z - 1)^+, & z \in \mathbb{R}, \end{cases}$$

where

$$\mathcal{L}_z v^* = \frac{\sigma^2}{2}\partial_{zz} v + \left(r - \frac{\sigma^2}{2}\right)\partial_z v - rv. \tag{3.5}$$

Denote the free boundary of problem (3.4) as $z_s(\tau)$, which is the counterpart of $y_s(\tau)$ and $z_s(\tau) = y_s(\tau) + (p^* - r)\tau$.

Step 2. We prove that for any $\tau > 0$, there holds

$$z_s(\tau) \geq z_1(\tau) = \ln\left[\frac{2L}{2r+\sigma^2}\left(1 - \frac{1}{1+r\tau}\right)\right]. \tag{3.6}$$

In fact, for any $T_0 > 0$, we will prove $z_s(T_0) \geq z_1(T_0)$.

Since $z_1(T_0) = \ln[\frac{2L}{2r+\sigma^2}(1 - \frac{1}{1+rT_0})]$, we define z_1 by

$$e^{z_1} = \frac{2L}{2r+\sigma^2}\left(1 - \frac{1}{1+rT_0}\right).$$

It follows that

$$e^{z_1} = \frac{2L}{2r+\sigma^2}\,\frac{rT_0}{1+rT_0} \leq \frac{LT_0}{1+rT_0}. \tag{3.7}$$

We define

$$W(z,\tau) = \begin{cases} \frac{L}{1+rT_0}(T_0 - \tau), & (z,\tau) \in (-\infty, z_1\,] \times [\,0, T_0\,], \\ e^z - \frac{1}{\alpha}e^{\alpha(z-z_1)+z_1} - \frac{L}{1+rT_0}\tau, & (z,\tau) \in (z_1, +\infty) \times [\,0, T_0\,], \end{cases} \tag{3.8}$$

where $\alpha = -\frac{2r}{\sigma^2}$.

We claim that $W(x,\tau)$ possesses the following four properties:

(I) $W \in W^{2,1}_{p,\,loc}(\mathbb{R} \times (0,T_0)) \cap C(\mathbb{R} \times [\,0,T_0\,])$.

(II) $W \geq 0 \ \ \forall (z,\tau) \in \mathbb{R} \times (0,T_0\,]$.

(III) $W(z,0) \geq (e^z - 1)^+$.

(IV) $\partial_\tau W - \mathcal{L}_z W \geq -L$, $(z,\tau) \in \mathbb{R} \times [\,0,T_0\,]$.

Indeed, since

$$W(z_1 + 0, \tau) = e^{z_1}\left(1 - \frac{1}{\alpha}\right) - \frac{L}{1+rT_0}\tau = \frac{L}{1+rT_0}(T_0 - \tau), \tag{3.9}$$

$$\partial_z W(z_1 + 0, \tau) = e^{z_1} - e^{z_1} = 0. \tag{3.10}$$

Then property (I) follows. Next, we prove properties (II) and (III).

It is clear that $W(z,\tau) \geq 0$ for $(z,\tau) \in (-\infty, z_1\,] \times [\,0,T_0\,]$. Moreover, it can be seen that if $z > z_1$,

$$\partial_{zz} W = e^z - \alpha e^{\alpha(z-z_1)+z_1} \geq e^z > 0.$$

Combining (3.9) and (3.10), we know that for any $(z,\tau) \in (z_1, +\infty) \times [\,0,T_0\,]$, there hold

$$\partial_z W(z,\tau) > 0, \qquad W(z,\tau) > \frac{L}{1+rT_0}(T_0 - \tau) \geq 0.$$

Thus we get property (II). On the other hand, applying (3.7), we obtain

$$W(z,0)=\begin{cases} \frac{LT_0}{1+rT_0}\geq e^{z_1}>(e^z-1)^+, & z\leq z_1,\\ e^z-\frac{1}{\alpha}e^{\alpha(z-z_1)+z_1}>e^z>(e^z-1)^+, & z>z_1. \end{cases}$$

Then we have property (III). In the following, we prove property (IV).

In fact, notice that 1 and α are the characteristic roots of the ODE $\mathcal{L}_z W=0$. Then we see that for any $(z,\tau)\in(z_1,+\infty)\times(0,T_0]$,

$$\partial_\tau W-\mathcal{L}_z W=-\frac{L}{1+rT_0}-\frac{rL}{1+rT_0}\tau=-L\frac{1+r\tau}{1+rT_0}\geq -L.$$

On the other hand, for any $(z,\tau)\in(-\infty,z_1)\times(0,T_0]$, as above, we have

$$\partial_\tau W-\mathcal{L}_z W=-\frac{L}{1+rT_0}+\frac{rL}{1+rT_0}(T_0-\tau)\geq -\frac{L}{1+rT_0}-\frac{rL}{1+rT_0}\tau\geq -L.$$

Hence we have property (IV).

Properties (I)–(IV) indicate that W is a supersolution of problem (3.4). Hence $W\geq v^*$ in $\mathbb{R}\times[\,0,T\,]$, especially

$$0\leq v^*(z,T_0)\leq W(z,T_0)=0,\qquad z\leq z_1,$$

meaning that $z_s(T_0)\geq z_1(T_0)$. Then (3.6) follows.

Step 3. We prove that for any $\tau>0$,

$$z_s(\tau)\leq z_2(\tau)=\ln\left[\frac{2L}{2r+\sigma^2}\left(1+\frac{1}{L\tau}\right)\right]. \tag{3.11}$$

In fact, for any $T_0>0$, denote

$$w(z,\tau)=\begin{cases} -1+\frac{\tau}{T_0}, & (z,\tau)\in(-\infty,z_2\,]\times[\,0,T_0\,],\\ e^z-\frac{1}{\alpha}e^{\alpha(z-z_2)+z_2}-A+(-1+\frac{\tau}{T_0}), & (z,\tau)\in(z_2,+\infty)\times[\,0,T_0\,], \end{cases} \tag{3.12}$$

where $\alpha=-\frac{2r}{\sigma^2}$ and

$$e^{z_2}=\frac{2L+2/T_0}{2r+\sigma^2},\qquad\qquad A=\frac{\alpha-1}{\alpha}e^{z_2}=\frac{L+1/T_0}{r}.$$

We claim that $w(z,\tau)$ satisfies the following three properties:

(I) $w\in W^{2,1}_{p,\,loc}(\mathbb{R}\times(0,T))\cap C(\mathbb{R}\times[\,0,T\,])$.

(II) $w(z,0)\leq(e^z-1)^+$.

(III) $\partial_\tau w-\mathcal{L}_z w\leq -L$ if $w>0$ and $(z,\tau)\in\mathbb{R}\times(0,T_0\,]$.

Indeed, as in the previous step, property (I) is obvious. Next, we prove properties (II) and (III). Considering

$$w(z,0)=\begin{cases} -1<0<(e^z-1)^+, & z\leq z_2,\\ (e^z-1)-(\frac{1}{\alpha}e^{\alpha(z-z_2)+z_2}+A)<(e^z-1)-(\frac{1}{\alpha}e^{z_2}+A) & \\ \qquad =(e^z-1)-e^{z_2}<(e^z-1)\leq(e^z-1)^+, & z>z_2, \end{cases}$$

property (II) follows.

In the following, we prove property (Ⅲ). It is clear that $w(z,\tau) \le 0$ for any $(z,\tau) \in (-\infty, z_2\,] \times [\,0, T_0\,]$. On the other hand, for any $(z,\tau) \in (z_2, +\infty) \times (0, T_0]$, by the method in Step 2, we see that

$$\partial_\tau w - \mathcal{L}_z w = \frac{1}{T_0} - rA + r\left(-1 + \frac{\tau}{T_0}\right) \le \frac{1}{T_0} - rA = \frac{1}{T_0} - r\frac{L + 1/T_0}{r} = -L.$$

Hence we have property (Ⅲ).

Based on properties (I)–(Ⅲ), as in the proof of the uniqueness in Lemma 2.1, we can prove $w \le v^*$ in $\mathbb{R} \times [\,0, T\,]$. Moreover, for $z > z_2$,

$$w(z, T_0) = (e^z - e^{z_2}) - \frac{1}{\alpha} e^{z_2}(e^{\alpha(z - z_2)} - 1) > 0.$$

Thus $v^*(z, T_0) > 0$ for any $z > z_2$, and (3.11) follows.

Step 4. From (3.6) and (3.11), we see that

$$\ln\left[\frac{2L}{2r + \sigma^2}\left(1 - \frac{1}{1 + r\tau}\right)\right] \le z_s(\tau) \le \ln\left[\frac{2L}{2r + \sigma^2}\left(1 + \frac{1}{L\tau}\right)\right].$$

Since $z_s(\tau) = y_s(\tau) + (p^* - r)\tau$, then (3.2) follows. □

Remark. Let $\tau \to +\infty$ in (3.2). We can deduce that

$$y_s(\tau) + (p^* - r)\tau = z_s(\tau) \to \ln\left(\frac{2L}{2r + \sigma^2}\right) \qquad \text{as} \qquad \tau \to +\infty.$$

Moreover, combining $z_s(\tau) = y_s(\tau) + (p^* - r)\tau$ is continuous in $[\,0, T\,]$. We see that there exist two constants M_1 and M_2 such that $M_1 \le z_s(\tau) \le M_2$, which are independent of T.

4. Characterizations of the free boundary of problem (1.4). We have obtained some properties (Theorems 3.1 and 3.2) about the free boundary $y_s(\tau)$ of the new problem (2.2). In the following, we come back to consider the free boundary of the original problem (1.4).

Denote $x_s(\tau)$ as the free boundary between $\mathbf{NR}_x$ and $\mathbf{SR}_x$, which are, respectively, the counterparts of $\mathbf{NR}_y$ and $\mathbf{SR}_y$ in the transformation (2.1). From the transformation (2.1), we have

$$V(x,\tau) = v(x + (r - q - p^*)\tau, \tau), \qquad x_s(\tau) = y_s(\tau) + (q + p^* - r)\tau.$$

Theorems 3.1 and 3.2 imply Theorem 4.1

THEOREM 4.1. (1) $x_s(\tau) \in C[\,0, T\,] \cap C^\infty(0, T\,]$ *with* $x_s(0) = 0$ *and satisfies*

$$\ln\left[\frac{2L}{2r + \sigma^2}\left(1 - \frac{1}{1 + r\tau}\right)\right] \le x_s(\tau) - q\tau \le \ln\left[\frac{2L}{2r + \sigma^2}\left(1 + \frac{1}{L\tau}\right)\right]. \tag{4.1}$$

(2) *If* $q = 0$ *and* $r \ge L$, $x_s(\tau) = y_s(\tau)$ *is strictly decreasing (see Figure* 4*).*

(3) *If* $q > 0$, $x_s(\tau) \to +\infty$ *as* $\tau \to +\infty$ *(see Figure* 5*).*

LEMMA 4.2. *If* $q = 0$ *and* $r < L \le r + \sigma^2/2$ *(see Figure* 6*), then*

$$x_s(\tau) \le 0.$$

Proof. Denote

$$w(x,\tau) = \begin{cases} 0, & (x,\tau) \in (-\infty, 0\,] \times [\,0, T\,], \\ C_1 e^x + C_2 e^{\alpha x} - \frac{L}{r}, & (x,\tau) \in (0, +\infty) \times [\,0, T\,], \end{cases} \tag{4.2}$$

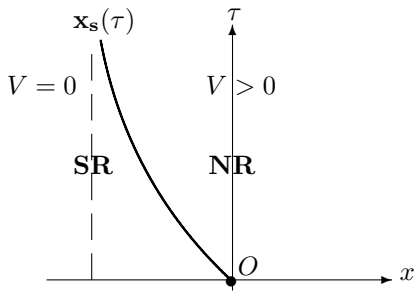


FIG. 4. $q = 0$ *and* $r \geq L$.

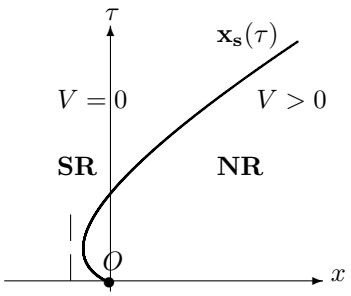


FIG. 5. $q > 0$.

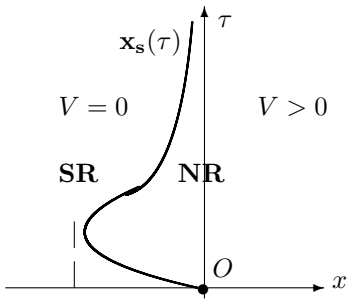


FIG. 6. $q = 0$ *and* $r < L \leq r + \sigma^2/2$.

where

$$\alpha = -2r/\sigma^2, \qquad C_1 = \frac{-\alpha L}{r(1-\alpha)} = \frac{2L}{2r+\sigma^2} \leq 1, \qquad C_2 = \frac{L}{r(1-\alpha)}. \tag{4.3}$$

We claim that $w(x,\tau)$ satisfies the following three properties:

$$\begin{cases} \text{(I) } w \in W^{2,1}_{p,\,loc}(\mathbb{R}\times(0,T)) \cap C(\mathbb{R}\times[\,0,T\,]), \\ \text{(II) } 0 \leq w(x,0) \leq (e^x-1)^+, \\ \text{(III) } \partial_\tau w - \mathcal{L}_x w = -L \quad \text{if } w > 0 \text{ and } (x,\tau) \in \mathbb{R}\times(0,T\,]. \end{cases} \tag{4.4}$$

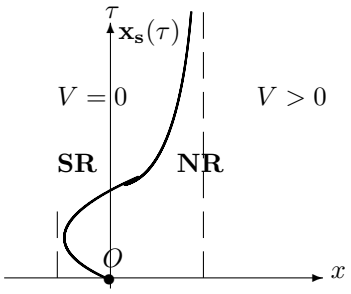


FIG. 7. $q = 0$ *and* $L > r + \sigma^2/2$.

Indeed, from (4.2) and (4.3), we have

$$w(0+0,\tau) = C_1 + C_2 - \frac{L}{r} = 0, \qquad \partial_x w(0+0,\tau) = C_1 + \alpha C_2 = 0. \tag{4.5}$$

Then we get property (I). Next, we prove property (II). Indeed, from (4.2), we have

$$\partial_{xx} w(x,\tau) = C_1 e^x + C_2 \alpha^2 e^{\alpha x} > 0 \qquad \forall x > 0.$$

This fact and (4.5) imply $\partial_x w(x,\tau) \geq 0$ and $w(x,\tau) \geq 0$ for any $x > 0$. Moreover, since as $x > 0$, there hold, by $C_1 \leq 1$,

$$\partial_x w(x,\tau) - \partial_x (e^x - 1) = C_1 e^x + C_2 \alpha e^{\alpha x} - e^x < 0 \qquad \text{and} \qquad w(0,0) = 0,$$

we have $w(x,\tau) < e^x - 1$ for any $x > 0$. Then, combining this result with (4.2), we get conclusion (II).

In the following, we prove property (III). Notice that 1 and α are the characteristic roots of ODE $\mathcal{L}_x w = 0$; then we have, if $(x,\tau) \in (0,+\infty) \times [\,0, T_0\,]$,

$$\partial_\tau w - \mathcal{L}_x w = -r\frac{L}{r} = -L.$$

Based on properties (I)–(III), as in the proof of the uniqueness in Lemma 2.1, we can prove $w \leq V$. Since $w > 0$ for any $x > 0$, so $V > 0$ for any $x > 0$; therefore, $x_s(\tau) \leq 0$. □

LEMMA 4.3. *If* $q = 0$ *and* $L > r + \sigma^2/2$, *then* $x_s(\tau) \leq \ln \frac{2L}{2r+\sigma^2}$ *(see Figure* 7*).*

Proof. Denote

$$w(x,\tau) = \begin{cases} 0, & (x,\tau) \in (-\infty, x_0\,] \times [\,0, T\,], \\ e^x - \frac{1}{\alpha} e^{\alpha(x-x_0)+x_0} - \frac{L}{r}, & (x,\tau) \in (x_0, +\infty) \times [\,0, T\,], \end{cases} \tag{4.6}$$

where

$$\alpha = \frac{-2r}{\sigma^2}, \qquad e^{x_0} = \frac{\alpha L}{r(\alpha - 1)} = \frac{2L}{2r + \sigma^2} > 1. \tag{4.7}$$

The proof is analogous to that of Lemma 4.2. We claim that $w(x,\tau)$ still satisfies the three properties in (4.4). Indeed, from (4.6) and (4.7), we have

$$w(x_0+0,\tau) = \frac{\alpha - 1}{\alpha} e^{x_0} - \frac{L}{r} = 0, \quad \partial_x w(x_0+0,\tau) = e^{x_0} - \alpha \frac{e^{x_0}}{\alpha} = 0. \tag{4.8}$$

Then we get property (I).

Next, from (4.6), we have

$$\partial_{xx} w(x,\tau) = e^x - \alpha e^{\alpha(x-x_0)+x_0} > 0 \qquad \forall\, (x,\tau) \in (x_0, +\infty) \times [\,0, T\,],$$

which, combined with (4.8), implies $\partial_x w(x,\tau) > 0$ and $w(x,\tau) > 0$ for any $x > x_0$. Moreover, since $x_0 > 0$ by $e^{x_0} > 1$, there hold, if $x > x_0$,

$$\partial_x w(x,\tau) - \partial_x (e^x - 1) = -e^{\alpha(x-x_0)+x_0} < 0 \qquad \text{and} \qquad w(x_0, 0) = 0 < e^{x_0} - 1;$$

hence, $w(x,\tau) < e^x - 1$ for any $x > x_0$. Combining this result with (4.6), we get conclusion (II).

Finally, as in the proof of Lemma 4.2, we can check

$$\partial_\tau w - \mathcal{L}_x w = -r\frac{L}{r} = -L, \qquad \forall\, (x,\tau) \in (x_0, +\infty) \times [\,0, T_0\,];$$

hence $w \le V$. Since $w > 0$ for any $x > x_0$, then $V > 0$ for any $x > x_0$; therefore $x_s(\tau) \le x_0 = \ln \frac{2L}{2r+\sigma^2}$. □

LEMMA 4.4. *There exists a $\tau_0 > 0$ such that $V(0,\tau) > 0$ for any $0 < \tau < \tau_0$ (see Figures 4–7), i.e.,*

$$\{x = 0,\ 0 < \tau < \tau_0\} \subset \mathbf{NR}, \qquad\qquad x_s(\tau) < 0 \quad \textit{for} \quad 0 < \tau < \tau_0.$$

Proof. The solution $V(x,t)$ of (1.4) satisfies

$$\text{(4.9)} \quad \begin{cases} \partial_\tau V - \mathcal{L}_x V \ge -L, & (x,\tau) \in \mathbb{R} \times (0, T\,], \\ V(x,0) \ = (e^x - 1)^+, & x \in \mathbb{R}. \end{cases}$$

Consider the Cauchy problem

$$\text{(4.10)} \quad \begin{cases} \partial_\tau V - \mathcal{L}_x V = -L, & (x,\tau) \in \mathbb{R} \times (0, T\,], \\ V_1(x,0) \ = (e^x - 1)^+, & x \in \mathbb{R}. \end{cases}$$

Applying the comparison principle to problems (4.9) and (4.10), we see that $V(x,\tau) \ge V_1(x,\tau)$. If we can prove that there exists a $\tau_0 > 0$ such that $V_1(0,\tau) > 0$ for $0 < \tau < \tau_0$, then the result of the theorem is an immediate result. To do this we define

$$V_1(x,\tau) = V_2(x,\tau) + \frac{L}{r}(e^{-r\tau} - 1).$$

Then $V_2(x,\tau)$ satisfies

$$\text{(4.11)} \qquad \begin{cases} \partial_\tau V_2 - \mathcal{L}_x V_2 = 0, & (x,\tau) \in \mathbb{R} \times (0, T\,], \\ V_2(x,0) \ = (e^x - 1)^+, & x \in \mathbb{R}. \end{cases}$$

This is a Cauchy problem which the price of the standard European call option satisfies. Its solution has an explicit formula (see [11]):

$$\text{(4.12)} \qquad V_2(x,\tau) = e^{x-q\tau} N(\widehat{d_1}) - e^{-r\tau} N(\widehat{d_2})\ ,$$

where

$$N(\widehat{d_1}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\widehat{d_1}} e^{-\eta^2/2} d\eta,$$

$$\widehat{d_1} = \frac{x + (r - q + \sigma^2/2)\tau}{\sigma\sqrt{\tau}}, \quad \widehat{d_2} = \frac{x + (r - q - \sigma^2/2)\tau}{\sigma\sqrt{\tau}}.$$

Now we check

$$\partial_\tau V_2(0,\tau) = -qe^{-q\tau}N(d_1) + e^{-q\tau}n(d_1)\frac{r-q+\sigma^2/2}{2\sigma\sqrt{\tau}} + re^{-r\tau}N(d_2)$$
$$- e^{-r\tau}n(d_2)\frac{r-q-\sigma^2/2}{2\sigma\sqrt{\tau}} \geq -q + e^{-q\tau}n(d_1)\frac{\sigma}{2\sqrt{\tau}},$$

where

$$n(d_1) = e^{-d_1^2/2}/\sqrt{2\pi}, \qquad d_1 = \frac{(r-q+\sigma^2/2)\sqrt{\tau}}{\sigma}, \qquad d_2 = \frac{(r-q-\sigma^2/2)\sqrt{\tau}}{\sigma},$$

and we had utilized the equality $e^{-q\tau}n(d_1) = e^{-r\tau}n(d_2)$. As $\tau \to 0^+$, we have

$$d_1 \to 0, \quad n(d_1) \to 1, \quad \partial_\tau V_2(0,\tau) \to +\infty.$$

It follows that $\partial_\tau V_1(0,\tau) \to +\infty$ as $\tau \to 0^+$. Since $V_1(0,0) = 0$ and $V_1 \in C(\mathbb{R}\times[\,0,T\,])$, then we see that there exists a $\tau_0 > 0$ such that $V_1(0,\tau) > 0$ for any $0 < \tau < \tau_0$. ∎

From Theorem 4.1 and Lemmas 4.2–4.4, we can conclude the following theorem.

THEOREM 4.5. (1) *If $q > 0$, then $x_s(\tau)$ is not monotonic; moreover, $x_s(\tau)$ changes its sign (see Figure* 5*) and*

$$\lim_{\tau\to+\infty}[x_s(\tau) - q\tau] = \ln\frac{2L}{2r+\sigma^2}.$$

(2) *If $q = 0$ and $r < L \leq r + \sigma^2/2$, then $x_s(\tau) \leq 0$ (see Figure* 6*) and*

$$\lim_{\tau\to+\infty} x_s(\tau) = \ln\frac{2L}{2r+\sigma^2}.$$

(3) *If $q = 0$ and $L > r + \sigma^2/2$, then $x_s(\tau)$ is not monotonic; moreover, $x_s(\tau)$ changes its sign (see Figure* 7*) and $x_s(\tau) \leq \ln\frac{2L}{2r+\sigma^2}$ with*

$$\lim_{\tau\to+\infty} x_s(\tau) = \ln\frac{2L}{2r+\sigma^2}.$$

Proof. (1) From Theorem 4.1, we see that $x_s(0) = 0$ and $x_s(\tau) \to +\infty$ as $\tau \to +\infty$; moreover, by Lemma 4.4, there exists a $\tau_0 > 0$ such that $x_s(\tau) < 0$ for any $0 < \tau < \tau_0$, and hence $x_s(\tau)$ is not monotonic and changes sign. Inequality (4.1) implies

$$\lim_{\tau\to+\infty}(x_s(\tau) - q\tau) = \ln\frac{2L}{2r+\sigma^2}.$$

(2) and (3) are easily obtained as above, so we omit their proofs. ∎

Remark on the monotonicity of $x_s(\tau)$ in (2) *of Theorem* 4.5. According to the result of asymptotic expansion of $S_s(\tau) = e^{x_s(\tau)}$ close to expire in (1.2), we know that

$$x_s(\tau) \sim -\sigma(-\tau\ln\tau)^{1/2}(1+o(1)); \quad x_s'(\tau) \sim -\infty \qquad \text{as } \tau = T - t \to 0^+.$$

So if τ is small enough, $x_s(\tau)$ is monotonic decreasing. Combining this result with (4.1), we guess that $x_s(\tau)$ should not be monotonic. But we cannot prove this conjecture.

5. The monotonicity of the free boundary $x_s(\ )$ with respect to the parameters r, q, and L. In this section, we have the following theorem.

THEOREM 5.1. *$x_s(\tau)$ is decreasing with respect to r and increasing with respect to q and L.*

Proof. We divide the proof into three steps.

Step 1. We prove

$$\partial_x V - V \geq 0. \tag{5.1}$$

For the proof of (5.1), we come back to the approximation problem (2.11) and denote

$$W^* = \partial_y v_{\varepsilon,n} - v_{\varepsilon,n} + 2\varepsilon.$$

We claim that $W^* \geq 0$; otherwise, we suppose W^* achieves its negative minimum at $(x_0, \tau_0) \in \overline{\Omega_T^n}$.

If n is large enough, from the definition of π_ε and the initial and boundary values in (2.11) and (2.13), we have

$$\begin{cases} W^*(-n,\tau) \geq 0 - 0 + 2\varepsilon > 0 & \text{(by (2.14))}, \\ W^*(n,\tau) \geq e^{n+(p^*-r)\tau} - e^{n+(p^*-r)\tau} + 2\varepsilon > 0 & \text{(by (2.12))}, \\ W^*(y,0) = \partial_y \pi_\varepsilon(e^y - 1) - \pi_\varepsilon(e^y - 1) + \varepsilon \geq 0 - \varepsilon + 2\varepsilon > 0 & \text{if } e^y - 1 \leq \varepsilon, \\ W^*(y,0) > e^y - (e^y - 1) + 2\varepsilon > 0 & \text{if } e^y - 1 > \varepsilon. \end{cases}$$

Hence, from (2.3), we deduce that

$$(x_0,\tau_0) \in \Omega_T^n \qquad \text{and} \qquad \partial_\tau W^*(x_0,\tau_0) - \mathcal{L}_y W^*(x_0,\tau_0) \leq 0. \tag{5.2}$$

On the other hand, (2.14) and $W^*(x_0,\tau_0) < 0$ imply $v_{\varepsilon,n}(x_0,\tau_0) > 2\varepsilon$. Then

$$\beta_\varepsilon(v_{\varepsilon,n}(x_0,\tau_0)) = 0, \qquad\qquad \beta'_\varepsilon(v_{\varepsilon,n}(x_0,\tau_0)) = 0.$$

Combining (2.11) and (2.13), we see that at the point (x_0,τ_0)

$$\partial_\tau W^* - \mathcal{L}_y W^* = L + 2r\varepsilon > 0,$$

which contradicts (5.2). Hence,

$$W^* \geq 0 \qquad \text{and} \qquad \partial_y v_{\varepsilon,n} - v_{\varepsilon,n} \geq -2\varepsilon.$$

Taking $\varepsilon \to 0^+$ and $n \to \infty$, we obtain

$$\partial_y v - v \geq 0.$$

Eventually, the transformation (2.1) implies (5.1).

Step 2. Suppose that V_1 is the solution to (1.4), where r is r_1, and that V_2 is the solution to (1.4), where r is r_2 and $r_1 > r_2$. By (5.1), we deduce $\partial_x V_1 - V_1 \geq 0$. Then V_1 satisfies

$$\begin{cases} \partial_\tau V_1 - \frac{\sigma^2}{2}\partial_{xx} V_1 - (r_2 - q - \frac{\sigma^2}{2})\partial_x V_1 + r_2 V_1 = -L + (r_1 - r_2)(\partial_x V_1 - V_1) \geq -L \\ \qquad\qquad \text{if } V_1 > 0 \quad \text{and} \quad (x,\tau) \in \mathbb{R} \times (0,T\,], \\ \partial_\tau V_1 - \frac{\sigma^2}{2}\partial_{xx} V_1 - (r_2 - q - \frac{\sigma^2}{2})\partial_x V_1 + r_2 V_1 \geq -L + (r_1 - r_2)(\partial_x V_1 - V_1) = -L \\ \qquad\qquad \text{if } V_1 = 0 \quad \text{and} \quad (x,\tau) \in \mathbb{R} \times (0,T\,], \\ V_1(x,0) = (e^x - 1)^+, \qquad\qquad x \in \mathbb{R}. \end{cases}$$

By the comparison principle for the solution of variational inequalities with respect to nonhomogeneous terms, we have $V_1 \geq V_2$. If we denote $\mathbf{NR_1}$, $\mathbf{NR_2}$ to be the nontransaction region of V_1, V_2, respectively, then $\mathbf{NR_1} \supset \mathbf{NR_2}$. Hence $x_s(\tau)$ is decreasing with respect to r.

Step 3. As in Step 2, from (2.21), we can deduce that $x_s(\tau)$ is increasing with respect to q and L. □

6. Numerical methods and results. Starting from problem (1.4), we have

$$\begin{cases} \min\{\ \partial_\tau V - \frac{\sigma^2}{2}\partial_{xx}V - (r-q-\frac{\sigma^2}{2})\partial_x V + rV + LV\ \} = 0, & x \in \mathbb{R},\ \tau \in (0,T\,], \\ V(x,0)\ = (e^x-1)^+, & x \in \mathbb{R}. \end{cases} \tag{6.1}$$

Given mesh size $\triangle x,\ \triangle\tau > 0,\ V_j^n = V(j\triangle x, n\triangle\tau)$ represents the value of numerical approximation at $(j\triangle x,\ n\triangle\tau)$. Then the PDE is changed into the following difference equation:

$$\begin{cases} \min\left\{ \frac{V_j^n - V_j^{n-1}}{\triangle\tau} - \frac{\sigma^2}{2}\frac{V_{j+1}^{n-1} - 2V_j^{n-1} + V_{j-1}^{n-1}}{\triangle x^2} \right. \\ \qquad \left. - (r-q-\frac{\sigma^2}{2})\frac{V_{j+1}^{n-1} - V_{j-1}^{n-1}}{2\triangle x} + rV_j^n + L, V_j^n \right\} = 0, \\ V_j^0\ = (e^{j\triangle x}-1)^+. \end{cases} \tag{6.2}$$

This means

$$\begin{cases} V_j^n = \max\left\{ \frac{1}{1+r\triangle\tau}[(1 - \frac{\sigma^2\triangle\tau}{\triangle x^2})V_j^{n-1} + \frac{\sigma^2 + (r-q-\sigma^2/2)\triangle x}{2\triangle x^2}\triangle\tau V_{j+1}^{n-1} \right. \\ \qquad\qquad \left. + \frac{\sigma^2 - (r-q-\sigma^2/2)\triangle x}{2\triangle x^2}\triangle\tau V_{j-1}^{n-1} - L\triangle\tau],\ 0 \right\}, \\ V_j^0\ = \max\{0,\ e^{j\triangle x} - 1\}. \end{cases} \tag{6.3}$$

Choosing $\sigma^2\triangle\tau(\triangle x)^{-2} = 1$, we have

$$\begin{cases} V_j^n = \max\left\{ \frac{1}{1+r\triangle\tau}[(\frac{1}{2} + \frac{r-q-\sigma^2/2}{2\sigma}\sqrt{\triangle\tau})V_{j+1}^{n-1} \right. \\ \qquad\qquad \left. + (\frac{1}{2} - \frac{r-q-\sigma^2/2}{2\sigma}\sqrt{\triangle\tau})V_{j-1}^{n-1} - L\triangle\tau], 0 \right\}, \\ V_j^0 = \max\{0,\ e^{j\triangle x} - 1\}. \end{cases} \tag{6.4}$$

Denote $u = e^{\sigma\sqrt{\triangle\tau}}$, $d = u^{-1}$, $\rho = e^{r\triangle\tau}$, $p = (\rho e^{-q\triangle\tau} - d)(u-d)^{-1}$.

Then it is clear, as $\triangle\tau \to 0^+$, that the following hold:

$$\frac{1}{2} + \frac{r-q-\sigma^2/2}{2\sigma}\sqrt{\triangle\tau} = p + o(\triangle\tau), \qquad \frac{1}{1+r\triangle\tau} = \frac{1}{\rho} + O(\triangle\tau^2).$$

Neglecting a higher order of $\sqrt{\triangle\tau}$, we obtain

$$\begin{cases} V_j^n = \max\{\ \frac{1}{\rho}[\ pV_{j+1}^{n-1} + (1-p)V_{j-1}^{n-1} - L\triangle\tau],\ 0\ \}, \\ V_j^0\ = \max\{\ 0,\ u^j - 1\}. \end{cases} \tag{6.5}$$

Consider the point $(x,\tau) = (j\triangle x, n\triangle\tau)$. Then

$$V_j^n = V(x,\tau), \quad V_j^{n-1} = V(x, \tau - \triangle\tau),$$
$$V_{j+1}^{n-1} = V(x+\triangle x, \tau - \triangle\tau), \quad V_{j-1}^{n-1} = V(x-\triangle x, \tau - \triangle\tau).$$

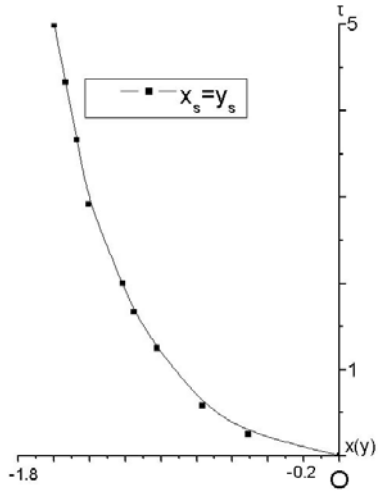


FIG. 8. $q = 0$ *and* $r \geq L$.

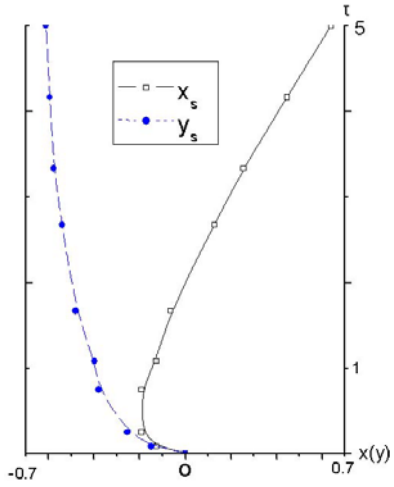


FIG. 9. $q > 0$.

Then we get Figures 8–11.

Figure 8 shows a plot of the optimal stop boundaries $x_s(\tau)$ $(y_s(\tau))$ as a function of time τ when $q = 0,\ r \geq L$. The parameter values used in the calculations are $r = 0.4,\ q = 0,\ \sigma = 0.7,\ L = 0.1,\ T = 5,\ n = 600$. $x_s(\tau),\ y_s(\tau)$ are the free boundaries of problems (1.4) and (2.2), respectively. Observe that $x_s(0) = y_s(\tau) = 0,\ -1.587 \leq x_s(\tau) = y_s(\tau) \leq 0$, and $x_s(\tau), y_s(\tau)$ are strictly decreasing with respect to τ. The numerical result is consistent with that of our proof (see Figure 8).

Figure 9 shows a plot of the optimal stop boundaries $x_s(\tau)$ $(y_s(\tau))$ as a function of time τ when $q > 0$. The parameter values used in the calculations are $r = 0.25,\ q = 0.2,\ \sigma = 0.7,\ L = 0.3,\ T = 5,\ n = 600$. $x_s(\tau),\ y_s(\tau)$ are the free boundaries of problems (1.4) and (2.2), respectively. Observe that $x_s(0) = y_s(\tau) = 0,\ x_s(\tau) \geq -0.1917$, and $x_s(\tau)$ is not monotonic and $y_s(\tau)$ is decreasing. The numerical result is consistent with that of our proof (see Figure 9).

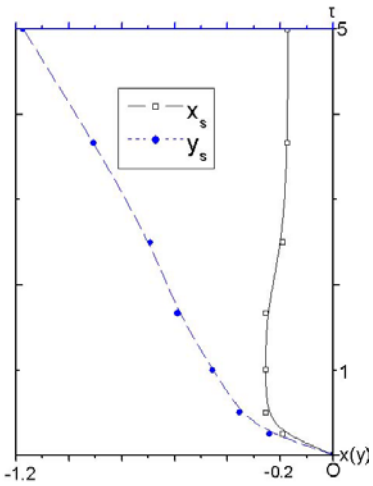


FIG. 10. $q = 0$ *and* $r < L \le r + \sigma^2/2$.

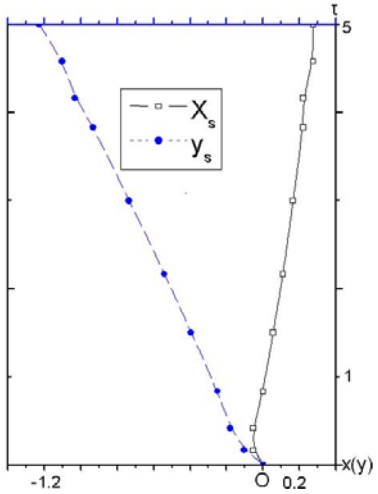


FIG. 11. $q = 0$ *and* $L > r + \sigma^2/2$.

Figure 10 shows a plot of the optimal stop boundaries $x_s(\tau)$ $(y_s(\tau))$ as a function of time τ when $q = 0$ and $r < L \le r + \sigma^2/2$. The parameter values used in the calculations are $r = 0.1,\ q = 0,\ \sigma = 0.7,\ L = 0.3,\ T = 5,\ n = 600$. $x_s(\tau),\ y_s(\tau)$ are the free boundaries of problems (1.4) and (2.2), respectively. Observe that $x_s(0) = y_s(\tau) = 0,\ -0.2556 \le x_s(\tau) \le -0.1278$, and $y_s(\tau)$ is decreasing. The numerical result is consistent with that of our proof (see Figure 10).

Figure 11 shows a plot of the optimal stop boundaries $x_s(\tau)$ $(y_s(\tau))$ as a function of time τ when $q = 0$ and $L > r+\sigma^2/2$. The parameter values used in the calculations are $r = 0.2,\ q = 0,\ \sigma = 0.6,\ L = 0.5,\ T = 5,\ n = 600$. $x_s(\tau),\ y_s(\tau)$ are the free boundaries of problems (1.4) and (2.2), respectively. Observe that $x_s(0) = y_s(\tau) = 0,\ -0.0548 \le x_s(\tau) \le 0.2739$, and $x_s(\tau)$ is not monotonic and $y_s(\tau)$ is decreasing. The numerical result is consistent with that of our proof (see Figure 11).

Appendix. Formulation of the model. A European call option is a contract which gives the owner the right but not the obligation to buy an asset at a fixed price K at the expiry date T. Because the holder of the option stands to make a profit without risking a loss, the holder must pay some premium for the option.

In a conventional European call option contract (see [2]), the holder pays the premium entirely up front and acquires the right. In a continuous installment European call option contract, the holder pays a smaller up-front premium and then a constant stream of installments at a certain rate per unit time. However, the holder can choose at any time to stop making installment payments by stopping the option contract.

There are some papers about install options, such as [7], [8], [1], [6]. Particularly, there are a variational inequality model in [7] and some numerical results about the model. In the following, we deduce a parabolic variational inequality model using their idea.

We consider a standard model for perfect market, continuous trading, no-arbitrage opportunity, a constant interest rate $r > 0$, and an asset with constant continuous dividend yield $q \geq 0$ with price S following a geometric Brownian motion

$$dS = \mu S dt + \sigma S dB, \tag{A.1}$$

where B is a Wiener process on a risk neutral probability space, $\mu > 0$ is the expected return rate, and $\sigma > 0$ is the constant volatility.

Let $C(S, t)$ denote the value of a European continuous installment call option and let L^* be the continuous install rate. Applying Itô's formula to $C(S, t)$ and combining (A.1), we obtain the dynamics

$$dC = \left(\frac{\partial C}{\partial t} + \frac{\sigma^2}{2} S^2 \frac{\partial C}{\partial S^2} + \mu S \frac{\partial C}{\partial S} \right) dt + \sigma S \frac{\partial C}{\partial S} dB. \tag{A.2}$$

We construct the $\triangle$-hedging portfolio consisting of one continuous installment option and an amount $-\triangle$ asset. The value of this portfolio is

$$\Pi = C - \triangle S. \tag{A.3}$$

Then its dynamics is given by

$$d\Pi = dC - \triangle S q dt - \triangle dS - L^* dt \leq r \Pi dt = r(C - \triangle S) dt. \tag{A.4}$$

Combining (A.1) and (A.3), we get

$$\begin{aligned} 0 &\geq d\Pi - r(C - \triangle S) dt \\ &= \left(\frac{\partial C}{\partial t} + \frac{\sigma^2}{2} S^2 \frac{\partial C}{\partial S^2} + \mu S \left(\frac{\partial C}{\partial S} - \triangle \right) + (r - q) \triangle S - rC - L^* \right) dt \\ &\quad + \sigma S \left(\frac{\partial C}{\partial S} - \triangle \right) dB. \end{aligned} \tag{A.5}$$

Setting $\triangle = \partial_S C$ the coefficient of dB vanishes. The portfolio is instantaneously riskless; then we see that $C(S, t)$ satisfies

$$\partial_t C + \frac{\sigma^2}{2} S^2 \partial_{SS} C + (r - q) S \partial_S C - rC \leq L^*. \tag{A.6}$$

Moreover, in the nontransaction region, $C(S, t)$ satisfies

$$\partial_t C + \frac{\sigma^2}{2} S^2 \partial_{SS} C + (r - q) S \partial_S C - rC = L^*. \tag{A.7}$$

At expiry date T, if $S > K$, the holder of the option will buy the asset with price S at K and can make a profit $S - K$. On the other hand, if $S \le K$, he will not exercise the option and stand any loss. Hence, the value of the option is $(S-K)^+$ at expiry date T, that is,

$$C(S, T) = (S - K)^+. \tag{A.8}$$

Because the owner must keep paying premiums to keep the option alive, if S is small enough at some t, the present value of the expected pay-off may be less than the present value of the remaining installments; then the holder would allow the option to lapse and stop paying installment payments. Hence, the value of the option is 0 in the case in which (S, τ) lies in the stop region.

It is clear that $C(S, t) \ge 0$ in nontransaction region. Otherwise, he can choose to stop the option for increased profit; hence we have

$$C(S, t) \ge 0. \tag{A.9}$$

From the above deduction, we see that the following equality holds in the stop region and the nontransaction region:

$$\left[\partial_t C + \frac{\sigma^2}{2} S^2 \partial_{SS} C + (r - q) S \partial_S C - rC - L^*\right] C(S, \tau) = 0. \tag{A.10}$$

Then we see that the value of the European continuous installment call option $C(S, t)$ satisfies (A.6), (A.8)–(A.10), that is, (1.1). Moreover, from the smooth fit conditions [12], we know that C, $\partial_S C$ are continuous.

REFERENCES

[1] G. Alobaidi, R. Mallier, and S. Deakin, *Laplace transforms and installment options*, Math. Models Methods Appl. Sci., 8 (2004), pp. 1167–1189.

[2] F. Black and M. Scholes, *The pricing of options and corporate liabilities*, J. Political Economy, 81 (1973), pp. 637–659.

[3] A. Blanchet, J. Dolbeault, and R. Monneau, *On the continuity of the time derivative of the solution to the parabolic obstacle problem with variable coefficients*, J. Math. Pures Appl. (9), 85 (2006), pp. 371–414.

[4] L. A. Caffarelli, *The regularity of free boundaries in higher dimensions*, J. Acta Math., 139 (1977), pp. 135–184.

[5] L. A. Caffarelli, A. Petrosyan, and H. Shahgholian, *Regularity of a free boundary in parabolic potential theory*, J. Amer. Math. Soc., 17 (2004), pp. 827–869.

[6] P. Ciurlia and I. Roko, *Valuation of American continuous-installment option*, Comput. Economics, 25 (2005), pp. 143–165.

[7] M. Davis, W. Schachermayer, and R. Tompkins, *Pricing, no-arbitrage bounds and robust hedging of installment options*, Quantitative Finance, 6 (2001), pp. 597–610.

[8] M. Davis, W. Schachermayer, and R. Tompkins, *Installment options and static hedging*, J. Risk Fin., 3 (2002), pp. 46–52.

[9] A. Friedman, *Parabolic variational inequalities in one space dimension and smoothness of the free boundary*, J. Funct. Anal., 18 (1975), pp. 151–176.

[10] A. Friedman, *Variational Principle and Free Boundary Problems*, John Wiley & Sons, New York, 1982.

[11] L. Jiang, *Mathematical Modeling and Methods of Option Pricing*, World Scientific, River Edge, NJ, 2005.

[12] R. C. Merton, *Theory of rational option pricing*, Bell J. Econom. and Management Sci., 4 (1973), pp. 141–184.

[13] K. Tao, *On Aleksandrov, Bakel'man type maximum principle for second order parabolic equations*, Comm. Partial Differential Equations, 10 (1985) pp. 543–553.

SIAM J. MATH. ANAL.
Vol. 40, No. 1, pp. 327–364

DETERMINATION OF THE SPECTRAL GAP IN THE KAC MODEL FOR PHYSICAL MOMENTUM AND ENERGY-CONSERVING COLLISIONS*

ERIC A. CARLEN†, JEFFREY S. GERONIMO†, AND MICHAEL LOSS†

Abstract. The Kac model describes the local evolution of a gas of N particles with *three*-dimensional velocities by a random walk in which the steps correspond to binary collisions that conserve momentum as well as energy. The state space of this walk is a sphere of dimension $3N-4$. The Kac conjecture concerns the spectral gap in the one-step transition operator Q for this walk. In this paper, we compute the exact spectral gap. As in previous work by Carlen, Carvalho, and Loss, where a lower bound on the spectral gap was proved, we use a method that relates the spectral properties of Q to the spectral properties of a simpler operator P, which is simply an average of certain noncommuting projections. The new feature is that we show how to use a knowledge of certain eigenfunctions and eigenvalues of P to determine spectral properties of Q, instead of simply using the spectral gap for P to bound the spectral gap for Q, inductively in N, as in previous work. The methods developed here can be applied to many other high-dimensional stochastic process, as we shall explain. We also use some deep results on Jacobi polynomials to obtain the required spectral information on P, and we show how the identity through which Jacobi polynomials enter our problem may be used to obtain new bounds on Jacobi polynomials.

Key words. Kac model, orthogonal polynomials, spectral gap

AMS subject classifications. 33C20, 82C40, 76P05

DOI. 10.1137/070695423

1. The Markov transition operator Q for the Kac walk.

Let X_N be the N particle state space consisting of N-tuples $\vec{v} = (v_1, \dots, v_N)$ of vectors v_j in $\mathbb{R}^3$ with

$$\sum_{j=1}^{N} |v_j|^2 = 1 \qquad \text{and} \qquad \sum_{j=1}^{N} v_j = 0.$$

We think of a point $\vec{v}$ as specifying the velocities of N particles and shall consider a random walk on X_N that was introduced by Kac [7]. At each step of this random walk, $\vec{v}$ is updated due to the effect of a binary collision that conserves energy and momentum—hence the constraints defining X_N.

To specify this walk in more detail, we consider a collision in which particles i and j collide. Suppose that v_i^* and v_j^* are the postcollisional velocities, while v_i and v_j are the precollisional velocities. Then by momentum conservation, the center of mass velocity is conserved; i.e.,

$$v_i^* + v_j^* = v_i + v_j.$$

Furthermore, by energy conservation, i.e., $|v_i^*|^2 + |v_j^*|^2 = |v_i|^2 + |v_j|^2$, and the parallelogram law, it follows that

$$|v_i^* - v_j^*| = |v_i - v_j|.$$

*Received by the editors July 18, 2007; accepted for publication (in revised form) December 21, 2007; published electronically April 23, 2008.

http://www.siam.org/journals/sima/40-1/69542.html

†School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 (carlen@math.gatech.edu, geronimo@math.gatech.edu, loss@math.gatech.edu). The first and third authors' work was partially supported by U.S. National Science Foundation grant DMS-060037. The second author's work was partially supported by U.S. National Science Foundation grant DMS-0500641.

This leads to a natural parameterization of all of the possible binary collision outcomes that conserve energy and momentum: The parameter σ is a unit vector in S^2, and, when particles i and j collide, one updates $\vec{v} \to \vec{v}^* = R_{i,j,\sigma}(\vec{v})$, where

$$\begin{aligned} v_i^* &= \frac{v_i + v_j}{2} + \frac{|v_i - v_j|}{2}\sigma, \\ v_j^* &= \frac{v_i + v_j}{2} - \frac{|v_i - v_j|}{2}\sigma, \\ v_k^* &= v_k \quad \text{for } k \neq i, j. \end{aligned} \tag{1.1}$$

The *Kac walk* on X_N is a random walk in which the steps are such binary collisions between pairs of particles. At each step, one picks a pair (i, j), $i < j$ uniformly at random, and also a unit vector σ in S^2. One then makes the update described in (1.1). Of course it remains to specify the probabilistic rule according to which σ should be selected. In the physics being modeled here, the likelihood of selecting a particular σ will depend only on the resulting *scattering angle* θ, which is the angle between $v_i^* - v_j^*$ and $v_i - v_j$. In the parameterization above, this is the angle between σ and $v_i - v_j$. That is,

$$\cos(\theta) = \sigma \cdot \frac{v_i - v_j}{|v_i - v_j|}.$$

The *scattering rate function* b is a nonnegative integrable function on $[-1, 1]$ with

$$\frac{1}{2}\int_{-1}^{1} b(u)\mathrm{d}u = 1.$$

Then for any $v_i \neq v_j$, and with $\mathrm{d}\sigma$ being the uniform probability measure on S^2,

$$\int_{S^2} b\left(\sigma \cdot \frac{v_i - v_j}{|v_i - v_j|}\right)\mathrm{d}\sigma = 1. \tag{1.2}$$

(If $v_i = v_j$, the collision has no effect and can be ignored.) One selects σ according to the probability density that is integrated in (1.2).

There are several choices of b of particular interest. One is the *uniform redirection model*, given by $b(x) = 1$ for all $-1 \leq x \leq 1$. In this case, the new direction of the relative velocity σ is chosen uniformly from S^2.

Another is the *Morgenstern model* [10], [11], or the *uniform reflection model*: For any unit vector $\omega \in S^2$, let H_ω be the reflection given by

$$H_\omega(v) = v - 2(v \cdot \omega)\omega.$$

In the uniform reflection model, one updates the relative velocity according to

$$v_i - v_j \to H_\omega(v_i - v_j) = v_i^* - v_j^*,$$

with ω chosen uniformly. The relation between ω and σ is given by $\sigma = H_\omega((v_i - v_j)/|v_i - v_j|)$, and, by computing the Jacobian of the map $\omega \to \sigma$, one finds

$$b(x) = \frac{1}{\sqrt{2}\sqrt{1 - x}}.$$

Both of these belong to the one-parameter family

$$b_\alpha(x) = (1-\alpha)2^\alpha(1-x)^{-\alpha}. \tag{1.3}$$

By leaving the particular choice of b open, this completes the specification of the steps in the Kac walk. For more detail and background, see [7] and [3].

The main object of study here is the spectrum of the one-step transition operator Q for this random walk, and the manner in which this spectrum depends on N. Q is defined as follows: Let $\vec{v}_n$ be the state of the process after the nth step. The one-step Markov transition operator Q is given by taking the conditional expectation

$$Q\phi(\vec{v}) = \mathrm{E}(\phi(\vec{v}_{n+1}) \mid \vec{v}_n = \vec{v})$$

for any continuous function ϕ on X_N.

From the above description, one deduces the formula

$$Q\phi(\vec{v}) = \begin{pmatrix} N \\ 2 \end{pmatrix}^{-1} \sum_{i<j} \int_{S^2} \phi(R_{i,j,\sigma}(\vec{v}))b\left(\sigma \cdot \frac{v_i - v_j}{|v_i - v_j|}\right) \mathrm{d}\sigma. \tag{1.4}$$

Let σ_N denote the uniform probability measure on X_N, which is the normalized measure induced on X_N as a manifold embedded in $\mathbb{R}^{3N}$.

For any two unit vectors σ and ω, one sees from (1.1) that

$$R_{i,j,\sigma}(R_{i,j,\omega}\vec{v}) = R_{i,j,\sigma}\vec{v}.$$

From this and the fact that the measure $\mathrm{d}\sigma_N$ is invariant under $\vec{v} \mapsto R_{i,j,\sigma}\vec{v}$, it follows that, for any continuous functions ϕ and ψ on X_N,

$$\int_{X_N} \psi(\vec{v})Q\phi(\vec{v})\mathrm{d}\sigma_N = \int_{X_N}\int_{S^2}\int_{S^2} \psi(R_{i,j,\omega}\vec{v})\phi(R_{i,j,\sigma}\vec{v})b(\omega\cdot\sigma)\mathrm{d}\omega\mathrm{d}\sigma\mathrm{d}\sigma_N.$$

It follows that Q is a self-adjoint Markov operator on $L^2(X_N, \sigma_N)$. Moreover, it is clearly a Markov operator; that is, in addition to being self-adjoint, Q is positivity preserving and $Q1 = 1$.

The motivation for considering the spectral properties of Q stems from a theorem of Kac [7] that relates the continuous time version of the Kac walk to the nonlinear Boltzmann equation. For the details, see [7] or [3]. Let $\vec{v}(t)$ denote the random variable giving the state of the system at time t for the process run in continuous time with the jumps taking place in a Poisson stream with the mean time between jumps being $1/N$. Then the equation describing the evolution of the probability law of $\vec{v}(t)$ is called the *Kac master equation*: If the initial law on X_N has a density F_0, then the law at time t has a density $F(\vec{v}, t)$ satisfying

$$\frac{\partial}{\partial t}F(\vec{v},t) = N(Q-I)F(\vec{v},t), \qquad \text{with} \qquad F(\vec{v},0) = F_0(\vec{v}).$$

The solution $F(\vec{v}, t)$ is of course given by

$$F(\vec{v},t) = e^{t\mathcal{L}}F_0(\vec{v}),$$

where

$$\mathcal{L} = N(Q-I).$$

Since Q is a self-adjoint Markov operator, its spectrum lies in the interval $[-1,1]$, and since $Q1 = 1$, the constant function is an eigenfunction with eigenvalue 1. It is easily seen that, as long as $b(x)$ is strictly positive on a neighborhood of $x = 1$, the eigenvalue 1 of Q has multiplicity one. It then follows that the spectrum of $\mathcal{L}$ lies in $[-2N, 0]$ and that 0 is an eigenvalue of multiplicity one. We impose this assumption on b throughout what follows.

The Kac conjecture for this stochastic process pertains to the spectral gap

$$\Delta_N = \inf\left\{ -\int_{X_N} \phi(\vec{v})\mathcal{L}\phi(\vec{v})\mathrm{d}\sigma_N \;\middle|\; \int_{X_N} \phi^2(\vec{v})\mathrm{d}\sigma_N = 1,\ \int_{X_N} \phi(\vec{v})\mathrm{d}\sigma_N = 0 \right\}$$

and states that

$$\liminf_{N\to\infty} \Delta_N > 0.$$

This was proved by Carlen, Carvalho, and Loss [3], but without an explicit lower bound. Kac also made a similar conjecture for a simplified model with one-dimensional velocities and no conservation of momentum. For this model, the conjecture was first proved by Janvresse [6], though her approach provided no explicit lower bound. The sharp bound for the simplified model was first established in [2]. See Maslen [9] for a representation theoretic approach.

The main goal of the present paper is to compute *exactly* $\liminf_{N\to\infty} \Delta_N$. We shall be able to do this under an easily checked condition relating Δ_2 and the quantities

$$B_1 = \frac{1}{2}\int_{-1}^{1} xb(x)\mathrm{d}x \qquad \text{and} \qquad B_2 = \frac{1}{2}\int_{-1}^{1} x^2 b(x)\mathrm{d}x. \tag{1.5}$$

The condition, given in (1.6) below, will turn out to be satisfied when b is given by b_α, as in (1.3), for all $0 \le \alpha \le 7/9$.

THEOREM 1.1. *Suppose that $B_2 > B_1$ and that*

$$\Delta_2 \ge \frac{20}{9}(1 - B_2). \tag{1.6}$$

Then for all $N \ge 3$,

$$\Delta_N = (1 - B_2)\frac{N}{(N-1)}. \tag{1.7}$$

Moreover, the eigenspace is three-dimensional and is spanned by the functions

$$\phi(\vec{v}) = \sum_{j=1}^{N} |v_j|^2 v_j^\alpha, \qquad \alpha = 1, 2, 3, \tag{1.8}$$

where v_j^α denotes the αth component of v_j.

As we shall see in the next section, for many choices of b, including the b_α, there is a simple monotonicity of the eigenvalues of Q for $N = 2$ which ensures that the eigenfunction providing the gap comes from a first degree polynomial and thus that

$$\Delta_2 = 2(1 - B_1). \tag{1.9}$$

When (1.9) is satisfied, the condition (1.6) reduces to $(1 - B_1)/(1 - B_2) > 10/9$.

Next, notice that the eigenfunctions listed in (1.8) are symmetric under permutation of the particle indices. Indeed, the operator Q commutes with such permutations, so that the subspace of functions with this symmetry is invariant. As explained in [7] and [3], it is the spectrum of Q on this subspace that is relevant for the study of the Boltzmann equation.

Moreover, notice that, in the collision rules (1.1), exchanging v_i^* and v_j^* has the same effect as changing σ to $-\sigma$. For this reason, if one's primary object of interest is the Boltzmann equation, one may freely assume that b is a symmetric function on $[-1,1]$, since then replacing $b(x)$ by $(b(x)+b(-x))/2$ will have no effect on the spectrum of Q on the invariant subspace of symmetric functions or on the corresponding Boltzmann equation. (See the introduction of [4] for more discussion of this point in the context of the Boltzmann equation.) If B is symmetric, then $B_1 = 0$, and we do have $B_1 < B_2$.

However, it is interesting that the Kac conjecture holds without restriction to the symmetric subspace and that the methods developed here can be used to determine the spectral gap even when b is not symmetric, and the eigenfunctions corresponding to the gap eigenvalue are not symmetric.

When b is not symmetric, it may happen that $B_1 \le B_2$. We shall give examples of this below. The next theorem gives the spectral gap and the eigenfunctions whenever $\Delta_2 = 2(1-B_1)$, regardless of whether $B_1 < B_2$ or $B_2 < B_1$. However, it gives the exact value of Δ_N only for $N \ge 7$. Since we are interested in large values of N, this is fully satisfactory. Indeed, it is remarkable that the two theorems show that already at relatively small values of N, the behavior of the system is very close, qualitatively and quantitatively to the behavior in the large N limit.

THEOREM 1.2. *Suppose that $\Delta_2 = 2(1-B_1)$. Then, for all $N \ge 7$,*

$$\Delta_N = \min\{\ (1-B_1),\ (1-B_2)\ \}\frac{N}{(N-1)}. \tag{1.10}$$

Moreover, if $B_2 > B_1$, the eigenspace is three-dimensional and is spanned by the functions

$$\phi(\vec{v}) = \sum_{j=1}^{N} |v_j|^2 v_j^\alpha, \qquad \alpha = 1,2,3, \tag{1.11}$$

where v_j^α denotes the αth component of v_j.

On the other hand, if $B_2 < B_1$, the eigenspace is spanned by the functions of the form

$$|v_i|^2 - |v_j|^2 \qquad \textit{and} \qquad v_i^\alpha - v_j^\alpha, \quad \alpha = 1,2,3, \tag{1.12}$$

for all $i < j$.

Finally, if $B_1 = B_2$, the eigenspace is spanned by both of the sets of functions listed in (1.8) *and* (1.12) *together.*

For the family of collision rates introduced so far, the b_α, one may apply Theorem 1.1, as we have indicated, but only for $\alpha \le 7/9$. As we shall see in section 2, Theorem 1.2 applies for all $0 \le \alpha < 1$ and in this case gives $\Delta_N = (N/N-1)(1-B_2)$ for $N \ge 7$. In order to illustrate the case in which Theorem 1.2 yields the gap $\Delta_N = (N/N-1)(1-B_1)$, we introduce

$$\widetilde{b}_\alpha(x) = 2(\alpha+1)1_{[0,1]}(x)x^\alpha, \qquad \alpha \ge 0. \tag{1.13}$$

Since $x^2 < x$ on $(0,1)$, it is clear that $B_2 < B_1$ for all α in this case. We show at the end of section 2 that, at least for $0 \le \alpha \le 1$, $\Delta_2 = 2(1 - B_1)$, so that Theorem 1.2 applies in these cases.

The method of proof is quite robust, and in section 10 we shall describe how it may be extended to determine the spectral gap of Q for still other choices of b that are not covered by Theorems 1.1 and 1.2.

The method of proof of these theorems relies on a basic strategy introduced in [3] but which is extended significantly here. The strategy consists of exploiting an inductive link between the spectral gap of Q and the one of an operator P, an average over projections introduced in section 3. In fact,

$$\Delta_N \ge \frac{N}{N-1}(1-\mu_N)\Delta_{N-1}, \tag{1.14}$$

where $1 - \mu_N$ is the gap of P. The eigenvalues of P are much easier to compute than the ones of Q since the range of P consists of sums of functions of single variables v_j.

In the case of the original model treated by Kac, one is in the happy circumstance that Q has a single gap eigenfunction ϕ which is also the gap eigenfunction of P for all N, *and*, when this is used as a trial function in the derivation of (1.14), one sees that (1.14) actually holds with equality, giving an identity relating Δ_N and Δ_{N-1}. Thus, starting at $N = 2$, where the gap can be easily calculated, the above formula yields a lower bound on Δ_N that turns out to be exact. The model treated in this paper does not have this simplifying feature, even when the gap eigenfunctions of Q are also the gap eigenfunctions of P. Nevertheless, the ideas that lead to (1.14) can be used in such a way that we can still calculate the gap of Q exactly. Very briefly, here is how.

Let $\mu_N^* < \mu_N$ be any number, and assume that there are finitely many eigenvalues $\mu_N^* \le \mu_N^{(m)} \le \cdots \le \mu_N^{(1)} \le \mu_N$ of P. Denote the corresponding eigenspaces by E_j. Let V_j be the smallest invariant subspace of Q that contains E_j. Lemma 4.1 in section 4 provides us with the following dichotomy: *Either*

$$\Delta_N \ge \frac{N}{N-1}(1-\mu_N^*)\Delta_{N-1} \tag{1.15}$$

or else

$$\text{the gap of } Q \text{ is the same as the gap of } Q \text{ restricted to } \oplus_{j=1}^m V_j. \tag{1.16}$$

If the threshold has been chosen so that the lower bound on Δ_N provided by (1.15) is at least as large as the upper bound on Δ_N provided by some trial function in $\oplus_{j=1}^m V_j$, then Δ_N is the gap of Q restricted to $\oplus_{j=1}^m V_j$. As we shall see, the V_j are finite-dimensional, so determining the gap of Q on $\oplus_{j=1}^m V_j$ is a tractable problem. In this case we have determined the exact value of Δ_N.

To proceed to the determination of Δ_N for all large N, one needs a strategy for choosing the threshold $\mu_N^\star$. The lower the value of $\mu_N^\star$ that is chosen, the stronger the bound (1.15) will be, but also the higher the value of m will be. The basis for the choice of $\mu_N^\star$ is a trial function calculation, providing a guess $\widetilde{\Delta}_N$ for the value of Δ_N. Indeed, natural trial functions can often be chosen on the basis of physical considerations. (The spectrum of the linearized Boltzmann equation is the source in the case at hand.) To show that the guess is correct, so that $\widetilde{\Delta}_N = \Delta_N$, we are led to choose $\mu_N^\star$ so that

$$\widetilde{\Delta}_N \le \frac{N}{N-1}(1-\mu_N^*)\widetilde{\Delta}_{N-1}. \tag{1.17}$$

Since $\widetilde{\Delta}_{N-1} \geq \Delta_{N-1}$, this forces us into the second alternative in the dichotomy discussed above, so that the gap eigenfunction for N particles lies in $\oplus_{j=1}^{m} V_j$. Indeed, if the physical intuition behind the guess was correct, the trial function leading to $\widetilde{\Delta}_N$ will lie in $\oplus_{j=1}^{m} V_j$ and yield the gap.

Choosing $\mu_N^\star$ small enough that (1.17) is satisfied might in principle lead to a value of m that depends on N. However, in the case at hand, we are fortunate and can work with a choice of $\mu_N^\star$ that leads to a fixed and small value of m but for which (1.17) is satisfied for all sufficiently large values of N—hence the restriction to $N \geq 7$ in Theorem 1.2.

As will be clear from this summary of the strategy, the determination of the spectrum of P is the main technical step that must be accomplished. As we mentioned before, this is relatively simple, compared to the determination of the spectrum of Q, since the range of P consists of functions that are a sum of functions of a single variable.

For this reason, we can reduce the study of the spectrum of P to that of a much simpler Markov operator K acting on functions on the unit ball B in $\mathbb{R}^3$. In the analysis of K, we shall draw on some deep results on Jacobi polynomials [8], [12]. In fact, it turns out that the connection between our eigenvalue problems and pointwise bounds on Jacobi polynomials is through a simple identity, and applications of this identity can be made in both directions: We not only use bounds on Jacobi polynomials to bound eigenvalues, we shall use simple eigenvalue estimates to sharpen certain best known bounds on Jacobi polynomials, as we briefly discuss in section 11.

First, however, we deal with a simpler technical problem, the computation of the spectral gap of Q for $N = 2$.

2. The spectral gap for $N = 2$. For $N = 2$, the state space X_2 consists of pairs $(v, -v)$, with $v \in \mathbb{R}^3$ satisfying $|v|^2 = 1/2$. Note that for $N = 2$ the collision rules (1.1) reduce to

$$v_1^* = \sigma/\sqrt{2} \qquad \text{and} \qquad v_2^* = -\sigma/\sqrt{2},$$

since $v_1 + v_2 = 0$.

The map $(v, -v) \mapsto \sqrt{2}v$ identifies X_2 with the unit sphere S^2, and the measure $\mathrm{d}\sigma_2$ on X_2 with $\mathrm{d}\sigma$ on S^2. Thus, we may think of Q as operating on functions on S^2. In this representation, we have the formula

$$Q\phi(u) = \int_{S^2} \phi(\sigma) b(u \cdot \sigma) \mathrm{d}\sigma.$$

Notice that if R is any rotation of $\mathbb{R}^3$

$$\begin{aligned}(Q\phi)(Ru) &= \int_{S^2} \phi(\sigma) b(Ru \cdot \sigma) \mathrm{d}\sigma \\ &= \int_{S^2} \phi(R\sigma) b(Ru \cdot R\sigma) \mathrm{d}\sigma \\ &= \int_{S^2} \phi(R\sigma) b(u \cdot \sigma) \mathrm{d}\sigma = Q(\phi \circ R)(u).\end{aligned}$$

That is, $(Q\phi) \circ R = Q(\phi \circ R)$, and this means that for each n the space of spherical harmonics of degree n is an invariant subspace of Q, contained in an eigenspace of Q. In turn, this means that we can determine the spectrum of Q by computing its

action on the zonal spherical harmonics, i.e., those of the form $P_n(e \cdot u)$, where e is any fixed unit vector in $\mathbb{R}^3$ and P_n is the nth degree Legendre polynomial. Now, for any function $\phi(u)$ of the form $\phi(u) = f(e \cdot u)$,

$$Q\phi(u) = \int_{S^2} \phi(\sigma \cdot e) b(\sigma \cdot u) \mathrm{d}\sigma.$$

We choose coordinates in which u and e span the x, z plane with

$$u = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \qquad \text{and} \qquad e = \begin{bmatrix} \sqrt{1-t^2} \\ 0 \\ t \end{bmatrix},$$

so that $t = u \cdot e$. Then with

$$\sigma = \begin{bmatrix} \sin\theta\cos\psi \\ \sin\theta\sin\psi \\ \sin\theta \end{bmatrix},$$

$$Q\phi(u) = \mathcal{Q}f(e \cdot u),$$

where

$$\begin{aligned} \mathcal{Q}f(t) &= \frac{1}{4\pi} \int_0^\pi \int_0^{2\pi} f(t\cos\theta + \sqrt{1-t^2}\sin\theta\cos\varphi) b(\cos\theta) \sin\theta \mathrm{d}\theta \mathrm{d}\varphi \\ &= \frac{1}{4\pi} \int_0^\pi \int_{-1}^1 f(ts + \sqrt{1-t^2}\sqrt{1-s^2}\cos\varphi) b(s) \mathrm{d}s \mathrm{d}\varphi. \end{aligned} \tag{2.1}$$

Now, if f is any eigenfunction of $\mathcal{Q}$, with $\mathcal{Q}f = \lambda f$, then by evaluating both sides at $t = 1$, we have $\lambda f(1) = \frac{1}{2}\int_0^\pi \int_{-1}^1 f(s) b(s) \mathrm{d}s$. Thus, the eigenvalue is given by

$$\lambda = \frac{1}{2} \int_{-1}^1 \frac{f(s)}{f(1)} b(s) \mathrm{d}s.$$

As we have observed above, the eigenfunctions of $\mathcal{Q}$ are the Legendre polynomials. Thus, if P_n is the Legendre polynomial of nth degree with the standard normalization $P_n(1) = 1$, and λ_n is the corresponding eigenvalue,

$$\lambda_n = \frac{1}{2} \int_{-1}^1 P_n(s) b(s) \mathrm{d}s. \tag{2.2}$$

This explicit formula enables one to easily compute Δ_2. For example, we can now easily prove the following.

LEMMA 2.1. *When $b(x) = b_\alpha(x)$, as in* (1.3)*, then $1 - B_2 < 1 - B_1$ for all $\alpha < 1$, and moreover*

$$\Delta_2 = 2(1 - \lambda_1) = \frac{4(1-\alpha)}{2-\alpha} = (1 - B_2)(3 - \alpha), \tag{2.3}$$

so that (1.6) *is satisfied for all α with $0 \le \alpha \le 7/9$.*

Proof. By using Rodrigues's formula [13, equation 4.3.1]

$$P_n(x) = \frac{1}{2^n n!} \frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^2 - 1)$$

and integration by parts, one computes

$$\lambda_n = (1-\alpha)\frac{(\alpha)_n}{(1-\alpha)_{n+1}} = \frac{(\alpha)_n}{(2-\alpha)_n},$$

where $(\alpha)_n = \alpha(\alpha+1)(\alpha+2)\dots(\alpha+n-1)$. Notice that, for all $0 \le \alpha < 1$, λ_n decreases as n increases, so with the collision rate given by b_α,

$$\Delta_2 = 2(1-\lambda_1) = \frac{4(1-\alpha)}{2-\alpha}. \tag{2.4}$$

Next, one computes

$$1 - B_1 = \frac{2(1-\alpha)}{(2-\alpha)} \qquad \text{and} \qquad 1 - B_2 = \frac{4(1-\alpha)}{(2-\alpha)(3-\alpha)}.$$

Since $2 > 4/(3-\alpha)$ for $\alpha < 1$, $1 - B_2 < 1 - B_1$ for all $\alpha < 1$. Moreover, from this computation, one readily obtains (2.3) and the statement concerning (1.6). □

In particular, the condition (1.6) is satisfied in both the uniform redirection model ($\alpha = 0$) and the Morgenstern model ($\alpha = 1/2$). Thus in these cases we have the exact spectral gaps

$$\begin{aligned} \Delta_N &= \frac{2}{3}\frac{N}{N-1} \qquad \text{for the uniform redirection model,} \\ \Delta_N &= \frac{8}{15}\frac{N}{N-1} \qquad \text{for the Morgenstern model.} \end{aligned} \tag{2.5}$$

We close this section with a remark that may provide a useful perspective on what follows. In determining the spectral gap of Q for $N = 2$, general symmetry conditions told us right away what all of the eigenfunctions were. A less obvious, though still simple, argument then provided us with the explicit formula (2.2) for all of the eigenvalues. There is one last hurdle to cross: There are infinitely many eigenvalues given by (2.2), and for a general b, we cannot determine which is the second largest by computing them all explicitly. What was particularly nice about b_α is that in this case the eigenvalues of Q were monotone-decreasing:

$$\lambda_{n+1} \le \lambda_n.$$

For other choices of b, this need not be the case. However, there are ways to use pointwise bounds on Legendre polynomials to reduce the problem of determining Δ_2 to the computation of a *finite* number of eigenvalues by using (2.2). For example, one has the classical bound (see [13, Theorem 7.3.3]):

$$|P_n(x)|^2 < \frac{2}{n\pi}\frac{1}{\sqrt{1-x^2}}. \tag{2.6}$$

As long as $b(x)(1-x^2)^{-1/4}$ is integrable, this gives an upper bound on λ_n that is proportional to $n^{-1/2}$: Define

$$\tilde{\lambda}_n = \left(\frac{1}{8\pi n}\right)^{1/2} \int_{-1}^{1} b(x)(1-x^2)^{-1/4}\mathrm{d}x.$$

Then let n_0 be the least value of n such that $\tilde{\lambda}_n \le \lambda_1$. Then the second largest eigenvalue of Q is

$$\max_{1 \le n \le n_0} \lambda_n.$$

We illustrate this by showing that, for the rate function $\widetilde{b}_\alpha$ introduced in (1.13), $\Delta_2 = 2(1 - B_1)$ at least for $0 \le \alpha \le 1$. (Of course, the integrals in (2.2) can be computed exactly in this case; see [5, section 7.231, page 822]. However, we prefer to illustrate the use of (2.6).)

By (2.6) and (2.2),

$$\begin{aligned} |\lambda_n| &\le (\alpha+1)\left(\int_0^1 x^{2\alpha}\mathrm{d}x\right)^{1/2}\left(\int_0^1 P_n(x)^2\mathrm{d}x\right)^{1/2} \\ &< \frac{\alpha+1}{\sqrt{2\alpha+1}}\frac{1}{\sqrt{n}}. \end{aligned} \tag{2.7}$$

Also, by (2.2), $\lambda_1 = B_1 = (\alpha+1)/(\alpha+2)$. Comparison of the formulas shows that, for $0 \le \alpha \le 1$, $\lambda_n < \lambda_1$ for all $n > 4$. Thus it suffices to check that $\lambda_j < \lambda_1$ for $j = 2, 3$, and 4 by direct computation with (2.2). By doing so, one finds that this is the case. Hence, Theorem 1.2 applies and yields $\Delta_N = (N/N-1)(1-B_1)$ for $N \ge 7$.

Further calculation would extend this result to higher values of α. Notice that, as α tends to infinity, $\widetilde{b}_\alpha(x)$ is more and more concentrated at $x = 1$, which corresponds to $\theta = 0$. Thus, for large values of α, $\widetilde{b}_\alpha$ represents a "grazing collision model."

For $N > 2$, the operator Q is much more complicated, and direct determination of the spectrum is not feasible. Instead, we use an inductive procedure involving an auxiliary operator that we now introduce.

3. The average of projections operator P and its relation to Q. A simple convexity argument shows that, for each j,

$$\sup\{|v_j|^2 \ : \ \vec{v} \in X_N\} = \frac{N-1}{N}.$$

For each j, define $\pi_j(\vec{v})$ by

$$\pi_j(\vec{v}) = \sqrt{\frac{N}{N-1}}v_j,$$

so that π_j maps X_N onto the unit ball B in $\mathbb{R}^3$.

For any function ϕ in $L^2(X_N, \mathrm{d}\sigma_N)$, and any j with $1 \le j \le N$, define $P_j(\phi)$ to be the orthogonal projection of ϕ onto the subspace of $L^2(X_N, \mathrm{d}\sigma_N)$ consisting of square integrable functions that depend on $\vec{v}$ through v_j alone. That is, $P_j(\phi)$ is the unique element of $L^2(X_N, \mathrm{d}\sigma_N)$ of the form $f(\pi_j(\vec{v}))$ such that

$$\int_{X_N} \phi(\vec{v})g(\pi_j(\vec{v}))\mathrm{d}\sigma_N = \int_{X_N} f(\pi_j(\vec{v}))g(\pi_j(\vec{v}))\mathrm{d}\sigma_N$$

for all continuous functions g on B.

The *average of projections operator* P is then defined through

$$P = \frac{1}{N}\sum_{j=1}^{N} P_j.$$

If the individual projections P_j all commuted with one another, then the spectrum of P would be very simple: The eigenvalues of each P_j are 0 and 1. Moreover, $P_j\phi = \phi$ if and only if ϕ depends only on v_j so that it cannot then also satisfy $P_k\phi = \phi$ for

$k \neq j$, unless ϕ is constant. It would then follow that the eigenvalues of P would be 0, $1/N$, and 1, with the last having multiplicity one.

However, the individual projections P_j do not commute with one another, due to the nature of the constraints defining X_N.

We now define

$$\mu_N = \sup\left\{ \int_{X_N} \phi(\vec{v}) P\phi(\vec{v}) \mathrm{d}\sigma_N \;\middle|\; \int_{X_N} \phi^2(\vec{v}) \mathrm{d}\sigma_N = 1,\ \int_{X_N} \phi(\vec{v}) \mathrm{d}\sigma_N = 0 \right\}. \tag{3.1}$$

The P operator is simpler than the Q operator in that if ϕ is any eigenfunction of P with a nonzero eigenvalue, then clearly ϕ has the form

$$\phi = \sum_{j=1}^{N} f_j \circ \pi_j$$

for some functions $f_1, \ldots, f_N$ on B. For $N \geq 4$, most of the eigenfunctions of Q have a more complicated structure. Nonetheless, there is a close relation between the spectra of Q and P, as we now explain.

To do this, we need a more explicit formula for P, such as the formula (1.4) that we have for Q. The key to computing $P_j\phi$ is a factorization formula [3] for the measure σ_N. Define a map $T_N : X_{N-1} \times B \to X_N$ as follows:

$$T_N(\vec{y}, v) = \left(\alpha(v) y_1 - \frac{1}{\sqrt{N^2 - N}} v, \ldots, \alpha(v) y_{N-1} - \frac{1}{\sqrt{N^2 - N}} v, \sqrt{\frac{N-1}{N}} v \right), \tag{3.2}$$

where

$$\alpha^2(v) = 1 - |v|^2.$$

This map induces coordinates $(\vec{y}, v)$ on X_N, and, in terms of these coordinates, one has the integral factorization formula

$$\int_{X_N} \phi(\vec{v}) \mathrm{d}\sigma_N = \frac{|S^{3N-7}|}{|S^{3N-4}|} \int_B \left[\int_{X_{N-1}} \phi(T_N(\vec{y}, v)) \mathrm{d}\sigma_{N-1} \right] (1 - |v|^2)^{(3N-8)/2} \mathrm{d}v.$$

It follows from this and the definition of P_N that

$$P_N \phi(\vec{v}) = f \circ \pi_N(\vec{v}),$$

where

$$f(v) = \int_{X_{N-1}} \phi(T_N(\vec{y}, v)) \mathrm{d}\sigma_{N-1}.$$

For $j < N$, one has analogous formulas for T_j and P_j, except the roles of v_N and v_j are interchanged.

Next, we make the definition for Q that is analogous to (3.1) for P: Define λ_N by

$$\lambda_N = \sup\left\{ \int_{X_N} \phi(\vec{v}) Q\phi(\vec{v}) \mathrm{d}\sigma_N \;\middle|\; \int_{X_N} \phi^2(\vec{v}) \mathrm{d}\sigma_N = 1,\ \int_{X_N} \phi(\vec{v}) \mathrm{d}\sigma_N = 0 \right\}. \tag{3.3}$$

With this explicit formula in hand, and the definitions of μ_N and λ_N, we come to the fundamental fact relating P and Q.

LEMMA 3.1. *For any square integrable function ϕ on X_N that is orthogonal to the constants,*

$$\langle \phi, Q\phi \rangle \leq \lambda_{N-1} \|\phi\|_2^2 + (1 - \lambda_{N-1}) \langle \phi, P\phi \rangle, \tag{3.4}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product on $L^2(X_N, \sigma_N)$.

Proof. To bound $\langle \phi, Q\phi \rangle$ in terms of λ_{N-1}, define, for $1 \leq k \leq N$, the operator $Q^{(k)}$ by

$$Q^{(k)}\phi(\vec{v}) = \begin{pmatrix} N-1 \\ 2 \end{pmatrix}^{-1} \sum_{i<j, i\neq k, j\neq k} \int_{S^2} \phi(R_{i,j,\sigma}(\vec{v})) \mathrm{d}\sigma.$$

That is, we leave out collisions involving the kth particle and average over the rest. Clearly,

$$Q = \frac{1}{N} \sum_{k=1}^{N} Q^{(k)}.$$

Therefore, for any ϕ in $L^2(X_N, \sigma_N)$,

$$\langle \phi, Q\phi \rangle = \frac{1}{N} \sum_{k=1}^{N} \langle \phi, Q^{(k)}\phi \rangle.$$

By using the coordinates $(\vec{y}, v)$ induced by the map $T_k : X_{N-1} \times B \to X_N$, it is easy to see that, for $i \neq k, j \neq k$, $R_{i,j,\sigma}$ acts only on the $\vec{y}$ variable. That is, for such i and j,

$$R_{i,j,\sigma}(T_k(\vec{y}, v)) = T_k(R_{i,j,\sigma}(\vec{y}), v).$$

Thus, if we hold v fixed as a parameter, we can think of $(Q^{(k)}\phi)(T_k(\vec{y}, v))$ as resulting from applying the $N-1$ dimensional version of Q to ϕ with v_k held fixed.

To estimate λ_N, we need to estimate $\langle \phi, Q\phi \rangle$ when ϕ is orthogonal to the constants. When ϕ is orthogonal to the constants, and we fix v, the function

$$\vec{y} \mapsto \phi(T_k(\vec{y}, v))$$

is not, in general, orthogonal to the constants on X_{N-1}. However, we can correct for that by adding and subtracting $P_k\phi$. Therefore

$$\begin{aligned} \langle (\phi - P_k\phi), Q^{(k)}(\phi - P_k\phi) \rangle &\leq \lambda_{N-1} \|\phi - P_k\phi\|_2^2 \\ &= \lambda_{N-1}(\|\phi\|_2^2 - \|P_k\phi\|_2^2) \\ &= \lambda_{N-1}(\|\phi\|_2^2 - \langle \phi, P_k\phi \rangle). \end{aligned} \tag{3.5}$$

Then since $Q^{(k)}P_k\phi = P_k\phi$ and since $P_k\phi$ is orthogonal to $\phi - P_k\phi$,

$$\begin{aligned} \langle \phi, Q^{(k)}\phi \rangle &= \langle ((\phi - P_k\phi) + P_k\phi) Q^{(k)} ((\phi - P_k\phi) + P_k\phi) \rangle \\ &= \langle (\phi - P_k\phi), Q^{(k)}(\phi - P_k\phi) \rangle + \langle P_k\phi, P_k\phi \rangle \\ &= \langle (\phi - P_k\phi), Q^{(k)}(\phi - P_k\phi) \rangle + \langle \phi, P_k\phi \rangle \\ &\leq \lambda_{N-1}(\|\phi\|_2^2 - \langle \phi P_k\phi \rangle) + \langle \phi P_k\phi \rangle . \end{aligned} \tag{3.6}$$

By averaging over k, we have (3.4). ∎

Lemma 3.1 was used as follows in [3]: Any trial function ϕ for λ_N is a valid trial function for μ_N, so that

$$\lambda_N \leq \lambda_{N-1} + (1 - \lambda_{N-1})\mu_N. \tag{3.7}$$

Then since $\Delta_N = N(1 - \lambda_N)$, we have

$$\Delta_N \geq \frac{N}{N-1}(1 - \mu_N)\Delta_{N-1}. \tag{3.8}$$

Therefore, with $a_N = \frac{N}{N-1}(1 - \mu_N)$, for all $N \geq 3$,

$$\Delta_N \geq \left(\prod_{j=3}^{N} a_j \right) \Delta_2.$$

Thus, one route to proving a lower bound on Δ_N is to prove an upper bound on μ_N and hence a lower bound on a_N. This route led to a sharp lower bound for Δ_N—the exact value—for the one dimension Kac model investigated in [2]. However, it would not lead to a proof of Theorem 1.1. The reasons for this are worth pointing out before we proceed.

As we shall see below, the eigenspace of P with the eigenvalue μ_N—the gap eigenspace of P—is spanned by the functions specified in (1.8). Granted this, and granted Theorem 1.1, whenever condition (1.6) is satisfied:

$$\text{For } (1 - B_2) < (1 - B_1), \quad Q\phi = \lambda_N \phi \quad \Rightarrow \quad P\phi = \mu_N \phi,$$

while

$$\text{for } (1 - B_1) < (1 - B_2), \quad Q\phi = \lambda_N \phi \quad \Rightarrow \quad P\phi \neq \mu_N \phi.$$

In the second case $(1-B_1) < (1-B_2)$, the mismatch between the gap eigenspaces for Q and P means that equality cannot hold in (3.7), and hence the recursive relation (3.8) cannot possibly yield exact results in this case.

Moreover, even in the first case $(1 - B_2) < (1 - B_1)$, where there is a match between the gap eigenspaces of Q and P, there *still* will not be equality in (3.7). The reasons for this are more subtle: The inequality (3.7) comes from the key estimate (3.6). By considering (3.6), one sees that equality will hold there if and only if

$$Q^{(k)}(\phi - P_k\phi) = \lambda_{N-1}(\phi - P_k\phi)$$

for each k, where $(\phi - P_k\phi)$ is regarded as a function on X_{N-1} through the change of variables $T_k : (X_{N-1}, B) \to X_N$ that was introduced just before Lemma 3.1.

However, if ϕ is in the gap eigenspace for Q on X_N, Theorem 1.1 tells us that it is a linear combination of the three functions specified in (1.8), all of which are homogeneous of degree 3 in v. Because of the translation in (3.2), which is due to momentum conservation, $(\phi - P_k\phi)$ is regarded as a function on X_{N-1} that will *not* be homogeneous of degree 3—it will contain lower order terms. Hence $(\phi - P_k\phi)$ will not be in the gap eigenspace for $Q^{(k)}$.

The main result of the next section provides a way to use more detailed spectral information about P to sharpen the recursive estimate so that we do obtain the exact results announced in Theorem 1.1.

4. How to use more detailed spectral information on P to determine the spectral gap of Q. The following lemma is the key to using (3.4) to obtain sharp results for the model considered here.

LEMMA 4.1. *For any $N \geq 3$, let $\mu_N^\star$ be a number with*

$$\mu_N^\star < \mu_N$$

such that there are only finitely eigenvalues of P between $\mu_N^\star$ and μ_N:

$$\mu_N^\star \leq \mu_N^{(m)} < \cdots < \mu_N^{(1)} < \mu_N.$$

Let $\mu_N^{(0)}$ denote μ_N, and then, for $j = 0, \ldots, m$, let E_j denote the eigenspace of P corresponding to $\mu_N^{(j)}$. Let V_j denote the smallest invariant subspace of Q that contains E_j. Let ν_j be the largest eigenvalue of Q on V_j.

Then either

$$\lambda_N = \max\{\nu_0, \ldots, \nu_m\} \tag{4.1}$$

or else

$$\Delta_N \geq \frac{N}{N-1}(1 - \mu_N^\star)\Delta_{N-1}. \tag{4.2}$$

If $\mu_N^\star = \mu_N^{(m)}$, then we have the same alternative except with strict inequality in (4.2).

Proof. If $\lambda_N > \max\{\nu_0, \ldots, \nu_m\}$, then in the variational principle for λ_N we need only consider functions ϕ that are orthogonal to the constants and also in the orthogonal complement of each of the V_j. This means also that ϕ belongs to the orthogonal complement of each of the E_j. But then

$$\langle \phi, P\phi \rangle \leq \mu_N^\star \|\phi\|_2^2.$$

By using this estimate in (3.4), we have (4.2). Moreover, if $\mu_N^\star = \mu_N^{(m)}$, then strict inequality must hold in the last inequality. □

Lemma 4.1 gives us the dichotomy between (1.15) and (1.16) that plays a key role in the strategy described in the introduction. To put this strategy into effect, we must first carry out a more detailed investigation of the spectrum of P. The main result of the next section reduces the investigation of the spectrum of P to the study of a simpler operator—the *correlation operator* K, which is a Markov operator on functions on the unit ball B in $\mathbb{R}^3$.

5. The correlation operator K and its relation to P. While Q and P are both operators on spaces of functions of a large number of variables, the problem of computing the eigenvalues of P reduces to the problem of computing the eigenvalues of an operator on functions on B, the unit ball in $\mathbb{R}^3$.

First, define the measure ν_N on B to be the "push forward" of σ_N under the map π_j. That is, for any continuous function f on B,

$$\int_B f(v)\mathrm{d}\nu_N = \int_{X_N} f(\pi_j(\vec{v}))\mathrm{d}\sigma_N.$$

By the permutation invariance of σ_N, this definition does not depend on the choice of j. By direct calculation [3], one finds that

$$\mathrm{d}\nu_N(v) = \frac{|S^{3N-7}|}{|S^{3N-4}|}(1 - |v|^2)^{(3N-8)/2}\mathrm{d}v. \tag{5.1}$$

Now define the self-adjoint operator K on $L^2(B, \mathrm{d}\nu_N)$ through the following quadratic form:

$$\langle f, Kf\rangle_{L^2(\nu)} = \int_{X_N} f(\pi_1(\vec{v}))f(\pi_2(\vec{v}))\mathrm{d}\sigma_N \tag{5.2}$$

for all f in $L^2(B, \mathrm{d}\nu_N)$. Equivalently,

$$(Kf)\circ\pi_1 = P_1(f\circ\pi_2). \tag{5.3}$$

Note that, by the permutation invariance of σ_N, one can replace the pair $(1,2)$ of indices by any other pair of distinct indices without affecting the operator K defined by (5.3). This is the *correlation operator*.

To see the relation between the spectra of P and the spectra of K, suppose that ϕ is an eigenfunction of P that is symmetric under permutation of the particle indices. (These symmetric eigenfunctions are the ones that are significant in the physical application.) Then since any vector in the image of P has the form $\sum_{j=1}^N f_j\circ\pi_j$ for functions $f_1,\dots,f_N$ on B, we must have, for ϕ symmetric,

$$\phi = \sum_{j=1}^N f\circ\pi_j. \tag{5.4}$$

Now we ask: For which choices of f will ϕ given by (5.4) be an eigenfunction of P? To answer this, note that, by (5.3),

$$P_k\phi = f\circ\pi_k + \sum_{j=1, j\neq k}^N P_k(f\circ\pi_j). \tag{5.5}$$

Therefore, from (5.5) and the definition of K, $P_k\phi = f\circ\pi_k + (N-1)(Kf)\circ\pi_k$. Thus, by averaging over k,

$$P\phi = \frac{1}{N}\phi + \frac{N-1}{N}\sum_{j=1}^N (Kf)\circ\pi_j. \tag{5.6}$$

In the case $Kf = \kappa f$, this reduces to

$$P\phi = \frac{1}{N}(1+(N-1)\kappa)\phi,$$

and thus eigenfunctions of K yield eigenfunctions of P. It turns out that all symmetric eigenfunctions arise in exactly this way and that all eigenfunctions, symmetric or not, arise in a similar way, specified in the next lemma.

LEMMA 5.1. *Let V be the orthogonal complement in $L^2(X_N, \sigma_N)$ of the kernel of P. There is a complete orthonormal basis of V consisting of eigenfunctions ϕ of P of one of the two forms:*

(i) *For some eigenfunction f of K, $\phi = \sum_{k=1}^N f\circ\pi_k$. In this case, if $Kf = \kappa f$, then $P\phi = \mu\phi$, where*

$$\mu = \frac{1}{N}\left(1+(N-1)\kappa\right). \tag{5.7}$$

(ii) *For some eigenfunction f of K, and some pair of indices $i < j$, $\phi = f \circ \pi_i - f \circ \pi_j$. In this case, if $Kf = \kappa f$, then $P\phi = \mu\phi$, where*

$$\mu = \frac{1-\kappa}{N}. \tag{5.8}$$

Proof. Suppose that ϕ is an eigenfunction of P with nonzero eigenvalue μ, and ϕ is orthogonal to the constants. By the permutation invariance we may assume that either ϕ is invariant under permutations or that there is some pair permutation, which we may as well take to be $\sigma_{1,2}$, such that $\phi \circ \sigma_{1,2} = -\phi$. We will treat these two cases separately.

First, suppose that ϕ is symmetric. We have already observed that, in this case, the recipe $\phi = \sum_{j=1}^{N} f \circ \pi_j$, with f an eigenfunction of K, yields symmetric eigenfunctions of P. We now show that all symmetric eigenfunctions of P on V have this form.

First, simply because such a ϕ is in the image of P and is symmetric, ϕ must have the form (5.4). It remains to show that f must be an eigenfunction of K. Then by (5.6), $\mu\phi = P\phi$ becomes

$$\mu \sum_{k=1}^{N} f \circ \pi_k = \frac{1}{N} \sum_{k=1}^{N} (f + (N-1)Kf) \circ \pi_k.$$

Apply P_1 to both sides to obtain

$$\frac{1}{N} \left([f + (N-1)Kf] + (N-1)K \left[f + (N-1)Kf \right] \right) = \mu(f + (N-1)Kf),$$

which is

$$\frac{1}{N} \left(I + (N-1)K \right)^2 f = \mu(I + (N-1)K)f. \tag{5.9}$$

Since $\mu \neq 0$, f is not in the null space of either $I + (N-1)K$ or $(I + (N-1)K)^2$. It then follows from (5.9) that

$$\frac{1}{N} \left(I + (N-1)K \right) f = \mu f.$$

Thus, when ϕ is symmetric, there is an eigenfunction f of K with eigenvalue κ such that $\phi = \sum_{k=1}^{N} f \circ \pi_k$ and

$$\mu = \frac{1}{N} \left(1 + (N-1)\kappa \right).$$

We next consider the case in which

$$\phi \circ \sigma_{1,2} = -\phi.$$

Note that

$$P_k(\phi \circ \sigma_{1,2}) = P_k\phi = 0$$

whenever k is different from both 1 and 2. It follows that

$$\frac{1}{N} \sum_{k=1}^{N} P_k\phi = \frac{1}{N} \left(P_1\phi + P_2\phi \right).$$

The right-hand side is of the form $f(v_1) - f(v_2)$, and hence ϕ must have this form if it is an eigenvector. By taking $\phi = f \circ \pi_1 - f \circ \pi_2$ we have

$$\frac{1}{N}\sum_{k=1}^{N} P_k\phi = \frac{1}{N}\left((f - Kf)\circ\pi_1 - (f - Kf)\circ\pi_2\right).$$

Hence when $P\phi = \mu\phi$ and ϕ is antisymmetric as above, there is an eigenvalue κ of K such that

$$\mu = \frac{1-\kappa}{N}.$$

This proves the second part. □

Lemma 5.1 reduces the computation of the spectrum of P to the computaton of the spectrum of K. We undertake this in the next three sections.

6. Explicit form of the correlation operator K. For any two functions f and g on B that are square integrable with respect to ν_N, consider the bilinear form $\int_{X_N} f(\pi_1(\vec{v}))g(\pi_2(\vec{v}))\mathrm{d}\sigma_N$. It is easily seen from (5.3) that

$$\langle f, Kg\rangle = \int_{X_N} f(\pi_1(\vec{v}))g(\pi_2(\vec{v}))\mathrm{d}\sigma_N,$$

where $\langle\cdot,\cdot\rangle$ is the inner product on $L^2(B,\nu_N)$.

By computing the right-hand side using the factorization formula (3.2), but for T_1 instead of T_N, one finds, for $N > 3$:

$$Kg(v) = \frac{|S^{3N-10}|}{|S^{3N-7}|}\int_B g\left(\frac{\sqrt{N^2-2N}}{N-1}\sqrt{1-|v|^2}y - \frac{1}{N-1}v\right)(1-|y|^2)^{(3N-11)/2}\mathrm{d}y.$$

The explicit form of K is slightly different for $N = 3$. We can see this different form as a limiting case, if we make the dimension a continuous fact. The following way of doing this will be convenient later on.

For $\alpha > -1$, define the constant C_α by

$$C_\alpha = \left(\int_B (1-|y|^2)^\alpha \mathrm{d}y\right)^{-1},$$

so that, for

$$\alpha = \frac{3N-8}{2},$$

$$\mathrm{d}\nu_N(v) = C_\alpha(1-|y|^2)^\alpha\mathrm{d}y,$$

and then

$$Kg(v) = C_{\alpha-3/2}\int_B g\left(\frac{\sqrt{N^2-2N}}{N-1}\sqrt{1-|v|^2}y - \frac{1}{N-1}v\right)(1-|y|^2)^{\alpha-3/2}\mathrm{d}y.$$

Now, as N approaches 3, $\alpha - 3/2$ approaches -1. Then the measure $C_\alpha(1-|y|^2)^\alpha\mathrm{d}y$ concentrates more and more on the boundary of the ball B, so that, in the limit, it becomes the uniform measure on S^2. Understood in this way, the formula remains valid at $\alpha = 1/2$, i.e., at $N = 3$.

It is clear that K is a self-adjoint Markov operator on $L^2(B,\nu_N)$ and that 1 is an eigenvalue of multiplicity one. With more effort, there is much more that can be said; the spectrum of K can be completely determined.

7. The spectrum of K and ratios of Jacobi polynomials. In studying the spectrum of the correlation operator, it is in fact natural and useful to study a wider family of operators of this type. Fix any $\alpha > 1/2$ and any numbers a and b such that

$$a^2 + b^2 = 1.$$

Then define the generalized correlation operator, still simply denoted by K, through

$$Kg(v) = C_{\alpha-3/2} \int_B g\left(a\sqrt{1-|v|^2}y + bv\right)(1-|y|^2)^{\alpha-3/2}\mathrm{d}y. \tag{7.1}$$

Notice that, as v and y range over B, the maximum of $|a\sqrt{1-|v|^2}y + bv|$ occurs when ay and bv are parallel. In that case,

$$|a\sqrt{1-|v|^2}y+bv| = |a||y|\sqrt{1-|v|^2}+|b||v| \le (a^2+b^2)^{1/2}((1-|v|^2)|y|^2+|v|^2)^{1/2} \le 1.$$

Thus, as v and y range over B, so does

$$u(y,v) = a\sqrt{1-|v|^2}y + bv, \tag{7.2}$$

and $g(a\sqrt{1-|v|^2}y + bv)$ is well defined for any function g on B. Thus, K is well defined.

Now when

$$a = \frac{\sqrt{N^2-2N}}{N-1} \qquad \text{and} \qquad b = -\frac{1}{N-1}, \tag{7.3}$$

we know that K is self-adjoint because in that case it is defined in terms of a manifestly symmetric bilinear form. We shall show here that K is always self-adjoint for all $a^2 + b^2 = 1$ and that the eigenvalues of K are given by an explicit formula involving ratios of Jacobi polynomials.

To explain this, we fix some terminology and notation. For any numbers $\alpha > -1$ and $\beta > -1$, $P_n^{(\alpha,\beta)}$ denotes the nth degree polynomial in the sequence of orthogonal polynomials on $[-1,1]$ for the measure

$$(1-x)^\alpha(1+x)^\beta\mathrm{d}x$$

and is referred to as the nth degree Jacobi polynomial for (α,β). As is well known, $\{P_n^{(\alpha,\beta)}\}_{n\ge0}$ is a complete orthogonal basis for $L^2([-1,1],(1-x)^\alpha(1+x)^\beta\mathrm{d}x)$.

Of course, what we have said so far specifies $P_n^{(\alpha,\beta)}$ only up to a multiplicative constant. One common normalization is given by Rodrigues's formula

$$P_n^{(\alpha,\beta)}(x) = \frac{(-1)^n}{2^n n!}(1-x)^{-\alpha}(1+x)^{-\beta}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left((1-x)^{\alpha+n}(1+x)^{\beta+n}\right).$$

For this normalization,

$$P_n^{(\alpha,\beta)}(1) = \begin{pmatrix} n+\alpha \\ n \end{pmatrix} \qquad \text{and} \qquad P_n^{(\alpha,\beta)}(-1) = \begin{pmatrix} n+\beta \\ n \end{pmatrix}. \tag{7.4}$$

LEMMA 7.1. *Fix any $\alpha > 1/2$ and any numbers a and b such that $a^2+b^2 = 1$, and define K through the formula* (7.1). *Then K is a self-adjoint Markov operator, and*

the spectrum of K consists of eigenvalues $\kappa_{n,\ell}$ enumerated by nonnegative integers n and ℓ; these eigenvalues are given by the explicit formula

$$\kappa_{n,\ell} = \frac{P_n^{(\alpha,\beta)}(-1+2b^2)}{P_n^{(\alpha,\beta)}(1)} b^\ell, \tag{7.5}$$

where $\beta = \ell + 1/2$ and α is the parameter α entering into the definition of K.

Proof. To see that K is self-adjoint, we write it as a bilinear form and change variables to reveal the symmetry. The change of variable that we make is naturally $(y,v) \to (u,v)$, with $u(y,v)$ given by (7.2). From (7.2), one computes $y = (u - bv)/(a\sqrt{1-|v|^2})$, so that

$$\begin{aligned} 1 - |y|^2 &= \frac{a^2 - a^2|v|^2 - |u|^2 - b^2|v|^2 + 2bu\cdot v}{a^2(1-|v|^2)} \\ &= \frac{b^2 - (|u|^2 + |v|^2) + 2bu\cdot v}{a^2(1-|v|^2)}. \end{aligned} \tag{7.6}$$

The Jacobian is easy to work out, and one finds that $\mathrm{d}u\mathrm{d}v = a^3(1-|v|^2)^{3/2}\mathrm{d}y\mathrm{d}v$, so that

$$\begin{aligned} &\int_B f(v) K g(v) C_\alpha (1-|v|^2)^\alpha \mathrm{d}v \\ &= \int_B \int_B f(v) g(u) a^{-2\alpha} \left[a^2 - (|u|^2 + |v|^2) + 2bu\cdot v\right]_+^{\alpha - 3/2} C_{\alpha-3/2} \mathrm{d}u\mathrm{d}v. \end{aligned} \tag{7.7}$$

This shows that the operator K is self-adjoint on $L^2(B, C_\alpha(1-|v|^2)^\alpha)$ for all $\alpha \geq 1/2$ and all a and b with $a^2 + b^2 = 1$.

Our next goal is to prove the eigenvalue formula (7.5). This shall follow from several simple properties of K.

First, K commutes with rotations in $\mathbb{R}^3$. That is, if R is a rotation on $\mathbb{R}^3$, it is evident that

$$K(g \circ R) = (Kg) \circ R.$$

Hence we may restrict our search for eigenfunctions g of K to functions of the form

$$g(v) = h(|v|)|v|^\ell \mathcal{Y}_{\ell,m}(v/|v|)$$

for some function h on $[0,\infty)$ and some spherical harmonic $\mathcal{Y}_{\ell,m}$.

Second, for each $n \geq 0$, K preserves the space of polynomials of degree n. To see this, notice that any monomial in $\sqrt{1-|v|^2}y$ that is of odd degree is annihilated when integrated against $(1-|y|^2)^{\alpha-3/2}\mathrm{d}y$, and any even monomial in $\sqrt{1-|v|^2}y$ is a polynomial in v.

By combining these two observations, we see that K has a complete basis of eigenfunctions of the form

$$g_{n,\ell,m}(v) = h_{n,\ell}(|v|^2)|v|^\ell \mathcal{Y}_{\ell,m}(v/|v|),$$

where $h_{n,\ell}$ is a polynomial of degree n.

To determine these polynomials, we use the fact that K is self-adjoint, so that the eigenfunctions $g_{n,\ell,m}$ can be taken to be orthogonal. In particular, for any two

distinct positive integers n and p, the eigenfunctions $g_{n,\ell,m}$ and $g_{p,\ell,m}$ are orthogonal in $L^2(B, C_\alpha(1-|v|^2)^\alpha)$. Hence for each ℓ, and for $n \neq p$,

$$\int_{|v|\leq 1} h_{n,\ell}(|v|^2) h_{p,\ell}(|v|^2)(1-|v|^2)^\alpha |v|^{2\ell} \mathrm{d}v = 0.$$

By taking $r = |v|^2$ as a new variable, we have

$$\int_0^1 h_{n,\ell}(r) h_{p,\ell}(r)(1-r)^\alpha r^{\ell+1/2} \mathrm{d}r = 0.$$

This is the orthogonality relation for a family of Jacobi polynomials in one standard form, and this identifies the polynomials $h_{n,\ell}$. A more common standard form, and one that is used in the sources to which we shall refer, is obtained by the change of variable $t = 2r - 1$, so that the variable t ranges over the interval $[-1, 1]$. Then for $\alpha, \beta > -1$, $P_n^{(\alpha,\beta)}(t)$ is the nth degree orthogonal polynomial for the weight $(1-t)^\alpha(1+t)^\beta$. With the variables t and $|v|^2$ related as above, i.e.,

$$t = 2|v|^2 - 1,$$

$$h_{n,\ell}(|v|^2) = P_n^{(\alpha,\beta)}(t)$$

for

$$\beta = \ell + \frac{1}{2}.$$

Now that we have all of the eigenfunctions determined, a further observation gives us a simple formula for the eigenvalues. Consider any eigenfunction g with eigenvalue κ, so that $Kg(v) = \kappa g(v)$. Let $\hat{e}$ be any unit vector in R^3. Then since g is a polynomial and hence continuous,

$$\begin{aligned} \lim_{t\to 1} Kg(t\hat{e}) &= \lim_{t\to 1} \int_B g\left(a\sqrt{1-t^2}y + bt\hat{e}\right) C_{\alpha-3/2}(1-|y|)^{\alpha-3/2}\mathrm{d}y \\ &= g\left(b\hat{e}\right), \end{aligned} \tag{7.8}$$

since $K1 = 1$. By combining this with $Kg(v) = \kappa g(v)$, we have

$$g\left(b\hat{e}\right) = \kappa g(\hat{e}).$$

Now consider any eigenfunction $g_{n,\ell,m}$ of the form given above, and let $\kappa_{n,\ell}$ be the corresponding eigenvalue, which will not depend on m. Then by taking any $\hat{e}$ so that $\mathcal{Y}_{\ell,m}(\hat{e}) \neq 0$, we have

$$\kappa_{n,\ell} = \frac{h_{n,\ell}(b^2)}{h_{n,\ell}(1)} b^\ell. \tag{7.9}$$

By changing variables as above to express this as a ratio of Jacobi polynomials, we finally have proved (7.5). □

One might expect the largest eigenvalues of K to correspond to eigenfunctions that are polynomials of low degree. After all, in a system of orthogonal polynomials, those with high degree will have many changes of sign, and one might expect considerable

cancellation when applying an averaging operator, such as K, to them. Therefore, let us compute the $\kappa_{n,\ell}$ for low values of n and ℓ. We find from (7.5), by using the value $b = -1/(N-1)$ from (7.3), that

$$\kappa_{0,1} = \kappa_{1,0} = \frac{-1}{N-1}, \tag{7.10}$$

so that $\kappa_{n,\ell}$ is negative for $n + \ell = 1$. For $n + \ell = 2$, we find from (7.5) that

$$\begin{aligned} \kappa_{1,1}(N) &= \frac{5N-3}{3(N-1)^3}, \\ \kappa_{2,0}(N) &= \frac{(N-3)(15N^2-15N+4)}{3(3N-4)(N-1)^4}, \\ \kappa_{0,2}(N) &= \frac{1}{(N-1)^2}. \end{aligned} \tag{7.11}$$

Evidently, for large N,

$$\kappa_{0,2}(N) = \frac{1}{N^2} + \mathcal{O}\left(\frac{1}{N^3}\right),$$

while

$$\kappa_{1,1}(N) = \frac{5}{3N^2} + \mathcal{O}\left(\frac{1}{N^3}\right) \qquad \text{and} \qquad \kappa_{0,2}(N) = \frac{5}{3N^2} + \mathcal{O}\left(\frac{1}{N^3}\right).$$

Thus, one might expect that, at least for large values of N, 1, $\kappa_{1,1}(N)$, $\kappa_{2,0}(N)$, and $\kappa_{0,2}(N)$ are the four largest eigenvalues of K and that $\kappa_{0,1} = \kappa_{1,0}$ is the most negative, with all other eigenvalues of K lying strictly between these. We shall show in the next section that this is indeed the case for all $N \geq 4$ and that 1 and $\kappa_{1,1}$ are the two largest eigenvalues of K for all $N \geq 3$.

When we use Lemma 5.1 to convert this to spectral information on P, we find that $\kappa_{0,1}$, $\kappa_{1,0}$, and $\kappa_{0,2}$ all correspond to the same eigenvalues of P, namely,

$$\frac{1}{N}\left(1 + \frac{1}{N-1}\right) = \frac{1}{N}\left(1 + (N-1)\frac{1}{(N-1)^2}\right) = \frac{1}{N-1}.$$

This is the eigenvalue of P that shall play the role of $\mu_N^{(m)}$ in our application of Lemma 4.1.

Let us conclude this section by recording a number of useful calculations that can be made by using (7.5).

For $N = 3$, we have

$$\kappa_{1,1}(3) = \frac{1}{2} > \kappa_{2,2}(3) = \frac{13}{40} > \kappa_{0,2}(3) = \frac{1}{4} > \kappa_{2,0}(3) = 0. \tag{7.12}$$

For $N = 4$, we have

$$\kappa_{1,1}(4) = \frac{17}{81} > \kappa_{0,2}(4) = \frac{1}{9} > \kappa_{2,0}(4) = \frac{23}{243}. \tag{7.13}$$

For $N = 5$, we have

$$\kappa_{1,1}(5) = \frac{11}{96} > \kappa_{2,0}(5) = \frac{19}{264} > \kappa_{0,2}(5) = \frac{1}{16}. \tag{7.14}$$

In each case, the second largest eigenvalue after 1, among the ones listed, is $\kappa_{1,1}$. In the next section we shall see that the list is not misleading: $\kappa_{1,1}$ is the gap eigenvalue. However, note that the third largest eigenvalue comes from different values of n and ℓ for each of $N = 3$, $N = 4$, and $N = 5$. As we shall see, things do settle down for $N \geq 5$; the third largest eigenvalue does turn out to be $\kappa_{2,0}$ in all such cases.

LEMMA 7.2. *For all $N \geq 5$, $\kappa_{1,1}(N) > \kappa_{2,0}(N) > \kappa_{0,2}(N)$.*

Proof. From (7.11),

$$\kappa_{2,0}(N) - \kappa_{0,2}(N) = \frac{2N(3N^2 - 15N + 8)}{3(3N-4)(N-1)^4}.$$

A simple calculation shows that the roots of the polynomial in the numerator are less than 5, so that $\kappa_{2,0}(N) > \kappa_{0,2}(N)$ for $N \geq 5$. A similar argument applied to $\kappa_{1,1}(N) - \kappa_{2,0}(N)$ yields the conclusion of the lemma. □

Our goal in the next section is to show that, for all $N \geq 4$, there are no eigenvalues $\kappa_{n,\ell}$ with $n + \ell > 2$ that are larger than the ones listed above and that, for $N = 3$, the three largest eigenvalues are $1 = \kappa_{0,0} > 1/2 = \kappa_{1,1} > 13/40 = \kappa_{2,2}$. However, since there is no simple monotonicity in $n + \ell$, this shall require some detailed estimate on ratios of Jacobi polynomials.

We shall also need to know that in all cases $\kappa_{0,1} = \kappa_{1,0} = -1/(N-1)$ is the most negative eigenvalue. This will tell us the four largest eigenvalues of P for $N \geq 4$ and the three largest for $n = 3$, and this shall turn out to be enough to prove the main result, Theorem 1.1.

Finally, the value of $\kappa_{2,2}(N)$ will play an important role in the proof of Theorem 1.2, and so we record the expression here:

$$\kappa_{2,2}(N) = \frac{21N^3 - 60N^2 + 27N - 4}{(3N-4)(N-1)^6}. \tag{7.15}$$

8. The determination of the spectrum of K. The main result in this section is the following theorem.

THEOREM 8.1. *For $N \geq 5$ and all n and ℓ with $n + \ell > 2$,*

$$-\frac{1}{N-1} \leq \kappa_{n,\ell}(N) < \kappa_{0,2}(N). \tag{8.1}$$

For $N = 4$ and all n and ℓ with $n + \ell > 2$,

$$-\frac{1}{N-1} \leq \kappa_{n,\ell}(4) < \kappa_{2,0}(4). \tag{8.2}$$

For $N = 3$ and all n and ℓ with $n + \ell > 0$, except for $n = 1, \ell = 1$,

$$-\frac{1}{N-1} \leq \kappa_{n,\ell}(3) \leq \kappa_{2,2}(3) = \frac{13}{40}. \tag{8.3}$$

We present the proof at the end of this section after a number of preparatory lemmas. These lemmas rest on two deep results about Jacobi polynomials. One is a formula due to Koornwinder [8] (see also [1, p. 31]) that was already applied in [3].

For all $-1 \leq x \leq 1$, all n, and all $\alpha > \beta$,

$$\frac{J_n^{(\alpha,\beta)}(x)}{J_n^{(\alpha,\beta)}(1)} = \int_0^\pi \int_0^1 \left[\frac{1 + x - (1-x)r^2}{2} + i\sqrt{1-x^2}\, r\cos(\theta)\right]^n \mathrm{d}m_{\alpha,\beta}(r,\theta), \tag{8.4}$$

where

$$m_{\alpha,\beta}(r,\theta) = c_{\alpha,\beta}(1-r^2)^{\alpha-\beta-1} r^{2\beta+1} (\sin\theta)^{2\beta}\, \mathrm{d}r\mathrm{d}\theta,$$

and $c_{\alpha,\beta}$ is a normalizing constant that makes $\mathrm{d}m_{\alpha,\beta}$ a probability measure.

Koornwinder's bound is very useful for obtaining uniform control in n for given ℓ and N. But since in Lemma 7.1

$$\alpha = \frac{3N-8}{2} \quad \text{and} \qquad \beta = \ell + \frac{1}{2}, \tag{8.5}$$

we can apply (8.4) only when

$$\ell < \ell^\star = \frac{3N-9}{2}. \tag{8.6}$$

As in [3], one may use this formula to show the following.

LEMMA 8.2. *For all ℓ with $2 \le \ell < \ell^\star$, and all $n > 0$ and $N \ge 3$,*

$$|\kappa_{n,\ell}(N)| < \frac{1}{(N-1)^2} = \kappa_{0,2}(N).$$

Note that, while this lemma does not address the case $n = 0$, this is not a problem: We have the explicit formula

$$\kappa_{0,\ell} = \left(\frac{-1}{N-1}\right)^\ell. \tag{8.7}$$

To handle large values of ℓ, we need another deep result, which is a uniform pointwise bound on the *orthonormal* Jacobi polynomials that was obtained by Nevai, Erdelyi, and Magnus [12]. Let $p_n^{\alpha,\beta}$ be the orthonormal Jacobi polynomial of degree n with a positive leading coefficient for the weight $w(x) = (1-x)^\alpha(1+x)^\beta$. It was shown in [12] that, for all $\alpha \ge -1/2$ and $\beta \ge -1/2$ and all nonnegative integers n,

$$\max_{x\in[-1,1]} \sqrt{1-x^2}\, w(x) p_n^{\alpha,\beta}(x)^2 \le \frac{2e(2+\sqrt{\alpha^2+\beta^2})}{\pi}. \tag{8.8}$$

Of course, we could use the orthonormal Jacobi polynomials in the ratio formula (7.5), since any normalization factor would cancel out in the ratio. However, the exact formula (7.4) for the denominator in (7.5) is simplest in the other normalization. Hence we need the relation between $p_n^{\alpha,\beta}$ and $P_n^{\alpha,\beta}$, which is given by $p_n^{\alpha,\beta} = l_n P_n^{\alpha,\beta}$, where

$$l_n = \left(\frac{2n+\alpha+\beta+1}{2^{\alpha+\beta+1}} \frac{\Gamma(n+1)\Gamma(n+\alpha+\beta+1)}{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}\right)^{1/2}.$$

Therefore

$$\frac{P_n^{\alpha,\beta}(x)^2}{P_n^{\alpha,\beta}(1)^2} \le \frac{1}{l_n^2} \frac{2e\Gamma(n+1)^2\Gamma(\alpha+1)^2(2+\sqrt{\alpha^2+\beta^2})}{\sqrt{1-x^2}\, w(x)\pi\Gamma(n+\alpha+1)^2}. \tag{8.9}$$

At this point it is perhaps worth noting that, since the spectrum of K lies in $[-1,1]$, any upper bound on its eigenvalues by a number larger than one is vacuous. This implies that for certain regions the identity (7.9) will provide a stronger bound than (8.9). We shall return to this point at the end of the paper.

Substituting $x = -1 + \frac{2}{(N-1)^2}$, $\beta = \ell + \frac{1}{2}$, and $\alpha = \frac{3}{2}N - 4$ in (8.9) and then multiplying by $\frac{1}{(N-1)^{2l}}$ yields

$$\kappa_{n,\ell}(N) \le \tilde{\kappa}_{n,\ell}(N), \tag{8.10}$$

where

$$\tilde{\kappa}^2_{n,\ell}(N) = \frac{2e}{\pi} g_1(n,\ell,N) g_2(N) g_3(n,N) g_4(n,\ell,N), \tag{8.11}$$

with

$$\begin{aligned} g_1(n,\ell,N) &= \left(\frac{4 + \sqrt{9N^2 - 48N + 65 + 4\ell^2 + 4\ell}}{3N + 4n + 2\ell - 5} \right), \\ g_2(N) &= \left(\frac{(N-1)^2}{N(N-2)} \right)^{(3N-7)/2}, \\ g_3(n,N) &= \frac{\Gamma(n+1)\Gamma\left(\frac{3}{2}N - 3\right)}{\Gamma\left(n + \frac{3}{2}N - 3\right)}, \\ g_4(n+\ell,N) &= \frac{(N-1)^2 \Gamma\left(n + \ell + \frac{3}{2}\right) \Gamma\left(\frac{3}{2}N - 3\right)}{\Gamma\left(n + \ell + \frac{3}{2}N - \frac{5}{2}\right)}. \end{aligned} \tag{8.12}$$

Our goal now is to extract a reasonably tight upper bound for $\tilde{\kappa}_{n,\ell}(N)$ with as much monotonicity in n, ℓ, and N as possible. The next lemmas address this goal.

LEMMA 8.3. *For $\ell \ge 0$, $N \ge 3$, and $n \ge 0$,*

$$g_1(n,\ell,N) \le \left(\frac{4}{3N + 4n + 2\ell - 5} + 1 \right), \tag{8.13}$$

where the right-hand side is clearly decreasing in n, ℓ, and N.

Proof. Note that, for $n \ge 0$, $\ell \ge 0$, and $N \ge 3$,

$$\frac{\sqrt{9N^2 - 48N + 65 + 4\ell^2 + 4\ell}}{3N + 4n + 2\ell - 5} \le 1 \tag{8.14}$$

since then

$$\begin{aligned} &(3N + 4n + 2\ell - 5)^2 - (9N^2 - 48N + 65 + 4\ell^2 + 4\ell) \\ &\quad = (24N - 40)n + (12N - 24)\ell + 16n^2 + 16n\ell + 18N - 40 > 0. \end{aligned} \tag{8.15}$$

∎

LEMMA 8.4. *For $N \ge 4$, $g_2(N)$ is a decreasing function of N.*

Proof. Let $h(x) = (1 - 1/x^2)^{2-3x/2}$, so that $g_2(N) = h(N-1)$. By computing the derivative of $\ln(h(x))$, one finds that it is negative for $x \ge 3$. ∎

LEMMA 8.5. *For $n \ge 0$ and $N \ge 3$, $g_3(n,N)$ is a decreasing function of n and N.*

Proof. For n a nonzero integer

$$\frac{\Gamma(n+1)\Gamma\left(\frac{3}{2}N - 3\right)}{\Gamma\left(n + \frac{3}{2}N - 3\right)} = \frac{n}{n + \frac{3}{2}N - 4} \, \frac{n-1}{n + \frac{3}{2}N - 5} \cdots \frac{1}{\frac{3}{2}N - 3}. \tag{8.16}$$

Since each factor is less than 1 for $N \ge 3$ and is a decreasing function of N, the assertion follows. ∎

LEMMA 8.6. *For $N \geq 3$, $g_4(n+\ell, N)$ is a decreasing function of $n+\ell$, with*

$$\lim_{n+\ell\to\infty} g_4(n+\ell, N) = 0. \tag{8.17}$$

Moreover, for $n+\ell \geq \ell^\star = 3(N-3)/2$,

$$g_4(n,\ell,N) \leq \frac{(N-1)^2\Gamma\left(\frac{3}{2}N-3\right)^2}{\Gamma(3N-7)} \leq f(N), \tag{8.18}$$

where

$$f(N) = \frac{(N-1)^2\sqrt{\pi}(\frac{3}{2}N-4)}{2^{3N-8}}. \tag{8.19}$$

Finally, for $N \geq 5$, $(N-1)^4 f(N)$ is a decreasing function of N.

Proof. Since

$$\frac{\Gamma\left(n+\ell+\frac{5}{2}\right)}{\Gamma\left(n+\ell+\frac{3}{2}N-\frac{3}{2}\right)}\,\frac{\Gamma\left(n+\ell+\frac{3}{2}N-\frac{5}{2}\right)}{\Gamma\left(n+\ell+\frac{3}{2}\right)} = \frac{\left(n+\ell+\frac{3}{2}\right)}{n+\ell+\frac{3}{2}N-\frac{5}{2}} < 1 \tag{8.20}$$

for $N \geq 3$, it follows that, for fixed nonnegative integers N,

$$\frac{\Gamma\left(n+\ell+\frac{3}{2}\right)}{\Gamma\left(n+\ell+\frac{3}{2}N-\frac{5}{2}\right)} \tag{8.21}$$

is a decreasing function of $n+\ell$. Hence, for $n+\ell \geq 3(N-3)/2$,

$$\frac{\Gamma\left(n+\ell+\frac{3}{2}\right)}{\Gamma\left(n+\ell+\frac{3}{2}N-\frac{5}{2}\right)} \leq \frac{\Gamma\left(\frac{3}{2}N-3\right)}{\Gamma\left(3N-7\right)}.$$

This together with the definition of g_4 proves the first inequality in (8.18). Use of the duplication formula for the Γ function yields

$$\frac{\Gamma\left(\frac{3}{2}N-3\right)^2}{\Gamma(3N-7)} = \frac{\sqrt{\pi}\Gamma\left(\frac{3}{2}N-3\right)}{2^{3N-8}\Gamma\left(\frac{3}{2}N-\frac{7}{2}\right)} = \frac{\sqrt{\pi}\left(\frac{3}{2}N-4\right)\Gamma\left(\frac{3}{2}N-4\right)}{2^{3N-8}\Gamma\left(\frac{3}{2}N-\frac{7}{2}\right)} < \frac{\sqrt{\pi}\left(\frac{3}{2}N-4\right)}{2^{3N-8}}.$$

This implies the second inequality in (8.18). A check of the logarithmic derivative of $(N-1)^4 f(N)$ shows that it is negative for $N \leq 5$. ∎

Now, by combining the results in the last four lemmas, we have, for $N \geq 3$ and $n+\ell \geq \ell^\star = 3(N-3)/2$,

$$\tilde{\kappa}^2_{n,\ell}(N) \leq \hat{\kappa}^2_{n,l}(N) \leq \kappa^2(N), \tag{8.22}$$

where

$$\hat{\kappa}^2_{n,l}(N) = \frac{2e}{\pi}\left(\frac{4}{3N+4n+2\ell-5}+1\right) g_2(N)g_3(n,N)g_4(n,l,N) \tag{8.23}$$

and

$$\kappa^2(N) = \frac{2e}{\pi}\left(\frac{4}{6N-14}+1\right) g_2(N)f(N), \tag{8.24}$$

where g_2, g_3, and f are given by (8.12) and (8.19).

We are now ready to prove the main theorem of this section.

Proof of Theorem 8.1. First, we take care of large values of N. By Lemmas 8.4 and 8.6, $(N-1)^4\kappa(N)$ is a decreasing function of N for $N \geq 5$. Direct computation shows that at $N = 12$ this quantity is less than one. Hence for $N \geq 12$, $\kappa(N) \leq (N-1)^{-4} = \kappa_{0,2}^2$. For $\ell \geq \ell^\star$, so that (8.22) is satisfied, this proves (8.1) for $N \geq 12$. On the other hand, if $2 \leq \ell < \ell^\star$, we have this from Lemma 8.2 or (8.7). Thus, in any case, (8.1) is valid for $N \geq 12$.

For $4 \leq N \leq 11$, we again use Lemma 8.2 or (8.7) for $2 \leq \ell < \ell^*$ and computation of $\hat{\kappa}_{n,\ell}$. By (8.17), for each such N there is a finite value $k(N)$ so that we need only consider values of $n + \ell < k(N)$. By checking these cases, we obtain (8.1) and (8.2).

We finally turn to $N = 3$, which requires the greatest amount of computation. First for $n = 0$, we have from (8.7)

$$\kappa_{0,\ell}(3) = \left(\frac{-1}{2}\right)^\ell$$

so $\kappa_{0,1}(3) = -1/2$ and $|\kappa_{0,\ell}(3)| < 1/3$ for $\ell \geq 2$.

The exact forms of the eigenvalues are simple enough to be useful for $n = 1$ and 2 as well. We have

$$\kappa_{1,\ell}(N) = \frac{(-1)^{\ell+1}}{3}\frac{[2\ell N + 3(N-1)]}{(N-1)^{\ell+2}}$$

and

$$\kappa_{2,\ell}(N) = \frac{(-1)^\ell((4\ell^2+16\ell+15)N^3 - (8\ell^2+44\ell+60)N^2 + (49+16\ell)N - 12)}{3(3N-4)(N-1)^{\ell+4}}.$$

By specializing to $N = 3$,

$$\kappa_{1,\ell}(3) = (-1)^{\ell+1}\frac{\ell+1}{2^{\ell+1}}$$

so that $|\kappa_{1,\ell}(3)| \leq 3/8$ for $\ell \geq 2$. Likewise, for $N = 3$,

$$\kappa_{2,\ell}(3) = \frac{\ell}{20}\frac{(7+3\ell)}{2^\ell}(-1)^\ell,$$

which implies that

$$|\kappa_{2,\ell}(3)| \leq |\kappa_{22}(3)| = \frac{13}{40}. \tag{8.25}$$

For higher values of n, we estimate $\kappa_{n,\ell}^2$ by means of $\hat{\kappa}_{n,\ell}^2$. Since $\ell^\star = 0$ for $N = 3$, we may use Lemma 8.6 for all ℓ, and then by (8.17), for each fixed n, there is a maximal value $\ell(n)$ that needs to be considered and a maximum value of n that needs to be considered. Table 1 gives the values of n, ℓ, and $\hat{\kappa}_{n,\ell}^2, (3)$ when $\tilde{\kappa}_{n,\ell}^2(3) < 1/4$. The monotonicity of $\kappa_{n,\ell}^2(3)$ in n and l shows that $\hat{\kappa}_{n,\ell}^2(3) \leq \hat{\kappa}_{n_0,\ell_0}^2(3)$ for $n \geq n_0$ and $\ell \geq \ell_0$ where (n_0, ℓ_0) is chosen from the table. The remaining values can be computed

TABLE 1

n	ℓ	$\hat{\kappa}^2_{n,\ell}$	n	ℓ	$\hat{\kappa}^2_{n,\ell}$	n	ℓ	$\hat{\kappa}^2_{n,\ell}$	n	ℓ	$\hat{\kappa}^2_{n,\ell}$
3	1253	0.10562	20	210	0.10547	37	90	0.10543	54	34	0.10514
4	989	0.10562	21	198	0.10559	38	86	0.10531	55	31	0.10538
5	817	0.10556	22	188	0.10546	39	82	0.105277	56	29	0.10506
6	694	0.10561	23	178	0.10552	40	78	0.10528	57	26	0.10538
7	604	0.10555	24	169	0.10549	41	74	0.10537	58	24	0.10511
8	533	0.10560	25	161	0.10537	42	70	0.10551	59	21	0.10551
9	477	0.10558	26	153	0.10540	43	67	0.10523	60	19	0.10529
10	431	0.10558	27	145	0.10558	44	63	0.10552	61	17	0.10509
11	393	0.10555	28	138	0.10561	45	60	0.10534	62	14	0.10560
12	360	0.10561	29	132	0.10542	46	57	0.10521	63	12	0.10545
13	333	0.10548	30	126	0.10534	47	54	0.10512	64	10	0.10534
14	308	0.10558	31	120	0.10540	48	50	0.10562	65	8	0.10523
15	287	0.10554	32	114	0.10554	49	48	0.10508	66	6	0.10514
16	268	0.10556	33	109	0.10543	50	45	0.10512	67	4	0.10509
17	251	0.10558	34	104	0.10540	51	42	0.10521	68	2	0.10506
18	236	0.10554	35	99	0.10546	52	39	0.10535	69	0	0.10503
19	222	0.10560	36	94	0.10562	53	36	0.10552			

from the exact formula for $\kappa_{n,\ell}(3)$ from (7.1), and the results are all consistent with (8.3). □

9. The determination of the spectrum of P. For given values of N, n, and ℓ, let $\mu_{n,\ell}(N)$ be the eigenvalue of P corresponding to the eigenvalue $\kappa_{n,\ell}(N)$ of K through Theorem 8.1, where we use (5.7) if $\kappa_{n,\ell}(N) > 0$ and use (5.8) if $\kappa_{n,\ell}(N) < 0$. (This is the relevant choice, as we are concerned with the largest eigenvalues of P.)

By consulting the calculations in (7.10) for $n + \ell = 1$, and in (7.12), (7.13), and (7.14) for $n + \ell = 2$, and finally the bounds in Theorem 8.1 for $n + \ell > 2$, we see that for all $N \geq 3$ the largest eigenvalues of K is $\kappa_{1,1}$, and the least (most negative) is $\kappa_{0,1} = \kappa_{1,0}$. Thus, by turning to Lemma 5.1 and using the positive eigenvalue in (5.7) and the negative one in (5.8), we see that the positive one yields the greater value for each N. Thus, the gap eigenvalue of P, μ_N, is given by

$$\mu_N = \mu_{1,1}(N) = \frac{3N-1}{3(N-1)^2}. \tag{9.1}$$

Use of this result in (3.8) would yield a strictly positive lower bound on Δ_N, uniform in N, but, as we have said above, it would not yield the exact lower bound. To obtain this, we now carry out the strategy outlined in the introduction.

First, we combine Lemma 5.1 and Theorem 8.1 to produce the information necessary for the application of Lemma 4.1. We must now make a choice of the thresholds $\mu^\star_N$ that appear in Lemma 4.1. The choice we shall make is based on trial function computations with Q that suggest that the gap eigenfunctions are the ones specified in Theorem 1.1.

Notice that in Theorem 1.1 the formula given for Δ_N is of the form $C\frac{N}{N-1}$ for some constant C. This value can be guessed by computing the eigenvalues of Q on the invariant subspace of polynomials of degree 4 or less in the v_j. If we are to prove this guess correct by using (4.2) of Lemma 4.1, we require a value of $\mu_N^\star$ such that

$$\frac{N}{N-1}(1-\mu_N^\star)\frac{N-1}{N-2} \geq \frac{N}{N-1}, \tag{9.2}$$

at least for $N \geq 4$. (The guess is valid only for $N-1 \geq 3$. For $N-1=2$, there is a different value of Δ_2 which has been determined already in section 2.)

The largest value of $\mu_N^\star$ that will satisfy (9.2) is

$$\mu_N^\star = \frac{1}{N-1} \qquad \text{for} \qquad N \geq 4. \tag{9.3}$$

This turns out to be an eigenvalue of P: Indeed, we have found in (7.11) that $\kappa_{0,2} = 1/(N-1)^2$. Furthermore, we have found in (7.10) that $\kappa_{0,1} = \kappa_{1,0} = -1/(N-1)$. By using the first of these results in (5.7) of Lemma 5.1 and the second in (5.8), we find that

$$\mu_{0,2} = \mu_{1,0} = \mu_{0,1} = \frac{1}{N-1}.$$

For $N=3$ we need to make a different choice, as the spectrum of Q is quite different for $N=2$ and for $N \geq 3$. The choice that will work is $\mu_3^\star = \mu_{2,2}(3)$. Since $\kappa_{2,2}(3) = 13/40$, we have from Lemma 5.1 that $\mu_{2,2}(3) = \frac{1}{3}(1+2(13/40)) = \frac{11}{20}$. Thus,

$$\mu_3^\star = \frac{11}{20}. \tag{9.4}$$

Now, to apply Lemma 4.1, we need the eigenspaces of P for the eigenvalues μ satisfying $1 > \mu > \mu_N^\star$. By Theorem 8.1 and (7.13), for $N=3$ and $N=4$, there is just one such eigenvalue, namely, $\mu_{1,1}(4)$, the gap eigenvalue, and for $N \geq 5$, there are two: $\mu_{1,1}(N)$ and $\mu_{2,0}(N)$.

Let $E_{n,\ell}$ be the eigenspace of P corresponding to the eigenvalue $\mu_{n,\ell}(N)$. For all values of n and ℓ with $n+\ell \leq 2$, we have determined the corresponding eigenfunctions of K and thus, through Lemma 5.1, the corresponding eigenfunctions of P. Thus, we have the following explicit descriptions of the $E_{n,\ell}$ for $n+\ell \leq 2$.

First, for $n+\ell = 1$, the eigenvalues of K are negative, and so by Lemma 5.1, the eigenfunctions are antisymmetric. If we are concerned only with the spectrum of Q on the subspace of symmetric functions (which is all that is of significance for Kac's application to the Boltzmann equation), we can ignore these eigenspaces. However, they turn out to be very simple. The $n=0, \ell=1$ eigenfunctions of K are degree one spherical harmonics, and the $n=1, \ell=0$ eigenfunctions of K are degree one Jacobi polynomials in $|v|^2$. Hence

$$E_{0,1} \quad \text{is spanned by the functions} \quad v_i^\alpha - v_j^\alpha, \alpha = 1,2,3, \quad \text{and } i<j, \tag{9.5}$$

while

$$E_{1,0} \quad \text{is spanned by the functions} \quad |v_i|^2 - |v_j|^2, i<j. \tag{9.6}$$

Next, for $n+\ell = 2$, the eigenvalues of K are positive, and so by Lemma 5.1, the eigenfunctions are symmetric. The $n=0, \ell=2$ eigenfunctions of K are degree two spherical harmonics and so have the form

$$f_{0,2}(v) = \sum_{\alpha,\beta=1}^{3} A_{\alpha,\beta} v^\alpha v^\beta$$

for some traceless symmetric 3×3 matrix A. Hence, by Lemma 5.1,

$$E_{0,2} \quad \text{is spanned by the functions} \quad \sum_{j=1}^{N} f_{0,2}(v_j), \tag{9.7}$$

with $f_{0,2}$ given as above.

For $n = 1$, $\ell = 1$, the eigenfunctions of K are the product of a degree one spherical harmonic and a degree one Jacobi polynomial in $|v|^2$. When we sum over the particles, the constant term in the Jacobi polynomial drops out due to the momentum constraint, and we see that

$$E_{1,1} \quad \text{is spanned by the functions} \quad \sum_{j=1}^{N} f_{1,1}(v_j), \tag{9.8}$$

where

$$f_{1,1}(v) = |v|^2 v^\alpha, \quad \alpha = 1, 2, 3.$$

Finally, for $n = 2$, $\ell = 0$, the eigenfunction of K is a degree two Jacobi polynomial in $|v|^2$. After summing on the particles, the linear term can be absorbed into the constant by the energy constraint, and so we see that

$$E_{2,0} \quad \text{is spanned by the function} \quad \sum_{j=1}^{N} f_{2,0}(v_j), \tag{9.9}$$

where

$$f_{2,0}(v) = |v|^4 - \int_B |v|^4 \mathrm{d}\nu_N.$$

We close this section with a lemma that we shall need to prove Theorem 1.2. There we shall need to know the next largest eigenvalue of P below $\max_{n+\ell \leq 2} \mu_{n,\ell}(N)$. One might guess that this occurs for some values of n and ℓ with $n + \ell = 3$, but this is not the case: By (8.3) of Theorem 8.1, and (7.12), we see that for $N = 3$ the most negative eigenvalue of K is $-1/2$, and by Lemma 5.1, this corresponds to the eigenvalue $1/2$ of P. On the other hand, the largest eigenvalue of K apart from $\kappa_{1,1}(3)$ is $\kappa_{2,2}(3) = 13/40$. This corresponds to the eigenvalue $11/20$ of P. Since $11/20 > 1/2$, we do indeed have

$$\sup_{n+\ell>2} \mu_n, \ell(3) = \mu_{2,2}(3) = \frac{11}{20}.$$

It seems likely, on the basis of computations that we have made, that in fact

$$\sup_{n+\ell>2} \mu_n, \ell(N) = \mu_{2,2}(N) \tag{9.10}$$

for all $N \geq 3$. However, for the proof of Theorem 1.2, all that we require is the following.

LEMMA 9.1. *For $N = 3, 4, 5, 6$, and 7, (9.10) is true.*

Proof. Note that the case $N = 3$ has already been proved in the remarks above. To deal with the other cases, we proceed essentially as in the proof of Theorem 8.1, using (8.23) to reduce the number of cases to be checked to a finite number and then checking these. We will therefore be brief in our remarks on the remaining cases.

Perhaps the most important point to recall is that (8.23) is valid for $n + \ell \geq 3(N-3)/2$. Since the right-hand side evaluates to zero for $N = 3$, we could use it without restriction. For $N = 7$, though, $3(N-3)/2$ evaluates to 6, and so we may use (8.23) only for $n + \ell \geq 6$. So these cases must be checked by direct computation of the eigenvalues by using 7.1 and then converting these to eigenvalues of P by using Lemma 5.1.

Then, by using (8.23) for $n + \ell > 6$, one finds that

$$\kappa^2_{n,\ell}(7) < \kappa^2_{2,2}(7)$$

unless $0 \leq n \leq 6$ and $0 \leq \ell \leq 27$. By computing the rest of the eigenvalues of P in this 6 by 27 rectangle, we find that the stated result is true for $N = 7$.

A similar analysis takes care of $N = 4$, $N = 5$, and $N = 6$. □

We shall not need to know the corresponding eigenfunctions in our application of Lemma 9.1, since we will be concerned only with the eigenspaces of eigenvalues lying strictly above $\mu_{2,2}(N)$, and those have been determined already in this section.

10. The spectrum of Q on invariant subspaces containing eigenspaces of P. For each n and ℓ, let $V_{n,\ell}$ be the smallest invariant subspace of Q containing $E_{n,\ell}$. As we shall see, for $n + \ell \leq 2$, $V_{n,\ell} = E_{n,\ell}$ except for $n = 2$, $\ell = 0$, in which case $V_{2,0}$ is two-dimensional, while $E_{1,0}$ is one-dimensional. This is established in the next lemma, which also specifies the spectrum of Q on these invariant subspaces. The eigenvalues of course depend on the particular choice of b in the definition of Q, but in a very simple way: The dependence on b is only through the quantities $(1 - B_1)$ and $(1 - B_2)$, where B_j is the jth moment of b, as defined in (1.5).

LEMMA 10.1. *Every nonzero function in $E_{0,1}$ and in $E_{1,0}$ is an eigenfunction of Q with eigenvalue*

$$\lambda^Q_{0,1} = \lambda^Q_{1,0} = 1 - (1 - B_1)\frac{1}{N-1}, \tag{10.1}$$

so that $V_{0,1} = E_{0,1}$ and $V_{1,0} = E_{1,0}$.

Every nonzero function in $E_{1,1}$ is an eigenfunction of Q with the eigenvalue

$$\lambda^Q_{1,1} = 1 - (1 - B_2)\frac{1}{(N-1)}, \tag{10.2}$$

so that $V_{1,1} = E_{1,1}$.

Furthermore, every nonzero function in $E_{0,2}$ is an eigenfunction of Q with the eigenvalue

$$\lambda^Q_{0,2} = 1 - (1 - B_2)\frac{3}{2(N-1)}, \tag{10.3}$$

so that $V_{0,2} = E_{0,2}$.

Finally, while $V_{2,0}$ is larger than $E_{2,0}$, there are only two eigenvalues of Q in $V_{2,0}$. These are

$$1 - (1 - B_2)\left(\frac{1}{N(N-1)}\left[(2N-1) \pm \sqrt{N^2 - 3N + 1}\right]\right). \tag{10.4}$$

For all $N \geq 3$, the largest of these eigenvalues is $\lambda^Q_{1,1}$.

Before beginning the proof, we note that, if $(1/2)b(x)\mathrm{d}x$ is a Dirac mass at $x = 1$, the collisions are all trivial (zero scattering angle), and thus $Q = I$ in this case. But also in this case $(1 - B_1) = (1 - B_2) = 0$, so all of the eigenvalues $\lambda^Q_{n,\ell}$ listed above are 1—as they must be for $Q = I$.

Proof. We begin with the last case, $n = 2$, $\ell = 0$, which is the most involved. Consider the function

$$\phi = \sum_{i=1}^{N} |v_i|^4, \tag{10.5}$$

and note that $\phi - \int_{X_N} \phi \mathrm{d}\sigma_N$ spans $E_{2,0}$, as we have noted above.

One simple way to calculate $Q\phi$ is to take advantage of the permutation symmetry of Q: Define the symmetrization operator $\mathcal{S}$ by

$$\mathcal{S}f(v_1, \ldots, v_N) = \frac{1}{N!} \sum_{\pi} f(v_{\pi(1)}, \ldots, v_{\pi(N)}),$$

where the sum runs over all permutations π of $\{1, \ldots, N\}$. Then it is easy to see that

$$\mathcal{S}(|v_1|^4) = \frac{1}{N}\phi.$$

Thus, since $\mathcal{S}Q = Q\mathcal{S}$,

$$Q\phi = N\mathcal{S}Q(|v_1|^4).$$

One now directly calculates $Q(|v_1|^4)$ and then symmetrizes. In carrying out the calculation, we make use of the following.

LEMMA 10.2. *Let c and d be any two vectors in $\mathbb{R}^3$, and let e be any unit vector in $\mathbb{R}^3$. Then with B_1 and B_2 defined as in* (1.5)*, we have the following identities:*

$$\int_{S^2} (c \cdot \sigma) b(e \cdot \sigma) \mathrm{d}\sigma = (c \cdot e) B_1$$

and

$$\int_{S^2} (c \cdot \sigma)(d \cdot \sigma) b(e \cdot \sigma) \mathrm{d}\sigma = (c \cdot d) \frac{1 - B_2}{2} + (e \cdot c)(e \cdot d) \frac{3B_2 - 1}{2}.$$

Proof. We choose coordinates in which c and e span the x, z plane with

$$e = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \qquad \text{and} \qquad c = \begin{bmatrix} c^1 \\ 0 \\ c^3 \end{bmatrix}.$$

Then with

$$\sigma = \begin{bmatrix} \sin\theta \cos\psi \\ \sin\theta \sin\psi \\ \sin\theta \end{bmatrix},$$

the computations are easily accomplished. □

Now to compute $Q\phi$, go back to the definition of Q given in (1.4), and note first of all that with $\eta(\vec{v}) = |v_1|^4$, unless $i = 1$,

$$\eta(R_{i,j,\sigma}(\vec{v})) = \eta(\vec{v}).$$

Hence

$$Q\eta(\vec{v}) = \left(1 - \frac{2}{N}\right)\eta(\vec{v}) + \frac{2}{N(N-1)}\sum_{j=2}^{N}\int_{S^2}\eta(R_{1,j,\sigma}(\vec{v}))b\left(\sigma\cdot\frac{v_i - v_j}{|v_i - v_j|}\right)\mathrm{d}\sigma.$$

Then from (1.1),

$$\begin{aligned}\eta(R_{1,j,\sigma}(\vec{v})) &= \left|\frac{v_1 + v_j}{2} + \frac{|v_1 - v_j|}{2}\sigma\right|^4 \\ &= \frac{1}{8}\left||v_1|^2 + |v_j|^2 + |v_1 - v_j|(v_1 + v_j)\cdot\sigma\right|^2 \\ &= \frac{1}{8}\Big((|v_1|^2 + |v_j|^2)^2 + 2(|v_1|^2 + |v_j|^2)|v_1 - v_j|(v_1 + v_j)\cdot\sigma \\ &\qquad + |v_1 - v_j|^2((v_1 + v_j)\cdot\sigma)^2\Big). \end{aligned} \tag{10.6}$$

Integrating over S^2 using Lemma 10.2 yields

$$\begin{aligned}\int_{S^2}\eta(R_{1,j,\sigma}(\vec{v}))b\left(\sigma\cdot\frac{v_i - v_j}{|v_i - v_j|}\right)\mathrm{d}\sigma &= \frac{1}{8}(|v_1|^2 + |v_j|^2)^2 \\ &\quad + \frac{1}{4}(|v_1|^2 + |v_j|^2)B_1(|v_1|^2 - |v_j|^2) \\ &\quad + |v_1 - v_j|^2|v_1 + v_j|^2\frac{1 - B_2}{16} \\ &\quad + ((v_1 - v_j)\cdot(v_1 + v_j))^2\frac{3B_2 - 1}{16}. \end{aligned} \tag{10.7}$$

The right-hand side simplifies to

$$\begin{aligned}&\frac{1}{8}(|v_1|^4 + |v_j|^4 + 2|v_1|^2|v_j|^2) + \frac{B_1}{4}(|v_1|^4 - |v_j|^4) \\ &\quad + (|v_1|^4 + |v_j|^4 + 2|v_1|^2|v_j|^2 - 4(v_1\cdot v_j)^2)\frac{1 - B_2}{16} \\ &\quad + (|v_1|^4 + |v_j|^4 - 2|v_1|^2|v_j|^2)\frac{3B_2 - 1}{16}. \end{aligned} \tag{10.8}$$

It is now a simple matter to carry out the sum on $j \geq 2$. By using the identities

$$\sum_{j=2}^{N}|v_j|^4 = \phi(\vec{v}) - |v_1|^4 \qquad \text{and} \qquad \sum_{j=2}^{N}|v_j|^2 = \phi(\vec{v}) - |v_1|^2$$

and the symmetrizing, one finds that

$$Q\phi = \phi - \frac{1 - B_2}{N}\left[\frac{N+1}{N-1}\phi + \frac{1}{(N-1)}\psi - \frac{2}{(N-1)}\right], \tag{10.9}$$

where

$$\psi = \sum_{i\neq j}(v_i\cdot v_j)^2. \tag{10.10}$$

For $N \geq 4$, the two functions ϕ and ψ are linearly independent, although for $N = 3$ they are not. In fact for $N = 3$, one has the identity

$$\psi = 2\phi - \frac{1}{2}. \tag{10.11}$$

Evidently, for $N \geq 4$, $\phi - \int_{X_N} \phi \mathrm{d}\sigma_N$ is not an eigenfunction of Q, so that $E_{0,2}$ is not an eigenspace of Q. We are required to compute $Q\psi$.

We again take advantage of the permutation symmetry and note that

$$\mathcal{S}((v_1 \cdot v_2)^2) = \frac{2}{N(N-1)}\psi \qquad \text{and} \qquad Q\psi = \frac{N(N-1)}{2}\mathcal{S}(Q(v_1 \cdot v_2)^2).$$

We carry out the calculation in the same way that we calculated $Q\phi$ and find that

$$Q\psi = \psi - \frac{1 - B_2}{N}\left[3\psi + \frac{(N-3)}{(N-1)}\phi - \frac{1}{N}\right]. \tag{10.12}$$

We see that the subspace spanned by $\phi - \int_{X_N} \phi \mathrm{d}\sigma_N$ and $\psi - \int_{X_N} \phi \mathrm{d}\sigma_N$ is invariant under Q. By using (10.12) and (10.9) we easily find that the two eigenvalues of $N(I - Q)$ on the two-dimensional space $V_{2,0}$ are the eigenvalues of

$$\frac{1 - B_2}{N-1}\begin{bmatrix} N+1 & 1 \\ N-3 & 3N-3 \end{bmatrix},$$

which are

$$\frac{1 - B_2}{N-1}\left[(2N-1) \pm \sqrt{N^2 - 3N + 1}\right].$$

The minus sign clearly gives the lesser of these and gives the gap for $N(I - Q)$ on $V_{2,0}$. From here, one easily deduces (10.4).

A further, much simpler calculation shows that the three functions

$$\psi^\alpha_{1,1} = \sum_{k=1}^{N} |v_k|^2 v^\alpha_k, \tag{10.13}$$

where α indexes the components, are also eigenfunctions of Q; more precisely,

$$Q\psi^\alpha_{1,1} = \left(1 - \frac{1 - B_2}{N-1}\right)\psi^\alpha_{1,1}. \tag{10.14}$$

Thus the unique eigenvalue of A on $V_{1,1}$ is

$$\lambda^Q_{1,1} = 1 - \frac{1 - B_2}{N-1}.$$

For $E_{0,2}$, a simple computation shows that the functions

$$\psi^{\alpha,\beta}_{0,2} = \sum_k v^\alpha_k v^\beta_k, \tag{10.15}$$

where $\alpha \neq \beta$ are indices for the components, are also eigenfunctions for Q; in fact,

$$Q\psi^{\alpha,\beta}_{0,2} = \left(1 - \frac{3(1 - B_2)}{2(N-1)}\right)\psi^{\alpha,\beta}_{0,2}. \tag{10.16}$$

Thus, $V_{0,2} = E_{0,2}$, and the unique eigenvalue of Q on this subspace is

$$\lambda^Q_{0,2} = 1 - \frac{1}{N-1}.$$

Finally, we consider the spectrum on Q on the eigenspaces of P corresponding to $n + \ell = 1$. In this case, as noted above, the eigenfunctions are antisymmetric, so that if we are concerned only with the spectrum of Q on the subspace of symmetric functions (which is all that is of significance for Kac's application to the Boltzmann equation), we can ignore these eigenspaces. However, if we define $\eta_{0,1}(\vec{v}) = v_1 - v_2$ and $\eta_{1,0}(\vec{v}) = |v_1|^2 - |v_2|^2$, we find, as above, that

$$Q\eta_{1,0} = \left(1 - (1 - B_1)\frac{1}{N-1}\right)\eta_{1,0} \qquad \text{and} \tag{10.17}$$

$$Q\eta_{0,1} = \left(1 - (1 - B_1)\frac{1}{N-1}\right)\eta_{0,1}. \qquad \square \tag{10.18}$$

Now that we have all of our eigenvalues, we need to order them. By a simple comparison, we determine that for all N the largest eigenvalue of Q on our three invariant subspaces with $n + \ell = 2$ is $\lambda^Q_{1,1}$. This is true for all choices of b, since the only dependence on b in these eigenvalues is a common factor of $(1 - B_2)$.

It is worth noting, however, that for large N

$$\lambda^Q_{1,1} = 1 - (1 - B_2)\frac{1}{N} + \mathcal{O}\left(\frac{1}{N^2}\right) \qquad \text{and} \qquad \lambda^Q_{2,0} = 1 - (1 - B_2)\frac{1}{N} + \mathcal{O}\left(\frac{1}{N^2}\right),$$

so that these eigenvalues merge as N tends to infinity. Still, for all finite N,

$$\lambda^Q_{1,1} = 1 - (1 - B_2)\frac{1}{N-1}$$

is strictly larger.

Next, the invariant subspaces of Q with $n + \ell = 1$ are also eigenspaces of Q with the eigenvalue

$$\lambda^Q_{0,1} = \lambda^Q_{1,0} = 1 - (1 - B_1)\frac{1}{N-1}.$$

In summary, the largest eigenvalue of Q on the invariant subspaces $V_{n,\ell}$ in $L^2(X_N, \mathrm{d}\sigma_N)$ with $n + \ell = 1, 2$ and $N \geq 3$ is either

$$1 - (1 - B_2)\frac{1}{N-1} \qquad \text{or} \qquad 1 - (1 - B_1)\frac{1}{N-1},$$

depending on which of these is larger. In particular,

$$\Delta_3 \leq \min\{(1 - B_2), (1 - B_1)\}\frac{3}{2}. \tag{10.19}$$

With the above arguments we have all of the ingredients needed to prove Theorem 1.1.

Proof of Theorem 1.1. First, we wish to apply Lemma 4.1 to estimate Δ_3 in terms of Δ_2. In (9.4), we have set $\mu^\star_3 = 11/20$, and with this choice of the threshold, we have seen that there is just one eigenvalue of P between $\mu^\star_3$ and 1, namely, the gap

eigenvalue $\mu_3 = \mu_{1,1}(3) = \mu_{0,1}(3) = \mu_{1,0}(3)$. Thus, from Lemma 4.1 and the eigenvalue computations in Lemma 10.1, either the gap eigenvalue of Q for $N = 3$ is

$$\max\left\{\ 1-(1-B_2)\frac{1}{N-1},\ 1-(1-B_1)\frac{1}{N-1}\ \right\} \tag{10.20}$$

or else

$$\Delta_3 \geq \frac{3}{2}\left(1-\frac{11}{20}\right)\Delta_2 = \frac{27}{40}\Delta_2. \tag{10.21}$$

If (10.20) does give the gap eigenvalue of Q for $N = 3$, then

$$\Delta_3 = \min\{\ (1-B_1),\ (1-B_2)\ \}\frac{3}{2}. \tag{10.22}$$

Since, according to Lemma 4.1, at least one of (10.21) and (10.22) is true, the condition

$$\frac{27}{40}\Delta_2 \geq \frac{3}{2}\min\{\ (1-B_1),\ (1-B_2)\ \} \tag{10.23}$$

and (10.19) ensure that (10.22) is true and thus give us the gap eigenvalue for $N = 3$. Note that the condition (10.23) is equivalent to the condition (1.6) in Theorem 1.1.

Now we proceed by induction. For any $n \geq 4$, assume that

$$\Delta_{N-1} = \min\{\ (1-B_1),\ (1-B_2)\ \}\frac{N-1}{N-2}. \tag{10.24}$$

In (9.3) we have set

$$\mu_N^\star = \frac{1}{N-1}$$

for all $N \geq 4$, and we have seen that the only eigenvalues μ of P with $1 > \mu \geq \mu_N^\star$ are the gap eigenvalue $\mu_N = \mu_{1,1}(N) = \mu_{0,1}(N) = \mu_{1,0}(N)$, and for $N > 4$, $\mu_{2,0}(N)$. Thus, by Lemma 4.1 and the eigenvalue computations in Lemma 10.1, either the gap eigenvalue of Q for N is

$$\max\left\{\ 1-(1-B_2)\frac{1}{N-1},\ 1-(1-B_1)\frac{1}{N-1}\ \right\} \tag{10.25}$$

or else

$$\Delta_N > \frac{N}{N-1}\left(1-\frac{1}{N-1}\right)\Delta_{N-1}. \tag{10.26}$$

There is strict inequality in (10.26) since all remaining eigenvalues of P not taken into account in (10.25) are strictly less than $\mu_N^\star$. By the inductive hypothesis (10.24) yields

$$\Delta_N > \min\{\ (1-B_1),\ (1-B_2)\ \}\frac{N}{N-1}.$$

This is impossible, as the trial functions leading to (10.25) yield the upper bound

$$\Delta_N \leq \min\{\ (1-B_1),\ (1-B_2)\ \}\frac{N}{N-1}. \tag{10.27}$$

Thus equality holds in (10.27), which completes the proof of the inductive step. Because of the strict inequality in (10.26), the only eigenfunctions with the gap eigenvalue are found in the invariant subspaces considered here, i.e., in the $V_{n,\ell}$ with $0 < n+\ell \leq 2$. By the results of Lemma 10.1, this yields the statement in Theorem 1.1 concerning the gap eigenfunctions of Q. □

Proof of Theorem 1.2. We proceed as in the previous proof except that, for low values of N, we shall use a different choice for the threshold $\mu_N^\star$, namely,

$$\mu_N^\star = \mu_{2,2}(N). \tag{10.28}$$

We know from Lemma 9.1 that for all $N \leq 7$

$$\mu_{n,\ell}(N) \leq \mu_{2,2}(N) \quad \text{for all} \quad n+\ell > 2.$$

Thus, at least for such N, the only eigenvalues μ of P with $\mu > \mu_N^\star = \mu_{2,2}(N)$ are those with $n+\ell \leq 2$. We have already computed the gap for Q on the invariant subspaces containing these eigenvalues, and we have found that the gap in these subspaces is

$$\widetilde{\Delta}_N = \min\{\ (1-B_1),\ (1-B_2)\ \}\frac{N}{N-1}.$$

If for any $N_0 \geq 3$ it turns out that $\widetilde{\Delta}_{N_0} = \Delta_{N_0}$, the gap on the whole space, then we can switch from that point onwards to the use of $\mu_N^\star = \mu_{0,2}(N)$ as in the proof of Theorem 1.1 to show that $\widetilde{\Delta}_N = \Delta_N$ for all $N \geq N_0$ and that the eigenfunctions are exactly as claimed for all $N > N_0$.

We now show that it is always the case that $\widetilde{\Delta}_{N_0} = \Delta_{N_0}$ for some $N_0 \leq 7$. To do this, pick any value $N_1 \geq 4$, and suppose that for $3 \leq j \leq N_1$ we have

$$\Delta_j < \min\{\ (1-B_1),\ (1-B_2)\ \}\frac{j}{j-1}. \tag{10.29}$$

Then by Lemmas 4.1 and 9.1, by using the value $\mu_j^\star = \mu_{2,2}(j)$, for $3 \leq j \leq N_1$, we have

$$\Delta_{N_1} \geq \frac{N_1}{2}\prod_{j=3}^{N_1}(1-\mu_{2,2}(j))\Delta_2.$$

By using the hypothesis $\Delta_2 = 2(1-B_1)$, we have

$$\Delta_{N_1} \geq \frac{N_1}{2}\prod_{j=3}^{N_1}(1-\mu_{2,2}(j))2(1-B_1).$$

Of course we can rewrite this as

$$\Delta_{N_1} \geq \frac{N_1}{2}\left(\prod_{j=4}^{N_1}(1-\mu_{0,2}(j))\right)(1-\mu_{2,2}(3))\left(\prod_{j=4}^{N_1}\frac{(1-\mu_{2,2}(j))}{(1-\mu_{0,2}(j))}\right)2(1-B_1)$$

$$= \frac{N_1}{N_1-1}\left(\prod_{j=4}^{N_1}\frac{(1-\mu_{2,2}(j))}{(1-\mu_{0,2}(j))}\right)\frac{9}{10}(1-B_1), \tag{10.30}$$

since, as in the last proof, $(1-\mu_{2,2}(3)) = 9/20$, and

$$\frac{N_1}{2}\prod_{j=4}^{N_1}(1-\mu_{0,2}(j)) = \frac{N_1}{N_1-1}.$$

Now, by direct computation, we find that

$$\prod_{j=4}^{7}\frac{(1-\mu_{2,2}(j))}{(1-\mu_{0,2}(j))} = \frac{558018643}{495720000} > \frac{10}{9}.$$

For $N_1 \geq 7$, this would lead to

$$\Delta_{N_1} > \frac{N_1}{N_1-1}(1-B_1),$$

and this is impossible, since we have a trial function showing that the gap cannot be so large. Hence it must be that (10.29) is false for some $j \leq 7$. By what we have said above, from this point onward, we can proceed as in the proof of Theorem 1.1, and we obtain Theorem 1.2. □

While the results presented here cover a very wide range of models, it is possible to come up with choices of b for which $\Delta_2 \neq 2(1-B_1)$. If one found a need to deal with such an example, one might have to go deeper into the spectrum of P. It is very likely that Lemma 9.1 holds for all $N \geq 3$, based on extensive computations. These computations also show that, as N increases, $\mu_{2,1}(N)$ comes very close to $\mu_{2,2}(N)$, so that to get much more leverage one would need to compute all of the eigenvalues of Q on the smallest invariant subspaces of Q that contain both of these eigenspaces of P. This could be done by using the methods illustrated above, but the computations would be considerably more involved than the ones we have presented in this section. Thus, having treated a wide range of models, we shall conclude our discussion of Q here. In the brief final section, we discuss a point we raised earlier concerning bounds on Jacobi polynomials.

11. Bounds on Jacobi polynomials. As alluded to in section 8, the identity (7.9), together with the trivial bound on the $|\kappa_{n,\ell}| \leq 1$, which comes from the fact that K is a Markov operator, will for certain regions provide a stronger bound than (8.8), the bound of Nevai, Erdelyi, and Magnus. We close this section by showing how (7.9) can be used to obtain better bounds.

To begin, write

$$\begin{aligned} & b^{2\beta-1}\frac{P_n^{\alpha,\beta}(-1+2b^2)}{P_n^{\alpha,\beta}(1)} \\ & \leq \frac{2e}{\pi}\frac{\Gamma(n+1)}{b(1-b^2)^{\alpha+1/2}}\frac{2+\sqrt{\alpha^2+\beta^2}}{2n+\alpha+\beta+1}\frac{\Gamma(n+\beta+1)}{\Gamma(n+\alpha+\beta+1)}\frac{\Gamma(\alpha+1)^2}{\Gamma(n+\alpha+1)}, \end{aligned} \tag{11.1}$$

where $\beta = l+1/2$, with l an integer. In regions where the right-hand side of the above equation becomes larger than one, the simple bound

$$b^{2\beta-1}\frac{P_n^{\alpha,\beta}(-1+2b^2)}{P_n^{\alpha,\beta}(1)} \leq 1$$

becomes stronger. In the region $2n+1 < \alpha < \beta$, we find $\frac{2+\sqrt{\alpha^2+\beta^2}}{2n+\alpha+\beta+1} > \frac{1}{4}$. This plus Stirling's formula with the remainder yields

$$\frac{2e}{\pi}\frac{\Gamma(n+1)}{b(1-b^2)^{\alpha+1/2}}\frac{2+\sqrt{\alpha^2+\beta^2}}{2n+\alpha+\beta+1}\frac{\Gamma(n+\beta+1)}{\Gamma(n+\alpha+\beta+1)}\frac{\Gamma(\alpha+1)^2}{\Gamma(n+\alpha+1)}$$
$$> \frac{e^n}{\sqrt{2\pi}}\frac{\Gamma(n+1)}{b(1-b^2)^{\alpha+1/2}}\frac{\alpha^{\alpha+1/2-n}\beta^{-\alpha}}{(1+\frac{\alpha}{\beta})^{n+\alpha+\beta+1}}\frac{(1+\frac{1}{\alpha})^{2\alpha+1}}{(1+\frac{n+1}{\alpha})^{n+\alpha+1/2}(1+\frac{n+1}{\alpha+\beta})^{\beta}} * r$$
$$> \frac{e^n}{\sqrt{2\pi}}\frac{\Gamma(n+1)}{b(1-b^2)^{\alpha+1/2}}\frac{\alpha^{\alpha+1/2-n}\beta^{-\alpha}}{2^{2n+2\alpha+2\beta+3/2}}r,$$

where $r = (1-\frac{1}{12(n+\alpha+\beta+1)})(1-\frac{1}{12(n+\alpha+1)})$ and n is assumed to be fixed. Choosing $b(1-b^2)^{\alpha+1/2}$ so that the last inequality is greater than one provides a region where the combination of (7.9) and $|\kappa_{n,\ell}| \le 1$ does better than (7.9). It would be interesting to obtain better bounds on $|\kappa_{n,\ell}|$ by direct analysis of K and to use these to sharpen the argument just made.

Acknowledgment. We thank Doron Lubinsky for valuable discussions concerning (8.8), the bound of Nevai, Erdelyi, and Magnus, and related results.

REFERENCES

[1] R. Askey, *Orthogonal Polynomials and Special Functions*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 21, SIAM, Philadelphia, 1975.
[2] E. Carlen, M. C. Carvalho, and M. Loss, *Many–body aspects of approach to equilibrium*, J. EDP, 11 (2000), pp. 1–12.
[3] E. Carlen, M. Carvalho, and M. Loss, *Determination of the spectral gap for Kac's master equation and related stochastic evolution*, Acta Math., 191 (2003), pp. 1–54.
[4] E. Carlen and X. Lu, *Fast and slow convergence to equilibrium Maxwellian molecules via Wild sums*, J. Statist. Phys., 112 (2003), pp. 59–134.
[5] I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals Series and Products*, Academic Press, New York, 1965.
[6] E. Janvresse, *Spectral Gap for Kac's model of Boltzmann Equation*, Ann. Probab., 29 (2001), pp. 288–304.
[7] M. Kac, *Foundations of kinetic theory*, in Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability, J. Neyman, ed., Vol. 3, University of California, Berkeley, 1956, pp. 171–197.
[8] T. H. Koornwinder, *The addition formula for Jacobi polynomials.* I, *summary of results*, Indag. Math., 34 (1972), pp. 188–191.
[9] D. Maslen, *The eigenvalues of Kac's master equation*, Math. Z., 243 (2003), pp. 291–331.
[10] D. Morgenstern, *General existence and uniqueness proof for spatially homogeneous solutions of the Maxwell-Boltzmann equation in the case of Maxwellian Molecules*, Proc. Natl. Acad. Sci. USA, 40 (1954), pp. 719–721.
[11] D. Morgenstern, *Analytical studies related to the Maxwell-Boltzmann equation*, J. Ration. Mech. Anal., 4 (1955), pp. 154–183.
[12] P. Nevai, T. Erdélyi, and A. Magnus, *Generalized Jacobi weights, Christoffel functions, and Jacobi polynomials*, SIAM J. Math. Anal., 25 (1994), pp. 602–614.
[13] G. Szegö, *Orthogonal Polynomials*, Amer. Math. Soc. Colloq. Publ. 23, AMS, Providence, RI, 1939.

SIAM J. MATH. ANAL.
Vol. 40, No. 1, pp. 365–381

ORBITAL STABILITY OF BOUND STATES OF SEMICLASSICAL NONLINEAR SCHRÖDINGER EQUATIONS WITH CRITICAL NONLINEARITY*

TAI-CHIA LIN† AND JUNCHENG WEI‡

Abstract. We consider the orbital stability of single-spike bound states of semiclassical nonlinear Schrödinger equations with critical nonlinearity and a trap potential. Due to the effect of the trap potential, we derive the asymptotic expansion formulas and obtain the necessary conditions for orbital stability and instability of single-spike bound states. Our argument is applied to two-component systems of nonlinear Schrödinger equations with a common trap potential, cubic nonlinearity in two spatial dimensions. The orbital stability of bound states with spikes of these systems is investigated. Our results show the existence of stable spikes in two-dimensional Bose–Einstein condensates.

Key words. orbital stability, spike, trap potential

AMS subject classifications. 35J50, 35Q55

DOI. 10.1137/070683842

1. Introduction.

The nonlinear Schrödinger (NLS) equation with a trap potential is central to the understanding of many physical phenomena. For example, it has become a well-known model referred to as the Gross–Pitaevskii equation governing the evolution of Bose–Einstein condensates (BEC) given by

$$-i\hbar\frac{\partial\psi}{\partial t} = \frac{\hbar^2}{2m}\triangle\psi - V_{trap}\psi - \mu|\psi|^2\psi \tag{1.1}$$

for $x \in \mathbb{R}^N$, $N \leq 3$, and $t > 0$, where $\psi = \psi(x,t) \in \mathbb{C}$ is the wavefunction of BEC and $V_{trap} = V_{trap}(x)$ is the trap potential. Also, $\hbar$ is the Planck constant, m is the atom mass, and $\mu \sim 4\pi\frac{\hbar^2}{2m}a$, where a denotes the s-wave scattering length.

In BEC, spikes may occur when the s-wave scattering length is negative and large. Due to Feshbach resonance, the s-wave scattering length of a single condensate can be tuned over a very large range by adjusting the externally applied magnetic field. As the s-wave scattering length of a single condensate is negative and large enough, the interactions of atoms are strongly attractive and the associated condensate tends to increase its density at the center of the trap potential in order to lower the interaction energy (cf. [28]). Under the effect of trap potentials, spikes of BEC are observable by physical experiments (cf. [10]) so there must be stability to ensure spikes appearing in the condensate wavefunction (cf. [7]). In [24], stable bright solitons (spikes) of BEC can be observed by numerical simulations, provided that the strength of the trap potential exceeds a threshold value. Here we want to develop mathematical theorems to support the existence of stable spikes in BEC.

*Received by the editors February 27, 2007; accepted for publication (in revised form) August 15, 2007; published electronically April 25, 2008.
http://www.siam.org/journals/sima/40-1/68384.html
†Department of Mathematics, National Taiwan University, Taipei, 106 Taiwan (tclin@math.ntu.edu.tw). The research of this author was partially supported by a research grant from NSC of Taiwan.
‡Department of Mathematics, The Chinese University of Hong Kong, Shatin, Hong Kong (wei@math.cuhk.edu.hk). The research of this author was partially supported by an Earmarked Grant from RGC of Hong Kong.

To get spikes in BEC, we may assume the s-wave scattering length a, i.e., μ is negative and large. Setting $h^2 = \hbar^2/(2m\mu)$, $V_{trap}(x) = \mu^{-1} V(x)$, and suitable time scale, (1.1) with negative and large μ can be equivalent to a semiclassical NLS given by

$$-ih\frac{\partial \psi}{\partial t} = h^2 \triangle \psi - V\psi + |\psi|^2\psi, \quad x \in \mathbb{R}^N,\ t > 0, \tag{1.2}$$

where $0 < h \ll 1$ is a small parameter (semiclassical limit) and $V = V(x)$ is a smooth nonnegative function. We may generalize (1.2) to an NLS equation having the form

$$-ih\frac{\partial \psi}{\partial t} = h^2 \triangle \psi - V\psi + |\psi|^{p-1}\psi, \quad x \in \mathbb{R}^N,\ t > 0, \tag{1.3}$$

with critical nonlinearity

$$p = 1 + \frac{4}{N}\,, \quad N \geq 1. \tag{1.4}$$

In particular, when $N = 2$, (1.3) is exactly the same as (1.2).

Bound states of (1.3) are of the form $\psi(x,t) = e^{i\lambda t/h}u(x)$, where $\lambda > 0$ and u satisfies the following nonlinear elliptic equation:

$$h^2 \triangle u - (V + \lambda)u + u^p = 0\,, \quad u \in H^1(\mathbb{R}^N)\,, \quad u > 0 \quad \text{in} \quad \mathbb{R}^N\,, \tag{1.5}$$

with zero Dirichlet boundary condition, i.e., $u(x) \to 0$ as $|x| \to \infty$.

In the case when $V(x) \equiv 0$, for any $\lambda > 0$, problem (1.5) admits a unique radially symmetric ground state, which is stable for any $\lambda > 0$ if $p < 1 + \frac{4}{N}$ and unstable for any $\lambda > 0$ if $p \geq 1 + \frac{4}{N}$ (see [5], [6], and [36]).

When $V(x) \not\equiv 0$, the existence of single- or multiple-spike solutions of (1.5) was first established by Floer and Weinstein [11] in the one-dimensional case, i.e., $N = 1$ and $1 < p < 5$, and later extended by Oh [25], [26] to the higher-dimensional case, i.e., $N \geq 2$ and $1 < p < \frac{N+2}{N-2}$ under the condition that the trap potential V has nondegenerate critical points. When the trap potential V becomes degenerate, there have been many works in recent years. The reader may refer to [1], [3], [17], [8], [9] [19], [29], [30], [32], [33], and the references therein.

The trap potential V may also play a crucial role in the orbital (dynamic) stability of single-spike bound states. As the trap potential V is switched off, it is well known that all bound states of (1.3) with the condition (1.4) are orbitally unstable if the dimension $N = 2$ (cf. [37]). To stabilize bound states, one has to turn on the trap potential. However, in general, some nonzero trap potentials may still cause dynamic instability in BEC. For instance, one may find bending-wave instability of vortex ring dynamics under some nonzero trap potentials (cf. [18]). Consequently, to get the dynamic stability of single-spike bound states, we have to choose trap potentials properly. For suitable trap potentials, Oh [26] and Grillakis, Shatah, and Strauss [14] proved that when $N = 1$, the single-spike bound state (concentrating at local nondegenerate minimum of the trap potential V) is stable if $1 < p < 1 + \frac{4}{N}$ and unstable if $p > 1 + \frac{4}{N}$. Generically, the case of $p = 1 + \frac{4}{N}$ is left open and referred to as a critical case in the literature. In this paper, we give an affirmative answer for such a case by studying the orbital stability and instability of single-spike bound states when the trap potential V has nondegenerate critical points.

In [25] and [26], a single-spike bound state solution u_h of (1.5) can be obtained, provided the trap potential V is of class $(V)_a$ and fulfills other conditions. Hereafter,

we set u_h as a single-spike bound state constructed in [25] and [26] and satisfying (1.5). Of course, the trap potential V is also of class $(V)_a$ and fulfills other conditions in [25] and [26]. Hence $\psi_h(x,t) = e^{i\lambda t/h}u_h(x)$ may form an orbit of (1.3). From [14], the orbital stability of ψ_h's is defined as follows: For all $\epsilon > 0$, there exists $\delta > 0$ such that if $\|\psi_0 - u_h\|_{H^1} < \delta$ and ψ is a solution of (1.3) in some interval $[0, t_0)$ with $\psi|_{t=0} = \psi_0$, then $\psi(t,\cdot)$ can be extended to a solution in $0 \le t < \infty$ and

$$\sup_{0<t<\infty} \inf_{s\in\mathbb{R}} \|\psi(\cdot,t) - \psi_h(\cdot,s)\|_{H^1} < \epsilon\,.$$

Otherwise, the orbit ψ_h is called orbitally unstable. To check the orbital stability of ψ_h, we use the linearized operator defined by

$$L_h = h^2\triangle - (V+\lambda) + pu_h^{p-1}, \quad p = 1 + \frac{4}{N}\,. \tag{1.6}$$

Observe that u_h may depend on the variable λ. Moreover, we assume u_h to be nondegenerate due to [16]. Let $n(L_h)$ be the number of positive eigenvalues of L_h and

$$d(\lambda) = \int_{\mathbb{R}^N} \left[\frac{h^2}{2}|\nabla u_h|^2 + \frac{1}{2}(V+\lambda)u_h^2 - \frac{1}{p+1}u_h^{p+1}\right]. \tag{1.7}$$

Assume that d is nondegenerate, i.e., $d'' \neq 0$. Let $p(d'') = 1$ if $d'' > 0$ and $p(d'') = 0$ if $d'' < 0$. According to the general theory of orbital stability of bound states (cf. [14], [15]), u_h is orbitally stable if $n(L_h) = p(d'')$ and orbitally unstable if $n(L_h) - p(d'')$ is odd (see page 309 of [15]).

It is worth noting that if $V \equiv C$ and $p = 1 + \frac{4}{N}$, then $d''(\lambda) = 0$, i.e., the function d becomes degenerate, where C is a positive constant. Consequently, we may assume that the trap potential V has nondegenerate critical points in order to derive the asymptotic expansion formulas for the operator L_h and the function d as the parameter h goes to zero. These formulas are crucial to obtaining the orbital stability and instability of single-spike bound states as follows.

THEOREM 1.1. *Let N be a positive integer and $p = 1 + \frac{4}{N}$. For $0 < h < 1$, let u_h be a bound state of* (1.3) *concentrated at a nondegenerate critical point x_0 of the potential V such that $\Delta V(x_0) \neq 0$. Let m denote the number of negative eigenvalues of the matrix $(\nabla^2 V(x_0))$. Suppose the parameter h is sufficiently small. Then u_h is orbitally stable if x_0 is a nondegenerate local minimum point of the potential V. Furthermore, u_h is orbitally unstable if $m - \frac{1}{2}(1 + \frac{\Delta V(x_0)}{|\Delta V(x_0)|})$ is even.*

Remark. In [12], Fukuizumi considered the orbital stability of the standing wave solution to (1.3) in the critical case $p = 1 + \frac{4}{N}$ under the following conditions:

$$h = 1, \quad V(x) = V(|x|), u(x,t) = u(|x|,t). \tag{1.8}$$

He studied the stability of the standing wave solution $e^{i\lambda t}u(x)$ for λ large (*in the radially symmetric class*). After suitable scaling, the standing wave solution satisfies

$$\epsilon^2\Delta u - (1 + \epsilon^2 V(|x|))u + u^p = 0, u > 0, \quad \text{where } \epsilon^2 = \frac{1}{\lambda}. \tag{1.9}$$

There is a slightly subtle difference between (1.5) and (1.9). However, the main difference is that we consider the *full functional space* here. The method of proving Theorem 1.1 can also be applied to obtain a more general result to Fukuizumi's problem for general $V(x)$.

Another motivation of studying (1.3) in the critical case may come from two-component systems of NLS equations which describe a double condensate, i.e., a binary mixture of BEC (cf. [28]). To get stable spikes of a double condensate with two spatial dimensions, we study orbitally stable bound states with spikes of a two-component system of NLS equations given by

$$\begin{cases} -ih\dfrac{\partial \Phi}{\partial t} = h^2\triangle\Phi - V\Phi + |\Phi|^2\Phi + \beta|\Psi|^2\Phi\,, \\ -ih\dfrac{\partial \Psi}{\partial t} = h^2\triangle\Psi - V\Psi + |\Psi|^2\Psi + \beta|\Phi|^2\Psi \end{cases} \tag{1.10}$$

for $x \in \mathbb{R}^2$ and $t > 0$, where $V = V(x)$ is a smooth nonnegative function, $\beta \in \mathbb{R}$ is a nonzero constant, and $0 < h \ll 1$ is a small parameter. Bound states of (1.10) are of the form $\Phi(x,t) = e^{i\lambda t/h}u(x)$ and $\Psi(x,t) = e^{i\lambda t/h}v(x)$, where (u,v) satisfies the following nonlinear elliptic system:

$$\begin{cases} h^2\triangle u - (V+\lambda)u + u^3 + \beta uv^2 = 0, & x \in \mathbb{R}^2, \\ h^2\triangle v - (V+\lambda)v + v^3 + \beta u^2 v = 0, & x \in \mathbb{R}^2, \\ u(x), v(x) > 0, & u, v \in H^1(\mathbb{R}^2). \end{cases} \tag{1.11}$$

Note that in $\mathbb{R}^2$, the nonlinearity u^3, v^3 are a critical nonlinearity by the simple algebra $p = 3 = 1 + \frac{4}{N}$ with $N = 2$. The system (1.11) admits a bound state solution $(\frac{1}{\sqrt{1+\beta}}u_h, \frac{1}{\sqrt{1+\beta}}u_h)$, where $\beta > -1$ and u_h satisfies (1.5). Generically, such a solution may be neither a unique positive solution nor a ground state solution of the system (1.11). Thus the stability problem is nontrivial. Here we want to get the orbital stability of such a solution using suitable trap potentials V's. To study the orbital stability of such a bound state solution, we set the linearized operator of (1.10) around $(\frac{1}{\sqrt{1+\beta}}u_h, \frac{1}{\sqrt{1+\beta}}u_h)$ given by

$$\mathbb{L}_h \begin{pmatrix} \phi \\ \psi \end{pmatrix} = \begin{pmatrix} h^2\triangle\phi - (V+\lambda)\phi + \frac{3+\beta}{1+\beta}u_h^2\phi + \frac{2\beta}{1+\beta}u_h^2\psi \\ h^2\triangle\psi - (V+\lambda)\psi + \frac{3+\beta}{1+\beta}u_h^2\psi + \frac{2\beta}{1+\beta}u_h^2\phi \end{pmatrix}. \tag{1.12}$$

Furthermore, we also need a function defined as follows:

$$\begin{aligned} d(\lambda_1,\lambda_2) = &\int_{\mathbb{R}^2} \frac{h^2}{2}|\nabla u_{h,\lambda_1,\lambda_2}|^2 + \frac{V(x)+\lambda+\lambda_1}{2}u_{h,\lambda_1,\lambda_2}^2 - \frac{1}{4}\int_{\mathbb{R}^2} u_{h,\lambda_1,\lambda_2}^4 \\ &+ \int_{\mathbb{R}^2} \frac{h^2}{2}|\nabla v_{h,\lambda_1,\lambda_2}|^2 + \frac{V(x)+\lambda+\lambda_2}{2}v_{h,\lambda_1,\lambda_2}^2 - \frac{1}{4}\int_{\mathbb{R}^2} v_{h,\lambda_1,\lambda_2}^4 \\ &- \frac{\beta}{2}\int_{\mathbb{R}^2} u_{h,\lambda_1,\lambda_2}^2 v_{h,\lambda_1,\lambda_2}^2\,, \end{aligned} \tag{1.13}$$

where $(u_{h,\lambda_1,\lambda_2}, v_{h,\lambda_1,\lambda_2})$ is the solution of

$$\begin{cases} h^2\triangle u - (V+\lambda+\lambda_1)u + u^3 + \beta uv^2 = 0 \quad \text{in } \mathbb{R}^2\,, \\ h^2\triangle v - (V+\lambda+\lambda_2)v + v^3 + \beta u^2 v = 0 \quad \text{in } \mathbb{R}^2\,, \end{cases} \tag{1.14}$$

such that $(u_{h,\lambda_1,\lambda_2}, v_{h,\lambda_1,\lambda_2}) \to (\frac{1}{\sqrt{1+\beta}}u_h, \frac{1}{\sqrt{1+\beta}}u_h)$ as $|\lambda_1| + |\lambda_2| \to 0$.

Suppose the solution $(\frac{1}{\sqrt{1+\beta}}u_h, \frac{1}{\sqrt{1+\beta}}u_h)$ is nondegenerate, i.e., the operator $\mathbb{L}_h$ has no zero eigenvalue. Let $n(\mathbb{L}_h)$ denote the positive eigenvalues of $\mathbb{L}_h$, and set p as the number of positive eigenvalues of the Hessian matrix $(\nabla^2 d(0,0))$. From [14] and [15], we know that the solution $(\frac{1}{\sqrt{1+\beta}}u_h, \frac{1}{\sqrt{1+\beta}}u_h)$ is orbitally stable if $n(\mathbb{L}_h) = p$ and orbitally unstable if $n(\mathbb{L}_h) - p$ is odd. The parameter β may affect the orbital stability of the solution $(\frac{1}{\sqrt{1+\beta}}u_h, \frac{1}{\sqrt{1+\beta}}u_h)$. Now we state our result as follows.

THEOREM 1.2. *For $0 < h < 1$, let u_h be a single-spike solution concentrated at a local minimum point of the function V. Suppose the parameter h is sufficiently small. Then $(\frac{1}{\sqrt{1+\beta}}u_h, \frac{1}{\sqrt{1+\beta}}u_h)$ is an orbitally stable solution to* (1.10) *if $0 < \beta \neq 1$.*

Remark. The orbital instability of $(\frac{1}{\sqrt{1+\beta}}u_h, \frac{1}{\sqrt{1+\beta}}u_h)$ for $-1 < \beta < 0$ can also be investigated. However, the condition is quite complicated so we may omit it here. On the other hand, as $\beta = 1$, the system (1.11) may have infinitely many solutions with the form $(u, v) = (w, \eta w)$ for $\eta \neq 0$, where w is the solution of $h^2 \triangle w - (V + \lambda)w + (1 + \eta^2)w^3 = 0$ in $\mathbb{R}^2$. This may provide a reason to ignore the case $\beta = 1$ in Theorem 1.2.

For the existence of other bound states to the system (1.11), the reader may refer to [2], [4], [13], [20], [21], [23], [31], and the references therein. Our result here seems to be the first in studying the orbital stability of (1.11) with a trapping potential.

The argument of Theorem 1.2 is applicable to studying another two-component system of NLS equations having symbiotic bright solitons (cf. [22] and [27]) given by

$$\begin{cases} -ih\dfrac{\partial \Phi}{\partial t} = h^2 \triangle \Phi - V\Phi - |\Phi|^2\Phi + \beta|\Psi|^2\Phi\,, \\ -ih\dfrac{\partial \Psi}{\partial t} = h^2 \triangle \Psi - V\Psi - |\Psi|^2\Psi + \beta|\Phi|^2\Psi \end{cases} \tag{1.15}$$

for $x \in \mathbb{R}^2$ and $t > 0$, where $V = V(x)$ is a smooth nonnegative function, $\beta \in \mathbb{R}$ is a nonzero constant, and $0 < h \ll 1$ is a small parameter. It is remarkable that the coefficients of the terms $|\Phi|^2\Phi$ and $|\Psi|^2\Psi$ of the system (1.15) have opposite sign to those of the system (1.10). As for the system (1.11), bound states of (1.15) are of the form $\Phi(x,t) = e^{i\lambda t/h}u(x)$ and $\Psi(x,t) = e^{i\lambda t/h}v(x)$, where (u, v) satisfies the following nonlinear elliptic system:

$$\begin{cases} h^2 \triangle u - (V + \lambda)u - u^3 + \beta u v^2 = 0, & x \in \mathbb{R}^2, \\ h^2 \triangle v - (V + \lambda)v - v^3 + \beta u^2 v = 0, & x \in \mathbb{R}^2, \\ u(x), v(x) > 0, & u, v \in H^1(\mathbb{R}^2). \end{cases} \tag{1.16}$$

It is easy to check that the system (1.16) has a solution $(\frac{1}{\sqrt{\beta-1}}u_h, \frac{1}{\sqrt{\beta-1}}u_h)$ for $\beta > 1$. As for Theorem 1.2, we may have the following corollary.

COROLLARY 1.3. *For $0 < h < 1$, let u_h be a single-spike solution concentrated at a local minimum point of the function V. Suppose the parameter h is sufficiently small. Then $(\frac{1}{\sqrt{\beta-1}}u_h, \frac{1}{\sqrt{\beta-1}}u_h)$ is an orbitally stable solution to* (1.15) *if $\beta > 1$.*

The proof of Corollary 1.3 is quite similar to that of Theorem 1.2 so we may neglect the detailed proof here.

The rest of this paper is organized as follows. In section 2, we figure out the properties of u_h and the spectrum of the linearized operator L_h as the parameter h goes to zero. Then we state the proof of Theorem 1.1 in section 3. Finally, we provide the proof of Theorem 1.2 in section 4.

2. Properties of u_h. In this section, we study the properties of u_h, which is a single-spike solution concentrated at a nondegenerate critical point x_0 of $V(x)$. Let x_h be the unique local maximum point of u_h. So $x_h \to x_0$. Let us recall the following results of Grossi [16].

LEMMA 2.1.

(1) $x_h = x_0 + \mathrm{o}(h)$;

(2) u_h *is unique and nondegenerate, i.e.,* L_h *has no zero eigenvalue.*

Proof. (1) follows from Lemma 5.4 of [16], while (2) follows from Theorem 1.1 of [16]. □

We need the following two lemmas. The first one is an asymptotic behavior of u_h.

LEMMA 2.2.

$$u_h(x_h + hy) = (V(x_h) + \lambda)^{\frac{1}{p-1}}\, w(\sqrt{V(x_h) + \lambda}\, y) + h^2\phi_0 + \mathrm{o}(h^2), \tag{2.1}$$

where w *is the unique positive solution of*

$$\triangle w - w + w^p = 0, w(0) = \max_{y\in\mathbb{R}^N} w(y), w > 0 \quad in \quad \mathbb{R}^N, w \to 0 \ as \ |y| \to +\infty, \tag{2.2}$$

ϕ_0 *satisfies*

$$\triangle\phi_0 - (V(x_h) + \lambda)\phi_0 + pw_{x_h}^{p-1}\phi_0 - \frac{1}{2}\sum_{i,j} V_{ij}(x_0) y_i y_j w_{x_h} = 0 \tag{2.3}$$

with

$$w_{x_h}(y) := (V(x_h) + \lambda)^{\frac{1}{p-1}}\, w(\sqrt{V(x_h) + \lambda}\, y). \tag{2.4}$$

Proof. Note that for fixed s, $w_s(y)$ satisfies

$$\triangle w_s - (V(s) + \lambda)w_s + w_s^p = 0. \tag{2.5}$$

Let $\phi_h(y) = u_h(x_h + hy) - w_{x_h}(y)$. Then $|\phi_h| \to 0$ uniformly and ϕ_h satisfies

$$\triangle\phi_h - (V(x_h + hy) + \lambda)\phi_h + pw_{x_h}^{p-1}\phi_h + N(\phi_h) - (V(x_h + hy) - V(x_h))w_{x_h} = 0, \tag{2.6}$$

where $N(\phi_h) = (w_{x_h} + \phi_h)^p - w_{x_h}^p - pw_{x_h}^{p-1}\phi_h$. Note that $\nabla\phi_h(0) = 0$ and

$$\begin{aligned}(V(x_h + hy) - V(x_h)) &= (\nabla V(x_h))hy + \frac{1}{2}\sum_{i,j} V_{ij}(x_h)h^2 y_i y_j + \mathrm{O}(h^3|y|^3)\\ &= \mathrm{o}(h^2)|y| + \frac{1}{2}\sum_{i,j} V_{ij}(x_0)h^2 y_i y_j + \mathrm{o}(h^2|y|^2).\end{aligned} \tag{2.7}$$

Here we have used Lemma 2.1(1).

Now we claim that $|\phi_h| \le ch^2$. In fact, suppose not. We may assume that $|\phi_h|_{L^\infty} h^{-2} \to \infty$. Let $\widetilde{\phi}_h = \frac{\phi_h}{|\phi_h|_{L^\infty}}$. Then $\widetilde{\phi}_h$ satisfies

$$\triangle\widetilde{\phi}_h - (V + \lambda)\widetilde{\phi}_h + pw_{x_h}^{p-1}\widetilde{\phi}_h + \frac{N(\phi_h)}{|\phi_h|_{L^\infty}} - \frac{(V(x_h + hy) - V(x_h))w_{x_h}}{|\phi_h|_{L^\infty}} = 0. \tag{2.8}$$

Note that by (2.7),

$$\frac{|V(x_h+hy)-V(x_h)||w_{x_h}|}{|\phi_h|_{L^\infty}} \le \frac{h^2|y|^2|w_{x_h}|}{|\phi_h|_{L^\infty}} \le o(1)|y|^2|w_{x_h}| \tag{2.9}$$

for $|y| \ge 1$. Let y_h be the global maximum point of $\widetilde{\phi}_h$, i.e., $\widetilde{\phi}_h(y_h) = \max_y \frac{\phi_h(y)}{|\phi_h|_{L^\infty}} = 1$. Then by (2.8) and (2.9) and the maximum principle, we have $|y_h| \le C$. Here we have used the fact that $V \ge 0$ and $\lambda > 0$.

By the usual elliptic regularity theory, we may take a subsequence $\widetilde{\phi}_h \to \overline{\phi}_0$, where $\overline{\phi}_0$ satisfies

$$\triangle\overline{\phi}_0 - (V(x_0)+\lambda)\overline{\phi}_0 + pw_{x_0}^{p-1}\overline{\phi}_0 = 0,\ \nabla\overline{\phi}_0(0) = 0. \tag{2.10}$$

Since $\nabla\overline{\phi}_0(0) = 0$, we see that $\overline{\phi}_0 = \sum_{j=1}^N c_j \frac{\partial w_{x_0}}{\partial y_j}$, and hence $c_j = 0$. Consequently, $\overline{\phi}_0 \equiv 0$. This may contradict the fact that $1 = \widetilde{\phi}_h(y_h) \to \overline{\phi}_0(y_0)$ for some y_0. Therefore $|\phi_h| \le ch^2$. Now we let $\phi_h = h^2\phi_0 + h^2\overline{\phi}_h$. Then, as for the previous argument, we may have $\overline{\phi}_h = o(1)$ and complete the proof of Lemma 2.2. □

As in Proposition 3.6 of [19], one may get two lemmas, as follows.

LEMMA 2.3. *For each $s \in \mathbb{R}^N$, the map*

$$L_s\phi := \triangle\phi - (V(s)+\lambda)\phi + pw_s^{p-1}\phi \tag{2.11}$$

is invertible from $K_s^\perp$ to $C_s^\perp$, where

$$K_s^\perp = \left\{\phi \in H^2(\mathbb{R}^N) \,\middle|\, \int_{\mathbb{R}^N} \phi\frac{\partial w_s}{\partial y_j}(y)\,dy = 0\,, j = 1,\dots,N\right\} \subset H^2(\mathbb{R}^N),$$

$$C_s^\perp = \left\{\phi \in L^2(\mathbb{R}^N) \,\middle|\, \int_{\mathbb{R}^N} \phi\frac{\partial w_s}{\partial y_j}(y)\,dy = 0\,, j = 1,\dots,N\right\} \subset L^2(\mathbb{R}^N).$$

LEMMA 2.4. *The map*

$$L_0\phi := \triangle\phi - (V(x_0)+\lambda)\phi + pw_{x_0}^{p-1}\phi \tag{2.12}$$

admits the following eigenvalues:

$$\lambda_1 > 0\,,\ \lambda_2 = \cdots = \lambda_{N+1} = 0\,,\ \lambda_{N+2} < 0,$$

where the kernel of L_0 is spanned by $\frac{\partial w_{x_0}}{\partial y_j}$, $j = 1,\dots,N$.

Our main result in this section is the following.

THEOREM 2.5. *The eigenvalue problem*

$$L_h\psi_h = \lambda_h\psi_h \tag{2.13}$$

admits eigenvalues

$$\lambda_{h,1} > \lambda_{h,2} > \cdots > \lambda_{h,N+1} > \lambda_{h,N+2}, \tag{2.14}$$

satisfying, as $h \to 0$, $\lambda_{h,1} \to \lambda_1 > 0$, $\lambda_{h,N+2} \to \lambda_{N+2} < 0$, and

$$\frac{\lambda_{h,j}}{h^2} \to c_0\nu_{j-1},\ j = 2,\dots,N+1, \tag{2.15}$$

where c_0 is a negative constant and ν_j's are eigenvalues of the Hessian matrix $(\nabla^2 V(x_0))$.

Proof. We follow the proofs given in section 5 of [35]. Assume that $\|\psi_h\|_{L^2} = 1$. It is easy to see that for eigenvalues $\lambda_h \in [\frac{1}{2}\lambda_{N+2}, \frac{1}{2}\lambda_1]$, as $h \downarrow 0$, $\lambda_h \to \lambda_j$ for some j, where λ_j's are given in Lemma 2.4. Now we focus on the case $\lambda_{h,j} \to 0$, i.e., $\lambda_h \to 0$ as $h \downarrow 0$. Then the corresponding eigenfunctions can be written as

$$\psi_h(x_h + hy) = \sum_{j=1}^{N} c_j \frac{\partial w_{x_h}}{\partial y_j}(y) + \psi_h^{\perp}(y), \tag{2.16}$$

where $\int_{\mathbb{R}^N} \frac{\partial w_{x_h}}{\partial y_j}\psi_h^{\perp}(y)\,dy = 0$, $j = 1, 2, \ldots, N$. Hence by (2.13) and (2.16), $\psi_h^{\perp}$ may satisfy

(2.17)

$$\begin{aligned}&\triangle\psi_h^{\perp} - (V(x_h + hy) + \lambda)\psi_h^{\perp} + pw_{x_h}^{p-1}(y)\psi_h^{\perp} + p(u_h^{p-1}(x_h + hy) - w_{x_h}^{p-1}(y))\psi_h^{\perp}\\ &\quad + \sum_j c_j L_h \frac{\partial w_{x_h}}{\partial y_j} = \lambda_h \left(\sum_j c_j \frac{\partial w_{x_h}}{\partial y_j} + \psi_h^{\perp}\right).\end{aligned}$$

Using (2.1) and (2.7) of Lemma 2.2, we have

$$\begin{aligned}L_h \frac{\partial w_{x_h}}{\partial y_j} &= \triangle\left(\frac{\partial w_{x_h}}{\partial y_j}\right) - (V(x_h + hy) + \lambda)\frac{\partial w_{x_h}}{\partial y_j} + pu_h^{p-1}(x_h + hy)\frac{\partial w_{x_h}}{\partial y_j}\\ &= (V(x_h) - V(x_h + hy))\frac{\partial w_{x_h}}{\partial y_j} + p(u_h^{p-1}(x_h + hy) - w_{x_h}^{p-1}(y))\frac{\partial w_{x_h}}{\partial y_j}\\ &= O(h^2). \qquad (2.18)\end{aligned}$$

From Lemma 2.3, the map $L_{x_h} = \triangle - (V(x_h) + \lambda) + pw_{x_h}^{p-1}$ is invertible in the space $K_{x_h}^{\perp}$. Thus by (2.1), (2.7), and (2.18), (2.17) may give

$$||\psi_h^{\perp}||_{H^2} \le c(h^2 + |\lambda_h|)\sum_j |c_j|\,. \tag{2.19}$$

Now we set $z_j(y) = \frac{\partial w_{x_h}}{\partial y_j}(y)$ for $j = 1, \ldots, N$. Then multiplying (2.17) by z_k and integrating over $\mathbb{R}^N$, it is obvious that

(2.20)

$$\int_{\mathbb{R}^N} \left(L_h\psi_h^{\perp}\right) z_k\,dy + \sum_j c_j \int_{\mathbb{R}^N} \left(L_h \frac{\partial w_{x_h}}{\partial y_j}\right) z_k\,dy = \lambda_h \left(\sum_j c_j \int_{\mathbb{R}^N} z_j z_k\right) dy\,.$$

Here we have used the fact that $\psi_h^{\perp} \in K_{x_h}^{\perp}$. Using (2.18), (2.19), $\lambda_h = \mathrm{o}(1)$, and integration by parts, we obtain

$$\int_{\mathbb{R}^N} \left(L_h\psi_h^{\perp}\right) z_k = \int_{\mathbb{R}^N} \psi_h^{\perp} L_h z_k = \mathrm{o}(h^2) \tag{2.21}$$

and

$$\int_{\mathbb{R}^N} (L_h z_j)\, z_k = \int_{\mathbb{R}^N} (V(x_h) - V(x_h + hy)) z_j z_k + p \int_{\mathbb{R}^N} \left(u_h^{p-1} - w_{x_h}^{p-1} \right) z_j z_k$$

$$:= I_1 + I_2, \tag{2.22}$$

where

$$I_1 = \int_{\mathbb{R}^N} (V(x_h) - V(x_h + hy)) z_j z_k$$

$$= \mathrm{o}(h^2) - \frac{h^2}{2} \sum_{l,m} V_{lm}(x_h) \int_{\mathbb{R}^N} y_l y_m z_j z_k$$

$$= -\frac{h^2}{2} V_{jk}(x_h) \int_{\mathbb{R}^N} y_j z_j y_k z_k \,. \tag{2.23}$$

Here we have used (2.7) to get (2.23). For I_2, we use Lemma 2.2:

$$I_2 = p(p-1)h^2 \int_{\mathbb{R}^N} \phi_0 w_{x_h}^{p-2} z_j z_k + \mathrm{o}(h^2)$$

$$= -h^2 \int_{\mathbb{R}^N} L_0 \left(\frac{\partial^2 w}{\partial y_j \partial y_k} \right) \phi_0 + \mathrm{o}(h^2)$$

$$= -h^2 \int_{\mathbb{R}^N} (L_0 \phi_0) \frac{\partial^2 w}{\partial y_j \partial y_k} + \mathrm{o}(h^2)$$

$$= -\frac{h^2}{2} \sum_{l,m} V_{lm}(x_h) \int_{\mathbb{R}^N} y_l y_m w_{x_h} \frac{\partial^2 w}{\partial y_j \partial y_k} + \mathrm{o}(h^2)$$

$$= \frac{h^2}{2} \sum_{l,m} V_{lm}(x_h) \int_{\mathbb{R}^N} \frac{\partial}{\partial y_j} (y_l y_m w_{x_h}) z_k + \mathrm{o}(h^2)$$

$$= \frac{h^2}{2} V_{jk}(x_h) \int_{\mathbb{R}^N} y_j y_k z_j z_k - \frac{h^2}{2} V_{jk}(x_h) \int_{\mathbb{R}^N} w_{x_h}^2 + \mathrm{o}(h^2)\,. \tag{2.24}$$

Here we have used the following identity:

$$\int_{\mathbb{R}^N} \frac{\partial}{\partial y_j} (y_l y_m w_{x_h}) z_k = \int_{\mathbb{R}^N} \delta_{jl} y_m w_{x_h} z_k + \int_{\mathbb{R}^N} \delta_{jm} y_l w_{x_h} z_k + \int_{\mathbb{R}^N} y_l y_m z_j z_k$$

$$= \frac{1}{2} \int_{\mathbb{R}^N} \delta_{jl} y_m \frac{\partial}{\partial y_k} (w_{x_h}^2)$$

$$+ \frac{1}{2} \int_{\mathbb{R}^N} \delta_{jm} y_l \frac{\partial}{\partial y_k} (w_{x_h}^2) + \int_{\mathbb{R}^N} y_l y_m z_j z_k$$

$$= -(\delta_{jl}\delta_{km} + \delta_{jm}\delta_{lk}) \frac{1}{2} \int_{\mathbb{R}^N} w_{x_h}^2 + \int_{\mathbb{R}^N} y_l y_m z_j z_k.$$

Combining (2.23) and (2.24), we have

$$I_1 + I_2 = -\frac{h^2}{2} V_{jk}(x_h) \int_{\mathbb{R}^N} w_{x_0}^2 + \mathrm{o}(h^2). \tag{2.25}$$

Substituting (2.21) and (2.25) into (2.20), we may obtain $\lambda_j/h^2 \to c_0\nu_j$ for $j = 1, \ldots, N$, where $c_0 = -\frac{\int_{\mathbb{R}^N} w_{x_0}^2\, dy}{\int_{\mathbb{R}^N} z_j^2\, dy}$ is a negative constant. The rest of the proof follows from a perturbation result, similar to pages 1473–1474 of [35]. We may omit the details here. □

From Theorem 2.5, we may deduce the following.

THEOREM 2.6. *u_h is smooth in λ. Moreover, let $R_h = \frac{\partial u_h}{\partial \lambda}$. Then R_h satisfies*

$$L_h R_h - u_h = 0 \tag{2.26}$$

and

$$R_h = \sum_{j=1}^{N} c_j^h z_j + R_0 + \mathrm{o}(1)\,, \tag{2.27}$$

where $R_0 = \frac{\partial}{\partial \lambda} w_{x_h} = (V(x_h) + \lambda)^{-1}(\frac{1}{p-1} w_{x_h} + \frac{1}{2} y \cdot \nabla w_{x_h})$ and $|c_j^h| = \mathrm{O}(1)$ for $j = 1, \ldots, N$.

Proof. Since u_h is unique and L_h is invertible, it is easy to see that u_h is smooth in λ and R_h satisfies (2.26). Now we decompose R_h as

$$R_h = \sum_{j=1}^{N} c_j^h z_j + R_0 + \overline{R}_h,$$

where $\int_{\mathbb{R}^N} z_j \overline{R}_h = 0,\ j = 1, \ldots, N$. Then $\overline{R}_h$ satisfies

$$L_h \overline{R}_h - u_h + L_h R_0 + \sum_{j=1}^{N} c_j^h L_h z_j = 0. \tag{2.28}$$

As for the proof of Theorem 2.5, we have

$$\|\overline{R}_h\|_{H^2} \le c \left(|c_j^h| h^2 + \|L_h R_0 - u_h\|_{L^2} \right). \tag{2.29}$$

From (2.4) and (2.5), it is easy to check that $R_0 = \frac{\partial}{\partial \lambda} w_{x_h} = (V(x_h) + \lambda)^{-1}(\frac{1}{p-1} w_{x_h} + \frac{1}{2} y \cdot \nabla w_{x_h})$ and $L_{x_h} R_0 = w_{x_h}$ by differentiating (2.5) with respect to λ. Hence

$$L_h R_0 - u_h = p(u_h^{p-1}(x_h + hy) - w_{x_h}^{p-1}(y)) R_0 - (V(x_h + hy) - V(x_h))\, R_0 + w_{x_h} - u_h\,.$$

Consequently, by Lemma 2.2 and (2.7), we obtain

$$\|L_h R_0 - u_h\|_{L^2} = \mathrm{O}(h^2)\,, \tag{2.30}$$

and then by (2.29),

$$\|\overline{R}_h\|_{H^2} = (1 + |c_j^h|)\mathrm{O}(h^2)\,. \tag{2.31}$$

To estimate c_j^h's, we may multiply (2.28) by z_k and integrate over $\mathbb{R}^N$. Then

$$\sum_{j=1}^{N} c_j^h \int_{\mathbb{R}^N} (L_h z_j) z_k + \int_{\mathbb{R}^N} (L_h R_0 - u_h) z_k + \int_{\mathbb{R}^N} (L_h \overline{R}_h) z_k = 0\,. \tag{2.32}$$

Hence by (2.22) and (2.25), (2.32) may imply

$$|c_j^h| \le \frac{C}{h^2} \left(\left| \int_{\mathbb{R}^N} (L_h R_0 - u_h) z_k \right| + \left| \int_{\mathbb{R}^N} (L_h \overline{R}_h) z_k \right| \right). \tag{2.33}$$

Using integration by parts and (2.18), we have

$$\int_{\mathbb{R}^N} (L_h \overline{R}_h) z_k = \int_{\mathbb{R}^N} (L_h z_k) \overline{R}_h = \|\overline{R}_h\|_{L^2}\, \mathrm{O}(h^2)\,. \tag{2.34}$$

Therefore by (2.30), (2.31), (2.33), and (2.34), we may obtain $|c_j^h| = \mathrm{O}(1)$ and complete the proof. □

3. Proof of Theorem 1.1. Let $p = 1 + \frac{4}{N}$. By Theorem 2.5, L_h has $m+1$ positive eigenvalues and no zero eigenvalue, where m is the number of negative eigenvalues of the matrix $(\nabla^2 V(x_0))$. Let us now compute $d''(\lambda)$.

From (1.7), it is easy to get

$$d'(\lambda) = \frac{1}{2} \int_{\mathbb{R}^N} u_h^2$$

and hence

$$d''(\lambda) = \int_{\mathbb{R}^N} \frac{\partial u_h}{\partial \lambda} u_h = \int_{\mathbb{R}^N} R_h u_h.$$

By direct computations,

$$L_h \left(\frac{1}{p-1} u_h + \frac{1}{2} hy \cdot \nabla u_h \right) = \frac{1}{2} hy \cdot \nabla V(x_h + hy) u_h + (V(x_h + hy) + \lambda) u_h. \tag{3.1}$$

Consequently,

$$\begin{aligned}
(V(x_h) + \lambda) \int_{\mathbb{R}^N} R_h u_h &= \int_{\mathbb{R}^N} R_h (V(x_h) - V(x_h + hy)) u_h \\
&\quad + \int_{\mathbb{R}^N} R_h (V(x_h + hy) + \lambda) u_h \\
&= \int_{\mathbb{R}^N} R_h \left(V(x_h) - V(x_h + hy) - \frac{1}{2} hy \cdot \nabla V(x_h + hy) \right) u_h \\
&\quad + \int_{\mathbb{R}^N} R_h L_h \left(\frac{1}{p-1} u_h + \frac{1}{2} hy \cdot \nabla u_h \right). \qquad (3.2)
\end{aligned}$$

Using integration by parts and (2.26), we may obtain

$$\begin{aligned}
\int_{\mathbb{R}^N} R_h L_h \left(\frac{1}{p-1} u_h + \frac{1}{2} hy \cdot \nabla u_h \right) &= \int_{\mathbb{R}^N} (L_h R_h) \left(\frac{1}{p-1} u_h + \frac{1}{2} hy \cdot \nabla u_h \right) \\
&= \int_{\mathbb{R}^N} u_h \left(\frac{1}{p-1} u_h + \frac{1}{2} hy \cdot \nabla u_h \right) \\
&= \left(\frac{1}{p-1} - \frac{N}{4} \right) \int_{\mathbb{R}^N} u_h^2 = 0\,, \qquad (3.3)
\end{aligned}$$

since $p = 1 + \frac{4}{N}$. So by (2.7), (3.2), (3.3), Lemma 2.2, and Theorem 2.6, we have

$$
\begin{aligned}
d''(\lambda) &= \frac{1}{V(x_h)+\lambda}\int_{\mathbb{R}^N} R_h\left(V(x_h) - V(x_h+hy) - \frac{1}{2}hy\cdot\nabla V(x_h+hy)\right)u_h \\
&= \frac{h^2}{V(x_h)+\lambda}\int_{\mathbb{R}^N} R_h\left[\sum_{i,j} -V_{ij}(x_h)y_iy_j\right]u_h + \mathrm{o}(h^2) \\
&= \frac{h^2}{V(x_h)+\lambda}\int_{\mathbb{R}^N} \left(\sum_l c_l^h z_l + R_0 + \mathrm{o}(1)\right)\left(\sum_{i,j} -V_{ij}(x_h)y_iy_j\right) \\
&\quad (w_{x_h} + \mathrm{O}(h^2)) + \mathrm{o}(h^2) \\
&= -\frac{h^2}{V(x_h)+\lambda}\sum_i V_{ii}(x_h)\int_{\mathbb{R}^N} R_0 y_i^2 w_{x_h} + \mathrm{o}(h^2) \\
&= -\frac{h^2}{(V(x_h)+\lambda)^2}\sum_i V_{ii}(x_h)\int_{\mathbb{R}^N}\left(\frac{1}{p-1}w_{x_h} + \frac{1}{2}y\cdot\nabla w_{x_h}\right)y_i^2 w_{x_h} + \mathrm{o}(h^2) \\
&= -\frac{h^2}{(V(x_h)+\lambda)^2}\left(\frac{1}{p-1} - \frac{N+2}{4}\right)\sum_i V_{ii}(x_h)\int_{\mathbb{R}^N} y_i^2 w_{x_h}^2 + \mathrm{o}(h^2) \\
&= \frac{h^2}{2(V(x_0)+\lambda)^2}\Delta V(x_0)\int_{\mathbb{R}^N} y_i^2 w_{x_0}^2 + \mathrm{o}(h^2) \quad \left(\text{because } p = 1+\frac{4}{N}\right). \qquad (3.4)
\end{aligned}
$$

If x_0 is a local minimum point, then $m = 0$ and $n(L_h) = 1$. Since the Hessian matrix $(\nabla^2 V(x_0))$ is positive definite, then

$$
d''(\lambda) > 0, \quad p(d''(\lambda)) = 1, \qquad (3.5)
$$

which implies that u_h is orbitally unstable by the orbital stability criteria of [14] and [15].

If x_0 is not a local minimum, then $m \geq 1$ and $n(L_h) \geq 2$. In this case, by the formula (3.4), $p(d''(\lambda)) = \frac{1}{2}(1 + \frac{\Delta V(x_0)}{|\Delta V(x_0)|})$. By the instability criteria of [15], we conclude that u_h is orbitally unstable if $m - \frac{1}{2}(1 + \frac{\Delta V(x_0)}{|\Delta V(x_0)|})$ is even. This completes the proof of Theorem 1.1.

4. Proof of Theorem 1.2. Let $(\frac{1}{\sqrt{1+\beta}}u_h, \frac{1}{\sqrt{1+\beta}}u_h)$ be a solution of (1.11). The linearized operator of (1.10) around $(\frac{1}{\sqrt{1+\beta}}u_h, \frac{1}{\sqrt{1+\beta}}u_h)$ is

$$
\mathbb{L}_h\begin{pmatrix}\phi\\ \psi\end{pmatrix} = \begin{pmatrix} h^2\triangle\phi - (V(x)+\lambda)\phi + \frac{3+\beta}{1+\beta}u_h^2\phi + \frac{2\beta}{1+\beta}u_h^2\psi \\ h^2\triangle\psi - (V(x)+\lambda)\psi + \frac{3+\beta}{1+\beta}u_h^2\psi + \frac{2\beta}{1+\beta}u_h^2\phi \end{pmatrix}. \qquad (4.1)
$$

We first define a sequence of numbers $\beta_j \in (-1, 0)$. By Lemma 4.2 of [34], the eigenvalue problem

$$
\triangle\psi - (V(x_0)+\lambda)\psi + \mu w_{x_0}^2\psi = 0 \qquad (4.2)
$$

admits eigenvalues

$$\mu_1 = 1,\ \mu_2 = \cdots = \mu_{N+1} = 3,\ \mu_{N+2} > 3. \tag{4.3}$$

We then define β_j by

$$\beta_j = \frac{3-\mu_j}{1+\mu_j}, \quad j = 1, 2, \ldots . \tag{4.4}$$

The following lemma shows the nondegeneracy of $\mathbb{L}$.

LEMMA 4.1. *$\mathbb{L}_h$ has no zero eigenvalue if $\beta \neq \beta_j, j = 1, 2, \ldots$.*

Proof. Let $\mathbb{L}_h \binom{\phi}{\psi} = \binom{0}{0}$. Then by an orthonormal transformation it is equivalent to

$$L_{h,1}\widetilde{\phi} = 0, \tag{4.5}$$

$$L_{h,2}\widetilde{\psi} = 0, \tag{4.6}$$

where $L_{h,1} = L_h$, $L_{h,2} = h^2\triangle - (V(x_h + hy) + \lambda) + \frac{3-\beta}{1+\beta}u_h^2$. By Theorem 2.5, we may conclude that $\widetilde{\phi} = 0$. It remains to consider (4.6). As $h \to 0$, (4.6) may tend to the limiting equation given by

$$\triangle\psi - (V(x_0) + \lambda)\psi + \frac{3-\beta}{1+\beta}w_{x_0}^2\psi = 0. \tag{4.7}$$

Since $\beta \neq \beta_j$, i.e., $\frac{3-\beta}{1+\beta} \neq \mu_j$, then by Lemma 4.2 of [34], (4.7) has only a trivial solution. Therefore, $\widetilde{\psi} = 0$ and we may complete the proof. ∎

The next lemma computes the Morse index of $(\frac{1}{\sqrt{1+\beta}}u_h, \frac{1}{\sqrt{1+\beta}}u_h)$. Here the Morse index is defined to be the number of positive eigenvalues of $\mathbb{L}_h$, which is just $n(\mathbb{L}_h)$.

LEMMA 4.2. *If $-1 < \beta < 0$ and $\beta \notin \{\beta_2, \ldots, \beta_j, \ldots\}$, then the Morse index of $(\frac{1}{\sqrt{1+\beta}}u_h, \frac{1}{\sqrt{1+\beta}}u_h)$ is at least $N+2$. If $0 < \beta < 1$, then the Morse index of $(\frac{1}{\sqrt{1+\beta}}u_h, \frac{1}{\sqrt{1+\beta}}u_h)$ is two. If $\beta > 1$, then the Morse index of $(\frac{1}{\sqrt{1+\beta}}u_h, \frac{1}{\sqrt{1+\beta}}u_h)$ is one.*

Proof. The eigenvalue problem $\mathbb{L}_h \binom{\phi}{\psi} = \bar{\lambda}\binom{\phi}{\psi}$ can be decomposed to

$$L_{h,1}\widetilde{\phi} = \bar{\lambda}\widetilde{\phi}, \tag{4.8}$$

$$L_{h,2}\widetilde{\psi} = \bar{\lambda}\widetilde{\psi}. \tag{4.9}$$

By Theorem 2.5, $L_{h,1}$ has only one positive eigenvalue. It remains to consider the spectrum of $L_{h,2}$. If $\beta < 0$, then the eigenvalue problem

$$\triangle\psi - (V(x_0) + \lambda)\psi + \frac{3-\beta}{1+\beta}w_{x_0}^2\psi = \bar{\lambda}\psi \tag{4.10}$$

has at least $N+1$ positive eigenvalues. We may define a space of functions by

$$\mathbf{V} = \operatorname{span}\left\{ w_{x_0}, \frac{\partial w_{x_0}}{\partial y_j},\ j = 1, \ldots, N, \right\}.$$

Since $-1 < \beta < 0$, we have $\frac{3-\beta}{1+\beta} > 3$. Hence

$$\int_{\mathbb{R}^N} \left[|\nabla\phi|^2 + (V(x_0) + \lambda)\phi^2 - \frac{3-\beta}{1+\beta}w_{x_0}^2\phi^2 \right] < 0 \quad \forall \phi \in \mathbf{V}. \tag{4.11}$$

Thus by the variational characterization of the eigenvalues of (4.10), we see that $\lambda_{N+1} > 0$. Moreover, by the perturbation argument, (4.9) has at least $N+1$ positive eigenvalues.

So when $\beta < 0$, $\mathbb{L}_h$ has at least $N+2$ positive eigenvalues.

When $0 < \beta < 1$, $1 < \frac{3-\beta}{1+\beta} < 3$, (4.10) has only one positive eigenvalue. So the Morse index is two.

When $\beta > 1$, (4.10) has no positive eigenvalue. So the Morse index is one. □

Since $\mathbb{L}_h$ is invertible, $(\frac{1}{\sqrt{1+\beta}}u_h, \frac{1}{\sqrt{1+\beta}}u_h)$ is nondegenerate. Thus the system

$$\begin{cases} h^2\triangle u - (V(x)+\lambda+\lambda_1)u + u^3 + \beta u v^2 = 0 & \text{in } \mathbb{R}^N, \\ h^2\triangle v - (V(x)+\lambda+\lambda_2)v + v^3 + \beta u^2 v = 0 & \text{in } \mathbb{R}^N \end{cases} \tag{4.12}$$

has a solution $(u_{h,\lambda_1,\lambda_2}, v_{h,\lambda_1,\lambda_2})$ satisfying

$$(u_{h,\lambda_1,\lambda_2}, v_{h,\lambda_1,\lambda_2}) = \left(\frac{1}{\sqrt{1+\beta}}u_h, \frac{1}{\sqrt{1+\beta}}u_h\right) + \mathrm{O}\left((|\lambda_1|+|\lambda_2|)h^{-2}\right) \tag{4.13}$$

as $|\lambda_1| + |\lambda_2| \ll 1$.

Let us define

$$\begin{aligned} d(\lambda_1,\lambda_2) = &\int_{\mathbb{R}^N} \frac{h^2}{2}|\nabla u_{h,\lambda_1,\lambda_2}|^2 + \frac{V(x)+\lambda+\lambda_1}{2}u_{h,\lambda_1,\lambda_2}^2 - \frac{1}{4}\int_{\mathbb{R}^N} u_{h,\lambda_1,\lambda_2}^4 \\ &+ \int_{\mathbb{R}^N} \frac{h^2}{2}|\nabla v_{h,\lambda_1,\lambda_2}|^2 + \frac{V(x)+\lambda+\lambda_2}{2}v_{h,\lambda_1,\lambda_2}^2 - \frac{1}{4}\int_{\mathbb{R}^N} v_{h,\lambda_1,\lambda_2}^4 \\ &- \frac{\beta}{2}\int_{\mathbb{R}^N} u_{h,\lambda_1,\lambda_2}^2 v_{h,\lambda_1,\lambda_2}^2. \end{aligned} \tag{4.14}$$

It is easy to see that

$$\frac{\partial d}{\partial \lambda_1} = \frac{1}{2}\int_{\mathbb{R}^N} u_{h,\lambda_1,\lambda_2}^2, \quad \frac{\partial^2 d}{\partial \lambda_1^2} = \int_{\mathbb{R}^N} u_{h,\lambda_1,\lambda_2}\frac{\partial u_{h,\lambda_1,\lambda_2}}{\partial \lambda_1},$$

$$\frac{\partial d}{\partial \lambda_2} = \frac{1}{2}\int_{\mathbb{R}^N} v_{h,\lambda_1,\lambda_2}^2, \quad \frac{\partial^2 d}{\partial \lambda_2^2} = \int_{\mathbb{R}^N} v_{h,\lambda_1,\lambda_2}\frac{\partial v_{h,\lambda_1,\lambda_2}}{\partial \lambda_2},$$

$$\frac{\partial^2 d}{\partial \lambda_1 \partial \lambda_2} = \int_{\mathbb{R}^N} u_{h,\lambda_1,\lambda_2}\frac{\partial u_{h,\lambda_1,\lambda_2}}{\partial \lambda_2}.$$

Now we may define functions as

$$\Phi_1 = \left.\frac{\partial u_{h,\lambda_1,\lambda_2}}{\partial \lambda_1}\right|_{(\lambda_1,\lambda_2)=(0,0)} \quad \text{and} \quad \Psi_1 = \left.\frac{\partial u_{h,\lambda_1,\lambda_2}}{\partial \lambda_2}\right|_{(\lambda_1,\lambda_2)=(0,0)}.$$

Then by (4.12), $\binom{\Phi_1}{\Psi_1}$ satisfies

$$\mathbb{L}_h \begin{pmatrix} \Phi_1 \\ \Psi_1 \end{pmatrix} = \begin{pmatrix} u_{h,0,0} \\ 0 \end{pmatrix}. \tag{4.15}$$

Similarly, if we set $\Phi_2 = \frac{\partial v_{h,\lambda_1,\lambda_2}}{\partial \lambda_1}|_{(\lambda_1,\lambda_2)=(0,0)}$ and $\Psi_2 = \frac{\partial v_{h,\lambda_1,\lambda_2}}{\partial \lambda_2}|_{(\lambda_1,\lambda_2)=(0,0)}$, then by (4.12), we have

$$\Psi_2 = \Phi_1, \quad \Phi_2 = \Psi_1. \tag{4.16}$$

Let $\mathbb{B} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$. Then (4.15) is equivalent to

$$\mathbb{B}\mathbb{L}_h \begin{pmatrix} \Phi_1 \\ \Psi_1 \end{pmatrix} = \begin{pmatrix} L_{h,1}(\Phi_1 + \Psi_1) \\ L_{h,2}(\Phi_1 - \Psi_1) \end{pmatrix} = \begin{pmatrix} u_{h,0,0} \\ u_{h,0,0} \end{pmatrix}. \tag{4.17}$$

So

$$\Phi_1 + \Psi_1 = R_{h,1} \quad \text{and} \quad \Phi_1 - \Psi_1 = R_{h,2}, \tag{4.18}$$

where

$$R_{h,1} = \frac{1}{\sqrt{1+\beta}} R_h \quad \text{and} \quad R_{h,2} = L_{h,2}^{-1}\left(\frac{1}{\sqrt{1+\beta}} u_h\right). \tag{4.19}$$

Now we compute the Hessian matrix

$$(\nabla^2 d)\big|_{(\lambda_1,\lambda_2)=(0,0)} = \begin{pmatrix} \frac{1}{\sqrt{1+\beta}} \int_{\mathbb{R}^N} u_h \Phi_1 & \frac{1}{\sqrt{1+\beta}} \int_{\mathbb{R}^N} u_h \Psi_1 \\ \frac{1}{\sqrt{1+\beta}} \int_{\mathbb{R}^N} u_h \Psi_1 & \frac{1}{\sqrt{1+\beta}} \int_{\mathbb{R}^N} u_h \Phi_1 \end{pmatrix},$$

$$\begin{aligned} \mathbb{B}\,(\nabla^2 d)\big|_{(\lambda_1,\lambda_2)=(0,0)}\,\mathbb{B}^T &= \begin{pmatrix} \frac{1}{\sqrt{1+\beta}} \int_{\mathbb{R}^N} u_h R_{h_1} & \frac{1}{\sqrt{1+\beta}} \int_{\mathbb{R}^N} u_h R_{h_1} \\ \frac{1}{\sqrt{1+\beta}} \int_{\mathbb{R}^N} u_h R_{h_2} & -\frac{1}{\sqrt{1+\beta}} \int_{\mathbb{R}^N} u_h R_{h_2} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \\ &= \begin{pmatrix} \frac{2}{\sqrt{1+\beta}} \int_{\mathbb{R}^N} u_h R_{h_1} & 0 \\ 0 & \frac{2}{\sqrt{1+\beta}} \int_{\mathbb{R}^N} u_h R_{h_2} \end{pmatrix}. \end{aligned}$$

By the results in section 3, $\frac{2}{\sqrt{1+\beta}} \int_{\mathbb{R}^N} u_h R_{h_1} = \frac{2}{1+\beta} \int_{\mathbb{R}^N} u_h R_h > 0$. It is enough to compute $\sqrt{1+\beta} \int_{\mathbb{R}^N} u_h R_{h_2} = \int_{\mathbb{R}^N} u_h L_{h,2}^{-1}(u_h)$. Note that as $h \to 0^+$,

$$\int_{\mathbb{R}^N} u_h L_{h,2}^{-1}(u_h) \to \int_{\mathbb{R}^N} w_{x_0} L_\mu^{-1}(w_{x_0}),$$

where

$$L_\mu \phi = \triangle \phi - (V(x_0) + \lambda)\phi + \mu w_{x_0}^2 \phi \tag{4.20}$$

with $\mu = \frac{3-\beta}{1+\beta}$.

Let $\rho(\mu) = \int_{\mathbb{R}^N} w_{x_0} L_\mu^{-1}(w_{x_0})$ and ϕ_μ the unique solution of $\triangle \phi_\mu - (V(x_0)+\lambda)\phi_\mu + \mu w_{x_0}^2 \phi_\mu = w_{x_0}$, i.e., $L_\mu \phi_\mu = w_{x_0}$ for $\mu \neq \mu_j, j = 1, 2, \ldots$. Then $\frac{\partial \phi_\mu}{\partial \mu}$ satisfies

$$L_\mu\left(\frac{\partial \phi_\mu}{\partial \mu}\right) = -w_{x_0}^2 \phi_\mu, \quad \text{i.e.,} \quad \frac{\partial \phi_\mu}{\partial \mu} = -L_\mu^{-1}(w_{x_0}^2 \phi_\mu).$$

Hence

$$\begin{aligned} \rho'(\mu) &= \int_{\mathbb{R}^N} w_{x_0} \frac{\partial \phi_\mu}{\partial \mu} = -\int_{\mathbb{R}^N} w_{x_0} L_\mu^{-1}(w_{x_0}^2 \phi_\mu) \\ &= -\int_{\mathbb{R}^N} (L_\mu^{-1} w_{x_0})(w_{x_0}^2 \phi_\mu) \\ &= -\int_{\mathbb{R}^N} w_{x_0}^2 \phi_\mu^2 < 0 \quad \text{for} \quad \mu \neq \mu_j\,, j = 1, 2, \ldots, \end{aligned}$$

i.e.,

$$\rho'(\mu) < 0 \quad \text{for} \quad \mu \neq \mu_j\,, j = 1, 2, \dots. \tag{4.21}$$

Due to (4.3), ρ is smooth on $(-\infty, 1) \cup (1, 3) \cup (3, \infty) \setminus \{\mu_j : j = N + 2, N + 3, \dots\}$. On the other hand, as $\mu \to 3$,

$$\phi_\mu \to L_0^{-1} w_{x_0} = \frac{1}{2} w_{x_0} + \frac{1}{2} y \cdot \nabla w_{x_0}, \tag{4.22}$$

$$\rho(\mu) \to \int_{\mathbb{R}^2} w_{x_0} \left(\frac{1}{2} w_{x_0} + \frac{1}{2} y \cdot \nabla w_{x_0} \right) = 0. \tag{4.23}$$

Here we have used the fact that $N = 2$ and $p = 3$. Thus for $1 < \mu < 3$, $\rho(\mu) > 0$. This implies that for $0 < \beta < 1$, $\int_{\mathbb{R}^2} u_h R_{h,2} > 0$ and thus $(\nabla^2 d(0, 0))$ has *two* positive eigenvalues.

Now we consider $\mu \in (-\infty, 1)$, i.e., $\beta > 1$. By the standard maximal principle, $\phi_\mu < 0$ in $\mathbb{R}^2$ for $\mu < 0$. Consequently, $\rho(\mu) < 0$ for $\mu < 0$. Hence by (4.21), $\rho(\mu) < 0$ for $\mu \in (-\infty, 1)$, i.e., $\beta > 1$. This implies that $\int_{\mathbb{R}^2} u_h R_{h,2} < 0$ and thus $(\nabla^2 d(0, 0))$ has only *one* positive eigenvalue.

In conclusion, we see that the matrix $(\nabla^2 d(0, 0))$ has two positive eigenvalues when $0 < \beta < 1$ and one positive eigenvalue when $\beta > 1$. That is $p = 2$ when $0 < \beta < 1$ and $p = 1$ when $\beta > 1$. By Lemmas 4.1 and 4.2, we also deduce that $n(\mathbb{L}_h) = 2$ when $0 < \beta < 1$ and $n(\mathbb{L}_h) = 1$ when $\beta > 1$. Hence for $\beta > 0, \beta \neq 1$, we have $n(\mathbb{L}_h) = p$. Therefore, we conclude that $(u_{h,0,0}, v_{h,0,0}) = (\frac{1}{\sqrt{1+\beta}} u_h, \frac{1}{\sqrt{1+\beta}} v_h)$ is orbitally stable if $0 < \beta,\ \beta \neq 1$. This completes the proof of Theorem 1.2.

Acknowledgment. We thank the referees for bringing our attention to [12] and [13] and for many helpful suggestions.

REFERENCES

[1] A. Ambrosetti, M. Badiale, and S. Cingolani, *Semiclassical states of nonlinear Schrödinger equations*, Arch. Ration. Mech. Anal., 140 (1997), pp. 285–300.

[2] A. Ambrosetti and E. Colorado, *Bound and ground states of coupled nonlinear Schrödinger equations*, C. R. Math. Acad. Sci. Paris, 342 (2006), pp. 453–458.

[3] A. Ambrosetti, A. Malchiodi, and W.-M. Ni, *Singularly perturbed elliptic equations with symmetry: Existence of solutions concentrating on spheres, Part* I, Comm. Math. Phys., 235 (2003), pp. 427–466.

[4] T. Bartsch, Z.-Q. Wang, and J. Wei, *Bound states for a coupled Schrödinger system*, J. Fixed Point Theory Appl., 2 (2007), pp. 353–367.

[5] H. Berestycki and T. Cazenave, *Instabilité des états stationnaires dans les équations de Schrödinger et de Klein-Gordon non linéaires*, C. R. Acad. Sci. Paris Sér. I Math., 293 (1981), pp. 489–492.

[6] T. Cazenave and P. L. Lions, *Orbital stability of standing waves for some nonlinear Schrödinger equations*, Comm. Math. Phys., 85 (1982), pp. 549–561.

[7] S. L. Cornish, S. T. Thompson, and C. E. Wieman, *Formation of bright matter-wave solitons during the collapse of Bose–Einstein condensates*, Phys. Rev. Lett., 96 (2006), article 170401.

[8] M. del Pino and P. Felmer, *Semiclassical states for nonlinear Schrödinger equation*, J. Funct. Anal., 149 (1997), pp. 245–265.

[9] M. del Pino and P. Felmer, *Semiclassical states of nonlinear Schrödinger equations: A variational reduction method*, Math. Ann., 324 (2002), pp. 1–32.

[10] E. A. Donley, N. R. Claussen, S. L. Cornish, J. L. Roberts, E. A. Cornell, and C. E. Wieman, *Dynamics of collapsing and exploding Bose–Einstein condensates*, Nature, 19 (2001), pp. 295–299.

[11] A. Floer and A. Weinstein, *Nonspreading wave packets for the cubic Schrödinger equation with a bounded potential*, J. Funct. Anal., 69 (1986), pp. 397–408.
[12] R. Fukuizumi, *Stability of standing waves for nonlinear Schrödinger equations with critical power nonlinearity and potentials*, Adv. Differential Equations, 10 (2005), pp. 259–276.
[13] R. Fukuizumi and L. Jeanjean, *Stability of standing waves for a nonlinear Schrödinger equation with a repulsive Dirac delta potential*, Discrete Contin. Dyn. Syst. to appear.
[14] M. Grillakis, J. Shatah, and W. Strauss, *Stability theory of solitary waves in the presence of symmetry* I, J. Funct. Anal., 74 (1987), pp. 160–197.
[15] M. Grillakis, J. Shatah, and W. Strauss, *Stability theory of solitary waves in the presence of symmetry* II, J. Funct. Anal., 94 (1990), pp. 308–348.
[16] M. Grossi, *On the number of single-peaked solutions of the nonlinear Schrödinger equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 19 (2002), pp. 261–280.
[17] C. Gui, *Existence of multi-bump solutions for nonlinear Schrödinger equations via variational method*, Comm. Partial Differential Equations, 21 (1996), pp. 787–820.
[18] T. L. Horng, S. C. Gou, and T. C. Lin, *Bending-wave instability of a vortex ring in a trapped Bose–Einstein condensate*, Phys. Rev. A, 74 (2006), article 041603.
[19] X. Kang and J. Wei, *On interacting bumps of semiclassical states of nonlinear Schrödinger equations*, Adv. Differential Equations, 5 (2000), pp. 899–928.
[20] T.-C. Lin and J.-C. Wei, *Ground state of N coupled Nonlinear Schrödinger Equations in R^n, $n \leq 3$*, Comm. Math. Phys., 255 (2005), pp. 629–653.
[21] T. C. Lin and J. Wei, *Spikes in two-component systems of nonlinear Schrödinger equations with trapping potentials*, J. Differential Equations, 229 (2006), pp. 538–569.
[22] T. C. Lin and J. Wei, *Symbiotic bright solitary wave solutions of coupled nonlinear Schödinger equations*, Nonlinearity, 19 (2006), pp. 2755–2773.
[23] L. A. Maia, E. Montefusco, and B. Pellacci, *Positive solutions for a weakly coupled nonlinear Schrödinger system*, J. Differential Equations, 229 (2006), pp. 743–767.
[24] D. Mihalache, D. Mazilu, F. Lederer, B. A. Malomed, L.-C. Crasovan, Y. V. Kartashov, and L. Torner, *Stable three-dimensional solitons in attractive Bose–Einstein condensates loaded in an optical lattice*, Phys. Rev. A, 72 (2005), article 021601.
[25] Y.-G. Oh, *On positive multi-bump states of nonlinear Schrödinger equation under multiple well potentials*, Comm. Math. Phys., 131 (1990), pp. 223–253.
[26] Y.-G. Oh, *Stability of semiclassical bound state of nonlinear Schrödinger equations with potentials*, Comm. Math. Phys., 121 (1989), pp. 11–33.
[27] V. M. Perez-Garcia and J. Belmonte Beitia, *Symbiotic solitons in heteronuclear multicomponent Bose–Einstein condensates*, Phys. Rev. A, 72 (2005), article 033620.
[28] L. Pitaevskii and S. Stringari, *Bose–Einstein Condensation*, Oxford University Press, Oxford, 2003.
[29] P. Rabinowitz, *On a class of nonlinear Schrödinger equations*, Z. Angew. Math. Phys., 43 (1992), pp. 270–291.
[30] B. Sirakov, *Standing wave solutions of the nonlinear Schrödinger equation in $\mathbb{R}^N$*, Ann. Mat. Pura Appl. (4), 4 (2002), pp. 73–83.
[31] B. Sirakov, *Least energy solitary waves for a system of nonlinear Schrödinger equations*, Comm. Math. Phys., to appear.
[32] X. Wang, *On concentration of positive bound states of nonlinear Schrödinger equations*, Comm. Math. Phys., 153 (1993), pp. 229–243.
[33] Z.-Q. Wang, *Existence and symmetry of multi-bump solutions for nonlinear Schrödinger equations*, J. Differential Equations, 159 (1999), pp. 102–137.
[34] J. Wei, *On the construction of single-peaked solutions to a singularly perturbed elliptic Dirichlet problem*, J. Differential Equations, 129 (1996), pp. 315–333.
[35] J. Wei, *On the interior spike layer solutions for some singular perturbation problems*, Proc. Roy. Soc. Edinburgh Sect. A, 128 (1998), pp. 849–874.
[36] M. Weinstein, *Nonlinear Schrödinger equations and sharp interpolation estimates*, Comm. Math. Phys., 87 (1983), pp. 567–576.
[37] M. I. Weinstein, *The Connection between Finite and Infinite-Dimensional Dynamical Systems*, AMS, Providence, RI, 1989.

SIAM J. MATH. ANAL.
Vol. 40, No. 1, pp. 382–412

FORCED VIBRATIONS OF A NONHOMOGENEOUS STRING*

PIETRO BALDI† AND MASSIMILIANO BERTI‡

Abstract. We prove existence of vibrations of a nonhomogeneous string under a nonlinear time periodic forcing term in the case in which the forcing frequency avoids resonances with the vibration modes of the string (nonresonant case). The proof relies on a Lyapunov–Schmidt reduction and a Nash–Moser iteration scheme.

Key words. nonlinear wave equations, periodic solutions, small divisors, Nash–Moser iteration scheme

AMS subject classifications. 35B10, 35L70, 58C15

DOI. 10.1137/060665038

1. Introduction.

In this paper we study forced vibrations of a nonhomogeneous string,

$$\begin{cases} \rho(x)u_{tt} - (p(x)u_x)_x = \mu f(x, \omega t, u), \\ u(0,t) = u(\pi,t) = 0, \end{cases} \tag{1}$$

where $\rho(x) > 0$ is the mass per unit length, $p(x) > 0$ is the modulus of elasticity multiplied by the cross-sectional area (see [15, p. 291]), $\mu > 0$ is a parameter, and the nonlinear forcing term $f(x, \omega t, u)$ is $(2\pi/\omega)$-periodic in time (i.e., $f(x, \cdot, u)$ is 2π-periodic).

Equation (1) is a nonlinear model also for propagation of waves in nonisotropic media describing seismic phenomena; see, e.g., [2].

We look for $(2\pi/\omega)$-time periodic solutions $u(x,t)$ of (1).

This problem has received wide attention since the pioneering paper of Rabinowitz [26] dealing with the weakly nonlinear homogeneous string with $\rho(x) = p(x) = 1$, μ small, and $\omega = 1$. In this case the forcing frequency ω enters in resonance with the proper eigenfrequencies $\omega_j = j \in \mathbb{N}$ of the string.

For functions 2π-periodic in time and satisfying spatial Dirichlet boundary conditions, the spectrum $\{-l^2 + j^2,\ l \in \mathbb{Z},\ j \geq 1\}$ of the D'Alembertian operator $\partial_{tt} - \partial_{xx}$ possesses the zero eigenvalue with infinite multiplicity (for $|l| = j$), but the other eigenvalues are well separated. The corresponding infinite-dimensional bifurcation problem is solved in [26] for nonlinearities f which are monotone in u; see [7] for nonmonotone f.

Subsequently many other results, both of bifurcation and of a global nature ($\mu = 1$), have been obtained, still for rational forcing frequencies $\omega \in \mathbb{Q}$, relying on the separation properties of the spectrum; see, e.g., [27, 28, 14, 31, 4].

When the forcing frequency $\omega \in \mathbb{R}\backslash\mathbb{Q}$ is irrational (nonresonant case) the situation is completely different. Indeed the wave operator $\omega^2\partial_{tt} - \partial_{xx}$ does not possess the

*Received by the editors July 13, 2006; accepted for publication (in revised form) September 14, 2007; published electronically April 25, 2008. This work was supported by MURST under the national project "Variational Methods and Nonlinear Differential Equations."
http://www.siam.org/journals/sima/40-1/66503.html
†SISSA, via Beirut 2-4, 34014 Trieste, Italy (baldi@sissa.it).
‡Dipartimento di Matematica e Applicazioni "R. Caccioppoli," Università di Napoli "Federico II," via Cintia, 80126 Napoli, Italy (m.berti@unina.it).

zero eigenvalue, but its spectrum $\{-\omega^2 l^2 + j^2,\ l \in \mathbb{Z},\ j \geq 1\}$ accumulates to zero for almost every ω. This is a "small divisors problem."

We underline that this "small divisors" phenomenon arises naturally for more realistic model equations like (1) where the density $\rho(x)$ and the modulus of elasticity $p(x)$ are not constant. Indeed in this case the eigenfrequencies ω_j of the string are no longer integer numbers, having the asymptotic expansion

$$\omega_j^2 = \frac{j^2}{c^2} + b + O\Big(\frac{1}{j}\Big) \tag{2}$$

with suitable constants c, b depending on ρ, p; see (66).

If $\omega = m/n \in \mathbb{Q}$, good separation properties of the spectrum can be recovered when $p(x) = \rho(x)$ (so $c = 1$) and assuming the extra condition $b \neq 0$; see [3, 29]. Indeed in this case the linear spectrum

$$-\omega^2 l^2 + \omega_j^2 = -\omega^2 l^2 + j^2 + b + O\Big(\frac{1}{j}\Big)$$

possesses at most finitely many zero eigenvalues and the remaining part of the spectrum is far away from zero. On the other hand, if $b = 0$, the eigenvalues with $(l, j) \in (n, m)\mathbb{N}$ tend to zero (also in the case $\omega \in \mathbb{Q}$!).

Existence of weak solutions in the nonresonant case was proved by Acquistapace [1] for $\rho = 1$, μ small, and for a zero measure set of forcing frequencies ω for which the eigenvalues $-\omega^2 l^2 + \omega_j^2$ are far away from zero. These frequencies are essentially the numbers whose continued fraction expansion is bounded; see [30].

For a similar zero measure set of frequencies, McKenna [23] has obtained some result when $\mu = 1$, $\rho = p = 1$, and $f(x, t, u) = g(u) + h(x, t)$ with g uniformly Lipschitz, via a fixed point argument; see also [5]; for related results using variational methods see [18, 10].

Existence of classical solutions of (1) for a positive measure set of frequencies was proved by Plotnikov and Yungerman [24] for the homogeneous string $\rho = p = 1$, μ small, and f monotone in u. This monotonicity condition allows one to control the constant coefficient in the asymptotic expansion of the eigenvalues (like b in (2)) of some perturbed linearized operator.

Recently Fokam [19] has proved existence of classical periodic solutions for large frequencies ω in a set of asymptotically full measure for the homogeneous string $\rho = p = 1$ plus a potential, $\mu = 1$ and $f = u^3 + h(x, t)$ with h a trigonometric polynomial odd in time and space.

In the present paper we prove existence of classical solutions of the nonhomogeneous string (1) for every $\rho(x)$, $p(x) > 0$ for general nonlinear terms $f(x, \omega t, u)$, and for (μ, ω) belonging to a large measure Cantor set B_γ, when the ratio μ/ω is small. Our Theorem 1 covers both the case $\mu \to 0$ (weak forcing) and the case $\omega \to +\infty$ (rapid forcing).

In the limit $\mu/\omega \to 0$ the solution we find tends to a static equilibrium $v(x)$ with smaller, zero average oscillations $w(x, t)$ of amplitude $O(\mu/\omega)$; see (13), (14), and Figure 1. The nonlinearity f selects such v through the infinite-dimensional bifurcation equation (10), which possesses nondegenerate solutions under natural assumptions on f; see Hypothesis (V). This problem is not present in [19], where, thanks to the symmetry assumptions on f, there is no bifurcation equation.

Considering the structure of the expected solution, it is natural to attack the problem via a Lyapunov–Schmidt decomposition.

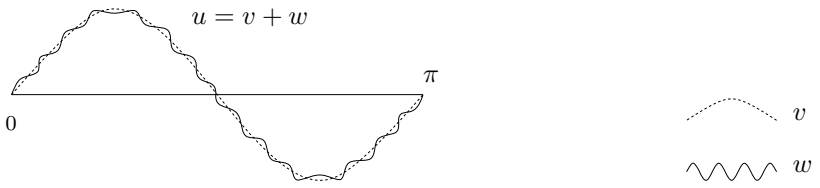


FIG. 1. *The solution* $u(x,t) = v(x) + w(x,t)$ *of* (1).

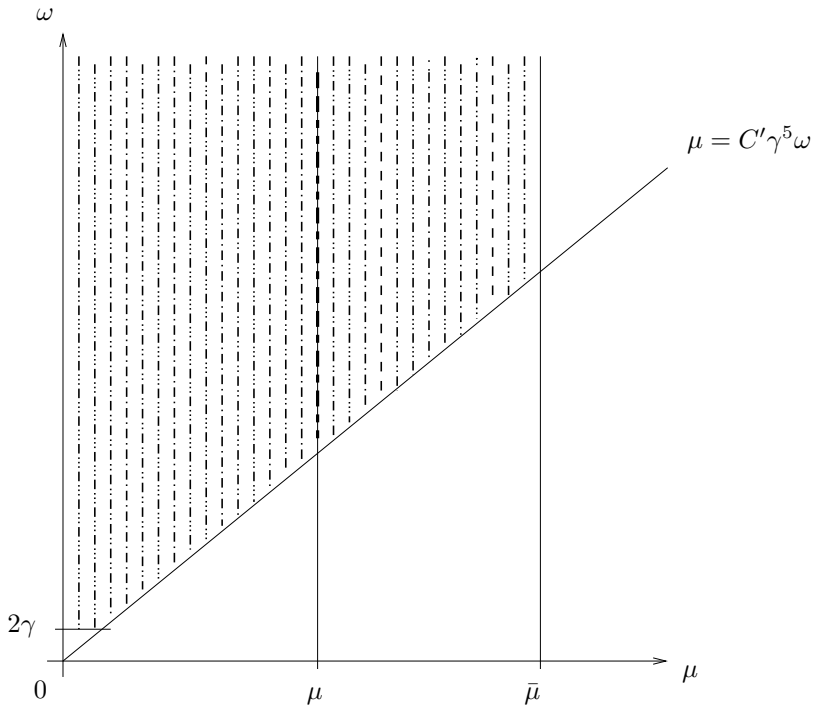


FIG. 2. *The Cantor set* B_γ.

In the range equation (to find w) a small divisors problem arises, and we solve it with a Nash–Moser-type iterative scheme. The inversion of the "linearized operators"—which is the core of any Nash–Moser scheme—is obtained adapting the techniques of [8] to the present time-dependent case (section 6). This method is also reminiscent of the approach of Kuksin (unpublished) explained by Bourgain in [13, pp. 90–94]. See also the works of Craig and Wayne [16, 17] and Bourgain [11, 12] for related techniques.

It is in the solution of the range equation where the interaction between the forcing frequency ω and the normal modes of oscillations of the string linearized at different positions (approximating better and better the final string configuration) appears.

The set B_γ of "nonresonant" parameters (μ, ω) for which we find a solution of the range equation (and then of (1)) is constructed avoiding these primary resonances. In particular the forcing frequency ω must not enter in resonance with the normal frequencies of oscillations of the string linearized at the limit solution; see (15). At the end of the construction we obtain a large measure Cantor set B_γ which looks like Figure 2. Outside this set the effect of resonance phenomena shall in general destroy the existence of periodic solutions like those found in Theorem 1.

Finally we recall that related existence results of periodic and quasi-periodic solutions for autonomous Hamiltonian PDEs have been obtained via KAM-type techniques since the pioneering works of Kuksin [21] and Wayne [32]; see also [22] and the references therein.

We now present rigorously our results.

1.1. Main result. After a time rescaling we look for 2π-periodic solutions of

$$\begin{cases} \omega^2\rho(x)u_{tt} - (p(x)u_x)_x = \mu f(x,t,u), \\ u(0,t) = u(\pi,t) = 0, \end{cases} \tag{3}$$

where $\mu \in [0,\bar\mu]$ for some given $\bar\mu > 0$, under the 2π-periodic forcing term

$$f(x,t,u) = \sum_{l\in\mathbb{Z}} f_l(x,u)e^{ilt} = f_0(x,u) + \bar f(x,t,u), \tag{4}$$

where

$$\bar f(x,t,u) := \sum_{l\neq 0} f_l(x,u)e^{ilt}\,.$$

We suppose that f is analytic in (t,u); more precisely,

$$f(x,t,u) = \sum_{l\in\mathbb{Z},\, k\in\mathbb{N}} f_{lk}(x)\, u^k\, e^{ilt},$$

where $f_{lk}(x) \in H^1((0,\pi);\mathbb{C})$, $f_{-l,k} = f_{lk}^*$ (the symbol z^* denotes the complex conjugate of $z\in\mathbb{C}$), and we assume the following hypothesis on the decay of the coefficients $\|f_{lk}\|_{H^1}$.

Hypothesis (F). *There exist* $2\sigma_0 > 0$, $r > 0$ *such that*

$$\sum_{l\in\mathbb{Z}} \|f_{lk}\|_{H^1}^2 (1+l^2)e^{(2\sigma_0)2|l|} := C_k^2(f) < \infty \quad \textit{and} \quad \sum_{k=0}^{+\infty} C_k(f)\, r^k < \infty\,.$$

For example, trigonometric polynomials in t and polynomials in u, namely,

$$f(x,t,u) = \sum_{|l|\le L, 0\le k\le K} f_{lk}(x)\, u^k\, e^{ilt} \tag{5}$$

for some $L, K \in \mathbb{N}$, satisfy Hypothesis (F) for every σ_0, r.

Remark 1. We notice that if $f(x,t,0) = \sum_{l\in\mathbb{Z}} f_{l0}(x)\, e^{ilt} \neq 0$, then (3) does not possess the trivial solution $u = 0$.

We look for periodic solutions of (3) in the Hilbert space

$$X_{\sigma,s} := \Big\{ u : \mathbb{T} \to H_0^1((0,\pi);\mathbb{R}),\ u(x,t) = \sum_{l\in\mathbb{Z}} u_l(x)\, e^{ilt}, \quad u_l \in H_0^1((0,\pi);\mathbb{C}),$$

$$u_{-l} = u_l^*, \quad \|u\|_{\sigma,s}^2 := \sum_{l\in\mathbb{Z}} \|u_l\|_{H^1}^2 (1+l^{2s})e^{2\sigma|l|} < \infty \Big\}$$

of 2π-periodic-in-time functions valued in $H^1((0,\pi);\mathbb{R})$ which have a bounded analytic extension on the complex strip $|\mathrm{Im}\, t| < \sigma$ with trace function on $|\mathrm{Im}\, t| = \sigma$ belonging to $H^s(\mathbb{T}; H^1((0,\pi);\mathbb{C}))$.

For $s > 1/2$, $X_{\sigma,s}$ is a multiplicative Banach algebra:

$$\|uv\|_{\sigma,s} \le c_s \|u\|_{\sigma,s} \|v\|_{\sigma,s} \quad \forall\, u, v \in X_{\sigma,s} \tag{6}$$

with

$$c_s := 2^s \Big(\sum_{n \in \mathbb{Z}} \frac{1}{1+n^{2s}} \Big)^{1/2}; \tag{7}$$

see, e.g., [6]. We shall use the notation X_σ, resp., $\| \, \|_\sigma$, for $X_{\sigma,1}$, resp., $\| \, \|_{\sigma,1}$.

1.2. The Lyapunov–Schmidt reduction. To find solutions of (3) we implement the Lyapunov–Schmidt reduction according to the decomposition

$$X_{\sigma,s} = V \oplus (W \cap X_{\sigma,s}),$$

where

$$V := H^1_0(0,\pi)\,, \qquad W := \Big\{ w = \sum_{l \neq 0} w_l(x)\, e^{ilt} \in X_{0,s} \Big\},$$

writing every $u \in X_{\sigma,s}$ as $u = u_0(x) + \sum_{l \neq 0} u_l(x)\, e^{ilt}$.

Projecting (3) with

$$u = v + w\,, \quad v \in V,\ w \in W\,,$$

yields

$$\begin{cases} -(pv')' = \mu \Pi_V f(v+w) & \text{(bifurcation equation)}, \\ L_\omega w = \mu \Pi_W f(v+w) & \text{(range equation)}, \end{cases} \tag{8}$$

where Π_V, Π_W denote the projectors, $f(u)(x,t) := f(x,t,u(x,t))$, and

$$L_\omega u := \omega^2 \rho(x) u_{tt} - (p(x) u_x)_x\,.$$

We shall find solutions of (8) when μ/ω is small. In this limit w tends to 0 and the bifurcation equation reduces to the time-independent equation

$$-(pv')' = \mu f_0(v) \tag{9}$$

because, by (4), for $w = 0$

$$\Pi_V f(v) = \Pi_V f_0(x, v(x)) + \Pi_V \bar f(x,t,v(x)) = f_0(v)\,.$$

The infinite-dimensional "0th order bifurcation equation" (9) is a second order ODE, which, under natural conditions on f_0, possesses nondegenerate solutions satisfying the boundary conditions $v(0) = v(\pi) = 0$.

Hypothesis (V). *The problem*

$$\begin{cases} -(p(x)v'(x))' = \mu f_0(x, v(x)), \\ v(0) = v(\pi) = 0 \end{cases} \tag{10}$$

admits a solution $\bar v \in H^1_0(0,\pi)$ which is nondegenerate, namely, the linearized equation

$$-(ph')' = \mu\, f_0'(\bar v) h$$

possesses in $H^1_0(0,\pi)$ only the trivial solution $h = 0$.

We note that, for $\mu = 0$, the trivial solution $\bar{v} = 0$ is nondegenerate, so, by the implicit function theorem, Hypothesis (V) is automatically satisfied for μ small. We deal also with the case μ not small; see, for example, Lemmas 2 and 3.

By the implicit function theorem, Hypothesis (V) implies the existence of a smooth map

$$(\mu, w) \mapsto v(\mu, w) \in V$$

such that $v(\mu, w)$ solves the bifurcation equation in (8); see Lemma 4.

Remark 2. For a discussion about the difficulties caused by a degenerate solution, we refer to [9].

Let λ_j denote the eigenvalues of the Sturm–Liouville problem

$$\begin{cases} -(p(x)y'(x))' = \lambda\rho(x)y(x), \\ y(0) = y(\pi) = 0 \end{cases} \tag{11}$$

and $\omega_j := \sqrt{\lambda_j}$. These are the frequencies of the free vibrations of the string (note that all the eigenvalues λ_j are positive). Physically, it is the sequence of the fundamental tone ω_1 and all its overharmonics $\omega_2, \omega_3, \dots$ which compose the musical note of the string.

For $\gamma \in (0, 1)$ we define

$$A_\gamma := \Big\{(\mu,\omega) \in (\mu_1,\mu_2) \times (\gamma, +\infty) \ : \ \frac{\mu}{\omega} < C'\gamma^5, \quad |\omega l - \omega_j| > \frac{\gamma}{l^\tau} \quad \forall\, l = 1, \dots, N_0, \quad j \geq 1\Big\}, \tag{12}$$

where ω_j are given by (11), and (μ_1, μ_2), $N_0 \in \mathbb{N}$, $C' > 0$ shall be fixed in the next theorem.

THEOREM 1 (existence). *Suppose $p(x), \rho(x) > 0$ are of class $H^3(0, \pi)$, f satisfies Hypothesis* (F)*, and Hypothesis* (V) *holds for some $\mu_0 \in [0, \bar{\mu}]$.*

Fix $\tau \in (1, 2)$, $\gamma \in (0, 1)$. There exist a neighborhood (μ_1, μ_2) of μ_0, $N_0 \in \mathbb{N}$, positive constants C, C' (depending on $\rho, p, f, \bar{\mu}, \bar{v}, \tau$), a map

$$\tilde{w} \in C^\infty(A_\gamma, X_{\sigma_0/2} \cap W),$$

and a Cantor set $B_\gamma \subset A_\gamma$ of positive measure such that, for all $(\mu, \omega) \in B_\gamma$,

$$\tilde{u}(\mu, \omega) := v(\mu, \tilde{w}(\mu, \omega)) + \tilde{w}(\mu, \omega) \in V \oplus (W \cap X_{\sigma_0/2}) \tag{13}$$

is a classical solution of (3) *and satisfies*

$$\tilde{u}(\cdot, t) \in H^3(0, \pi) \cap H^1_0(0, \pi) \quad \forall t \in \mathbb{R}\,.$$

The Cantor set B_γ is defined in (15) *and satisfies the measure estimate* (56)*.*

Furthermore, for all $(\mu, \omega) \in A_\gamma$ the following estimates hold:

$$\|\tilde{w}(\mu, \omega)\|_{\sigma_0/2} \leq C\frac{\mu}{\gamma\omega}\,, \quad \|v(\mu, \tilde{w}(\mu, \omega)) - v(\mu, 0)\|_{H^1} \leq C\frac{\mu}{\gamma\omega}, \tag{14}$$

and $\|v(\mu, 0) - \bar{v}\|_{H^1} \leq C|\mu - \mu_0|$.

The neighborhood (μ_1, μ_2) of μ_0 is fixed in Lemma 4, the integer N_0 is fixed in Lemma 9, and the constant C' is fixed in Lemma 13.

Estimate (14) shows how close the solution $\tilde u$ is to the static configuration $v(\mu,0)$; see Figure 1.

Remark 3. We underline that the function $\tilde w(\mu,\omega)$, as well as $\tilde u(\mu,\omega)$, is defined for all the values of the parameters $(\mu,\omega)\in A_\gamma$ and not only on the Cantor set B_γ ($\tilde w(\mu,\omega)$ is introduced in Lemma 11). What is true is that if $(\mu,\omega)\in B_\gamma$, then $\tilde w(\mu,\omega)$ solves the range equation; see Theorem 3. As a consequence, if $(\mu,\omega)\in B_\gamma$, then $\tilde u(\mu,\omega)$ solves (3).

The Cantor set B_γ is explicitly defined by

$$B_\gamma := \Big\{(\mu,\omega)\in(\mu_1,\mu_2)\times(2\gamma,+\infty)\,:\,|\omega l-\omega_j|>\frac{2\gamma}{l^\tau}\quad\forall\, l=1,\dots,N_0,\quad j\geq 1,$$

$$\frac{\mu}{\omega}<C'\gamma^5,\quad \Big|\omega l-\frac{j}{c}\Big|>\frac{2\gamma}{l^\tau},\quad |\omega l-\tilde\omega_j(\mu,\omega)|>\frac{2\gamma}{l^\tau}\quad\forall\, l,j\geq 1\Big\},\tag{15}$$

where

$$c:=\frac{1}{\pi}\int_0^\pi\Big(\frac{\rho(x)}{p(x)}\Big)^{1/2}dx\tag{16}$$

and $\tilde\lambda_j(\mu,\omega):=\tilde\omega_j^2(\mu,\omega)$ denote the eigenvalues of the Sturm–Liouville problem

$$\begin{cases}-(py')'-\mu\Pi_V f'(\tilde u(\mu,\omega))\,y=\lambda\rho y,\\ y(0)=y(\pi)=0\,.\end{cases}\tag{17}$$

Note that B_γ is constructed by means of the function $\tilde u(\mu,\omega)$, which is defined for all $(\mu,\omega)\in A_\gamma$; see Remark 3.

Remark 4. If some $\tilde\lambda_j(\mu,\omega)$ is negative, then $\tilde\omega_j(\mu,\omega)=i\sqrt{|\tilde\lambda_j(\mu,\omega)|}$ is a purely imaginary complex number and the nonresonance conditions in (15) are trivially satisfied.

The Cantor set B_γ is large in a measure theoretical sense; see section 4.3. In particular, for all $\mu\in(\mu_1,\mu_2)$,

$$S(\mu):=\big\{\omega:(\mu,\omega)\in\cup_{\gamma\in(0,1)}B_\gamma\big\}$$

has asymptotically full measure at $\omega\to+\infty$, i.e.,

$$\lim_{\omega\to+\infty}|S(\mu)\cap(\omega,\omega+1)|=1\,.\tag{18}$$

Analogously,

$$S(\omega):=\big\{\mu:(\mu,\omega)\in\cup_{\gamma\in(0,1)}B_\gamma\big\}$$

satisfies, for all $\omega'>0$ and for all $\gamma'\in(0,1)$,

$$\lim_{\mu\to0}\Big|\Big\{\omega\in(\omega',\omega'+1):\ \frac{|S(\omega)\cap(0,\mu)|}{\mu}\ \geq 1-\gamma'\Big\}\Big|=1\,.\tag{19}$$

Finally we discuss the regularity of the solution $\tilde u(x,t)$ found in Theorem 1 with respect to x (by construction $\tilde u$ is analytic with respect to t).

THEOREM 2 (regularity). *Assume the hypotheses of Theorem* 1. *In addition, suppose that, for some* $m\geq 3$,

$$\rho(x)\in H^m(0,\pi)\,,\quad p(x)\in H^{m+1}(0,\pi)\,,\quad f_{lk}(x)\in H^m(0,\pi)\ \ \forall\, l,k\tag{20}$$

and, for some $r_m > 0$,

$$\sum_{l\in\mathbb{Z},\, k\geq 0} \|f_{lk}\|_{H^m} r_m^k < +\infty\,. \tag{21}$$

If $\|\tilde u(\cdot\,,t)\|_{H^1} r_m^{-1}$ *is small enough, then*

$$\tilde u(\cdot\,,t) \in H^{m+2}(0,\pi)\cap H_0^1(0,\pi)\,. \tag{22}$$

This conclusion holds true, for example, when $f_0(x,0) = d_u f_0(x,0) = 0$ *and* $\mu/\gamma\omega$ *is small enough.*

Note that the regularity (22) requires no skewsymmetry assumptions on f and requires just a smallness condition for the H^1 norm of $\tilde u(\cdot\,,t)$.

Remark 5. If $f(x,t,u)$ is a trigonometric polynomial in t and a polynomial in u as in (5), then the series in (21) is a finite sum. Therefore the conclusion (22) is true without smallness conditions for $\tilde u$.

In particular, if $\rho(x)$, $p(x)$, $f_{lk}(x)$ are C^∞ (for example, $f = \cos x \cos t\,(1+u^2)$), then the solution $\tilde u$ is C^∞ also in the variable x (the above f does not satisfy the skewsymmetry assumption (24)).

The subtle problem to prove Theorem 2 is that, because of Dirichlet boundary conditions, the Sobolev regularity of a function with respect to x is *not* characterized by the rapid decaying properties of the Fourier coefficients (unless we assume skewsymmetry assumptions on the nonlinearity and restrict solutions to $u(x,t)$ odd in x; see Remark 6). Theorem 2 is proved in section 5 via bootstrap arguments.

1.3. Outline of the proof. In section 2 we prove that, under assumption (F), the composition operator induced by the nonlinearity f on $X_{\sigma,s}$ is an analytic map.

In section 3, we find a solution $v(\mu,w)$ of the infinite-dimensional bifurcation equation in (8). Thanks to assumption (V) (which is verified on several examples in Lemmas 2 and 3), $v(\mu,w)$ is obtained in Lemma 4 by a standard implicit function theorem.

In section 4 we solve the range equation by means of an iterative Nash–Moser implicit function theorem. The final theorem, Theorem 3, is proved in several steps.

In section 4.1 we find inductively a sequence of approximate solutions $w_n(\mu,\omega)$ defined on smaller and smaller subsets A_n of the parameters (μ,ω) (see (39)). The reason for these "excisions" is to avoid resonance phenomena in order to prove the invertibility of the linearized operators obtained at each step of the iteration; see conditions (33)–(34) in Lemma 7.

In section 4.2 we extend these approximate solutions $w_n(\mu,\omega)$ to C^∞-functions $\tilde w_n(\mu,\omega)$ defined for all the values of the parameters (μ,ω) and converging (superexponentially fast) to a C^∞ map $\tilde w$ defined for all (μ,ω); see Lemma 11. It is in proving the regularity of w_n with respect to the parameters (μ,ω) that we find it convenient to define the approximate solutions w_n as exact solutions of (41) (with $k=n$); see Remark 9.

In Lemma 12, we prove that the Cantor set B_γ, defined in (15) by means of $\tilde w$, is contained in A_n (which depends on w_{n-1}) for each n. This is a consequence of the superexponentially fast convergence of $\tilde w_n$ to $\tilde w$; see (52).

In section 4.3 we prove that B_γ is a large set in a measure theoretical sense.

In all the previous steps we have to assume smallness conditions for μ/ω. The most restrictive one is $\mu/\omega < C'\gamma^5$ in the definition (12) of A_γ.

In section 5 we conclude the proof of the existence Theorem 1, and we prove the regularity Theorem 2.

In section 6 we study the key step for the inversion of the linearized operators. Lemma 7 is obtained by a variant of the techniques developed in [8]. In particular, the key estimate on the small divisors of Lemma 18 is reminiscent of the method of Kuksin explained in [13, pp. 90–94].

Notation. The symbols K, K_i, K_i' shall denote positive constants depending only on $\rho, p, f, \bar\mu, \bar v, \tau$.

2. Regularity of the composition operator. We first prove the analyticity of the composition operator

$$u(x,t) \mapsto f(x,t,u(x,t))$$

induced by f on $X_{\sigma,s}$.

By the Banach algebra property (6) of $X_{\sigma,s}$ the composition operator

$$u \mapsto u^k \qquad \forall k \in \mathbb{N}$$

is an analytic map from $X_{\sigma,s}$ into itself. Thanks to the rapid decay of the coefficients $\|f_{lk}\|_{H^1}$ assumed in Hypothesis (F), this property holds true also for the composition operator $f(u)$.

LEMMA 1. *Let f satisfy assumption* (F). *For every $\sigma \in [0,\sigma_0]$, $s > 1/2$, the composition operator f is analytic on the ball $\{u \in X_{\sigma,s} : \|u\|_{\sigma,s} < r/c_s\}$, where c_s is defined in* (7).

Proof. First note that

$$\sum_{l\in\mathbb{Z}} \|u_l\|_\infty \le \sqrt{\frac{\pi}{2}} \sum_{l\in\mathbb{Z}} \|u_l\|_{H^1} \le \sqrt{\frac{\pi}{2}} \Big(\sum_{l\in\mathbb{Z}} \|u_l\|_{H^1}^2 (1+l^{2s}) \Big)^{1/2} \Big(\sum_{l\in\mathbb{Z}} \frac{1}{1+l^{2s}} \Big)^{1/2}$$

so $\|u\|_\infty \le c_s \|u\|_{\sigma,s}$ for all $u \in X_{\sigma,s}$, $\sigma \ge 0$, $s > 1/2$, and $f(x,t,u(x,t))$ is well-defined.

By definition of the norm $\|\cdot\|_{\sigma,s}$, there exists $C := C(\sigma_0, s) > 0$ such that for all $\sigma \in [0,\sigma_0]$ and for all $k \in \mathbb{N}$,

$$\big\| \sum_{l\in\mathbb{Z}} f_{lk}(x) e^{ilt} \big\|_{\sigma,s} \le C \, \big\| \sum_{l\in\mathbb{Z}} f_{lk}(x) e^{ilt} \big\|_{2\sigma_0,1} .$$

Next

$$\big\| \sum_{l\in\mathbb{Z}} f_{lk}(x) e^{ilt} \big\|_{2\sigma_0,1}^2 = \sum_{l\in\mathbb{Z}} \|f_{lk}\|_{H^1}^2 (1+l^2) e^{(2\sigma_0)2|l|} =: C_k^2(f) < +\infty$$

by assumption (F). Therefore $\sum_{l\in\mathbb{Z}} f_{lk}(x) e^{ilt} \in X_{\sigma,s}$ and

$$\big\| \sum_{l\in\mathbb{Z}} f_{lk}(x) e^{ilt} \big\|_{\sigma,s} \le C \, C_k(f) . \tag{23}$$

Using the algebra property (6) of $X_{\sigma,s}$ and (23)

$$\begin{aligned} \|f(u)\|_{\sigma,s} &\le \sum_{k=0}^\infty \big\| \big(\sum_{l\in\mathbb{Z}} f_{lk}(x) e^{ilt} \big) u^k \big\|_{\sigma,s} \\ &\le \sum_{k=0}^\infty c_s \big\| \sum_{l\in\mathbb{Z}} f_{lk}(x) e^{ilt} \big\|_{\sigma,s} \, \|u^k\|_{\sigma,s} \\ &\le C \sum_{k=0}^\infty C_k(f) \, (c_s \|u\|_{\sigma,s})^k < C \sum_{k=0}^\infty C_k(f) \, r^k < +\infty \end{aligned}$$

for $c_s\|u\|_{\sigma,s} < r$, by (F) again.

The analyticity of the composition operator f with respect to $\|\cdot\|_{\sigma,s}$ follows from the properties of the power series as explained in [25, Appendix A]. □

We emphasize that the analyticity of f as a map in $X_{\sigma,s}$ is not an assumption but follows from (F).

Remark 6. If $f(x,t,u)$ admits an analytic extension, which is 2π-periodic in x and skewsymmetric, namely,

$$f(-x,t,-u) = -f(x,t,u)\,, \tag{24}$$

then the Dirichlet problem on $[0,\pi]$ is equivalent to the 2π-periodic problem within the space of all functions odd in x. In this case also the spatial regularity is characterized by the decay properties of the Fourier coefficients. Therefore we could look for analytic solutions of (3),

$$u(x,t) = \sum_{l\in\mathbb{Z}} u_l(x)e^{ilt},$$

which are periodic and odd in x, more precisely with

$$u_l(x) \in Y := \Big\{ y(x) = \sum_{j\geq 1} y_j \sin(jx) \;:\; \sum_{j\geq 1} |y_j|^2 j^{2b} e^{2aj} < +\infty \Big\}$$

for some $a \geq 0$, $b > 1/2$. Without the oddness assumption (24) the composition operator f does not map this subspace into itself. It is for this reason that we consider the space $X_{\sigma,s}$ of functions valued in $H^1_0(0,\pi)$: also without (24), f sends $X_{\sigma,s}$ into itself (Lemma 1).

Throughout this paper we shall use spaces $X_{\sigma,s}$ with $\sigma \in [\sigma_0/2, \sigma_0]$ and $s \in \mathcal{S} := \{1,\, 1-\frac{\tau-1}{2},\, 1+\frac{(\tau-1)\tau}{2-\tau}\}$. So we choose $\bar{c} := \max_{s\in\mathcal{S}} c_s$ as a multiplicative algebra constant for all spaces $X_{\sigma,s}$.

By Lemma 1, f is analytic in the ball

$$\Big\{ u \in X_{\sigma,s} : \|u\|_{\sigma,s} < R_0 := \frac{r}{\bar{c}} \Big\}$$

and $f, f', f'', \ldots$ are bounded, uniformly in σ, s.

3. The bifurcation equation. Now we give some examples in which Hypothesis (V) holds.

LEMMA 2. *Suppose $f_0(x,u) = u^m$ for $m \geq 3$ odd and $p(x) \equiv 1$. Then, for all μ, there exists an unbounded sequence of nondegenerate solutions v_n of* (10).

Proof. All the solutions of the autonomous equation $-v'' = \mu v^m$ are periodic and can be parametrized by their energy

$$E = \frac{1}{2}\, v'^2 + \frac{\mu}{m+1}\, v^{m+1}\,.$$

Let T_E denote the period of the solution v_E. We can suppose $v_E(0) = 0$, so $v'_E(0) = \sqrt{2E}$. The other boundary condition $v_E(\pi) = 0$ is satisfied iff

$$k\frac{T_E}{2} = \pi \quad \text{for some } \; k \in \mathbb{N}\,. \tag{25}$$

By symmetry and energy conservation $v_E(T_E/4) = [(m+1)E/\mu]^{\frac{1}{m+1}}$. So

$$T_E = 4\int_0^{\left[\frac{(m+1)E}{\mu}\right]^{\frac{1}{m+1}}} \Big[2\Big(E - \frac{\mu\, x^{m+1}}{m+1}\Big)\Big]^{-1/2} dx$$

$$= \frac{4(m+1/\mu)^{\frac{1}{m+1}}}{E^{\frac{1}{2}-\frac{1}{m+1}}} \int_0^1 \frac{dy}{\sqrt{2(1-y^{m+1})}} =: \frac{C(m,\mu)}{E^{\frac{1}{2}-\frac{1}{m+1}}}$$

by the change of variable $y = x\,[E(m+1)/\mu]^{-\frac{1}{m+1}}$, and (25) is satisfied at infinitely many energy levels. Let $\bar E > 0$ such that $T_{\bar E} = 2\pi/k$ and denote the solution $\bar v := v_{\bar E}$.

Let us prove that $\bar v$ is nondegenerate. Any solution h of the linearized equation at $\bar v$,

$$-h''(x) = \mu m\, \bar v^{\,m-1}(x)\, h(x)\,, \tag{26}$$

can be written as $h = A\bar v' + B\beta$, $A, B \in \mathbb{R}$, because $\bar v'(x)$ and $\beta(x) := (\partial_E v_E)_{|E=\bar E}(x)$ are solutions of (26); they are independent because $\bar v'(0) \neq 0$ while $\beta(0) = 0$. If $h(0) = 0$, then $A = 0$. We claim that $\beta(\pi) \neq 0$; as a consequence, if $h(\pi) = 0$, then $B = 0$, and so $h = 0$, i.e., $\bar v$ is nondegenerate. To prove that $\beta(\pi) \neq 0$, we differentiate at $\bar E$ the identity $v_E(kT_E/2) = 0$,

$$\beta(\pi) + \bar v'(\pi)(\partial_E T_E)_{|E=\bar E} = 0\,.$$

Since $\bar v'(\pi) = (-1)^k\sqrt{2E} \neq 0$ and $\partial_E T_E \neq 0$, we get $\beta(\pi) \neq 0$. □

LEMMA 3. *If $f_0(x,0) = d_u f_0(x,0) = 0$, then $\bar v = 0$ is a nondegenerate solution of (10) for every μ.*

Proof. The linearized equation $-(ph')' = 0$, $h(0) = h(\pi) = 0$ has only the trivial solution. □

When Hypothesis (V) holds at some $(\mu_0, \bar v)$, we solve first the bifurcation equation in (8) using the standard implicit function theorem. We find, for every w small enough and μ in a neighborhood of μ_0, a unique solution $v(\mu, w)$ of the bifurcation equation.

LEMMA 4 (solution of the bifurcation equation). *There exist $0 < R < R_0$, a neighborhood $[\mu_1, \mu_2]$ of μ_0, and a C^∞ map*

$$[\mu_1, \mu_2] \times \big\{w \in W \cap X_{\sigma,s} \,:\, \|w\|_{\sigma,s} < R\big\} \to V, \quad (\mu, w) \mapsto v(\mu, w)$$

such that $v(\mu, w)$ solves the bifurcation equation in (8).

Proof. The linear operator

$$h \mapsto -(ph')' - \mu_0 d_v \Pi_V f(v)[h] = -(ph')' - \mu_0\, f_0'(v)\, h$$

is invertible on $H^1_0(0,\pi)$ by Hypothesis (V). Then we apply the implicit function theorem. □

Remark 7. The solutions of the 0th order bifurcation equation (10) found in Lemmas 2 and 3 are nondegenerate for every μ, so, in that case, we can continue $v(\mu, w)$ for all $[\mu_1, \mu_2] = [0, \bar\mu]$.

We denote by $\lambda_j(\mu, w) := \omega_j^2(\mu, w)$ the eigenvalues of the Sturm–Liouville problem

$$\begin{cases} -(py')' - \mu \Pi_V f'(v(\mu, w) + w)\, y = \lambda \rho y, \\ y(0) = y(\pi) = 0\,. \end{cases} \tag{27}$$

LEMMA 5. *The eigenvalues of* (27) *satisfy the continuity property*

$$|\lambda_j(\mu,w)-\lambda_j(\mu',w')|\le K\big(|\mu-\mu'|+\|w-w'\|_{\sigma,s}\big) \tag{28}$$

for some constant $K>0$ *independent of* j.

Proof. For the proof of the lemma, see the appendix. □

The nondegeneracy of $\bar v=v(\mu_0,0)$ means that $\lambda_j(\mu_0,0)\neq 0$ for all j. By (28),

$$\delta_0:=\inf\Big\{\big|\lambda_j(\mu,w)\big|:\ j\ge 1,\ \ \mu\in[\mu_1,\mu_2],\ \ \|w\|_{\sigma_0/2}\le R\Big\}>0, \tag{29}$$

taking, if necessary, $|\mu_2-\mu_1|$ and R smaller in Lemma 4.

Note also that the index j_0 of the smallest positive eigenvalue is constant, independently of (μ,w).

4. Solution of the range equation. It remains to solve the range equation

$$L_\omega w=\mu\Pi_W\mathcal{F}(\mu,w), \tag{30}$$

where

$$\mathcal{F}(\mu,w):=f(v(\mu,w)+w)\,.$$

By Lemmas 1 and 4, $\mathcal{F}$ is C^∞ and bounded, together with its derivatives, on $[\mu_1,\mu_2]\times B_R$, where $B_R:=\{w\in W\cap X_{\sigma,s}\ :\ \|w\|_{\sigma,s}<R\}$.

4.1. The Nash–Moser recursive scheme. We define the sequence of finite-dimensional subspaces

$$W^{(n)}:=\Big\{w=\sum_{1\le|l|\le N_n}w_l(x)e^{ilt}\Big\}\subset W,$$

where

$$N_n:=N_0 2^n$$

and $N_0\in\mathbb{N}$ will be fixed in Lemma 9. We also set

$$W^{(n)\perp}:=\Big\{w=\sum_{|l|>N_n}w_l(x)e^{ilt}\in W\Big\}$$

and denote by P_n, resp., $P_n^\perp$, the projection on $W^{(n)}$, resp., $W^{(n)\perp}$. Note that $P_n\circ\Pi_W=P_n$.

LEMMA 6 (smoothing estimate). *For* $w\in W^{(n)\perp}$, *if* $0<\sigma''<\sigma'$,

$$\|w\|_{\sigma'',s}\le\exp[-(\sigma'-\sigma'')N_n]\,\|w\|_{\sigma',s}\,. \tag{31}$$

Proof. It follows from the definition of the norms $\|\cdot\|_{\sigma,s}$ and $W^{(n)\perp}$; see, e.g., [16, 8]. □

The key property for the construction of the iterative sequence is the invertibility of the linear operator

$$\begin{aligned}\mathcal{L}_n(w)h&:=-L_\omega h+\mu P_n[d_w\mathcal{F}(\mu,w)h]\\&=-L_\omega h+\mu P_n\big[f'(v(\mu,w)+w)\big(h+d_wv(\mu,w)[h]\big)\big]\quad\forall\,h\in W^{(n)}\,.\end{aligned} \tag{32}$$

LEMMA 7 (inversion of the linear problem). *Let $\omega > 0$, $\tau \in (1,2)$, $\gamma \in (0,1)$, $\gamma < \omega$, and $\sigma \in (0, \sigma_0]$. Assume the "Melnikov" nonresonance conditions*

$$\left|\omega l - \frac{j}{c}\right| > \frac{\gamma}{l^\tau} \qquad \forall\, l = 1, 2, \dots, N_n, \quad \forall\, j \geq 1, \tag{33}$$

where c is defined in (16)*, and*

$$|\omega^2 l^2 - \lambda_j(\mu, w)| > \frac{\gamma\omega}{l^{\tau-1}} \qquad \forall\, l = 1, 2, \dots, N_n,\ j \geq 1, \tag{34}$$

where $\lambda_j(\mu, w)$ are the eigenvalues of (27)*.*

Let $u := v(\mu, w) + w$. There exist K_1, K_1' such that if

$$\frac{\mu}{\gamma^3\omega} \left\| \Pi_W f'(u) \right\|_{\sigma, 1+\frac{\tau(\tau-1)}{2-\tau}} < K_1' \,, \tag{35}$$

then $\mathcal{L}_n(w)$ is invertible and

$$\|\mathcal{L}_n(w)^{-1} h\|_\sigma \leq \frac{K_1 N_n^{\tau-1}}{\gamma\omega} \|h\|_\sigma \qquad \forall\, h \in W^{(n)} \,. \tag{36}$$

Proof. For the proof of the lemma, see section 6. □

Remark 8. The condition $\omega > 0$ means that (1) is nonautonomous. Indeed, if $\omega = 0$, the nonlinearity $f(x, \omega t, u) = f(x, 0, u)$ is independent of t.

For $\vartheta := 3\sigma_0/\pi^2$ we define the sequence

$$\sigma_{n+1} := \sigma_n - \frac{\vartheta}{(n+1)^2} \,, \qquad \sigma_0 > \sigma_1 > \sigma_2 > \cdots > \frac{\sigma_0}{2} \,. \tag{37}$$

Let A_0 denote the open set

$$A_0 := \Big\{ (\mu, \omega) \in (\mu_1, \mu_2) \times (\gamma, +\infty) \ :\ |\omega l - \omega_j| > \frac{\gamma}{l^\tau} \ \ \forall\, l = 1, \dots, N_0,\ j \geq 1 \Big\},$$

where ω_j are defined by (11).

LEMMA 8 (approximate solution). *There exist K_2, K_2' such that if $(\mu, \omega) \in A_0$ and $\mu N_0^{\tau-1}/\gamma\omega < K_2'$, then there exists a solution $w_0 := w_0(\mu, \omega) \in W^{(0)}$ of*

$$L_\omega w_0 = \mu P_0 \mathcal{F}(\mu, w_0)$$

satisfying $\|w_0\|_{\sigma_0} \leq \mu K_2 N_0^{\tau-1}/\gamma\omega$.

Proof. By definition of A_0, the eigenvalues of $(1/\rho)L_\omega$ satisfy

$$|\omega^2 l^2 - \lambda_j| > \frac{\gamma\omega}{l^{\tau-1}} \qquad \forall\, l = 1, 2, \dots, N_0, \quad j \geq 1 \,,$$

so L_ω is invertible on $W^{(0)}$ and, for some K,

$$\|L_\omega^{-1} h\|_{\sigma_0} \leq \frac{K N_0^{\tau-1}}{\gamma\omega} \|h\|_{\sigma_0} \qquad \forall\, h \in W^{(0)} \,. \tag{38}$$

Then we look for a solution $w_0 \in W^{(0)}$ of $w_0 = \mu L_\omega^{-1} P_0 \mathcal{F}(\mu, w_0)$. The right-hand side term is a contraction in $\{\|w_0\|_{\sigma_0} < R\}$ if $\mu N_0^{\tau-1}/\gamma\omega$ is sufficiently small. □

Given $w_n \in W^{(n)}$, $\|w_n\|_{\sigma_n} < R$, and $A_n \subseteq A_0$, we define

$$A_{n+1} := \Big\{ (\mu,\omega) \in A_n \,:\, \big|\omega l - \omega_j(\mu,w_n)\big| > \frac{\gamma}{l^\tau}\,, \quad \Big|\omega l - \frac{j}{c}\Big| > \frac{\gamma}{l^\tau} \qquad \forall\, l = 1,2,\dots,N_{n+1},\ \ j \ge 1 \Big\} \subseteq A_n, \tag{39}$$

where $\lambda_j(\mu,w_n) = \omega_j^2(\mu,w_n)$ are defined in (27) with $w = w_n$.

In Lemma 8 we have constructed $h_0 := w_0$ for $(\mu,\omega) \in A_0$. Next, we proceed by induction. By means of w_0 we define the set A_1 as above, and we find $w_1 := h_0 + h_1 \in W^{(1)}$ for every $(\mu,\omega) \in A_1$ by Lemma 9 below. Then we define A_2, we find $w_2 \in W^{(2)}$, and so on. The main goal of the construction is to prove that, at the end of the recurrence, the set of parameters $(\mu,\omega) \in \cap_n A_n$ is actually a large set (see Lemmas 12 and 13).

LEMMA 9 (inductive step). *Fix $\chi := 3/2$. There exist $N_0 \in \mathbb{N}$ (depending only on $\rho, p, f, \bar\mu, \bar v, \tau$) and $K_3' \le K_2'/N_0^{\tau-1}$ with the following property.*

Suppose that $h_i \in W^{(i)}$ for all $i = 0,\dots,n$ satisfy

$$\|h_i\|_{\sigma_i} < \frac{\mu K_3 N_0^{\tau-1}}{\gamma\omega} \exp(-\chi^i), \tag{40}$$

where $K_3 := eK_2$ and K_2 is the constant in Lemma 8; for all $k = 0,\dots,n$, let $w_k := h_0 + \cdots + h_k$ satisfy $\|w_k\|_{\sigma_k} < R$ and

$$L_\omega w_k = \mu P_k \mathcal{F}(\mu, w_k) \tag{41}$$

and suppose that $(\mu,\omega) \in A_n$, where A_{i+1} is constructed by means of w_i as shown above.

If $(\mu,\omega) \in A_{n+1}$ and $\mu/\gamma^3\omega < K_3'$, then there exists $h_{n+1} \in W^{(n+1)}$ satisfying

$$\|h_{n+1}\|_{\sigma_{n+1}} < \frac{\mu K_3 N_0^{\tau-1}}{\gamma\omega}\, \exp(-\chi^{n+1}) \tag{42}$$

such that $w_{n+1} = w_n + h_{n+1}$ verifies $\|w_{n+1}\|_{\sigma_{n+1}} < R$ and

$$L_\omega w_{n+1} = \mu P_{n+1} \mathcal{F}(\mu, w_{n+1})\,. \tag{43}$$

Proof. In short $\mathcal{F}(w) := \mathcal{F}(\mu,w)$ and $D\mathcal{F}(w) := d_w\mathcal{F}(\mu,w)$. Equation (43) for $w_{n+1} = w_n + h_{n+1}$ is $L_\omega[w_n + h_{n+1}] = \mu P_{n+1}\mathcal{F}(w_n + h_{n+1})$.

By assumption, w_n satisfies (41) for $k = n$, namely, $L_\omega w_n = \mu P_n \mathcal{F}(w_n)$, so the equation for h_{n+1} can be written as

$$\mathcal{L}_{n+1}(w_n)h_{n+1} + \mu(P_{n+1} - P_n)\mathcal{F}(w_n) + \mu P_{n+1} Q = 0, \tag{44}$$

where, as defined in (32), $\mathcal{L}_{n+1}(w_n)h_{n+1} := -L_\omega h_{n+1} + \mu P_{n+1} D\mathcal{F}(w_n)h_{n+1}$, and Q denotes the quadratic remainder

$$Q = Q(w_n, h_{n+1}) := \mathcal{F}(w_{n+1}) - \mathcal{F}(w_n) - D\mathcal{F}(w_n)h_{n+1}\,.$$

Step 1. *Inversion of $\mathcal{L}_{n+1}(w_n)$.* We verify the assumptions of Lemma 7. By definition of A_{n+1}, ω satisfies (33). If $\lambda_j(\mu,w_n) < 0$, then $|\omega^2 l^2 - \lambda_j(\mu,w_n)| \ge \omega^2 l^2 > \gamma\omega/l^{\tau-1}$ because $\omega > \gamma$. If $\lambda_j(\mu,w_n) > 0$, we have

$$|\omega^2 l^2 - \lambda_j(\mu,w_n)| \ge |\omega l - \omega_j(\mu,w_n)|\omega l > \frac{\gamma\omega}{l^{\tau-1}} \qquad \forall\, l = 1,\dots,N_{n+1}$$

because $(\mu,\omega) \in A_{n+1}$. In both cases the nonresonance condition (34) holds.

To verify (35) we need an estimate for w_n. Let $\eta := \tau(\tau-1)/(2-\tau)$ and $\alpha > 0$. Using the elementary inequality

$$\frac{1+l^{2(1+\eta)}}{1+l^2}\cdot\frac{e^{2(\sigma-\alpha)|l|}}{e^{2\sigma|l|}} \le \frac{2l^{2\eta}}{e^{2\alpha|l|}} \le 2\max_{y>0}(y^{2\eta}e^{-2\alpha y}) = 2\Big(\frac{\eta}{\alpha e}\Big)^{2\eta} \quad \forall\, l\neq 0\,,$$

we deduce

$$\|h_i\|_{\sigma_{n+1},1+\eta} \le \frac{C_\eta}{(\sigma_i-\sigma_{n+1})^\eta}\,\|h_i\|_{\sigma_i},$$

where $C_\eta := \sqrt{2}(\eta/e)^\eta$. Since $\sigma_i-\sigma_{n+1} \ge \sigma_i-\sigma_{i+1}$ for every $i \le n$,

$$\|w_n\|_{\sigma_{n+1},1+\eta} \le \sum_{i=0}^{n}\|h_i\|_{\sigma_{n+1},1+\eta} \le C_\eta\sum_{i=0}^{n}\frac{\|h_i\|_{\sigma_i}}{(\sigma_i-\sigma_{i+1})^\eta} \le S_\eta\,\frac{\mu K_3 N_0^{\tau-1}}{\gamma\omega},$$

using (40) where $S_\eta := (C_\eta/\vartheta^\eta)\sum_{i=0}^{+\infty}(i+1)^{2\eta}\exp(-\chi^i) < +\infty$. If

$$\frac{S_\eta\,\mu K_3 N_0^{\tau-1}}{\gamma\omega} < R,$$

then

$$\|f'(u_n)\|_{\sigma_{n+1},1+\eta} \le K$$

for some K, where $u_n := v(\mu,w_n)+w_n$. Hence Hypothesis (35) is verified for $\mu/\gamma^3\omega$ sufficiently small.

Analogously we get $\|w_n\|_{\sigma_n} < R$ if $\mu N_0^{\tau-1}/\gamma\omega$ is small enough.

By Lemma 7 the operator $\mathcal{L}_{n+1}(w_n)$ is invertible on $W^{(n+1)}$ and

$$\|\mathcal{L}_{n+1}(w_n)^{-1}h\|_{\sigma_{n+1}} \le \frac{K_1 N_{n+1}^{\tau-1}}{\gamma\omega}\,\|h\|_{\sigma_{n+1}} \quad \forall h\in W^{(n+1)}\,. \tag{45}$$

Equation (44) amounts to the fixed point problem

$$h_{n+1} = -\mu\mathcal{L}_{n+1}(w_n)^{-1}\big[(P_{n+1}-P_n)\mathcal{F}(w_n)+P_{n+1}Q\big] := \mathcal{G}(h_{n+1})$$

for $h_{n+1}\in W^{(n+1)}$.

Step 2. $\mathcal{G}$ *is a contraction.* We prove that $\mathcal{G}$ is a contraction on the ball $B_{n+1} := \{\|h\|_{\sigma_{n+1}} < r_{n+1}\}$, where $r_{n+1} := (\mu K_3 N_0^{\tau-1}/\gamma\omega)\exp(-\chi^{n+1})$, implying (42). By (31)

$$\|(P_{n+1}-P_n)\mathcal{F}(w_n)\|_{\sigma_{n+1}} \le \|\mathcal{F}(w_n)\|_{\sigma_n}\exp[-(\sigma_n-\sigma_{n+1})N_n]\,.$$

Since $\|w_n\|_{\sigma_n} < R$, we have $\|Q\|_{\sigma_{n+1}} \le K\|h_{n+1}\|^2_{\sigma_{n+1}}$. Hence, by (45),

$$\|\mathcal{G}(h_{n+1})\|_{\sigma_{n+1}} \le K\,\frac{\mu N_{n+1}^{\tau-1}}{\gamma\omega}\Big(\exp[-(\sigma_n-\sigma_{n+1})N_n] + \|h_{n+1}\|^2_{\sigma_{n+1}}\Big)\,.$$

Therefore $\mathcal{G}(B_{n+1}) \subseteq B_{n+1}$ if

$$\frac{\mu K N_{n+1}^{\tau-1}}{\gamma\omega}\,\exp[-(\sigma_n-\sigma_{n+1})N_n] < \frac{r_{n+1}}{2}\,, \qquad \frac{\mu K N_{n+1}^{\tau-1}}{\gamma\omega}\,r_{n+1}^2 < \frac{r_{n+1}}{2}\,. \tag{46}$$

By the definition of σ_n in (37) and $N_n := N_0 2^n$, the first inequality is verified for every $n \geq 0$ if $\sigma_0 N_0$ is greater than a constant depending only on χ, K, K_3. The second inequality is verified for every $n \geq 0$ if $\mu N_0^{\tau-1}/\gamma\omega$ is small enough.

The estimate for $\|\mathcal{G}h - \mathcal{G}k\|$, $h, k \in B_{n+1}$ is similar. The lemma now follows from the contraction mapping theorem. □

Remark 9. In the previous scheme h_{n+1} is found as an exact solution of (44). We find this convenient to prove the regularity of h_{n+1} with respect to the parameters (μ, ω) in Lemma 10. However, other schemes are possible. For example, we could define h_{n+1} as a solution of the linearized equation $\mathcal{L}_{n+1}(w_n)h + \mu(P_{n+1} - P_n)\mathcal{F}(w_n) = 0$.

COROLLARY 1 (existence). *Suppose* $A_\infty := \cap_{n\geq 0} A_n \neq \emptyset$. *If* $(\mu, \omega) \in A_\infty$ *and* $\mu/\gamma^3\omega < K_3'$, *then*

$$w_\infty(\mu, \omega) := \sum_{n\geq 0} h_n(\mu, \omega) \in W \cap X_{\sigma_0/2}$$

is a solution of the range equation (30) *satisfying* $\|w_\infty\|_{\sigma_0/2} \leq K_\infty \mu/\gamma\omega$ *for some* K_∞.

Proof. Since w_n solves (41) for $k = n$,

$$-L_\omega w_n + \mu \Pi_W f(u_n) = \mu P_n^\perp f(u_n) \in W^{(n)\perp},$$

where $u_n := v(\mu, w_n) + w_n$. By (31)

$$\lim_{n\to+\infty} \| - L_\omega w_n + \mu f(u_n)\|_{\sigma_0/2} \leq \lim_{n\to+\infty} K \exp[-(\sigma_n - \sigma_0/2)N_n] = 0\,.$$

Since $w_n \to w_\infty$ in $\|\cdot\|_{\sigma_0/2}$ also $f(u_n) \to f(u_\infty)$ in the same norm, while $L_\omega w_n \to L_\omega w_\infty$ in the sense of distributions. So w_∞ is a weak solution of the range equation (30). □

Remark 10. We shall prove, as a consequence of Lemma 12 and section 4.3, that A_∞ is actually a positive measure set. One possible way to prove it uses the Whitney extension of w_∞ of section 4.2.

4.2. Whitney C^∞ extension. The functions h_n constructed in Lemmas 8 and 9 depend smoothly on the parameters (μ, ω).

LEMMA 10. *There exist* K_4 *and* $K_4' \leq K_3'$ *such that the maps*

$$h_i : A_i \cap \{\mu/\gamma^3\omega < K_4'\} \to W^{(i)}$$

are C^∞ *and*

$$\|\partial_\omega h_i(\mu, \omega)\|_{\sigma_i} \leq \frac{K_4 \mu}{\gamma^2 \omega} \exp(-\chi_0^i), \quad \|\partial_\mu h_i(\mu, \omega)\|_{\sigma_i} \leq \frac{K_4}{\gamma\omega} \exp(-\chi_0^i),$$

where $\chi_0 := (1 + \chi)/2 = 5/4$.

Proof. Since $w_0 = \mu L_\omega^{-1} P_0 \mathcal{F}(\mu, w_0)$, by the implicit function theorem the map w_0 is C^∞. Differentiating the identity $L_\omega(L_\omega^{-1}h) = h$ with respect to ω, by (38) we get $\|\partial_\omega L_\omega^{-1} h\|_{\sigma_0} \leq (K/\gamma^2\omega)\,\|h\|_{\sigma_0}$. For $\mu/\gamma\omega$ small,

$$\|\partial_\omega w_0\|_{\sigma_0} \leq \frac{K\mu}{\gamma^2\omega}\,.$$

Differentiating with respect to μ we get $\|\partial_\mu w_0\|_{\sigma_0} \leq K'/\gamma\omega$ for some K'.

By induction, suppose that h_i depends smoothly on $(\mu,\omega)\in A_i$ for every $i=0,\dots,n$. For $(\mu,\omega)\in A_{n+1}$, by (43), h_{n+1} is a solution of

$$-L_\omega h_{n+1}+\mu P_{n+1}[\mathcal{F}(w_n+h_{n+1})-\mathcal{F}(w_n)]+\mu(P_{n+1}-P_n)\mathcal{F}(w_n)=0\,. \tag{47}$$

By the implicit function theorem $h_{n+1}\in C^\infty$ once we prove that

$$\mathcal{L}_{n+1}(w_{n+1})[z]:=-L_\omega z+\mu P_{n+1}D\mathcal{F}(w_n+h_{n+1})[z]$$

is invertible. By (45), $\mathcal{L}_{n+1}(w_n)$ is invertible. Hence it is sufficient that

$$\Big\|\mathcal{L}_{n+1}^{-1}(w_n)(\mathcal{L}_{n+1}(w_{n+1})-\mathcal{L}_{n+1}(w_n))\Big\|_{\sigma_{n+1}}<\frac{1}{2}\,,$$

which holds true for $\mu^2/\gamma\omega$ small enough; indeed, by (42),

$$\|\mathcal{L}_{n+1}(w_{n+1})-\mathcal{L}_{n+1}(w_n)\|_{\sigma_{n+1}}\le K\mu\|h_{n+1}\|_{\sigma_{n+1}}\le\frac{\mu^2K'N_0^{\tau-1}}{\gamma\omega}\,\exp(-\chi^{n+1})\,.$$

Finally (45) implies

$$\|\mathcal{L}_{n+1}(w_{n+1})^{-1}\|_{\sigma_{n+1}}\le\frac{2K_1N_{n+1}^{\tau-1}}{\gamma\omega}\,. \tag{48}$$

Differentiating (47) with respect to ω

$$\begin{aligned}\mathcal{L}_{n+1}(w_{n+1})[\partial_\omega h_{n+1}]&=2\omega\rho(x)(h_{n+1})_{tt}-\mu(P_{n+1}-P_n)D\mathcal{F}(w_n)\partial_\omega w_n\\&\quad-\mu P_{n+1}\big[D\mathcal{F}(w_n+h_{n+1})-D\mathcal{F}(w_n)\big]\partial_\omega w_n\end{aligned} \tag{49}$$

and, using (48) and (31),

$$\begin{aligned}\|\partial_\omega h_{n+1}\|_{\sigma_{n+1}}\le\frac{KN_{n+1}^{\tau-1}}{\gamma\omega}\Big(&\omega N_{n+1}^2\|h_{n+1}\|_{\sigma_{n+1}}+\frac{\mu\|\partial_\omega w_n\|_{\sigma_n}}{\exp[(\sigma_n-\sigma_{n+1})N_n]}\\&+\mu\|h_{n+1}\|_{\sigma_{n+1}}\|\partial_\omega w_n\|_{\sigma_n}\Big)\,.\end{aligned}$$

We note that $\|\partial_\omega w_n\|_{\sigma_n}\le\sum_{i=0}^n\|\partial_\omega h_i\|_{\sigma_i}$. Using (46) the sequence $a_n:=\|\partial_\omega h_n\|_{\sigma_n}$ satisfies

$$\begin{aligned}a_{n+1}&\le\frac{KN_{n+1}^{\tau-1}}{\gamma\omega}\Big(\omega N_{n+1}^2r_{n+1}+\frac{\omega\gamma r_{n+1}}{N_{n+1}^{\tau-1}}\sum_{i=0}^n a_i+\mu r_{n+1}\sum_{i=0}^n a_i\Big)\\&\le b_{n+1}\Big(1+\sum_{i=0}^n a_i\Big),\quad\text{where }\ b_{n+1}:=\frac{K\mu}{\gamma^2\omega}\,N_{n+1}^{\tau+1}\,\exp(-\chi^{n+1}),\end{aligned}$$

recalling that $r_{n+1}=(\mu K/\gamma\omega)\exp(-\chi^{n+1})$. By induction, for $K\mu/\omega\gamma^2<1$, we have $a_n\le 2b_n$ and

$$\|\partial_\omega h_{n+1}\|_{\sigma_{n+1}}\le\frac{K\mu}{\gamma^2\omega}\,N_{n+1}^{\tau+1}\exp(-\chi^{n+1})\le\frac{K'\mu}{\gamma^2\omega}\,\exp(-\chi_0^{n+1}),$$

where $\chi_0:=(1+\chi)/2$. It follows that $\|\partial_\omega w_{n+1}\|_{\sigma_{n+1}}\le K\mu/\gamma^2\omega$.

Differentiating (47) with respect to μ we obtain the estimate for $\partial_\mu h_{n+1}$. □

Define, for $\nu_0 > 0$,

$$A_n^* := \Big\{(\mu,\omega) \in A_n \,:\, \mathrm{dist}\big((\mu,\omega),\, \partial A_n\big) > \frac{\nu_0\gamma^4}{N_n^3}\Big\}, \tag{50}$$

$$\tilde{A}_n := \Big\{(\mu,\omega) \in A_n \,:\, \mathrm{dist}\big((\mu,\omega),\, \partial A_n\big) > \frac{2\nu_0\gamma^4}{N_n^3}\Big\} \subset A_n^* .$$

LEMMA 11 (Whitney extension). *There exists a C^∞ map*

$$\tilde{w} : A_0 \cap \Big\{(\mu,\omega) : \frac{\mu}{\gamma^3\omega} < K_4^{'}\Big\} \to W \cap X_{\sigma_0/2}$$

satisfying

$$\|\tilde{w}(\mu,\omega)\|_{\sigma_0/2} \le \frac{K_5\mu}{\gamma\omega}, \tag{51}$$

$$\|\partial_\omega\tilde{w}(\mu,\omega)\|_{\sigma_0/2} \le \frac{C(\nu_0)\mu}{\gamma^5\omega}, \quad \|\partial_\mu\tilde{w}(\mu,\omega)\|_{\sigma_0/2} \le \frac{C(\nu_0)}{\gamma^5\omega}$$

for some K_5 and for some $C(\nu_0) > 0$, such that, for $(\mu,\omega) \in \tilde{A}_\infty := \cap_{n\ge 0}\tilde{A}_n$, $\tilde{w}(\mu,\omega)$ solves the range equation (30).

Moreover there exists a sequence of C^∞ maps

$$\tilde{w}_n : A_0 \cap \Big\{(\mu,\omega) : \frac{\mu}{\gamma^3\omega} < K_4^{'}\Big\} \to W^{(n)}$$

such that $\tilde{w}_n(\mu,\omega) = w_n(\mu,\omega)$ for $(\mu,\omega) \in \tilde{A}_n$, and

$$\|\tilde{w}(\mu,\omega) - \tilde{w}_n(\mu,\omega)\|_{\sigma_0/2} \le \frac{K_5\mu}{\gamma\omega}\,\exp(-\chi^n). \tag{52}$$

Proof. Let $\varphi : \mathbb{R}^2 \to \mathbb{R}^+$ be a C^∞-function supported in the open ball $B(0,1)$ of center 0 and radius 1 and with $\int_{\mathbb{R}^2}\varphi = 1$. Let $\varphi_n : \mathbb{R}^2 \to \mathbb{R}^+$ be the mollifier

$$\varphi_n(x) := \frac{N_n^6}{\nu_0^2\gamma^8}\,\varphi\Big(\frac{N_n^3}{\nu_0\gamma^4}\,x\Big).$$

$\mathrm{Supp}\,(\varphi_n) \subset B(0,\nu_0\gamma^4/N_n^3)$ and $\int_{\mathbb{R}^2}\varphi_n = 1$. We define $\psi_n : \mathbb{R}^2 \to \mathbb{R}$ as

$$\psi_n(x) := \big(\varphi_n * \chi_{A_n^*}\big)(x) = \int_{\mathbb{R}^2}\varphi_n(y-x)\chi_{A_n^*}(y)\,dy,$$

where $\chi_{A_n^*}$ is the characteristic function of the set A_n^*. ψ_n is C^∞,

$$|D\psi_n(x)| \le \int_{\mathbb{R}^2}|D\varphi_n(x-y)|\,\chi_{A_n^*}(y)\,dy \le \frac{N_n^3}{\nu_0\gamma^4}\,C, \tag{53}$$

where $C := \int_{\mathbb{R}^2}|D\varphi|\,dy$,

$$0 \le \psi_n(x) \le 1, \qquad \mathrm{supp}(\psi_n) \subset A_n, \qquad \psi_n(x) = 1 \;\; \forall\, x \in \tilde{A}_n .$$

We define, for $(\mu,\omega)\in A_0$, the C^∞-functions

$$\tilde h_n(\mu,\omega):=\begin{cases}\psi_n(\mu,\omega)h_n(\mu,\omega) & \text{if } (\mu,\omega)\in A_n,\\ 0 & \text{if } (\mu,\omega)\notin A_n,\end{cases}$$

and

$$\tilde w_n(\mu,\omega):=\sum_{i=0}^{n}\tilde h_i\,,\qquad \tilde w(\mu,\omega):=\sum_{i\ge 0}\tilde h_i,$$

which is a series if $(\mu,\omega)\in A_\infty:=\cap_{n\ge0}A_n$.

The estimate for $\|\tilde w\|_{\sigma_0/2}$ follows by $\|\tilde h_i\|_{\sigma_i}\le\|h_i\|_{\sigma_i}$ (because $0\le\psi_i\le1$) and (40). The estimates for the derivatives in (51) follow by differentiating the product $\tilde h_i=\psi_i h_i$ and using (53), (40), and Lemma 10. Similarly it follows that $\tilde w$ is in C^∞; see [8] for details.

For $(\mu,\omega)\in\tilde A_n$, $\psi_n(\mu,\omega)=1$, implying $\tilde w_n=w_n$. As a consequence, for $(\mu,\omega)\in\tilde A_\infty:=\cap_{n\ge0}\tilde A_n$, by Corollary 1, $\tilde w=w_\infty$ solves (30).

Finally, using (40),

$$\|\tilde w-\tilde w_n\|_{\sigma_0/2}\le\sum_{i\ge n+1}\|\tilde h_i\|_{\sigma_i}\le\sum_{i\ge n+1}\frac{K\mu}{\gamma\omega}\exp(-\chi^i)\le\frac{K'\mu}{\gamma\omega}\exp(-\chi^n)\,.\qquad\square$$

In the next lemma we fix the constant ν_0 introduced in (50).

LEMMA 12. *There exist $\nu_0>0$ and $K_5'\le K_4'$ such that if $\mu/\gamma^3\omega<K_5'$, then*

$$B_\gamma\subseteq\tilde A_n\subset A_n\quad\forall\, n\ge0,$$

where B_γ is defined in (15) *taking $C'\le K_5'$.*

Proof. The proof is by induction. Let $(\mu,\omega)\in B_\gamma$. Then $(\mu,\omega)\in\tilde A_0$ if A_0 contains the closed ball of center (μ,ω) and radius $2\nu_0\gamma^4/N_0^3$. Let (ω',μ') belong to such a ball. Then, for all $l=1,\dots,N_0$,

$$|\omega' l-\omega_j|\ge|\omega l-\omega_j|-|\omega-\omega'|l>\frac{2\gamma}{l^\tau}-\frac{2\nu_0\gamma^4}{N_0^3}\,l\ge\frac{\gamma}{l^\tau}$$

if $\nu_0\le1/2$.

Suppose now that $B_\gamma\subseteq\tilde A_n$ and let $(\mu,\omega)\in B_\gamma$. To prove that $(\mu,\omega)\in\tilde A_{n+1}$, we have to show that the closed ball of center (μ,ω) and radius $2\nu_0\gamma^4/N_{n+1}^3$ is contained in A_{n+1}. Let (μ',ω') belong to such a ball. The nonresonance condition on $|\omega' l-j/c|$ is verified, as above, for $\nu_0\le1/2$. For the other condition, we denote in short $\omega_j^n(\mu',\omega'):=\omega_j(\mu',w_n(\mu',\omega'))$ (see (27) for the definition of $\omega_j(\mu,w)$). It results, for all $l=1,\dots,N_{n+1}$, in

$$\begin{aligned}|\omega' l-\omega_j^n(\mu',\omega')|&\ge|\omega l-\tilde\omega_j(\mu,\omega)|-|\omega-\omega'|l-|\omega_j^n(\mu',\omega')-\tilde\omega_j(\mu,\omega)|\\&>\frac{2\gamma}{l^\tau}-\frac{2\nu_0\gamma^4 l}{N_{n+1}^3}-|\omega_j^n(\mu',\omega')-\tilde\omega_j(\mu,\omega)|\\&>\frac{3\gamma}{2l^\tau}-|\omega_j^n(\mu',\omega')-\tilde\omega_j(\mu,\omega)|\end{aligned}\tag{54}$$

if $\nu_0 \le 1/4$. Now we estimate the last term

$$|\omega_j^n(\mu',\omega') - \tilde\omega_j(\mu,\omega)| = \frac{|\lambda_j^n(\mu',\omega') - \tilde\lambda_j(\mu,\omega)|}{|\tilde\omega_j(\mu,\omega)| + |\omega_j^n(\mu',\omega')|} \le \frac{|\lambda_j^n(\mu',\omega') - \tilde\lambda_j(\mu,\omega)|}{\sqrt{\delta_0}}$$

by (29), both for $j < j_0$ and for $j \ge j_0$. By the comparison principle (28)

$$\delta_0^{-1/2}|\lambda_j^n(\mu',\omega') - \tilde\lambda_j(\mu,\omega)| \le K|\mu-\mu'| + K\|w_n(\mu',\omega') - \tilde w(\mu,\omega)\|_{\sigma_0/2}\,.$$

By Lemma 10, $\|\partial_\omega w_n\|_{\sigma_0/2}$, $\|\partial_\mu w_n\|_{\sigma_0/2} \le K/\gamma^2\omega$ for some other K, and being $\omega,\omega' > \gamma$,

$$K\|w_n(\mu',\omega') - w_n(\mu,\omega)\|_{\sigma_0/2} \le \frac{K'}{\gamma^3}\,\frac{\nu_0\gamma^4}{N_{n+1}^3} < \frac{\gamma}{8l^\tau} \quad \forall l = 1,\dots,N_{n+1}$$

if ν_0 is small enough ($1 < \tau < 2$). On the other hand, since $(\mu,\omega) \in \tilde A_n$ we have $w_n(\mu,\omega) = \tilde w_n(\mu,\omega)$ (Lemma 11) and, by (52),

$$K\|w_n(\mu,\omega) - \tilde w(\mu,\omega)\|_{\sigma_0/2} \le \frac{K'\mu}{\gamma\omega}\,\exp(-\chi^n) < \frac{\gamma}{8l^\tau} \quad \forall l = 1,\dots,N_{n+1}$$

for $\mu/\gamma^2\omega$ sufficiently small. By (54), collecting the previous estimates,

$$|\omega' l - \omega_j^n(\mu',\omega')| > \frac{\gamma}{l^\tau} \quad \forall l = 1,\dots,N_{n+1}$$

and (μ',ω') belongs to A_{n+1}. □

4.3. Measure of the Cantor set B_γ. In the following $R := (\mu',\mu'') \times (\omega',\omega'')$ denotes a rectangle contained in the region $\{(\mu,\omega) \in [\mu_1,\mu_2] \times (2\gamma,+\infty) : \mu < K_6'\gamma^5\omega\}$. Furthermore we consider $\omega'' - \omega'$ as a fixed quantity ("of order 1").

LEMMA 13. *There exist K_6 and $K_6' \le K_5'$ such that, taking $C' \le K_6'$ in the definition* (15) *of B_γ, for all $\mu \in (\mu_1,\mu_2)$ the section*

$$S_\gamma(\mu) := \{\omega : (\mu,\omega) \in B_\gamma\}$$

satisfies the measure estimate

$$|S_\gamma(\mu) \cap (\omega',\omega'')| \ge (1 - K_6\gamma)(\omega'' - \omega')\,. \tag{55}$$

As a consequence, for every $R := (\mu',\mu'') \times (\omega',\omega'')$

$$|B_\gamma \cap R| \ge |R|\,(1 - K_6\gamma)\,. \tag{56}$$

Proof. We consider the inequalities $|\omega l - \tilde\omega_j(\mu,\omega)| > 2\gamma/l^\tau$ in the definition of B_γ. The analogous inequalities for $|\omega l - \omega_j|$ and $|\omega l - j/c|$ are simpler because j/c and ω_j do not depend on (μ,ω).

The complementary set we have to estimate is

$$\mathcal{C} := \bigcup_{l,j\ge 1} \mathcal{R}_{lj},$$

where $\mathcal{R}_{lj} := \{\omega \in (\omega',\omega'') : |l\omega - \tilde\omega_j(\mu,\omega)| \le 2\gamma/l^\tau\}$.

We claim that

$$\left|\partial_\omega \tilde\omega_j(\mu,\omega)\right| \le \frac{K\mu}{\gamma^5\omega}\,. \tag{57}$$

Indeed, by the same arguments as in the proof of Lemma 12 and the comparison principle (28), we have

$$|\tilde\omega_j(\mu,\omega)-\tilde\omega_j(\mu,\omega')| \le K\|\tilde w(\mu,\omega)-\tilde w(\mu,\omega')\|_{\sigma_0/2} \le \frac{K\mu}{\gamma^5\omega}\,|\omega-\omega'|$$

using (51). As a consequence of (57)

$$\partial_\omega\big(l\omega-\tilde\omega_j(\mu,\omega)\big) \ge l-\frac{K\mu}{\gamma^5\omega} \ge \frac{l}{2} \qquad \forall l\ge 1$$

for $\mu/\gamma^5\omega$ small enough; we deduce $|\mathcal{R}_{lj}| \le 4\gamma/l^{\tau+1}$.

Furthermore the set $\mathcal{R}_{lj}$ is nonempty only if

$$\omega' l-\frac{2\gamma}{l^\tau} < \tilde\omega_j(\mu,\omega) < \omega'' l+\frac{2\gamma}{l^\tau}\,.$$

So, for every fixed l, the number of indices j such that $\mathcal{R}_{lj}\neq\emptyset$ is

$$\sharp\{j\} \le \frac{1}{\delta}\left(l(\omega''-\omega')+\frac{4\gamma}{l^\tau}\right)+1 \le Kl(\omega''-\omega'),$$

where

$$\delta := \inf\Big\{\,|\tilde\omega_{j+1}(\mu,\omega)-\tilde\omega_j(\mu,\omega)|\,:\,j\ge 1,\ (\mu,\omega)\in B_\gamma\Big\}\,.$$

For $\|\tilde w\|_{\sigma_0/2} \le K'\mu/\gamma\omega < R$ we have $\delta\ge\delta_1$, where

$$\delta_1 := \inf\Big\{\big|\omega_{j+1}(\mu,w)-\omega_j(\mu,w)\big|\,:\,j\ge 1\,,\ \mu\in[\mu_1,\mu_2],\ \|w\|_{\sigma_0/2}\le R\Big\} > 0, \tag{58}$$

as proved in the appendix.

In conclusion, the measure of the complementary set is

$$|\mathcal{C}| \le \sum_{l=1}^{+\infty}\frac{4\gamma}{l^{\tau+1}}\,K\,l(\omega''-\omega') \le K'(\omega''-\omega')\gamma$$

and (55) is proved. Integrating on (μ',μ'') we obtain (56). □

By Fubini's theorem also the section $S_\gamma(\omega)$ is large for ω in a large set.

LEMMA 14. *Let*

$$S_\gamma(\omega) := \{\mu\,:\,(\mu,\omega)\in B_\gamma\}\,.$$

For every $R := (\mu',\mu'')\times(\omega',\omega'')$, $\gamma'\in(0,1)$ *we obtain*

$$\left|\left\{\omega\in(\omega',\omega'')\,:\,\frac{|S_\gamma(\omega)\cap(\mu',\mu'')|}{\mu''-\mu'}\ge 1-\gamma'\right\}\right| \ge (\omega''-\omega')\Big(1-K_6\frac{\gamma}{\gamma'}\Big)\,. \tag{59}$$

Proof. Consider

$$\Omega^+ := \big\{\omega\in(\omega',\omega'')\,:\,|S_\gamma(\omega)\cap(\mu',\mu'')|\ge(\mu''-\mu')(1-\gamma')\big\}\,,$$

$$\Omega^- := \big\{\omega\in(\omega',\omega'')\,:\,|S_\gamma(\omega)\cap(\mu',\mu'')|<(\mu''-\mu')(1-\gamma')\big\}\,.$$

Using Fubini's theorem

$$|B_\gamma \cap R| = \int_{\omega'}^{\omega''} |S_\gamma(\omega) \cap (\mu', \mu'')|\, d\omega$$

$$= \int_{\Omega^+} |S_\gamma(\omega) \cap (\mu', \mu'')|\, d\omega + \int_{\Omega^-} |S_\gamma(\omega) \cap (\mu', \mu'')|\, d\omega$$

$$\leq (\mu'' - \mu')|\Omega^+| + (\mu'' - \mu')(1 - \gamma')|\Omega^-|\,. \tag{60}$$

By (56), $|B_\gamma \cap R| \geq (\omega'' - \omega')(\mu'' - \mu')(1 - K_6\gamma)$ and therefore, by (60),

$$(\omega'' - \omega')(1 - K_6\gamma) \leq |\Omega^+| + (1 - \gamma')|\Omega^-| = (\omega'' - \omega') - \gamma'|\Omega^-| \tag{61}$$

because $|\Omega^+| + |\Omega^-| = \omega'' - \omega'$. Then

$$|\Omega^-| \leq (\omega'' - \omega')K_6 \frac{\gamma}{\gamma'}$$

and, by the first inequality in (61), $|\Omega^+| \geq (\omega'' - \omega')(1 - K_6\gamma/\gamma')$, which is (59). ∎

Inequalities (55) and (59) imply the measure estimates (18)–(19).

The main conclusions of this section are summarized in the following theorem, which follows by Lemmas 11, 12, and 13.

THEOREM 3 (solution of the range equation). *There exist* $\tilde{w} \in C^\infty(A_\gamma,\ W \cap X_{\sigma_0/2})$ *satisfying* (51) *and the large (see* (56)*) Cantor set* B_γ *defined in* (15) *such that, for every* $(\mu, \omega) \in B_\gamma$*, the function* $\tilde{w}(\mu, \omega)$ *solves the range equation* (30)*.*

5. Proof of Theorems 1 and 2.

Proof of Theorem 1. By Theorem 3 for all $(\mu, \omega) \in B_\gamma$ the function $\tilde{w}(\mu, \omega) \in X_{\sigma_0/2}$ solves the range equation (30). By Lemma 4, $v(\mu, \tilde{w}(\mu, \omega))$ solves the bifurcation equation in (8), and therefore

$$\tilde{u} := v(\mu, \tilde{w}(\mu, \omega)) + \tilde{w}(\mu, \omega) \in X_{\sigma_0/2}$$

is a solution of (3). Estimates (14) follow by (51).

Since $\tilde{u}$ solves

$$-(p(x)\tilde{u}_x)_x = \mu f(x, t, \tilde{u}) - \omega^2 \rho(x)\tilde{u}_{tt} \tag{62}$$

we deduce

$$-(p(x)\tilde{u}_x(t, x))_x \in H^1(0, \pi) \qquad \forall t \in \mathbb{R}\,.$$

This implies, since $p(x) \in H^3(0, \pi)$, that

$$\tilde{u}(t, x) \in H^3(0, \pi) \cap H_0^1(0, \pi) \subset C^2(0, \pi) \quad \forall t \in \mathbb{R}\,. \qquad ∎$$

Proof of Theorem 2. For every fixed t, by the algebra property of H^m

$$\big\|f(x, t, u(x, t))\big\|_{H^m} \leq \sum_{l,k} \|f_{lk}(x)u^k(x)\|_{H^m} \leq K \sum_{l,k} \|f_{lk}\|_{H^m}\, \|u^k\|_{H^m}$$

for some $K > 0$.

Using the Gagliardo–Nirenberg-type inequality

$$\|u^k\|_{H^m} \le (C_m\|u\|_{H^1})^{k-1}\|u\|_{H^m}$$

valid for every $u \in H^1_0 \cap H^m$ (see, e.g., [26, 20]), we get

$$\big\|f(x,t,u(x,t))\big\|_{H^m} \le K\|u\|_{H^m}\sum_{l,k}\|f_{lk}\|_{H^m}\big(C_m\|u\|_{H^1}\big)^{k-1}, \tag{63}$$

which is convergent for $\|u\|_{H^1} < r_m/C_m$ by (21).

The solution $\tilde u$ satisfies (62) and $\tilde u(\cdot\,,t) \in H^3(0,\pi)$ for all t.

By assumption $\|\tilde u\|_{H^1} < r_m/C_m$. By induction, assume $\tilde u(\cdot\,,t) \in H^k$ for $k = 3,\dots,m$. Hence $\tilde u_{tt}(\cdot\,,t) \in H^k$ and $\rho(x)\tilde u_{tt}(\cdot\,,t) \in H^k$ because $\rho \in H^m$. Furthermore, by (63), $f(x,t,\tilde u) \in H^k$. We deduce, by (62), that $p(x)\tilde u_x \in H^{k+1}$. Finally $\tilde u \in H^{k+2}$ because $p \in H^{m+1}$.

If $f_0(x,0) = d_u f_0(x,0) = 0$, then, by Lemma 3, we can take $v(\mu,0) = 0$ for all μ. Therefore, by (14),

$$\|\tilde u(t,\cdot)\|_{H^1} \le \|\tilde u\|_{\sigma_0/2} \le \frac{2C\mu}{\gamma\omega} \quad \forall t\,,$$

and, for $\mu/\gamma\omega$ small enough, we deduce the regularity in (22). □

6. Inversion of the linearized problem. Here we prove Lemma 7. Decomposing in Fourier series

$$f'(u) = \sum_{k\in\mathbb{Z}} a_k(x)e^{ikt}$$

we write, for all $h = \sum_{1\le|l|\le N_n} h_l(x)e^{ilt} \in W^{(n)}$,

$$\begin{aligned}
-L_\omega h + \mu P_n[f'(u)h] &= \sum_{1\le|l|\le N_n}\big[\omega^2 l^2\rho h_l + \partial_x\big(p\,\partial_x h_l\big)\big]e^{ilt} \\
&\quad + \mu P_n\Big[\Big(\sum_{k\in\mathbb{Z}} a_k e^{ikt}\Big)\Big(\sum_{1\le|l|\le N_n} h_l e^{ilt}\Big)\Big] \\
&= \sum_{1\le|l|\le N_n}\big[\omega^2 l^2\rho h_l + \partial_x\big(p\,\partial_x h_l\big) + \mu a_0 h_l\big]e^{ilt} \\
&\quad + \mu \sum_{|l|,|k+l|\in\{1,\dots,N_n\},\,k\neq 0} a_k\,h_l\,e^{i(k+l)t}\,.
\end{aligned}$$

Hence $\mathcal{L}_n(w)$ defined in (32) can be decomposed as

$$\mathcal{L}_n(w)h = \rho\,\big(Dh + M_1 h + M_2 h\big), \tag{64}$$

where

$$\begin{aligned}
Dh &:= \frac{1}{\rho}\sum_{|l|=1}^{N_n}\big[\omega^2 l^2\rho\,h_l + \big(p\,h_l'\big)' + \mu a_0\,h_l\big]e^{ilt}, \\
M_1 h &:= \frac{\mu}{\rho}\sum_{|l|,|k|\in\{1,\dots,N_n\},\,l\neq k} a_{k-l}\,h_l\,e^{ikt}, \qquad (65)\\
M_2 h &:= \frac{\mu}{\rho}\,P_n\big[f'(u)\,d_w v(\mu,w)[h]\big]\,.
\end{aligned}$$

Note that D is a diagonal operator in time Fourier basis. To study the eigenvalues of D, we use Sturm–Liouville-type techniques.

LEMMA 15 (Sturm–Liouville). *The eigenvalues $\lambda_j(\mu,w)$ of the Sturm–Liouville problem* (27) *form a strictly increasing sequence which tends to $+\infty$. Every $\lambda_j(\mu,w)$ is simple and the following asymptotic formula holds:*

$$\lambda_j(\mu,w)=\frac{j^2}{c^2}+b+M(\mu,w)+r_j(\mu,w),\qquad |r_j(\mu,w)|\le\frac{K}{j} \tag{66}$$

for all $j\ge 1$, $(\mu,w)\in[\mu_1,\mu_2]\times B_R$, where

$$c:=\frac{1}{\pi}\int_0^\pi\Big(\frac{\rho}{p}\Big)^{1/2}dx,\qquad b:=\frac{1}{4\pi c}\int_0^\pi\left[\frac{(\rho p)'}{\rho\sqrt[4]{\rho p}}\right]'\frac{1}{\sqrt[4]{\rho p}}\,dx,$$

$$M(\mu,w):=-\frac{\mu}{c\pi}\int_0^\pi\frac{\Pi_V f'(v(\mu,w)+w)}{\sqrt{\rho p}}\,dx\,.$$

The eigenfunctions $\varphi_j(\mu,w)$ of (27) *form an orthonormal basis of $L^2(0,\pi)$ with respect to the scalar product $(y,z)_{L^2_\rho}:=c^{-1}\int_0^\pi yz\rho\,dx$. For K big enough*

$$(y,z)_{\mu,w}:=\frac{1}{c}\int_0^\pi p\,y'z'+\big[K\rho-\mu\Pi_V f'(v(\mu,w)+w)\big]\,yz\,dx$$

defines an equivalent scalar product on $H^1_0(0,\pi)$ and

$$K'\|y\|_{H^1}\le\|y\|_{\mu,w}\le K''\,\|y\|_{H^1}\quad\forall y\in H^1_0\,. \tag{67}$$

$\varphi_j(\mu,w)$ is also an orthogonal basis of $H^1_0(0,\pi)$ with respect to the scalar product $(\cdot,\cdot)_{\mu,w}$ and, for $y=\sum_{j\ge1}\hat y_j\,\varphi_j(\mu,w)$,

$$\|y\|^2_{L^2_\rho}=\sum_{j\ge1}\hat y_j^2\,,\qquad \|y\|^2_{\mu,w}=\sum_{j\ge1}\hat y_j^2\big(\lambda_j(\mu,w)+K\big)\,. \tag{68}$$

Proof. For the proof of the lemma, see the appendix. □

We develop

$$Dh=\sum_{1\le|l|\le N_n}D_l h_l e^{ilt},$$

where

$$D_l z:=\frac{1}{\rho}\left[\omega^2l^2\rho z+\big(p\,z'\big)'+\mu a_0 z\right]\quad\forall\,z\in H^1_0(0,\pi)$$

and $a_0=\Pi_V f(v(\mu,w)+w)$.

By Lemma 15 each D_l is diagonal with respect to the basis $\varphi_j(\mu,w)$:

$$z=\sum_{j=1}^{+\infty}\hat z_j\,\varphi_j(\mu,w)\in H^1_0(0,\pi)\quad\Rightarrow\quad D_l z=\sum_{j=1}^{+\infty}\big(\omega^2l^2-\lambda_j(\mu,w)\big)\hat z_j\,\varphi_j(\mu,w)\,.$$

LEMMA 16. *Suppose all the eigenvalues $\omega^2l^2-\lambda_j(\mu,w)$ are not zero. Then*

$$|D_l|^{-1/2}z:=\sum_{j=1}^{+\infty}\frac{\hat z_j\,\varphi_j(\mu,w)}{\sqrt{|\omega^2l^2-\lambda_j(\mu,w)|}}$$

satisfies

$$\big\| |D_l|^{-1/2} z \big\|_{H^1} \leq \frac{K}{\sqrt{\alpha_l}} \|z\|_{H^1} \quad \forall z \in H^1_0(0,\pi), \tag{69}$$

where $\alpha_l := \min_{j\geq 1} |\omega^2 l^2 - \lambda_j(\mu, w)| > 0$.

Proof. By (68) $\| |D_l|^{-1/2} z\|^2_{\mu,w} \leq (1/\alpha_l) \|z\|^2_{\mu,w}$. Hence (69) follows by the equivalence of the norms (67). □

LEMMA 17 (inversion of D). *Assume the nonresonance condition* (34). *Then* $|D|^{-1/2} : W^{(n)} \to W^{(n)}$ *defined by*

$$|D|^{-1/2} h := \sum_{1\leq |l|\leq N_n} |D_l|^{-1/2} h_l \, e^{ilt}$$

satisfies

$$\| |D|^{-1/2} h\|_{\sigma,s} \leq \frac{K}{\sqrt{\gamma\omega}} \|h\|_{\sigma, s+\frac{\tau-1}{2}} \leq \frac{K N_n^{\frac{\tau-1}{2}}}{\sqrt{\gamma\omega}} \|h\|_{\sigma,s} \quad \forall\, h \in W^{(n)}\,.$$

Proof. By (69) and $\alpha_{-l} = \alpha_l \geq \gamma\omega/|l|^{\tau-1}$

$$\begin{aligned}
\| |D|^{-1/2} h\|^2_{\sigma,s} &= \sum_{1\leq |l|\leq N_n} \| |D_l|^{-1/2} h_l\|^2_{H^1} (1+l^{2s}) e^{2\sigma|l|} \\
&\leq \sum_{1\leq |l|\leq N_n} \frac{K^2 |l|^{\tau-1}}{\gamma\omega} \|h_l\|^2_{H^1} (1+l^{2s}) e^{2\sigma|l|} \\
&\leq \frac{K'}{\gamma\omega} \|h\|^2_{\sigma, s+\frac{\tau-1}{2}}
\end{aligned}$$

because $|l|^{\tau-1}(1+l^{2s}) < 2(1+|l|^{2s+\tau-1})$ for all $|l| \geq 1$. □

To prove the invertibility of $\mathcal{L}_n(w)$ we write (64) as

$$\mathcal{L}_n(w) = \rho |D|^{1/2} \big(U + T_1 + T_2\big) |D|^{1/2}, \tag{70}$$

where

$$\begin{cases} U := |D|^{-1/2}\, D\, |D|^{-1/2}, \\ T_i := |D|^{-1/2}\, M_i\, |D|^{-1/2}\,, \quad i = 1,2\,. \end{cases} \tag{71}$$

With respect to the basis $\varphi_j(\mu, w)\, e^{ilt}$ the operator U is diagonal and its (l,j)th eigenvalue is $\mathrm{sign}(\omega^2 l^2 - \lambda_j(\mu,w)) \in \{\pm 1\}$, implying that the operator norm is

$$\|U\|_\sigma := \sup_{\|h\|_\sigma \leq 1} \|Uh\|_\sigma = 1\,. \tag{72}$$

The smallness of T_1 requires an analysis of the small divisors. Formula (66) implies, by Taylor expansion, the asymptotic dispersion relation

$$\Big| \omega_j(\mu, w) - \frac{j}{c} \Big| \leq \frac{K}{j}, \tag{73}$$

and there exists K such that, for every $x \geq 0$,

$$|x^2 - \lambda_{j^*}(\mu, w)| = \min_{j \geq 1} |x^2 - \lambda_j(\mu, w)| \quad \Rightarrow \quad j^* \geq Kx\,. \tag{74}$$

LEMMA 18 (analysis of the small divisors). *Assume the nonresonance conditions* (33)–(34) *and* $\omega > \gamma$. *Then for all* $|k|, |l| \in \{1, \dots, N_n\}$, $k \neq l$,

$$\alpha_l \alpha_k \geq \left(\frac{K\gamma^3 \omega}{|k-l|^{\frac{\tau(\tau-1)}{2-\tau}}} \right)^2,$$

where $\alpha_l := \min_{j \geq 1} |\omega^2 l^2 - \lambda_j(\mu, w)|$.

Proof. Since $\alpha_{-l} = \alpha_l$ for all l, we can suppose $l, k \geq 1$.

We distinguish two cases, if k, l are close to or far from each other. Let $\beta := (2-\tau)/\tau \in (0,1)$.

Case 1. Let $2|k-l| > (\max\{k,l\})^\beta$. By (34)

$$\alpha_k \alpha_l \geq \frac{(\gamma\omega)^2}{(kl)^{\tau-1}} \geq \frac{(\gamma\omega)^2}{(\max\{k,l\})^{2(\tau-1)}} \geq \frac{C(\gamma\omega)^2}{|k-l|^{\frac{2(\tau-1)}{\beta}}}\,.$$

Case 2. Let $0 < 2|k-l| \leq (\max\{k,l\})^\beta$. In this case $2k \geq l \geq k/2$. Indeed, if $k > l$, then $2(k-l) \leq k^\beta$, so $2l \geq 2k - k^\beta \geq k$ because $\beta \in (0,1)$—analogously if $l > k$.

Let i, resp., j, be an integer which realizes the minimum α_k, resp., α_l, and write in short $\lambda_j(\mu) := \lambda_j(\mu, w)$, $\omega_j(\mu) := \omega_j(\mu, w)$.

If both $\lambda_i(\mu), \lambda_j(\mu) \leq 0$, then $\alpha_l \geq \omega^2 l^2$, $\alpha_k \geq \omega^2 k^2$, $\alpha_l \alpha_k \geq \omega^4 > \gamma^2 \omega^2$.

If only $\lambda_j(\mu) \leq 0$, then $\alpha_l \alpha_k \geq \gamma \omega^3 l^2 / k^{\tau-1} \geq 2^{1-\tau} \gamma \omega^3 \geq 2^{1-\tau} \gamma^2 \omega^2$.

The really resonant cases happen if $\lambda_i(\mu), \lambda_j(\mu) > 0$. Suppose, for example, that $\max\{k,l\} = k$. By (73), $|\omega_j(\mu) - (j/c)| \leq K/j$, and, by (74), $i \geq K\omega k$, $j \geq K\omega l$. Hence, using also (33),

$$\begin{aligned} |(\omega k - \omega_i(\mu)) - (\omega l - \omega_j(\mu))| &= \big|\omega(k-l) - (\omega_i(\mu) - \omega_j(\mu))\big| \\ &\geq \Big|\omega(k-l) - \frac{i-j}{c}\Big| - \frac{K}{\omega l} - \frac{K}{\omega k} \\ &\geq \frac{\gamma}{(k-l)^\tau} - \frac{3K}{\omega k} \geq \frac{2^\tau \gamma}{k^{\beta\tau}} - \frac{3K}{\omega k} \end{aligned}$$

because $2(k-l) \leq k^\beta$, $2l \geq k$. Since $\beta\tau < 1$ and $k \leq 2l$,

$$\big|(\omega k - \omega_i(\mu)) - (\omega l - \omega_j(\mu))\big| \geq \frac{1}{2}\Big(\frac{\gamma}{k^{\beta\tau}} + \frac{\gamma}{l^{\beta\tau}}\Big) \quad \forall\, k \geq \Big(\frac{K}{\omega\gamma}\Big)^{\frac{1}{1-\beta\tau}} =: k^*\,.$$

We reach the same conclusion if $\max\{k,l\} = l$. It follows that, for $\max\{k,l\} \geq k^*$, there holds $|\omega k - \omega_i(\mu)| \geq \gamma/2k^{\beta\tau}$ or $|\omega l - \omega_j(\mu)| \geq \gamma/2l^{\beta\tau}$. Suppose $|\omega k - \omega_i(\mu)| \geq \gamma/2k^{\beta\tau}$. Then

$$\alpha_k = |\omega^2 k^2 - \omega_i^2(\mu)| \geq |\omega k - \omega_i(\mu)| \omega k \geq \frac{\gamma\omega}{2}\, k^{1-\beta\tau}\,.$$

Since $l \leq 2k$, for α_l we can use (34),

$$\alpha_k \alpha_l \geq \frac{\gamma\omega k^{1-\beta\tau}}{2} \frac{\gamma\omega}{l^{\tau-1}} \geq \frac{\gamma^2\omega^2}{2^\tau}\, k^{2-\tau-\beta\tau} = \frac{\gamma^2\omega^2}{2^\tau}$$

because $2 - \tau - \beta\tau = 0$.

On the other hand, if $\max\{k,l\} < k^* = (K/\omega\gamma)^{1/(\tau-1)}$, we can use (34) for both k, l:

$$\alpha_k\alpha_l \geq \frac{(\gamma\omega)^2}{(kl)^{\tau-1}} > \frac{(\gamma\omega)^2}{(k^*)^{2(\tau-1)}} = (\gamma\omega)^2 \left(\frac{\omega\gamma}{K}\right)^{\frac{1}{\tau-1}\,2(\tau-1)} > \frac{\gamma^6\omega^2}{K^2}$$

(using $\omega > \gamma$). Since $\gamma < 1$, taking the minimum for all these cases concludes the proof. □

LEMMA 19 (estimate of T_1). *Assume the nonresonance conditions* (33)–(34), $\omega > \gamma$, *and* $\Pi_W f'(u) = \sum_{l\neq 0} a_l(x)e^{ilt} \in X_{\sigma,1+\frac{\tau(\tau-1)}{2-\tau}}$. *There exists* K *such that*

$$\|T_1 h\|_\sigma \leq \frac{K\mu}{\gamma^3\omega}\,\|\Pi_W f'(u)\|_{\sigma,1+\frac{\tau(\tau-1)}{2-\tau}}\,\|h\|_\sigma \quad \forall h \in W^{(n)}\,.$$

Proof. For all $h \in W^{(n)}$, $T_1 h = \sum_{1\leq|k|\leq N_n} (T_1 h)_k\, e^{ikt}$, where

$$\begin{aligned}(T_1 h)_k &= |D_k|^{-1/2}\big(M_1|D|^{-1/2}h\big)_k \\ &= |D_k|^{-1/2}\left[\sum_{1\leq|l|\leq N_n,\, l\neq k} \mu\,\frac{a_{k-l}}{\rho}\,|D_l|^{-1/2}h_l\right].\end{aligned}$$

Setting $A_m := \|a_m/\rho\|_{H^1}$, using (69) and Lemma 18, we obtain

$$\|(T_1 h)_k\|_{H^1} \leq K\,\mu \sum_{1\leq|l|\leq N_n,\, l\neq k} \frac{A_{k-l}}{\sqrt{\alpha_k}\sqrt{\alpha_l}}\,\|h_l\|_{H^1} \;\leq\; \frac{K\mu}{\gamma^3\omega}\,S_k, \tag{75}$$

where

$$S_k := \sum_{|l|\leq N_n,\, l\neq k} A_{k-l}\,|k-l|^{\frac{\tau(\tau-1)}{2-\tau}}\|h_l\|_{H^1}\,.$$

By (75) we get, defining $S(t) := \sum_{|k|=1}^{N_n} S_k\, e^{ikt}$,

$$\begin{aligned}\|T_1 h\|_\sigma^2 &= \sum_{|k|=1}^{N_n} \|(T_1 h)_k\|_{H^1}^2 (1+k^2)e^{2\sigma|k|} \\ &\leq \left(\frac{K\mu}{\gamma^3\omega}\right)^2 \sum_{|k|=1}^{N_n} S_k^2(1+k^2)e^{2\sigma|k|} = \left(\frac{K\mu}{\gamma^3\omega}\right)^2 \|S\|_\sigma^2\,.\end{aligned}$$

Since $S = P_n(\varphi\psi)$ with $\varphi(t) := \sum_{l\in\mathbb{Z}} A_l|l|^{\frac{\tau(\tau-1)}{2-\tau}}\, e^{ilt}$ and $\psi(t) := \sum_{|l|=1}^{N_n} \|h_l\|_{H^1}\, e^{ilt}$

$$\|T_1 h\|_\sigma \leq \frac{K\mu}{\gamma^3\omega}\,\|\varphi\|_\sigma\|\psi\|_\sigma \leq \frac{K\mu}{\gamma^3\omega}\,\Big\|\Pi_W f'(u)\Big\|_{\sigma,1+\frac{\tau(\tau-1)}{2-\tau}}\,\|h\|_\sigma$$

because $\|\varphi\|_\sigma \leq 2\|\Pi_W f'(u)\|_{\sigma,1+\frac{\tau(\tau-1)}{2-\tau}}$ and $\|\psi\|_\sigma = \|h\|_\sigma$. □

LEMMA 20 (estimate of T_2). *Suppose that* $\Pi_W f'(u) \in X_{\sigma,1+\frac{\tau-1}{2}}$. *Then*

$$\|T_2 h\|_\sigma \leq \frac{K\mu}{\gamma\omega}\,\|\Pi_W f'(u)\|_{\sigma,1+\frac{\tau-1}{2}}\,\|h\|_\sigma \qquad \forall\, h \in W^{(n)}$$

for some K.

Proof. By the definitions (71) and (65) and by Lemma 17,

$$\begin{aligned}\|T_2h\|_\sigma &\le \frac{K}{\sqrt{\gamma\omega}}\,\|M_2|D|^{-1/2}h\|_{\sigma,1+\frac{\tau-1}{2}}\\ &\le \frac{K'\mu}{\sqrt{\gamma\omega}}\,\|\Pi_W f'(u)\|_{\sigma,1+\frac{\tau-1}{2}}\,\big\|d_w v(\mu,w)\big[\,|D|^{-1/2}h\big]\,\big\|_{\sigma,1+\frac{\tau-1}{2}}\\ &= \frac{K'\mu}{\sqrt{\gamma\omega}}\,\|\Pi_W f'(u)\|_{\sigma,1+\frac{\tau-1}{2}}\,\big\|d_w v(\mu,w)\big[\,|D|^{-1/2}h\big]\,\big\|_{H^1}\end{aligned}$$

because $d_w v(\mu,w)[|D|^{-1/2}h] \in V$. By Lemmas 4 and 17

$$\big\|d_w v(\mu,w)\big[\,|D|^{-1/2}h\big]\,\big\|_{H^1} \le K\,\|\,|D|^{-1/2}h\|_{\sigma,1-\frac{\tau-1}{2}} \le \frac{K}{\sqrt{\gamma\omega}}\,\|h\|_{\sigma,1},$$

implying the thesis. □

Proof of Lemma 7. By (72), $\|U\|_\sigma = 1$. If

$$\|T_1+T_2\|_\sigma < \frac{1}{2}\,, \tag{76}$$

then by Neumann series $U+T_1+T_2$ is invertible in $(W^{(n)},\|\ \|_\sigma)$ and

$$\|(U+T_1+T_2)^{-1}\|_\sigma < 2\,.$$

By Lemmas 19 and 20, condition (76) is verified if

$$\|T_1\|_\sigma \le \frac{K\mu}{\gamma^3\omega}\,\|\Pi_W f'(u)\|_{\sigma,1+\frac{\tau(\tau-1)}{2-\tau}} < \frac{1}{4} \tag{77}$$

and

$$\|T_2\|_\sigma \le \frac{K\mu}{\gamma\omega}\,\|\Pi_W f'(u)\|_{\sigma,1+\frac{\tau-1}{2}} \le \frac{K\mu}{\gamma^3\omega}\,\|\Pi_W f'(u)\|_{\sigma,1+\frac{\tau(\tau-1)}{2-\tau}} < \frac{1}{4} \tag{78}$$

(we recall that $\gamma \in (0,1)$ and $(\tau-1)/2 < \tau(\tau-1)/(2-\tau)$ because $\tau > 1$). Both conditions (77) and (78) are satisfied if

$$\frac{\mu}{\gamma^3\omega}\,\|\Pi_W f'(u)\|_{\sigma,1+\frac{\tau(\tau-1)}{2-\tau}} < \frac{1}{4K} =: K_1'\,,$$

which is condition (35). Hence, inverting (70)

$$\mathcal{L}_n(w)^{-1}h = |D|^{-1/2}\big(U+T_1+T_2\big)^{-1}|D|^{-1/2}\Big(\frac{h}{\rho}\Big)\,,$$

which, using Lemma 17, yields (36). □

7. Appendix.

Proof of Lemma 15. Let $a(x) \in L^2(0,\pi)$. Under the "Liouville change of variable"

$$x = \psi(\xi) \;\Leftrightarrow\; \xi = g(x), \qquad g(x) := \frac{1}{c}\int_0^x \Big(\frac{\rho(s)}{p(s)}\Big)^{1/2} ds\,, \tag{79}$$

we have that $(\lambda, y(x))$ satisfies

$$\begin{cases} -(p(x)y'(x))' + a(x)y(x) = \lambda\rho(x)y(x), \\ y(0) = y(\pi) = 0 \end{cases} \tag{80}$$

iff $(\nu, z(\xi))$ satisfies

$$\begin{cases} -z''(\xi) + [q(\xi) + \alpha(\xi)]\, z(\xi) = \nu z(\xi), \\ z(0) = z(\pi) = 0, \end{cases} \tag{81}$$

where

$$\nu = c^2\lambda, \qquad r(x) = \sqrt[4]{p(x)\,\rho(x)}, \qquad z(\xi) = y(\psi(\xi))\, r(\psi(\xi)),$$
$$\alpha(\xi) = c^2\, \frac{a(\psi(\xi))}{\rho(\psi(\xi))}\,, \qquad q(\xi) = c^2 Q(\psi(\xi)), \qquad Q = \frac{p}{\rho}\,\frac{r''}{r} + \frac{1}{2}\left(\frac{p}{\rho}\right)'\frac{r'}{r}\,.$$

By [25, Theorem 4 in Chapter 2, p. 35], the eigenvalues of (81) form an increasing sequence ν_j satisfying the asymptotic expansion

$$\nu_j = j^2 + \frac{1}{\pi}\int_0^\pi (q+\alpha)\, d\xi \; - \frac{1}{\pi}\int_0^\pi \cos(2j\xi)\big(q(\xi)+\alpha(\xi)\big)\, d\xi + r_j, \quad |r_j| \le \frac{C}{j},$$

where $C := C(\|q+\alpha\|_{L^2})$ is a positive constant. Moreover every ν_j is simple [25, Theorem 2, p. 30].

Since p, ρ are positive and belong to H^3, if $a \in H^1$, then $q, \alpha \in H^1$. Integrating by parts, $|\int_0^\pi \cos(2j\xi)(q+\alpha)\, d\xi| \le \|q+\alpha\|_{H^1}/j$ and so

$$\nu_j = j^2 + \frac{1}{\pi}\int_0^\pi (q+\alpha)\, d\xi + r_j', \quad |r_j'| \le \frac{C'}{j}$$

for some $C' := C'(\|q+\alpha\|_{H^1})$. Dividing by c^2 and using the inverse Liouville change of variable, we obtain the formula for the eigenvalues $\lambda_j(a)$ of (80),

$$\lambda_j(a) = \frac{j^2}{c^2} \; + \; \frac{1}{\pi c}\int_0^\pi \frac{Q\sqrt{\rho}}{\sqrt{p}}\, dx \; + \; \frac{1}{\pi c}\int_0^\pi \frac{a}{\sqrt{\rho p}}\, dx \; + \; r_j(a), \quad |r_j(a)| \le \frac{C}{j} \tag{82}$$

for some $C(\rho, p, \|a\|_{H^1}) > 0$. Formula (66) follows for $a(x) = -\mu\Pi_V f'(v(\mu, w) + w)(x)$ and some algebra.

By [25, Theorem 7, p. 43], the eigenfunctions of (81) form an orthonormal basis for L^2. Applying the Liouville change of variable (79) in the integrals, the eigenfunctions $\varphi_j(a)$ of (80) form an orthonormal basis for L^2 with respect to the scalar product $(\cdot\,,\cdot\,)_{L^2_\rho}$.

Finally, since $\varphi_j := \varphi_j(a)$ solves

$$-(p\varphi_j')' + (K\rho + a)\varphi_j = (\lambda_j(a) + K)\rho\varphi_j\,,$$

multiplying by φ_i and integrating by parts gives

$$(\varphi_j, \varphi_i)_{\mu,w} = \delta_{i,j}(\lambda_j(a) + K),$$

and (68) follows (note that $\lambda_j(a) + K > 0$ for all j and for K large enough). □

Proof of Lemma 5. Let $a, b \in H^1(0,\pi)$ and consider $\alpha := c^2 a(\psi)/\rho(\psi)$, $\beta := c^2 b(\psi)/\rho(\psi)$ constructed as above via the Liouville change of variable (79). By [25, p. 34], for every j

$$|\lambda_j(a) - \lambda_j(b)| = \frac{1}{c^2}|\nu_j(\alpha) - \nu_j(\beta)| \le \frac{1}{c^2}\|\alpha - \beta\|_\infty \le K\|a - b\|_{H^1}, \tag{83}$$

and (28) follows by the mean value theorem because $\mu\Pi_V f(v(\mu, w) + w)$ has bounded derivatives on bounded sets. □

Proof of (58). By the asymptotic formula (73)

$$\min_{j\ge 1} |\omega_{j+1}(\mu, w) - \omega_j(\mu, w)| \ge \frac{1}{c} - \frac{2K}{j} > \frac{1}{2c}$$

if $j > 4Kc$, uniformly in $\mu \in [\mu_1, \mu_2]$, $w \in B_R$. For $1 \le j \le 4Kc$ the minimum

$$m_j := \min_{(\mu,w)\in[\mu_1,\mu_2]\times B_R} |\omega_{j+1}(\mu, w) - \omega_j(\mu, w)|$$

is attained because $a \mapsto \lambda_j(a)$ is a compact function on H^1 by the compact embedding $H^1(0,\pi) \hookrightarrow L^\infty(0,\pi)$ and by (83) (see also [25, Theorem 3, pp. 31 and 34]). Each $m_j > 0$ because all the eigenvalues λ_j are simple. □

REFERENCES

[1] P. Acquistapace, *Soluzioni periodiche di un'equazione iperbolica non lineare*, Boll. Un. Mat. Ital. B (5), 13 (1976), pp. 760–777.

[2] A. Bamberger, G. Chavent, and P. Lailly, *About the stability of the inverse problem in 1D wave equations—Applications to the interpretation of seismic profiles*, Appl. Math. Optim., 5 (1979), pp. 1–47.

[3] V. Barbu and N. H. Pavel, *Periodic solutions to nonlinear one dimensional wave equation with x-dependent coefficients*, Trans. Amer. Math. Soc., 349 (1997), pp. 2035–2048.

[4] T. Bartsch, Y. H. Ding, and C. Lee, *Periodic solutions of a wave equation with concave and convex nonlinearities*, J. Differential Equations, 153 (1999), pp. 121–141.

[5] J. Berkovits and J. Mawhin, *Diophantine approximation, Bessel functions and radially symmetric periodic solutions of semilinear wave equations in a ball*, Trans. Amer. Math. Soc., 353 (2001), pp. 5041–5055.

[6] M. Berti, *Nonlinear Oscillations of Hamiltonian PDEs*, Progr. Nonlinear Differential Equations Appl. 74, H. Brézis, ed., Birkhäuser, Boston, 2008.

[7] M. Berti and L. Biasco, *Forced vibrations of wave equations with non-monotone nonlinearities*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 23 (2006), pp. 437–474.

[8] M. Berti and P. Bolle, *Cantor families of periodic solutions for completely resonant nonlinear wave equations*, Duke Math. J., 134 (2006), pp. 359–419.

[9] M. Berti and P. Bolle, *Cantor families of periodic solutions for wave equations via a variational principle*, Adv. Math., 217 (2008), pp. 1671–1727.

[10] M. Berti and M. Procesi, *Quasi-periodic solutions of completely resonant forced wave equations*, Commun. Partial Differential Equations, 31 (2006), pp. 959–985.

[11] J. Bourgain, *Construction of quasi-periodic solutions for Hamiltonian perturbations of linear equations and applications to nonlinear PDE*, Int. Math. Res. Not., 11 (1994), pp. 475–497.

[12] J. Bourgain, *Periodic solutions of nonlinear wave equations*, in Harmonic Analysis and Partial Differential Equations, Chicago Lectures in Math., University of Chicago Press, Chicago, IL, 1999, pp. 69–97.

[13] J. Bourgain, *Nonlinear Schrödinger equations*, in Hyperbolic Equations and Frequency Interactions (Park City, UT, 1995), IAS/Park City Math. Ser. 5, AMS, Providence, RI, 1999, pp. 3–157.

[14] H. Brézis and L. Nirenberg, *Forced vibrations for a nonlinear wave equation*, Comm. Pure Appl. Math., 31 (1978), pp. 1–30.

[15] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. I, Interscience Publishers, New York, 1953.

[16] W. Craig, *Problèmes de Petits Diviseurs dans les Équations aux Dérivées Partielles*, Panor Synthèses 9, Société Mathématique de France, Paris, 2000.
[17] W. Craig and E. Wayne, *Newton's method and periodic solutions of nonlinear wave equations*, Comm. Pure Appl. Math., 46 (1993), pp. 1409–1498.
[18] R. de la Llave, *Variational methods for quasi-periodic solutions of partial differential equations*, in Hamiltonian Systems and Celestial Mechanics (Pátzcuaro, 1998), World Sci. Monogr. Ser. Math. 6, World Scientific, River Edge, NJ, 2000, pp. 214–228.
[19] J.-M. Fokam, *Forced Vibrations via Nash-Moser Iteration*, Ph.D. thesis, University of Texas, Austin, TX.
[20] A. Friedman, *Partial Differential Equations*, Robert E. Krieger, Huntington, NY, 1976.
[21] S. Kuksin, *Hamiltonian perturbations of infinite-dimensional linear systems with imaginary spectrum*, Funktsional. Anal. i Prilozhen., 21 (1987), pp. 22–37.
[22] S. Kuksin, *Analysis of Hamiltonian PDEs*, Oxford Lecture Ser. Math. Appl. 19, Oxford University Press, New York, 2000.
[23] P. J. McKenna, *On solutions of a nonlinear wave question when the ratio of the period to the length of the intervals is irrational*, Proc. Amer. Math. Soc., 93 (1985), pp. 59–64.
[24] P. I. Plotnikov and L. N. Yungerman, *Periodic solutions of a weakly nonlinear wave equation with an irrational relation of period to interval length*, Differential Equations, 24 (1988), pp. 1059–1065.
[25] J. Pöschel and E. Trubowitz, *Inverse Spectral Theory*, Academic Press, Orlando, FL, 1987.
[26] P. Rabinowitz, *Periodic solutions of nonlinear hyperbolic partial differential equations*, Comm. Pure Appl. Math., 20 (1967), pp. 145–205.
[27] P. Rabinowitz, *Time periodic solutions of nonlinear wave equations*, Manuscripta Math., 5 (1971), pp. 165–194.
[28] P. Rabinowitz, *Free vibration of a semilinear wave equation*, Comm. Pure Appl. Math, 31 (1978), pp. 31–68.
[29] I. A. Rudakov, *Periodic solutions of a nonlinear wave equation with nonconstant coefficients*, Math. Notes, 76 (2004), pp. 395–406.
[30] W. M. Schmidt, *Diophantine Approximation*, Lecture Notes in Math. 785, Springer-Verlag, Berlin, 1980.
[31] K. Tanaka, *Infinitely many periodic solutions for the equation:* $u_{tt} - u_{xx} \pm |u|^{p-1}u = f(t,x)$, Trans. Amer. Math. Soc., 307 (1988), pp. 615–645.
[32] E. Wayne, *Periodic and quasi-periodic solutions of nonlinear wave equations via KAM theory*, Comm. Math. Phys., 127 (1990), pp. 479–528.

SIAM J. MATH. ANAL.
Vol. 40, No. 1, pp. 413–441

NORMAL FORMS, QUASI-INVARIANT MANIFOLDS, AND BIFURCATIONS OF NONLINEAR DIFFERENCE-ALGEBRAIC EQUATIONS*

R. BEARDMORE† AND K. WEBSTER†

Abstract. We study the existence of quasi-invariant manifolds in a neighborhood of a fixed point of the *difference*-algebraic equation (ΔAE) $F(z_n, z_{n+1}) = 0$, where $F : \mathbb{R}^{2m} \to \mathbb{R}^m$ is a smooth map satisfying $F(0,0) = 0$. We demonstrate the existence of quasi-invariant manifolds on which one can define forward and backward orbits of the ΔAE under mild assumptions on its linearization at the fixed point $z = 0$. Indeed, by assuming this linearization to be a regular matrix pencil, one obtains a functional equation satisfied by invariant manifolds which can be solved using an extension of the contraction mapping to spaces that satisfy an interpolation property. If the ΔAE under study is permitted to depend smoothly on a parameter, we then obtain a Neimark–Sacker bifurcation theorem as a corollary that can be deduced from the existence of a normal form for nonlinear ΔAEs.

Key words. invariant manifolds, bifurcations, difference-algebraic equations

AMS subject classifications. 39B72, 39A11, 37G99

DOI. 10.1137/050638618

1. Introduction.

The purpose of this paper is to provide an analysis of the invariant manifolds and bifurcations found in a class of *difference*-algebraic equations (ΔAEs) of the form

$$F(z_n, z_{n+1}) = 0; \tag{1.1}$$

the nomenclature and chosen acronym for (1.1) have been taken from [4, 22]. We assume that $F(= F(z, \bar{z})) : \mathbb{R}^{2m} \to \mathbb{R}^m$ is a smooth map satisfying $F(0,0) = 0$ and say that (1.1) is *singular* because the partial derivative $d_{\bar{z}}F(0,0)$ is not an isomorphism from $\mathbb{R}^m$ to itself. The purpose of the first part of this paper is to provide conditions under which (1.1) has suitably defined invariant manifolds that contain the fixed point $z = 0$, where the main difficulty to overcome in this analysis is the fact that forward orbits of (1.1) are not necessarily uniquely defined in a neighborhood of the fixed point.

The second part of the paper utilizes the existence of the aforementioned invariant manifolds to investigate the presence of bifurcations in ΔAEs in the sense that by extending F to be a C^k-mapping of the form $F : \mathbb{R}^{2m} \times \mathbb{R} \to \mathbb{R}^{2m}$, where $k \geq 5$, we examine the structure of invariant sets in the family of ΔAEs

$$F(z_n, z_{n+1}, \mu) = 0. \tag{1.2}$$

Our rationale is taken from bifurcation theory for maps which leads to the following question. If $F(0,0,\mu) = 0$ for all μ in some interval and the one-parameter family of matrix pencils

$$\mathcal{P}(\mu) := (A(\mu), B(\mu)) := (d_{\bar{z}}F(0,0,\mu), d_z F(0,0,\mu)),$$

*Received by the editors August 22, 2005; accepted for publication (in revised form) August 28, 2007; published electronically May 7, 2008.
http://www.siam.org/journals/sima/40-1/63861.html
†Department of Mathematics, Imperial College, 180 Queen's Gate, London, SW7 2AZ, United Kingdom (r.beardmore@ic.ac.uk, k.webster@ic.ac.uk). The first author was supported by Nuffield Foundation grant NAL/00511/G. The second author was supported by EPSRC grant GR/S/17215/01.

where F has $(z, \overline{z}, \mu)$ as its argument, is such that $\mathcal{P}(\mu_0)$ has a finite eigenvalue of unit modulus in the complex plane, does an invariant set of (1.2) bifurcate from the fixed point $z = 0$ at $\mu = \mu_0$?

Due to the lack of forward uniqueness we modify what we mean by the term *invariant*, which we do by using the prefixed *quasi-invariant*, and say that a set $\mathcal{Q} \subset F^{-1}\{0\}$ is quasi-invariant for (1.1) if, for $(z, \overline{z}) \in \mathcal{Q}$, there is a subsequent iterate $(\overline{z}, \overline{\overline{z}}) \in \mathcal{Q}$ for some $\overline{\overline{z}} \in \mathbb{R}^m$.

The paper is organized in the following way: The remainder of section 1 briefly covers the linear prerequisites for (1.1). Section 2 then provides some motivating applications. Section 3 presents the basic definitions of how (1.1) defines a local dynamical system and gives the invariant manifold equation of fixed points of (1.1). Section 4 provides a reformulation of the invariant manifold equation from section 3 as a nonlinear fixed-point problem in suitable Banach spaces which then is shown to have a solution in section 4.3. Section 5 gives a normal form for (1.1) with and without the presence of a bifurcation parameter. This section concludes with theorems that can be deduced using these normal forms, giving bifurcation results for (1.1) when that parameter is included. Finally, section 6 finishes the paper with a series of examples.

1.1. The linear case: Kronecker normal form. As a precursor to the analysis of the nonlinear problem (1.1), consider the linear case

$$Bz_n + Az_{n+1} = 0, \tag{1.3}$$

where $A, B : \mathbb{R}^m \to \mathbb{R}^m$ are linear maps and A is singular. In order to discuss the behavior of (1.3) we first introduce a normal form for matrix pencils.

When A is singular, the matrix pencil (A, B) is said to be *regular* if there is an $\omega \in \mathbb{C}$ such that $\det(\omega A + B) \neq 0$. The following result is well known for regular matrix pencils (see [7, 3]): There are complementary subspaces $K_1 \simeq \mathbb{R}^p, K_2 \simeq \mathbb{R}^q \subset \mathbb{R}^m$ such that $p + q = m$ and nonsingular linear mappings P, Q on $\mathbb{R}^m$, $L : K_1 \to K_1$, and $N : K_2 \to K_2$ such that

$$PAQ = \begin{pmatrix} I_p & 0 \\ 0 & N \end{pmatrix}, \quad PBQ = \begin{pmatrix} L & 0 \\ 0 & I_q \end{pmatrix}; \tag{1.4}$$

I_p and I_q are identities on K_1 and K_2, respectively. Moreover, there is a $\nu \geq 1$ such that $N^\nu = 0$, and ν is said to be the Kronecker index of (A, B).

The Kronecker normal form (KNF) in (1.4) can be used to rewrite (1.3) as a coupled system of difference equations

$$Lu_n + u_{n+1} = 0, \quad v_n + Nv_{n+1} = 0, \tag{1.5}$$

which has the solution $u_n = (-L)^n u_0$ and $v_n \equiv 0$ for all n, and thus (1.3) *has* a quasi-invariant subspace that arises from the quasi-invariant space $\{(u, v) : v = 0\}$ associated with (1.5). It is the presence of the former that we shall exploit in the remainder of the paper to study nonlinear perturbations of (1.5) that arise from a consideration of problems of the form (1.1).

1.2. Notation. If we define the spectrum of a matrix pencil to be

$$\sigma(A, B) = \{\lambda \in \mathbb{C} : \det(\lambda A + B) = 0\},$$

then $\sigma(A, B) = -\sigma(L)$ (*note the minus sign*), and p as defined within the KNF above coincides with the number of finite eigenvalues of (A, B), where eigenvalues are

counted according to their algebraic multiplicity. The matrix pencil (A, B) is said to be *hyperbolic* if $\sigma(A, B)$ is nonempty and contains no elements of unit modulus; otherwise, it is said to be *elliptic*. We shall also write $\rho(A, B) = \sup\{|\lambda| : \lambda \in \sigma(A, B)\}$ and denote the spectral radius of any linear mapping L by $\rho(L)$. Throughout we shall use # to denote the cardinality of a set of eigenvalues, counted according to algebraic multiplicity.

We shall use $BL(X, Y)$ to denote the space of continuous linear maps from one normed linear space X to another Y, even when X and Y are finite-dimensional. We shall use $B_\epsilon(x)$ for the open ball of radius ϵ about x, and $B_\epsilon(x; X)$ will specify that this ball is contained in the space X. If $L \in BL(X, Y)$, we shall denote the usual operator norm by $\|L\|_{BL(X,Y)}$, which is given by $\sup\{\|Lx\|_Y : x \in X, \|x\|_X = 1\}$. If the context is clear, we shall simply write $\|L\|$, and $BL(X)$ is also used for $BL(X, X)$. Throughout, if $F : X \to Y$ is a nonlinear mapping, then $dF(x) \in BL(X, Y)$ shall denote the Fréchet derivative, and when acting on $h \in X$ it will be written with square brackets, as in $dF(x)[h]$. Similarly, $d^2F(x)[h, k]$ denotes the second derivative, and this is bilinear in $[h, k]$.

If n is a positive integer, we shall use $\mathcal{O}_n(x)$ on occasion to denote any mapping, H, say, with the property that $\lim_{x\to 0} \|H(x)\|/\|x\|^n$ exists.

2. Motivation. There are several problems from control theory and numerical analysis that lead to discrete systems where the relationship between the current and future states of a system are not explicit; see [12, 10, 6, 14] for examples.

2.1. Discretized differential-algebraic equations. In [11] the authors apply a Runge–Kutta method to solve a differential-algebraic boundary-value problem arising from an optimal control problem, yielding a nonlinear *difference-algebraic equation* where the control plays the role of an implicit variable. For example, using a forward-Euler method to discretize the differential-algebraic equation (DAE)

$$\dot{x} = f(x, y), 0 = g(x, y) \qquad ((x(0), y(0)) \text{ given})$$

yields the ΔAE

$$x_{n+1} = x_n + hf(x_n, y_n), 0 = g(x_n, y_n) \qquad ((x_0, y_0) \text{ given}),$$

where h is a small parameter. A singularity in this context occurs when the partial derivative $d_y g(x, y)$ is singular on some subset of $g^{-1}\{0\}$.

Over the past decade a great deal of attention has been devoted to singular DAEs

$$F(z, \dot{z}) = 0, \tag{2.1}$$

where $d_{\dot{z}} F(z, \dot{z})$ changes rank on some set, and our study of (1.1) can be viewed as an extension of the work undertaken on (2.1) to the discrete-time case.

It is well known that (2.1) supports a range of singular and regular behavior, including *impasse points* and *pseudoequilibria* [16, 19, 17, 20]. However, not a great deal of the current literature is devoted to the study of *bifurcations* of DAEs nor to the unfolding of singularities in DAEs, such as the image and kernel singularities defined in [23]. One reason for this is the difficulty of proving a suitable center manifold theorem that can cope with the kind of singularities peculiar to DAE. We do note, however, that a Hopf bifurcation theorem is presented in [9] for systems of the form $\dot{x} = f(x, \dot{x}, \alpha)$, where α is a bifurcation parameter, $x \in \mathbb{R}^m$, and f is *nonexpansive* with respect to $\dot{x}$, a case that does include certain DAE singularities; a Hopf bifurcation

theorem for regular DAEs can be found in [15]. We also note that there are results in the DAE literature that yield the existence of an invariant manifold containing an equilibrium point; for regular DAEs see [18], and for singular DAEs see [24, 2].

2.2. Output-nulling control. Take a discrete dynamical system of the form

$$x_{n+1} = f(x_n, u_n),\ y_n = g(x_n, u_n), \tag{2.2}$$

where (x_n) is a sequence of states, (u_n) are controls, and (y_n) is a sequence of outputs. One may ask whether there is an admissible control that nullifies or fixes the outputs: Given (y_n), does there exist a sequence pair $((x_n),(u_n))$ satisfying (2.2)? The resulting equation is an infinite-dimensional system of equations that has a structure reminiscent of the semiexplicit, index-1 DAE (for terminology, see [3]).

2.3. Optimal control. The preprint [14] is relevant to the present work as it presents an invariant manifold result which leads to the existence of a control for the following variational problem with an infinite horizon:

$$\min_{(u_k)} \left\{ \sum_{k=0}^{\infty} \ell(x_k, u_k) : x_{k+1} = f(x_k, u_k), x_0 \in \mathbb{R}^m, u_k \in B_\epsilon(0; \mathbb{R}^p) \right\}$$

such that $f(0,0) = 0$ and $\ell(0,0) = 0$. An optimal orbit satisfies the first-order optimality conditions given by the quasi-linear, *implicit* difference equation

$$x_{k+1} = f(x_k, u_k), \tag{2.3}$$

$$\frac{\partial H}{\partial x}(x_k, u_k, \lambda_{k+1}) = \lambda_k, \tag{2.4}$$

$$0 = \frac{\partial H}{\partial u}(x_k, u_k, \lambda_{k+1}), \tag{2.5}$$

where H is the Hamiltonian $H(x,u,\lambda) = \lambda^T f(x,u) + \ell(x,u)$. In [14] the author demonstrates the existence of a *stable manifold* associated with (2.3)–(2.5) which has a dimension that coincides with the number of eigenvalues of the linearization about its fixed point, and the existence of this stable manifold then provides the necessary optimal control. The obstacle treated in [14] is the existence of a zero closed-loop eigenvalue which is analogous to the type of singularity treated in this paper. One can see the resemblance of (2.3)–(2.5) to (1.5) in that state variables (x_k) propagate forwards in time in (2.3)–(2.5), whereas adjoint variables (λ_k) propagate backwards, a property shared by (1.5).

3. A functional equation for quasi-invariant manifolds. Let us now define in what sense we expect (1.1) to induce a dynamical system. An element $z \in \mathbb{R}^m$ is said to be a *fixed point* of (1.1) if $F(z,z) = 0$. If z denotes the first argument of F and $\overline{z}$ the second, as in $F(z,\overline{z})$, then we define the following conditions:

(A1) $z = 0$ is a fixed point of (1.1): $F(0,0) = 0$,
(A2) $\det(d_{\overline{z}} F(0,0)) = 0$, and
(A3) there is a $\xi \in \mathbb{C}$ such that $\det(d_z F(0,0) + \xi d_{\overline{z}} F(0,0)) \neq 0$.

Throughout we make use of the matrix pencil (A,B), where

$$A := d_{\overline{z}} F(0,0) \quad \text{and} \quad B := d_z F(0,0), \tag{3.1}$$

and (A3) is the condition that (A,B) is regular. We shall assume that $F \in C^k(\mathbb{R}^{2m}, \mathbb{R}^m)$ for $k > 3$ and seek *local* and *global orbits* of the ΔAE (1.1) in the following sense.

DEFINITION 1. *A sequence* $(z_n)_{j=0}^{J}$ *is said to be a* J-orbit *of* (1.1) *for* $J \in \mathbb{N}$ *if*

$$F(z_j, z_{j+1}) = 0 \quad \textit{for} \quad 0 \le j \le J-1 \quad \textit{and} \quad 2 \le J < \infty,$$

and a local orbit *is a* J*-orbit for some* $J \ge 2$. *If there is a* $z_2 \in \mathbb{R}^m$ *such that* $(z_0, z_1; z_2)$ *is a local* 2*-orbit, then we say that the initial condition* (z_0, z_1) supports *this orbit. A sequence* $(z_n)_{j=0}^{\infty}$ *is said to be a* global orbit *of* (1.1) *as it is a* J*-orbit for each* $J \ge 2$.

Following the terminology used for DAEs, a pair (z_0, z_1) such that $F(z_0, z_1) = 0$ is said to be *consistent*, and if this pair supports some orbit, then it is said to be a consistent initial condition. We could also have analogously defined backward orbits for $J \le -2$, but we omit this for brevity. Note that initial conditions lie in $\mathbb{R}^{2m}$ and not $\mathbb{R}^m$, a property that is analogous to DAEs whereby initial positions *and* certain initial derivatives must be provided in order to obtain the existence of solutions.

We now give the definition which stipulates how we expect (1.1) to induce a dynamical system.

DEFINITION 2. (1.1) *induces a local dynamical system on a manifold* $\mathcal{M} \subset F^{-1}\{0\} \subset \mathbb{R}^{2m}$ *which contains the origin of* $\mathbb{R}^{2m}$ *if there is a ball* $\mathcal{M}_r := \mathcal{M} \cap B_r(0; \mathbb{R}^{2m})$ *such that for each* $(z, \overline{z}) \in \mathcal{M}_r$ *there is a* unique $(\overline{z}, \overline{\overline{z}}) \in \mathcal{M}$. *If these conditions hold,* $\mathcal{M}$ *is said to be* a solution manifold of (1.1).

This definition ensures that every point $(z, \overline{z}) \in \mathcal{M}_r$ supports the nontrivial 3-orbit $(z, \overline{z}; \overline{\overline{z}})$ and that the point $\overline{\overline{z}} \in \mathbb{R}^m$ is uniquely determined if we are to impose the requirement that $(\overline{z}, \overline{\overline{z}}) \in \mathcal{M}$.

DEFINITION 3. *A set* $\mathcal{Q} \subset F^{-1}\{0\} \subset \mathbb{R}^{2m}$ *is said to be* quasi-*invariant if, for each* $(z, \overline{z}) \in \mathcal{Q}$, there exists *a* $\overline{\overline{z}} \in \mathbb{R}^m$ *such that* $(\overline{z}, \overline{\overline{z}}) \in \mathcal{Q}$.

As an aside, note that (1.1) induces a trivial dynamical system on the quasi-invariant set $\{(0,0)\}$ by virtue of (A1), even if assumption (A3) fails. Note also that a solution manifold $\mathcal{M}$ is not necessarily unique; it is the local orbit within $\mathcal{M}$ that must be uniquely determined. Indeed, there may well be many possible choices for $\overline{\overline{z}}$ in order to keep the orbit on $F^{-1}\{0\}$, many of which may not be elements of $\mathcal{M}$.

3.1. The functional equation. Our strategy for locating quasi-invariant manifolds of (1.1) is to study a functional equation obtained in an analogous manner to the center-manifold equation from the theory of invariant manifolds for maps. The solution of this equation then provides the manifold $\mathcal{M}$ needed to form a local dynamical system for (1.1). We shall show in Theorem 1 that one can find a linear space $K_1 \subset \mathbb{R}^{2m}$ with an associated locally defined, differentiable map $\varphi : K_1 \to K_1$, a manifold $\mathcal{M} \subset F^{-1}\{0\}$, and a local diffeomorphism $\theta : K_1 \to \mathcal{M}$ such that

$$F(\theta(u)) = 0 \Longrightarrow F(\theta(\varphi(u))) = 0.$$

As a result, we will be able to ensure that (1.1) induces a local dynamical system on $\mathcal{M}$ essentially by iterating the map φ. This simply means that if $(z, \overline{z}) = \theta(u)$ is a consistent initial condition, then $(\overline{z}, \overline{\overline{z}}) = \theta(\varphi(u))$ and $(\overline{\overline{z}}, \overline{\overline{\overline{z}}}) = \theta(\varphi(\varphi(u)))$ provide subsequent iterates of (1.1).

Returning to (1.1), let us change the form of the problem by setting $w_n = z_{n+1}$, so that along an orbit of (1.1) we have

$$z_{n+1} = w_n, \tag{3.2}$$

$$0 = Bz_n + Aw_n + \Phi(w_n, z_n), \tag{3.3}$$

where Φ is the C^k function which satisfies $\Phi(0,0) = 0, d\Phi(0,0) = 0$ and which is defined by

$$F(z,w) - Bz - Aw := \Phi(w,z).$$

The problem of finding an initial condition which is consistent, (z_0, w_0), say, is of an algebraic nature, whereas the problem of finding an orbit which is supported by this initial condition is a *dynamic* problem. This means that the problem of finding a manifold of orbits of a ΔAE will lead not to an algebraic equation that one could tackle using an elementary version of the implicit function theorem but instead to a functional equation.

Let us now obtain this functional equation. By applying the condition that (A, B) is a regular matrix pencil (condition (A3)), it follows that

$$(\mathcal{A}, \mathcal{B}) := \left(\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & I \\ B & A \end{pmatrix} \right) \tag{3.4}$$

is also a regular matrix pencil. If we define the vector $W_n = (z_n, w_n) \in \mathbb{R}^{2m}$, using (3.2)–(3.3), (1.1) can be written in the semilinear form

$$\mathcal{A}W_{n+1} = \mathcal{B}W_n + \Psi(W_n),$$

where Ψ is the C^k-mapping $\Psi(W) := (0, \Phi(W))$, so that $\Psi(0) = 0$ and $d\Psi(0) = 0$. There are mappings P and Q that put $(\mathcal{A}, \mathcal{B})$ in Kronecker normal form, and, by setting $W_n = QX_n$, we may write (3.2)–(3.3) in the form

$$[P\mathcal{A}Q]\, X_{n+1} = [P\mathcal{B}Q]\, X_n + P\Psi(QX_n), \tag{3.5}$$

where the terms in square brackets are in normal form:

$$\begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix} X_{n+1} = \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix} X_n + P\Psi(QX_n).$$

Consequently, there are linear spaces $K_1 \simeq R^p$ and $K_2 \simeq \mathbb{R}^q$ such that $K_1 \oplus K_2 \simeq \mathbb{R}^{2m}$ and $X_n = (u_n, v_n) \in K_1 \oplus K_2$, where (u_n, v_n) satisfies the difference equation in normal form

$$\text{(NF)} \qquad \begin{cases} u_{n+1} &= Cu_n + f(u_n, v_n), \\ Nv_{n+1} &= \quad v_n - g(u_n, v_n). \end{cases}$$

We now ask that there is a manifold given by the graph of some function h on which one can solve (NF) uniquely in a neighborhood of the fixed point $(u, v) = (0, 0)$ in the sense that $v_n = h(u_n)$ holds along orbits. This imposes the two conditions

$$u_{n+1} = Cu_n + f(u_n, h(u_n)) \quad \text{and} \quad Nh(u_{n+1}) = h(u_n) - g(u_n, h(u_n))$$

on h, and it follows that the local orbit (u_n, v_n) of (NF) can be found if h satisfies the functional equation

$$h(u) = Nh(Cu + f(u, h(u))) + g(u, h(u)), \qquad h(0) = 0,\ dh(0) = 0, \tag{3.6}$$

for all u in some neighborhood of the origin in $\mathbb{R}^p$. The boundary conditions in (3.6) ask first that the fixed point $(u, v) = (0, 0)$ of (NF) lies on the graph of h and then that this graph is tangent to the quasi-invariant subspace obtained on setting $f = 0$ and $g = 0$ in (NF).

3.2. Further preliminaries.

3.2.1. Perturbation of eigenvalues. For completeness we have included the following two preliminary results regarding the spectra of one-parameter families of matrix pencils, which are mappings of the form

$$\mathcal{P} : (-1, 1) \to BL(\mathbb{R}^m) \times BL(\mathbb{R}^m); \mu \mapsto (A(\mu), B(\mu)).$$

If we define the family of analytic functions

$$f_\mu(\omega) = \det(\omega A(\mu) + B(\mu)),$$

the multiplicity of an eigenvalue of a matrix pencil is then the multiplicity of the corresponding zero of $f_\mu(\cdot)$, which is at most m. The identity

$$\frac{d}{d\omega} f_\mu(\omega) = f_\mu(\omega) \mathrm{tr}[(\omega A(\mu) + B(\mu))^{-1} A(\mu)], \tag{3.7}$$

whenever this inverse is defined, can be used to obtain the following two lemmas.

LEMMA 1 (C^1-dependence of eigenvalues). *Suppose that $\mathcal{P}(\mu) := (A(\mu), B(\mu))$ is a C^1-parameterized family of real matrix pencils, with $\mu \in (-1, 1)$, such that $\mathcal{P}(0)$ is a regular matrix pencil. An element $\lambda_0 \in \sigma(\mathcal{P}(0))$ is said to be an* algebraically simple eigenvalue *of $\mathcal{P}(0)$ if*

$$\ker(\lambda_0 A(0) + B(0)) = \langle x_0 \rangle \quad \text{and} \quad x_0 \notin \mathrm{ran}(\lambda_0 A(0) + B(0)).$$

If λ_0 is an algebraically simple eigenvalue of $\mathcal{P}(0)$, then there is a C^1-parameter family of algebraically simple eigenvalues $\lambda(\mu) \in \mathbb{C}$ of $\mathcal{P}(\mu)$ such that $\lambda(0) = \lambda_0$, with a corresponding C^1 family of unit eigenvectors $x(\mu)$, with $x(0) = x_0$.

Proof. This follows from the implicit function theorem applied to the system $F(\lambda, x, \mu) = (0, 0)$, where $F(\lambda, x, \mu) := \big[(\lambda A(\mu) + B(\mu))x, \|x\|_2^2 - 1\big]$. ∎

LEMMA 2 (C^0-dependence of eigenvalues). *Suppose that $\mathcal{P}(\mu) := (A(\mu), B(\mu))$ is a C^0-parameterized family of real matrix pencils, with $\mu \in (-1, 1)$, such that $\mathcal{P}(0)$ is regular. If λ_0 is an eigenvalue of $\mathcal{P}(0)$ of algebraic multiplicity l, then it is isolated in the complex plane, and for each $\epsilon > 0$ there is a $\delta > 0$ such that if $|\mu| < \delta$, then $\mathcal{P}(\mu)$ has l eigenvalues (counted according to algebraic multiplicity) in the disk $D(\lambda_0, \epsilon)$.*

Proof. As $f_0(\cdot)$ does not vanish identically because $\mathcal{P}(0)$ is regular by assumption, neither can $f_\mu(\cdot)$ for sufficiently small μ. The *isolatedness* of eigenvalues of $\mathcal{P}(\mu)$ is a consequence of the fact that analytic functions have isolated zeros. Now by using (3.7) we integrate around a closed circle in the complex plane with center $\omega = \lambda_0$ and radius ϵ, from where

$$\#\{\sigma(\mathcal{P}(\mu)) \cap D(\lambda_0, \epsilon)\} = \frac{1}{2\pi i} \oint_{\partial D(\lambda_0, \epsilon)} \mathrm{tr}[(\omega A(\mu) + B(\mu))^{-1} A(\mu)] d\omega,$$

where $D(\lambda_0, \epsilon)$ is an open disk of radius ϵ about λ_0 in the complex plane. This quantity is integer-valued and depends continuously on μ, and the result now follows. ∎

3.2.2. Notation. From this point we shall identify the linear space K_1 from the KNF with $\mathbb{R}^p$ and K_2 with $\mathbb{R}^q$; now let $|\cdot|_p$ and $|\cdot|_q$ denote norms on $\mathbb{R}^p$ and $\mathbb{R}^q$, and let $\Omega_\delta = \{u \in \mathbb{R}^p : |u|_p < \delta\}$. We also assume that the unit sphere $\partial\Omega_1$ is a C^∞ manifold.

Let $C^0(\overline{\Omega}_\delta, \mathbb{R}^q)$ be the Banach space of continuous maps on $\overline{\Omega}_\delta$ with norm $\|h\|_{C^0} = \sup_{u\in\overline{\Omega}_\delta} |h(u)|_q$. Similarly, let $C^j(\overline{\Omega}_\delta, \mathbb{R}^q)$ be the space of all j-times continuously differentiable functions on $\overline{\Omega}_\delta$ with norm $\|h\|_{C^j} = \max_{0\le i\le j} \sup_{u\in\overline{\Omega}_\delta} \|d^i h(u)\|_{C^0}$, where d^i denotes the ith (Fréchet) derivative, so that $d^j h(u)$ is a j-linear form which we denote $[k_1, \dots, k_j] \to d^j h(u)[k_1, \dots, k_j]$. Consequently, we have the norm of a higher derivative given by the formula

$$\|d^j h\|_{C^0} = \sup_{u\in\overline{\Omega}_\delta} \sup_{|k_i|_p\le 1} |d^j h(u)[k_1, \dots, k_j]|_q. \tag{3.8}$$

If M is any multilinear form on a linear space Z and $z \in Z$, then $M[z]^{(k)}$ is shorthand for $M[z, z, \dots, z]$.

From the smoothness of the unit sphere $\partial\Omega_1$, it follows that the embedding of $C^{j+1}(\overline{\Omega}_\delta)$ into $C^j(\overline{\Omega}_\delta)$ is compact, so that if $(h_n) \subset C^{j+1}(\overline{\Omega}_\delta)$ is bounded in the norm of the latter space, there is a subsequence (h_{n_k}) which converges in $C^j(\overline{\Omega}_\delta)$ to some element of $C^j(\overline{\Omega}_\delta)$. We shall also make limited use of the Hölder spaces, which we denote by $C^{j+\alpha}(\overline{\Omega}_\delta)$ whenever j is an integer and $0 < \alpha < 1$, recalling the compact embedding $C^{j+\alpha}(\overline{\Omega}_\delta) \subset C^{j+\beta}(\overline{\Omega}_\delta)$ if $\alpha > \beta$.

It can be somewhat notationally cumbersome to include all of the references to the underlying spaces in all of the norms that we use, so we shall limit their use and expect that the precise meaning can be taken from context.

4. Solving the fixed-point problem (3.6). It is not (NF) that we shall seek to solve directly, but we make the substitution

$$u = \epsilon\tilde{u}, v = \epsilon\tilde{v}$$

in (NF) to give (after removal of the tildes for clarity)

$$\left\{ \begin{array}{rl} u_{n+1} & = Cu_n + \epsilon^{-1} f(\epsilon u_n, \epsilon v_n), \\ Nv_{n+1} & = \quad v_n - \epsilon^{-1} g(\epsilon u_n, \epsilon v_n). \end{array} \right. \tag{NF}_\epsilon$$

As the functions f and g are higher than linear order at the origin, $(\mathrm{NF})_\epsilon$ is in fact smooth with respect to variations in ϵ.

Let us define the one-parameter family of C^k functions f_ϵ and g_ϵ (with C^{k-1} dependence on ϵ) by

$$\mathsf{f}_\epsilon(u, v) = \epsilon^{-1} f(\epsilon u, \epsilon v) \quad \text{and} \quad \mathsf{g}_\epsilon(u, v) = \epsilon^{-1} g(\epsilon u, \epsilon v),$$

respectively. For $j \in \mathbb{N}$ we also have

$$d^j \mathsf{f}_\epsilon(u, v) = \epsilon^{j-1} d^j f(\epsilon u, \epsilon v) \quad \text{and} \quad d^j \mathsf{g}_\epsilon(u, v) = \epsilon^{j-1} d^j g(\epsilon u, \epsilon v), \tag{4.1}$$

whenever these derivatives are defined.

Now seek an invariant manifold $\mathcal{M}$ of $(\mathrm{NF})_\epsilon$ given by a graph on which

$$v_n = h(u_n),$$

and then $\mathcal{M}$ can be realized as such a graph if there is a solution of the nonlinear functional equation

$$\left\{ \begin{array}{l} h(u) = Nh(Cu + \mathsf{f}_\epsilon(u, h(u))) + \mathsf{g}_\epsilon(u, h(u)), \\ h(0) = 0, dh(0) = 0. \end{array} \right. \tag{4.2}$$

4.1. Preliminary estimates. The following are simple but essential estimates on the derivatives of f and g. By the mean-value inequality and the fact that the mapping $(u,v) \mapsto (f(u,v), g(u,v))$ and its derivative vanish at $(u,v) = (0,0)$, there exists an ℓ $(= \ell(\delta, r)) > 0$ such that

$$|f(u,v)|_p \le \ell \|(u,v)\|^2, \quad |g(u,v)|_q \le \ell \|(u,v)\|^2, \tag{4.3}$$

and

$$\|df(u,v)\| \le \ell \|(u,v)\|, \quad \|dg(u,v)\| \le \ell \|(u,v)\|, \tag{4.4}$$

whenever $|u|_p \le \delta, |v|_q \le r$, where here and throughout we use the norm

$$\|(u,v)\| = \max(|u|_p, |v|_q) \qquad (\forall (u,v) \in \mathbb{R}^p \times \mathbb{R}^q).$$

Here $\|df(u,v)\|$ and $\|dg(u,v)\|$ both refer to induced operator norms, treating $df(u, v)$ and $dg(u,v)$ as linear mappings. By using the mean-value inequality we obtain

$$\|d^2 g(u,v) - d^2 g(0,0)\| \le \sup_{|u|_p \le \delta, |v|_q \le r} \|d^3 g(u,v)\| \|(u,v)\|,$$

and the triangle inequality gives

$$\|d^2 g(u,v)\| \le \|d^2 g(0,0)\| + \overline{\ell} \|(u,v)\| \qquad (|u|_p \le \delta, |v|_q \le r),$$

where $\overline{\ell}$ $(= \overline{\ell}(\delta, r)) = \sup_{|u|_p \le \delta, |v|_q \le r} \|d^3 g(u,v)\|$, whence

$$\|d^2 \mathsf{g}_\epsilon(u,v)\| \le \epsilon \left(\|d^2 g(0,0)\| + \epsilon \overline{\ell} \|(u,v)\| \right) \qquad (|u|_p \le \delta, |v|_q \le r). \tag{4.5}$$

An analogous inequality holds for f and f_ϵ:

$$\|d^2 \mathsf{f}_\epsilon(u,v)\| \le \epsilon \left(\|d^2 f(0,0)\| + \epsilon \overline{\ell} \|(u,v)\| \right) \qquad (|u|_p \le \delta, |v|_q \le r). \tag{4.6}$$

It is the $O(\epsilon)$ size of these quantities that will be important later.

4.2. Introducing a cutoff function. It is not (4.2) that we shall seek to solve directly, but we must employ a cutoff function to rewrite (4.2) in a fixed-point form that is amenable to a Picard iteration. This is not the case at present because if we were to define a nonlinear operator acting on h by the right-hand side of (4.2), there is no reason for it or its iterates to be well-defined on a suitable function space.

For any $\delta > 0$, there is a cutoff function $\psi \in C^\infty(\mathbb{R}^p)$ such that

$$\psi(u) = \begin{cases} u & \text{if} \quad |u|_p \le \delta/2, \\ 0 & \text{if} \quad |u|_p \ge 3\delta/2 \end{cases}$$

and such that $|\psi(u)|_p \le \delta$. By using this cutoff we define a Nemitskii operator π as follows:

$$\pi(h)(u) = \psi(Cu + \mathsf{f}_\epsilon(u, h(u))) \qquad (\forall u \in \overline{\Omega}_\delta, h : \overline{\Omega}_\delta \subset \mathbb{R}^p \to \mathbb{R}^q), \tag{4.7}$$

so that

$$\pi : C^0(\overline{\Omega}_\delta, \mathbb{R}^q) \to C^0(\overline{\Omega}_\delta, \mathbb{R}^p),$$

and the inequality $|\pi(h)(u)|_p \leq \delta$ holds pointwise. Moreover, by the C^k-regularity of f, it follows that π is itself a C^k-mapping for each $\epsilon > 0$ fixed, with Fréchet derivative

$$d\pi(h)[k](\cdot) = d\psi(C\cdot + \mathsf{f}_\epsilon(\cdot, h))[d_v \mathsf{f}_\epsilon(\cdot, h)[k]] \qquad (\forall h, k \in C^0(\overline{\Omega}_\delta, \mathbb{R}^q)). \tag{4.8}$$

Also note the following estimate, which will be important later.

LEMMA 3. *Suppose that $h \in C^0(\overline{\Omega}_\delta, \mathbb{R}^q)$ satisfies $\|h\|_{C^0} \leq r$, then*

$$\|d\pi(h)\|_{BL(C^0)} \leq \epsilon \ \cdot \|\psi\|_{C^1} \ell \max(\delta, r), \tag{4.9}$$

and there are constants $\kappa_1, \kappa_2 > 0$, depending on ℓ, δ, and r, but not on ϵ, such that

$$\|d^2\pi(h)\|_{BL(C^0)\times BL(C^0)} \leq \epsilon \|\psi\|_{C^2} \left(\kappa_1(\ell, \delta, r) + \epsilon\kappa_2(\ell, \delta, r)\right). \tag{4.10}$$

Proof. By using (4.8) we obtain

$$\begin{aligned}
\sup_{k\in C^0, \|k\|_{C^0}=1} \|d\pi(h)[k]\|_{C^0} &= \sup_{k\in C^0, \|k\|_{C^0}=1} \|d\psi(C\cdot + \mathsf{f}_\epsilon(\cdot, h))[d_v \mathsf{f}_\epsilon(\cdot, h)[k]]\|_{C^0} \\
&\leq \|\psi\|_{C^1} \cdot \sup_{k\in C^0, \|k\|_{C^0}=1} \|d_v \mathsf{f}_\epsilon(\cdot, h)[k]\|_{C^0} \\
&\leq \|\psi\|_{C^1} \cdot \sup_{|u|_p\leq\delta, |v|_q\leq r} \|d_v \mathsf{f}_\epsilon(u, v)\| \\
&\leq \|\psi\|_{C^1} \cdot \ell \cdot \sup_{|u|_p\leq\delta, |v|_q\leq r} \|(\epsilon u, \epsilon v)\| \text{ (by (4.1) and (4.3))},
\end{aligned}$$

and the first part follows. The second part follows from

$$d^2\pi(h)[k_1, k_2] = d^2\psi(C\cdot + \mathsf{f}_\epsilon(\cdot, h))[d_v \mathsf{f}_\epsilon[k_1], d_v \mathsf{f}_\epsilon[k_2]] + d\psi[d^2_{vv} \mathsf{f}_\epsilon(\cdot, h)[k_1, k_2]],$$

so that, for $i = 1, 2$,

$$\begin{aligned}
\sup_{k_i\in C^0, \|k_i\|_{C^0}=1} \|d^2\pi(h)[k_1, k_2]\|_{C^0} &\leq \sup_{u\in\mathbb{R}^p} \|d\psi(u)\| \cdot \|d^2_{vv} \mathsf{f}_\epsilon(\cdot, h)\| \\
&\quad + \sup_{u\in\mathbb{R}^p} \|d^2\psi(u)\| \cdot \|d_v \mathsf{f}_\epsilon(\cdot, h)\|^2 \\
&\leq \|\psi\|_{C^2} \left(\sup_{|u|_p\leq\delta, |v|_q\leq r} \|d^2_{vv} \mathsf{f}_\epsilon(u, v)\| + \sup_{|u|_p\leq\delta, |v|_q\leq r} \|d_v \mathsf{f}_\epsilon(u, v)\|^2 \right) \\
&\leq \|\psi\|_{C^2} \left(\epsilon\|d^2 f(0, 0)\| + \epsilon^2\overline{\ell} \max(\delta, r) + \sup_{|u|_p\leq\delta, |v|_q\leq r} \|d_v f(\epsilon u, \epsilon v)\|^2 \right) \quad \text{(by (4.5))} \\
&\leq \|\psi\|_{C^2} \left(\epsilon\|d^2 f(0, 0)\| + \epsilon^2\overline{\ell} \max(\delta, r) + \sup_{|u|_p\leq\delta, |v|_q\leq r} \|(\epsilon u, \epsilon v)\|^2 \ell^2 \right) \quad \text{(by (4.3))},
\end{aligned}$$

and the result follows directly from here. ∎

In order to solve (4.2), we now tackle the following nonlinear fixed-point problem:

$$h = Nh(\pi(h)) + \mathsf{G}_\epsilon(h), \tag{4.11}$$

where $\mathsf{G}_\epsilon : C^0(\overline{\Omega}_\delta, \mathbb{R}^q) \to C^0(\overline{\Omega}_\delta, \mathbb{R}^q)$ is the Nemitskii operator

$$\mathsf{G}_\epsilon(h)(u) := \mathsf{g}_\epsilon(u, h(u)).$$

Notice that the cutoff ψ has been used in (4.11), but, because ψ coincides with the identity on some balls around the origin, solutions of (4.11) will satisfy (4.2) on this ball.

This construction ensures that $h(\pi(h)) \in C^0(\overline{\Omega}_\delta, \mathbb{R}^q)$ whenever $h \in C^0(\overline{\Omega}_\delta, \mathbb{R}^q)$, and we may, as a result, define the operator

$$T : C^0(\overline{\Omega}_\delta, \mathbb{R}^q) \to C^0(\overline{\Omega}_\delta, \mathbb{R}^q); \quad h \mapsto h(\pi(h)),$$

noting that the operators $T : C^1(\overline{\Omega}_\delta, \mathbb{R}^q) \to C^1(\overline{\Omega}_\delta, \mathbb{R}^q)$ and $T : C^2(\overline{\Omega}_\delta, \mathbb{R}^q) \to C^2(\overline{\Omega}_\delta, \mathbb{R}^q)$ are also well-defined as the restrictions of T to various subspaces of $C^0(\overline{\Omega}_\delta, \mathbb{R}^q)$ as f and g are C^3 functions.

However, it is not (4.11) that we shall solve, but we exploit the nilpotency of N to bring the *functional part* of (4.2) and (4.11), that is, $Nh(\pi(h))$, into a higher-order contribution to the problem. However, because $g(u, v)$ is a second-, or possibly higher-order function, the operator G_ϵ contains no linear terms, and this will help us to obtain a contractive sequence by iterating T.

By way of example, let us suppose that $N \neq 0$ but $N^2 = 0$. If there exists a solution of $h = Nh(\pi(h)) + \mathsf{G}_\epsilon(h)$, there results $Nh = N\mathsf{G}_\epsilon(h)$, and therefore $Nh(\pi(h)) = N\mathsf{G}_\epsilon(h(\pi(h)))$. In this case we find that h must satisfy

$$h = N\mathsf{G}_\epsilon(h(\pi(h))) + \mathsf{G}_\epsilon(h), \tag{4.12}$$

and one observes that the *functional part* of the equation (that is, $h(\pi(h))$) now sits inside a higher-order term (and not a linear one as in (4.11)). Conversely, if h satisfies (4.12), then $Nh = N\mathsf{G}_\epsilon(h)$, so that $N\mathsf{G}_\epsilon(h(\pi(h))) = Nh(\pi(h))$ because $N^2 = 0$, and $h = Nh(\pi(h)) + \mathsf{G}_\epsilon(h)$ follows.

When the nilpotency index of the map N is arbitrary we extend this idea in the following lemma, where here and in the remainder we shall write

$$\mathcal{G}_\epsilon(h) = \sum_{j=0}^{\nu-1} N^j \mathsf{G}_\epsilon(T^j(h)),$$

and

$$T^{j+1} = T(T^j), \qquad \text{where} \qquad T^0 = I,$$

and the latter denotes the identity on $C^0(\overline{\Omega}_\delta, \mathbb{R}^q)$.

LEMMA 4. *Suppose that $N^\nu = 0$ but $N^{\nu-1} \neq 0$; then h is a solution of* (4.11) *if it is a solution of the fixed-point problem*

$$h = \mathcal{G}_\epsilon(h). \tag{4.13}$$

Proof. Let us suppose that h is a solution of (4.13); then

$$Nh = N\mathcal{G}_\epsilon(h) = N\left[\sum_{j=0}^{\nu-1} N^j \mathsf{G}_\epsilon(T^j(h))\right] = \sum_{j=0}^{\nu-1} N^{j+1} \mathsf{G}_\epsilon(T^j(h)),$$

and so, because $N^\nu = 0$, we obtain

$$\begin{aligned} Nh(\pi(h)) &= \sum_{j=0}^{\nu-1} N^{j+1} \mathsf{G}_\epsilon(T^j(h(\pi(h)))) \\ &= \sum_{j=0}^{\nu-1} N^{j+1} \mathsf{G}_\epsilon(T^{j+1}(h)) = \sum_{j=1}^{\nu-1} N^j \mathsf{G}_\epsilon(T^j(h)) = \mathcal{G}_\epsilon(h) - \mathsf{G}_\epsilon(h) = h - \mathsf{G}_\epsilon(h), \end{aligned}$$

which therefore provides a solution of (4.11) as required. ∎

4.3. The main result. Our strategy for solving (4.13), and hence (4.2), is to show that $\mathcal{G}_\epsilon$ satisfies a *refined* Banach contraction theorem of the type given in [26, p. 286]. The idea that we employ several times is encapsulated in the following idea. Consider Banach spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ such that $x_0 \in Y$ and $Y \subset X$, and moreover $\mathcal{T}$ is a mapping satisfying $\mathcal{T} : Y \to Y$ and $\mathcal{T} : X \to X$. Now, if there is an $r > 0$ and a $\kappa \in [0, 1)$ such that

1. $\mathcal{T} : \overline{B}_r(x_0; X) \to \overline{B}_r(x_0; X)$,
2. $\|\mathcal{T}(y) - \mathcal{T}(y')\|_X \leq \kappa \|y - y'\|_X$ for all $y, y' \in Y \cap \overline{B}_r(x_0; X)$,

then $\mathcal{T}$ has a fixed point $y^* \in \overline{B}_r(x_0; X)$. In addition, if

3. $\mathcal{T} : \overline{B}_\rho(x_0; Y) \to \overline{B}_\rho(x_0; Y)$ and
4. there is an interpolating Banach space Z such that $Y \subset Z$ with compact embedding and $Z \subset X$ with continuous embedding,

then $y^* \in Z$.

The point here is that we cannot ensure that $y^* \in Y$, although one still obtains a fixed point in some space from the standard iteration scheme. By using this idea one can prove an existence and regularity result for (4.2), where we have in mind $x_0 = 0, \mathcal{T} = \mathcal{G}_\epsilon, Y = C^{k+1}, Z = C^{k+\alpha}$, and $X = C^k$, where $\alpha \in (0, 1)$. We begin by providing the details to cover the cases $k = 0$ and $k = 1$.

The following theorem is the main result of this paper from which the invariant manifold and bifurcation theorems are deduced.

THEOREM 1. *Let $\alpha \in [0, 1)$. There exists an $\epsilon_0 > 0$ such that, for each $\epsilon \in (0, \epsilon_0)$,* (4.13) *has a solution $h \in C^{1+\alpha}(\overline{\Omega}_\delta, \mathbb{R}^q)$; moreover $h(0) = 0$ and $dh(0) = 0$.*

Proof. Let X_r be the C^0-closed ball of radius r about zero in $C^0(\overline{\Omega}_\delta, \mathbb{R}^q)$ and Y_r the C^1-closed ball of radius r about zero in $C^1(\overline{\Omega}_\delta, \mathbb{R}^q)$. Let $h_0 \in Y_r$, and define a sequence

$$h_{n+1} := \mathcal{G}_\epsilon(h_n).$$

We shall show that we can choose ϵ such that (h_n) is well-defined, contractive, and hence Cauchy in X_r, and it therefore converges. Throughout the remainder of the proof we shall use the positive constant

$$n_* := \sum_{j=0}^{\nu-1} \|N\|^j_{BL(\mathbb{R}^q)}.$$

We now give a proof of Theorem 1 in four short steps, each placing a stronger restriction on ϵ relative to the fixed choice of δ and r to ensure that $\mathcal{G}_\epsilon$ contracts when acting on C^1 functions, measured in the C^0 norm.

Step 1. If $\epsilon < r(\ell n_* \max(\delta, r)^2)^{-1} =: \epsilon_1$, then $\mathcal{G}_\epsilon : X_r \to X_r$.

Proof. Suppose that $h \in X_r$, and then $T(h) = h(\pi(h))$ is continuous as h is. Moreover if $\|h\|_{C^0} \leq r$, then $\|T(h)\|_{C^0} = \|h(\pi(h))\|_{C^0} \leq r$; similarly $\|T^j(h)\|_{C^0} \leq r$ for all $0 \leq j \leq \nu - 1$. By definition,

$$\begin{aligned}\|\mathcal{G}_\epsilon(h)\|_{C^0} &\leq \sum_{j=0}^{\nu-1} \|N\|^j \|\mathsf{G}_\epsilon(T^j(h))\|_{C^0} \leq \sum_{j=0}^{\nu-1} \|N\|^j \sup_{\|\overline{h}\|_{C^0} \leq r} \|\mathsf{G}_\epsilon(\overline{h})\|_{C^0} \\ &\leq n_* \sup_{|u|_p \leq \delta, |v|_q \leq r} |\mathsf{g}_\epsilon(u, v)|_q = n_* \sup_{|u|_p \leq \delta, |v|_q \leq r} \epsilon^{-1} |g(\epsilon u, \epsilon v)|_q.\end{aligned}$$

From (4.3) it follows that

$$\|\mathcal{G}_\epsilon(h)\|_{C^0} \leq \epsilon\, \ell n_* \sup_{|u|_p \leq \delta, |v|_q \leq r} \|(u, v)\|^2 = \epsilon\, \ell n_* \max(\delta, r)^2.$$

By assumption, $\epsilon \ell n_* \max(\delta, r)^2 < r$, and Step 1 is complete. ∎

Step 2. If $\epsilon < \min\{\epsilon_1, [r\ell \max(\delta, r)\|\psi\|_{C^1}]^{-1}, r[n_*(1+r)\ell \max(\delta, r)]^{-1}\} =: \epsilon_2$, then $\mathcal{G}_\epsilon : Y_r \to Y_r$.

Proof. Suppose that $h \in Y_r$ so that $h \in X_r$, then $T^j(h)$ is differentiable on $\overline{\Omega}_\delta$ as h is; moreover throughout the remainder of the proof of Step 2 we shall write $H := \mathcal{G}_\epsilon(h)$ for brevity. It then follows that

$$dH(u) = \sum_{j=0}^{\nu-1} N^j \left[d_u \mathbf{g}_\epsilon(u, T^j(h)(u)) + d_v \mathbf{g}_\epsilon(u, T^j(h)(u))[d_u(T^j(h))(u)] \right], \tag{4.14}$$

and we now need to estimate $\|dH(u)\|_{C^0}$. By using the fact that $\|h\|_{C^0} \le r$, since $T : X_r \to X_r$ from Step 1, we obtain $|T^j(h)(u)|_q = |h(\pi(h(\ldots))\ldots)|_q \le r$, and therefore

$$\begin{aligned}
\sup_{|u|_p \le \delta} \|dH(u)\|_{BL(\mathbb{R}^p, \mathbb{R}^q)} &\le \sum_{j=0}^{\nu-1} \|N\|^j \left(\|d_u \mathbf{g}_\epsilon(u, T^j(h)(u))\| \right. \\
&\qquad \left. + \|d_v \mathbf{g}_\epsilon(u, T^j(h)(u))\| \|d_u(T^j(h))(u)\| \right) \\
&\le n_* \sup_{0 \le j \le \nu-1} \left(\|d_u \mathbf{g}_\epsilon(u, T^j(h)(u))\| + \|d_v \mathbf{g}_\epsilon(u, T^j(h)(u))\| \|d_u(T^j(h))(u)\| \right) \\
&\le n_* \sup_{\substack{0 \le j \le \nu-1 \\ |u|_p \le \delta, |v|_q \le r}} \left(\|d_u \mathbf{g}_\epsilon(u, v)\| + \|d_v \mathbf{g}_\epsilon(u, v)\| \|d_u(T^j(h))(u)\| \right) \\
&\le \epsilon\ n_* \ell \max(r, \delta) \left(1 + \sup_{0 \le j \le \nu-1, |u|_p \le \delta} \|d_u(T^j(h))(u)\| \right).
\end{aligned}$$

Now we estimate the final bracketed term in the latter expression. The linear mapping obtained from differentiating $T^j(h)(u)$ with respect to u is

$$d_u(T^j(h))(u) = d_u(h(\pi(h(\ldots \pi(h) \ldots)))),$$

which can be written as the recurrence

$$d_u(T^j(h))(u) = dh(\pi(T^{j-1}(h))(u)) d\pi(T^{j-1}(h)(u)) \cdot d_u(T^{j-1}(h)(u)), \tag{4.15}$$

where, by definition, $d_u(T^0(h))(u) = d_u(h)(u) = dh(u)$. By taking C^0 norms and setting

$$\xi_j := \|d_u(T^j(h))(u)\|_{BL(\mathbb{R}^p, \mathbb{R}^q)},$$

we obtain the relation

$$\xi_j \le \|dh\|_{C^0} \sup_{\|\overline{h}\|_{C^0} \le r} \|d\pi(\overline{h})\|_{BL(C^0)} \cdot \xi_{j-1},$$

where $\xi_0 = \|dh\|_{C^0} \le \|h\|_{C^1} \le r$. From (4.9) of Lemma 3, we find that

$$\xi_j \le \epsilon \ell r \max(\delta, r) \|\psi\|_{C^1} \cdot \xi_{j-1} \le (\epsilon \ell r \max(\delta, r) \|\psi\|_{C^1})^j\ \xi_0 \le r$$

because $\epsilon \ell r \max(\delta, r) \|\psi\|_{C^1} < 1$ by assumption. As a result, the inequality

$$\begin{aligned}
\sup_{|u|_p \le \delta} \|dH(u)\|_{BL(\mathbb{R}^p, \mathbb{R}^q)} &\le \epsilon\ n_* \ell \max(r, \delta) \left(1 + \sup_{0 \le j \le \nu-1} \xi_j \right) \\
&\le \epsilon n_* \ell \max(r, \delta)(1 + r) \le r
\end{aligned}$$

also now follows from the assumption of the claim, and we have proven that $h \in Y_r$, as required. ∎

Step 3. If $\epsilon < \epsilon_2$ and we define the $O(\epsilon)$ quantity $\kappa(\epsilon)$ by

$$\kappa(\epsilon) := \epsilon \cdot \ell \max(\delta, r) \sum_{i=0}^{\nu-1} (r\epsilon \cdot \|\psi\|_{C^1} \ell \max(\delta, r))^i,$$

then $\|h_{n+1} - h_n\|_{C^0} \le \kappa(\epsilon)\|h_n - h_{n-1}\|_{C^0}$. As a result, (h_n) is Cauchy in X_r if ϵ is further restricted so that $\kappa(\epsilon) < 1$, and so (h_n) converges in X_r.

Proof. For brevity, let us write h in place of h_{n+1} and k for h_n after setting $h_0 = 0 \in Y_r$ and $h_{n+1} = \mathcal{G}_\epsilon(h_n)$. Now

$$\begin{aligned}\|\mathcal{G}_\epsilon(h) - \mathcal{G}_\epsilon(k)\|_{C^0} &\le \sum_{j=0}^{\nu-1} \|N\|^j \|\mathsf{G}_\epsilon(T^j(h)) - \mathsf{G}_\epsilon(T^j(k))\|_{C^0} \\ &\le n_* \sup_{0\le j\le \nu-1} \|\mathsf{g}_\epsilon(u, T^j(h)(u)) - \mathsf{g}_\epsilon(u, T^j(k)(u))\|_{C^0}.\end{aligned}$$

The mean-value inequality yields

$$\begin{aligned}&|\mathsf{g}_\epsilon(u, T^j(h)(u)) - \mathsf{g}_\epsilon(u, T^j(k)(u))|_q \\ &\qquad \le \sup_{z\in[T^j(h)(u), T^j(k)(u)]} \|d_v \mathsf{g}_\epsilon(u, z)\| \|T^j(h) - T^j(k)\|_{C^0} \\ &\qquad \le \sup_{|u|_p\le\delta, |v|_q\le r} \|d_v \mathsf{g}_\epsilon(u, v)\| \|T^j(h) - T^j(k)\|_{C^0} \qquad \text{(using Step 1)} \\ &\qquad \le \epsilon\ell \max(r, \delta) \cdot \|T^j(h) - T^j(k)\|_{C^0} \qquad \text{(by (4.3))},\end{aligned} \tag{4.16}$$

where, for any $z_1, z_2 \in \mathbb{R}^p$, the generalized interval from z_1 to z_2 is given by

$$[z_1, z_2] := \{\lambda z_1 + (1-\lambda) z_2 : 0 \le \lambda \le 1\}.$$

So let us define $\chi_j := \|T^j(h) - T^j(k)\|_{C^0}$, and we estimate χ_j as follows:

$$\begin{aligned}|T^j(h)(u) - T^j(k)(u)|_q &= |h(\pi(T^{j-1}(h))) - k(\pi(T^{j-1}(k)))|_q \\ &\le |h(\pi(T^{j-1}(h))) - k(\pi(T^{j-1}(h)))|_q \\ &\qquad + |k(\pi(T^{j-1}(h))) - k(\pi(T^{j-1}(k)))|_q \\ \implies \qquad \chi_j &\le \|h - k\|_{C^0} + \|k(\pi(T^{j-1}(h))) - k(\pi(T^{j-1}(k)))\|_{C^0}.\end{aligned}$$

However, from Step 2 we have $\|k\|_{C^1} \le r$, and therefore

$$|k(u) - k(u')|_q \le r|u - u'|_p \qquad \forall u, u' \in \overline{\Omega}_\delta.$$

Using the fact that π is a Fréchet differentiable mapping on $C^0(\overline{\Omega}_\delta, \mathbb{R}^q)$ with norm bounded according to (4.9), an application of the mean-value inequality gives

$$\chi_j \le \|h - k\|_{C^0} + r \sup_{\|\overline{h}\|_{C^1}\le r} \|d\pi(\overline{h})\|_{BL(C^0)} \cdot \chi_{j-1},$$

so that $\chi_j \le \|h - k\|_{C^0} + \epsilon \cdot r\|\psi\|_{C^1} \ell \max(\delta, r)\chi_{j-1}$, where $\chi_0 = \|h - k\|_{C^0}$. The discrete Gronwall inequality now gives

$$\chi_j \le \|h - k\|_{C^0} \sum_{i=0}^{\nu-1} (r\epsilon \|\psi\|_{C^1} \ell \max(\delta, r))^i,$$

and with (4.16) we have the desired inequality

$$\|\mathcal{G}_\epsilon(h) - \mathcal{G}_\epsilon(k)\|_{C^0} \le \kappa(\epsilon) \cdot \|h - k\|_{C^0}.$$

The standard contraction argument now shows that $(h_n) \subset Y_r$ is Cauchy in X_r as claimed, and there therefore exists an $h \in X_r$ such that $h_n \overset{C^0}{\to} h$ as $n \to \infty$. ∎

Since $(h_n) \subset Y_r$, it follows that there is a subsequence (n_j) such that $h_{n_j} \overset{C^\alpha}{\to} \overline{h}$ as $j \to \infty$ for some $\overline{h} \in X_r \cap C^\alpha(\overline{\Omega}_\delta)$, and we deduce that $\overline{h} = h \in C^\alpha(\overline{\Omega}_\delta)$. By restricting ϵ further we can actually ensure that h is differentiable, as follows.

Step 4. There is an $\epsilon_3 > 0$ such that $h \in C^{1+\alpha}(\overline{\Omega}_\delta)$ whenever $\epsilon < \min(\epsilon_2, \epsilon_3)$ and $\kappa(\epsilon) < 1$.

Proof. Let $h \in C^2(\overline{\Omega}_\delta)$ satisfy $\|h\|_{C^2} \le r$, and recall that $H := \mathcal{G}_\epsilon(h)$. From (4.14) we obtain

$$\begin{aligned} d^2H(u) = \sum_{j=0}^{\nu-1} N^j \{ & d^2_{uu}\mathrm{g}_\epsilon(u, T^j(h)(u)) + 2d^2_{uv}\mathrm{g}_\epsilon(u, T^j(h)(u))[I, d_u(T^j(h)(u))] \\ & + d^2_{vv}\mathrm{g}_\epsilon(u, T^j(h)(u))[d_u(T^j(h)(u)), d_u(T^j(h)(u))] \\ & + d_v\mathrm{g}_\epsilon(u, T^j(h)(u))[d^2_{uu}(T^j(h)(u))]\}. \end{aligned}$$

In seeking a bound on $\|H\|_{C^2}$, we now examine the term $d^2_{uu}(T^j(h)(u))$ more closely as bounds on the remaining elements of d^2H can be obtained from Steps 1 and 2. By applying the chain rule (4.15) we obtain the recurrence

$$\begin{aligned} d^2_{uu}(T^j(h)(u)) &= d_u\left\{dh(\pi(T^{j-1}(h))(u))d\pi(T^{j-1}(h)(u)) \cdot d_u(T^{j-1}(h)(u))\right\} \\ &= d^2h(\pi(T^{j-1}(h))(u))[d\pi(T^{j-1}(h)(u)) \cdot d_u(T^{j-1}(h)(u))]^{(2)} \\ &\quad + dh(\pi(T^{j-1}(h))(u))[d^2\pi(T^{j-1}(h)(u))[d_u(T^{j-1}(h)(u))]^{(2)}] \\ &\quad + dh(\pi(T^{j-1}(h))(u))[d\pi(T^{j-1}(h)(u))[d^2_{uu}(T^{j-1}(h)(u))]], \end{aligned}$$

and taking norms gives

$$\begin{aligned} \|d^2_{uu}(T^j(h)(u))\|_{C^0} \le & \|d^2h\|_{C^0}\|d\pi(h)\|^2\|d_u(T^j(h)(u))\|_{C^0} \\ & + \|dh\|_{C^0}\|d^2\pi(h)\|\|d_u(T^j(h)(u))\|^2_{C^0} \\ & + \|dh\|_{C^0}\|d\pi(h)\|\|d^2_{uu}(T^j(h)(u))\|_{C^0}. \end{aligned}$$

If we write

$$\eta_j := \|d^2_{uu}(T^j(h)(u))\|_{C^0},$$

and use the fact that $\xi_j = \|d_u(T^j(h))(u)\|_{BL(\mathbb{R}^p,\mathbb{R}^q)} \le r$ for all $0 \le j \le \nu - 1$, which was established in Step 2, we obtain the difference inequality

$$\eta_j \le r^3\|d\pi(h)\|^2_{C^0} + r^2\|d^2\pi(h)\|_{C^0} + r\|d\pi\|_{C^0} \cdot \eta_{j-1}$$

such that $\eta_0 \le r$ by definition. There results, for $j \ge 1$,

$$\eta_j \le (r\|d\pi(h)\|_{C^0})^j\,\eta_0 + \left(r^3\|d\pi(h)\|^2_{C^0} + r^2\|d^2\pi(h)\|_{C^0}\right)\sum_{i=0}^{j-1}(r\|d\pi(h)\|_{C^0})^i,$$

and the bounds (4.9) and (4.10) show that $\eta := \max_{0\le j\le\nu-1}\eta_j$ has an $O(1)$ dependence on ϵ in the sense that there is an $M > 0$ such that $\eta \le M$ whenever $0 \le \epsilon \le 1$. (In fact, one can choose M to be less than r if ϵ is sufficiently small.)

We now obtain

$$\begin{aligned}\|d^2H\|_{C^0} &\leq n_* \sup_{0\leq j\leq \nu-1} \big\{ \|d^2_{uu}\mathsf{g}_\epsilon(u,T^j(h)(u))\|_{C^0} + 2\xi_j\|d^2_{uv}\mathsf{g}_\epsilon(u,T^j(h)(u))\|_{C^0} \\ &\qquad + \xi_j^2\|d^2_{vv}\mathsf{g}_\epsilon(u,T^j(h)(u))\|_{C^0} + \eta_j\|d_v\mathsf{g}_\epsilon(u,T^j(h)(u))\|_{C^0}\big\} \\ &\leq n_* \sup_{\|\overline{h}\|_{C^0}\leq r, 0\leq j\leq \nu-1} \big\{ \|d^2_{uu}\mathsf{g}_\epsilon(u,\overline{h})\|_{C^0} + 2r\|d^2_{uv}\mathsf{g}_\epsilon(u,\overline{h})\|_{C^0} \\ &\qquad + r^2\|d^2_{vv}\mathsf{g}_\epsilon(u,\overline{h})\|_{C^0} + \eta\|d_v\mathsf{g}_\epsilon(u,\overline{h})\|_{C^0}\big\}, \end{aligned} \tag{4.17}$$

where η has been used to bound the last term in (4.17). By using (4.5) to estimate the second derivative terms we find that

$$\begin{aligned}\|d^2H\|_{C^0} \leq \epsilon\cdot n_* \sup_{0\leq j\leq \nu-1} \big\{ &\|d^2g(0,0)\| + \epsilon\overline{\ell}\max(r,\delta) \\ &+ 2r(\|d^2g(0,0)\| + \epsilon r\overline{\ell}\max(r,\delta)) \\ &+ r^2(\|d^2g(0,0)\| + \epsilon\overline{\ell}\max(r,\delta)) + \eta\ell\max(r,\delta)\big\}. \end{aligned} \tag{4.18}$$

It is immediate from (4.18) and Steps 1 and 2 that a suitably small choice of ϵ ensures that $\|H\|_{C^2}\leq r$ whenever $\|h\|_{C^2}\leq r$. As a result, if we impose the following restriction on the initial guess for a fixed point of $\mathcal{G}_\epsilon$:

$$h_0 \in Z_r := \{h\in C^2(\overline{\Omega}_\delta,\mathbb{R}^q) : \|h\|_{C^2}\leq r\},$$

then the C^0-convergent sequence from Step 3 also satisfies $(h_n)\subset Z_r$. This means that a $C^{1+\alpha}$-convergent subsequence can now be extracted from (h_n), so that the convergence of h_n to h actually occurs in $C^{1+\alpha}$ and h therefore lies in this smoother space. □

We have shown that there is a differentiable solution of (4.2) on a sufficiently small ball around the origin, but there remains to prove the last part of Theorem 1 regarding the behavior of h at the origin. So let h be a C^1 solution of (4.2) on some ball $\overline{\Omega}_\delta$, and put $\zeta = h(0)$. It follows that ζ is a solution of the algebraic equation

$$-\zeta + Nh(\mathsf{f}_\epsilon(0,\zeta)) + \mathsf{g}_\epsilon(0,\zeta) = 0, \tag{4.19}$$

and (4.19) has solution $\zeta = 0$. As the linearization of the left-hand side of (4.19) at $\zeta = 0$ is a multiple of the identity, the inverse function theorem ensures that $\zeta = 0$ is the only solution of (4.19) in some neighborhood of zero, and this ensures that $h(0) = 0$. Since any solution of (4.13) provides one of (4.2), we can differentiate (4.2) with respect to u and set $u = 0$; this gives

$$dh(0) = Ndh(0)[C],$$

but then we can continue in an inductive manner to deduce that

$$Ndh(0)[C] = N^2dh(0)[C^2] = \cdots = N^\nu dh(0)[C^\nu] = 0,$$

as N is nilpotent. We find that $dh(0) = 0$, and this concludes the proof of Theorem 1. □

The question of maximal regularity of a solution of (4.2), or equivalently (4.13), is not addressed, although the method of proof used in Theorem 1 can be continued by restricting ϵ further as required to show that $\mathcal{G}_\epsilon$ maps the ball of radius r in $C^j(\overline{\Omega}_\delta,\mathbb{R}^p)$

into itself. This ensures that the sequence (h_n) constructed in the proof of Theorem 1 converges to h in as strong a C^j norm as we like, provided that f and g are sufficiently smooth. We cannot, however, be sure that the resulting solution h lies in $C^\infty(\overline{\Omega}_\delta, \mathbb{R}^p)$ because the interval in which ϵ must reside so as to obtain a C^j fixed point of $\mathcal{G}_\epsilon$ could shrink indefinitely as j grows.

Theorem 1 does not ensure the existence of a continuous fixed point of (4.2) which is not C^α for some $0 < \alpha < 1$. If we take a continuous initial guess for a solution, $h_0 \in C^0(\overline{\Omega}_\delta, \mathbb{R}^q)$, say, then, although there is a sequence of continuous functions defined by $h_{n+1} = \mathcal{G}_\epsilon(h_n)$, there is no reason to suspect that the sequence of iterates (h_n) will satisfy the property of being a C^0-contractive sequence. On the other hand, a suitably small C^1 initial guess will lead to C^α convergence of the resulting iterates for any $\alpha \in [0, 1)$.

4.4. Stability implies uniqueness. There are some simple cases where uniqueness and smoothness can be easily established. The most obvious is where $N = 0$, and then Theorem 1 can be proven using the elementary implicit function theorem. Another occurs when the matrix denoted C that arises from the Kronecker normal form of $(\mathcal{A}, \mathcal{B})$ in (3.5) satisfies $\|C\| < 1$ in the norm induced by $|\cdot|_p$. In this case the cutoff function ψ used above is not needed in order to obtain a well-defined operator π. If we define the Nemitskii operator

$$\pi(h)(u) = Cu + \mathsf{f}_\epsilon(u, h(u)),$$

and $h \in C^0(\overline{\Omega}_\delta, \mathbb{R}^q)$ satisfies $\|h\|_{C^0} \le r$, then $|Cu + \mathsf{f}_\epsilon(u, h(u))|_p \le \|C\||u|_p + |\mathsf{f}_\epsilon(u, h(u))|_p \le \|C\|\delta + \epsilon \cdot \ell \max(\delta, r)^2$. In order for $h(\pi(h))$ to be well-defined, we now need only to choose ϵ such that $\|C\|\delta + \epsilon \cdot \ell \max(\delta, r)^2 \le \delta$, which can be done. The proof of Theorem 1 then goes through with this minor modification, and the resulting fixed point of $\mathcal{G}_\epsilon$ is unique in the space of continuously differentiable functions.

4.5. Polynomial approximation. Let us remove the dependence of (4.2) on ϵ for clarity and return to the fixed-point problem (3.6) directly, which we recall defines a nonlinear operator G via

$$h(u) = Nh(Cu + f(u, h(u))) + g(u, h(u)) =: (Gh)(u), \quad \text{where } h(0) = 0, dh(0) = 0. \tag{4.20}$$

PROPOSITION 1. *Suppose that $h \in C^k(\overline{\Omega}_\delta, \mathbb{R}^q)$ is a solution of* (4.20), *and suppose that $\overline{h} \in C^k(\overline{\Omega}_\delta, \mathbb{R}^q)$ satisfies*

$$\overline{h}(u) - G(\overline{h})(u) = e(u), \quad \overline{h}(0) = 0,$$

where e is a given C^k function that satisfies $e(0) = 0, de(0) = 0, \ldots, d^k e(0) = 0$. Then

$$|h(u) - \overline{h}(u)|_q = o(|u|_p^k) \text{ as } u \to 0.$$

Proof. If we define the function $\Delta := h - \overline{h}$, then we have to prove that

$$\Delta(0) = 0, d\Delta(0) = 0, \ldots, d^k\Delta(0) = 0,$$

and the result then follows from the basic properties of the derivative. Clearly $\Delta(0) = 0$, and differentiating (4.20) with respect to u gives

$$\begin{aligned} dh(u) = Ndh(Cu + f(u, h(u)))[C + d_u f(u, h(u)) + d_v f(u, h(u))[dh(u)]] \\ + d_u g(u, h(u)) + d_v g(u, h(u))[dh(u)], \end{aligned}$$

and similarly,

$$d\overline{h}(u) = Nd\overline{h}(Cu + f(u, \overline{h}(u)))[C + d_u f(u, \overline{h}(u)) + d_v f(u, \overline{h}(u))[d\overline{h}(u)]] \\ + d_u g(u, \overline{h}(u)) + d_v g(u, \overline{h}(u))[d\overline{h}(u)] + de(u).$$

As a result, because $h(0) = \overline{h}(0) = 0$, we find that

$$d\Delta(0) = Nd\Delta(0)[C] + de(0) = Nd\Delta(0)[C],$$

by the assumption that $de(0) = 0$, so that

$$d\Delta(0) = Nd\Delta(0)[C] = N^2 d\Delta(0)[C^2] = \cdots = N^\nu d\Delta(0)[C^\nu] = 0.$$

Continuing in a similar vein, we obtain

$$d^2h(u) = Nd^2h(Cu + f(u, h(u)))[C + d_u f(u, h(u)) \\ + d_v f(u, h(u))[dh(u)]]^{(2)} + d^2_{uu} g(u, h(u)) + 2d^2_{uv} g(u, h(u))[I, dh(u)] \\ + d^2_{vv} g(u, h(u))[dh(u), dh(u)] + d_v g(u, h(u))[d^2h(u)],$$

with a similar expression for $d^2\overline{h}(u)$, with the additional presence of the term $d^2e(u)$. We find that

$$d^2h(0) = Nd^2h(0)[C, C] + d^2_{uu} g(0, 0), \tag{4.21}$$

and, because $d^2e(0) = 0$, (4.21) also holds with $h(0)$ replaced by $\overline{h}(0)$. We deduce that the bilinear form $d^2\Delta(0)$ satisfies

$$d^2\Delta(0)[X, Y] = Nd^2\Delta(0)[CX, CY] \qquad (\forall X, Y \in \mathbb{R}^p),$$

so that

$$d^2\Delta(0)[X, Y] = N^\nu d^2\Delta(0)[C^\nu X, C^\nu Y] = 0 \qquad (\forall X, Y \in \mathbb{R}^p).$$

We omit the details, but by continuing inductively and assuming $d^j e(0) = 0$, one obtains the result that the j-linear form $d^j\Delta(0)$ satisfies

$$d^j\Delta(0)[X_1, X_2, \ldots, X_j] = Nd^j\Delta(0)[CX_1, CX_2, \ldots, CX_j]$$

for all $X_1, \ldots, X_j \in \mathbb{R}^p$. The nilpotency of N now ensures that the latter quantity is zero. □

A simple corollary to Proposition 1 is that if (4.20) has two infinitely differentiable solutions h and $\overline{h}$ defined on some neighborhood of zero, then they agree beyond all orders at zero:

$$\lim_{u \to 0} \frac{|h(u) - \overline{h}(u)|_q}{|u|_p^k} = 0 \qquad (\forall k \geq 1).$$

5. Applications.

5.1. Nonlinear normal form. The first application of Theorem 1 is the following result, which says that (1.1) induces a local dynamical system on a manifold in the sense of Definition 2. Recall the definition of the matrix pencil (A, B) via

$$(A, B) := (d_{\overline{z}}F(0,0), d_z F(0,0)),$$

where F has $(z, \overline{z})$ as its argument, and note that the following matrix pencil defined on $\mathbb{R}^{2m}$:

$$(\mathcal{A}, \mathcal{B}) := \left(\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & I \\ B & A \end{pmatrix} \right),$$

satisfies $\sigma(\mathcal{A}, \mathcal{B}) = -\sigma(A, B)$.

THEOREM 2 (nonlinear Kronecker normal form). *Let $\alpha \in [0, 1)$, and suppose that* (A1)–(A3) *hold; then there is a linear space K_1 and a $C^{1+\alpha}$-manifold $\mathcal{M}$ modeled on K_1 such that* $\dim(K_1) = \#\sigma(A, B)$ *and*

1. *for all $(z, w) \in \mathcal{M}$ there results $F(z, w) = 0$,*
2. *there is an r' such that for each $(z, w) \in \mathcal{M}$, with $\|(z, w)\| < r'$, there is a unique $(\overline{z}, \overline{w}) \in \mathcal{M}$ such that $w = \overline{z}$, and*
3. *there is a $C^{1+\alpha}$-diffeomorphism $\theta : K_1 \to \mathcal{M}$ such that if $(z, w) = \theta(u)$ for some $u \in K_1$, then $(\overline{z}, \overline{w}) = \theta(\varphi(u))$, where*

$$\varphi(u) = Cu + p(u)$$

for some linear map $C : K_1 \to K_1$ that satisfies $\sigma(C) = \sigma(A, B)$. Moreover, $p : B_\delta(0; K_1) \to K_1$ satisfies $p(0) = 0$ and $d_u p(0) = 0$.

Proof. From Theorem 1, first identify the linear space K_1 from the KNF of $(\mathcal{A}, \mathcal{B})$ with $\mathbb{R}^p$ and K_2 with $\mathbb{R}^q$. Now define the $C^{1+\alpha}$-graph

$$\hat{\mathcal{M}} := \{(u, h(u)) \in \mathbb{R}^p \oplus \mathbb{R}^q : u \in \Omega_\delta\},$$

and let the r-neighborhood of zero in $\hat{\mathcal{M}}$ be $\hat{\mathcal{M}}_r := \{(u, h(u)) \in \mathbb{R}^p \oplus \mathbb{R}^q : |u|_p < r\}$ whenever $r < \delta$.

As a consequence of Theorem 1, there is an $r > 0$ such that for each $(u, v) = (u, h(u)) \in \hat{\mathcal{M}}_r$ we can find a pair $(\overline{u}, \overline{v})$ given by $(\overline{u}, h(\overline{u})) \in \hat{\mathcal{M}}$ such that (NF) is satisfied:

$$\overline{u} = Cu + f(u, v), \quad N\overline{v} = v - g(u, v),$$

where C is obtained from the KNF of $(\mathcal{A}, \mathcal{B})$ so that $\sigma(C) = -\sigma(\mathcal{A}, \mathcal{B}) = \sigma(A, B)$.

Let $\varphi(u) := Cu + f(u, h(u))$, and for each $u \in \mathbb{R}^p$ of sufficiently small norm set

$$\theta(u) := Q(u, h(u)) \qquad \text{and} \qquad \mathcal{M} \equiv Q(\hat{\mathcal{M}}), \quad \mathcal{M}_r \equiv Q(\hat{\mathcal{M}}_r),$$

where the linear map Q is taken from the KNF of $(\mathcal{A}, \mathcal{B})$ from the discussion that immediately follows (3.4). The map θ then provides a local diffeomorphism between Ω_δ and $\mathcal{M}$; moreover both

$$F(\theta(u)) = 0 \qquad \text{and} \qquad F(\theta(\varphi(u))) = 0$$

follow by the construction of h. The following diagram illustrates how a map is induced on $\mathcal{M}$ in this way:

$$\begin{array}{ccc} \mathcal{M}_r & \longrightarrow & \mathcal{M} \\ \theta^{-1}\downarrow & & \uparrow\theta \\ \Omega_r & \xrightarrow{\varphi} & \Omega_\delta \end{array},$$

where $\theta^{-1}(Q(u,h(u))) = u$, and this concludes the proof. ∎

The following immediate corollaries of Theorem 1 provide some information regarding the stable and unstable behavior of orbits in a neighborhood of a fixed point of (1.1).

COROLLARY 1. *Suppose that F satisfies* (A1)–(A3) *and (A,B) is a regular matrix pencil with $\rho(A,B)<1$; then there is a $C^{1+\alpha}$-solution manifold $\mathcal{M}$ of* (1.1) *containing* 0 *such that each $(z,\overline{z})\in\mathcal{M}$ supports a global orbit $(z_n)_{n=0}^{\infty}$ with $(z_n,z_{n+1})\in\mathcal{M}$, $z_0=z$, $z_1=\overline{z}$, and $\lim_{n\to\infty} z_n = 0$.*

Proof. The orbit is constructed by iterating the map $\varphi(u) = Cu + p(u)$ given in part 3 of Theorem 2: Because $\rho(A,B)<1$ we have $\rho(C)<1$ so that φ is a contraction near the origin in some norm, and the result follows. ∎

The following are the natural definitions of stable and unstable sets associated with fixed points of (1.1); note that they are subsets of $\mathbb{R}^{2m}$ and not $\mathbb{R}^m$.

DEFINITION 4. *The set*

$$W^s_{\mathrm{loc}}(0) := \{(z,\overline{z})\in\mathbb{R}^{2m} : \exists \text{ global orbit } (z_n)_{n=0}^{\infty}, z_0=z, z_1=\overline{z}, \lim_{n\to\infty} z_n = 0\}$$

is the local stable set associated with the zero fixed point of (1.1), *and*

$$W^u_{\mathrm{loc}}(0) := \{(z,\overline{z})\in\mathbb{R}^{2m} : \exists \text{ global orbit } (z_n)_{n=0}^{-\infty}, z_{-1}=z, z_0=\overline{z}, \lim_{n\to-\infty} z_n = 0\}$$

is the local unstable set.

In case $\sigma(A,B)$ contains elements outside the unit disk, one can apply the stable manifold theorem for maps to φ in Theorem 2 to give the following result.

COROLLARY 2. *Suppose that* (A1)–(A3) *are satisfied and (A,B) possesses n_s eigenvalues in the open unit disk; then* (1.1) *possesses a subset of the local stable set which is a differentiable manifold of dimension n_s.*

There is an analogous corollary to show that the unstable set is nonempty and contains a manifold of dimension n_u, where n_u is the number of elements of $\sigma(A,B)$ lying outside the closed unit disk. This result is obtained by applying Corollary 2 to (1.1) but with time running *backwards*.

COROLLARY 3. *Suppose that* (A1)–(A3) *are satisfied and (B,A) possesses n_u eigenvalues in the open unit disk; then* (1.1) *possesses a subset of the local unstable set which is a differentiable manifold of dimension n_u.*

Proof. Let us rewrite (1.1) in the form

$$F(z_{n-1}, z_n) = 0 \tag{5.1}$$

to emphasize the fact that we are seeking an orbit that propagates backwards in time, with (z_{-1}, z_0) given. The linearization of (5.1) is of the form

$$Bz_{n-1} + Az_n$$

and if $\det B \neq 0$, then we may locally solve (5.1) for $z_{n-1} = f(z_n)$, and then one can apply the stable manifold theorem for maps to this.

On the other hand if $\det B = 0$, then conditions (A1)–(A3), appropriately modified by exchanging the roles of z and $\overline{z}$ because time is flowing backwards, still apply to (5.1) because (B, A) is a regular matrix pencil due to the fact that (A, B) is regular. The result then follows from Corollary 2. ∎

As A is a singular mapping it follows that the finite spectrum of (B, A) satisfies

$$\sigma(B, A) = (\sigma(A, B)\backslash\{0\})^{-1} \cup \{0\},$$

whether or not B is singular, and hence $\sigma(B, A)$ contains zero so that $n_u \geq 1$. An unstable manifold therefore always exists for (1.1) under conditions (A1)–(A3).

5.2. Bifurcation theorems. We now consider a C^k-mapping $F : \mathbb{R}^{2m} \times \mathbb{R} \to \mathbb{R}^{2m}$, where $k \geq 5$, and examine the family of difference equations

$$F(z_n, z_{n+1}, \mu) = 0. \tag{5.2}$$

Define the one-parameter family of matrix pencils

$$\mathcal{P}(\mu) := (A(\mu), B(\mu)) := (d_{\overline{z}}F(0, 0, \mu), d_z F(0, 0, \mu)),$$

where F has $(z, \overline{z}, \mu)$ as its argument.

THEOREM 3 (parameterized nonlinear KNF). *Suppose that* $(0, 0)$ *is a fixed point of* (5.2) *for all* $\mu \in \mathbb{R}$ *and that* $\mathcal{P}(0)$ *is a regular matrix pencil. Then there is a linear space* K_1 *and a* $C^{1+\alpha}$*-parameter family of* $C^{1+\alpha}$*-manifolds* $\mathcal{M}_\mu$ *modeled on* K_1 *such that* $\dim(K_1) = \#\sigma(\mathcal{P}(0))$ *and*

1. *for all* $(z, w) \in \mathcal{M}_\mu$ *there results* $F(z, w, \mu) = 0$,
2. *there is an* r' *(independent of* μ*) such that for each* $(z, w) \in \mathcal{M}_\mu$*, with* $\|(z, w)\| < r'$*, there is a unique* $(\overline{z}, \overline{w}) \in \mathcal{M}_\mu$ *such that* $w = \overline{z}$*, and*
3. *there is a* $C^{1+\alpha}$*-parameter family of* $C^{1+\alpha}$*-diffeomorphisms* $\theta_\mu : K_1 \to \mathcal{M}_\mu$ *such that if* $(z, w) = \theta_\mu(u)$ *for some* $u \in K_1$*, then* $(\overline{z}, \overline{w}) = \theta_\mu(\varphi(u, \mu))$*, where*

$$\varphi(u, \mu) = C(\mu)u + p(u, \mu).$$

Moreover, $C(\cdot)$ *is a* $C^{1+\alpha}$*-parameter family of maps in* $BL(K_1)$ *and, for some* $\delta > 0$*,* $p : B_\delta(0; K_1) \times B_\delta(0; \mathbb{R}) \to K_1$ *satisfies*

$$p(0, \mu) \equiv 0, \;\; d_u p(0, \mu) \equiv 0.$$

4. *If* $\lambda : [-\delta, \delta] \to \mathbb{C}$ *is a continuous (and so bounded) curve, then* $\lambda(\mu) \in \sigma(\mathcal{P}(\mu))$ *for all* $\mu \in [-\delta, \delta]$ *if and only if* $\lambda(\mu) \in \sigma(C(\mu))$ *for all* $\mu \in [-\delta, \delta]$.

Proof. Consider the suspended difference equation

$$\text{(S)} \qquad \begin{cases} \mu_{n+1} = \mu_n, \\ z_{n+1} = w_n, \\ 0 = F(z_n, w_n, \mu_n). \end{cases}$$

Let us write

$$F(z, w, \mu) = A(\mu)w + B(\mu)z + \mathcal{F}(z, w, \mu),$$

where $F(0, 0, \mu) = 0$, $d_z F(0, 0, \mu) = B(\mu)$, and $d_w F(0, 0, \mu) = A(\mu)$, and then consider the following matrix pencil on $\mathbb{R}^{2m}$:

$$(\mathcal{A}, \mathcal{B}(\mu)) := \left(\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & I \\ B(\mu) & A(\mu) \end{pmatrix} \right).$$

This satisfies $\sigma(\mathcal{A}, \mathcal{B}(\mu)) = -\sigma(\mathcal{P}(\mu))$ and is a regular matrix pencil when $\mu = 0$, and we can exploit this fact using the resulting Kronecker normal form to put (S) into a normal form. If we set $\mathbf{z} = (z, w)$, then we may write (S) as

$$\begin{aligned} \overline{\mu} &= \mu, \\ \mathcal{A}\overline{\mathbf{z}} &= \mathcal{B}(\mu)\mathbf{z} + \mathcal{F}(\mathbf{z}, \mu), \end{aligned}$$

where an overbar is used to denote a forward iterate. Now there is a matrix pair (P, Q) such that $P\mathcal{A}Q = \left(\begin{smallmatrix} I_{K_1} & 0 \\ 0 & N \end{smallmatrix}\right)$ and $P\mathcal{B}(0)Q = \left(\begin{smallmatrix} C & 0 \\ 0 & I_{K_2} \end{smallmatrix}\right)$, where $Q : \mathbb{R}^{2m} \to K_1 \oplus K_2 = \mathbb{R}^{p+q}$ and N is nilpotent. With $(u, v) := \mathbf{w} = Q\mathbf{z} \in K_1 \oplus K_2$, we obtain

$$\overline{\mu} = \mu, \tag{5.3}$$

$$\overline{u} = \alpha(\mu)u + \beta(\mu)v + \mathcal{G}_1(u, v, \mu), \tag{5.4}$$

$$N\overline{v} = \gamma(\mu)u + \delta(\mu)v + \mathcal{G}_2(u, v, \mu), \tag{5.5}$$

where

$$\alpha : K_1 \to K_1, \quad \beta : K_2 \to K_1, \quad \gamma : K_1 \to K_2, \quad \text{and} \quad \delta : K_2 \to K_2$$

are differentiable linear maps in μ, and

$$\alpha(0) = C, \quad \beta(0) = 0, \quad \gamma(0) = 0, \quad \text{and} \quad \delta(0) = I_{K_2},$$

where $C \in BL(K_1)$ is provided by the KNF of $\mathcal{P}(0)$ and $\sigma(C) = -\sigma(\mathcal{A}, \mathcal{B}(0)) = \sigma(\mathcal{P}(0))$. Moreover, $\mathcal{G}_1$ and $\mathcal{G}_2$ represent $\mathcal{O}_2(u, v)$-functions parameterized by μ.

Seeking an invariant manifold on which $v = h(u, \mu)$, we put (5.3)–(5.5) into the form

$$\overline{\mu} = \mu, \tag{5.6}$$

$$\overline{u} = Cu + \mathcal{O}_2(u, v, \mu), \tag{5.7}$$

$$N\overline{v} = v + \mathcal{O}_2(u, v, \mu), \tag{5.8}$$

where $\mathcal{O}_2(u, v, \mu)$ denotes a function of (u, v, μ) which vanishes to second or higher order at the origin. From Theorem 1 we obtain a local invariant manifold of (5.6)–(5.7) on which $v = h(u, \mu)$. Moreover $h(0, 0) = 0, dh(0, 0) = 0$, and we may assume that h is C^1.

It follows that $h(u, \mu)$ satisfies the functional equation

$$Nh(\alpha u + \beta h(u, \mu) + \mathcal{G}_1(u, h(u, \mu), \mu)) = \gamma(\mu)u + \delta h(u, \mu) + \mathcal{G}_2(u, h(u, \mu), \mu),$$

and if we set $x = h(0, \mu)$, then x satisfies the equation

$$Nh(\beta x + \mathcal{G}_1(0, x, \mu), \mu) = \delta(\mu)x + \mathcal{G}_2(0, x, \mu).$$

The latter is an algebraic equation for x which we denote $a(x, \mu) = 0$; moreover $a(0, \mu) = 0$ holds for all μ near 0, whence $h(0, \mu) \equiv 0$. In addition, a short calculation shows that

$$d_x a(0, \mu) = \delta(\mu) - N d_u h(0, \mu)[\beta(\mu)],$$

which is an identity mapping when $\mu = 0$. The implicit function theorem now ensures that $x = 0$ is the only solution of $a(x, \mu) = 0$ for all μ near 0.

The functional equation satisfied by $d_u h(u,\mu)$ is then

$$-Nd_u h(\alpha u+\beta h+\mathcal{G}_1(u,h,\mu))[\alpha+\beta d_u h+d_u\mathcal{G}_1+d_v\mathcal{G}_1\cdot d_u h] \\ +\gamma+\delta d_u h+d_u\mathcal{G}_2+d_v\mathcal{G}_2\cdot d_u h=0,$$

where various arguments have been omitted for brevity. If we write τ for the linear map $d_u h(0,\mu)$, then

$$\gamma(\mu)+\delta(\mu)\tau=N\tau\cdot[\alpha(\mu)+\beta(\mu)\tau].$$

This is a Riccati equation for τ that can be solved near $\mu=0$ for τ as a function of μ using the implicit function theorem and the properties of α,β,γ, and δ. The result that $\tau(0)=0$ then follows because N is nilpotent and $\tau(0)=N\tau(0)[C]$.

We are now in a position to define the one-parameter family of matrices, denoted by $C(\mu)$ in the statement of the theorem, namely,

$$C(\mu):=\alpha(\mu)+\beta(\mu)\tau(\mu),$$

so that $C(0)=C$. If $C(\mu)$ has an eigenvalue λ, say, then

$$(\alpha+\beta\tau)w=\lambda w \implies \gamma+\delta\tau=N\tau\cdot[\lambda w],$$

whence

$$\begin{bmatrix}\alpha & \beta\\ \gamma & \delta\end{bmatrix}\begin{bmatrix}w\\ \tau w\end{bmatrix}=\lambda\begin{bmatrix}w\\ N\cdot\tau w\end{bmatrix},$$

and therefore

$$-\lambda\in\sigma\left(\begin{bmatrix}I & 0\\ 0 & N\end{bmatrix},\begin{bmatrix}\alpha & \beta\\ \gamma & \delta\end{bmatrix}\right)=\sigma(\mathcal{A},\mathcal{B}(\mu))=-\sigma(\mathcal{P}(\mu)).$$

We have deduced that

$$\sigma(\alpha(\mu)+\beta(\mu)\tau(\mu))\subseteq\sigma(\mathcal{P}(\mu)),$$

but the left-hand side of this inclusion has $\dim(K_1)$ elements counted according to algebraic multiplicity, whereas the right-hand side may have more unless, that is, $\mu=0$, in which case the inclusion is replaced by an equality because $\sigma(\alpha(0))=\sigma(C)=\sigma(\mathcal{P}(0))$.

As a result, if $\lambda(\mu)\in C([-\delta,\delta],\mathbb{C})$ is a continuous path of eigenvalues of $\mathcal{P}(\mu)$, then by virtue of the fact that $\lambda(0)\in\sigma(\alpha(\mu)+\beta(\mu)\tau(\mu)|_{\mu=0})$, it follows by the continuous dependence of eigenvalues on μ and by counting their location in the complex plane that $\lambda(\mu)\in\sigma(\alpha(\mu)+\beta(\mu)\tau(\mu)|_{\mu\neq 0})$. □

Conclusion 4 of Theorem 3 is not equivalent to saying that a locus of eigenvalues of $\mathcal{P}(\mu)$ is necessarily a locus of eigenvalues of $C(\mu)$. This is because $\mathcal{P}(\mu)$ may have other eigenvalues for small μ which become unbounded as μ tends to zero, and such curves cannot correspond to eigenvalues of $C(\mu)$. This happens, for instance, in the singularity-induced bifurcation theorem from [1] because an eigenvalue has a pole with respect to the bifurcation parameter.

5.2.1. Existence of bifurcations for (1.1). One can easily prove a period-doubling bifurcation theorem for (1.1) without having recourse to the nonlinear normal form given in Theorem 2, but we now include this result for completeness.

THEOREM 4. *Suppose that* (5.2) *satisfies the following hypotheses:*

1. $1 \notin \sigma(\mathcal{P}(0))$ *and* $F(0,0,\mu) \equiv 0$,
2. $\ker(-A(0)+B(0)) = \langle k \rangle$, *so that* $-1 \in \sigma(\mathcal{P}(0))$, *and*
3. $B'(0)k \notin \mathrm{im}(-\mathrm{A}(0)+\mathrm{B}(0))$.

Then $\mu = 0$ *is a period-doubling bifurcation point for* (5.2) *from the trivial solution* $z = 0$.

Proof. Consider the algebraic equation $G(z,w,\mu) = 0$, where $G : \mathbb{R}^{2m+1} \to \mathbb{R}^{2m}$ is given by

$$G(z,w,\mu) := \begin{bmatrix} F(z,w,\mu) \\ F(w,z,\mu) \end{bmatrix},$$

and moreover G has the trivial solution branch. Also define $\overline{G}(z,\mu) := F(z,z,\mu)$, so that

$$d_{(z,w)}G(0,0,\mu) = \begin{bmatrix} A(\mu) & B(\mu) \\ B(\mu) & A(\mu) \end{bmatrix},$$

and $d_z\overline{G}(0,\mu) = A(\mu)+B(\mu)$. By assumption, $A(0)+B(0)$ is invertible, and therefore $d_z\overline{G}(0,\mu)$ is an invertible map for small $|\mu|$, so that if $G(z,w,\mu) = 0$ near $\mu = 0$, then $z \neq w$ unless $z = w = 0$. The theorem now follows from the simple eigenvalue bifurcation theorem applied to G at $\mu = 0$, noting that the kernel of $d_{(z,w)}G(0,0,0)$ is $(k,-k)^T$. ∎

One can of course formulate a similarly straightforward fold bifurcation for (1.1) in an entirely analogous fashion. However, the following theorem relies on Theorem 3 in a nontrivial way.

THEOREM 5 (Neimark–Sacker bifurcation). *Suppose that* (5.2) *has the fixed point* $z = 0$ *for all* $\mu \in \mathbb{R}$ *and that* $\lambda(\mu) \in \sigma(\mathcal{P}(\mu))$ *is a curve which satisfies the following:*

1. $\mathcal{P}(0)$ *is a regular matrix pencil;*
2. $|\lambda(0)| = 1$, *and* $\lambda(0)$ *is an algebraically simply eigenvalue of* $\mathcal{P}(0)$*;*
3. $\lambda(0)^n \neq 1$ *for* $n \in \{1,2,3,4\}$*;*
4. $\frac{d}{d\mu}|\lambda(\mu)|\big|_{\mu=0} \neq 0$.

Then modulo a further nonresonance condition[1] *there is a half-interval* $J \subset \mathbb{R}$ *containing* 0 *in its closure such that* (5.2) *possesses a quasi-invariant circle* $\Gamma_\mu \subset \mathbb{R}^{2m}$ *for all* $\mu \in J$. *Moreover, if* $\mathrm{diam}(\Gamma_\mu) = \sup\{\|z-w\| : z,w \in \Gamma_\mu\}$, *then* $\lim_{\mu\to 0} \mathrm{diam}(\Gamma_\mu) = 0$.

Proof. Theorem 5 follows immediately from the Neimark–Sacker bifurcation for maps applied to $\varphi(u,\mu)$ from Theorem 3 (part 3). ∎

6. Examples. *Example* 1 (output-nulling control problem). The results in this paper give sufficient conditions for a positive answer to the following question:

(Q) *Given* $f(= f(x,u)) : \mathbb{R}^{n+m} \to \mathbb{R}^n, g(= g(x)) : \mathbb{R}^n \to \mathbb{R}^m$ *such that* $f(0,0) = 0, g(0) = 0$, *does there exist a sequence of states* (x_n) *given by the iterates of* f *and controls* (u_n) *such that* $g(x_n) \equiv 0$ *and* $x_n \to 0$ *as* $n \to \infty$*?*

Thus, we seek a global orbit of the ΔAE

$$x_{n+1} = f(x_n,u_n),\ g(x_n) = 0.$$

[1] See [21] or [25, p. 376], where the open condition "$a \neq 0$" is given, and this requires the computation of third-order terms in the normal form for this bifurcation.

A necessary condition for the existence of such a solution can be obtained by substituting the dynamic part of the problem into the constraint, to give the hidden constraint

$$g(f(x_n, u_n)) = 0 \qquad (\forall n \geq 1).$$

Hence, provided the function $g(f(x,u))$ has an invertible partial u-derivative at $(x,u) = (0,0)$, by the stable manifold theorem the response to (**Q**) is affirmative if the spectrum of the x-derivative of $f(x, U(x))$, also evaluated at $(x,u) = (0,0)$, contains an element of the open unit disk. Here $U(x)$ denotes the solution of the equation $g(f(x,U)) = 0$ given locally by the implicit function theorem. Note for a moment that if the stated u-derivative $dg(0)d_uf(0,0)$ is invertible, it follows that the matrix pencil

$$(A,B) := \left(\begin{pmatrix} I_x & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} d_xf & d_uf \\ dg & 0 \end{pmatrix} \right)\Bigg|_{(x,u)=(0,0)}$$

is regular, and thus (A1)–(A3) hold for this problem.

However, by using Theorem 1 one can dispense with the condition that $dg(0)d_uf(0,0)$ is invertible. In fact, let us assume that $K := \ker(dg(0)d_uf(0,0)) \neq \{0\}$. In this case, the matrix $\lambda A + B$ is invertible if $-\lambda \notin \sigma(d_xf(0,0))$ and $dg(\lambda I_x + d_xf)^{-1}d_uf$ is also invertible at $(x,u) = (0,0)$. However, for λ large we have

$$\begin{aligned} \lambda^2 dg(\lambda I_x + d_xf)^{-1}d_uf &= \lambda dg(I_x + \lambda^{-1}d_xf)^{-1}d_uf \\ &= \lambda dg(I_x - \lambda^{-1}d_xf + O(\lambda^{-2}))d_uf \\ &= \lambda\ dg \cdot d_uf - dg \cdot d_xf \cdot d_uf + O(\lambda^{-1}), \end{aligned}$$

evaluating all of the stated derivatives at $(x,u) = (0,0)$. As a result, if the weaker condition holds that the pencil $(dg \cdot d_uf, dg \cdot d_xf \cdot d_uf)$ is regular, then (A,B) is regular and (A1)–(A3) still apply. The response to (Q) is again affirmative if $\sigma(A,B)$ contains at least one member of the open unit disk.

In fact one can show that output-nulling control problems are well-posed in sequence spaces as follows. First consider the linear problem

$$Az_{n+1} + Bz_n = \Gamma_n, \tag{6.1}$$

where $n \in \mathbb{N}$ and (Γ_n) is a given sequence in

$$\ell^\infty_{\mathbb{N}}(\mathbb{R}^m) = \left\{ (z_n)_{n\in\mathbb{N}} : z_n \in \mathbb{R}^m, \sup_{n\in\mathbb{N}} \|z_n\|_{\mathbb{R}^m} < \infty \right\},$$

but $\det A = 0$. If (A,B) is a regular matrix pencil with index ν, the KNF allows us to write (6.1) in the form

$$u_{n+1} = Cu_n + \alpha_n, \tag{6.2}$$

$$Nv_{n+1} = v_n + \beta_n, \tag{6.3}$$

where $(u_n, v_n) \in \mathbb{R}^{p+q}$. In order to solve (6.1), let us consider the linear operator $I - N\sigma$ on a space of sequences $\ell^\infty_{\mathbb{N}}(\mathbb{R}^q)$. We take linear maps $T \in BL(\mathbb{R}^q)$ to act pointwise on $\ell^\infty_{\mathbb{N}}(\mathbb{R}^q)$, so

$$T(w_n)_{n\in\mathbb{N}} = (Tw_n)_{n\in\mathbb{N}} \qquad (\forall (w_n)_{n\in\mathbb{N}} \in \ell^\infty_{\mathbb{N}}(\mathbb{R}^q)),$$

and we define σ the forward-shift map by

$$\sigma(w_n)_{n\in\mathbb{N}} = (w_{n+1})_{n\in\mathbb{N}} \qquad (\forall (w_n)_{n\in\mathbb{N}} \in \ell^\infty_{\mathbb{N}}(\mathbb{R}^q)).$$

Any such T will commute with σ, and one can see by a direct multiplication that

$$(I - N\sigma)\sum_{i=0}^{\nu-1} N^i\sigma^i = I,$$

where I denotes the identity on $\ell^\infty_{\mathbb{N}}(\mathbb{R}^q)$.

One can solve (6.2) in a suitably weighted sequence space if no restrictions are to be placed on the spectrum of (A, B). Equation (6.3) can also be solved:

$$\mathbf{v} = -(I - N\sigma)^{-1}\beta = -\sum_{i=0}^{\nu-1} N^i\sigma^i\beta,$$

where $\mathbf{v} = (v_i)_{i\in\mathbb{N}}$ and $\beta = (\beta_i)_{i\in\mathbb{N}}$, whence

$$v_n = -\sum_{i=0}^{\nu-1} N^i\beta_{n+i}.$$

So from a temporal point of view the current values of the state depend on future values of the input, but (6.1) is still well-posed in a sequence space as $\mathbf{v}$ depends continuously on β.

This means that a second- or higher-order nonlinear perturbation of (6.1),

$$Az_{n+1} + Bz_n + \mathcal{F}(z_n) = \Gamma_n, \tag{6.4}$$

say, where $\mathcal{F}(0) = 0, d\mathcal{F}(0) = 0$, can be written as an infinite-dimensional problem

$$(A\sigma + B)\mathbf{z} + \mathcal{F}(\mathbf{z}) = \mathbf{\Gamma} \tag{6.5}$$

in $\ell^\infty_{\mathbb{N}}(\mathbb{R}^m)$, and one can apply the inverse function theorem to solve locally for small-norm solutions of the form $\mathbf{z} = \mathbf{z}(\mathbf{\Gamma})$, where $\mathbf{z}(\mathbf{0}) = \mathbf{0}$.

This solution can be found via the Picard iteration $\mathbf{z}(\mathbf{\Gamma}) = \lim_{n\to\infty} \mathbf{y}^{(n)}$, where $\mathbf{y}^{(0)} = \mathbf{0}$ and

$$\mathbf{y}^{(n+1)} = -(A\sigma + B)^{-1}[\mathcal{F}(\mathbf{y}^{(n)}) - \mathbf{\Gamma}].$$

As a result, writing the solution sequence $\mathbf{z}(\mathbf{\Gamma})$ as $(z_n(\mathbf{\Gamma}))_{n\in\mathbb{N}}$, it is clear that the nonlinear perturbation will have the effect of making each z_n depend on infinitely many elements of the sequence $\mathbf{\Gamma}$, unless ν happens to equal 1.

This effect has been observed before in the literature in the context of delay DAEs [5, 13], where it is noted in the former reference that linear systems of delay DAEs can act like advanced systems when their index is two or higher. The problem (6.5) is displaying exactly this behavior.

Example 2. This example serves to illustrate how we can use Theorem 1 to deduce qualitative similarities between a DAE and its discrete counterpart. Take the DAE

$$\dot{x} = f(x, y), \tag{6.6}$$

$$g(x, y) = 0, \tag{6.7}$$

subject to $x \in \mathbb{R}^n, y \in \mathbb{R}^m$ with an equilibrium at the origin, so that $f(0,0) = 0, g(0,0) = 0$. Let us impose a singularity of the form

$$\ker(d_y g(0,0)) = \langle k \rangle \quad \text{such that} \quad d_y f(0,0) d_x g(0,0) k \notin \operatorname{im}(d_y g(0,0)), \tag{6.8}$$

and the equilibrium solution is isolated:

$$\det \begin{pmatrix} d_x f(0,0) & d_y f(0,0) \\ d_x g(0,0) & d_y g(0,0) \end{pmatrix} \neq 0.$$

From [2] it is known that (6.6)–(6.7) has an invariant manifold W of dimension $n-1$ that contains the origin and intersects the singularity in an $n-2$-dimensional manifold of pseudoequilibria.

Now consider the forward-Euler method in state-space form [8, p. 375] applied to (6.6)–(6.7), resulting in the difference equation

$$x_{i+1} - x_i = h f(x_i, y_i), \tag{6.9}$$

$$g(x_{i+1}, y_{i+1}) = 0. \tag{6.10}$$

Using Theorem 1, in order to to show that (6.9)–(6.10) possesses a quasi-invariant manifold of solutions W_h that contains the origin and has dimension $n-1$, we need only show that (A1)–(A3) hold, which entails showing that the derivative at the origin of (6.9)–(6.10) is a regular matrix pencil. Hence we seek a $\xi \in \mathbb{C}$ such that

$$\det \left(\xi \begin{bmatrix} I & 0 \\ d_x g & d_y g \end{bmatrix} + \begin{bmatrix} I + h d_x f & h d_y f \\ 0 & 0 \end{bmatrix} \right) \Bigg|_{(x,y)=(0,0)} \neq 0.$$

However, the conditions in (6.8) ensure the existence of such a ξ for any fixed $h \neq 0$, and the existence of W_h follows. The dimension of W_h comes from counting the number of finite eigenvalues of the linearization of (6.9)–(6.10) which is given in [1] as $n-1$.

Example 3. Consider again (6.6)–(6.7) but now with a parameter α

$$\dot{x} = f(x, y, \alpha), g(x, y, \alpha) = 0, \tag{6.11}$$

and suppose that $(x, y) = (0, 0)$ is an equilibrium locus for all $\alpha \in \mathbb{R}$. Now suppose that the conditions for a Hopf bifurcation are formally satisfied by (6.11) at $\alpha = \alpha_0$: $\omega(\alpha) \in \sigma(M, -L(\alpha))$ is a locus of algebraically simple eigenvalues, where

$$M = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \; L(\alpha) = \begin{bmatrix} d_x f(0,0,\alpha) & d_y f(0,0,\alpha) \\ d_x g(0,0,\alpha) & d_y g(0,0,\alpha) \end{bmatrix};$$

and $\omega(\alpha_0) = i\omega_0$ is an eigenvalue of $(M, -L(\alpha))$ such that $\frac{d}{d\alpha} \Re(\omega(\alpha)) \big|_{\alpha=\alpha_0}$ is nonzero. Also, no other eigenvalues of the regular matrix pencil $(M, -L(\alpha_0))$ have a zero real part.

Then, modulo an open condition on the third-order derivatives of the smooth functions f and g, we can show that the backward-Euler method

$$x_{i+1} - x_i = h f(x_{i+1}, y_{i+1}), \tag{6.12}$$

$$-h \cdot g(x_{i+1}, y_{i+1}) = 0, \tag{6.13}$$

satisfies the conditions of the Neimark–Sacker bifurcation theorem for all sufficiently small $h > 0$. Note that $(M - hL(\alpha), -M)$ is the linearization of (6.12)–(6.13) about the zero fixed point and is a regular matrix pencil for $h > 0$ and $\alpha \approx \alpha_0$ such that

$$(1 - h\omega(\alpha))^{-1} \in \sigma(M - hL(\alpha), -M) \qquad (\forall \alpha).$$

In order to apply Theorem 5 to (6.12)–(6.13) we first note that the eigenvalue locus $(1 - h\omega(\alpha))^{-1}$ has unit length if and only if $1 - h\omega(\alpha)$ has unit length. If we define functions $R(\alpha)$ and $I(\alpha)$ by $\omega(\alpha) = R(\alpha) + iI(\alpha)$, then $|1 - h\omega(\alpha)| = 1$ if and only if $(1 - hR(\alpha))^2 + h^2 I(\alpha)^2 = 1$, which holds when

$$h\left(-R(\alpha) + \frac{h}{2}(R(\alpha)^2 + I(\alpha)^2)\right) = 0. \tag{6.14}$$

As a result, we define the function $b(\alpha, h) := -R(\alpha) + \frac{h}{2}(R(\alpha)^2 + I(\alpha)^2)$ and note that $b(\alpha, h) = 0$ for $h > 0$ ensures that the linearization of (6.12)–(6.13) at $(0, 0)$ has an eigenvalue of unit modulus. Now $b(\alpha_0, 0) = 0$ and $\frac{\partial b}{\partial \alpha}(\alpha_0, 0) = -R'(\alpha_0) \neq 0$ by assumption, and, as a result, one may solve $b(\alpha, h) = 0$ locally using the implicit function theorem for $\alpha = \alpha(h)$ such that $\alpha(0) = \alpha_0$.

From this calculation one can show that the numerical scheme (6.12)–(6.13) has a quasi-invariant circle for α in some half-neighborhood of $\alpha(h) = \alpha_0 + \frac{h\omega_0^2}{R'(\alpha)} + O(h^2)$ provided $h > 0$ is sufficiently small.

Note that no assumption is made regarding the invertibility of $d_y g(0, 0, 0)$. If this mapping were invertible in addition to the formal conditions given above for Hopf bifurcation, then the existence of a locus of periodic solutions of (6.11) could be deduced. However, without the invertibility of $d_y g(0, 0, 0)$, it is not known whether a Hopf bifurcation occurs in (6.11), but (6.12)–(6.13) has a locus of invariant circles nevertheless.

6.1. Concluding remark. The results in this paper can be used to show that second-order problems of the form

$$F(z_n, z_{n+1}, z_{n+2}) = 0 \tag{6.15}$$

have quasi-invariant manifolds provided that the appropriate matrix pencil is regular simply by rewriting (6.15) as a first-order problem. However, the methods of this paper do not easily extend to the study of invariant manifolds of the system that one would like to study if (1.1) had a period-2 orbit $(z, w, z, w, \ldots)$, namely, the system

$$F(z_n, z_{n+1}) = 0, \tag{6.16}$$

$$F(z_{n+1}, z_{n+2}) = 0, \tag{6.17}$$

where $F(z, w) = 0$ and $F(w, z) = 0$, with $w \neq z$. Another approach is needed to answer the question of whether overdetermined systems of this type have any invariant manifolds associated with them.

REFERENCES

[1] R. E. Beardmore, *The singularity-induced bifurcation and its Kronecker normal form*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 126–137.

[2] R. E. Beardmore and R. Laister, *The flow of a DAE near a singular equilibrium*, SIAM J. Matrix Anal. Appl., 24 (2002), pp. 106–120.

[3] K. E. Brenan, S. L. Campbell, and L. R. Petzold, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, North-Holland, Amsterdam, 1989.
[4] R. Bru, C. Coll, and E. Sanchez, *Structural properties of positive linear time-invariant difference-algebraic equations*, Linear Algebra Appl., 349 (2002), pp. 1–10.
[5] S. L. Campbell, *Singular linear systems of differential equations with delays*, Appl. Anal., 11 (1980), pp. 129–136.
[6] T. Fliegner, Ü. Kotta, and H. Nijmeijer, *Solvability and right-inversion of implicit nonlinear discrete-time systems*, SIAM J. Control Optim., 34 (1996), pp. 2092–2115.
[7] F. R. Gantmacher, *Theory of Matrices*, Vol. 2, Chelsea, New York, 1977.
[8] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations*, 2nd revised ed., Springer Ser. Comput. Math. II, Springer, New York, 2002.
[9] T. Kaczynski and W. Krawcewicz, *A local Hopf bifurcation theorem for a certain class of implicit differential equations*, Canad. Math. Bull., 36 (1993), pp. 183–189.
[10] P. E. Kloeden and P. Marín-Rubio, *Weak pullback attractors of non-autonomous difference inclusions*, J. Difference Equ. Appl., 9 (2003), pp. 489–502.
[11] J. Laurent-Varin, F. Bonnans, N. Berend, C. Talbot, and M. Haddou, *On the refinement of discretization for optimal control problems*, in Proceedings of the 16th IFAC Symposium on Automatic Control in Aerospace, St. Petersburg, 2004.
[12] J.-Y. Lin and Z.-H. Yang, *A discrete optimal control for descriptor systems*, IEEE Trans. Automat. Control, 34 (1989), pp. 177–181.
[13] H. Logeman, *Destabilizing effects of small time delays on feedback-controlled descriptor systems*, Linear Algebra Appl., (1998), pp. 131–153.
[14] C. Navasca, *Local stable manifold for the bidirectional discrete-time dynamics*, SIAM J. Control Optim., submitted.
[15] P. J. Rabier, *The Hopf bifurcation theorem for quasilinear differential-algebraic equations*, Comput. Methods Appl. Mech. Engrg., 170 (1999), pp. 355–371.
[16] P. J. Rabier and W. C. Rheinboldt, *On impasse points of quasilinear differential-algebraic equations*, Math. Anal. Appl., 181 (1994), pp. 429–454.
[17] P. J. Rabier and W. C. Rheinboldt, *Theoretical and Numerical Analysis of Differential-Algebraic Equations*, Handbook of Numerical Analysis, Vol. VIII, Elsevier Science, New York, 2003, Part 4, pp. 183–540.
[18] S. Reich, *On the qualitative behaviour of DAEs*, Circuits Systems Signal Proc., 14 (1995), pp. 427–443.
[19] G. Reißig and H. Boche, *On singularities of autonomous implicit ordinary differential equations*, IEEE CAS I, 50 (2003), pp. 922–931.
[20] R. Riaza and P. J. Zufiria, *Stability of singular equilibria in quasilinear implicit differential equations*, J. Differential Equations, 171 (2001), pp. 24–53.
[21] R. J. Sacker, *On Invariant Surfaces and Bifurcation of Periodic Solutions of Ordinary Differential Equations*, Ph.D. thesis, New York University, Courant Institute of Mathematical Sciences, 1964.
[22] G. Söderlind, *Remarks on the stability of high-index DAEs with respect to parametric perturbations*, Computing, 49 (1992), pp. 303–314.
[23] J. Sotomayor and M. Zhitomirskii, *Impasse singularities of differential systems of the form* $A(x)\dot{x} = f(x)$, J. Differential Equations, 169 (2001), pp. 567–587.
[24] V. Venkatasubramanian, *Singularity induced bifurcation in the Van Der Pol oscillator*, IEEE Trans. Circuits Syst. I Fund. Theory Appl., 41 (1994), pp. 765–769.
[25] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Texts Appl. Math., Springer, New York, 1990.
[26] E. Zeidler, *Nonlinear Analysis: Linear Monotone Operators*, Vol. II, Springer, New York, 1990.

SIAM J. MATH. ANAL.
Vol. 40, No. 2, pp. 443–474

ON THE WELL-POSEDNESS FOR THE VISCOUS SHALLOW WATER EQUATIONS*

QIONGLEI CHEN†, CHANGXING MIAO†, AND ZHIFEI ZHANG‡

Abstract. In this paper, we prove the existence and uniqueness of the solutions for the two-dimensional viscous shallow water equations with low regularity assumptions on the initial data as well as the initial height bounded away from zero.

Key words. shallow water equations, well-posedness, Bony's paraproduct decomposition, weight Besov space

AMS subject classifications. 35Q35, 76D

DOI. 10.1137/060660552

1. Introduction.

In this paper, we study the two-dimensional (2D) viscous shallow water equations with a more general diffusion,

$$\begin{cases} h_t + \operatorname{div}(hu) = 0, \\ h(u_t + u\cdot\nabla u) - \nu\nabla\cdot(hD(u)) - \nu\nabla(h\operatorname{div}(u)) + h\nabla h = 0, \\ u(0,\cdot) = u_0,\ h(0,\cdot) = h_0, \end{cases} \tag{1.1}$$

where $h(t,x)$ is the height of fluid surface, $u(t,x) = (u_1(t,x), u_2(t,x))$ is the horizontal velocity vector field, $D(u) = \frac{1}{2}(\nabla u + \nabla u^t)$ is the deformation tensor, and $\nu > 0$ is the viscous coefficient.

Recently, the viscous shallow water equations have been widely studied by mathematicians; see the review paper [4]. Bui [5] proved the local existence and uniqueness of classical solutions to the Cauchy–Dirichlet problem for shallow water equations with initial data h_0, u_0 in Hölder spaces as well as h_0 bounded away from a vacuum. Kloeden [17] and Sundbye [20] independently proved global existence and uniqueness of classical solutions to the Cauchy–Dirichlet problem in Sobolev spaces. Later, Sundbye [21] also proved global existence and uniqueness of classical solutions to the Cauchy problem. However, for all of the above results (except those of [5]), the authors only consider the case when the initial data h_0 is a small perturbation of some positive constant $\bar{h}_0$ and u_0 is small in some sense. Very recently, Wang and Xu [23] proved the local well-posedness of the Cauchy problem in Sobolev spaces for the large data u_0 and h_0 closing to $\bar{h}_0$. More precisely, they obtained the following result.

THEOREM 1.1 (see [23]). *Let $\bar{h}_0$ be a strictly positive constant and $s > 2$. Assume that*

(i) $(u_0, h_0 - \bar{h}_0) \in H^s(\mathbb{R}^2) \otimes H^s(\mathbb{R}^2)$;

(ii) $\|h_0 - \bar{h}_0\|_{H^s} \ll \bar{h}_0$.

*Received by the editors May 23, 2006; accepted for publication (in revised form) September 17, 2007; published electronically May 16, 2008.

http://www.siam.org/journals/sima/40-2/66055.html

†Institute of Applied Physics and Computational Mathematics, P.O. Box 8009, Beijing 100088, People's Republic of China (chen_qionglei@iapcm.ac.cn, miao_changxing@iapcm.ac.cn). The research of these authors was partially supported by the NSF of China (grants 10701012 and 10725102).

‡School of Mathematical Science, Peking University, Beijing 100871, People's Republic of China (zfzhang@math.pku.edu.cn). This author's research was supported by the NSF of China (grant 10601002).

Then there exist a positive time T and a unique solution (u,h) of (1.1) such that

$$u,\ h-\bar{h}_0\in L^\infty([0,T],H^s),\quad \nabla u\in L^2([0,T];H^s). \tag{1.2}$$

Moreover, there exists a strictly positive constant c such that if

$$\|u_0\|_{H^s}+\|h_0-\bar{h}_0\|_{H^s}\leq c, \tag{1.3}$$

then we can choose $T=+\infty$.

One purpose of this paper is to study the well-posedness of (1.1) for the initial data with the minimal regularity. For the incompressible Navier–Stokes equations, such research has been initiated by Fujita and Kato [16]; see also [6, 7, 18] for other relevant results. They proved local well-posedness for the incompressible Navier–Stokes equations in the scaling invariant space. The scaling invariance means that if (u,p) is a solution of the incompressible Navier–Stokes equations with initial data $u_0(x)$, then

$$u_\lambda(t,x)\triangleq\lambda u(\lambda^2t,\lambda x),\quad p_\lambda(t,x)\triangleq\lambda^2p(\lambda^2t,\lambda x) \tag{1.4}$$

is also a solution of the incompressible Navier–Stokes equations with $u_{0,\lambda}\triangleq\lambda u_0(\lambda x)$. Obviously, $\dot{H}^{\frac{d}{2}-1}(\mathbb{R}^d)$ is a scaling invariant space under the scaling of (1.4), i.e.,

$$\|u_\lambda\|_{\dot{H}^{\frac{d}{2}-1}}=\|u\|_{\dot{H}^{\frac{d}{2}-1}}.$$

Equations (1.1) have no scaling invariance like the incompressible Navier–Stokes equations. However, due to the similarity of the structure between (1.1) and the incompressible Navier–Stokes equations, we still solve (1.1) for initial data whose regularity fits the scaling of (1.4). It should be pointed out that Danchin was the first to consider the same problem for the compressible Navier–Stokes equations, and some ideas of this paper are motivated by [11].

The second purpose of this paper is to prove the local well-posedness of (1.1) under the more natural assumption that the initial height is bounded away from zero. For the initial data with slightly higher regularity, this can be easily obtained by modifying the argument of Danchin [13]. However, for the initial data with low regularity, his method is not applicable anymore, since the proof of [13] relies on the fact that some profits can be gained from the inclusion map $B^s\hookrightarrow L^\infty$ in the case of $s>\frac{d}{2}$. For this reason, we have to introduce some kind of weighted Besov space E^s_T (see section 3), which is crucial to eliminating the condition that the initial height h_0 is close to $\bar{h}_0$. One important observation is that the E^s_T norm of the solution is small for small time T.

Before stating our main result, let us first introduce some notations and definitions. Choose a radial function $\varphi\in\mathcal{S}(\mathbb{R}^d)$ such that

$$\operatorname{supp}\varphi\subset\left\{\xi\in\mathbb{R}^d;\frac{5}{6}\leq|\xi|\leq\frac{12}{5}\right\},\quad \sum_{k\in\mathbb{Z}}\varphi(2^{-k}\xi)=1,\quad \xi\in\mathbb{R}^d\setminus\{0\}.$$

Here $\varphi_k(\xi)=\varphi(2^{-k}\xi)$, $k\in\mathbb{Z}$.

Definition 1.1. *Let $k\in\mathbb{Z}$. The Littlewood–Paley projection operators Δ_k and S_k are defined as follows:*

$$\Delta_kf=\varphi(2^{-k}D)f,\quad S_kf=\sum_{j\leq k-1}\Delta_jf\quad \textit{for}\quad f\in\mathcal{S}'(\mathbb{R}^d).$$

We denote the space $\mathcal{Z}'(\mathbb{R}^d)$ by the dual space of $\mathcal{Z}(\mathbb{R}^d) = \{f \in \mathcal{S}(\mathbb{R}^d);\ D^\alpha \hat{f}(0) = 0; \forall \alpha \in \mathbb{N}^d \text{ multi-index}\}$; it also can be identified by the quotient space of $\mathcal{S}'(\mathbb{R}^d)/\mathcal{P}$ with the polynomials space $\mathcal{P}$. The formal equality

$$f = \sum_{k\in\mathbb{Z}} \Delta_k f$$

holds true for $f \in \mathcal{Z}'(\mathbb{R}^d)$ and is called the homogeneous Littlewood–Paley decomposition. It has nice properties of quasi-orthogonality: with our choice of φ,

$$\Delta_j\Delta_k f = 0 \quad \text{if} \quad |j-k| \geq 2 \quad \text{and} \quad \Delta_j(S_{k-1}\Delta_k f) = 0 \quad \text{if} \quad |j-k| \geq 4. \tag{1.5}$$

DEFINITION 1.2. *Let $s \in \mathbb{R}$, $1 \leq p, r \leq +\infty$. The homogeneous Besov space $\dot{B}^s_{p,r}$ is defined by*

$$\dot{B}^s_{p,r} = \{f \in \mathcal{Z}'(\mathbb{R}^d) :\ \|f\|_{\dot{B}^s_{p,r}} < +\infty\},$$

where

$$\|f\|_{\dot{B}^s_{p,r}} = \begin{cases} \displaystyle\Big(\sum_{k\in\mathbb{Z}} 2^{ksr}\|\Delta_k f\|_p^r\Big)^{\frac{1}{r}} & \text{for} \quad r < +\infty, \\ \displaystyle\sup_{k\in\mathbb{Z}} 2^{ks}\|\Delta_k f\|_p & \text{for} \quad r = +\infty. \end{cases}$$

If $p = r = 2$, $\dot{B}^s_{2,2} = \dot{H}^s$, and if $d = 2$, we have $\dot{B}^1_{2,1} \hookrightarrow L^\infty$ and

$$\|f\|_\infty \leq C\|f\|_{\dot{B}^1_{2,1}}.$$

We refer to [8, 22] for more details.

In addition to the general time-space space such as $L^\rho(0,T;\dot{B}^s_{p,r})$, we introduce a useful mixed time-space homogeneous Besov space $\widetilde{L}^\rho_T(\dot{B}^s_{p,r})$, which was initiated in [10] and will be used in the proof of the uniqueness.

DEFINITION 1.3. *Let $s \in \mathbb{R}$, $1 \leq p, r, \rho \leq +\infty$, $0 < T \leq +\infty$. The mixed time-space homogeneous Besov space $\widetilde{L}^r_T(\dot{B}^s_{p,q})$ is defined by*

$$\widetilde{L}^\rho_T(\dot{B}^s_{p,r}) = \{f \in \mathcal{Z}'(\mathbb{R}^{d+1}) :\ \|f\|_{\widetilde{L}^\rho_T(\dot{B}^s_{p,r})} < +\infty\},$$

where

$$\|f\|_{\widetilde{L}^\rho_T(\dot{B}^s_{p,r})} = \Big\|2^{ks}\Big(\int_0^T \|\Delta_k f(t)\|_p^\rho dt\Big)^{\frac{1}{\rho}}\Big\|_{\ell^r}.$$

Using the Minkowski inequality, it is easy to verify that

$$L^\rho_T(\dot{B}^s_{p,r}) \subseteq \widetilde{L}^\rho_T(\dot{B}^s_{p,r}) \quad \text{if} \quad \rho \leq r \quad \text{and} \quad \widetilde{L}^\rho_T(\dot{B}^s_{p,r}) \subseteq L^\rho_T(\dot{B}^s_{p,r}) \quad \text{if} \quad \rho \geq r.$$

Next, we introduce a hybrid-index Besov space which plays an important role in the study of compressible fluids and was initiated in [11, 12].

DEFINITION 1.4. *Let s, $\sigma \in \mathbb{R}$, and set*

$$\|f\|_{\widetilde{B}^{s,\sigma}_2} \triangleq \sum_{k\leq 0} 2^{ks}\|\Delta_k f\|_2 + \sum_{k>0} 2^{k\sigma}\|\Delta_k f\|_2.$$

Let $m=-[\frac{d}{2}+1-s]$; we define

$$\widetilde{B}_2^{s,\sigma}(\mathbb{R}^d)=\{f\in\mathcal{S}'(\mathbb{R}^d):\|f\|_{\widetilde{B}_2^{s,\sigma}}<+\infty\}\quad \textit{if}\quad m<0,$$

$$\widetilde{B}_2^{s,\sigma}(\mathbb{R}^d)=\{f\in\mathcal{S}'(\mathbb{R}^d)/\mathcal{P}_m:\|f\|_{\widetilde{B}_2^{s,\sigma}}<+\infty\}\quad \textit{if}\quad m\geq 0,$$

where $\mathcal{P}_m$ denotes the set of polynomials of degree $\leq m$.

Throughout this paper, we will denote $\dot{B}^s_{2,1}$ by B^s and $\widetilde{B}^{s,\sigma}_2$ by $\widetilde{B}^{s,\sigma}$. The following facts can be easily verified by using the definition of $\widetilde{B}^{s,\sigma}$:

(i) $\widetilde{B}^{s,s}=\dot{B}^s_{2,1}$.

(ii) If $s\leq\sigma$, then $\widetilde{B}^{s,\sigma}=\dot{B}^s_{2,1}\cap\dot{B}^\sigma_{2,1}$. Otherwise, $\widetilde{B}^{s,\sigma}=\dot{B}^s_{2,1}+\dot{B}^\sigma_{2,1}$.

Now we state our main result as follows.

THEOREM 1.2. *Let $\bar{h}_0$ be a positive constant. Assume that*

(i) $(u_0,h_0-\bar{h}_0)\in B^0(\mathbb{R}^2)\otimes\widetilde{B}^{0,1}(\mathbb{R}^2)$;

(ii) $h_0\geq\bar{h}_0$.

Then there exist a positive time T and a unique solution (u,h) of (1.1) such that

$$\begin{aligned}&u\in C([0,T];B^0)\cap L^1(0,T;B^2),\\ &h-\bar{h}_0\in C([0,T];\widetilde{B}^{0,1})\cap L^1(0,T;\widetilde{B}^{2,1}),\quad h\geq\frac{1}{2}\bar{h}_0.\end{aligned}\tag{1.6}$$

Moreover, there exists a strictly positive constant c such that if

$$\|u_0\|_{B^0}+\|h_0-\bar{h}_0\|_{\widetilde{B}^{0,1}}\leq c,\tag{1.7}$$

then we can choose $T=+\infty$.

The structure of this paper is as follows. In section 2, we recall some useful multilinear estimates in the Besov spaces. In section 3, we prove the existence of the solution. In section 4, we prove the uniqueness of the solution. Finally, in the appendix, we prove some multilinear estimates in the weighted Besov spaces.

Throughout the paper, C denotes various "harmless" large finite constants, and c denotes various "harmless" small constants. We shall sometimes use $X\lesssim Y$ to denote the estimate $X\leq CY$ for some constant C. We denote $\|\cdot\|_p$ by the L^p norm of a function.

2. Multilinear estimates in the Besov spaces. Let us first recall Bony's paraproduct decomposition.

DEFINITION 2.1. *We shall use the following Bony paraproduct decomposition (see* [1, 3]*):*

$$fg=T_fg+T_gf+R(f,g),\tag{2.1}$$

with

$$T_fg=\sum_{k\in\mathbb{Z}}S_{k-1}f\Delta_kg\quad\textit{and}\quad R(f,g)=\sum_{k\in\mathbb{Z}}\sum_{|k'-k|\leq1}\Delta_kf\Delta_{k'}g.\tag{2.2}$$

Next, let us recall some useful lemmas and multilinear estimates in the Besov spaces.

LEMMA 2.1 (Bernstein's inequality). *Let $1\leq p\leq q\leq+\infty$. Assume that $f\in\mathcal{S}'(\mathbb{R}^d)$; then for any $\gamma\in\mathbb{Z}^d$, there exist constants C_1, C_2 independent of f, j*

such that

$$\mathrm{supp}\hat{f} \subseteq \{|\xi| \le A_0 2^j\} \Rightarrow \|\partial^\gamma f\|_q \le C_1 2^{j|\gamma|+jd(\frac{1}{p}-\frac{1}{q})}\|f\|_p,$$

$$\mathrm{supp}\hat{f} \subseteq \{A_1 2^j \le |\xi| \le A_2 2^j\} \Rightarrow \|f\|_p \le C_2 2^{-j|\gamma|} \sup_{|\beta|=|\gamma|} \|\partial^\beta f\|_p.$$

The proof can be found in [8].

PROPOSITION 2.2. *If $s > 0$, $f, g \in B^s \cap L^\infty$, then $fg \in B^s \cap L^\infty$ and*

$$\|fg\|_{B^s} \le C(\|f\|_\infty \|g\|_{B^s} + \|g\|_\infty \|f\|_{B^s}). \tag{2.3}$$

If s_1, $s_2 \le \frac{d}{2}$ such that $s_1 + s_2 > 0$, $f \in B^{s_1}$, and $g \in B^{s_2}$, then $fg \in B^{s_1+s_2-\frac{d}{2}}$ and

$$\|fg\|_{B^{s_1+s_2-\frac{d}{2}}} \le C\|f\|_{B^{s_1}}\|g\|_{B^{s_2}}. \tag{2.4}$$

If $|s| < \frac{d}{2}$, $1 \le r \le +\infty$, $f \in \dot{B}^s_{2,r}$, and $g \in B^{\frac{d}{2}}$, then $fg \in \dot{B}^s_{2,r}$ and

$$\|fg\|_{\dot{B}^s_{2,r}} \le C\|f\|_{\dot{B}^s_{2,r}}\|g\|_{B^{\frac{d}{2}}}. \tag{2.5}$$

If $s \in (-\frac{d}{2}, \frac{d}{2}]$, $f \in B^s$, and $g \in \dot{B}^{-s}_{2,\infty}$, then $fg \in \dot{B}^{-\frac{d}{2}}_{2,\infty}$ and

$$\|fg\|_{\dot{B}^{-\frac{d}{2}}_{2,\infty}} \le C\|f\|_{B^s}\|g\|_{\dot{B}^{-s}_{2,\infty}}. \tag{2.6}$$

If $1 \le \rho_1, \rho_2, \rho \le \infty, s \in (-\frac{d}{2}, \frac{d}{2}]$, $f \in \widetilde{L}^{\rho_1}_T(B^s)$, and $g \in \widetilde{L}^{\rho_2}_T(\dot{B}^{-s}_{2,\infty})$, then there holds

$$\|fg\|_{\widetilde{L}^\rho_T(\dot{B}^{-\frac{d}{2}}_{2,\infty})} \le C\|f\|_{\widetilde{L}^{\rho_1}_T(B^s)}\|g\|_{\widetilde{L}^{\rho_2}_T(\dot{B}^{-s}_{2,\infty})}, \tag{2.7}$$

where $\frac{1}{\rho_1} + \frac{1}{\rho_2} = \frac{1}{\rho}$.

Proof. For the sake of simplicity, we present only the proof of (2.4); the others can be deduced in the same way (see also [14, 19]). By Bony's paraproduct decomposition and the property of quasi-orthogonality (1.5), for fixed $j \in \mathbb{Z}$, we write

$$\begin{aligned}\Delta_j(fg) &= \sum_{|k-j|\le 3} \Delta_j(S_{k-1}f\Delta_k g) + \sum_{|k-j|\le 3} \Delta_j(S_{k-1}g\Delta_k f) \\ &\quad + \sum_{k\ge j-2}\sum_{|k-k'|\le 1} \Delta_j(\Delta_k f \Delta_{k'} g) \\ &\triangleq I + II + III.\end{aligned}$$

Thanks to the definition of Besov space B^s, we have

$$\begin{aligned}\|fg\|_{B^{s_1+s_2-\frac{d}{2}}} &\le \Big(\sum_{j\in\mathbb{Z}} 2^{(s_1+s_2-\frac{d}{2})j}\|I\|_2\Big) + \cdots + \Big(\sum_{j\in\mathbb{Z}} 2^{(s_1+s_2-\frac{d}{2})j}\|III\|_2\Big) \\ &\triangleq I' + II' + III'. \end{aligned} \tag{2.8}$$

It suffices to estimate the above three terms separately. Using Young's inequality and Lemma 2.1, we have

$$\begin{aligned}\|\Delta_j(S_{k-1}f\Delta_k g)\|_2 &\lesssim \|S_{k-1}f\|_\infty\|\Delta_k g\|_2 \lesssim \sum_{k'\leq k-2}\|\Delta_{k'}f\|_\infty\|\Delta_k g\|_2\\ &\lesssim \sum_{k'\leq k-2}2^{k's_1}\|\Delta_{k'}f\|_2 2^{k'(\frac{d}{2}-s_1)}\|\Delta_k g\|_2\\ &\lesssim \|f\|_{B^{s_1}}\|\Delta_k g\|_2 2^{k(\frac{d}{2}-s_1)},\end{aligned}$$

where we have used the fact $s_1 \leq \frac{d}{2}$ in the last inequality. Hence, we get

$$\begin{aligned}I' &\lesssim \|f\|_{B^{s_1}}\sum_{j\in\mathbb{Z}}2^{(s_1+s_2-\frac{d}{2})j}\sum_{|k-j|\leq 3}2^{k(\frac{d}{2}-s_1)}\|\Delta_k g\|_2\\ &\lesssim \|f\|_{B^{s_1}}\sum_{|\ell|\leq 3}2^{-(s_1+s_2-\frac{d}{2})\ell}\sum_{j\in\mathbb{Z}}2^{s_2(j+\ell)}\|\Delta_{j+\ell}g\|_2 \lesssim \|f\|_{B^{s_1}}\|g\|_{B^{s_2}}. \qquad (2.9)\end{aligned}$$

Similarly, using the fact $s_2 \leq \frac{d}{2}$, we can obtain

$$II' \lesssim \|f\|_{B^{s_1}}\|g\|_{B^{s_2}}. \qquad (2.10)$$

Now we turn to estimate III'. From Lemma 2.1 and the Hölder inequality, it follows that

$$\|\Delta_j(\Delta_k f\Delta_{k'}g)\|_2 \lesssim 2^{j\frac{d}{2}}\|\Delta_k f\Delta_{k'}g\|_1 \lesssim 2^{j\frac{d}{2}}\|\Delta_k f\|_2\|\Delta_{k'}g\|_2.$$

So, we get by the Minkowski inequality that for $s_1+s_2>0$

$$\begin{aligned}III' &\lesssim \sum_{j\in\mathbb{Z}}2^{(s_1+s_2-\frac{d}{2})j}2^{j\frac{d}{2}}\Big(\sum_{k\geq j-2}\sum_{|k-k'|\leq 1}\|\Delta_k f\|_2\|\Delta_{k'}g\|_2\Big)\\ &\lesssim \sum_{\ell\geq -2}2^{-(s_1+s_2)\ell}\sum_{j\in\mathbb{Z}}2^{s_1(j+\ell)}\|\Delta_{j+\ell}f\|_2\|g\|_{B^{s_2}} \lesssim \|f\|_{B^{s_1}}\|g\|_{B^{s_2}}. \qquad (2.11)\end{aligned}$$

Summing up (2.8)–(2.11), we get the desired inequality (2.4). □

PROPOSITION 2.3. (1) *Let $s>0$. Assume that $F\in W^{[s]+2,\infty}_{loc}(\mathbb{R}^d)$ such that $F(0)=0$. Then there exists a constant $C(s,d,F)$ such that if $u\in B^s\cap L^\infty$, it holds that*

$$\|F(u)\|_{B^s} \leq C(1+\|u\|_\infty)^{[s]+1}\|u\|_{B^s}; \qquad (2.12)$$

and if $u\in \dot{B}^s_{2,\infty}\cap L^\infty$, it holds that

$$\|F(u)\|_{\dot{B}^s_{2,\infty}} \leq C(1+\|u\|_\infty)^{[s]+1}\|u\|_{\dot{B}^s_{2,\infty}}. \qquad (2.13)$$

(2) *Assume that $G\in W^{[\frac{d}{2}]+3,\infty}_{loc}(\mathbb{R}^d)$ such that $G'(0)=0$. Then there exists a functions $C(s,d,G)$ such that if $-\frac{d}{2}<s\leq\frac{d}{2}$, $u,v\in B^{\frac{d}{2}}\cap L^\infty$ and $u-v\in B^s$, it holds that*

$$\|G(u)-G(v)\|_{B^s} \leq C(\|u\|_\infty,\|v\|_\infty)(\|u\|_{B^{\frac{d}{2}}}+\|v\|_{B^{\frac{d}{2}}})\|u-v\|_{B^s}; \qquad (2.14)$$

and if $|s| < \frac{d}{2}$, $u, v \in B^{\frac{d}{2}} \cap L^\infty$ *and* $u - v \in \dot{B}^s_{2,\infty}$, *it holds that*

$$\|G(u) - G(v)\|_{\dot{B}^s_{2,\infty}} \le C(\|u\|_\infty, \|v\|_\infty)(\|u\|_{B^{\frac{d}{2}}} + \|v\|_{B^{\frac{d}{2}}})\|u - v\|_{\dot{B}^s_{2,\infty}}. \tag{2.15}$$

Proof. We can refer to [2, 19] for the proof of (1). For (2), we refer to [11, 15]. For example, we write

$$G(u) - G(v) = (u - v)\int_0^1 G'(v + \tau(u - v))d\tau.$$

Then it follows from (2.5) that for $|s| < \frac{d}{2}$

$$\|G(u) - G(v)\|_{\dot{B}^s_{2,\infty}} \le C\|u - v\|_{\dot{B}^s_{2,\infty}}\|G'(v + \tau(u - v))\|_{B^{\frac{d}{2}}},$$

which together with (2.12) implies (2.15). □

PROPOSITION 2.4. *Let* A *be a homogeneous smooth function of degree* m. *Assume that* $-\frac{d}{2} < s_1, t_1, s_2, t_2 \le 1 + \frac{d}{2}$. *Then it holds that if* $k \ge 1$,

$$\begin{aligned} &|(A(D)\Delta_k(v \cdot \nabla f), A(D)\Delta_k f)| \\ &\lesssim \alpha_k 2^{-k(s_2 - m)}\|v\|_{B^{\frac{d}{2}+1}}\|f\|_{\widetilde{B}^{s_1,s_2}}\|A(D)\Delta_k f\|_2; \end{aligned} \tag{2.16}$$

if $k \le 0$,

$$\begin{aligned} &|(A(D)\Delta_k(v \cdot \nabla f), A(D)\Delta_k f)| \\ &\lesssim \alpha_k 2^{-k(s_1 - m)}\|v\|_{B^{\frac{d}{2}+1}}\|f\|_{\widetilde{B}^{s_1,s_2}}\|A(D)\Delta_k f\|_2; \end{aligned} \tag{2.17}$$

if $k \ge 1$,

$$\begin{aligned} &|(A(D)\Delta_k(v \cdot \nabla f), \Delta_k g) + (\Delta_k(v \cdot \nabla g), A(D)\Delta_k f)| \\ &\lesssim \alpha_k\|v\|_{B^{\frac{d}{2}+1}}(2^{-kt_2}\|g\|_{\widetilde{B}^{t_1,t_2}}\|A(D)\Delta_k f\|_2 + 2^{-k(s_2 - m)}\|f\|_{\widetilde{B}^{s_1,s_2}}\|\Delta_k g\|_2); \end{aligned} \tag{2.18}$$

and if $k \le 0$,

$$\begin{aligned} &|(A(D)\Delta_k(v \cdot \nabla f), \Delta_k g) + (\Delta_k(v \cdot \nabla g), A(D)\Delta_k f)| \\ &\lesssim \alpha_k\|v\|_{B^{\frac{d}{2}+1}}(2^{-kt_1}\|g\|_{\widetilde{B}^{t_1,t_2}}\|A(D)\Delta_k f\|_2 + 2^{-k(s_1 - m)}\|f\|_{\widetilde{B}^{s_1,s_2}}\|\Delta_k g\|_2), \end{aligned} \tag{2.19}$$

where $\sum_{k\in\mathbb{Z}} \alpha_k \le 1$.

For the proof we refer to [12].

3. Existence. In this section, we prove the existence of the solution for the 2D viscous shallow water equations. Without loss of generality, we assume that $\bar{h}_0 = 1$ and $\nu = 1$. Replacing h by $h + 1$ in (1.1), we rewrite (1.1) as

$$\begin{cases} h_t + \mathrm{div} u + \mathrm{div}(hu) = 0, \\ u_t - (\nabla \cdot D(u) + \nabla \mathrm{div}\, u) + u \cdot \nabla u - \frac{\nabla h}{1+h}(D(u) + \mathrm{div}\, u) + \nabla h = 0, \\ u(0, \cdot) = u_0, \quad h(0, \cdot) = h_0. \end{cases} \tag{3.1}$$

3.1. The linearized system. In this subsection, we consider the linearized system of (3.1):

$$\begin{cases} h_t + v\cdot\nabla h + \mathrm{div} u = \mathcal{H}, \\ u_t - (\nabla\cdot D(u) + \nabla \mathrm{div}\, u) + v\cdot\nabla u + \nabla h = \mathcal{G}, \\ u(0,\cdot) = u_0, \quad h(0,\cdot) = h_0. \end{cases} \tag{3.2}$$

Let us first introduce some definitions. Set

$$e_k^r(t) \triangleq (1 - e^{-cr2^{2k}t})^{\frac{1}{r}}, \quad \omega_k(t) = \sum_{\widetilde{k}\geq k} 2^{-(\widetilde{k}-k)}(e_{\widetilde{k}}^1(t) + e_{\widetilde{k}}^2(t)),$$

where c is a positive constant which will be determined later. We remark that

$$\omega_k(t) \leq C \quad \text{for any } k\in\mathbb{Z},$$

which will be constantly used in the following.

Definition 3.1. *Let $s\in\mathbb{R}$ and $T>0$. The function space E_T^s is defined by*

$$E_T^s = \{f\in\mathcal{Z}'((0,T)\times\mathbb{R}^d) : \|f\|_{E_T^s} < +\infty\},$$

where

$$\|f\|_{E_T^s} \triangleq \sum_{k\in Z} 2^{ks}\omega_k(T)\|\Delta_k f\|_{L_T^\infty(L^2)}.$$

Definition 3.2. *Let $s_1, s_2\in\mathbb{R}$ and $T>0$. The function space $\widetilde{E}_T^{s_1,s_2}$ is defined by*

$$\widetilde{E}_T^{s_1,s_2} = \{f\in\mathcal{Z}'((0,T)\times\mathbb{R}^d) : \|f\|_{\widetilde{E}_T^{s_1,s_2}} < +\infty\},$$

where

$$\|f\|_{\widetilde{E}_T^{s_1,s_2}} \triangleq \sum_{k\leq 0} 2^{ks_1}\omega_k(T)\|\Delta_k f\|_{L_T^\infty(L^2)} + \sum_{k\geq 1} 2^{ks_2}\omega_k(T)\|\Delta_k f\|_{L_T^\infty(L^2)}.$$

Remark 3.1. If $s_1\leq s_2$, then $\widetilde{E}_T^{s_1,s_2} = E_T^{s_1}\cap E_T^{s_2}$. Otherwise, $\widetilde{E}_T^{s_1,s_2} = E_T^{s_1} + E_T^{s_2}$.

Let (u,h) be a smooth solution of (3.2). We want to establish the following a priori estimates for (h,u):

$$\begin{aligned} &\|u\|_{L_T^1(B^2)} + \|u\|_{L_T^2(B^1)} + \|h\|_{\widetilde{E}_T^{0,1}} \\ &\quad\leq C\sum_{k\in\mathbb{Z}}\omega_k(T)E_k(0) + C\sum_{k\in\mathbb{Z}}\omega_k(T)\|\Delta_k\mathcal{G}(t)\|_{L_T^1(L^2)} \\ &\qquad + C\sum_{k\geq 1}\omega_k(T)\|\nabla\Delta_k\mathcal{H}(t)\|_{L_T^1(L^2)} + C\sum_{k<1}\omega_k(T)\|\Delta_k\mathcal{H}(t)\|_{L_T^1(L^2)} \\ &\qquad + C\|u\|_{L_T^2(B^1)}\|v\|_{L_T^2(B^1)} + C\|h\|_{\widetilde{E}_T^{0,1}}\|v\|_{L_T^1(B^2)} \end{aligned} \tag{3.3}$$

and

$$\begin{aligned} &\|u\|_{L_T^\infty(B^0)} + \|h\|_{L_T^\infty(\widetilde{B}^{0,1})} + \|h\|_{L_T^1(\widetilde{B}^{2,1})} \\ &\leq E_0 + C\left(\|\mathcal{H}\|_{L_T^1(\widetilde{B}^{0,1})} + \|\mathcal{G}\|_{L_T^1(B^0)} + \int_0^T V'(t)(\|u(t)\|_{B^0} + \|h(t)\|_{\widetilde{B}^{0,1}})dt\right), \end{aligned} \tag{3.4}$$

where $V(t) = \|v(t')\|_{L^1_t(B^2)}$ and

$$E_0 = \sum_{k\in\mathbb{Z}} E_k(0), \quad E_k(t) = \begin{cases} E_{hk}(t), & k \geq 1, \\ E_{lk}(t), & k < 1, \end{cases}$$

with

$$E^2_{hk}(t) = \frac{1}{2}\|u_k(t)\|^2_2 + \|\nabla h_k(t)\|^2_2 + (u_k(t), \nabla h_k(t)),$$

$$E^2_{lk}(t) = \frac{1}{2}\|u_k(t)\|^2_2 + \frac{1}{2}\|h_k(t)\|^2_2 + \frac{1}{8}(u_k(t), \nabla h_k(t)).$$

Let us begin with the proof of (3.3) and (3.4). Set

$$u_k = \Delta_k u, \quad h_k = \Delta_k h, \quad \mathcal{H}^k = \Delta_k \mathcal{H}, \quad \mathcal{G}^k = \Delta_k \mathcal{G}.$$

Then we get by applying the operator Δ_k to (3.2) that

$$\text{(3.5)} \qquad \begin{cases} \partial_t h_k + \Delta_k(v \cdot \nabla h) + \mathrm{div} u_k = \mathcal{H}_k, \\ \partial_t u_k - (\nabla \cdot D(u_k) + \nabla \mathrm{div}\, u_k) + \Delta_k(v \cdot \nabla u) + \nabla h_k = \mathcal{G}_k, \\ u_k(0, \cdot) = \Delta_k u_0, \quad h_k(0, \cdot) = \Delta_k h_0. \end{cases}$$

Multiplying the second equation of (3.5) by u_k, and integrating the resulting equation over $\mathbb{R}^2$, we obtain

$$\text{(3.6)} \quad \frac{1}{2}\frac{d}{dt}\|u_k\|^2_2 + \frac{1}{2}\|\nabla u_k\|^2_2 + \frac{3}{2}\|\mathrm{div} u_k\|^2_2 + (\nabla h_k, u_k) = (\mathcal{G}_k, u_k) - (\Delta_k(v \cdot \nabla u), u_k).$$

In the following, we will deal with the high frequency and the low frequency of h in a different manner.

High frequencies: $k \geq 1$. First, applying ∇ to the first equation of (3.5) and multiplying it by ∇h_k, then integrating the resulting equation over $\mathbb{R}^2$, we obtain

$$\text{(3.7)} \quad \frac{1}{2}\frac{d}{dt}\|\nabla h_k\|^2_2 + (\nabla \mathrm{div} u_k, \nabla h_k) = (\nabla \mathcal{H}_k, \nabla h_k) - (\nabla \Delta_k(v \cdot \nabla h), \nabla h_k).$$

Second, applying the operator ∇ to the first equation of (3.5) and taking the L^2 product of the resulting equation with u_k, then taking the L^2 product of the second equation of (3.5) with ∇h_k, we get by summing them up that

$$\begin{aligned} &\frac{d}{dt}(u_k, \nabla h_k) - \|\mathrm{div} u_k\|^2_2 - 2(\nabla \mathrm{div} u_k, \nabla h_k) + \|\nabla h_k\|^2_2 \\ &\text{(3.8)} \qquad = (\nabla \mathcal{H}_k, u_k) + (\mathcal{G}_k, \nabla h_k) - (\nabla \Delta_k(v \cdot \nabla h), u_k) - (\Delta_k(v \cdot \nabla u), \nabla h_k), \end{aligned}$$

where we used the fact that

$$(\nabla \cdot D(u_k) + \nabla \mathrm{div}\, u_k, \nabla h_k) = 2(\nabla \mathrm{div} u_k, \nabla h_k).$$

Then we get by summing up (3.6), (3.7)×2, and (3.8) that

$$\begin{aligned}
&\frac{d}{dt}\left[\frac{1}{2}\|u_k\|_2^2+\|\nabla h_k\|_2^2+(u_k,\nabla h_k)\right]\\
&\qquad+\left[\|\nabla h_k\|_2^2+\frac{1}{2}\|\nabla u_k\|_2^2+\frac{1}{2}\|\mathrm{div}u_k\|_2^2+(\nabla h_k,u_k)\right]\\
&=\Big[(\nabla\mathcal{H}_k,u_k)+2(\nabla\mathcal{H}_k,\nabla h_k)+(\mathcal{G}_k,u_k)+(\mathcal{G}_k,\nabla h_k)\Big]\\
&\quad-(\Delta_k(v\cdot\nabla u),u_k)-2(\nabla\Delta_k(v\cdot\nabla h),\nabla h_k)\\
&\quad-\Big[(\nabla\Delta_k(v\cdot\nabla h),u_k)+(\Delta_k(v\cdot\nabla u),\nabla h_k)\Big]\\
&\triangleq I+II+III+IV. \qquad (3.9)
\end{aligned}$$

Note that

$$(u_k,\nabla h_k)\le\frac{1}{3}\|u_k\|_2^2+\frac{3}{4}\|\nabla h_k\|_2^2;$$

hence, we get by the definition of E_{hk} that

$$\frac{1}{6}(\|u_k\|_2^2+\|\nabla h_k\|_2^2)\le E_{hk}^2\le 2(\|u_k\|_2^2+\|\nabla h_k\|_2^2). \qquad (3.10)$$

Similarly, using the fact that $\frac{5}{6}2^k\ge\frac{5}{3}$ and (3.10), we have

$$\|\nabla h_k\|_2^2+\frac{1}{2}\|\nabla u_k\|_2^2+\frac{1}{2}\|\mathrm{div}u_k\|_2^2+(\nabla h_k,u_k)\ge\frac{1}{8}E_{hk}^2. \qquad (3.11)$$

By summing up (3.9)–(3.11), we obtain

$$\frac{d}{dt}E_{hk}^2+cE_{hk}^2\le C|I+II+III+IV|. \qquad (3.12)$$

In order to obtain (3.3), we use Lemma 5.1 to deal with the right-hand terms of (3.12). First, we get by using the Cauchy–Schwarz inequality and (3.10) that

$$|I|\le C(\|\nabla\mathcal{H}_k(t)\|_2+\|\mathcal{G}_k(t)\|_2)E_{hk}. \qquad (3.13)$$

From Lemma 5.1 and (3.10), it follows that

$$|II+III+IV|\le C(\|\mathcal{F}_k^1(t)\|_2+\|\widetilde{\mathcal{F}}_k^0(t)\|_2)E_{hk}. \qquad (3.14)$$

By summing up (3.12), and (3.13)–(3.14), we obtain

$$\frac{d}{dt}E_{hk}+cE_{hk}\le C\Big(\|\nabla\mathcal{H}_k(t)\|_2+\|\mathcal{G}_k(t)\|_2+\|\mathcal{F}_k^1(t)\|_2+\|\widetilde{\mathcal{F}}_k^0(t)\|_2\Big), \qquad (3.15)$$

which implies that

$$\begin{aligned}
\|E_{hk}(t)\|_{L_T^\infty}\le&E_{hk}(0)+C\Big(\|\nabla\mathcal{H}_k(t)\|_{L_T^1(L^2)}+\|\mathcal{G}_k(t)\|_{L_T^1(L^2)}\\
&+\|\mathcal{F}_k^1(t)\|_{L_T^1(L^2)}+\|\widetilde{\mathcal{F}}_k^0(t)\|_{L_T^1(L^2)}\Big). \qquad (3.16)
\end{aligned}$$

Furthermore, by (5.4) and (5.5), it holds that

$$\sum_{k\in\mathbb{Z}}\omega_k(T)\Big(\|\mathcal{F}^1_k(t)\|_{L^1_T(L^2)}+\|\widetilde{\mathcal{F}}^0_k(t)\|_{L^1_T(L^2)}\Big)$$
$$\leq C\Big(\|u\|_{L^2_T(B^1)}\|v\|_{L^2_T(B^1)}+\|h\|_{E^1_T}\|v\|_{L^1_T(B^2)}\Big).$$

Multiplying $\omega_k(T)$ on both sides of (3.16), then summing up the resulting equation over $k\geq 1$, we obtain

$$\begin{aligned}\sum_{k\geq1}\omega_k(T)\|E_{hk}(t)\|_{L^\infty_T}\leq&\sum_{k\geq1}\omega_k(T)E_{hk}(0)\\&+C\sum_{k\geq1}\omega_k(T)\Big(\|\nabla\mathcal{H}_k(t)\|_{L^1_T(L^2)}+\|\mathcal{G}_k(t)\|_{L^1_T(L^2)}\Big)\\&+C\Big(\|u\|_{L^2_T(B^1)}\|v\|_{L^2_T(B^1)}+\|h\|_{E^1_T}\|v\|_{L^1_T(B^2)}\Big).\end{aligned}\tag{3.17}$$

Next, we use the decay effect of the parabolic operators to estimate $\|u\|_{L^2_T(B^1)\cap L^1_T(B^2)}$. It follows from (3.6) and Lemma 5.1 that

$$\frac{d}{dt}\|u_k\|_2+c2^{2k}\|u_k\|_2\leq C(\|\nabla h_k(t)\|_2+\|\mathcal{G}_k(t)\|_{L^2}+\|\widetilde{\mathcal{F}}^0_k(t)\|_{L^2}),$$

which implies that

$$\|u_k\|_2\leq e^{-ct2^{2k}}\|u_k(0)\|_2+Ce^{-ct2^{2k}}*_t\Big(\|\nabla h_k(t)\|_2+\|\mathcal{G}_k(t)\|_2+\|\widetilde{\mathcal{F}}^0_k(t)\|_2\Big),$$

where the sign $*$ denotes the convolution of functions defined in $\mathbb{R}^+$; more precisely,

$$e^{-ct2^{2k}}*_t f\triangleq\int_0^t e^{-c(t-\tau)2^{2k}}f(\tau)d\tau.$$

Taking the L^r norm for $r=1,2$ with respect to t, we get by using Young's inequality that

$$\|u_k\|_{L^r_T(L^2)}\leq C2^{-2k/r}e^r_k(T)\Big(\|u_k(0)\|_2+\|\nabla h_k\|_{L^1_T(L^2)}+\|\mathcal{G}_k\|_{L^1_T(L^2)}+\|\widetilde{\mathcal{F}}^0_k\|_{L^1_T(L^2)}\Big),$$

which together with (5.5) implies that

$$\begin{aligned}\sum_{k\geq1}\Big(2^{2k}\|u_k\|_{L^1_T(L^2)}+2^k\|u_k\|_{L^2_T(L^2)}\Big)\leq&C\sum_{k\geq1}\omega_k(T)\|u_k(0)\|_2\\&+C\sum_{k\geq1}\omega_k(T)\Big(\|\nabla h_k\|_{L^1_T(L^2)}+\|\mathcal{G}_k\|_{L^1_T(L^2)}\Big)+C\|u\|_{L^2_T(B^1)}\|v\|_{L^2_T(B^1)},\end{aligned}\tag{3.18}$$

where we used the fact that

$$e^1_k(T)+e^2_k(T)\leq\omega_k(T).$$

On the other hand, it follows from (3.15) that

$$\|E_{hk}\|_2\leq e^{-ct}E_{hk}(0)+Ce^{-ct}*_t\Big(\|\nabla\mathcal{H}_k(t)\|_2+\|\mathcal{G}_k(t)\|_2+\|\mathcal{F}^1_k(t)\|_2+\|\widetilde{\mathcal{F}}^0_k(t)\|_2\Big).$$

Taking the L^1 norm with respect to t, we get by using Young's inequality that

$$\|E_{hk}\|_{L^1_T} \leq C(1-e^{-cT})E_{hk}(0) + C(1-e^{-cT})\Big(\|\nabla\mathcal{H}_k(t)\|_{L^1_T(L^2)} + \|\mathcal{G}_k(t)\|_{L^1_T(L^2)}$$

$$+ \|\mathcal{F}^1_k(t)\|_{L^1_T(L^2)} + \|\widetilde{\mathcal{F}}^0_k(t)\|_{L^1_T(L^2)}\Big). \tag{3.19}$$

Note that for $k \geq 1$

$$1 - e^{-ct} \leq 1 - e^{-ct2^{2k}} \leq \omega_k(t),$$

which together with (3.19) and Lemma 5.1 gives

$$\sum_{k\geq 1}\|E_{hk}\|_{L^1_T} \leq C\sum_{k\geq 1}\omega_k(T)E_{hk}(0) + C\sum_{k\geq 1}\omega_k(T)\Big(\|\nabla\mathcal{H}_k(t)\|_{L^1_T(L^2)} + \|\mathcal{G}_k(t)\|_{L^1_T(L^2)}\Big)$$

$$+ C\Big(\|u\|_{L^2_T(B^1)}\|v\|_{L^2_T(B^1)} + \|h\|_{E^1_T}\|v\|_{L^1_T(B^2)}\Big). \tag{3.20}$$

Plugging (3.20) into (3.18), we obtain

$$\sum_{k\geq 1}\Big(2^{2k}\|u_k\|_{L^1_T(L^2)} + 2^k\|u_k\|_{L^2_T(L^2)}\Big)$$

$$\leq C\sum_{k\geq 1}\omega_k(T)E_{hk}(0) + C\sum_{k\geq 1}\omega_k(T)\Big(\|\nabla\mathcal{H}_k(t)\|_{L^1_T(L^2)} + \|\mathcal{G}_k(t)\|_{L^1_T(L^2)}\Big)$$

$$+ C\Big(\|u\|_{L^2_T(B^1)}\|v\|_{L^2_T(B^1)} + \|h\|_{E^1_T}\|v\|_{L^1_T(B^2)}\Big). \tag{3.21}$$

On the other hand, in order to obtain (3.4), we use Proposition 2.4 to deal with the right-hand terms of (3.12). Applying (2.16) with $s_1 = s_2 = 0$ to II, (2.16) with $s_1 = 0, s_2 = 1$ to III, and (2.18) with $t_1 = t_2 = 0, s_1 = 0, s_2 = 1$ to IV, we obtain

$$|II + III + IV| \leq CE_{hk}\alpha_k V'(t)(\|u\|_{B^0} + \|h\|_{\widetilde{B}^{0,1}}), \tag{3.22}$$

with $\sum_{k\in\mathbb{Z}}\alpha_k \leq 1$ and $V(t) = \|v(t')\|_{L^1_t(B^2)}$. From (3.13) and (3.22), it follows that

$$\frac{d}{dt}E_{hk} + cE_{hk} \leq C\Big(\|\nabla\mathcal{H}_k(t)\|_2 + \|\mathcal{G}_k(t)\|_2 + \alpha_k V'(t)(\|u\|_{B^0} + \|h\|_{\widetilde{B}^{0,1}})\Big),$$

from which a similar proof of (3.21) ensures that

$$\sum_{k\geq 1}\Big(\|E_{hk}\|_{L^1_T} + \|E_{hk}\|_{L^\infty_T}\Big) \leq C\sum_{k\geq 1}E_{hk}(0)$$

$$+C\Big(\|\mathcal{H}\|_{L^1_T(\widetilde{B}^{0,1})} + \|\mathcal{G}\|_{L^1_T(B^0)} + \int_0^T V'(t)(\|u(t)\|_{B^0} + \|h(t)\|_{\widetilde{B}^{0,1}})dt\Big). \tag{3.23}$$

Low frequencies: $k < 1$. Multiplying the first equation of (3.5) by h_k, we get by integrating the resulting equation over $\mathbb{R}^2$ that

$$\frac{1}{2}\frac{d}{dt}\|h_k\|_2^2 + (\mathrm{div}u_k, h_k) = (\mathcal{H}_k, h_k) - (\Delta_k(v\cdot\nabla h), h_k). \tag{3.24}$$

Summing up (3.6), (3.8) $\times \frac{1}{8}$, and (3.24), we obtain

$$\begin{aligned}
&\frac{d}{dt}\left[\frac{1}{2}\|u_k\|_2^2+\frac{1}{2}\|h_k\|_2^2+\frac{1}{8}(u_k,\nabla h_k)\right]\\
&\qquad+\left[\frac{1}{8}\|\nabla h_k\|_2^2+\frac{1}{2}\|\nabla u_k\|_2^2+\frac{11}{8}\|\mathrm{div} u_k\|_2^2-\frac{1}{4}(\nabla \mathrm{div}\, u_k,\nabla h_k)\right]\\
&=\left[\frac{1}{8}(\nabla\mathcal{H}_k,u_k)+(\mathcal{H}_k,h_k)+(\mathcal{G}_k,u_k)+\frac{1}{8}(\mathcal{G}_k,\nabla h_k)\right]\\
&\quad-(\Delta_k(v\cdot\nabla u),u_k)-(\Delta_k(v\cdot\nabla h),h_k)\\
&\quad-\frac{1}{8}\Big[(\nabla\Delta_k(v\cdot\nabla h),u_k)+(\Delta_k(v\cdot\nabla u),\nabla h_k)\Big]\\
&\triangleq I+II+III+IV. \qquad (3.25)
\end{aligned}$$

Note that $2^k\le 1$. We get by the Cauchy–Schwarz inequality that

$$\frac{1}{8}(u_k,\nabla h_k)\le\frac{3}{10}\|u_k\|_2\|h_k\|_2\le\frac{1}{4}\|u_k\|_2^2+\frac{1}{4}\|h_k\|_2^2;$$

hence, we get by the definition of E_{lk} that

$$\frac{1}{4}(\|u_k\|_2^2+\|h_k\|_2^2)\le E_{lk}^2\le 2(\|u_k\|_2^2+\|h_k\|_2^2). \qquad (3.26)$$

Similarly, we can prove

$$\frac{1}{4}(\nabla\mathrm{div}\, u_k,\nabla h_k)\le\frac{3}{5}\|\mathrm{div} u_k\|_2\|\nabla h_k\|_2\le\frac{9}{10}\|\nabla u_k\|_2^2+\frac{1}{10}\|\nabla h_k\|_2^2,$$

which together with (3.26) implies that

$$\begin{aligned}
&\frac{1}{8}\|\nabla h_k\|_2^2+\frac{1}{2}\|\nabla u_k\|_2^2+\frac{11}{8}\|\mathrm{div} u_k\|_2^2-\frac{1}{4}(\nabla\mathrm{div}\, u_k,\nabla h_k)\\
&\qquad\ge\frac{1}{160}2^{2k}(\|u_k\|_2^2+\|h_k\|_2^2)\ge\frac{1}{320}2^{2k}E_{lk}^2. \qquad (3.27)
\end{aligned}$$

By summing up (3.25)–(3.27), we obtain

$$\frac{d}{dt}E_{lk}^2+c2^{2k}E_{lk}^2\le C|I+II+III+IV|. \qquad (3.28)$$

In order to obtain (3.3), we use Lemma 5.1 to estimate the right-hand terms of (3.28). Using the fact that $2^k\le 1$, we get by the Cauchy–Schwarz inequality and (3.26) that

$$|I|\le C(\|\mathcal{H}_k(t)\|_2+\|\mathcal{G}_k(t)\|_2)E_{lk}. \qquad (3.29)$$

Using Lemma 5.1 and (3.26), we have

$$|II+III+IV|\le C(\|\mathcal{F}_k^1(t)\|_2+\|\mathcal{F}_k^0(t)\|_2+\|\widetilde{\mathcal{F}}_k^0(t)\|_2)E_{lk}. \qquad (3.30)$$

By summing up (3.28)–(3.30), we obtain

$$\frac{d}{dt}E_{lk}+c2^{2k}E_{lk}\leq C\Big(\|\mathcal{H}_k(t)\|_2+\|\mathcal{G}_k(t)\|_2+\|\mathcal{F}^1_k(t)\|_2+\|\mathcal{F}^0_k(t)\|_2+\|\widetilde{\mathcal{F}}^0_k(t)\|_2\Big),$$

which implies that

$$\begin{aligned}E_{lk}&\leq e^{-c2^{2k}t}E_{lk}(0)\\&\quad+Ce^{-c2^{2k}t}*_t\Big(\|\mathcal{H}_k(t)\|_2+\|\mathcal{G}_k(t)\|_2+\|\mathcal{F}^1_k(t)\|_2+\|\mathcal{F}^0_k(t)\|_2+\|\widetilde{\mathcal{F}}^0_k(t)\|_2\Big).\end{aligned}$$

Taking the L^r norm with respect to t, we get by using Young's inequality that

$$\begin{aligned}\|E_{lk}\|_{L^r_T}&\leq C2^{-2k/r}e^r_k(T)\Big(E_{lk}(0)+\|\mathcal{H}_k(t)\|_{L^1_T(L^2)}+\|\mathcal{G}_k(t)\|_{L^1_T(L^2)}\\&\qquad+\|\mathcal{F}^1_k(t)\|_{L^1_T(L^2)}+\|\mathcal{F}^0_k(t)\|_2+\|\widetilde{\mathcal{F}}^0_k(t)\|_{L^1_T(L^2)}\Big),\end{aligned}$$

from which, together with Lemma 5.1, it follows that

$$\begin{aligned}&\sum_{k<1}\omega_k(T)\|E_{lk}\|_{L^\infty_T}\\&\quad\leq C\sum_{k<1}\omega_k(T)E_{lk}(0)+\sum_{k<1}\omega_k(T)(\|\mathcal{H}_k(t)\|_{L^1_T(L^2)}+\|\mathcal{G}_k(t)\|_{L^1_T(L^2)})\\&\quad\quad+C\Big(\|u\|_{L^2_T(B^1)}\|v\|_{L^2_T(B^1)}+\|h\|_{\widetilde{E}^{0,1}_T}\|v\|_{L^1_T(B^2)}\Big)\end{aligned}\tag{3.31}$$

and

$$\begin{aligned}&\sum_{k<1}(2^{2k}\|E_{lk}\|_{L^1_T}+2^k\|E_{lk}\|_{L^2_T})\\&\quad\leq C\sum_{k<1}\omega_k(T)E_{lk}(0)+\sum_{k<1}\omega_k(T)(\|\mathcal{H}_k(t)\|_{L^1_T(L^2)}+\|\mathcal{G}_k(t)\|_{L^1_T(L^2)})\\&\quad\quad+C\Big(\|u\|_{L^2_T(B^1)}\|v\|_{L^2_T(B^1)}+\|h\|_{\widetilde{E}^{0,1}_T}\|v\|_{L^1_T(B^2)}\Big).\end{aligned}\tag{3.32}$$

On the other hand, in order to obtain (3.4), we use Proposition 2.4 to deal with the right-hand terms of (3.28). Applying (2.17) with $s_1=s_2=0$ to II, (2.17) with $s_1=0,s_2=1$ to III, and (2.19) with $t_1=t_2=0,s_1=0,s_2=1$ to IV, we obtain

$$|II+III+IV|\leq CE_{lk}\alpha_kV'(t)(\|u\|_{B^0}+\|h\|_{\widetilde{B}^{0,1}}),\tag{3.33}$$

with $\sum_{k\in\mathbb{Z}}\alpha_k\leq 1$ and $V(t)=\|v(t')\|_{L^1_t(B^2)}$. From (3.32) and (3.36), it follows that

$$\frac{d}{dt}E_{lk}+c2^{2k}E_{lk}\leq C\Big(\|\mathcal{H}_k(t)\|_2+\|\mathcal{G}_k(t)\|_2+\alpha_kV'(t)(\|u\|_{B^0}+\|h\|_{\widetilde{B}^{0,1}})\Big),$$

which, together with a similar proof of (3.21), ensures that

$$\begin{aligned}&\sum_{k<1}\big(2^{2k}\|E_{lk}\|_{L^1_T}+\|E_{lk}\|_{L^\infty_T}\big)\leq\sum_{k<1}E_{hk}(0)\\&\quad+C\left(\|\mathcal{H}\|_{L^1_T(\widetilde{B}^{0,1})}+\|\mathcal{G}\|_{L^1_T(B^0)}+\int_0^T V'(t)(\|u(t)\|_{B^0}+\|h(t)\|_{\widetilde{B}^{0,1}})dt\right).\end{aligned}\tag{3.34}$$

The completion of the a priori estimates. First, adding up (3.17), (3.21), (3.31), and (3.32) yields that

$$\begin{aligned}
&\|u\|_{L^1_T(B^2)} + \|u\|_{L^2_T(B^1)} + \|h\|_{\widetilde{E}^{0,1}_T} \\
&\quad \leq C\sum_{k\in\mathbb{Z}} \omega_k(T)E_k(0) + C\sum_{k\in\mathbb{Z}} \omega_k(T)\|\mathcal{G}_k(t)\|_{L^1_T(L^2)} \\
&\qquad + C\sum_{k\geq 1} \omega_k(T)\|\nabla\mathcal{H}_k(t)\|_{L^1_T(L^2)} + C\sum_{k<1} \omega_k(T)\|\mathcal{H}_k(t)\|_{L^1_T(L^2)} \\
&\qquad + C\|u\|_{L^2_T(B^1)}\|v\|_{L^2_T(B^1)} + C\|h\|_{\widetilde{E}^{0,1}_T}\|v\|_{L^1_T(B^2)}, \qquad (3.35)
\end{aligned}$$

where we used the fact that

$$\|h\|_{E^1_T} \leq C\|h\|_{\widetilde{E}^{0,1}_T}.$$

On the other hand, adding up (3.23) and (3.34) gives rise to

$$\begin{aligned}
&\|u\|_{L^\infty_T(B^0)} + \|h\|_{L^\infty_T(\widetilde{B}^{0,1})} + \|h\|_{L^1_T(\widetilde{B}^{2,1})} \\
&\leq E_0 + C\Big(\|\mathcal{H}\|_{L^1_T(\widetilde{B}^{0,1})} + \|\mathcal{G}\|_{L^1_T(B^0)} + \int_0^T V'(t)(\|u\|_{B^0} + \|h\|_{\widetilde{B}^{0,1}})dt\Big), \qquad (3.36)
\end{aligned}$$

which, together with the Gronwall inequality, implies that

$$\begin{aligned}
&\|u\|_{L^\infty_T(B^0)} + \|h\|_{L^\infty_T(\widetilde{B}^{0,1})} + \|h\|_{L^1_T(\widetilde{B}^{2,1})} \\
&\leq Ce^{C\|v\|_{L^1_T(B^2)}}\Big(E_0 + \|\mathcal{H}\|_{L^1_T(\widetilde{B}^{0,1})} + \|\mathcal{G}\|_{L^1_T(B^0)}\Big). \qquad (3.37)
\end{aligned}$$

Finally, let us remark that

$$E_0 \approx (\|h_0\|_{\widetilde{B}^{0,1}} + \|u_0\|_{B^0}).$$

3.2. The uniform estimate of the approximate sequence of solutions. In this subsection, we will construct the approximate solutions of (3.1) and present the uniform estimate of the approximate solutions. Let us first define the approximate sequence $(h^n, u^n)_{n\in\mathbb{N}}$ of (3.1) by the following system:

$$\begin{cases}
\partial_t h^{n+1} + u^n\cdot\nabla h^{n+1} + \mathrm{div}u^{n+1} = \mathcal{H}^n, \\
\partial_t u^{n+1} - (\nabla\cdot D(u^{n+1}) + \nabla\mathrm{div}\, u^{n+1}) + u^n\cdot\nabla u^{n+1} + \nabla h^{n+1} = \mathcal{G}^n, \\
(h^{n+1}, u^{n+1})|_{t=0} = \displaystyle\sum_{|k|\leq n+N} \Delta_k(h_0, u_0),
\end{cases} \qquad (3.38)$$

where

$$\mathcal{H}^n \triangleq -h^n\mathrm{div}u^n, \qquad \mathcal{G}^n \triangleq \frac{\nabla h^n}{1+h^n}\widetilde{\nabla}u^n, \quad \text{with} \quad \widetilde{\nabla}u^n = D(u^n) + \mathrm{div}u^n,$$

and N is a fixed large integer such that

$$1 + h^n(0) \geq \frac{3}{4} \qquad \text{for} \quad n \geq 1.$$

Set $(h^0,u^0)=(0,0)$ and solve the linear system. We can define $(h^n,u^n)_{n\in\mathbb{N}_0}$ by the induction. Next, we are going to prove by the induction that there exist positive constants η, K, and T such that the following bounds hold for all $n\in\mathbb{N}_0$:

$$1+h^n\geq\frac{1}{2}, \tag{3.39}$$

$$\|u^n\|_{L^1_T(B^2)\cap L^2_T(B^1)}+\|h^n\|_{\widetilde{E}^{0,1}_T}\leq\eta, \tag{3.40}$$

$$\|u^n\|_{L^\infty_T(B^0)}+\|h^n\|_{L^\infty_T(\widetilde{B}^{0,1})\cap L^1_T(\widetilde{B}^{2,1})}\leq KE_0. \tag{3.41}$$

Assume that (3.39)–(3.41) hold for (h^n,u^n). We need to prove that (3.39)–(3.41) also hold for (h^{n+1},u^{n+1}). Applying the a priori estimates (3.35) and (3.37) to (h^{n+1},u^{n+1}), we obtain

$$\begin{aligned}
&\|u^{n+1}\|_{L^1_T(B^2)}+\|u^{n+1}\|_{L^2_T(B^1)}+\|h^{n+1}\|_{\widetilde{E}^{0,1}_T}\\
&\quad\leq C\mathcal{Q}_0(T)+C\sum_{k\in\mathbb{Z}}\omega_k(T)\|\mathcal{G}^n_k(t)\|_{L^1_T(L^2)}+C\sum_{k\geq1}\omega_k(T)\|\nabla\mathcal{H}^n_k(t)\|_{L^1_T(L^2)}\\
&\qquad+C\sum_{k<1}\omega_k(T)\|\mathcal{H}^n_k(t)\|_{L^1_T(L^2)}+C\|u^{n+1}\|_{L^2_T(B^1)}\|u^n\|_{L^2_T(B^1)}\\
&\qquad+C\|h^{n+1}\|_{\widetilde{E}^{0,1}_T}\|u^n\|_{L^1_T(B^2)}
\end{aligned} \tag{3.42}$$

and

$$\begin{aligned}
&\|u^{n+1}\|_{L^\infty_T(B^0)}+\|h^{n+1}\|_{L^\infty_T(\widetilde{B}^{0,1})}+\|h^{n+1}\|_{L^1_T(\widetilde{B}^{2,1})}\\
&\qquad\leq Ce^{C\|u^n\|_{L^1_T(B^2)}}\Big(E_0+\|\mathcal{H}^n\|_{L^1_T(\widetilde{B}^{0,1})}+\|\mathcal{G}^n\|_{L^1_T(B^0)}\Big),
\end{aligned} \tag{3.43}$$

with

$$\mathcal{Q}_0(T)\triangleq\sum_{k\in\mathbb{Z}}\omega_k(T)E_k(0).$$

Thanks to (2.4), we have

$$\|\mathcal{H}^n\|_{B^0}\leq C\|h^n\|_{B^0}\|u^n\|_{B^2}\quad\text{and}\quad\|\mathcal{H}^n\|_{B^1}\leq C\|h^n\|_{B^1}\|u^n\|_{B^2},$$

which, together with the fact that $\widetilde{B}^{0,1}=B^0\cap B^1$, yields

$$\|\mathcal{H}^n\|_{L^1_T(\widetilde{B}^{0,1})}\leq C\|h^n\|_{L^\infty_T(\widetilde{B}^{0,1})}\|u^n\|_{L^1_T(B^2)}\leq CKE_0\eta. \tag{3.44}$$

We rewrite $\mathcal{G}^n$ as

$$\frac{\nabla h^n}{1+h^n}\widetilde{\nabla}u^n=(1+h^n)\nabla\bigg(\frac{h^n}{1+h^n}\bigg)\widetilde{\nabla}u^n.$$

Using (2.4) and (2.12), we get

$$\begin{aligned}
\|\mathcal{G}^n\|_{L^1_T(B^0)}&\leq C\bigg\|\nabla\bigg(\frac{h^n}{1+h^n}\bigg)\bigg\|_{L^\infty_T(B^0)}\|(1+h^n)\widetilde{\nabla}u^n\|_{L^1_T(B^1)}\\
&\leq C(1+\|h^n\|_{L^\infty_T(L^\infty)})^2\|h^n\|_{L^\infty_T(B^1)}(1+\|h^n\|_{L^\infty_T(B^1)})\|u^n\|_{L^1_T(B^2)}\\
&\leq C(1+\|h^n\|_{L^\infty_T(\widetilde{B}^{0,1})})^3\|h^n\|_{L^\infty_T(B^1)}\|u^n\|_{L^1_T(B^2)}\\
&\leq CKE_0(1+KE_0)^3\eta.
\end{aligned} \tag{3.45}$$

Plugging (3.44) and (3.45) into (3.43) yields

$$\begin{aligned}&\|u^{n+1}\|_{L^\infty_T(B^0)}+\|h^{n+1}\|_{L^\infty_T(\widetilde{B}^{0,1})}+\|h^{n+1}\|_{L^1_T(\widetilde{B}^{2,1})}\\&\leq Ce^{C\eta}\Big(E_0+KE_0(1+KE_0)^3\eta\Big).\end{aligned}\tag{3.46}$$

We take $\eta>0$ small enough and $K=4C$ such that

$$e^{C\eta}\leq 2,\quad K(1+KE_0)^3\eta\leq 1,\tag{$\Re_1$}$$

from which, together with (3.46), it follows that

$$\|u^{n+1}\|_{L^\infty_T(B^0)}+\|h^{n+1}\|_{L^\infty_T(\widetilde{B}^{0,1})}+\|h^{n+1}\|_{L^1_T(\widetilde{B}^{2,1})}\leq KE_0.$$

This proves (3.41) for (u^{n+1},h^{n+1}).

Next, we prove (3.40) for (u^{n+1},h^{n+1}). Applying Lemma 5.2 with $s_1=0$ and $s_2=1$, (2.4) with $s_1=s_2=1$, and Lemma 5.4 with $s=1$, we obtain

$$\begin{aligned}\sum_{k\in\mathbb{Z}}\omega_k(T)\|\mathcal{G}^n_k(t)\|_{L^1_T(L^2)}&\leq C\Big\|\nabla\Big(\frac{h^n}{1+h^n}\Big)\Big\|_{E^0_T}\|(1+h^n)\widetilde{\nabla}u^n\|_{L^1_T(B^1)}\\&\leq C(1+\|h^n\|_{L^\infty_T(L^\infty)})^3\|h^n\|_{E^1_T}(1+\|h^n\|_{L^\infty_T(B^1)})\|u^n\|_{L^1_T(B^2)}\\&\leq C(1+\|h^n\|_{L^\infty_T(\widetilde{B}^{0,1})})^4\|h^n\|_{\widetilde{E}^{0,1}_T}\|u^n\|_{L^1_T(B^2)}\\&\leq C(1+KE_0)^4\eta^2.\end{aligned}\tag{3.47}$$

On the other hand, we apply Lemma 5.2 with $s_1=0,s_2=1$ to get

$$\begin{aligned}&\sum_{k\geq1}\omega_k(T)\|\nabla\mathcal{H}^n_k(t)\|_{L^1_T(L^2)}+\sum_{k<1}\omega_k(T)\|\mathcal{H}^n_k(t)\|_{L^1_T(L^2)}\\&\leq C\sum_{k\in\mathbb{Z}}\omega_k(T)(\|\nabla h^n_k\|_{L^\infty_T(L^2)}+\|h^n_k\|_{L^\infty_T(L^2)})\|\mathrm{div}u^n\|_{L^1_T(B^1)}\\&\quad+C\sum_{k\in\mathbb{Z}}\omega_k(T)2^{2k}\|u^n_k\|_{L^1_T(L^2)}\|h^n\|_{L^\infty_T(\widetilde{B}^{0,1})}\\&\triangleq I+II.\end{aligned}$$

Obviously, we have

$$I\leq C\|h^n\|_{\widetilde{E}^{0,1}_T}\|u^n\|_{L^1_T(B^2)}\leq C\eta^2.\tag{3.48}$$

In order to estimate II, we first fix $k_0\geq1$ such that

$$\sum_{k\geq k_0}\|u_k(0)\|_2\leq\frac{\eta}{16CKE_0}.\tag{3.49}$$

Then we write

$$\begin{aligned}II&=C\sum_{k\geq k_0}\omega_k(T)2^{2k}\|u^n_k\|_{L^1_T(L^2)}\|h^n\|_{L^\infty_T(\widetilde{B}^{0,1})}+C\sum_{k\leq k_0}\omega_k(T)2^{2k}\|u^n_k\|_{L^1_T(L^2)}\|h^n\|_{L^\infty_T(\widetilde{B}^{0,1})}\\&\triangleq II_1+II_2.\end{aligned}$$

Using (3.18), (3.47), and (3.49), we obtain

$$\begin{aligned} II_1 &\leq CKE_0\Big[\sum_{k\geq k_0}\omega_k(T)\|u_k(0)\|_2+\sum_{k\geq k_0}\omega_k(T)\Big(\|\nabla h_k^n\|_{L^1_T(L^2)}+\|\mathcal{G}_k^{n-1}\|_{L^1_T(L^2)}\Big)\\ &\quad+\|u^n\|_{L^2_T(B^1)}\|u^{n-1}\|_{L^2_T(B^1)}\Big]\\ &\leq CKE_0\Big[\frac{\eta}{16CKE_0}+\sum_{k\geq k_0}\omega_k(T)\|\nabla h_k^n\|_{L^1_T(L^2)}+(1+KE_0)^4\eta^2\Big]. \end{aligned} \tag{3.50}$$

On the other hand, thanks to (3.19) and Lemma 5.1, we have

$$\begin{aligned} \sum_{k\geq k_0}\omega_k(T)\|\nabla h_k^n\|_{L^1_T(L^2)} &\leq C(1-e^{-cT})\sum_{k\geq k_0}\omega_k(T)E_{hk}(0)\\ &\quad+C(1-e^{-cT})\sum_{k\geq k_0}\omega_k(T)\\ &\quad\Big(\|\nabla\mathcal{H}_k^{n-1}(t)\|_{L^1_T(L^2)}+\|\mathcal{G}_k^{n-1}(t)\|_{L^1_T(L^2)}\Big)\\ &\quad+C\|u^n\|_{L^2_T(B^1)}\|u^{n-1}\|_{L^2_T(B^1)}+C\|h^n\|_{E^1_T}\|u^{n-1}\|_{L^2_T(B^1)}\\ &\leq C(1-e^{-cT})E_0+C(1-e^{-cT})KE_0\eta+C(1+KE_0)^4\eta^2, \end{aligned}$$

where we used (3.44) and (3.47) in the second inequality. Plugging the above inequality into (3.50) yields

$$II_1\leq CKE_0\Big[\frac{\eta}{16CKE_0}+(1-e^{-cT})(E_0+KE_0\eta)+(1+KE_0)^4\eta^2\Big]. \tag{3.51}$$

Note that for $k\leq k_0$, we can choose $T>0$ small enough such that

$$\omega_k(T)\leq\frac{1}{16CKE_0}, \tag{$\Re_2$}$$

so we get

$$|II_2|\leq\frac{\eta}{16}. \tag{3.52}$$

Plugging (3.47), (3.48), (3.51), and (3.52) into (3.42), we get

$$\begin{aligned} &\|u^{n+1}\|_{L^1_T(B^2)}+\|u^{n+1}\|_{L^2_T(B^1)}+\|h^{n+1}\|_{\widetilde{E}^{0,1}_T}\\ &\quad\leq C\mathcal{Q}_0(T)+\frac{\eta}{8}+C(1+KE_0)^5\eta^2+CKE_0(1-e^{-cT})(E_0+KE_0\eta)\\ &\quad\quad+C\eta(\|u^{n+1}\|_{L^2_T(B^1)}+\|h^{n+1}\|_{\widetilde{E}^{0,1}_T}). \end{aligned} \tag{3.53}$$

Note that $\mathcal{Q}_0(0)=0$. We can take T,η small enough such that

$$C\eta\leq\frac{1}{2},\quad C\mathcal{Q}_0(T)\leq\frac{\eta}{8},\quad C(1+KE_0)^5\eta<\frac{1}{8},\quad\text{and}$$

$$CKE_0(1-e^{-cT})(E_0+KE_0\eta)\leq\frac{\eta}{8}, \tag{$\Re_3$}$$

which, together with (3.53), gives

$$\|u^{n+1}\|_{L^1_T(B^2)} + \|u^{n+1}\|_{L^2_T(B^1)} + \|h^{n+1}\|_{\widetilde{E}^{0,1}_T} \leq \eta.$$

Finally, let us prove (3.39) for h^{n+1}. We rewrite the first equation of (3.38) as

$$\partial_t(1+h^{n+1}) + u^n \cdot \nabla(1+h^{n+1}) + \mathrm{div} u^{n+1} - \mathcal{H}^n = 0.$$

Then $1+h^{n+1}$ can be represented as

$$\begin{aligned}(1+h^{n+1})(t,x) = (1+h_0^{n+1})((\psi^n)_t^{-1}(x)) + \int_0^t \mathrm{div}\, u^{n+1}(\tau, \psi^n_\tau((\psi^n)_t^{-1}(x)))d\tau \\ + \int_0^t \mathcal{H}^n(\tau, \psi^n_\tau((\psi^n)_t^{-1}(x)))d\tau, \end{aligned} \tag{3.54}$$

where the flow map ψ^n_t is defined by

$$\begin{cases} \partial_t \psi^n_t(x) = u^n(t, \psi^n_t(x)), \\ \psi^n_t|_{t=0} = x. \end{cases}$$

Thanks to the inclusion map $B^1 \hookrightarrow L^\infty$ and (2.4), we get

$$\begin{aligned} &\int_0^t \|\mathrm{div}\, u^{n+1}(\tau, \psi^n_\tau((\psi^n)_t^{-1}(x)))\|_\infty d\tau \leq \|u^{n+1}\|_{L^1_t(B^2)} \leq \eta, \\ &\int_0^t \|\mathcal{H}^n(\tau, \psi_\tau(\psi_t^{-1}(x)))\|_\infty d\tau \leq \|h^n \mathrm{div} u^n\|_{L^1_t(B^1)} \\ &\qquad\qquad \leq C\|h^n\|_{L^\infty_t(\widetilde{B}^{0,1})}\|u^n\|_{L^1_t(B^2)} \leq CKE_0\eta, \end{aligned}$$

from which, together with (3.54), it follows that

$$1 + h^{n+1} \geq \frac{3}{4} - (1+CKE_0)\eta. \tag{3.55}$$

We take η small enough such that

$$(1+CKE_0)\eta \leq \frac{1}{4}, \tag{$\Re_4$}$$

which, together with (3.55), ensures that

$$1 + h^{n+1} \geq \frac{1}{2}.$$

So far, we have shown that T, η can be chosen small enough such that the assumptions ($\Re_1$)–($\Re_4$) hold, under which the approximate solutions $(u^n, h^n)_{n\in\mathbb{N}_0}$ are uniformly bounded in

$$\mathcal{E}_T \triangleq \Big(L^\infty_T(B^0) \cap L^1_T(B^2)\Big) \times \Big(L^\infty_T(\widetilde{B}^{0,1}) \cap L^1_T(\widetilde{B}^{2,1})\Big).$$

It should be pointed out that if $\|u_0\|_{B^0} + \|h_0\|_{\widetilde{B}^{0,1}}$ is small enough, we can take $T = +\infty$ such that the assumptions ($\Re_1$)–($\Re_4$) hold.

3.3. The existence of the solution. Now let us turn to prove the existence of the solution, and the standard compact arguments will be used. We should point out that in order to obtain the convergence of the terms $u^n\cdot\nabla u^{n+1}$ and $u^n\cdot\nabla h^{n+1}$ in (3.38), we need to show that both (u^n,h^n) and (u^{n+1},h^{n+1}) have the same limit. Indeed, following the argument in the proof of uniqueness (section 4), it can be proved that $(u^n,h^n)_n$ is a Cauchy sequence in some weak topology which particularly ensures the uniqueness of the limit for any subsequence of $(u^n,h^n)_n$. In section 3.2, we have showed that the approximate solutions $(h^n,u^n)_{n\in\mathbb{N}}$ satisfy (3.39)–(3.41), and without loss of generality, we can assume the following:

$$1+h^n\geq\frac{1}{2}, \tag{3.56}$$

$$\|u^n\|_{L^\infty_T(B^0)\cap L^1_T(B^2)}+\|h^n\|_{L^\infty_T(\widetilde{B}^{0,1})\cap L^1_T(\widetilde{B}^{2,1})}\leq KE_0. \tag{3.57}$$

Using the interpolation and the fact that $B^0\cap B^1=\widetilde{B}^{0,1}$, we have

$$\|h^n\|_{L^2_T(B^1)}\lesssim\|h^n\|^{\frac12}_{L^\infty_T(\widetilde{B}^{0,1})}\|h^n\|^{\frac12}_{L^1_T(\widetilde{B}^{2,1})},\quad \|u^n\|_{L^2_T(B^1)}\lesssim\|u^n\|^{\frac12}_{L^\infty_T(B^0)}\|u^n\|^{\frac12}_{L^1_T(B^2)},$$

$$\|h^n\|_{L^4_T(B^{\frac12})}\lesssim\|h^n\|^{\frac12}_{L^\infty_T(\widetilde{B}^{0,1})}\|h^n\|^{\frac12}_{L^2_T(B^1)},\quad \|u^n\|_{L^{\frac43}_T(B^{\frac32})}\lesssim\|u^n\|^{\frac14}_{L^\infty_T(B^0)}\|u^n\|^{\frac34}_{L^1_T(B^2)},$$

from which, together with (3.57), it follows that

$$\|h^n\|_{L^2_T(B^1)}+\|u^n\|_{L^2_T(B^1)}+\|h^n\|_{L^4_T(B^{\frac12})}+\|u^n\|_{L^{\frac43}_T(B^{\frac32})}\lesssim KE_0. \tag{3.58}$$

Now, we show that (h^n,u^n) is uniformly bounded in $C^{\frac12}_{loc}(B^0)\times C^{\frac14}_{loc}(B^{-\frac12})$. Using (2.4), (3.57), and (3.58), it is easy to verify that

$$\|u^n\cdot\nabla h^{n+1}\|_{L^2_T(B^0)}\lesssim\|u^n\|_{L^2_T(B^1)}\|h^{n+1}\|_{L^\infty_T(\widetilde{B}^{0,1})}\lesssim(KE_0)^2,$$

$$\|h^n\mathrm{div}u^n\|_{L^2_T(B^0)}\lesssim\|u^n\|_{L^2_T(B^1)}\|h^n\|_{L^\infty_T(\widetilde{B}^{0,1})}\lesssim(KE_0)^2,$$

from which, together with the first equation of (3.38), it follows that $\partial_t h^n$ is uniformly bounded in $L^2_T(B^0)$, which implies that h^n is uniformly bounded in $C^{\frac12}_{loc}(B^0)$. On the other hand, thanks to (2.4), (3.58), and (2.12), we have

$$\|u^n\cdot\nabla u^{n+1}\|_{L^{\frac43}_T(B^{-\frac12})}\lesssim\|u^n\|_{L^\infty_T(B^0)}\|u^{n+1}\|_{L^{\frac43}_T(B^{\frac32})}\lesssim(KE_0)^2,$$

$$\left\|\frac{\nabla h^n}{1+h^n}\widetilde{\nabla}u^n\right\|_{L^{\frac43}_T(B^{-\frac12})}\lesssim C(1+\|h^n\|_{L^\infty_T(B^1)})^3\|u^n\|_{L^{\frac43}_T(B^{\frac32})}\lesssim C(1+KE_0)^3KE_0,$$

from which, together with the second equation of (3.38), it follows that $\partial_t u^n$ is uniformly bounded in $L^{\frac43}_T(B^{-\frac12})$, which implies that u^n is uniformly bounded in $C^{\frac14}_{loc}(B^{-\frac12})$.

Next, we claim that the inclusions $B^0\cap B^1\hookrightarrow L^2$ and $B^{-\frac12}\cap B^0\hookrightarrow\dot{H}^{-\frac12}$ are locally compact. Indeed, these can be proved by noting that for $s'<s$, $\dot{H}^{s'}\cap\dot{H}^s\hookrightarrow\dot{H}^{s'}$ is locally compact and for $s\in\mathbb{R}$, $B^s\hookrightarrow\dot{H}^s$. Then, by the Arzelà–Ascoli theorem and

Cantor's diagonal process, there exist a subsequence (u^{n_k}, h^{n_k}) and a function (u, h) such that

$$(u^{n_k}, h^{n_k}) \to (u, h) \qquad \text{in} \quad C_{loc}(\dot{H}^{-\frac{1}{2}}_{loc}) \times C_{loc}(L^2_{loc}), \tag{3.59}$$

as $n_k \to \infty$. On the other hand, (u^{n_k}, h^{n_k}) is uniformly bounded in $\mathcal{E}_T$. Then there exists a subsequence (which is still denoted by (u^{n_k}, h^{n_k})) such that

$$(u^{n_k}, h^{n_k}) \rightharpoonup (u, h) \quad \text{in} \quad \mathcal{E}_T,$$

where "$\rightharpoonup$" denotes weak* convergence.

Finally, let us prove that (u, h) solves (1.1) in the sense of distribution. We need only prove that the nonlinear terms such as $u^n \cdot \nabla h^n$, $\frac{\nabla h^n}{1+h^n}\widetilde{\nabla} u^n$, etc., tend to the corresponding nonlinear terms in the sense of distribution. This can be done by using the uniform estimates of (u^n, h^n), (u, h) in $\mathcal{E}_T$ and the convergence result (3.59). Here, we show only the case of the term $Y(h^n)\widetilde{\nabla} u^n$ (where $Y(z) \triangleq \nabla z/(1+z)$); the other terms can be treated in the same way. For any test function $\theta \in C_0^\infty([0, T^*) \times \mathbb{R}^2)$, we write

$$\begin{aligned}
&\big\langle Y(h^n)\widetilde{\nabla} u^n - Y(h)\widetilde{\nabla} u,\, \theta\big\rangle \\
&\quad = \Big\langle (1+h^n)\nabla\Big(\frac{h^n}{1+h^n} - \frac{h}{1+h}\Big)\widetilde{\nabla} u^n,\, \theta\Big\rangle \\
&\qquad + \Big\langle (h^n - h)\nabla\Big(\frac{h}{1+h}\Big)\widetilde{\nabla} u^n,\, \theta\Big\rangle + \Big\langle (1+h)\nabla\Big(\frac{h}{1+h}\Big)\widetilde{\nabla}(u^n - u),\, \theta\Big\rangle \\
&\quad \triangleq I_1 + I_2 + I_3.
\end{aligned}$$

Thanks to (2.4) and (3.56), we have

$$\begin{aligned}
I_1 &\le \Big\| \frac{\psi(h^n - h)}{(1+h^n)(1+h)} \Big\|_2 \|\nabla((1+h^n)\widetilde{\nabla} u^n \theta)\|_2 \lesssim \|\psi(h^n - h)\|_2 \|(1+h^n)\widetilde{\nabla} u^n\|_{B^1} \\
&\lesssim \|\psi(h^n - h)\|_2 (1 + \|h^n\|_{\widetilde{B}^{0,1}}) \|u^n\|_{B^2},
\end{aligned}$$

where $\psi \in C_0^\infty([0, T^*) \times \mathbb{R}^2)$, and $\psi = 1$ on $\operatorname{supp}\theta$. For I_2, we have

$$\begin{aligned}
I_2 &\le \|\theta(h^n - h)\|_2 \Big\| \nabla\Big(\frac{h}{1+h}\Big)\widetilde{\nabla} u^n \Big\|_2 \lesssim \|\theta(h^n - h)\|_2 \|\nabla h\|_2 \|\nabla u^n\|_{L^\infty} \\
&\lesssim \|\theta(h^n - h)\|_2 \|h\|_{\widetilde{B}^{0,1}} \|u^n\|_{B^2}.
\end{aligned}$$

Using (3.56) and the interpolation, we get

$$\begin{aligned}
I_3 &\le \Big\| (1+h)\nabla\Big(\frac{h}{1+h}\Big) \Big\|_2 \|\widetilde{\nabla}(u^n - u)\theta\|_2 \lesssim (1 + \|h\|_\infty) \|\nabla h\|_2 \|(u^n - u)\theta\|_{\dot{H}^1} \\
&\lesssim (1 + \|h\|_{\widetilde{B}^{0,1}}) \|h\|_{B^1} \|u^n - u\|^{\frac{3}{5}}_{H^2} \|(u^n - u)\theta\|^{\frac{2}{5}}_{\dot{H}^{-\frac{1}{2}}}.
\end{aligned}$$

Thus, by (3.59), we get as $n \to 0$

$$\big\langle Y(h^n)\widetilde{\nabla} u^n - Y(h)\widetilde{\nabla} u,\, \theta\big\rangle \longrightarrow 0.$$

Following the argument in [11], we can also prove that (u, h) is continuous in time with values in $B^0 \times \widetilde{B}^{0,1}$.

4. Uniqueness. In this section, we will prove the uniqueness of the solution. First, let us recall some known results.

LEMMA 4.1 (Osgood's lemma). *Let ρ be a measurable positive function and γ a positive locally integrable function, each defined on the domain $[t_0, t_1]$. Let $\mu : [0,\infty) \to [0,\infty)$ be a continuous nondecreasing function, with $\mu(0) = 0$. Let $a \geq 0$, and assume that for all t in $[t_0, t_1]$,*

$$\rho(t) \leq a + \int_{t_0}^{t} \gamma(\tau)\mu(\rho(\tau))d\tau.$$

If $a > 0$, then

$$-\mathcal{M}(\rho(t)) + \mathcal{M}(a) \leq \int_{t_0}^{t} \gamma(\tau)d\tau, \quad \text{where} \quad \mathcal{M}(x) = \int_{x}^{1} \frac{d\tau}{\mu(\tau)}.$$

If $a = 0$ and $\mathcal{M} = \infty$, then $\rho \equiv 0$.

This lemma can be understood as a generalization of the classical Gronwall lemma and can be found in [8].

PROPOSITION 4.2. *Let $s \in (-\frac{d}{p}, 1 + \frac{d}{p})$ and $1 \leq p, r \leq +\infty$. Let v be a vector field such that $\nabla v \in L^1_T(\dot{B}^{\frac{d}{p}}_{p,r} \cap L^\infty)$. Assume that $f_0 \in \dot{B}^s_{p,r}$, $g \in L^1_T(\dot{B}^s_{p,r})$ and $f \in L^\infty_T(\dot{B}^s_{p,r}) \cap C([0,T]; \mathcal{S}')$ is the solution of*

$$\begin{cases} \partial_t f + v \cdot \nabla f = g, \\ f(0,x) = f_0. \end{cases}$$

Then there exists a constant $C(s,p,d)$ such that for $t \in [0,T]$

$$\|f\|_{\widetilde{L}^\infty_t(\dot{B}^s_{p,r})} \leq Ce^{CV(t)}\Big(\|f_0\|_{\dot{B}^s_{p,r}} + \int_0^t e^{-CV(\tau)}\|g(\tau)\|_{\dot{B}^s_{p,r}}d\tau\Big), \tag{4.1}$$

where $V(t) \triangleq \int_0^t \|\nabla v(\tau)\|_{\dot{B}^{\frac{d}{p}}_{p,r}\cap L^\infty} d\tau$. If $r < +\infty$, then f belongs to $C([0,T]; \dot{B}^s_{p,r})$.

The proof can be found in [15].

PROPOSITION 4.3. *Let $T > 0$, $s \in \mathbb{R}$, and $1 \leq q, r \leq +\infty$. Assume that $u_0 \in \dot{B}^s_{2,q}$, $g \in \widetilde{L}^1_T(\dot{B}^s_{2,q})$, and u is the solution of*

$$\begin{cases} \partial_t u - \nu\widetilde{\Delta}u = g, \\ u(0,x) = u_0, \end{cases}$$

where $\widetilde{\Delta}u = \nabla \cdot D(u) + \nabla \mathrm{div}\, u$. Then there exists a constant $C(s,d,\nu)$ such that

$$\begin{aligned}(r\nu)^{\frac{1}{r}}\|u\|_{\widetilde{L}^r_T(\dot{B}^{s+\frac{2}{r}}_{2,q})} \leq \Big(\sum_{k\in\mathbb{Z}}\big(1 - e^{-r\nu 2^{2k}T}\big)^{\frac{q}{r}} 2^{qks}\|\Delta_k u_0\|_2^q\Big)^{\frac{1}{q}} \\ + C\Big(\sum_{k\in\mathbb{Z}}\big(1 - e^{-r\nu 2^{2k}T}\big)^{\frac{q}{r}} 2^{qks}\|\Delta_k g\|^q_{L^1_T(L^2)}\Big)^{\frac{1}{q}}.\end{aligned} \tag{4.2}$$

If $q < +\infty$, then u belongs to $C([0,T]; \dot{B}^s_{2,q})$.

The proof is similar to the case when the diffusion term $\widetilde{\Delta}u$ is replaced by Δu. We refer the reader to [9] for details.

Now we introduce the logarithmic interpolation inequality (see [14]).

PROPOSITION 4.4. *For any* $1 \leq p, \rho \leq +\infty$, $s \in \mathbb{R}$, *and* $0 < \epsilon \leq 1$, *we have*

$$\|f\|_{\widetilde{L}^\rho_T(\dot{B}^s_{p,1})} \leq C\frac{\|f\|_{\widetilde{L}^\rho_T(\dot{B}^s_{p,\infty})}}{\epsilon}\log\left(e + \frac{\|f\|_{\widetilde{L}^\rho_T(\dot{B}^{s-\epsilon}_{p,\infty})} + \|f\|_{\widetilde{L}^\rho_T(\dot{B}^{s+\epsilon}_{p,\infty})}}{\|f\|_{\widetilde{L}^\rho_T(\dot{B}^s_{p,\infty})}}\right). \tag{4.3}$$

Now, let us prove the uniqueness of the solution of (3.1). Let (u_1, h_1), (u_2, h_2) $\in \big(L^\infty_T(B^0)\cap L^1_T(B^2)\big)\times L^\infty_T(\widetilde{B}^{0,1})$ be two solutions of (3.1) with the same initial data. The difference $\vartheta \triangleq h_2 - h_1$, $w \triangleq u_2 - u_1$ satisfies the following system:

$$\begin{cases} \partial_t\vartheta + u_2\cdot\nabla\vartheta = -\mathrm{div}w - w\nabla h_1 - \vartheta\mathrm{div}u_2 - h_1\mathrm{div}w, \\ \partial_t w - \nu\widetilde{\Delta}w = -\nabla\vartheta - u_2\cdot\nabla w - w\cdot\nabla u_1 + \nu(1+h_1)\nabla\Big(\dfrac{h_1}{1+h_1}\Big)\widetilde{\nabla}w \\ \qquad\qquad + \nu(1+h_1)\nabla\Big(\dfrac{h_2}{1+h_2} - \dfrac{h_1}{1+h_1}\Big)\widetilde{\nabla}u_2 + \nu\vartheta\nabla\Big(\dfrac{h_2}{1+h_2}\Big)\widetilde{\nabla}u_2, \\ \vartheta(0,x) = 0, \quad w(0,x) = 0. \end{cases} \tag{4.4}$$

Without loss of generality, we assume that there holds for sufficiently small T

$$1 + h_1 \geq \frac{1}{2}, \tag{4.5}$$

$$\|h_1\|_{\widetilde{E}^{0,1}_T} \leq \varepsilon, \tag{4.6}$$

where $\varepsilon > 0$ is small enough. Applying Proposition 4.2 to the first equation of (4.4) yields

$$\|\vartheta(t)\|_{\dot{B}^0_{2,\infty}} \lesssim \int_0^t e^{C(V_2(t)-V_2(\tau))}\|w\cdot\nabla h_1 + \vartheta\mathrm{div}u_2 + h_1\mathrm{div}w + \mathrm{div}w\|_{\dot{B}^0_{2,\infty}}d\tau, \tag{4.7}$$

with $V_2(t) \triangleq \int_0^t \|\nabla u_2\|_{\dot{B}^1_{2,\infty}\cap L^\infty}d\tau$. It follows from (2.5) with $s = 0$ that

$$\begin{aligned} \|w\cdot\nabla h_1\|_{\dot{B}^0_{2,\infty}} &\lesssim \|\nabla h_1\|_{\dot{B}^0_{2,\infty}}\|w\|_{B^1} \lesssim \|w\|_{B^1}\|h_1\|_{B^1}, \\ \|\vartheta\mathrm{div}u_2\|_{\dot{B}^0_{2,\infty}} &\lesssim \|\vartheta\|_{\dot{B}^0_{2,\infty}}\|u_2\|_{B^2}, \\ \|h_1\mathrm{div}w\|_{\dot{B}^0_{2,\infty}} &\lesssim \|\mathrm{div}w\|_{\dot{B}^0_{2,\infty}}\|h_1\|_{B^1} \lesssim \|w\|_{B^1}\|h_1\|_{B^1}, \end{aligned}$$

where we have used $B^1 \hookrightarrow \dot{B}^1_{2,\infty}$. Plugging the above estimates into (4.7), we get

$$\|\vartheta(t)\|_{\dot{B}^0_{2,\infty}} \lesssim \int_0^t e^{C(V_2(t)-V_2(\tau))}\Big[\|w\|_{B^1}(1+\|h_1\|_{B^1}) + \|\vartheta\|_{\dot{B}^0_{2,\infty}}\|u_2\|_{B^2}\Big]d\tau. \tag{4.8}$$

Recalling that $u^i \in L^1_T(B^2)$, we can take a $T \in (0,\infty)$ small enough so that

$$C\|u_2\|_{L^1_T(B^2)} \leq \frac{1}{4},$$

which together with (4.8) implies that for $t \leq T$

$$\|\vartheta\|_{L^\infty_t(\dot{B}^0_{2,\infty})} \lesssim \|w\|_{L^1_t(B^1)}(1 + \|h_1\|_{L^\infty_t(B^1)}). \tag{4.9}$$

Applying (4.3) to the term $\|w\|_{L^1_t(B^1)}$ yields

(4.10)
$$\|\vartheta\|_{L^\infty_t(\dot B^0_{2,\infty})} \lesssim \|w\|_{\widetilde L^1_t(\dot B^1_{2,\infty})} \log\left(e + \frac{\|w\|_{\widetilde L^1_t(\dot B^0_{2,\infty})} + \|w\|_{\widetilde L^1_t(\dot B^2_{2,\infty})}}{\|w\|_{\widetilde L^1_t(\dot B^1_{2,\infty})}}\right)(1 + \|h_1\|_{L^\infty_t(B^1)}).$$

Thanks to $B^s \hookrightarrow \dot B^s_{2,\infty}$ and $\widetilde B^{0,1} \hookrightarrow B^1$, we have

$$\|\vartheta\|_{L^\infty_t(\dot B^0_{2,\infty})} \lesssim \|w\|_{\widetilde L^1_t(\dot B^1_{2,\infty})} \log\left(e + \frac{W(t)}{\|w\|_{\widetilde L^1_t(\dot B^1_{2,\infty})}}\right) \tag{4.11}$$

with

$$W(t) \triangleq \sum_{i=1}^{2} \|u^i\|_{\widetilde L^1_t(B^0)} + \|u^i\|_{\widetilde L^1_t(B^2)},$$

and for finite t, $W(t) < +\infty$.

Next, we deal with the second equation of (4.4). We get by applying (2.7) with $s = 1$, $s = 0$, respectively, that

$$\|u_2 \cdot \nabla w\|_{\widetilde L^1_t(\dot B^{-1}_{2,\infty})} \lesssim \|u_2\|_{L^2_t(B^1)} \|w\|_{\widetilde L^2_t(\dot B^0_{2,\infty})}, \tag{4.12}$$

$$\|w \cdot \nabla u_1\|_{\widetilde L^1_t(\dot B^{-1}_{2,\infty})} \lesssim \|u_1\|_{L^2_t(B^1)} \|w\|_{\widetilde L^2_t(\dot B^0_{2,\infty})}. \tag{4.13}$$

We can deduce $h_i \in C(0,T;\mathbb{R}^2)$ $(i = 1, 2)$ from the fact $B^1 \hookrightarrow C$. Moreover, due to (4.5), we can assume $h_1(t,x) + 1 \geq \frac{1}{2}$ for all $t \leq T$, $x \in \mathbb{R}^2$. Since h_1, h_2 have the same initial data, from the continuity of h_2, there exists a $\widetilde T \leq T$ such that

$$h_2(x,t) + 1 \geq \frac{1}{4} \quad \text{for all} \quad t \in [0, \widetilde T], \quad x \in \mathbb{R}^2.$$

It follows from (2.6) with $s = 1$, (2.15), and $B^1 \hookrightarrow \dot B^1_{2,\infty} \cap L^\infty$ that

$$\begin{aligned}
&\Big\|(1+h_1)\nabla\Big(\frac{h_2}{1+h_2} - \frac{h_1}{1+h_1}\Big)\widetilde\nabla u_2\Big\|_{\dot B^{-1}_{2,\infty}} \\
&\quad\lesssim \Big\|(1+h_1)\nabla\Big(\frac{h_2}{1+h_2} - \frac{h_1}{1+h_1}\Big)\Big\|_{\dot B^{-1}_{2,\infty}} \|\widetilde\nabla u_2\|_{B^1} \\
&\quad\lesssim (1 + \|h_1\|_{B^1})\Big\|\frac{h_2}{1+h_2} - \frac{h_1}{1+h_1}\Big\|_{\dot B^0_{2,\infty}} \|u_2\|_{B^2} \\
&\quad\lesssim (1 + \|h_1\|_{B^1})(\|h_1\|_{B^1} + \|h_2\|_{B^1})\|\vartheta\|_{\dot B^0_{2,\infty}} \|u_2\|_{B^2},
\end{aligned}$$

which, together with $L^1_t(\dot B^{-1}_{2,\infty}) \subset \widetilde L^1_t(\dot B^{-1}_{2,\infty})$, yields

$$\begin{aligned}
&\Big\|(1+h_1)\nabla\Big(\frac{h_2}{1+h_2} - \frac{h_1}{1+h_1}\Big)\widetilde\nabla u_2\Big\|_{\widetilde L^1_t(\dot B^{-1}_{2,\infty})} \\
&\quad\lesssim \int_0^t (1 + \|h_1\|_{B^1})(\|h_1\|_{B^1} + \|h_2\|_{B^1})\|\vartheta\|_{\dot B^0_{2,\infty}} \|u_2\|_{B^2} d\tau.
\end{aligned} \tag{4.14}$$

Thanks to (2.6), (2.12), and $L^1_t(\dot B^{-1}_{2,\infty}) \subset \widetilde L^1_t(\dot B^{-1}_{2,\infty})$, we get

$$\Big\|\vartheta\nabla\Big(\frac{h_2}{1+h_2}\Big)\widetilde\nabla u_2\Big\|_{\widetilde L^1_t(\dot B^{-1}_{2,\infty})} \lesssim \int_0^t \|\vartheta\|_{\dot B^0_{2,\infty}}\|h_2\|_{B^1}\|u_2\|_{B^2}d\tau. \tag{4.15}$$

Thanks to Lemma 5.3 with $s_1 = s_2 = 0$, Lemma 5.4 with $s = 1$, and (2.5) with $s = 0$, we have

$$\begin{aligned}
&\sup_{k\in\mathbb{Z}}\omega_k(t)2^{-k}\Big\|\Delta_k\Big((1+h_1)\nabla\Big(\frac{h_1}{1+h_1}\Big)\widetilde\nabla w\Big)\Big\|_{L^1_t(L^2)}\\
&\quad\lesssim \Big\|\nabla\Big(\frac{h_1}{1+h_1}\Big)\Big\|_{E^0_t}\|(1+h_1)\widetilde\nabla w\|_{\widetilde L^1_t(\dot B^0_{2,\infty})}\\
&\quad\lesssim \Big\|\frac{h_1}{1+h_1}\Big\|_{E^1_t}(1+\|h_1\|_{L^\infty_t(B^1)})\|\nabla w\|_{\widetilde L^1_t(\dot B^0_{2,\infty})}\\
&\quad\lesssim \|h_1\|_{\widetilde E^{0,1}_t}(1+\|h_1\|_{L^\infty_t(\widetilde B^{0,1})})^4\|w\|_{\widetilde L^1_t(\dot B^1_{2,\infty})}.
\end{aligned}\tag{4.16}$$

In terms of Proposition 4.3, (4.12)–(4.16), and $\widetilde B^{0,1} \hookrightarrow B^1$, we finally obtain

$$\begin{aligned}
&\|w\|_{\widetilde L^1_t(\dot B^1_{2,\infty})} + \|w\|_{\widetilde L^2_t(\dot B^0_{2,\infty})}\\
&\lesssim \|u_2\|_{L^2_t(B^1)}\|w\|_{\widetilde L^2_t(\dot B^0_{2,\infty})} + \|u_1\|_{L^2_t(B^1)}\|w\|_{\widetilde L^2_t(\dot B^0_{2,\infty})}\\
&\quad + \|h_1\|_{\widetilde E^{0,1}_t}\big(1+\|h_1\|_{L^\infty_t(\widetilde B^{0,1})}\big)^4\|w\|_{\widetilde L^1_t(B^1_{2,\infty})}\\
&\quad + \int_0^t (1+\|h_1\|_{\widetilde B^{0,1}})(1+\|h_1\|_{\widetilde B^{0,1}}+\|h_2\|_{\widetilde B^{0,1}})(1+\|u_2\|_{B^2})\|\vartheta\|_{\dot B^0_{2,\infty}}d\tau.
\end{aligned}\tag{4.17}$$

Let us define

$$Z(t) \triangleq \|w\|_{\widetilde L^1_t(\dot B^1_{2,\infty})} + \|w\|_{\widetilde L^2_t(\dot B^0_{2,\infty})}.$$

Due to (4.6), if T is chosen small enough, then the first three terms on the right side of (4.17) can be absorbed by the left side $Z(t)$. Noting that $r\log(e+\frac{W(T)}{r})$ is increasing, from (4.11) and (4.17), it follows that

$$\begin{aligned}
Z(t) &\lesssim \int_0^t (1+W'(\tau))Z(\tau)\log\Big(e+\frac{W(\tau)}{Z(\tau)}\Big)d\tau\\
&\lesssim \int_0^t (1+W'(\tau))Z(\tau)\log\Big(e+\frac{W(T)}{Z(\tau)}\Big)d\tau.
\end{aligned}\tag{4.18}$$

It is easy to verify that

$$1+W'(\tau)\in L^1_{loc}(\mathbb{R}^+) \quad\text{and}\quad \int_0^1 \frac{dr}{r\log(e+\frac{W(T)}{r})} = +\infty.$$

Hence by the Osgood lemma, we have $Z \equiv 0$ on $[0,\widetilde T]$, i.e., $w \equiv 0$; then from (4.9), $\vartheta = h_2 - h_1 \equiv 0$. Then a standard continuous argument gives the uniqueness.

5. Appendix. In this appendix, we prove some multilinear estimates in the weighted Besov space.

LEMMA 5.1. *Let A be a homogeneous smooth function of degree m. Assume that $-\frac{d}{2} < \rho \leq \frac{d}{2}$. Then it holds that*

$$\Big|\big(A(D)\Delta_k(v\cdot\nabla h), A(D)\Delta_k h\big)\Big| \leq C\|\mathcal{F}_k^m(t)\|_2\|A(D)\Delta_k h\|_2, \tag{5.1}$$

$$\Big|\big(A(D)\Delta_k(v\cdot\nabla u), A(D)\Delta_k u\big)\Big| \leq C\|\widetilde{\mathcal{F}}_k^m(t)\|_2\|A(D)\Delta_k u\|_2, \tag{5.2}$$

and

$$\begin{aligned}&\Big|\big(A(D)\Delta_k(v\cdot\nabla h), \Delta_k u\big) + \big(\Delta_k(v\cdot\nabla u), A(D)\Delta_k h\big)\Big|\\ &\qquad\leq C\big(\|\mathcal{F}_k^m(t)\|_2 + \|\widetilde{\mathcal{F}}_k^0(t)\|_2\big)\big(\|\Delta_k u\|_2 + \|A(D)\Delta_k h\|_2\big),\end{aligned} \tag{5.3}$$

where $\mathcal{F}_k^m(t)$ and $\widetilde{\mathcal{F}}_k^m(t)$ satisfy

$$\sum_{k\in\mathbb{Z}}\omega_k(T)2^{k(\rho-m)}\|\mathcal{F}_k^m(t)\|_{L^1_T(L^2)} \leq C\|h\|_{E^\rho_T}\|v\|_{L^1_T(B^{\frac{d}{2}+1})}, \tag{5.4}$$

$$\sum_{k\in\mathbb{Z}}2^{k(\rho-m)}\|\widetilde{\mathcal{F}}_k^m(t)\|_{L^1_T(L^2)} \leq C\|u\|_{L^2_T(B^{\rho+1})}\|v\|_{L^2_T(B^{\frac{d}{2}})}. \tag{5.5}$$

Proof. Let us first prove (5.1). Using Bony's paraproduct decomposition, we write

$$\big(A(D)\Delta_k(v\cdot\nabla h), A(D)\Delta_k h\big) = \big(A(D)\Delta_k(T'_{\partial_j h}v^j), A(D)\Delta_k h\big) + J_k, \tag{5.6}$$

where

$$\begin{aligned}&T'_f g = T_f g + R(f,g), \quad \text{and}\\ &J_k = \sum_{|k'-k|\leq 3}([A(D)\Delta_k, S_{k'-1}v^j]\Delta_{k'}\partial_j h, A(D)\Delta_k h)\\ &\qquad + \sum_{|k'-k|\leq 3}((S_{k'-1}-S_{k-1})v^j A(D)\Delta_k\Delta_{k'}\partial_j h, A(D)\Delta_k h)\\ &\qquad + (S_{k-1}v^j A(D)\Delta_k\partial_j h, A(D)\Delta_k h).\end{aligned}$$

We get by integration by parts that

$$(S_{k-1}v^j A(D)\Delta_k\partial_j h, A(D)\Delta_k h) = -\frac{1}{2}\big(S_{k-1}\mathrm{div}\, vA(D)\Delta_k h, A(D)\Delta_k h\big).$$

Let us set

$$\begin{aligned}&\mathcal{F}^m_{k,0}(t) = A(D)\Delta_k(T'_{\partial_j h}v^j),\\ &\mathcal{F}^m_{k,1}(t) = \sum_{|k'-k|\leq 3}[A(D)\Delta_k, S_{k'-1}v^j]\Delta_{k'}\partial_j h,\\ &\mathcal{F}^m_{k,2}(t) = \sum_{|k'-k|\leq 3}(S_{k'-1}-S_{k-1})v^j A(D)\Delta_k\Delta_{k'}\partial_j h,\\ &\mathcal{F}^m_{k,3}(t) = -\frac{1}{2}S_{k-1}\mathrm{div}\, vA(D)\Delta_k h.\end{aligned}$$

By the Cauchy–Schwarz inequality, we get

$$\big(A(D)\Delta_k(v\cdot\nabla h), A(D)\Delta_k h\big) \le \|\mathcal{F}_k^m(t)\|_2 \|A(D)\Delta_k h\|_2,$$

with $\mathcal{F}_k^m(t) = \sum_{i=0}^3 \mathcal{F}_{k,i}^m(t)$. So, it remains to prove that $\mathcal{F}_k^m(t)$ satisfies (5.4). For the simplicity, we set

$$\widetilde{\Delta}_k = \sum_{|k'-k|\le 1} \Delta_{k'}, \quad \widetilde{\widetilde{\Delta}}_k = \sum_{|k'-k|\le 3} \Delta_{k'}.$$

Thanks to the definition of $\mathcal{F}_{k,0}^m(t)$ and Lemma 2.1, we have

$$\begin{aligned}\|\mathcal{F}_{k,0}^m(t)\|_{L^1_T(L^2)} &\le \sum_{|k'-k|\le 3} 2^{km}\|S_{k'-1}\partial_j h\|_{L^\infty_T(L^\infty)}\|\Delta_{k'}v^j\|_{L^1_T(L^2)} \\ &\quad + \sum_{k'\ge k-2} 2^{(m+\frac{d}{2})k}\|\Delta_k(\Delta_{k'}\partial_j h\widetilde{\Delta}_{k'}v^j)\|_{L^1_T(L^1)} \\ &\triangleq I + II.\end{aligned}$$

Thanks to Lemma 2.1, we have

$$\begin{aligned}2^{k(\rho-m)}I &\lesssim 2^{k\rho}\sum_{k'\le k+1} 2^{k'(1+\frac{d}{2})}\|\Delta_{k'}h\|_{L^\infty_T(L^2)}\|\widetilde{\widetilde{\Delta}}_k v\|_{L^1_T(L^2)} \\ &\lesssim \sum_{k'\le k+1} 2^{(k'-k)(1+\frac{d}{2}-\rho)}2^{k'\rho}\|\Delta_{k'}h\|_{L^\infty_T(L^2)}2^{k(1+\frac{d}{2})}\|\widetilde{\widetilde{\Delta}}_k v\|_{L^1_T(L^2)},\end{aligned}$$

from which, together with the definition of $\omega_k(T)$, it follows that

$$\begin{aligned}&\sum_{k\in\mathbb{Z}}\omega_k(T)2^{k(\rho-m)}I \\ &\lesssim \sum_{k'\in\mathbb{Z}} 2^{k'\rho}\|\Delta_{k'}h\|_{L^\infty_T(L^2)}\sum_{k\ge k'-1}\omega_k(T)2^{(k'-k)(1+\frac{d}{2}-\rho)}2^{k(1+\frac{d}{2})}\|\widetilde{\widetilde{\Delta}}_k v\|_{L^1_T(L^2)} \\ &\lesssim \sum_{k'\in\mathbb{Z}}\omega_{k'}(T)2^{k'\rho}\|\Delta_{k'}h\|_{L^\infty_T(L^2)}\sum_{k\ge k'-1}2^{(k'-k)(\frac{d}{2}-\rho)}2^{k(1+\frac{d}{2})}\|\widetilde{\widetilde{\Delta}}_k v\|_{L^1_T(L^2)} \\ &\lesssim \|h\|_{E^\rho_T}\|v\|_{L^1_T(B^{\frac{d}{2}+1})}, \qquad (5.7)\end{aligned}$$

where we used the assumption $\rho \le \frac{d}{2}$ in the last inequality. Set $e_k(T) = e_k^1(T)+e_k^2(T)$. Using Lemma 2.1, we also have

$$\begin{aligned}\omega_k(T)2^{k(\rho-m)}II &\lesssim \omega_k(T)2^{k(\rho+\frac{d}{2})}\sum_{k'\ge k-2}2^{k'}\|\Delta_{k'}h\|_{L^\infty_T(L^2)}\|\widetilde{\Delta}_{k'}v\|_{L^1_T(L^2)} \\ &\lesssim 2^{k(\rho+\frac{d}{2})}\sum_{k'\ge k-2}2^{k'}\|\Delta_{k'}h\|_{L^\infty_T(L^2)}\|\widetilde{\Delta}_{k'}v\|_{L^1_T(L^2)}\sum_{k'\ge\widetilde{k}\ge k}2^{-(\widetilde{k}-k)}e_{\widetilde{k}}(T) \\ &\quad + 2^{k(\rho+\frac{d}{2})}\sum_{k'\ge k-2}2^{k'}\|\Delta_{k'}h\|_{L^\infty_T(L^2)}\|\widetilde{\Delta}_{k'}v\|_{L^1_T(L^2)} \\ &\qquad \sum_{\widetilde{k}\ge k,\widetilde{k}\ge k'}2^{-(\widetilde{k}-k)}e_{\widetilde{k}}(T) \\ &\triangleq II_1 + II_2.\end{aligned}$$

Note that for $\widetilde{k} \leq k'$

$$e_{\widetilde{k}}(T) \leq e_{k'}(T) \leq \omega_{k'}(T),$$

from which, together with $\rho > -\frac{d}{2}$, we deduce that

$$\sum_{k\in Z} II_1 \lesssim \sum_{k'\in Z} \omega_{k'}(T)2^{k'\rho}\|\Delta_{k'}h\|_{L^\infty_T(L^2)}2^{k'(\frac{d}{2}+1)}\|\widetilde{\Delta}_{k'}v\|_{L^1_T(L^2)} \sum_{k\leq k'+2} 2^{(k-k')(\rho+\frac{d}{2})}$$

$$\lesssim \|h\|_{E^\rho_T}\|v\|_{L^1_T(B^{\frac{d}{2}+1})}. \tag{5.8}$$

Similarly, we can obtain

$$\begin{aligned}\sum_{k\in\mathbb{Z}} II_2 &\lesssim \sum_{k'\in\mathbb{Z}} 2^{k'\rho}\|\Delta_{k'}h\|_{L^\infty_T(L^2)} \sum_{k\leq k'+2} 2^{(k-k')(\frac{d}{2}+\rho)} \sum_{\widetilde{k}\geq k'} 2^{-(\widetilde{k}-k)}e_{\widetilde{k}}(T)\|v\|_{L^1_T(B^{\frac{d}{2}+1})}\\ &\lesssim \sum_{k'\in\mathbb{Z}} \omega_{k'}(T)2^{k'\rho}\|\Delta_{k'}h\|_{L^\infty_T(L^2)} \sum_{k\leq k'+2} 2^{(k-k')(\frac{d}{2}+\rho+1)}\|v\|_{L^1_T(B^{\frac{d}{2}+1})}\\ &\lesssim \|h\|_{E^\rho_T}\|v\|_{L^1_T(B^{\frac{d}{2}+1})}. \end{aligned} \tag{5.9}$$

By summing up (5.7)–(5.9), we obtain

$$\sum_{k\in\mathbb{Z}} \omega_k(T)2^{k(\rho-m)}\|\mathcal{F}^m_{k,0}(t)\|_{L^1_T(L^2)} \lesssim \|h\|_{E^\rho_T}\|v\|_{L^1_T(B^{\frac{d}{2}+1})}. \tag{5.10}$$

Note that $A(D)\Delta_k = 2^{km}\widetilde{\varphi}(2^{-k}D)$ with $\widetilde{\varphi}(\xi) = A(\xi)\varphi(\xi)$. Setting $\tilde{\theta} = \mathcal{F}^{-1}\tilde{\varphi}$, we get by using Taylor's formula that

$$\begin{aligned}\mathcal{F}^m_{k,1}(t) = \sum_{|k'-k|\leq 3} 2^{k(m-1)} \int_{\mathbb{R}^d}\int_0^1 \tilde{\theta}(y)\\ (y\cdot S_{k'-1}\nabla v^j(x-2^{-k}\tau y))\Delta_{k'}\partial_j h(x-2^{-k}y)d\tau dy,\end{aligned}$$

from which, together with Lemma 2.1, it follows that

$$\begin{aligned}\|\mathcal{F}^m_{k,1}(t)\|_{L^1_T(L^2)} &\lesssim 2^{k(m-1)} \sum_{|k'-k|\leq 3} \|S_{k'-1}\nabla v^j\|_{L^1_T(L^\infty)}\|\Delta_{k'}\partial_j h\|_{L^\infty_T(L^2)}\\ &\lesssim 2^{km} \sum_{|k'-k|\leq 3} \|\Delta_{k'}h\|_{L^\infty_T(L^2)}\|v\|_{L^1_T(B^{\frac{d}{2}+1})};\end{aligned}$$

thus, we get

$$\sum_{k\in\mathbb{Z}} \omega_k(T)2^{k(\rho-m)}\|\mathcal{F}^m_{k,1}(t)\|_{L^1_T(L^2)} \lesssim \|h\|_{E^\rho_T}\|v\|_{L^1_T(B^{\frac{d}{2}+1})}. \tag{5.11}$$

Thanks to the fact $|k'-k| \leq 3$ and Lemma 2.1, we have

$$\|(S_{k'-1}-S_{k-1})v^j A(D)\Delta_k\Delta_{k'}\partial_j h\|_{L^1_T(L^2)} \lesssim 2^{km}\|\Delta_k h\|_{L^\infty_T(L^2)}\|v\|_{L^1_T(B^{\frac{d}{2}+1})},$$

from which it follows that

$$\sum_{k\in\mathbb{Z}}\omega_k(T)2^{k(\rho-m)}\big(\|\mathcal{F}^m_{k,2}(t)\|_{L^1_T(L^2)}+\|\mathcal{F}^m_{k,3}(t)\|_{L^1_T(L^2)}\big)\lesssim\|h\|_{E^\rho_T}\|v\|_{L^1_T(B^{\frac{d}{2}+1}_)},\tag{5.12}$$

which, together with (5.10) and (5.11), yields (5.4).

Using the decomposition (5.6), with h instead of u, and Lemma 2.1, (5.2) can be easily proved. We omit the proof here. In order to prove (5.3), we use the decomposition

$$\big(A(D)\Delta_k(v\cdot\nabla h),\Delta_k u\big)+\big(\Delta_k(v\cdot\nabla u),A(D)\Delta_k h\big)=I_k+J_k,$$

with

$$\begin{aligned}
I_k&=\big(A(D)\Delta_k(T'_{\partial_j h}v^j),\Delta_k u\big)+\big(\Delta_k(T'_{\partial_j u}v^j),A(D)\Delta_k h\big)\\
&\triangleq\big(\mathcal{F}^m_{k,0}(t),\Delta_k u\big)+\big(\widetilde{\mathcal{F}}^0_{k,0}(t),A(D)\Delta_k h\big),\\
J_k&=\sum_{|k'-k|\le3}\Big([A(D)\Delta_k,S_{k'-1}v^j]\Delta_{k'}\partial_j h,\Delta_k u\Big)\\
&\quad+\Big((S_{k'-1}-S_{k-1})v^jA(D)\Delta_k\Delta_{k'}\partial_j h,\Delta_k u\Big)\\
&\quad+\sum_{|k'-k|\le3}\Big([\Delta_k,S_{k'-1}v^j]\Delta_{k'}\partial_j u,A(D)\Delta_k h\Big)\\
&\quad+\Big((S_{k'-1}-S_{k-1})v^j\Delta_k\Delta_{k'}\partial_j u,A(D)\Delta_k h\Big)\\
&\quad-\Big(S_{k-1}\mathrm{div}vA(D)\Delta_k h,\Delta_k u\Big)\\
&\triangleq\big(\mathcal{F}^m_{k,1}(t),\Delta_k u\big)+\big(\mathcal{F}^m_{k,2}(t),\Delta_k u\big)+\big(\widetilde{\mathcal{F}}^0_{k,1}(t),A(D)\Delta_k h\big)\\
&\quad+\big(\widetilde{\mathcal{F}}^0_{k,2},A(D)\Delta_k h\big)+\big(\mathcal{F}^m_{k,3}(t),\Delta_k u\big),
\end{aligned}$$

from which, using a similar proof of (5.4), we obtain (5.3). This completes the proof of Lemma 5.1. □

LEMMA 5.2. *Let $s_1\le\frac{d}{2}-1$, $s_2\le\frac{d}{2}$, and $s_1+s_2>0$. Then it holds that*

$$\sum_{k\in\mathbb{Z}}\omega_k(T)2^{k(s_1+s_2-\frac{d}{2})}\|\Delta_k(fg)\|_{L^1_T(L^2)}\le C\sum_{k\in\mathbb{Z}}\omega_k(T)2^{ks_1}\|\Delta_k f\|_{L^{r_1}_T(L^2)}\|g\|_{L^{r_2}_T(B^{s_2})},\tag{5.13}$$

where $1\le r_1,r_2\le\infty$ and $\frac{1}{r_1}+\frac{1}{r_2}=1$.

Proof. Using Bony's paraproduct decomposition, we write

$$\begin{aligned}
\Delta_k(fg)&=\sum_{|k'-k|\le3}\Delta_k(S_{k'-1}f\Delta_{k'}g)+\sum_{|k'-k|\le3}\Delta_k(S_{k'-1}g\Delta_{k'}f)\\
&\quad+\sum_{k'\ge k-2}\Delta_k(\Delta_{k'}f\widetilde{\Delta}_{k'}g)\triangleq I+II+III.
\end{aligned}$$

A similar proof of (5.7) ensures that for $s_1 \leq \frac{d}{2} - 1$

$$\sum_{k\in\mathbb{Z}} \omega_k(T) 2^{k(s_1+s_2-\frac{d}{2})} \|I\|_{L^1_T(L^2)} \lesssim \sum_{k\in\mathbb{Z}} \omega_k(T) 2^{ks_1} \|\Delta_k f\|_{L^{r_1}_T(L^2)} \|g\|_{L^{r_2}_T(B^{s_2})},$$

while II can be directly deduced for $s_2 \leq \frac{d}{2}$. On the other hand, a similar proof of (5.8) and (5.9) gives for $s_1 + s_2 > 0$

$$\sum_{k\in\mathbb{Z}} \omega_k(T) 2^{k(s_1+s_2-\frac{d}{2})} \|III\|_{L^1_T(L^2)} \lesssim \sum_{k\in\mathbb{Z}} \omega_k(T) 2^{ks_1} \|\Delta_k f\|_{L^{r_1}_T(L^2)} \|g\|_{L^{r_2}_T(B^{s_2})}.$$

This completes the proof of Lemma 5.2. □

Similarly, we can also prove the following lemma.

LEMMA 5.3. *Let $s_1 \leq \frac{d}{2} - 1$, $s_2 < \frac{d}{2}$, and $s_1 + s_2 \geq 0$. Then it holds that*

$$\sup_{k\in\mathbb{Z}} \omega_k(T) 2^{k(s_1+s_2-\frac{d}{2})} \|\Delta_k(fg)\|_{L^1_T(L^2)} \leq C\|f\|_{E^{s_1}_T} \|g\|_{\widetilde{L}^1_T(\dot{B}^{s_2}_{2,\infty})}. \tag{5.14}$$

LEMMA 5.4. *Let $s > 0$. Assume that $F \in W^{[s]+3,\infty}_{loc}(\mathbb{R}^d)$ with $F(0) = 0$. Then it holds that*

$$\|F(f)\|_{E^s_T} \leq C(1 + \|f\|_{L^\infty_T(L^\infty)})^{[s]+2} \|f\|_{E^s_T}. \tag{5.15}$$

Proof. We decompose $F(f)$ as

$$\begin{aligned} F(f) = \sum_{k'\in\mathbb{Z}} F(S_{k'+1}f) - F(S_{k'}f) &= \sum_{k'\in\mathbb{Z}} \Delta_{k'} f \int_0^1 F'(S_{k'}f + \tau\Delta_{k'}f) d\tau \\ &\triangleq \sum_{k'\in\mathbb{Z}} \Delta_{k'} f \, m_{k'}, \end{aligned}$$

where $m_{k'} = \int_0^1 F'(S_{k'}f + \tau\Delta_{k'}f)d\tau$. Furthermore, we write

$$\Delta_k F(f) = \sum_{k'<k} \Delta_k(\Delta_{k'} f \, m_{k'}) + \sum_{k'\geq k} \Delta_k(\Delta_{k'} f \, m_{k'}) \triangleq I + II.$$

By Lemma 2.1, we have

$$\begin{aligned} \|I\|_{L^\infty_T(L^2)} &\leq \sum_{k'<k} \|\Delta_k(\Delta_{k'} f \, m_{k'})\|_{L^\infty_T(L^2)} \\ &\leq \sum_{k'<k} 2^{-k|\alpha|} \sup_{|\gamma|=|\alpha|} \|D^\gamma \Delta_k(\Delta_{k'} f \, m_{k'})\|_{L^\infty_T(L^2)}, \end{aligned} \tag{5.16}$$

with α to be determined later. Note that for $|\gamma| \geq 0$, we have

$$\|D^\gamma m_{k'}\|_\infty \lesssim 2^{k'|\gamma|}(1 + \|f\|_\infty)^{|\gamma|} \|F'\|_{W^{|\gamma|,\infty}},$$

from which, together with (5.16), it follows that

$$2^{ks}\|I\|_{L^\infty_T(L^2)} \lesssim 2^{k(s-|\alpha|)} \sum_{k'<k} 2^{k'|\alpha|} \|\Delta_{k'} f\|_{L^\infty_T(L^2)} (1 + \|f\|_{L^\infty_T(L^\infty)})^{|\alpha|} \|F'\|_{W^{|\alpha|,\infty}}.$$

Thus, if we take $|\alpha| = [s] + 2$, we get

$$\begin{aligned}
&\sum_{k\in\mathbb{Z}} \omega_k(T) 2^{ks} \|I\|_{L^\infty_T(L^2)} \\
&\quad \lesssim \sum_{k'\in\mathbb{Z}} 2^{k's} \omega_{k'}(T) \|\Delta_{k'} f\|_{L^\infty_T(L^2)} \sum_{k>k'} 2^{(k-k')(s-|\alpha|+1)} (1 + \|f\|_{L^\infty_T(L^\infty)})^{|\alpha|} \|F'\|_{W^{|\alpha|,\infty}} \\
&\quad \lesssim (1 + \|f\|_{L^\infty_T(L^\infty)})^{[s]+2} \|F'\|_{W^{[s]+2,\infty}} \|f\|_{E^s_T}. \qquad (5.17)
\end{aligned}$$

Next, let us turn to the proof of II. We get by using Lemma 2.1 that

$$\|II\|_{L^\infty_T(L^2)} \lesssim \sum_{k'\geq k} \|\Delta_{k'} f\|_{L^\infty_T(L^2)}.$$

Then we write

$$\begin{aligned}
\sum_{k\in\mathbb{Z}} \omega_k(T) 2^{ks} \|II\|_{L^\infty_T(L^2)} &\lesssim \sum_{k\in\mathbb{Z}} 2^{ks} \sum_{k'\geq k} \|\Delta_{k'} f\|_{L^\infty_T(L^2)} \sum_{k'\geq \widetilde{k}\geq k} 2^{-(\widetilde{k}-k)} e_{\widetilde{k}}(T) \\
&\quad + \sum_{k\in\mathbb{Z}} 2^{ks} \sum_{k'\geq k} \|\Delta_{k'} f\|_{L^\infty_T(L^2)} \sum_{\widetilde{k}\geq k', \widetilde{k}\geq k} 2^{-(\widetilde{k}-k)} e_{\widetilde{k}}(T),
\end{aligned}$$

from which a similar proof of (5.8) and (5.9) ensures that

$$\sum_{k\in\mathbb{Z}} \omega_k(T) 2^{ks} \|II\|_{L^\infty_T(L^2)} \lesssim \|f\|_{E^s_T}. \qquad (5.18)$$

By summing up (5.17) and (5.18), we deduce the inequality (5.15). This completes the proof of Lemma 5.4. □

Acknowledgment. The authors thank the referees for their invaluable comments and suggestions which helped improve the paper greatly.

REFERENCES

[1] H. Bahouri and J.-Y. Chemin, *Equations de transport relatives à des champs des vecteurs non-lipschitziens et mécanique des fluides*, Arch. Rational Mech. Anal., 127 (1994), pp. 159–199.

[2] H. Bahouri and J.-Y. Chemin, *Équations d'ondes quasilinéaires et estimations de Strichartz*, Amer. J. Math., 121 (1999), pp. 1337–1377.

[3] J.-M. Bony, *Calcul symbolique et propagation des singularitiés pour les équations aux dérivées partielles non linéaires*, Ann. Sci. École Norm. Sup. (4), 14 (1981), pp. 209–246.

[4] D. Bresch, B. Desjardins, and G. Métivier, *Recent Mathematical Results and Open Problems about Shallow Water Equations*, preprint.

[5] A. T. Bui, *Existence and uniqueness of a classical solution of an initial boundary value problem of the theory of shallow waters*, SIAM J. Math. Anal., 12 (1981), pp. 229–241.

[6] M. Cannone, *Ondelettes, paraproduits et Navier-Stokes*, Nouveaux essais, Diderot éditeurs, Paris, 1995.

[7] M. Cannone, *Harmonic analysis tools for solving the incompressible Navier–Stokes equations*, in Handbook of Mathematical Fluid Dynamics, Vol. III, North-Holland, Amsterdam, 2004, pp. 161–244.

[8] J.-Y. Chemin, *Perfect Incompressible Fluids*, Oxford University Press, New York, 1998.

[9] J.-Y. Chemin, *Théorèmes d'unicité pour le système de Navier-Stokes tridimensionnel*, J. Anal. Math., 77 (1999), pp. 27–50.

[10] J.-Y. Chemin and N. Lerner, *Flot de champs de vecteurs non lipschitziens et équations de Navier-Stokes*, J. Differential Equations, 121 (1992), pp. 314–328.
[11] R. Danchin, *Global existence in critical spaces for compressible Navier-Stokes equations*, Invent. Math., 141 (2000), pp. 579–614.
[12] R. Danchin, *Global existence in critical spaces for flows of compressible viscous and heat-conductive gases*, Arch. Rational Mech. Anal., 160 (2001), pp. 1–39.
[13] R. Danchin, *Local theory in critical spaces for compressible viscous and heat-conductive gases*, Comm. Partial Differential Equations, 26 (2001), pp. 1183–1233.
[14] R. Danchin, *Density-dependent incompressible viscous fluids in critical spaces*, Proc. Roy. Soc. Edinburgh Sect. A, 133 (2003), pp. 1311–1334.
[15] R. Danchin, *On the uniqueness in critical spaces for compressible Navier-Stokes equations*, Nonlinear Differential Equations Appl., 12 (2005), pp. 111–128.
[16] H. Fujita and T. Kato, *On the Navier-Stokes initial value problem* I, Arch. Rational Mech. Anal., 16 (1964), pp. 269–315.
[17] P. E. Kloeden, *Global existence of classical solutions in the dissipative shallow water equations*, SIAM J. Math. Anal., 16 (1985), pp. 301–315.
[18] Y. Meyer, *Wavelets, Paraproducts and Navier-Stokes Equations. Current Developments in Mathematics*, International Press, Boston, MA, 1996.
[19] T. Runst and W. Sickel, *Sobolev spaces of fractional order, Nemytskij operators, and nonlinear partial differential equations*, de Gruyter Ser. Nonlinear Anal. Appl. 3, Walter de Gruyter, Berlin, 1996.
[20] L. Sundbye, *Global existence for the Dirichlet problem for the viscous shallow water equations*, J. Math. Anal. Appl., 202 (1996), pp. 236–258.
[21] L. Sundbye, *Global existence for the Cauchy problem for the viscous shallow water equations*, Rocky Mountain J. Math., 28 (1998), pp. 1135–1152.
[22] H. Triebel, *Theory of Function Spaces*, Monogr. Math. 78, Birkhäuser Verlag, Basel, Boston, Stuttgart, 1983.
[23] W.-K. Wang and C.-J. Xu, *The Cauchy problem for viscous shallow water equations*, Rev. Mat. Iberoamericana, 21 (2005), pp. 1–24.

SIAM J. MATH. ANAL.
Vol. 40, No. 2, pp. 475–490

BLOW-UP AND DECAY OF THE SOLUTION OF THE WEAKLY DISSIPATIVE DEGASPERIS–PROCESI EQUATION*

SHUYIN WU† AND ZHAOYANG YIN†

Abstract. In this paper, we mainly study several problems on the weakly dissipative Degasperis–Procesi equation. We first establish the local well-posedness of the equation, derive a precise blow-up scenario, and present two blow-up criteria for strong solutions to the equation. We then give the precise blow-up rate of blow-up solutions to the equation. We finally prove that the equation has global strong solutions and these global solutions decay to zero as time goes to infinity provided the potentials associated with their initial data are of one sign.

Key words. the weakly dissipative Degasperis–Procesi equation, blow-up, blow-up rate, global strong solutions, decay of solutions

AMS subject classifications. 35G25, 35L05

DOI. 10.1137/07070855X

1. Introduction.

Recently, Degasperis and Procesi [20] studied the following family of third-order dispersive conservation laws:

$$u_t + c_0 u_x + \gamma u_{xxx} - \alpha^2 u_{txx} = (c_1 u^2 + c_2 u_x^2 + c_3 u u_{xx})_x, \tag{1.1}$$

where α, c_0, c_1, c_2, and c_3 are real constants. They found [20] that there are only three equations that satisfy the asymptotic integrability condition within this family: the KdV equation, the Camassa–Holm equation, and the Degasperis–Procesi equation.

If $\alpha = c_2 = c_3 = 0$, then (1.1) becomes the well-known KdV equation which describes the unidirectional propagation of waves at the free surface of shallow water under the influence of gravity. In this model $u(t, x)$ represents the wave's height above a flat bottom, x is proportional to the distance in the direction of propagation, and t is proportional to the elapsed time. The KdV equation is completely integrable, and its solitary waves are solitons [21, 39]. The Cauchy problem of the KdV equation has been the subject of a number of studies, and a satisfactory local or global (in time) existence theory is now in hand (for example, see [31, 43]). It is shown that the KdV equation is globally well-posed for $u_0 \in L^2(\mathbb{R})$; cf. [43]. It is observed that the KdV equation does not accommodate wave breaking (by wave breaking we mean that the wave remains bounded, but its slope becomes unbounded in finite time [45]).

For $c_1 = -\frac{3}{2}c_3/\alpha^2$ and $c_2 = c_3/2$, (1.1) becomes the Camassa–Holm equation, modeling the unidirectional propagation of shallow water waves over a flat bottom. Again $u(t, x)$ stands for the fluid velocity at time t in the spatial x direction, and c_0 is a nonnegative parameter related to the critical shallow water speed [3, 22, 30]. The Camassa–Holm equation is also a model for the propagation of axially symmetric waves in hyperelastic rods [15, 18]. It has a bi-Hamiltonian structure [33, 27] and is completely integrable [3, 8]. Its solitary waves are smooth if $c_0 > 0$ and peaked in the limiting case $c_0 = 0$; see [4]. The orbital stability of the peaked solitons is proved

*Received by the editors November 19, 2007; accepted for publication (in revised form) January 30, 2008; published electronically May 16, 2008.
http://www.siam.org/journals/sima/40-2/70855.html

†Department of Computer Science, Sun Yet-sen University, 510275 Guangzhou, China (isswsy@mail.sysu.edu.cn, mcsyzy@mail.sysu.edu.cn). The second author's research was partially supported by the AvH Foundation, the NNSF of China (10531040), and the NSF of Guangdong Province.

in [16] and that of the smooth solitons in [17]. The explicit interaction of the peaked solitons is given in [1].

The Cauchy problem of the Camassa–Holm equation has been studied extensively. It has been shown that this equation is locally well-posed [9, 34, 42] for initial data $u_0 \in H^s(\mathbb{R})$, $s > \frac{3}{2}$. More interestingly, it has global strong solutions [7, 9] and also finite time blow-up solutions [7, 9, 10, 13]. On the other hand, it has global weak solutions in $H^1(\mathbb{R})$ [2, 11, 14, 47]. It is also known that if u is the solution of the Camassa–Holm equation with the initial data u_0 in $H^1(\mathbb{R})$, then we have the following a priori estimate:

$$\|u(t,\cdot)\|_{L^\infty(\mathbb{R})} \le \sqrt{2}\|u(t,\cdot)\|_{H^1(\mathbb{R})} \le \sqrt{2}\|u_0(\cdot)\|_{H^1(\mathbb{R})}$$

for all $t > 0$. The advantage of the Camassa–Holm equation in comparison with the KdV equation lies in the fact that the Camassa–Holm equation has peaked solitons and models wave breaking [4].

If $c_1 = -2c_3/\alpha^2$ and $c_2 = c_3$ in (1.1), then after rescaling, shifting the dependent variable, and applying a Galilean boost [19], we find the Degasperis–Procesi equation of the form

$$u_t - u_{txx} + 4uu_x = 3u_xu_{xx} + uu_{xxx}, \quad t > 0, \quad x \in \mathbb{R}. \tag{1.2}$$

The integrability of (1.2) was proved in [19] by constructing a Lax pair. It was also shown in [19] that (1.2) has a bi-Hamiltonian structure and admits exact peakon solutions which are analogous to the Camassa–Holm peakons.

The Degasperis–Procesi equation can be regarded as a model for the motion of shallow water waves, and its asymptotic accuracy is the same as for the Camassa–Holm shallow water equation [23, 29]. An inverse scattering approach for computing n-peakon solutions to (1.2) was presented in [37]. The traveling wave solutions to (1.2) were investigated in [44, 32]. The multisoliton solutions to (1.2) and their peakon limits were studied in [38].

The Cauchy problem of the Degasperis–Procesi equation is locally well-posed [48] for initial data $u_0 \in H^s(\mathbb{R})$ with $s > \frac{3}{2}$. Analogous to the Camassa–Holm equation, it not only has global strong solutions [35, 51] but also blow-up solutions in finite time [35, 48]. Meanwhile, the Degasperis–Procesi equation also has global weak solutions with initial data $u_0 \in H^1(\mathbb{R})$ (cf. [50, 51]) and global entropy weak solutions in the class of $L^1(\mathbb{R}) \cap BV(\mathbb{R})$ and the class of $L^2(\mathbb{R}) \cap L^4(\mathbb{R})$; cf. [5].

Despite the similarities to the Camassa–Holm equation, we would like to point out that these two equations are truly different. One of the important features of the Degasperis–Procesi equation is that it not only has peakon solitons [19] and periodic peakons [49] but also shock peakons [6, 36] and periodic shock waves [25]. On the other hand, the isospectral problem for the Degasperis–Procesi equation is the third-order equation in the Lax pair [19], while the isospectral problem for the Camassa–Holm equation is the second-order equation [3]. Another indication of the fact that there is no simple transformation of the Degasperis–Procesi equation into the Camassa–Holm equation is the entirely different form of conservation laws for these two equations [3, 19]. Furthermore, the Camassa–Holm equation is a reexpression of geodesic flow on the diffeomorphism group [12] or on the Bott–Virasoro group [40]. Up to now, no geometric derivation of the Degasperis–Procesi equation has been available.

Recently, several new global existence and blow-up results for strong solutions to the Degasperis–Procesi equation were presented in [35]. It is proved that the first blow-up must occur as wave breaking and shock waves possibly appear afterward [35].

Global weak solution and blow-up structure for this equation were investigated in [24]. Initial boundary value problems for the Degasperis–Procesi equation were also discussed in [26].

In general, it is quite difficult to avoid energy dissipation mechanisms in a real world. Ott and Sudan [41] ever investigated how the KdV equation was modified by the presence of dissipation and the effect of such dissipation on the solitary solution of the KdV equation. Ghidaglia [28] studied the long time behavior of solutions to the weakly dissipative KdV equation as a finite-dimensional dynamical system. Recently, Wu and Yin [46] discussed the blow-up, blow-up rate, and decay of the solution of the weakly dissipative periodic Camassa–Holm equation.

In this paper, we would like to consider the following dissipative Degasperis–Procesi equation:

$$u_t - u_{txx} + 4uu_x + L(u) = 3u_x u_{xx} + uu_{xxx}, \quad t > 0, \quad x \in \mathbb{R},$$

where $L(u)$ is a dissipative term, and L can be a differential operator or a quasi-differential operator according to different physical situations. We are interested in the effect of the weakly dissipative term on the Degasperis–Procesi equation. In particular, we study the following weakly dissipative Degasperis–Procesi equation:

$$\begin{cases} y_t + uy_x + 3u_x y + \lambda y = 0, \quad t > 0, \quad x \in \mathbb{R}, \\ y = u - u_{xx}, \quad t > 0, \quad x \in \mathbb{R}, \\ u(0,x) = u_0(x), \qquad x \in \mathbb{R}, \end{cases} \tag{1.3}$$

where $\lambda y = \lambda(1 - \partial_x^2)u$ is the weakly dissipative term, and $\lambda > 0$ is a constant.

We find that the behaviors of (1.3) are similar to the Degasperis–Procesi equation in a finite interval of time, such as the local well-posedness and the blow-up phenomena. But there are considerable differences between (1.3) and the Degasperis–Procesi equation in their long time behaviors. Global solution of (1.3) decays to zero as time goes to infinity provided the potential $y_0 = (1 - \partial_x^2)u_0$ is of one sign. This long time behavior is an important feature that the Degasperis–Procesi equation does not possess. It is known that the Degasperis–Procesi equation has peaked traveling wave solutions [19]. Theorem 4.1 in what follows shows that any global solution decays in the H^1-norm. This means that there are no traveling wave solutions of (1.3). This is also another considerable difference between (1.3) and the Degasperis–Procesi equation (1.2) in their long time behaviors.

It is very interesting that (1.3) has the same blow-up rate as the Degasperis–Procesi equation does when the blow-up occurs. This fact shows that the blow-up rate of the Degasperis–Procesi equation is not affected by the weakly dissipative term. But the occurrence of blow-up of (1.2) is affected by the dissipative parameter.

It should be noticed that the weakly dissipative term breaks the conservation laws of the following Degasperis–Procesi equation [19]:

$$E_1(u) = \int_{\mathbb{R}} y dx, \quad E_2(u) = \int_{\mathbb{R}} yv dx, \quad E_3(u) = \int_{\mathbb{R}} u^3 dx,$$

where $y = (1 - \partial_x^2)u$ and $v = (4 - \partial_x^2)^{-1}u$, which play an important role in the study of the Degasperis–Procesi equation [5, 24, 25, 35, 48, 49, 50, 51].

Notation. As above and henceforth, we denote by $*$ the convolution. We write $\hat{f}$ as the Fourier transform of f. We also use $(\ ,\ )$ to represent the standard inner product in $L^2(\mathbb{R})$. For $1 \le p \le \infty$, the norm in the Lebesgue space L^p will be written $\|\cdot\|_{L^p}$, while $\|\cdot\|_r$, $r \ge 0$, will stand for the norm in the classical Sobolev spaces $H^r(\mathbb{R})$.

2. Local well-posedness and blow-up. In this section, we establish the local well-posedness of (1.3), derive a precise blow-up scenario, and present two blow-up criteria for strong solutions to (1.3).

Note that if $p(x) = \frac{1}{2}e^{-|x|}$, $x \in \mathbb{R}$, then $(1-\partial_x^2)^{-1}f = p * f$ for all $f \in L^2(\mathbb{R})$ and $p * y = u$. Using this identity, we can rewrite (1.3) as follows:

$$\begin{cases} u_t + uu_x + \partial_x p * (\frac{3}{2}u^2) + \lambda u = 0, & t > 0, \quad x \in \mathbb{R}, \\ u(0,x) = u_0(x), & x \in \mathbb{R}. \end{cases} \tag{2.1}$$

Using Kato's semigroup approach along lines very similar to [48], one obtains the local well-posedness of (1.3) or (2.1).

THEOREM 2.1. *Given $u_0 \in H^r(\mathbb{R})$, $r > \frac{3}{2}$, there exist a maximal $T = T(\lambda, \|u_0\|_r) > 0$, and a unique solution u to* (1.3) *(or* (2.1)*), such that*

$$u = u(\cdot, u_0) \in C([0,T); H^r(\mathbb{R})) \cap C^1([0,T); H^{r-1}(\mathbb{R})),$$

and the solution depends continuously on the initial data; i.e., the mapping $u \to u(\cdot, u_0) : H^r(\mathbb{R}) \to C([0,T); H^r(\mathbb{R})) \cap C^1([0,T); H^{r-1}(\mathbb{R}))$ is continuous. Moreover, T may be chosen independent of r in the following sense. If

$$u = u(\cdot, u_0) \in C([0,T); H^r(\mathbb{R})) \cap C^1([0,T); H^{r-1}(\mathbb{R}))$$

to (1.3) *(or* (2.1)*), and if $u_0 \in H^{r'}(\mathbb{R})$ for some $r' \neq r, r' > \frac{3}{2}$, then*

$$u \in C([0,T); H^{r'}(\mathbb{R})) \cap C^1([0,T); H^{r'-1}(\mathbb{R}))$$

with the same T.

The following results are proved only with regard to $r = 3$, since we can obtain the same conclusion for the general case $r > \frac{3}{2}$ by using Theorem 2.1 and a simple density argument.

We now present a precise blow-up scenario for strong solutions to (1.3).

THEOREM 2.2. *Given $u_0 \in H^r(\mathbb{R})$, $r > \frac{3}{2}$, the solution of* (1.3) *(or* (2.1)*) blows up in a finite time $T > 0$ if and only if*

$$\liminf_{t \to T} \{\inf_{x \in \mathbb{R}} u_x(t,x)\} = -\infty.$$

Proof. Let $T > 0$ be the maximal time of existence of the solution u to (1.3) (or (2.1)) with initial data $u_0 \in H^3(\mathbb{R})$. By (1.3), we have

$$\begin{aligned} \frac{d}{dt}\int_{\mathbb{R}} y^2 dx &= 2\int_{\mathbb{R}} yy_t dx \\ &= -2\int_{\mathbb{R}} uyy_x dx - 6\int_{\mathbb{R}} u_x y^2 dx - 2\lambda \int_{\mathbb{R}} y^2 dx \\ &= -5\int_{\mathbb{R}} u_x y^2 dx - 2\lambda \int_{\mathbb{R}} y^2 dx. \end{aligned} \tag{2.2}$$

If $u_0 \in H^4(\mathbb{R})$, then we can obtain by (1.3)

$$\begin{aligned}(2.3)\quad \frac{d}{dt}\int_{\mathbb{R}} y_x^2 dx &= 2\int_{\mathbb{R}} y_x y_{xt} dx \\ &= -8\int_{\mathbb{R}} u_x y_x^2 dx - 2\int_{\mathbb{R}} u y_x y_{xx} dx - 6\int_{\mathbb{R}} u_{xx} y y_x dx - 2\lambda\int_{\mathbb{R}} y_x^2 dx \\ &= -7\int_{\mathbb{R}} u_x y_x^2 dx + 3\int_{\mathbb{R}} u_x y^2 dx - 2\lambda\int_{\mathbb{R}} y_x^2 dx.\end{aligned}$$

As for $u_0 \in H^3(\mathbb{R})$, we will show that (2.3) still holds. In fact, we can approximate u_0 in $H^3(\mathbb{R})$ by function $u_0^n \in H^4(\mathbb{R})$. Moreover, we write $u^n = u^n(\cdot, u_0^n)$ for the solution of (1.3) with initial data u_0^n.

By Theorem 2.1, we know that

$$u^n \in C([0,T_n); H^4(\mathbb{R})) \cap C^1([0,T_n); H^3(\mathbb{R})), \qquad n \geq 1,$$

$$y^n = u^n - u_{xx}^n \in C([0,T_n); H^2(\mathbb{R})) \cap C^1([0,T_n); H^1(\mathbb{R})), \qquad n \geq 1,$$

$u^n \to u$ in $H^3(\mathbb{R})$, and $T_n \to T$ as $n \to \infty$.

Due to $u_0^n \in H^4(\mathbb{R})$, we have by (2.3)

$$\frac{d}{dt}\int_{\mathbb{R}} (y_x^n)^2 dx = -7\int_{\mathbb{R}} u_x^n (y_x^n)^2 dx + 3\int_{\mathbb{R}} u_x^n (y^n)^2 dx - 2\lambda\int_{\mathbb{R}} (y_x^n)^2 dx.$$

Since $u^n \to u$ in $H^3(\mathbb{R})$ as $n \to \infty$, it follows that $u_x^n \to u_x$ in $L^\infty(\mathbb{R})$ as $n \to \infty$. Also note that $y^n \to y$ in $H^1(\mathbb{R})$ and $y_x^n \to y_x$ in $L^2(\mathbb{R})$ as $n \to \infty$. Letting n go to infinity in the above equation, we can easily deduce that (2.3) holds for $u_0 \in H^3(\mathbb{R})$.

Adding (2.2) and (2.3), we get

$$\begin{aligned}\frac{d}{dt}\left(\int_{\mathbb{R}} y^2 dx + \int_{\mathbb{R}} y_x^2 dx\right) &= -7\int_{\mathbb{R}} u_x y_x^2 dx - 2\int_{\mathbb{R}} u_x y^2 dx \\ &\quad - 2\lambda\left(\int_{\mathbb{R}} y^2 dx + \int_{\mathbb{R}} y_x^2 dx\right).\end{aligned} \tag{2.4}$$

If u_x is bounded from below on $[0,T)$, then there exists a positive constant k such that $u_x \geq -k$ on $[0,T)$. Thus, we get by (2.4) and Gronwall's inequality

$$\|y\|_1^2 \leq \exp\{(7k - 2\lambda)t\}\|y(0)\|_1^2.$$

The above inequality implies that the H^3-norm of the solution u of (1.3) does not blow up in finite time. □

Next, we present two blow-up criteria for (1.3) guaranteeing the occurrence of this phenomenon. Let us first prove several useful lemmas.

Consider the following differential equation:

$$\begin{cases} q_t = u(t,q), & t \in [0,T), \\ q(0,x) = x, & x \in \mathbb{R}. \end{cases} \tag{2.5}$$

The system (2.5) is essential in deriving invariance properties for solutions of the Degasperis–Procesi equation [25, 35, 51] and the Camassa–Holm equation [7, 9, 13]. It is thus natural to expect it would also be useful in the present context.

Applying classical results in the theory of ordinary differential equations, one can obtain the following two results on q which are crucial in the proof of global existence and blow-up solutions.

LEMMA 2.1 (see [51]). *Let $u_0 \in H^r(\mathbb{R})$, $r \geq 3$, and let $T > 0$ be the maximal existence time of the corresponding solution u to (1.3). Then (2.5) has a unique solution $q \in C^1([0,T) \times \mathbb{R}, \mathbb{R})$. Moreover, the map $q(t,\cdot)$ is an increasing diffeomorphism of $\mathbb{R}$ with*

$$q_x(t,x) = \exp\left(\int_0^t u_x(s,q(s,x))ds\right) > 0 \quad \forall (t,x) \in [0,T) \times \mathbb{R}.$$

LEMMA 2.2. *Let $u_0 \in H^r(\mathbb{R})$, $r \geq 3$, and let $T > 0$ be the maximal existence time of the corresponding solution u to (1.3). Setting $y := u - u_{xx}$, we have*

$$y(t,q(t,x))q_x^3(t,x) = y_0(x)\exp\{-\lambda t\} \quad \forall (t,x) \in [0,T) \times \mathbb{R}.$$

Proof. Differentiation of the system (2.5) with respect to x yields

$$\begin{cases} \frac{d}{dt}q_x = u_x(t,q)q_x, & t \in (0,T), \\ q_x(0,x) = 1, & x \in \mathbb{R}. \end{cases}$$

Let $g(t,x) = y(t,q(t,x))q_x^3(t,x)$. By Lemma 2.1, we infer from (1.3) and (2.5) that

$$\frac{d}{dt}g(t,x) = -\lambda g(t,x).$$

Integrating the above relation with respect to $t < T$ on $[0,t]$ yields the desired result. ∎

LEMMA 2.3. *If $u_0 \in H^r(\mathbb{R})$, $r > \frac{3}{2}$, then as long as the solution $u(t,x)$ given by Theorem 2.1 exists, we have*

$$\int_{\mathbb{R}} y(t,x)v(t,x)dx = \exp\{-2\lambda t\}\int_{\mathbb{R}} y_0(x)v_0(x)dx,$$

where $y(t,x) = u(t,x) - u_{xx}(t,x)$ and $v(t,x) = (4 - \partial_x^2)^{-1}u$. Moreover, we have

$$\|u(t)\|_{L^2}^2 \leq 4\exp\{-2\lambda t\}\|u_0\|_{L^2}^2.$$

Proof. Let $T > 0$ be the maximal time of existence of the solution u to (1.3) (or (2.1)) with initial data $u_0 \in H^3(\mathbb{R})$. By (1.3), we have

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}} yv\,dx &= \frac{1}{2}\int_{\mathbb{R}} y_t v\,dx + \frac{1}{2}\int_{\mathbb{R}} yv_t dx = \int_{\mathbb{R}} y_t\, v\,dx \\ &= -\int_{\mathbb{R}} vy_x u\,dx - 3\int_{\mathbb{R}} vyu_x dx - \lambda\int_{\mathbb{R}} yv\,dx \\ &= -\int_{\mathbb{R}} v(yu)_x\,dx - 2\int_{\mathbb{R}} vyu_x dx - \lambda\int_{\mathbb{R}} yv\,dx. \end{aligned}$$

Using the relations $y = u - u_{xx}$ and $4v - v_{xx} = u$, it yields

$$\begin{aligned}
\int_{\mathbb{R}} v(yu)_x\,dx &= -\int_{\mathbb{R}} v_x yu\,dx = -\int_{\mathbb{R}} v_x u^2\,dx + \int_{\mathbb{R}} v_x u u_{xx} dx \\
&= -\int_{\mathbb{R}} v_x u^2\,dx - \int_{\mathbb{R}} (v_x u)_x\, u_x\,dx \\
&= -\int_{\mathbb{R}} v_x u^2\,dx - \int_{\mathbb{R}} v_{xx} u u_x\,dx - \int_{\mathbb{R}} v_x u_x^2 dx \\
&= -\int_{\mathbb{R}} v_x u^2\,dx + \frac{1}{2}\int_{\mathbb{R}} v_{xxx} u^2\,dx - \int_{\mathbb{R}} v_x u_x^2 dx \\
&= -\int_{\mathbb{R}} v_x u^2\,dx + \frac{1}{2}\int_{\mathbb{R}} (4v_x - u_x) u^2\,dx - \int_{\mathbb{R}} v_x u_x^2 dx \\
&= \int_{\mathbb{R}} v_x u^2\,dx - \int_{\mathbb{R}} v_x u_x^2 dx.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
2\int_{\mathbb{R}} v\, y\, u_x\,dx &= 2\int_{\mathbb{R}} v u u_x\,dx - 2\int_{\mathbb{R}} v u_{xx} u_x dx \\
&= -\int_{\mathbb{R}} v_x\, u^2\,dx + \int_{\mathbb{R}} v_x u_x^2\,dx.
\end{aligned}$$

Combining the above three relations, we deduce that

$$\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}} yv\,dx = -\lambda\int_{\mathbb{R}} yv dx.$$

Consequently, this implies the first desired result. In view of the proved equality, it then follows that

$$\begin{aligned}
\|u(t)\|_{L^2}^2 = \|\hat{u}(t)\|_{L^2}^2 &\le 4\int_{\mathbb{R}} \frac{1+\xi^2}{4+\xi^2}|\hat{u}(t,\xi)|^2\,d\xi \\
&= 4(\hat{y}(t),\ \hat{v}(t)) = 4(y(t),\ v(t)) \\
&= 4\exp\{-2\lambda t\}(y_0,\ v_0) = 4\exp\{-2\lambda t\}(\hat{y}_0,\ \hat{v}_0) \\
&\le 4\exp\{-2\lambda t\}\int_{\mathbb{R}} \frac{1+\xi^2}{4+\xi^2}|\hat{u}_0(\xi)|^2\,d\xi \\
&\le 4\exp\{-2\lambda t\}\|\hat{u}_0\|_{L^2}^2 = 4\exp\{-2\lambda t\}\|u_0\|_{L^2}^2.
\end{aligned}$$

This completes the proof of the lemma. □

LEMMA 2.4. *Assume $u_0 \in H^r(\mathbb{R})$, $r > \frac{3}{2}$. Let T be the maximal existence time of the solution u to* (1.3) *guaranteed by Theorem* 2.1. *Then we have*

$$\|u(t)\|_{L^\infty} \le \exp\{-\lambda t\}\left(\frac{3}{\lambda}\|u_0\|_{L^2}^2 + \|u_0\|_{L^\infty}\right) \quad \forall t \in [0,T].$$

Proof. Let $T > 0$ be the maximal time of existence of the solution u to (1.3) with the initial data $u_0 \in H^3(\mathbb{R})$. By (2.5), we get

$$\begin{aligned}\frac{du(t,q(t,x))}{dt} &= u_t(t,q(t,x)) + u_x(t,q(t,x))\frac{dq(t,x)}{dt}\\ &= (u_t + uu_x)(t,q(t,x)).\end{aligned}$$

By (2.1), we have

$$u_t + uu_x = -3p * (uu_x) - \lambda u.$$

Note that

$$\begin{aligned}-3p * (uu_x) &= -\frac{3}{2}\int_{-\infty}^{+\infty} e^{-|x-\eta|}uu_\eta d\eta\\ &= -\frac{3}{2}\int_{-\infty}^{x} e^{-x+\eta}uu_\eta d\eta - \frac{3}{2}\int_{x}^{+\infty} e^{x-\eta}uu_\eta d\eta\\ &= \frac{3}{4}\int_{-\infty}^{x} e^{-|x-\eta|}u^2 d\eta - \frac{3}{4}\int_{x}^{+\infty} e^{-|x-\eta|}u^2 d\eta.\end{aligned}$$

It then follows that

$$-\frac{3}{4}\int_{q(t,x)}^{+\infty} e^{-|q(t,x)-\eta|}u^2 d\eta \le \frac{du(t,q(t,x))}{dt} + \lambda u(t,q(t,x)) \le \frac{3}{4}\int_{-\infty}^{q(t,x)} e^{-|q(t,x)-\eta|}u^2 d\eta.$$

It thus transpires that

$$\left|\frac{du(t,q(t,x))}{dt} + \lambda u(t,q(t,x))\right| \le \frac{3}{4}\int_{-\infty}^{+\infty} e^{-|q(t,x)-\eta|}u^2 d\eta \le \frac{3}{4}\|u(t)\|_{L^2}^2.$$

In view of Lemma 2.3, we have

$$\left|\frac{du(t,q(t,x))}{dt} + \lambda u(t,q(t,x))\right| \le 3\exp\{-2\lambda t\}\|u_0\|_{L^2}^2.$$

Integrating the above inequality with respect to $t < T$ on $[0,t]$ yields

$$|\exp\{\lambda t\}u(t,q(t,x)) - u_0(x)| \le \frac{3}{\lambda}\|u_0\|_{L^2}^2.$$

Thus,

$$\begin{aligned}|u(t,q(t,x))| &\le \|u(t,q(t,x))\|_{L^\infty}\\ &\le \exp\{-\lambda t\}\left(\frac{3}{\lambda}\|u_0\|_{L^2}^2 + \|u_0\|_{L^\infty}\right).\end{aligned} \tag{2.6}$$

Using the Sobolev embedding to ensure the uniform boundedness of $u_x(s,\eta)$ for $(s,\eta) \in [0,t]\times\mathbb{R}$ with $t \in [0,T)$, in view of Lemma 2.1, we get for every $t \in [0,T)$ a constant $C(t) > 0$ such that

$$e^{-C(t)} \le q_x(t,x) \le e^{C(t)}, \qquad x \in \mathbb{R}.$$

We deduce from the above equation that the function $q(t,\cdot)$ is strictly increasing on $\mathbb{R}$ with $\lim_{x\to\pm\infty} q(t,x) = \pm\infty$ as long as $t \in [0,T)$. Thus, by (2.6) we can obtain

$$\|u(t,x)\|_{L^\infty} = \|u(t,q(t,x))\|_{L^\infty} \le \exp\{-\lambda t\}\left(\frac{3}{\lambda}\|u_0\|_{L^2}^2 + \|u_0\|_{L^\infty}\right).$$

This completes the proof of Lemma 2.4. □

We are now in the position to present the first blow-up result.

THEOREM 2.3. *Let $u_0 \in H^r(\mathbb{R})$, $r > \frac{3}{2}$, and assume that there exists $x_0 \in \mathbb{R}$ such that*

$$u_0'(x_0) < -\frac{1}{2}\lambda - \frac{1}{2}\sqrt{\lambda^2 + 6\left(\frac{3}{\lambda}\|u_0\|_{L^2}^2 + \|u_0\|_{L^\infty}\right)^2}.$$

Then the corresponding solution of (1.3) *blows up in finite time.*

Proof. Let $T > 0$ be the existence time of the solution $u(t,\cdot)$ of (1.3) (or (2.1)) with the initial data $u_0 \in H^3(\mathbb{R})$. Differentiating (2.1) with respect to x, in view of $\partial_x^2 p * f = p * f - f$, we get

$$u_{tx} = -u_x^2 - uu_{xx} + \frac{3}{2}u^2 - p * \left(\frac{3}{2}u^2\right) - \lambda u_x. \tag{2.7}$$

Note that

$$\begin{aligned}\frac{du_x(t,q(t,x))}{dt} &= u_{xt}(t,q(t,x)) + u_{xx}(t,q(t,x))\frac{dq(t,x)}{dt}\\ &= u_{tx}(t,q(t,x)) + u(t,q(t,x))u_{xx}(t,q(t,x)).\end{aligned}$$

Thus, we have

$$\begin{aligned}\frac{du_x(t,q(t,x))}{dt} =& -u_x^2(t,q(t,x)) + \frac{3}{2}u^2(t,q(t,x))\\ &- p * \left(\frac{3}{2}u^2(t,q(t,x)\right) - \lambda u_x(t,q(t,x)).\end{aligned} \tag{2.8}$$

In view of $p * \left(\frac{3}{2}u^2\right)(t,q(t,x)) \ge 0$, we infer from (2.8) and Lemma 2.4 that

$$\begin{aligned}\frac{du_x(t,q(t,x))}{dt} &\le -u_x^2(t,q(t,x)) - \lambda u_x(t,q(t,x)) + \frac{3}{2}u^2(t,q(t,x))\\ &\le -u_x^2(t,q(t,x)) - \lambda u_x(t,q(t,x)) + \frac{3}{2}\left(\frac{3}{\lambda}\|u_0\|_{L^2}^2 + \|u_0\|_{L^\infty}\right)^2.\end{aligned}$$

Set $m(t) = u_x(t,q(t,x_0))$ and

$$\alpha^2 = \frac{3}{2}\left(\frac{3}{\lambda}\|u_0\|_{L^2}^2 + \|u_0\|_{L^\infty}\right)^2.$$

Then we obtain

$$\begin{aligned}\frac{dm(t)}{dt} &\le -m^2(t) - \lambda m(t) + \alpha^2\\ &= -\frac{1}{4}\left(2m(t) + \lambda - \sqrt{\lambda^2 + 4\alpha^2}\right)\left(2m(t) + \lambda + \sqrt{\lambda^2 + 4\alpha^2}\right).\end{aligned}$$

From the hypothesis, we have $m(0) < -\frac{1}{2}\lambda - \frac{1}{2}\sqrt{\lambda^2+4\alpha^2}$, and thus $\frac{dm}{dt}|_{t=0} < 0$. By continuity with respect to t of $m(t)$, we have $\frac{dm}{dt} < 0$ for all $t \in [0,T)$. Therefore, $m(t) < -\frac{1}{2}\lambda - \frac{1}{2}\sqrt{\lambda^2+4\alpha^2}$ for all $t \in [0,T)$. Thus, we can solve the above inequality to obtain

$$\frac{2m(0)+\lambda+\sqrt{\lambda^2+4\alpha^2}}{2m(0)+\lambda-\sqrt{\lambda^2+4\alpha^2}}\exp\{\sqrt{\lambda^2+4\alpha^2}\,t\}-1 \le \frac{2\sqrt{\lambda^2+4\alpha^2}}{2m(t)+\lambda-\sqrt{\lambda^2+4\alpha^2}} \le 0.$$

Since

$$0 < \frac{2m(0)+\lambda+\sqrt{\lambda^2+4\alpha^2}}{2m(0)+\lambda-\sqrt{\lambda^2+4\alpha^2}} < 1,$$

there exists T satisfying

$$T \le \frac{1}{\sqrt{\lambda^2+4\alpha^2}}\ln\left(\frac{2m(0)+\lambda-\sqrt{\lambda^2+4\alpha^2}}{2m(0)+\lambda+\sqrt{\lambda^2+4\alpha^2}}\right)$$

such that $\lim_{t\uparrow T} m(t) = -\infty$. Hence, the above theorem is proved according to Theorem 2.2. □

We give the second criterion that guarantees the blow-up of the solutions of (1.3).

THEOREM 2.4. *Assume that $u_0 \in H^r(\mathbb{R})$ $\left(r > \frac{3}{2}\right)$ is odd, and $u_0'(0) < -\lambda$. Then the corresponding solution of* (1.3) *blows up in finite time.*

Proof. Let $T > 0$ be the maximal time of existence of the solution u to (1.3) (or (2.1)) with initial data $u_0 \in H^3(\mathbb{R})$. As one can check, the function

$$v(t,x) := -u(t,-x), \quad t \in [0,T), \quad x \in \mathbb{R},$$

is also a solution of (1.3) in $C([0,T);H^3(\mathbb{R})\bigcap C^1([0,T);H^2(\mathbb{R})$ with initial data u_0. By uniqueness we conclude that $v \equiv u$, and therefore $u(t,\cdot)$ is odd for any $t \in [0,T)$. In particular, by continuity with respect to the spatial variable of u and u_{xx}, we get

$$u(t,0) = u_{xx}(t,0) = 0, \qquad t \in [0,T).$$

Define $h(t) := u_x(t,0)$ for $t \in [0,T)$. Note that $h \in C^1([0,T),\mathbb{R})$. From (2.7), we get

$$\begin{aligned}\frac{dh}{dt}(t) &= -h^2(t) - \lambda h(t) - p * \left(\frac{3}{2}u^2\right)\\ &\le -\left(h(t)+\lambda\right)h(t), \qquad t \in [0,T).\end{aligned}$$

From the hypothesis, we have $h(0) < -\lambda$. Therefore, $h(t) < -\lambda$ for all $t \in [0,T)$. Solving the above inequality, we get

$$1 - \frac{h(0)}{h(0)+\lambda}\exp\{-\lambda t\} \le \frac{\lambda}{h(t)+\lambda} \le 0.$$

Since

$$\frac{h(0)}{h(0)+\lambda} > 1,$$

we conclude that there exists T and

$$T \le \frac{1}{\lambda}\ln\frac{h(0)}{h(0)+\lambda}$$

such that $\lim_{t\uparrow T} h(t) = -\infty$. We complete the proof of the theorem by Theorem 2.2. □

3. Blow-up rate. In this section, we give more insight into the blow-up mechanism for the wave-breaking solutions to (1.3).

LEMMA 3.1 (see [10]). *Let $T > 0$ and $v \in C^1([0,T); H^2(\mathbb{R}))$; then for every $t \in [0,T)$ there exists at least one point $\xi(t) \in \mathbb{R}$ with*

$$m(t) := \inf_{x\in\mathbb{R}} v_x(t,x) = v_x(t,\xi(t)).$$

The function $m(t)$ is a.e. differentiable on $(0,T)$ with

$$\frac{dm}{dt} = v_{tx}(t,\xi(t)) \text{ a.e. on } (0,T).$$

THEOREM 3.1. *Let $u_0 \in H^r(\mathbb{R})$, $r > \frac{3}{2}$, and let $T > 0$ be the maximal existence time of the corresponding solution to (1.3). If T is finite, we have*

$$\lim_{t\to T}(T-t)\inf_{x\in\mathbb{R}} u_x(t,x) = -1.$$

Proof. Let $T > 0$ be the maximal time of existence of the solution u to (1.3) (or (2.1)) with initial data $u_0 \in H^3(\mathbb{R})$. We already know by Theorem 2.2 that

$$\lim_{t\to T}\inf\{\inf_{x\in\mathbb{R}} u_x(t,x)\} = -\infty. \tag{3.1}$$

We know that $u(x,t) \in C^1([0,T); H^2(\mathbb{R}))$ by Theorem 2.1; then we infer from Lemma 3.1 that, for every $t \in [0,T)$, there exists at least one point $\xi(t) \in \mathbb{R}$ with $u_x(t,\xi(t)) = \inf_{x\in\mathbb{R}} u_x(t,x)$. Let

$$m(t) = u_x(t,\xi(t)) = \inf_{x\in\mathbb{R}} u_x(t,x);$$

then $u_{xx}(t,\xi(t)) = 0$ for all $t \in [0,T)$. From (2.7) we have

$$\frac{dm}{dt} + m^2(t) + \lambda m(t) = \frac{3}{2}u^2(t,\xi(t)) - p * \left(\frac{3}{2}u^2(t,\xi(t))\right) \text{ a.e. on } (0,T). \tag{3.2}$$

By Lemmas 2.3–2.4, we have for $t \in [0,T)$ that

$$\begin{aligned}
&\left|\frac{3}{2}u^2(t,\xi(t)) - p * \left(\frac{3}{2}u^2(t,\xi(t))\right)\right| \\
\le &\left|\frac{3}{2}u^2(t,\xi(t))\right| + \left|\frac{3}{4}\int_{\mathbb{R}} e^{-|\xi(t)-\eta|}u^2(t,\eta)d\eta\right| \\
\le &\frac{3}{2}\|u\|_{L^\infty}^2 + \frac{3}{4}\|u\|_{L^2}^2 \\
\le &\frac{3}{2}\left(\frac{3}{\lambda}\|u_0\|_{L^2}^2 + \|u_0\|_{L^\infty}\right)^2 + 3\|u_0\|_{L^2}^2.
\end{aligned}$$

Set

$$\beta = \frac{3}{2}\left(\frac{3}{\lambda}\|u_0\|_{L^2}^2 + \|u_0\|_{L^\infty}\right)^2 + 3\|u_0\|_{L^2}^2.$$

We infer from (3.2) that

$$\left|\frac{dm}{dt}+m^2(t)+\lambda m(t)\right|\le\beta \text{ a.e. on } (0,T).$$

Hence,

$$-\beta-\frac{1}{4}\lambda^2\le\frac{dm}{dt}+\left(m(t)+\frac{1}{2}\lambda\right)^2\le\beta+\frac{1}{4}\lambda^2 \text{ a.e. on } (0,T). \tag{3.3}$$

Let $\varepsilon\in(0,1)$. Since $\lim_{t\to T}\inf(m(t)+\frac{1}{2}\lambda)=-\infty$ (by (3.1)), there is some $t_0\in(0,T)$ with $m(t_0)+\frac{1}{2}\lambda<0$ and

$$\left(m(t_0)+\frac{1}{2}\lambda\right)^2>\frac{1}{\varepsilon}\left(\beta+\frac{1}{4}\lambda^2\right).$$

We claim that

$$\left(m(t)+\frac{1}{2}\lambda\right)^2>\frac{1}{\varepsilon}\left(\beta+\frac{1}{4}\lambda^2\right),\qquad t\in[t_0,T). \tag{3.4}$$

In fact, since $m(t)$ is locally Lipschitz (it belongs to $W^{1,\infty}_{loc}(\mathbb{R})$ by Theorem 2.1), there is some $\delta>0$ such that

$$\left(m(t)+\frac{1}{2}\lambda\right)^2>\frac{1}{\varepsilon}\left(\beta+\frac{1}{4}\lambda^2\right),\qquad t\in(t_0,t_0+\delta).$$

From (3.3), we have

$$\frac{dm}{dt}<(\varepsilon-1)\left(m(t)+\frac{1}{2}\lambda\right)^2<0,\qquad t\in(t_0,t_0+\delta) \text{ a.e. on } (0,T).$$

Being locally Lipschitz, the function $m(t)$ is absolutely continuous. Therefore, by integrating the above relation on $[t_0,t_0+\delta]$, we obtain that $m(t_0+\delta)\le m(t_0)$. Thus,

$$m(t_0+\delta)+\frac{1}{2}\lambda\le m(t_0)+\frac{1}{2}\lambda<0.$$

By the above inequality, we have

$$\left(m(t_0+\delta)+\frac{1}{2}\lambda\right)^2\ge\left(m(t_0)+\frac{1}{2}\lambda\right)^2>\frac{1}{\varepsilon}\left(\beta+\frac{1}{4}\lambda^2\right).$$

The relation (3.4) is proved by a continuous extension.

A combination of (3.3) with (3.4) enables us to infer that

$$-1-\varepsilon<\frac{\frac{dm}{dt}}{\left(m(t)+\frac{1}{2}\lambda\right)^2}<-1+\varepsilon \text{ a.e. on } (t_0,T). \tag{3.5}$$

For $t\in(t_0,T)$, integrating (3.5) on (t,T), we obtain

$$-1-\varepsilon<\frac{1}{\left(m(t)+\frac{1}{2}\lambda\right)(T-t)}<-1+\varepsilon,\qquad t\in(t_0,T).$$

Letting ε go to zero, we obtain

$$\lim_{t\to T}\left[\left(m(t)+\frac{1}{2}\lambda\right)(T-t)\right]=-1,$$

that is,

$$\lim_{t\to T}(T-t)m(t)=-1.$$

This completes the proof of the theorem. □

Remark 3.1. Although the occurrence of blow-up of strong solutions to (1.3) is affected by the dissipative parameter (see Theorems 2.3–2.4), Theorem 3.1 shows that the blow-up rate of strong solutions to the Degasperis–Procesi equation [24] is not affected by the weakly dissipative term.

4. Global solution and its decay. In this section we will show that there exist global strong solutions to (1.3) and these global solutions decay to zero as time goes to infinity provided the initial data u_0 satisfy certain sign conditions.

THEOREM 4.1. *Assume $u_0\in H^r(\mathbb{R})$, $r>\frac{3}{2}$. If $y_0=u_0-u_{0,xx}$ does not change sign on $\mathbb{R}$, then* (1.3) *(or* (2.1)*) has a global strong solution*

$$u=u(\cdot,u_0)\in C([0,\infty);H^r(\mathbb{R}))\cap C^1([0,\infty);H^{r-1}(\mathbb{R})).$$

Moreover, the global solution decays to 0 *in the H^1-norm and the H^3-norm as time goes to infinity.*

Proof. Let $T>0$ be the maximal time of existence of the solution u to (1.3) (or (2.1)) with initial data $u_0\in H^3(\mathbb{R})$. We first consider the case where $y_0\geq 0$ on $\mathbb{R}$. If $y_0\geq 0$, then Lemmas 2.1–2.2 ensure that $y\geq 0$ for all $t\in[0,T)$. Using $u=p*y$ and the positivity of p, we infer that $u(t,\cdot)\geq 0$ for all $t\geq 0$. Note that

$$u(t,x)=\frac{e^{-x}}{2}\int_{-\infty}^{x}e^{\eta}y(t,\eta)d\eta+\frac{e^{x}}{2}\int_{x}^{\infty}e^{-\eta}y(t,\eta)d\eta$$

and

$$u_x(t,x)=-\frac{e^{-x}}{2}\int_{-\infty}^{x}e^{\eta}y(\eta)d\eta+\frac{e^{x}}{2}\int_{x}^{\infty}e^{-\eta}y(\eta)d\eta.$$

From the above two equations, we deduce that

$$\begin{aligned}u(t,x)+u_x(t,x)&=e^{x}\int_{x}^{\infty}e^{-\eta}y(t,\eta)d\eta,\\ u(t,x)-u_x(t,x)&=e^{-x}\int_{-\infty}^{x}e^{\eta}y(t,\eta)d\eta.\end{aligned}\tag{4.1}$$

By (4.1) and $y\geq 0$ for all $t\in[0,T)$, we obtain, for $t\in[0,T)$,

$$|u_x(t,x)|\leq u(t,x)\quad\forall(t,x)\in[0,T)\times\mathbb{R}.\tag{4.2}$$

By Lemma 2.4, we have

$$\begin{aligned}|u_x(t,x)|&\leq u(t,x)\\ &\leq\exp\{-\lambda t\}\left(\frac{3}{\lambda}\|u_0\|_{L^2}^2+\|u_0\|_{L^\infty}\right)\quad\forall(t,x)\in[0,T)\times\mathbb{R}.\end{aligned}\tag{4.3}$$

The above inequality and Theorem 2.2 imply that $T = \infty$. This proves that the solution u exists globally in time.

Multiplying (1.3) by u and integrating by parts, in view of (4.2) and Lemmas 2.3–2.4, we get

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}}\left(u^2+u_x^2\right)dx+\lambda\int_{\mathbb{R}}\left(u^2+u_x^2\right)dx\\
=&-4\int_{\mathbb{R}}u^2u_xdx+3\int_{\mathbb{R}}uu_xu_{xx}dx+\int_{\mathbb{R}}u^2u_{xxx}dx\\
=&-\frac{1}{2}\int_{\mathbb{R}}u_x^3dx\le\frac{1}{2}\int_{\mathbb{R}}u^3dx\le\frac{1}{2}\|u\|_{L^\infty}\int_{\mathbb{R}}u^2dx\\
\le&2\exp\{-3\lambda t\}\left(\frac{3}{\lambda}\|u_0\|_{L^2}^2+\|u_0\|_{L^\infty}\right)\|u_0\|_{L^2}^2.
\end{aligned}$$

Integrating the above inequality with respect to t, we have

$$\|u(t)\|_1^2\le\exp\{-2\lambda t\}\left(\frac{12}{\lambda^2}\|u_0\|_{L^2}^4+\frac{4}{\lambda}\|u_0\|_{L^2}^2\|u_0\|_{L^\infty}+\|u_0\|_1^2\right).$$

This shows that the corresponding global solution with $y_0\ge 0$ decays to 0 in the H^1-norm.

By (4.3), we obtain that $-u_x\le\frac{\lambda}{7}$ for sufficiently large t. This yields, in combination with (2.4) and Gronwall's inequality,

$$\|y\|_1^2\le e^{-\lambda t}\|y_0\|_1^2$$

for large t. The above inequality implies that the corresponding global solution with $y_0\ge 0$ decays to 0 in the H^3-norm. This completes the proof of the theorem with the assumption $y_0\ge 0$ on $\mathbb{R}$. In the case when $y_0(x)\le 0$ on $\mathbb{R}$, one can repeat the above proof to get the desired result. □

Remark 4.1. Note that the global solution to the Degasperis–Procesi equation does not generally decay to zero as time goes to infinity [35]. Theorem 4.1 shows that there is a considerable difference between (1.3) and the Degasperis–Procesi equation in their long time behaviors. More precisely, the energy dissipation will affect the long time behavior of global solutions to the Degasperis–Procesi equation.

Remark 4.2. It is well known that the Degasperis–Procesi equation has peaked traveling wave solutions. Theorem 4.1 shows that global H^3-solutions with y_0 of one sign decay in the H^1-norm and the H^3-norm. Lemma 2.4 shows that any global solution decays in the L^∞-norm. This means that there are no traveling wave solutions of the dissipative equation (1.3). This is also another considerable difference between (1.3) and the Degasperis–Procesi equation in their long time behaviors.

Acknowledgment. The authors thank the referees for their valuable comments and suggestions.

REFERENCES

[1] R. Beals, D. Sattinger, and J. Szmigielski, *Acoustic scattering and the extended Korteweg-de Vries hierarchy*, Adv. Math., 140 (1998), pp. 190–206.

[2] A. Bressan and A. Constantin, *Global conservative solutions of the Camassa-Holm equation*, Arch. Ration. Mech. Anal., 183 (2007), pp. 215–239.
[3] R. Camassa and D. Holm, *An integrable shallow water equation with peaked solitons*, Phys. Rev. Lett., 71 (1993), pp. 1661–1664.
[4] R. Camassa, D. Holm, and J. Hyman, *A new integrable shallow water equation*, Adv. in Appl. Mech., 31 (1994), pp. 1–33.
[5] G. M. Coclite and K. H. Karlsen, *On the well-posedness of the Degasperis-Procesi equation*, J. Funct. Anal., 233 (2006), pp. 60–91.
[6] G. M. Coclite, K. H. Karlsen, and N. H. Risebro, *Numerical schemes for computing discontinuous solutions of the Degasperis-Procesi equation*, IMA J. Numer. Anal., 28 (2008), pp. 80–105.
[7] A. Constantin, *Global existence of solutions and breaking waves for a shallow water equation: A geometric approach*, Ann. Inst. Fourier (Grenoble), 50 (2000), pp. 321–362.
[8] A. Constantin, *On the scattering problem for the Camassa-Holm equation*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 457 (2001), pp. 953–970.
[9] A. Constantin and J. Escher, *Global existence and blow-up for a shallow water equation*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 26 (1998), pp. 303–328.
[10] A. Constantin and J. Escher, *Wave breaking for nonlinear nonlocal shallow water equations*, Acta Math., 181 (1998), pp. 229–243.
[11] A. Constantin and J. Escher, *Global weak solutions for a shallow water equation*, Indiana Univ. Math. J., 47 (1998), pp. 1527–1545.
[12] A. Constantin and B. Kolev, *Geodesic flow on the diffeomorphism group of the circle*, Comment. Math. Helv., 78 (2003), pp. 787–804.
[13] A. Constantin and H. P. McKean, *A shallow water equation on the circle*, Comm. Pure Appl. Math., 52 (1999), pp. 949–982.
[14] A. Constantin and L. Molinet, *Global weak solutions for a shallow water equation*, Comm. Math. Phys., 211 (2000), pp. 45–61.
[15] A. Constantin and W. Strauss, *Stability of a class of solitary waves in compressible elastic rods*, Phys. Lett. A, 270 (2000), pp. 140–148.
[16] A. Constantin and W. A. Strauss, *Stability of peakons*, Comm. Pure Appl. Math., 53 (2000), pp. 603–610.
[17] A. Constantin and W. A. Strauss, *Stability of the Camassa-Holm solitons*, J. Nonlinear Sci., 12 (2002), pp. 415–422.
[18] H. H. Dai, *Model equations for nonlinear dispersive waves in a compressible Mooney-Rivlin rod*, Acta Mech., 127 (1998), pp. 193–207.
[19] A. Degasperis, D. D. Holm, and A. N. W. Hone, *A new integral equation with peakon solutions*, Theoret. and Math. Phys., 133 (2002), pp. 1463–1474.
[20] A. Degasperis and M. Procesi, *Asymptotic integrability*, in Symmetry and Perturbation Theory, A. Degasperis and G. Gaeta, eds., World Scientific, River Edge, NJ, 1999, pp. 23–37.
[21] P. G. Drazin and R. S. Johnson, *Solitons: An Introduction*, Cambridge University Press, Cambridge, UK, New York, 1989.
[22] H. R. Dullin, G. A. Gottwald, and D. D. Holm, *An integrable shallow water equation with linear and nonlinear dispersion*, Phys. Rev. Lett., 87 (2001), pp. 4501–4504.
[23] H. R. Dullin, G. A. Gottwald, and D. D. Holm, *Camassa-Holm, Korteweg-de Vries-5 and other asymptotically equivalent equations for shallow water waves*, Fluid Dynam. Res., 33 (2003), pp. 73–95.
[24] J. Escher, Y. Liu, and Z. Yin, *Global weak solutions and blow-up structure for the Degasperis-Procesi equation*, J. Funct. Anal., 241 (2006), pp. 457–485.
[25] J. Escher, Y. Liu, and Z. Yin, *Shock waves and blow-up phenomena for the periodic Degasperis-Procesi equation*, Indiana Univ. Math. J., 56 (2007), pp. 87–117.
[26] J. Escher and Z. Yin, *On the initial boundary value problems for the Degasperis-Procesi equation*, Phys. Lett. A, 368 (2007), pp. 69–76.
[27] B. Fuchssteiner and A. Fokas, *Symplectic structures, their Bäcklund transformation and hereditary symmetries*, Phys. D, 4 (1981), pp. 47–66.
[28] J. M. Ghidaglia, *Weakly damped forced Korteweg-de Vries equations behave as a finite dimensional dynamical system in the long time*, J. Differential Equations, 74 (1988), pp. 369–390.
[29] R. I. Ivanov, *Water waves and integrability*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 365 (2007), pp. 2267–2280.
[30] R. S. Johnson, *Camassa-Holm, Korteweg-de Vries and related models for water waves*, J. Fluid Mech., 455 (2002), pp. 63–82.
[31] C. Kenig, G. Ponce, and L. Vega, *Well-posedness and scattering results for the generalized*

Korteweg-de Vries equation via the contraction principle, Comm. Pure Appl. Math., 46 (1993), pp. 527–620.

[32] J. Lenells, *Traveling wave solutions of the Degasperis-Procesi equation*, J. Math. Anal. Appl., 306 (2005), pp. 72–82.

[33] J. Lenells, *Conservation laws of the Camassa-Holm equation*, J. Phys. A, 38 (2005), pp. 869–880.

[34] Y. Li and P. Olver, *Well-posedness and blow-up solutions for an integrable nonlinearly dispersive model wave equation*, J. Differential Equations, 162 (2000), pp. 27–63.

[35] Y. Liu and Z. Yin, *Global existence and blow-up phenomena for the Degasperis-Procesi equation*, Comm. Math. Phys., 267 (2006), pp. 801–820.

[36] H. Lundmark, *Formation and dynamics of shock waves in the Degasperis-Procesi equation*, J. Nonlinear Sci., 17 (2007), pp. 169–198.

[37] H. Lundmark and J. Szmigielski, *Multi-peakon solutions of the Degasperis-Procesi equation*, Inverse Problems, 19 (2003), pp. 1241–1245.

[38] Y. Matsuno, *Multisoliton solutions of the Degasperis-Procesi equation and their peakon limit*, Inverse Problems, 21 (2005), pp. 1553–1570.

[39] H. P. Mckean, *Integrable systems and algebraic curves*, in Global Analysis, Lecture Notes in Math. 755, Springer, Berlin, 1979, pp. 83–200.

[40] G. Misiolek, *A shallow water equation as a geodesic flow on the Bott-Virasoro group*, J. Geom. Phys., 24 (1998), pp. 203–208.

[41] E. Ott and R. N. Sudan, *Damping of solitary waves*, Phys. Fluids, 13 (1970), pp. 1432–1434.

[42] G. Rodriguez-Blanco, *On the Cauchy problem for the Camassa-Holm equation*, Nonlinear Anal., 46 (2001), pp. 309–327.

[43] T. Tao, *Low-regularity global solutions to nonlinear dispersive equations*, in Surveys in Analysis and Operator Theory (Canberra, 2001), Proc. Centre Math. Appl. Austral. Nat. Univ. 40, Australian National University, Canberra, Australia, 2002, pp. 19–48.

[44] V. O. Vakhnenko and E. J. Parkes, *Periodic and solitary-wave solutions of the Degasperis-Procesi equation*, Chaos Solitons Fractals, 20 (2004), pp. 1059–1073.

[45] G. B. Whitham, *Linear and Nonlinear Waves*, John Wiley and Sons, New York, 1980.

[46] S. Wu and Z. Yin, *Blow-up, blow-up rate and decay of the solution of the weakly dissipative Camassa-Holm equation*, J. Math. Phys., 47 (2006), 013504.

[47] Z. Xin and P. Zhang, *On the weak solutions to a shallow water equation*, Comm. Pure Appl. Math., 53 (2000), pp. 1411–1433.

[48] Z. Yin, *On the Cauchy problem for an integrable equation with peakon solutions*, Illinois J. Math., 47 (2003), pp. 649–666.

[49] Z. Yin, *Global existence for a new periodic integrable equation*, J. Math. Anal. Appl., 283 (2003), pp. 129–139.

[50] Z. Yin, *Global weak solutions to a new periodic integrable equation with peakon solutions*, J. Funct. Anal., 212 (2004), pp. 182–194.

[51] Z. Yin, *Global solutions to a new integrable equation with peakons*, Indiana Univ. Math. J., 53 (2004), pp. 1189–1210.

SIAM J. MATH. ANAL.
Vol. 40, No. 2, pp. 491–515

NONLINEAR STABILITY OF STATIONARY SOLUTIONS FOR SURFACE DIFFUSION WITH BOUNDARY CONDITIONS*

HARALD GARCKE†, KAZUO ITO‡, AND YOSHIHITO KOHSAKA§

Abstract. The volume-preserving fourth order surface diffusion flow has constant mean curvature hypersurfaces as stationary solutions. We show nonlinear stability of certain stationary curves in the plane which meet an exterior boundary with a prescribed contact angle. Methods include semigroup theory, energy arguments, geometric analysis, and variational calculus.

Key words. surface diffusion, nonlinear stability, energy method, variational calculus

AMS subject classifications. 35B35, 35G30, 35K55, 35R35, 53C44

DOI. 10.1137/070694752

1. Introduction.

The surface diffusion flow

$$V = -\Delta_S \kappa \tag{1.1}$$

is a geometrical evolution law which describes the surface dynamics for phase interfaces, when mass diffusion occurs only within the interface. Here V is the normal velocity of the evolving surface, Δ_S is the surface Laplacian, and κ is the mean curvature of the surface. The flow (1.1) was first proposed by Mullins [20] in works concerned with thermal grooving. A derivation of (1.1) within rational thermodynamics was given by Davi and Gurtin [7]. In [22], Cahn and Taylor showed that (1.1) is the H^{-1}-gradient flow of the area functional, and in [5], Cahn, Elliott, and Novick-Cohen used formal asymptotics to derive (1.1) as the sharp interface limit of the Cahn–Hilliard equation with degenerate mobility. Further, the motion given by (1.1) has the significant geometrical properties that for closed embedded hypersurfaces the enclosed volume is preserved and surface area decreases in time (see, e.g., [10, 12]). The evolution law (1.1) leads to a fourth order parabolic equation which is in contrast to the second order mean curvature flow $V = \kappa$. We remark that the mean curvature flow is also area decreasing but changes the enclosed volume.

In this paper we study the motion by surface diffusion for curves in cases where the interface intersects an external boundary. More precisely, we consider the following problem. Let Ω be an open bounded domain in $\mathbb{R}^2$. We look for evolving curves $\Gamma = \{\Gamma_t\}_{t\geq 0}$ lying in Ω with $\partial\Gamma \subset \partial\Omega$ and satisfying

$$V = -\kappa_{ss} \tag{1.2}$$

for all points on Γ_t with the boundary conditions

$$\begin{cases} \Gamma_t \perp \partial\Omega & (90^\circ\text{-angle condition}), \\ \kappa_s = 0 & (\text{no-flux condition}) \end{cases} \tag{1.3}$$

*Received by the editors June 18, 2007; accepted for publication (in revised form) February 4, 2008; published electronically May 28, 2008. This work was supported by the Regensburger Universitätsstiftung Hans Vielberth and the Research Fellowship of the Japan Society for the Promotion of Young Scientists.

http://www.siam.org/journals/sima/40-2/69475.html

†NWFI-Mathematik, Universitaet Regensburg, 93040 Regensburg, Germany (harald.garcke@mathematik.uni-regensburg.de).

‡Advanced Algorithm & Systems, Ebisu IS Building, Tokyo 150-0013, Japan (k.ito@aas-ri.co.jp).

§Muroran Institute of Technology, Muroran 050-8585, Japan (kohsaka@mmm.muroran-it.ac.jp).

at $\Gamma_t \cap \partial\Omega$, where a subscript s denotes differentiation with respect to the arc-length parameter of the evolving curve Γ_t. The boundary conditions (1.3) are the natural boundary conditions when viewing the flow as the H^{-1}-gradient flow of the length functional. It is not difficult to show that under the surface diffusion flow (1.2) with the boundary conditions (1.3) the areas enclosed by Γ_t and $\partial\Omega$ are preserved and the length of Γ_t decreases in time. We also find that an arc of a circle or a line segment is stationary under (1.2) and (1.3). Our goal in this paper is to show a nonlinear stability result for stationary solutions to (1.2) and (1.3). A proof of such a result is difficult due to the area-preserving property and due to the fact that highly nonlinear boundary conditions appear. We remark that for nonlinear boundary conditions satisfactory stability results are not available within the context of semigroup theory. We also remark that it is not possible to use methods based on maximum or comparison principles which have been used for mean curvature flow; see [8, 9].

For closed curves evolving by surface diffusion, Elliott and Garcke [10] showed a global existence result in the case that the initial curve is close to a circle. In addition, they proved nonlinear stability of circles under surface diffusion. Escher, Mayer, and Simonett [12] generalized the result in [10] to the higher-dimensional case. For evolving curves which come into contact with the outer boundary, Garcke, Ito, and Kohsaka [13] studied the linearized stability of stationary curves for (1.2) and (1.3). They derived a linearized stability criterion by extending the work for mean curvature flow of [8, 9, 15] to motion by surface diffusion. For three evolving curves with a triple junction in the case that the outer boundary $\partial\Omega$ is a rectangle [11, 16] or a triangle [17], global existence results when the initial curve is a small perturbation of a certain stationary curve have been shown. Also, nonlinear stability of this stationary curve can be shown.

Since the proof of nonlinear stability will depend heavily on the linear stability criterion derived in [13], we will now state it in detail. Let Γ^* be a stationary curve parameterized by X^* such that

$$\Gamma^* = \{X^*(\sigma) \mid \sigma \in [l_-, l_+]\},$$

where σ is the arc-length parameter along Γ^* and $X^*(l_\pm) \in \partial\Omega$. Further, we denote by κ^* the curvature of Γ^* and by $h^*_\pm$ the curvature of $\partial\Omega$ at $X^*(l_\pm)$, where we assume the sign convention that $h^*_\pm$ is negative if Ω is convex. Then the linearized stability criterion requires that

$$I^*[w,w] = \int_{l_-}^{l_+} \left\{ w_\sigma^2 - (\kappa^*)^2 w^2 \right\} d\sigma + h^*_+(w^2|_{\sigma=l_+}) + h^*_-(w^2|_{\sigma=l_-}) \tag{1.4}$$

is positive for all $w \in H^1(\Gamma^*)$ with mean value zero. In [13] this criterion was derived by studying the stability of the linearized problem. The same bilinear form also appears if one computes the second variation of the length functional taking boundary contacts into account; see, e.g., Vogel [23]. We refer the reader to section 7 of [13] for several examples in which the linearized stability criterion has been applied. In the papers [2, 3] numerical results on the stability of stationary solutions for surface diffusion are presented.

Our methods to obtain a nonlinear stability result are the following. First we introduce new curvilinear coordinates in order to derive an appropriate parameterization for which we can formulate (1.2) and (1.3) in a PDE setting. We then prove a local existence result, where the local existence time depends only on the $C^{2+\alpha}$-norm

$(0 < \alpha < 1)$ of the initial curve. This is very helpful for a global existence result because we need a priori estimates only up to two spatial derivatives. In fact, by applying an energy method as in [6, 10, 16, 17] to a resulting evolution equation for the curvature, we can derive an a priori estimate of the L^2-norm of κ_s, which implies the boundedness of the $C^{2+\alpha}$-norm $(0 < \alpha < 1/2)$ of the solution for $t > 0$. In the derivation of this a priori estimate, the linearized stability criterion developed in [13] is used. In addition, we need to understand the set of stationary solutions. We can use a result by Vogel [23] which guarantees that linearly stable stationary solutions are strict local minimizers of the length functional under an area constraint. We also show that in the neighborhood of the linearly stable stationary solution other stationary solutions can be represented as a one parameter family, where the parameter is the enclosed area. This implies that the linearly stable stationary solution is isolated, a fact which will be important in order to study the long time behavior of solutions.

This paper proceeds as follows. In section 2, a parameterization established in [13] is employed for the geometric evolution equation (1.2) with boundary conditions (1.3). As a consequence, we obtain a nonlinear fourth order parabolic PDE with nonlinear boundary conditions. We show a local existence result for this nonlinear parabolic problem. For the reader's convenience we show an essential part of the proof of the local existence result in Appendix A. In section 3, an evolution equation for the curvature is derived together with some geometric identities. The evolution equation for the curvature allows it to apply an energy method as in [6, 10, 16, 17]. In section 4, we first derive a priori estimates for the length of Γ_t and the L^2-norm of κ_s when Γ_t is close to a linearly stable stationary curve. These estimates imply the boundedness of the $C^{2+\alpha}$-norm $(0 < \alpha < 1/2)$ of the solution for $t > 0$, so that the global existence result is proven when the initial curve is close to a linearly stable stationary curve. Finally, in section 5, we show nonlinear stability of linearly stable stationary curves.

2. Local existence and uniqueness. In order to derive local existence and uniqueness for the geometric evolution equation (1.2) with the boundary conditions (1.3), we employ a parameterization which was established in [13]. We remark that our parameterization describes the curves close to a stationary curve by a modified distance function over a fixed interval and leads to a single PDE in contrast to parameterizations used, e.g., in Bronsard and Reitich [4] which involve vector-valued functions and lead to a system of PDEs. We also want to avoid arc-length parameterizations, as they lead to time-dependent parameter intervals, which is less convenient for semigroup theory. For the reader's convenience, we give a detailed derivation of our parameterization in the following.

Let $\Omega \subset \mathbb{R}^2$ be a domain such that

$$\Omega = \{x \in \mathbb{R}^2 \mid \psi(x) < 0\}, \quad \partial\Omega = \{x \in \mathbb{R}^2 \mid \psi(x) = 0\}$$

with a smooth function $\psi : \mathbb{R}^2 \to \mathbb{R}$ fulfilling $\nabla\psi(x) \neq 0$ for x with $\psi(x) = 0$.

Also, let Γ^* be a stationary curve under the flow (1.2) and (1.3); i.e., Γ^* has constant curvature κ^*. We now introduce an arc-length parameterization of Γ^* in the form

$$\Gamma^* = \{\Phi^*(\sigma) \mid \sigma \in [l_-, l_+]\},$$

where Φ^* is a mapping from $[l_-, l_+]$ to $\mathbb{R}^2$ and $l_+ - l_-$ is the total length of Γ^*. Note that we can extend Γ^* naturally either to the full circle when Γ^* is an arc of a circle

or to a straight line when Γ^* is a line segment. We set

$$\bar{l} := \begin{cases} \pi/|\kappa^*| & \text{if } \kappa^* \neq 0, \\ +\infty & \text{if } \kappa^* = 0; \end{cases}$$

i.e., $\bar{l}$ is the length of the extension of Γ^* to a half circle if $\kappa_* \neq 0$. Without loss of generality, we can assume $[l_-, l_+] \subset (-\bar{l}, \bar{l})$. Define

$$\begin{cases} \xi_+(q) := \max\{\sigma \in (-\bar{l}, \bar{l}) \mid \Phi^*(\sigma) + qN^*(\sigma) \in \Omega\}, \\ \xi_-(q) := \min\{\sigma \in (-\bar{l}, \bar{l}) \mid \Phi^*(\sigma) + qN^*(\sigma) \in \Omega\}, \end{cases}$$

where $N^*(\sigma)$ is a unit normal vector of Γ^* at σ and is obtained by rotating the unit tangent vector $T^*(\sigma)$ of Γ^* by $\pi/2$. Above, q is a parameter with $q \in (-\bar{d}, \bar{d})$ for a small and given $\bar{d} > 0$. It holds that $\psi(\Phi^*(\xi_\pm(q)) + qN^*(\xi_\pm(q))) = 0$ and $\xi_\pm(0) = l_\pm$. Using the implicit function theorem, we see that $\xi_+(q)$ and $\xi_-(q)$ are smooth. Let

$$\Psi(\sigma, q) := \Phi^*(\xi(\sigma, q)) + qN^*(\xi(\sigma, q))$$

with

$$\xi(\sigma, q) := \xi_-(q) + \frac{\sigma - l_-}{l_+ - l_-}(\xi_+(q) - \xi_-(q)).$$

It is not difficult to check that $\xi(l_\pm, q) = \xi_\pm(q)$ and $\xi(\sigma, 0) = \sigma$.

In addition, one derives that $\Psi : (l_-, l_+) \times (-\bar{d}, \bar{d}) \to \Omega$ parameterizes the intersection W of a tubular neighborhood around the extended Γ^* with Ω. We now consider functions $\rho : [l_-, l_+] \to (-\bar{d}, \bar{d})$ and obtain $\Psi(\sigma, \rho(\sigma)) \in W$ for $\sigma \in (l_-, l_+)$. Then we define $\Phi(\sigma) := \Psi(\sigma, \rho(\sigma))$ for $\sigma \in [l_-, l_+]$, which is a parameterization of a curve Γ. An evolving curve is now given by

$$\Gamma_t := \{\Phi(\sigma, t) \mid \sigma \in [l_-, l_+]\} \tag{2.1}$$

with $\Phi(\sigma, t) := \Psi(\sigma, \rho(\sigma, t))$ for a function $\rho = \rho(\sigma, t)$. We note that $|\rho(\sigma, t)| < \bar{d}$ guarantees that $\Phi(\sigma, t) = \Psi(\sigma, \rho(\sigma, t)) \in W$ for $\sigma \in (l_-, l_+)$ and $t > 0$. We remark that $\rho \equiv 0$ corresponds to the stationary curve Γ^*.

Let us now express (1.2) and (1.3) with the help of parameterizations which have the form (2.1). For the arc-length parameter s of Γ_t, we have

$$\frac{ds}{d\sigma} = |\Phi_\sigma| = \sqrt{|\Psi_\sigma|^2 + 2(\Psi_\sigma, \Psi_q)_{\mathbb{R}^2}\rho_\sigma + |\Psi_q|^2\rho_\sigma^2} \ =: J(\rho). \tag{2.2}$$

By $|\cdot|$ and $(\cdot, \cdot)_{\mathbb{R}^2}$ we denote the norm and the inner product in $\mathbb{R}^2$, respectively. Then we find

$$T = \frac{1}{J(\rho)}\Phi_\sigma, \quad N = \frac{1}{J(\rho)}R\Phi_\sigma,$$

where T and N are the unit tangent and unit normal to Γ_t, respectively, and R is the rotation by the angle $\pi/2$. The normal velocity V of Γ_t is given by

$$V = (\Phi_t, N)_{\mathbb{R}^2} = \frac{1}{J(\rho)}(\Phi_t, R\Phi_\sigma)_{\mathbb{R}^2} = \frac{1}{J(\rho)}(\Psi_q, R\Psi_\sigma)_{\mathbb{R}^2}\rho_t.$$

Further, the Laplace–Beltrami operator $\Delta(\rho)$ on Γ_t is given via (2.2) as

$$\Delta(\rho) = \partial_s^2 = \frac{1}{J(\rho)}\partial_\sigma\left(\frac{1}{J(\rho)}\partial_\sigma\right) = \frac{1}{(J(\rho))^2}\partial_\sigma^2 + \frac{1}{J(\rho)}\left(\partial_\sigma\frac{1}{J(\rho)}\right)\partial_\sigma. \tag{2.3}$$

Then the curvature κ of Γ_t can be derived by using $\Delta(\rho)$ as

$$\begin{aligned}\kappa(\rho) &= (\Delta(\rho)\Phi, N)_{\mathbb{R}^2} = \frac{1}{(J(\rho))^3}(\Phi_{\sigma\sigma}, R\Phi_\sigma)_{\mathbb{R}^2} \\ &= \frac{1}{(J(\rho))^3}\Big[(\Psi_q, R\Psi_\sigma)_{\mathbb{R}^2}\rho_{\sigma\sigma} + \big\{2(\Psi_{\sigma q}, R\Psi_\sigma)_{\mathbb{R}^2} + (\Psi_{\sigma\sigma}, R\Psi_q)_{\mathbb{R}^2}\big\}\rho_\sigma \\ &\quad + \big\{(\Psi_{qq}, R\Psi_\sigma)_{\mathbb{R}^2} + 2(\Psi_{\sigma q}, R\Psi_q)_{\mathbb{R}^2} + (\Psi_{qq}, R\Psi_q)_{\mathbb{R}^2}\rho_\sigma\big\}\rho_\sigma^2 \\ &\quad + (\Psi_{\sigma\sigma}, R\Psi_\sigma)_{\mathbb{R}^2}\Big].\end{aligned} \tag{2.4}$$

Furthermore, we note that the Neumann boundary condition $(\Phi_\sigma, T_{\partial\Omega})_{\mathbb{R}^2} = 0$ on $\partial\Omega$ is equivalent to the condition $(R\Phi_\sigma, \nabla\psi(\Phi))_{\mathbb{R}^2} = 0$ on $\partial\Omega$. Then we compute that the parameterization of the Neumann boundary condition is

$$(R\Psi_\sigma + R\Psi_q\rho_\sigma, \nabla\psi(\Psi))_{\mathbb{R}^2} = 0 \ \text{ at } \ \sigma = l_\pm.$$

As a consequence, we conclude that the problem (1.2) and (1.3) is represented by

$$\begin{cases} \rho_t = -L(\rho)\Delta(\rho)\kappa(\rho) & \text{for} \quad \sigma \in (l_-, l_+),\ t > 0, \\ (R\Psi_\sigma + R\Psi_q\rho_\sigma, \nabla\psi(\Psi))_{\mathbb{R}^2} = 0 & \text{at} \quad \sigma = l_\pm, \\ \partial_\sigma\kappa(\rho) = 0 & \text{at} \quad \sigma = l_\pm. \end{cases} \tag{2.5}$$

Here $L(\rho) := J(\rho)/(\Psi_q, R\Psi_\sigma)_{\mathbb{R}^2}$; $\Delta(\rho)$ and $\kappa(\rho)$ are given by (2.3) and (2.4), respectively.

Let $\mathcal{I} = [l_-, l_+]$ and $\mathcal{Q}_{t_0,t_1} = \mathcal{I}\times(t_0, t_1]$ for $0 \le t_0 < t_1 < \infty$. For $0 < \alpha < 1$, we define the function space

$$\mathcal{Y}(\overline{\mathcal{Q}_{t_0,t_1}}) = \{\rho \in C^{2+\alpha,0}(\overline{\mathcal{Q}_{t_0,t_1}}) \cap C^{4+\alpha,1}(\mathcal{Q}_{t_0,t_1}) \mid \|\rho\|_{\mathcal{Y}(\overline{\mathcal{Q}_{t_0,t_1}})} < \infty\}$$

with the norm

$$\begin{aligned}\|\rho\|_{\mathcal{Y}(\overline{\mathcal{Q}_{t_0,t_1}})} &= \sup_{t_0\le t\le t_1}\|\rho(\cdot,t)\|_{C^{2+\alpha}(\mathcal{I})} + \sup_{t_0<t\le t_1}(t-t_0)^{1/2}\|\partial_\sigma^4\rho(\cdot,t)\|_{C^\alpha(\mathcal{I})} \\ &\quad + \sup_{t_0<t\le t_1}(t-t_0)^{1/2}\|\rho_t(\cdot,t)\|_{C^\alpha(\mathcal{I})},\end{aligned}$$

where $\overline{\mathcal{Q}_{t_0,t_1}}$ is the closure of $\mathcal{Q}_{t_0,t_1}$.

Now we are ready to state a local existence theorem.

THEOREM 2.1 (local existence). *Let $\alpha \in (0,1)$ and let us assume that $\rho_0 \in C^{2+\alpha}(\mathcal{I})$ with $\|\rho_0\|_{C^0(\mathcal{I})} < \bar{d}$ fulfills*

$$(R\Psi_\sigma + R\Psi_q\rho_\sigma, \nabla\psi(\Psi))_{\mathbb{R}^2} = 0 \ \text{ at } \ \sigma = l_\pm.$$

Then there exists a $T_0 = T_0(1/\|\rho_0\|_{C^{2+\alpha}(\mathcal{I})}) > 0$ such that the problem (2.5) with $\rho(\cdot,0) = \rho_0$ has a unique solution in $\mathcal{Y}(\overline{\mathcal{Q}_{0,T_0}})$.

This theorem is proved by applying similar arguments as in [16]. Since we have to take care of the boundary conditions in a different way, we will sketch the proof in Appendix A.

Remark 2.2. Applying the local existence results used in [4] directly, we obtain solutions for $C^{4+\alpha}$-initial curves. But this makes it difficult to derive a global existence result because we need a priori estimates for higher order derivatives. Thus our local existence result is an improvement of the one obtained in [4].

Remark 2.3. By using a bootstrapping argument as in [16, Theorem 3.6, Remark 3.7], it can be shown that the solution ρ established in Theorem 2.1 is smooth for $t \in (0, T_0]$.

3. An evolution equation for curvature. In order to show nonlinear stability of solutions for which the linearized stability criterion of [13] is fulfilled, we apply an energy method similar to the one used in [6, 10, 16, 17]. For this approach it is important to derive an evolution equation for the curvature. Such an equation will be useful for the derivation of a priori estimates with the help of the linearized stability criterion.

For the above-mentioned purpose, we employ a parameterization of the evolving curve Γ_t by arc length contrary to the one stated in section 2. Let X be a smooth mapping so that $X(\cdot, t)$ is an arc-length parameterization of Γ_t, i.e.,

$$\Gamma_t := \{X(s,t) \,|\, s \in [r_-(t), r_+(t)]\}$$

for any $t > 0$, where r_+ and r_- are smooth in t. In particular, $X(r_\pm(t), t) \in \partial\Omega$ and $r_+(t) - r_-(t) = L[\Gamma_t]$, where $L[\Gamma_t]$ denotes the total length of Γ_t. Let $N\,(= N(s,t))$ be the unit normal vector of Γ_t, which is represented as

$$N(s,t) = \begin{pmatrix} \cos\theta(s,t) \\ \sin\theta(s,t) \end{pmatrix}.$$

Also, let $T\,(= T(s,t))$ and $\kappa\,(= \kappa(s,t))$ be the unit tangent vector of Γ_t and the curvature of Γ_t, respectively. Note that the unit tangent vector T is obtained by rotating the unit normal vector N by $-\pi/2$. Then, using $\theta_s = \kappa$, we have

$$\begin{cases} N_s = -\kappa T, \quad T_s = \kappa N, \\ N_t = -\theta_t T, \quad T_t = \theta_t N. \end{cases} \tag{3.1}$$

In addition, set

$$V := (X_t, N)_{\mathbb{R}^2}, \quad v := (X_t, T)_{\mathbb{R}^2}.$$

Note that V and v are the normal velocity and the tangent velocity of X, respectively. Then it follows that

$$X_t = VN + vT. \tag{3.2}$$

Differentiating (3.2) with respect to s and using (3.1), we have

$$\begin{aligned} X_{ts} &= V_s N + V N_s + v_s T + v T_s \\ &= (V_s + \kappa v) N + (-\kappa V + v_s) T. \end{aligned}$$

This implies the following lemma.

LEMMA 3.1. *Let X be a smooth arc-length parameterization as above. Then*

$$\theta_t = V_s + \kappa v, \quad v_s = \kappa V.$$

Proof. Since $X_{ts} = X_{st}$ and $X_s = T$, it follows from (3.1) that

$$\theta_t N = (V_s + \kappa v)N + (-\kappa V + v_s)T.$$

Thus we obtain the desired results. ∎

By Lemma 3.1, we have the following formula for the time derivative of curvature.

LEMMA 3.2. *Let X be a smooth arc-length parameterization as above. Then*

$$\kappa_t = V_{ss} + \kappa^2 V + \kappa_s v.$$

Proof. By $\theta_s = \kappa$ and Lemma 3.1, we derive

$$\kappa_t = \theta_{st} = \theta_{ts} = (V_s + \kappa v)_s = V_{ss} + \kappa v_s + \kappa_s v = V_{ss} + \kappa^2 V + \kappa_s v.$$

This completes the proof. ∎

By the assumption that Γ_t touches $\partial\Omega$ with the angle $\pi/2$, we have

$$\psi(X(r_\pm(t), t)) = 0, \quad (\nabla\psi(X), N)_{\mathbb{R}^2} = 0 \quad \text{at } s = r_\pm(t).$$

Then we derive the following lemma.

LEMMA 3.3. *Let X be a smooth arc-length parameterization as above. Then*

$$v(r_\pm(t), t) + r'_\pm(t) = 0.$$

Proof. Differentiating $\psi(X(r_\pm(t), t)) = 0$ with respect to t and using $(\nabla\psi(X), N)_{\mathbb{R}^2} = 0$ at $s = r_\pm(t)$, we have at $s = r_\pm(t)$

$$\begin{aligned} 0 &= (\nabla\psi(X), X_s r'_\pm + X_t)_{\mathbb{R}^2} = (\nabla\psi(X), X_s r'_\pm + VN + vT)_{\mathbb{R}^2} \\ &= (v + r'_\pm)(\nabla\psi(X), T)_{\mathbb{R}^2} = \pm\,(v + r'_\pm)|\nabla\psi(X)|. \end{aligned}$$

The last identity is derived with the help of $T = \pm\,\nabla\psi(X)/|\nabla\psi(X)|$ at $s = r_\pm(t)$. Since $|\nabla\psi(X)| \neq 0$, we obtain the desired result. ∎

Now we can present an evolution equation for the curvature.

PROPOSITION 3.4 (evolution equation for the curvature). *Let evolving curves $\Gamma = \{\Gamma_t\}_{t\geq 0}$ be lying in Ω with $\partial\Gamma \subset \partial\Omega$. Then a smooth solution of*

$$V = -\kappa_{ss} \quad \text{on} \quad \Gamma_t \tag{3.3}$$

with the boundary conditions

$$\begin{cases} \sphericalangle(\Gamma_t, \partial\Omega) = \pi/2 & \text{at} \quad \Gamma_t \cap \partial\Omega, \\ \kappa_s = 0 & \text{at} \quad \Gamma_t \cap \partial\Omega \end{cases} \tag{3.4}$$

fulfills for $t > 0$

$$\kappa_t = -\kappa_{ssss} - \kappa^2\kappa_{ss} + \kappa_s v \quad \text{on} \quad \Gamma_t \tag{3.5}$$

and

$$\begin{cases} \kappa_s = 0 & \text{at} \quad \Gamma_t \cap \partial\Omega, \\ (\partial_s \pm h_\pm)\kappa_{ss} = 0 & \text{at} \quad \Gamma_t \cap \partial\Omega. \end{cases} \tag{3.6}$$

Here $h_\pm$ is the curvature of $\partial\Omega$ at the points $X(r_\pm(t),t) \in \Gamma_t \cap \partial\Omega$ with the sign convention that $h_\pm \le 0$ if Ω is convex.

Proof. We immediately obtain (3.5) from (3.3) and Lemma 3.2. Next we show (3.6). Differentiating $(\nabla\psi(X), N)_{\mathbb{R}^2} = 0$ at $s = r_\pm(t)$ with respect to t and using (3.1), (3.2), Lemma 3.1, and Lemma 3.3, we have at $s = r_\pm(t)$

$$\begin{aligned}
0 &= ([D^2\psi(X)](X_s r'_\pm + X_t), N)_{\mathbb{R}^2} + (\nabla\psi(X), N_s r'_\pm + N_t)_{\mathbb{R}^2} \\
&= (v + r'_\pm)([D^2\psi(X)]T, N)_{\mathbb{R}^2} + V([D^2\psi(X)]N, N)_{\mathbb{R}^2} \\
&\quad - \kappa r'_\pm(\nabla\psi(X), T)_{\mathbb{R}^2} - \theta_t(\nabla\psi(X), T)_{\mathbb{R}^2} \\
&= V([D^2\psi(X)]T_{\partial\Omega}(X), T_{\partial\Omega}(X))_{\mathbb{R}^2} - V_s(\nabla\psi(X), T)_{\mathbb{R}^2} \\
&\quad - \kappa(v + r'_\pm)(\nabla\psi(X), T)_{\mathbb{R}^2} \\
&= V([D^2\psi(X)]T_{\partial\Omega}(X), T_{\partial\Omega}(X))_{\mathbb{R}^2} \mp V_s|\nabla\psi(X)|.
\end{aligned}$$

Here $D^2\psi$ is the Hessian matrix of ψ. Then we observe that

$$\kappa_{\partial\Omega}(X) = -\frac{1}{|\nabla\psi(X)|}([D^2\psi(X)]T_{\partial\Omega}(X), T_{\partial\Omega}(X))_{\mathbb{R}^2},$$

so that

$$V_s \pm h_\pm V = 0 \quad \text{at } \ s = r_\pm(t),$$

where $h_\pm$ are given by $h_\pm := \kappa_{\partial\Omega}(X(r_\pm(t),t))$. This completes the proof. ∎

4. A priori estimates and global existence. We now derive basic evolution formulas for length and $\int_{\Gamma_t} \kappa_s^2\, ds$.

LEMMA 4.1. *A smooth solution of* (3.3)–(3.4) *fulfills*

(i) $\dfrac{d}{dt}L[\Gamma_t] = -\displaystyle\int_{\Gamma_t} \kappa_s^2\, ds,$

(ii) $$\frac{d}{dt}\int_{\Gamma_t} \kappa_s^2\, ds = -2\left\{\int_{\Gamma_t} V_s^2\, ds - \int_{\Gamma_t} \kappa^2 V^2\, ds + h_+(V^2|_{s=r_+(t)}) + h_-(V^2|_{s=r_-(t)})\right\} + \int_{\Gamma_t} \kappa_s^2 \kappa V\, ds,$$

where $h_\pm$ is evaluated at $X(r_\pm(t),t)$.

Proof. Recalling $L[\Gamma_t] = r_+(t) - r_-(t)$ and using Lemmas 3.1 and 3.3, we have

$$\begin{aligned}
\frac{d}{dt}L[\Gamma_t] = r'_+(t) - r'_-(t) &= -v(r_+(t),t) + v(r_-(t),t) = -\int_{\Gamma_t} v_s\, ds \\
&= -\int_{\Gamma_t} \kappa V\, ds = \int_{\Gamma_t} \kappa\kappa_{ss}\, ds = -\int_{\Gamma_t} \kappa_s^2\, ds.
\end{aligned}$$

The last term is derived using integration by parts and $\kappa_s = 0$ at $\Gamma_t \cap \partial\Omega$.

In order to prove (ii), we compute

$$\int_{\Gamma_t} \kappa_s(\kappa_t)_s\, ds = \int_{\Gamma_t} \kappa_s(-\kappa_{ssss} - \kappa^2\kappa_{ss} + \kappa_s v)_s\, ds. \tag{4.1}$$

Since $\kappa_{ts} = \kappa_{st}$ and $\kappa_s = 0$ at $\Gamma_t \cap \partial\Omega$, we have

$$\text{(left-hand side of (4.1))} = \int_{\Gamma_t} \kappa_s \kappa_{st}\, ds = \frac{1}{2}\int_{\Gamma_t} (\kappa_s^2)_t\, ds = \frac{1}{2}\frac{d}{dt}\int_{\Gamma_t} \kappa_s^2\, ds.$$

On the other hand, by means of integration by parts and using (3.6), we derive

$$\begin{aligned}\text{(right-hand side of (4.1))} &= -\int_{\Gamma_t} \kappa_{ss}(-\kappa_{ssss} - \kappa^2\kappa_{ss} + \kappa_s v)\, ds \\ &= \int_{\Gamma_t} \kappa_{ss}\kappa_{ssss}\, ds + \int_{\Gamma_t} \kappa^2\kappa_{ss}^2\, ds - \int_{\Gamma_t} \kappa_{ss}\kappa_s v\, ds \\ &= -h_+(\kappa_{ss}^2|_{s=r_+(t)}) - h_-(\kappa_{ss}^2|_{s=r_-(t)}) - \int_{\Gamma_t} \kappa_{sss}^2\, ds \\ &\quad + \int_{\Gamma_t} \kappa^2\kappa_{ss}^2\, ds + \frac{1}{2}\int_{\Gamma_t} \kappa_s^2 v_s\, ds.\end{aligned}$$

Thus it follows from $V = -\kappa_{ss}$ and $v_s = \kappa V$ that

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_{\Gamma_t} \kappa_s^2\, ds &= -\left\{\int_{\Gamma_t} V_s^2\, ds - \int_{\Gamma_t} \kappa^2 V^2\, ds + h_+(V^2|_{s=r_+(t)}) + h_-(V^2|_{s=r_-(t)})\right\} \\ &\quad + \frac{1}{2}\int_{\Gamma_t} \kappa_s^2 \kappa V\, ds.\end{aligned}$$

This completes the proof. □

Let us define the bilinear form I as

$$I[w,w] = \int_{r_-}^{r_+} (w_s^2 - \kappa_{av}^2 w^2)\, ds + h_+(w^2|_{s=r_+}) + h_-(w^2|_{s=r_-}) \tag{4.2}$$

for $w \in H^1(\Gamma_t)$ with

$$\int_{r_-}^{r_+} w\, ds = 0.$$

Here s is the arc-length parameter along Γ_t, which belongs to the interval $[r_-, r_+]$ with $L[\Gamma_t] = r_+ - r_-$; $h_\pm$ is the curvature of $\partial\Omega$ at $\Gamma_t \cap \partial\Omega$; and κ_{av} is the averaged curvature of Γ_t defined as

$$\kappa_{av} = \frac{1}{L[\Gamma_t]}\int_{r_-}^{r_+} \kappa\, ds.$$

Since $V = -\kappa_{ss}$ and $\kappa_s = 0$ at $\Gamma_t \cap \partial\Omega$, it holds that

$$\int_{\Gamma_t} V\, ds = 0. \tag{4.3}$$

Then we can rewrite Lemma 4.1(ii) as

$$\frac{d}{dt}\int_{\Gamma_t} \kappa_s^2\, ds + 2I[V,V] = -2\int_{\Gamma_t} (\kappa_{av}^2 - \kappa^2)V^2\, ds + \int_{\Gamma_t} \kappa_s^2 \kappa V\, ds. \tag{4.4}$$

Remark 4.2. Although the 90°-angle condition in (1.3) is the natural boundary condition when considering the gradient flow of the length functional, we can also consider the case where the prescribed angle is not 90°. In this case, the angle condition is represented as

$$(\nabla\psi(X), N)_{\mathbb{R}^2} = |\nabla\psi(X)| \cos\Theta_\pm \quad \text{at } \ s = r_\pm(t)$$

for $\Theta_\pm \in (0, \pi)$. Then the identity in Lemma 3.3 is replaced by

$$v + r'_\pm = \mp V \cot\Theta_\pm \quad \text{at } \ s = r_\pm(t),$$

and the second boundary condition of (3.6) is replaced by

$$\big\{\partial_s + (\pm h_\pm \csc\Theta_\pm \mp \kappa \cot\Theta_\pm)\big\}\kappa_{ss} = 0 \quad \text{at } \ s = r_\pm(t).$$

Further, setting

$$I_\Theta[w, w] = \int_{r_-}^{r_+} (w_s^2 - \kappa_{av}^2 w^2)\, ds + \big\{h_+ \csc\Theta_+ - (\kappa|_{s=r_+}) \cot\Theta_+\big\}(w^2|_{s=r_+}) + \big\{h_- \csc\Theta_- - (\kappa|_{s=r_-}) \cot\Theta_-\big\}(w^2|_{s=r_-})$$

instead of I in (4.2), we have

$$\frac{d}{dt}\int_{\Gamma_t} \kappa_s^2\, ds + 2I_\Theta[V, V] = -2\int_{\Gamma_t} (\kappa_{av}^2 - \kappa^2) V^2\, ds + \int_{\Gamma_t} \kappa_s^2 \kappa V\, ds.$$

We remark that for the non-90°-angle condition the stationary solutions with the property that I_Θ is positive are strict local minimizers of an energy also involving a wetting energy at the boundary if one takes an area constraint into account (see Vogel [23]). On the other hand,

$$\frac{d}{dt} L[\Gamma_t] = -\int_{\Gamma_t} \kappa_s^2\, ds - (V|_{s=r_+(t)}) \cot\Theta_+ - (V|_{s=r_-(t)}) \cot\Theta_-,$$

so that the monotonicity of $L[\Gamma_t]$ with respect to t is no longer fulfilled.

The following lemmas will be crucial in order to derive an a priori estimate.

LEMMA 4.3. *A smooth solution of* (3.3)–(3.4) *fulfills*

(i) $\displaystyle\left|\int_{\Gamma_t} \kappa_s^2 \kappa \kappa_{ss}\, ds\right| \le \frac{1}{3} L[\Gamma_t] \|\kappa_s\|_{L^2(\Gamma_t)}^2 \|\kappa_{ss}\|_{L^2(\Gamma_t)}^2,$

(ii) $\|\kappa - \kappa_{av}\|_{C^0(\Gamma_t)} \le L[\Gamma_t]^{1/2} \|\kappa_s\|_{L^2(\Gamma_t)}.$

Proof. We first prove (i). Since $\kappa_s = 0$ at $\Gamma_t \cap \partial\Omega$, we get

$$\int_{\Gamma_t} \kappa_s^2 \kappa \kappa_{ss}\, ds = -\frac{1}{3}\int_{\Gamma_t} \kappa_s^4\, ds.$$

Then it follows that

$$\left|\int_{\Gamma_t} \kappa_s^4\, ds\right| \le \|\kappa_s\|_{L^2(\Gamma_t)}^2 \|\kappa_s\|_{L^\infty(\Gamma_t)}^2 \le L[\Gamma_t] \|\kappa_s\|_{L^2(\Gamma_t)}^2 \|\kappa_{ss}\|_{L^2(\Gamma_t)}^2.$$

The last term is derived by using a Poincaré inequality since $\kappa_s = 0$ at $\Gamma_t \cap \partial\Omega$.

Next we prove (ii). Since

$$\int_{\Gamma_t} (\kappa - \kappa_{av})\, ds = 0,$$

for each $t > 0$, there is a $r_0\,(= r_0(t)) \in (r_-(t), r_+(t))$ such that $\kappa(r_0, t) - \kappa_{av}(t) = 0$. This implies that

$$|\kappa(s,\cdot) - \kappa_{av}| = \left| \int_{r_0}^{s} (\kappa - \kappa_{av})_s \, ds \right| = \left| \int_{r_0}^{s} \kappa_s \, ds \right| \le \int_{\Gamma_t} |\kappa_s| \, ds \le L[\Gamma_t]^{1/2} \|\kappa_s\|_{L^2(\Gamma_t)}.$$

Thus we have the desired result. □

We remind the reader that for functions w_1, w_2 with mean values zero we can define the H^{-1}-inner product via

$$(w_1, w_2)_{-1} = \int_{l_-}^{l_+} u_{1,\sigma} u_{2,\sigma} \, d\sigma,$$

where u_i is the solution of $-u_{i,\sigma\sigma} = w_i$ in (l_-, l_+) and $u_{i,\sigma} = 0$ at $\sigma = l_\pm$. According to [13], the bilinear form I^* as stated in the introduction (see (1.4)) is positive, provided that the maximal eigenvalue λ for the linearized problem to (1.2) and (1.3) is negative. In [13] it was shown that $I^*[w, w] \ge (-\lambda)(w, w)_{-1}$ for all $w \in H^1(\Gamma^*)$ with mean value zero. We now want to derive a perturbation of this result. Let us denote $L = L[\Gamma]$ and $L^* = L[\Gamma^*]\,(= l_+ - l_-)$. Then we have the following lemma, which implies a lower bound for I when the parameters κ_{av}, $h_\pm$, and L are close to κ^*, $h_\pm^*$, and L^*, respectively.

LEMMA 4.4. (i) *Let λ be the maximal eigenvalue of the linearized problem. For $\varepsilon > 0$ there exists $\delta > 0$ such that*

$$I[w, w] > (-\lambda - \varepsilon)(w, w)_{-1}$$

for $w \in H^1(\Gamma)$ with mean value zero, provided that

$$|\kappa_{av} - \kappa^*| < \delta, \quad |h_\pm - h_\pm^*| < \delta, \quad |L - L^*| < \delta.$$

(ii) *There exists $\mu > 0$ such that*

$$\mu \|w_s\|^2_{L^2(\Gamma)} \le I[w, w] + (w, w)_{-1}$$

for $w \in H^1(\Gamma)$ with mean value zero.

Proof. The largest eigenvalue λ corresponding to the bilinear form I depends continuously on L, κ_{av}, and $h_\pm$. In the case that $L = L^*$, $\kappa_{av} = \kappa^*$, and $h_\pm = h_\pm^*$ we obtain (i) with $\varepsilon = 0$, and hence (i) follows from a straightforward perturbation argument; compare [13] for similar arguments. Arguing as in the proof of Lemma 5.3 in [13], we obtain (ii). □

It is significant to obtain a positive lower bound of $L[\Gamma]$ in terms of ρ. The following lemma implies that L^* is a local minimum of $L[\Gamma]$, provided that I^* is positive.

LEMMA 4.5. *Let Γ^* be a stationary curve such that the bilinear form I^* is positive and let $\rho \in C^1(\mathcal{I})$ be a function describing a curve Γ close to Γ^* as in section 2. Assume that a curve Γ encloses the same area as Γ^*. Then there exist constants $\bar{c}$, $\gamma^* > 0$ such that*

$$L[\Gamma] \ge L^* + \bar{c} \|\rho\|^2_{H^1(\mathcal{I})}$$

if $\|\rho\|_{C^1(\mathcal{I})} < \gamma^*$.

Proof. This follows as in the proof of Theorem 2.1 of Vogel [23] (see (2.14) and the inequality after (2.19) in [23]). ∎

By virtue of Lemma 4.5, we have an a priori estimate of $L[\Gamma_t]$ and can derive useful estimates concerning κ_{av} and $h_\pm$.

LEMMA 4.6. *Let the assumptions of Lemma* 4.5 *hold for a stationary curve* Γ^* *and all curves* Γ_t, $t \in [0,T]$, *described by* $\rho(t) \in C^1(\mathcal{I})$ *for the parameterization in section* 2. *Assume in particular that* $\|\rho(t)\|_{C^1(\mathcal{I})} < \gamma^*$ *for* $t \in [0,T]$, *where* γ^* *is as in Lemma* 4.5. *We then obtain the following:*

(i) $L[\Gamma_0] \geq L[\Gamma_t] \geq L^*$ *for all* $t \in [0,T]$.

(ii) *There exist* $K_1, K_2 > 0$ *such that for* $t \in [0,T]$

$$|\kappa_{av}(t) - \kappa^*| \leq K_1 |L[\Gamma_t] - L^*|, \quad |h_\pm(t) - h_\pm^*| \leq K_2 |L[\Gamma_t] - L^*|.$$

Proof. (i) follows from Lemma 4.1(i) and Lemma 4.5. To prove (ii), we compute

$$\kappa_{av} = \frac{1}{L[\Gamma_t]} \int_{\Gamma_t} \kappa \, ds = \frac{1}{L[\Gamma_t]} \int_{\Gamma_t} \theta_s \, ds = \frac{1}{L[\Gamma_t]} (\theta_+ - \theta_-).$$

A similar computation gives

$$\kappa^* = \frac{1}{L^*} (\theta_+^* - \theta_-^*).$$

Then we have

$$\begin{aligned} |\kappa_{av} - \kappa^*| &= \left| \frac{1}{L[\Gamma_t]} (\theta_+ - \theta_-) - \frac{1}{L^*} (\theta_+^* - \theta_-^*) \right| \\ &= \frac{1}{L[\Gamma_t] L^*} |L^* (\theta_+ - \theta_-) - L[\Gamma_t] (\theta_+^* - \theta_-^*)| \\ &\leq \left(\frac{1}{L^*} \right)^2 \Big\{ |L^* (\theta_+ - \theta_- - (\theta_+^* - \theta_-^*))| + |L^* - L[\Gamma_t]| |\theta_+^* - \theta_-^*| \Big\}. \end{aligned}$$

By means of the mean value theorem, the smoothness of $\partial\Omega$, and the $\pi/2$ angle condition, we see that the quantity $|\theta_+ - \theta_+^*| + |\theta_-^* - \theta_-|$ is estimated by $\|\rho\|_{C^0(\mathcal{I})}$. Using Lemma 4.5 and an embedding result, we obtain the first inequality in (ii).

Recall that $\kappa_{\partial\Omega}(X)$ is represented by

$$\kappa_{\partial\Omega}(X) = -\frac{1}{|\nabla\psi(X)|} ([D^2\psi(X)] T_{\partial\Omega}(X), T_{\partial\Omega}(X))_{\mathbb{R}^2}.$$

Since this expression does not depend on derivatives of ρ, the mean value theorem implies that the quantity $|h_\pm - h_\pm^*|$ is estimated by $\|\rho\|_{C^0(\mathcal{I})}$. Using Lemma 4.5 and an embedding result, we derive the second inequality in (ii). ∎

Using Lemma 4.4, we obtain the existence of constants $\delta^* > 0$ and $\mu^* > 0$ such that

$$I[w,w] > -\frac{\lambda}{2} (w,w)_{-1} + \mu^* \|w_s\|_{L^2(\Gamma_t)}^2 \tag{4.5}$$

for $w \in H^1(\Gamma_t)$ with mean value zero, provided that

$$|\kappa_{av}(t) - \kappa^*| < \delta^*, \quad |h_\pm(t) - h_\pm^*| < \delta^*, \quad |L[\Gamma_t] - L^*| < \delta^*. \tag{4.6}$$

We are now in a position to derive a priori estimates for solutions of (2.5) if the solution is close to Γ^*.

PROPOSITION 4.7. *Let the assumptions of Lemma* 4.5 *hold for a stationary curve* Γ^* *and a curve* Γ_t *described by* $\rho(t) \in C^1(\mathcal{I})$ *for the parameterization in section* 2. *Assume that for* $t \in (0, T]$

$$\|\rho(t)\|_{C^1(\mathcal{I})} < \gamma^* \quad \text{and} \quad |L[\Gamma_t] - L^*| \le \frac{\delta^*}{1 + K_1 + K_2} \,(=: \delta_1^*), \tag{4.7}$$

where γ^* *is as in Lemma* 4.5, K_1 *and* K_2 *are as in Lemma* 4.6, *and* δ^* *is as in* (4.6). *Then there is a constant* $\delta_1 > 0$ *such that, if* $\|\kappa_s(t)\|^2_{L^2(\Gamma_t)} < \delta_1$ *for* $t \in (0, T]$, *it holds that*

$$\|\kappa_s(t)\|^2_{L^2(\Gamma_t)} + \mu^* \int_{t_0}^t \|V_s(\tau)\|^2_{L^2(\Gamma_t)}\, d\tau \le \|\kappa_s(t_0)\|^2_{L^2(\Gamma_t)}$$

for $t \in [t_0, T]$ *with* $t_0 > 0$, *where* μ^* *is as in* (4.5).

Proof. By (4.4), we have

$$\begin{aligned}
&\frac{d}{dt}\|\kappa_s\|^2_{L^2(\Gamma_t)} + 2I[V, V] \\
&= -2\int_{\Gamma_t} (\kappa_{av}^2 - \kappa^2) V^2\, ds + \int_{\Gamma_t} \kappa_s^2 \kappa V\, ds \\
&= 2\int_{\Gamma_t} (\kappa - \kappa_{av})^2 V^2\, ds + 4\kappa_{av}\int_{\Gamma_t} (\kappa - \kappa_{av}) V^2\, ds + \int_{\Gamma_t} \kappa_s^2 \kappa V\, ds.
\end{aligned}$$

By virtue of (4.7) and Lemma 4.6(ii), we also see that $\kappa_{av}(t)$, $h_\pm(t)$, and $L[\Gamma_t]$ satisfy (4.6). Then it follows from Lemma 4.3, Lemma 4.6(i), and (4.5) that there exist $C_1, C_2 > 0$ such that

$$\begin{aligned}
&\frac{d}{dt}\|\kappa_s\|^2_{L^2(\Gamma_t)} + (-\lambda)\,(V, V)_{-1} + 2\mu^*\|V_s\|^2_{L^2(\Gamma_t)} \\
&\le C_1\|V\|^2_{L^2(\Gamma_t)}\|\kappa_s\|^2_{L^2(\Gamma_t)} + C_2(\delta^* + |\kappa^*|)\|V\|^2_{L^2(\Gamma_t)}\|\kappa_s\|_{L^2(\Gamma_t)}.
\end{aligned}$$

Since $\|V\|_{L^\infty(\Gamma_t)} \le C\|V_s\|_{L^2(\Gamma_t)}$ by virtue of (4.3), we derive $\|V\|_{L^2(\Gamma_t)} \le \widetilde{C}\|V_s\|_{L^2(\Gamma_t)}$. By means of this fact and $(-\lambda)\,(V, V)_{-1} \ge 0$, we are led to

$$\frac{d}{dt}\|\kappa_s\|^2_{L^2(\Gamma_t)} + \{2\mu^* - \widetilde{C}_1\|\kappa_s\|^2_{L^2(\Gamma_t)} - \widetilde{C}_2(\delta^* + |\kappa^*|)\|\kappa_s\|_{L^2(\Gamma_t)}\}\|V_s\|^2_{L^2(\Gamma_t)} \le 0. \tag{4.8}$$

Then we choose δ_1 such that

$$0 < \delta_1 < \min\left\{ \frac{\mu^*}{2\,\widetilde{C}_1}\,,\ \left(\frac{\mu^*}{2\,\widetilde{C}_2(\delta^* + |\kappa^*|)} \right)^2 \right\}.$$

Assuming $\|\kappa_s(t)\|^2_{L^2(\Gamma_t)} < \delta_1$ for $t \in (0, T]$, it follows that

$$\frac{d}{dt}\|\kappa_s(t)\|^2_{L^2(\Gamma_t)} + \mu^*\|V_s(t)\|^2_{L^2(\Gamma_t)} \le 0. \tag{4.9}$$

Integrating (4.9) with respect to t in the interval $[t_0, t]$, we derive the desired result. ∎

Now we arrive at the main result in this section.

THEOREM 4.8 (global existence). *Let Γ^* be a stationary curve such that the bilinear form I^* is positive. Also, let $\rho_0 \in C^{2+\alpha}(\mathcal{I})$ be a function describing a curve Γ_0, which is close to Γ^* as in section 2 and satisfies $\Gamma_0 \perp \partial\Omega$. Assume that a curve Γ_0 includes the same area as Γ^*. Then there exist constants $\gamma_0 > 0$ and $\delta_0 > 0$ such that if $\|\rho_0\|_{C^1(\mathcal{I})} < \gamma_0$ and $L[\Gamma_0] - L^* < \delta_0$, the problem (2.5) admits a unique global-in-time solution ρ with*

$$\|\rho(t)\|_{C^1(\mathcal{I})} < \gamma_0 \quad \text{and} \quad L[\Gamma_t] - L^* < \delta_0 \quad \text{for} \quad t \geq 0,$$

where Γ_t is the curve parameterized by $\Psi(\sigma, \rho(\sigma, t))$ in section 2.

Proof. Choose γ_0 and δ_0 satisfying

$$0 < \gamma_0 < \frac{\gamma^*}{2}, \quad 0 < \delta_0 < \frac{\delta_1^*}{2}, \tag{4.10}$$

where γ^* is as in Lemma 4.5 and δ_1^* is as in (4.7). Assume that the initial curve Γ_0 satisfies $\|\rho_0\|_{C^1(\mathcal{I})} < \gamma_0$ and $L[\Gamma_0] - L^* < \delta_0$. Then Lemma 4.5 and an embedding result imply that

$$\|\rho_0\|_{C^0(\mathcal{I})} \leq C(L[\Gamma_0] - L^*) < C\delta_0. \tag{4.11}$$

Further, Lemma 4.6(i) implies that for $t > 0$

$$L[\Gamma_t] - L^* \leq L[\Gamma_0] - L^* < \delta_0. \tag{4.12}$$

We now prove that $\|\kappa_s(t)\|^2_{L^2(\Gamma_t)} < \delta_1$ for each time t in the existence interval of the solution, where δ_1 is as in Proposition 4.7. Let $0 < \beta < \alpha < 1/2$. By Theorem 2.1, we can construct a unique local-in-time solution for $\rho_0 \in C^{2+\beta}(\mathcal{I})$ and obtain the estimate

$$\|\rho\|_{\mathcal{Y}(\overline{\mathcal{Q}_{0,T_0}})} \leq K_0, \tag{4.13}$$

where K_0 is a constant, which depends on $\|\rho_0\|_{C^{2+\beta}(\mathcal{I})}$ increasingly, and T_0 is the local existence time, which depends on $1/\|\rho_0\|_{C^{2+\beta}(\mathcal{I})}$ increasingly (for details, see Appendix B). According to the interpolation inequality for Hölder spaces and (4.11), we have

$$\|\rho_0\|_{C^{2+\beta}(\mathcal{I})} \leq C(\|\rho_0\|_{C^0(\mathcal{I})})^{\frac{\alpha-\beta}{2+\alpha}}(\|\rho_0\|_{C^{2+\alpha}(\mathcal{I})})^{\frac{2+\beta}{2+\alpha}} \leq \widetilde{C}\delta_0^{\frac{\alpha-\beta}{2+\alpha}}. \tag{4.14}$$

Set $t_0 := \delta_0^{\frac{\alpha-\beta}{2+\alpha}} > 0$. Then it follows from (4.13), (4.14), and the definition of $\mathcal{Y}(\overline{\mathcal{Q}_{0,T_0}})$ that there exist $C > 0$ and $\nu > 0$ such that

$$\|\kappa_s(t_0)\|^2_{L^2(\Gamma_t)} \leq C\delta_0^\nu.$$

Since $\|\rho(t)\|_{C^1(\mathcal{I})}$ is continuous for $t \in [0, T_0]$, we see that $\|\rho(t)\|_{C^1(\mathcal{I})} < \gamma^*$ for $t \in [0, T]$ with a $T \in (0, T_0]$. Further, by (4.10) and (4.12), we have $L[\Gamma_t] - L^* < \delta_1^*$ for $t > 0$. Choose δ_0 such that $t_0 < T$ and $C\delta_0^\nu < \delta_1$. Then, by applying a similar argument to [10, Proof of Theorem 6.1] together with Proposition 4.7, we obtain that $\|\kappa_s(t)\|^2_{L^2(\Gamma_t)} < \delta_1$ for $t \in [t_0, T]$.

Next we prove that $\|\rho(t)\|_{C^1(\mathcal{I})} < \gamma_0$ for $t \in [t_0, T]$. By Lemma 4.5 and (4.12), it holds that for $t \in [0, T]$

$$\bar{c}\|\rho(t)\|_{H^1(\mathcal{I})} \le L[\Gamma_t] - L^* < \delta_0. \tag{4.15}$$

Then, by the embedding inequality and (4.15), we see that $\|\rho(t)\|_{C^0(\mathcal{I})} \le C\delta_0$ for $t \in [0, T]$. On the other hand, it follows from Lemma 4.3(ii) and Lemma 4.6(ii) that there exists $C > 0$ such that for $t \in [t_0, T]$

$$\|\kappa(t)\|_{C^0(\Gamma_t)} \le \|\kappa(t) - \kappa_{av}(t)\|_{C^0(\Gamma_t)} + |\kappa_{av}(t) - \kappa^*| + |\kappa^*| \le C(\delta_1 + \delta_0) + |\kappa^*|. \tag{4.16}$$

Thus, by virtue of (4.15), (4.16), and $\|\kappa_s(t)\|^2_{L^2(\Gamma_t)} < \delta_1$ for $t \in [t_0, T]$, we derive the boundedness of $\|\rho(t)\|_{H^3(\mathcal{I})}$ for $t \in [t_0, T]$, which implies the boundedness of $\|\rho(t)\|_{C^{2+\alpha}(\mathcal{I})}$ for $\alpha \in (0, 1/2)$. Then, by the interpolation inequality for Hölder spaces, we have

$$\|\rho(t)\|_{C^1(\mathcal{I})} \le C(\|\rho(t)\|_{C^0(\mathcal{I})})^{\frac{1+\alpha}{2+\alpha}}(\|\rho(t)\|_{C^{2+\alpha}(\mathcal{I})})^{\frac{1}{2+\alpha}} \le \widetilde{C}\delta_0^{\frac{1+\alpha}{2+\alpha}}$$

for $t \in [t_0, T]$. Choosing δ_0 such that $\widetilde{C}\delta_0^{\frac{1+\alpha}{2+\alpha}} < \gamma_0$, we obtain $\|\rho(t)\|_{C^1(\mathcal{I})} < \gamma_0$ for $t \in [t_0, T]$.

Finally, let us derive the existence of a unique global-in-time solution. Repeating the above argument until the local existence time T_0, we see that Γ_t satisfies

$$\|\rho(t)\|_{C^1(\mathcal{I})} < \gamma_0, \quad L[\Gamma_t] - L^* < \delta_0, \quad \|\kappa_s(t)\|^2_{L^2(\Gamma_t)} < \delta_1 \tag{4.17}$$

for $t \in [t_0, T_0]$. This implies that Γ_{T_0} satisfies the same conditions as those fulfilled by Γ_0 and the boundedness of $\|\rho(T_0)\|_{C^{2+\alpha}(\mathcal{I})}$ for $\alpha \in (0, 1/2)$ is guaranteed. Thus, due to Theorem 2.1, the solution of (2.5) can be extended over $t = T_0$ by a fixed amount of time. Further, by applying the same argument as we did in the first half of this proof, we have the estimates (4.17) for each time t in the extended existence interval of the solution. This procedure can be iterated as many times as we want, so that a unique global-in-time solution of (2.5) with $\rho(\cdot, 0) = \rho_0$ can be obtained. ∎

5. Stability of stationary curves. The following theorem shows nonlinear stability of the stationary curve Γ^* when the bilinear form I^* is positive.

THEOREM 5.1 (nonlinear stability). *Let the assumption of Theorem* 4.8 *hold. Then*

$$\|\rho(t)\|_{H^3(\mathcal{I})} \to 0 \quad \textit{as} \quad t \to \infty\,.$$

Proof. We apply a method similar to the one used in [10, Proof of Theorem 6.4]. By Lemma 4.1(i), we see that

$$\int_0^\infty \|\kappa_s(\tau)\|^2_{L^2(\Gamma_\tau)}\, d\tau \le L[\Gamma_0].$$

This implies that for any $\varepsilon \in (0, \delta_1)$ there exists a sufficiently large $t_\varepsilon > 0$ such that

$$\|\kappa_s(t_\varepsilon)\|^2_{L^2(\Gamma_t)} < \varepsilon.$$

According to the proof of Theorem 4.8, it holds that $\|\kappa_s(t)\|^2_{L^2(\Gamma_t)} < \delta_1$ as long as the solution exists. Thus, applying Proposition 4.7 for $t \in [t_\varepsilon, \infty)$, we have

$$\|\kappa_s(t)\|^2_{L^2(\Gamma_t)} + \mu^* \int_{t_\varepsilon}^t \|V_s(\tau)\|^2_{L^2(\Gamma_t)}\, d\tau \le \|\kappa_s(t_\varepsilon)\|^2_{L^2(\Gamma_t)} < \varepsilon.$$

This means that

$$\|\kappa_s(t)\|^2_{L^2(\Gamma_t)} \to 0 \quad \text{as} \quad t \to \infty. \tag{5.1}$$

By (5.1) and Lemma 4.3(ii), we also see that

$$\|\kappa(\cdot,t) - \kappa_{av}(t)\|_{C^0(\Gamma_t)} \to 0 \quad \text{as} \quad t \to \infty. \tag{5.2}$$

On the other hand, by virtue of Lemma 4.5 and Lemma 4.6(i), we obtain the boundedness of $\|\rho(t)\|_{H^1(\mathcal{I})}$. Using Lemma 4.6 and (5.2), we also have the boundedness of $\|\kappa(t)\|_{L^2(\Gamma_t)}$. Then the boundedness of $\|\rho(t)\|_{H^1(\mathcal{I})}$ and $\|\kappa(t)\|_{L^2(\Gamma_t)}$ imply the boundedness of $\|\rho(t)\|_{H^2(\mathcal{I})}$. Since it follows from the boundedness of $\|\rho(t)\|_{H^2(\mathcal{I})}$ and (5.1) that $\|\rho(t)\|_{H^3(\mathcal{I})}$ is bounded, there exists a sequence $\{t_n\}_{n\in\mathbb{N}}$ and $\widetilde{\rho}$ such that

$$\rho(t_n) \to \widetilde{\rho} \quad \text{in} \quad C^{2+\alpha}(\mathcal{I}) \quad \text{as} \quad n \to \infty.$$

By virtue of (5.2), $\widetilde{\rho}$ satisfies $\widetilde{\kappa} - \widetilde{\kappa}_{av} = 0$. The solution of the problem

$$\kappa = \kappa_{av}, \quad \sphericalangle(\Gamma, \partial\Omega) = \pi/2, \quad \text{Area}\,[\Gamma] = \text{Area}\,[\Gamma^*]$$

is unique in the C^0-neighborhood of Γ^* and given by $\rho \equiv 0$ (see Theorem 5.2). Since $\widetilde{\rho}$ is a solution of this problem, we obtain $\widetilde{\rho} \equiv 0$. In particular, we get

$$L[\Gamma_{t_n}] \to L[\Gamma^*] = L^* \quad \text{as} \quad n \to \infty.$$

We remark that Γ_{t_n} and Γ^* are the curves described by $\rho = \rho(t_n)$ and $\rho \equiv 0$ for the parameterization in section 2, respectively. Then, by the fact that $L[\Gamma_t]$ decreases in time, we obtain that

$$L[\Gamma_t] \to L^* \quad \text{as} \quad t \to \infty.$$

Applying Lemma 4.5, we have

$$\|\rho(t)\|^2_{H^1(\mathcal{I})} \to 0 \quad \text{as} \quad t \to \infty.$$

Hence, using this fact together with both (5.1) and (5.2), we obtain the desired result. □

It remains to prove the following result. We refer the reader to Grosse-Brauckmann [14] for a similar proof in the case of a different boundary condition.

THEOREM 5.2. *Let Γ^* be a stationary curve such that the bilinear form I^* is positive and let Γ be a curve described by ρ for the parameterization in section 2. Then there exists a C^2-neighborhood of Γ^* such that $\rho \equiv 0$ is the unique solution of the problem*

$$\kappa = \kappa_{av}\,, \quad \sphericalangle(\Gamma, \partial\Omega) = \pi/2\,, \quad \text{Area}\,[\Gamma] = \text{Area}\,[\Gamma^*]. \tag{5.3}$$

Proof. We use the following implicit function theorem (see Zeidler [24, Theorem 4.B]).

Suppose that the following hold:

(i) *The mapping $F : U(x_0, y_0) \subset X \times Y \to Z$ is defined on an open neighborhood $U(x_0, y_0)$ of (x_0, y_0), and $F(x_0, y_0) = 0$, where X, Y, and Z are Banach spaces over $\mathbb{R}$.*

(ii) *F_y exists as the partial Fréchet derivative on $U(x_0, y_0)$ and*

$$F_y(x_0, y_0) : Y \to Z$$

is bijective.

(iii) *F and F_y are continuous at (x_0, y_0).*

Then the following holds true: There exist positive numbers r_0 and r such that, for every $x \in X$ satisfying $\|x - x_0\| < r_0$, there is exactly one $y(x) \in Y$ for which $\|y(x) - y_0\| \le r$ and $F(x, y(x)) = 0$.

We use this theorem for

$$X := \{\rho \in C^2(\mathcal{I}) \mid \rho = \text{const.}\},$$

$$Y := \left\{\rho \in C^2(\mathcal{I}) \;\middle|\; \int_{l_-}^{l_+} \rho \, d\sigma = 0\right\},$$

$$Z := \left\{\rho \in C^0(\mathcal{I}) \;\middle|\; \int_{l_-}^{l_+} \rho \, d\sigma = 0\right\} \times \mathbb{R}^2$$

and

$$F(m, u) := \left(\kappa - \kappa_{av},\ \sphericalangle(\partial\Omega, \Gamma_t)_+ - \frac{\pi}{2},\ \sphericalangle(\partial\Omega, \Gamma_t)_- - \frac{\pi}{2}\right),$$

where κ is computed for the curve that we get by taking $\rho = u + m$ in section 2. The expression $\sphericalangle(\partial\Omega, \Gamma_t)_\pm$ denotes the angles with the outer boundary at the two boundary points. The derivative $F_u(0, 0)$ is (by a similar computation as in [13]) given by

$$F_u(0,0)(v)$$

$$= \left((\partial_\sigma^2 + \kappa^2)v - \frac{1}{l_+ - l_-}\int_{l_-}^{l_+} (\partial_\sigma^2 + \kappa^2)v \, d\sigma,\ (\partial_\sigma + h_+)v(l_+),\ (\partial_\sigma - h_-)v(l_-)\right).$$

The fact that I^* is positive implies that $F_u(0, 0)$ is invertible (using regularity theory for ODEs). Straightforward computations show that F and F_u are continuous at $(0, 0)$.

Hence, for $m \in X$ small, we find exactly one $u(m)$ such that

$$F(m, u(m)) = 0.$$

Let us define

$$\rho_m = u(m) + m$$

and let Γ_m be a curve described by ρ_m for the parameterization in section 2. Then we have

$$\text{Area}[\Gamma_m] = \text{Area}[\Gamma^*] + \int_{l_-}^{l_+} (u(m) + m) \, d\sigma + \mathcal{O}(\|u(m) + m\|^2_{C^2(\mathcal{I})})$$

$$= \text{Area}[\Gamma^*] + (l_+ - l_-)m + \mathcal{O}(\|u(m) + m\|^2_{C^2(\mathcal{I})}).$$

This implies that for $m \neq 0$

$$|\,\mathrm{Area}[\Gamma_m] - \mathrm{Area}[\Gamma^*]\,| \neq 0 \tag{5.4}$$

if $\|(m, u(m))\|_{C^2(\mathcal{I})}$ is small enough. We now represent a solution ρ of (5.3) with $\|\rho\|_{C^2(\mathcal{I})}$ small as $\rho = u + m$, where $u = \rho - \rho_{av}$ and $m = \rho_{av}$ with

$$\rho_{av} = \frac{1}{l_+ - l_-} \int_{l_-}^{l_+} \rho \, d\sigma.$$

Then we see that $F(m, u) = 0$. Due to the area-preserving property and (5.4), we obtain $m = 0$ and $u \equiv 0$, which implies that $\rho \equiv 0$. This proves the theorem. □

Appendix A. Proof of Theorem 2.1. The problem (2.5) is an initial boundary value problem for a quasi-linear parabolic PDE which has the form

$$\begin{cases} \rho_t = -\dfrac{1}{(J(\rho))^4}\,\partial_\sigma^4\rho + a(\rho, \partial_\sigma\rho, \partial_\sigma^2\rho)\,\partial_\sigma^3\rho + f(\rho, \partial_\sigma\rho, \partial_\sigma^2\rho) & \text{in} \quad \mathcal{Q}_{0,T}, \\ b_1(\rho)\partial_\sigma\rho + g_1(\rho) = 0 & \text{at} \quad \sigma = l_\pm, \\ b_2(\rho, \partial_\sigma\rho)\,\partial_\sigma^3\rho + g_2(\rho, \partial_\sigma\rho, \partial_\sigma^2\rho) = 0 & \text{at} \quad \sigma = l_\pm, \\ \rho|_{t=0} = \rho_0 & \text{in} \quad \mathcal{I}, \end{cases} \tag{A.1}$$

where a, f, b_i, and g_i $(i = 1, 2)$ are smooth functions with respect to ρ, $\partial_\sigma\rho$, and $\partial_\sigma^2\rho$; and g_i $(i = 1, 2)$ satisfy $\|g_1(t)\|_{C^0(\mathcal{I})} = \mathcal{O}(\|\rho(t)\|_{C^0(\mathcal{I})})$ and $\|g_2(t)\|_{C^0(\mathcal{I})} = \mathcal{O}(\|\rho(t)\|_{C^{2+\alpha}(\mathcal{I})})$ when $\|\rho\|_{C^{2+\alpha}(\mathcal{I})} \to 0$. In order to prove Theorem 2.1, we apply a fixed point argument. Let

$$\mathcal{D} := \{\rho \in \mathcal{Y}(\overline{\mathcal{Q}_{0,T}}) \mid \rho(\cdot, 0) = \rho_0,\ \|\rho\|_{\mathcal{Y}(\overline{\mathcal{Q}_{0,T}})} \leq K\}$$

for positive constants K and T, and define a mapping $\mathcal{P}$ as

$$\mathcal{P} : \mathcal{D} \ni \bar{\rho} \mapsto \rho \in \mathcal{Y}(\overline{\mathcal{Q}_{0,T}}),$$

where ρ is the unique solution of the linearized problem

$$\begin{cases} \rho_t = \mathcal{A}\rho + F(\sigma, t) & \text{for} \quad (\sigma, t) \in \mathcal{Q}_{0,T}, \\ \mathcal{B}_1\rho = G_1(\sigma, t) & \text{at} \quad \sigma = l_\pm,\ t \in (0, T], \\ \mathcal{B}_2\rho = G_2(\sigma, t) & \text{at} \quad \sigma = l_\pm,\ t \in (0, T], \\ \rho(\sigma, 0) = \rho_0 & \text{for} \quad \sigma \in \mathcal{I}. \end{cases} \tag{A.2}$$

Here the linearized operators $\mathcal{A}$, $\mathcal{B}_1$, and $\mathcal{B}_2$ around the initial data $\rho_0 \in C^{2+\alpha}(\mathcal{I})$ are given by

$$\mathcal{A} = -\frac{1}{(J(\rho_0))^4}\,\partial_\sigma^4 + a(\rho_0, \partial_\sigma\rho_0, \partial_\sigma^2\rho_0)\,\partial_\sigma^3\,,$$
$$\mathcal{B}_1 = b_1(\rho_0)\,\partial_\sigma, \quad \mathcal{B}_2 = b_2(\rho_0, \partial_\sigma\rho_0)\,\partial_\sigma^3\,,$$

and for given $\bar{\rho} \in \mathcal{D}$

$$\begin{aligned}
F(\sigma,t) &= -\left\{ \frac{1}{(J(\bar{\rho}))^4} - \frac{1}{(J(\rho_0))^4} \right\} \partial_\sigma^4 \bar{\rho} \\
&\quad + \left\{ a(\bar{\rho}, \partial_\sigma \bar{\rho}, \partial_\sigma^2 \bar{\rho}) - a(\rho_0, \partial_\sigma \rho_0, \partial_\sigma^2 \rho_0) \right\} \partial_\sigma^3 \bar{\rho} \\
&\quad + f(\bar{\rho}, \partial_\sigma \bar{\rho}, \partial_\sigma^2 \bar{\rho}), \\
G_1(\sigma,t) &= -\left\{ b_1(\bar{\rho}) - b_1(\rho_0) \right\} \partial_\sigma \bar{\rho} - g_1(\bar{\rho}), \\
G_2(\sigma,t) &= -\left\{ b_2(\bar{\rho}, \partial_\sigma \bar{\rho}) - b_2(\rho_0, \partial_\sigma \rho_0) \right\} \partial_\sigma^3 \bar{\rho} - g_2(\bar{\rho}, \partial_\sigma \bar{\rho}, \partial_\sigma^2 \bar{\rho}).
\end{aligned}$$

The existence of a unique solution for the linearized problem (A.2) in $\mathcal{Y}(\overline{\mathcal{Q}_{0,T}})$ is proved by applying the optimal regularity theory for analytic semigroups to the linearized problem (A.2) (see [18]). If the mapping $\mathcal{P}$ is a contraction on $\mathcal{D}$ for suitable constants K and T depending on $\|\rho_0\|_{C^{2+\alpha}(\mathcal{I})}$, $\mathcal{P}$ has a unique fixed point in $\mathcal{D}$ which is a unique solution of the nonlinear problem (A.1). Thus we show that the mapping $\mathcal{P}$ is a contraction on $\mathcal{D}$. In order to prove this fact, the following lemma is crucial.

LEMMA A.1. (i) *Assume that $\bar{\rho} \in \mathcal{D}$ and that ρ is a solution of the linearized problem* (A.2). *Then there exist positive constants M_0 and N such that*

$$\|\rho\|_{\mathcal{Y}(\overline{\mathcal{Q}_{0,T}})} \le M_0 + NT^{\frac{\alpha}{4}}.$$

In particular, M_0 depends on $\|\rho_0\|_{C^{2+\alpha}(\mathcal{I})}$ increasingly, and N depends on K increasingly.

(ii) *Assume that $\bar{\rho}_1, \bar{\rho}_2 \in \mathcal{D}$ and that ρ_1, ρ_2 are solutions of the linearized problem* (A.2). *Then there exists a positive constant N such that*

$$\|\rho_1 - \rho_2\|_{\mathcal{Y}(\overline{\mathcal{Q}_{0,T}})} \le NT^{\frac{\alpha}{4}} \|\bar{\rho}_1 - \bar{\rho}_2\|_{\mathcal{Y}(\overline{\mathcal{Q}_{0,T}})}.$$

In particular, N depends on K increasingly.

A method to prove this lemma is to use the optimal regularity theory of analytic semigroups as in [18]. We prove this lemma in the next section.

Lemma A.1 implies that if we take

$$K = 2M_0, \quad T_0 = \min\left\{ \left(\frac{K}{2N} \right)^{4/\alpha}, \left(\frac{1}{2N} \right)^{4/\alpha} \right\},$$

it follows that for $T \le T_0$

$$\|\rho\|_{\mathcal{Y}(\overline{\mathcal{Q}_{0,T}})} \le K, \quad \|\rho_1 - \rho_2\|_{\mathcal{Y}(\overline{\mathcal{Q}_{0,T}})} \le \frac{1}{2} \|\bar{\rho}_1 - \bar{\rho}_2\|_{\mathcal{Y}(\overline{\mathcal{Q}_{0,T}})}.$$

This means that $\mathcal{P}$ maps $\mathcal{D}$ into itself and is a contraction on $\mathcal{D}$ for $T \le T_0$. Thus the proof of Theorem 2.1 is completed.

Appendix B. Proof of Lemma A.1. We prove only Lemma A.1(i). Applying a similar argument, we can also derive Lemma A.1(ii). It is convenient to introduce the following estimate without proof.

LEMMA B.1 (see [18, section 2]). *For $k \in \mathbb{N}$, β_1, $\beta_2 \in (0,1)$, and a sectorial operator A, there exists a constant $C = C(k, \beta_1, \beta_2, A)$ such that*

$$\|t^{k-\beta_1+\beta_2} A^k e^{tA}\|_{L(D_A(\beta_1,\infty), D_A(\beta_2,\infty))} \le C \quad \text{for} \quad 0 < t \le 1. \tag{B.1}$$

The statement also holds for $k = 0$, provided that $\beta_1 \le \beta_2$.

Define $X := C(\mathcal{I})$ and

$$D(A) := \{u \in C^4(\mathcal{I}) \mid \mathcal{B}_1 u(l_\pm) = \mathcal{B}_2 u(l_\pm) = 0\}.$$

Then $A : X \supset D(A) \ni u \mapsto \mathcal{A}u \in X$ is the realization of $\mathcal{A}$ in X. It is known that A is a sectorial operator in X (see [21]).

Let ρ be a unique solution of the linearized problem (A.2). In order to reduce the inhomogeneous problem to a homogeneous problem at the boundaries, we introduce an auxiliary function ζ defined as

$$\zeta(\sigma,t) := \left\{ \frac{(\sigma - l_-)G_1(l_-,t)}{b_1(\rho_0)\big|_{\sigma=l_-}} + \frac{(\sigma-l_-)^3 G_2(l_-,t)}{3!\, b_2(\rho_0, \partial_\sigma \rho_0)\big|_{\sigma=l_-}} \right\} \eta(\sigma)$$

$$+ \left\{ \frac{(\sigma - l_+)G_1(l_+,t)}{b_1(\rho_0)\big|_{\sigma=l_+}} + \frac{(\sigma-l_+)^3 G_2(l_+,t)}{3!\, b_2(\rho_0, \partial_\sigma \rho_0)\big|_{\sigma=l_+}} \right\} \hat{\eta}(\sigma),$$

where η, $\hat{\eta} \in C^\infty(\mathcal{I})$ are cut-off functions satisfying

$$\begin{cases} \eta'(\sigma) < 0, \quad \hat{\eta}'(\sigma) > 0 & \text{for} \quad \sigma \in (l_- + L^*/4, l_+ - L^*/4), \\ \eta(\sigma) \equiv 1, \quad \hat{\eta}(\sigma) \equiv 0 & \text{for} \quad \sigma \in [l_-, l_- + L^*/4], \\ \eta(\sigma) \equiv 0, \quad \hat{\eta}(\sigma) \equiv 1 & \text{for} \quad \sigma \in [l_+ - L^*/4, l_+]. \end{cases}$$

Then it follows that $\rho - \zeta$ fulfills homogeneous boundary conditions. Since A is sectorial, we represent $\rho - \zeta$ with the help of a variant of the variation of constants formula and the analytic semigroup e^{tA}. By a simple computation, we obtain for $0 \le t \le T$

$$\rho(\cdot,t) = \rho_1(\cdot,t) + \rho_2(\cdot,t) + \rho_3(\cdot,t),$$

where

$$\rho_1(\cdot,t) = e^{tA}\{\rho_0 - \zeta(\cdot,0)\},$$

$$\rho_2(\cdot,t) = \int_0^t e^{(t-r)A}\{F(\cdot,r) + \mathcal{A}\zeta(\cdot,r)\}\, dr,$$

$$\rho_3(\cdot,t) = -A\int_0^t e^{(t-r)A}\{\zeta(\cdot,r) - \zeta(\cdot,0)\}\, dr + \zeta(\cdot,0).$$

Applying the theory of analytic semigroups as in [18], we have (see below)

$$\text{(B.2)} \qquad \begin{cases} \|\rho_1\|_{\mathcal{Y}(\overline{\mathcal{Q}_{0,T}})} \le C_0 \, \|\rho_0 - \zeta(\cdot,0)\|_{D_A(\frac{2+\alpha}{4},\infty)}, \\ \|\rho_2\|_{\mathcal{Y}(\overline{\mathcal{Q}_{0,T}})} \le C_0 \sup\limits_{0<\delta<T} \delta^{\frac{1}{2}} \sup\limits_{t\in[\delta,T]} \|F(\cdot,t) + \mathcal{A}\zeta(\cdot,t)\|_{D_A(\frac{\alpha}{4},\infty)}, \\ \|\rho_3\|_{\mathcal{Y}(\overline{\mathcal{Q}_{0,T}})} \le C_0 + C_{0,K} T^{\frac{1}{4}}. \end{cases}$$

In particular, it is verified that a constant C_0 increases with $\|\rho_0\|_{C^{2+\alpha}(\mathcal{I})}$ and that a constant $C_{0,K}$ increases with $\|\rho_0\|_{C^{2+\alpha}(\mathcal{I})}$ and K. Once (B.2) is proven, it follows from characterization of interpolation spaces $D_A(\beta,\infty)$ (see, e.g., [1, 18, 19]) and

the definition of F that

$$\begin{aligned}\|\rho\|_{\mathcal{Y}(\overline{\mathcal{Q}_{0,T}})} &\le \|\rho_1\|_{\mathcal{Y}(\overline{\mathcal{Q}_{0,T}})} + \|\rho_2\|_{\mathcal{Y}(\overline{\mathcal{Q}_{0,T}})} + \|\rho_3\|_{\mathcal{Y}(\overline{\mathcal{Q}_{0,T}})}\\ &\le \widetilde{C}_0\|\rho_0 - \zeta(\cdot,0)\|_{C^{2+\alpha}(\mathcal{I})}\\ &\quad + \widetilde{C}_0 \sup_{0<\delta<T} \delta^{\frac{1}{2}} \sup_{t\in[\delta,T]} \|F(\cdot,t) + \mathcal{A}\zeta(\cdot,t)\|_{C^{\alpha}(\mathcal{I})}\\ &\quad + \widetilde{C}_0 + \widetilde{C}_{0,K}T^{\frac{1}{4}}\\ &\le M_0 + N_{0,K}T^{\frac{\alpha}{4}} + N_{0,K}T^{\frac{1}{4}},\end{aligned}$$

where $\widetilde{C}_0$ and M_0 depend on $\|\rho_0\|_{C^{2+\alpha}(\mathcal{I})}$ increasingly, and $\widetilde{C}_{0,K}$ and $N_{0,K}$ depend on $\|\rho_0\|_{C^{2+\alpha}(\mathcal{I})}$ and K increasingly. This completes the proof of Lemma A.1(i). Thus we give the proof of (B.2) in detail.

First let us explain about the estimates for ρ_1 and ρ_2. Using (B.1) with $k = 0$ and $\beta_1 = \beta_2 = (2+\alpha)/4$ to ρ_1, and with $k = 1$, $\beta_1 = (2+\alpha)/4$, and $\beta_2 = \alpha/4$ to $\partial\rho_1/\partial t = A\rho_1$, we are led to the estimate of ρ_1 easily. Since $F + \mathcal{A}\zeta \in L^\infty((0,T]; D_A(\frac{\alpha}{4},\infty))$, applying the same argument as [18, section 4.3.2] to ρ_2 in $[\varepsilon, T]$ ($\varepsilon \in (0,T)$), we have an estimate for ρ_2. Let us consider the estimate for ρ_3. Since ζ is less regular, we cannot derive the desired estimate for ρ_3 if we use only (B.1) to ρ_3 directly. Set

$$z(t) = \int_0^t e^{(t-r)A}\{\zeta(\cdot,r) - \zeta(\cdot,0)\}\,dr. \tag{B.3}$$

Then z satisfies

$$\rho_3(\cdot,t) = -Az(t) + \zeta(\cdot,0) = -\frac{d}{dt}z(t) + \zeta(\cdot,t),$$

$$\frac{d}{dt}\rho_3(\cdot,t) = -A\frac{d}{dt}z(t) = A\{\rho_3(\cdot,t) - \zeta(\cdot,t)\}.$$

This means that if we obtain the estimates for dz/dt, we have the desired estimates for ρ_3. In fact, the estimate for $\|\rho_3\|_{\mathcal{Y}(\overline{\mathcal{Q}_{0,T}})}$ is given by

$$\begin{aligned}\|\rho_3\|_{\mathcal{Y}(\overline{\mathcal{Q}_{0,T}})} &\le \|\zeta(\cdot,0)\|_{C^{2+\alpha}(\overline{\mathcal{Q}_{0,T}})} + \|\zeta(\cdot,t) - \zeta(\cdot,0)\|_{C^{2+\alpha}(\overline{\mathcal{Q}_{0,T}})}\\ &\quad + \sum_{i=1}^{3} \sup_{0<t<T} t^{\frac{1}{2}}\|\mathcal{A}\zeta(\cdot,t)\|_{C^{\alpha}(\overline{\mathcal{Q}_{0,T}})}\\ &\quad + \widetilde{C}\Big(\|\dot{z}(t)\|_{D_A(\frac{2+\alpha}{4},\infty)} + \sup_{0<\delta<T}\delta^{\frac{1}{2}} \sup_{t\in[\delta,T]}\|A\dot{z}(t)\|_{D_A(\frac{\alpha}{4},\infty)}\Big).\end{aligned}$$

Here and hereafter we use $\dot{z}$ instead of dz/dt to simplify the notation. For the function z, we have the following estimates.

LEMMA B.2. *Let z be a function represented by* (B.3). *Then there exists a constant N, which depends on $\|\rho_0\|_{C^{2+\alpha}(\mathcal{I})}$, α, and K, such that*

$$\begin{cases} \|\dot{z}(t)\|_{D_A(\frac{2+\alpha}{4},\infty)} \le NT^{\frac{1}{4}},\\ \sup\limits_{0<\delta<T}\delta^{\frac{1}{2}} \sup\limits_{t\in[\delta,T]}\|A\dot{z}(t)\|_{D_A(\frac{\alpha}{4},\infty)} \le NT^{\frac{1}{4}}. \end{cases} \tag{B.4}$$

Proof. The proof of the first estimate of (B.4) is similar to arguments in [16, Appendix]. We prove only the second estimate of (B.4). For $t \geq \varepsilon$ with $\varepsilon \in (0,T)$, we have

$$\dot{z}(t) = e^{(t-\varepsilon/2)A}\dot{z}(\varepsilon/2) + \int_{\varepsilon/2}^{t} Ae^{(t-r)A}\{\zeta(\cdot,r) - \zeta(\cdot,t)\}\,dr + e^{(t-\varepsilon/2)A}\{\zeta(\cdot,t) - \zeta(\cdot,\varepsilon/2)\}.$$

This implies that

$$\begin{aligned}\|A\dot{z}(t)\|_{D_A(\frac{\alpha}{4},\infty)} \leq\ & \|Ae^{(t-\varepsilon/2)A}\dot{z}(\varepsilon/2)\|_{D_A(\frac{\alpha}{4},\infty)} \\ & + \left\| \int_{\varepsilon/2}^{t} A^2 e^{(t-r)A}\{\zeta(\cdot,r) - \zeta(\cdot,t)\}\,dr \right\|_{D_A(\frac{\alpha}{4},\infty)} \\ & + \|Ae^{(t-\varepsilon/2)A}\{\zeta(\cdot,t) - \zeta(\cdot,\varepsilon/2)\}\|_{D_A(\frac{\alpha}{4},\infty)} \\ =:\ & I_1(t) + I_2(t) + I_3(t).\end{aligned}$$

Let us first derive the estimate of $I_1(t)$. It follows that for $t \geq \varepsilon$

$$I_1(t) \leq C_0(t-\varepsilon/2)^{-\frac{\alpha}{4}}\|A\dot{z}(\varepsilon/2)\| \leq C_0(\varepsilon/2)^{-\frac{\alpha}{4}}\|A\dot{z}(\varepsilon/2)\|. \tag{B.5}$$

Thus it is necessary to obtain an estimate of $\|A\dot{z}(t)\|$. Since $\dot{z}(0) = 0$, we see that

$$\|A\dot{z}(t)\| \leq \int_0^t \|A^2 e^{(t-r)A}\{\zeta(\cdot,r) - \zeta(\cdot,t)\}\|\,dr + \|Ae^{tA}\{\zeta(\cdot,t) - \zeta(\cdot,0)\}\|.$$

We now recall the definition of ζ. Then we have to estimate each term. We show the estimate only for the term including the function

$$\hat{\zeta}(\sigma,t) := (\sigma - l_-)^3 G_2(l_-,t)\eta(\sigma).$$

The idea for the estimation of the other terms is similar. Set

$$J_1(t) := \int_0^t \|A^2 e^{(t-r)A}\{\hat{\zeta}(\cdot,\sigma) - \hat{\zeta}(\cdot,t)\}\|\,dr,$$

$$J_2(t) := \|Ae^{tA}\{\hat{\zeta}(\cdot,t) - \hat{\zeta}(\cdot,0)\}\|.$$

Let us derive the estimate of $J_1(t)$. For $t > r$ we have

$$\begin{aligned}&|G_2(\cdot,t) - G_2(\cdot,r)| \\ &\leq |b_2(\bar{\rho}(\cdot,t),\partial_\sigma\bar{\rho}(\cdot,t)) - b_2(\rho_0,\partial_\sigma\rho_0)||\partial_\sigma^3\bar{\rho}(\cdot,t) - \partial_\sigma^3\bar{\rho}(\cdot,r)| \\ &\quad + |b_2(\bar{\rho}(\cdot,t),\partial_\sigma\bar{\rho}(\cdot,t)) - b_2(\bar{\rho}(\cdot,r),\partial_\sigma\bar{\rho}(\cdot,r))||\partial_\sigma^3\bar{\rho}(\cdot,r)| \\ &\quad + |g_2(\bar{\rho}(\cdot,t),\partial_\sigma\bar{\rho}(\cdot,t),\partial_\sigma^2\bar{\rho}(\cdot,t)) - g_2(\bar{\rho}(\cdot,r),\partial_\sigma\bar{\rho}(\cdot,r),\partial_\sigma^2\bar{\rho}(\cdot,r))| \\ &\leq C_K\big\{t^{\frac{1+\alpha}{4}}\cdot r^{-\frac{1}{2}}(t-r)^{\frac{1+\alpha}{4}} + r^{-\frac{1}{2}}(t-r)^{\frac{3+\alpha}{4}}\cdot r^{-\frac{1}{4}} + r^{-\frac{1}{2}}(t-r)^{\frac{2+\alpha}{4}}\big\}.\end{aligned}$$

This fact and characterization of interpolation spaces $D_A(\beta,\infty)$ imply that

$$\begin{aligned}
J_1(t) &\le C_0 \int_0^t (t-r)^{\frac{3}{4}-2} \|(\sigma - l_-)^3 \eta\|_{D_A(\frac{3}{4},\infty)} |G_2(l_-,t) - G_2(l_-,r)|\, dr \\
&\le C_{0,K} \int_0^t (t-r)^{\frac{3}{4}-2} \big\{ t^{\frac{1+\alpha}{4}} \cdot r^{-\frac{1}{2}} (t-r)^{\frac{1+\alpha}{4}} \\
&\qquad + r^{-\frac{1}{2}} (t-r)^{\frac{3+\alpha}{4}} \cdot r^{-\frac{1}{4}} + r^{-\frac{1}{2}} (t-r)^{\frac{2+\alpha}{4}} \big\}\, dr \\
&\le C_{0,K,\alpha} \, (t^{\frac{1+\alpha}{4}} + t^{\frac{1}{4}} + t^{\frac{1}{4}})\, t^{\frac{\alpha}{4}-\frac{1}{2}} \\
&\le \widetilde{C}_{0,K,\alpha} \, (t^{\frac{1+\alpha}{4}} + t^{\frac{1}{4}})\, t^{\frac{\alpha}{4}-\frac{1}{2}}.
\end{aligned}$$

Applying the similar argument to $J_2(t)$, we are led to

$$\begin{aligned}
J_2(t) &\le C_0\, t^{\frac{3}{4}-1} \|(\sigma - l_-)^3 \eta\|_{D_A(\frac{3}{4},\infty)} |G_2(l_-,t) - G_2(l_-,0)| \\
&\le C_{0,K}\, t^{\frac{3}{4}-1} \, (t^{\frac{1+\alpha}{4}} \cdot K t^{-\frac{1}{4}} + t^{\frac{\alpha}{4}}) \\
&\le \widetilde{C}_{0,K}\, t^{\frac{1}{4}} \cdot t^{\frac{\alpha}{4}-\frac{1}{2}}.
\end{aligned}$$

Since the estimates for the other terms are also obtained similarly, we have

$$\|A\dot{z}(t)\| \le C_{0,K,\alpha} T^{\frac{1}{4}} \cdot t^{\frac{\alpha}{4}-\frac{1}{2}}.$$

It follows from (B.5) that

$$I_1(t) \le C_{0,K,\alpha} T^{\frac{1}{4}} \cdot (\varepsilon/2)^{-\frac{1}{2}}.$$

Let us derive the estimate for $I_2(t)$. Set

$$w(t) := \int_{\varepsilon/2}^t A^2 e^{(t-r)A} \{\zeta(\cdot,r) - \zeta(\cdot,t)\}\, dr.$$

In order to obtain the estimate of $\|w\|_{D_A(\frac{\alpha}{4},\infty)}$, we recall the definition of $\|\cdot\|_{D_A(\frac{\alpha}{4},\infty)}$. Since the estimate of $\|w\|$ is similar to that of $J_1(t)$, we consider only the estimate of the seminorm. According to the definition, we see that

$$\begin{aligned}
[w]_{D_A(\frac{\alpha}{4},\infty)} &= \sup_{0<\tau<1} \|\tau^{1-\frac{\alpha}{4}} A e^{\tau A} w\| \\
&\le \sup_{0<\tau<1} \tau^{1-\frac{\alpha}{4}} \int_{\varepsilon/2}^t \|A^3 e^{(t+\tau-r)A} \{\zeta(\cdot,r) - \zeta(\cdot,t)\}\|\, dr.
\end{aligned}$$

We show the estimate only for the term including $\hat{\zeta}(\sigma,t)$. In fact we obtain

$$
\begin{aligned}
&\tau^{1-\frac{\alpha}{4}} \int_{\varepsilon/2}^{t} \|A^3 e^{(t+\tau-r)A}\{\hat{\zeta}(\cdot,r)-\hat{\zeta}(\cdot,t)\}\|\,dr \\
&\le C_0\,\tau^{1-\frac{\alpha}{4}} \int_{\varepsilon/2}^{t} (t+\tau-r)^{\frac{3}{4}-3}\|(\sigma-l_-)^3\eta\|_{D_A(\frac{3}{4},\infty)}|G_2(l_-,t)-G_2(l_-,r)|\,dr \\
&\le C_{0,K}\,\tau^{1-\frac{\alpha}{4}} \int_{\varepsilon/2}^{t} (t+\tau-r)^{\frac{3}{4}-3}\{t^{\frac{1+\alpha}{4}}\cdot(\varepsilon/2)^{-\frac{1}{2}}(t-r)^{\frac{1+\alpha}{4}} \\
&\qquad\qquad + (\varepsilon/2)^{-\frac{1}{2}}(t-r)^{\frac{3+\alpha}{4}}\cdot r^{-\frac{1}{4}} + (\varepsilon/2)^{-\frac{1}{2}}(t-r)^{\frac{2+\alpha}{4}}\}\,dr \\
&\le C_{0,K}\,\tau^{1-\frac{\alpha}{4}} \int_{\varepsilon/2}^{t} (t+\tau-r)^{\frac{\alpha}{4}-2}\,dr\cdot(t^{\frac{1+\alpha}{4}}+t^{\frac{1}{4}})\cdot(\varepsilon/2)^{-\frac{1}{2}} \\
&\quad + C_{0,K}\,\tau^{1-\frac{\alpha}{4}} \int_{\varepsilon/2}^{t} (t+\tau-r)^{\frac{\alpha}{4}-2}\,(r-\varepsilon/2)^{-\frac{1}{4}}\,dr\cdot(t-\varepsilon/2)^{\frac{1}{2}}\cdot(\varepsilon/2)^{-\frac{1}{2}} \\
&\le C_{0,K,\alpha}\,\tau^{1-\frac{\alpha}{4}}\cdot\tau^{\frac{\alpha}{4}-1}\{T^{\frac{1}{4}}+(t-\varepsilon/2)^{\frac{1}{4}}\}\cdot(\varepsilon/2)^{-\frac{1}{2}} \\
&\le C_{0,K,\alpha}\,T^{\frac{1}{4}}\cdot(\varepsilon/2)^{-\frac{1}{2}}.
\end{aligned}
$$

As a consequence, we are led to

$$I_2(t) \le C_{0,K,\alpha}\,T^{\frac{1}{4}}\cdot(\varepsilon/2)^{-\frac{1}{2}}.$$

The estimate of $I_3(t)$ is omitted, since we can readily obtain it by using (B.1) together with the estimate of $|G_2(\cdot,t)-G_2(\cdot,r)|$.

Consequently, we have

$$\|A\dot{z}(t)\|_{D_A(\frac{\alpha}{4},\infty)} \le C_{0,K,\alpha}T^{\frac{1}{4}}\cdot\varepsilon^{-\frac{1}{2}} \quad \text{for} \quad \varepsilon \le t \le T.$$

This completes the proof of the second estimate of (B.4). □

REFERENCES

[1] P. Acquistapace and B. Terreni, *Hölder classes with boundary conditions as interpolation spaces*, Math. Z., 195 (1987), pp. 451–471.

[2] J. W. Barrett, H. Garcke, and R. Nürnberg, *A parametric finite element method for fourth order geometric evolution equations*, J. Comput. Phys., 222 (2007), pp. 441–467.

[3] J. W. Barrett, H. Garcke, and R. Nürnberg, *On the variational approximation of combined second and fourth order geometric evolution equations*, SIAM J. Sci. Comput., 29 (2007), pp. 1006–1041.

[4] L. Bronsard and F. Reitich, *On three-phase boundary motion and the singular limit of a vector-valued Ginzburg-Landau equation*, Arch. Rational Mech. Anal., 124 (1993), pp. 355–379.

[5] J. W. Cahn, C. M. Elliott, and A. Novick-Cohen, *The Cahn-Hilliard equation with a concentration dependent mobility: Motion by minus the Laplacian of the mean curvature*, European J. Appl. Math., 7 (1996), pp. 287–301.

[6] X. Chen, *The Hele-Shaw problem and area-preserving curve-shortening motions*, Arch. Rational Mech. Anal., 123 (1993), pp. 117–151.

[7] F. Davi and M. Gurtin, *On the motion of a phase interface by surface diffusion*, Z. Angew. Math. Phys., 41 (1990), pp. 782–811.

[8] S.-I. Ei, M.-H. Sato, and E. Yanagida, *Stability of stationary interfaces with contact angle in a generalized mean curvature flow*, Amer. J. Math., 118 (1996), pp. 653–687.
[9] S.-I. Ei and E. Yanagida, *Stability of stationary interfaces in a generalized mean curvature flow*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 40 (1993), pp. 651–661.
[10] C. M. Elliott and H. Garcke, *Existence results for diffusive surface motion laws*, Adv. Math. Sci. Appl., 7 (1997), pp. 465–488.
[11] J. Escher, H. Garcke, and K. Ito, *Exponential stability for a mirror-symmetric three phase boundary motion by surface diffusion*, Math. Nachr., 257 (2003), pp. 3–15.
[12] J. Escher, U. F. Mayer, and G. Simonett, *The surface diffusion flow for immersed hypersurfaces*, SIAM J. Math. Anal., 29 (1998), pp. 1419–1433.
[13] H. Garcke, K. Ito, and Y. Kohsaka, *Linearized stability analysis of stationary solutions for surface diffusion with boundary conditions*, SIAM J. Math. Anal., 36 (2005), pp. 1031–1056.
[14] K. Grosse-Brauckmann, *Stable constant mean curvature surfaces minimize area*, Pacific J. Math., 175 (1996), pp. 527–534.
[15] R. Ikota and E. Yanagida, *A stability criterion for stationary curves to the curvature-driven motion with a triple junction*, Differential Integral Equations, 16 (2003), pp. 707–726.
[16] K. Ito and Y. Kohsaka, *Three phase boundary motion by surface diffusion: Stability of a mirror symmetric stationary solution*, Interfaces Free Bound., 3 (2001), pp. 45–80.
[17] K. Ito and Y. Kohsaka, *Three phase boundary motion by surface diffusion in triangular domain*, Adv. Math. Sci. Appl., 11 (2001), pp. 753–779.
[18] A. Lunardi, *Analytic Semigroups and Optimal Regularity in Parabolic Problems*, Birkhäuser, Basel, 1995.
[19] A. Lunardi, E. Sinestrari, and W. von Wahl, *A semigroup approach to the time dependent parabolic initial-boundary value problem*, Differential Integral Equations, 5 (1992), pp. 1275–1306.
[20] W. W. Mullins, *Theory of thermal grooving*, J. Appl. Phys., 28 (1957), pp. 333–339.
[21] H. B. Stewart, *Generation of analytic semigroups by strongly elliptic operators under general boundary conditions*, Trans. Amer. Math. Soc., 259 (1980), pp. 299–310.
[22] J. E. Taylor and J. W. Cahn, *Linking anisotropic sharp and diffuse surface motion laws via gradient flows*, J. Statist. Phys., 77 (1994), pp. 183–197.
[23] T. Vogel, *Sufficient conditions for capillary surfaces to be energy minima*, Pacific J. Math., 194 (2000), pp. 469–489.
[24] E. Zeidler, *Nonlinear Functional Analysis and its Applications* I, Springer-Verlag, New York, 1986.

SIAM J. MATH. ANAL.
Vol. 40, No. 2, pp. 516–539

DECAY ASYMPTOTICS OF THE VISCOUS CAMASSA–HOLM EQUATIONS IN THE PLANE*

CLAYTON BJORLAND†

Abstract. We consider the vorticity formulation of the two-dimensional viscous Camassa–Holm equations in the whole space. We establish global existence for solutions corresponding to initial data in L^1 and describe the large time behavior of solutions with sufficiently small and localized initial data. We calculate the rate at which such solutions approach an "unfiltered" Oseen vortex by computing the rate at which the solution of a scaled vorticity problem approaches the solution to a corresponding linearized equation.

Key words. Camassa–Holm, Navier–Stokes-α, vorticity, decay

AMS subject classifications. 35Q35, 76B03

DOI. 10.1137/070684070

1. Introduction.

The viscous Camassa–Holm equations (VCHE) are commonly written

$$\begin{aligned} v_t + u \cdot \nabla v + v \cdot (\nabla u)^T + \nabla p &= \triangle v, \\ H_\alpha^{-1}(u) = u - \alpha^2 \triangle u &= v, \\ \nabla \cdot u &= 0, \\ v(0) &= v_0. \end{aligned} \tag{1.1}$$

Here H_α is the Helmholtz operator with constant α defined by solving the PDE $u - \alpha^2 \triangle u = v$. This will be referred to as the *filter* associated with the VCHE. For a derivation of these equations from variational principles see [11] or [15]; for a derivation based on modifying the Navier–Stokes system see [7]. The significance of these equations is in the combination of a close relation to the famous Navier–Stokes equations and easily computable bounds on solutions. In particular, in dimensions 2–4, the VCHE admit smooth global solutions which satisfy a modified Kelvin circulation theorem where circulation is conserved around "loops" moving with the filtered flow. The filter is responsible for the smoothing effect on solutions, and the nonlinear term $v \cdot (\nabla u)^T$ brings the solution into compliance with the Kelvin circulation theorem. These solutions are well suited for numerical and analytic calculations and retain many properties displayed by solutions of the Navier–Stokes equations. For proofs of global existence and uniqueness in dimensions 2–4 see [2], [7], [8], [13], and [15]. Decay of energy and higher norms is considered in [2]. Numerical literature is outside the scope of this paper, and the reader is referred to [12] for a survey of the VCHE role in computational turbulence models and a more complete bibliography.

The relation between the VCHE and the Navier–Stokes equations is particularly visible in the vorticity form of the equations found by taking the curl ($\nabla \times v = \tilde{v}$) of

*Received by the editors March 1, 2007; accepted for publication (in revised form) February 7, 2008; published electronically May 28, 2008. This research was partially supported by the NSF under grant OISE-0630623.

http://www.siam.org/journals/sima/40-2/68407.html

†Department of Mathematics, UC Santa Cruz, Santa Cruz, CA 95064 (cbjorland@math.ucsc.edu).

(1.1). In two dimensions the vorticity form is

$$\tilde{v}_t + u \cdot \nabla \tilde{v} = \triangle \tilde{v}, \tag{1.2}$$

$$B(H_\alpha(\tilde{v})) = u,$$

$$\tilde{v}(0) = \tilde{v}_0.$$

In the above equation, B represents convolution with the well-known Biot–Savart kernel which reconstructs the velocity from the vorticity; see (2.1). The aim of this paper is to further explore the relationship between solutions of the Navier–Stokes equation and the VCHE by describing the way a solution of (1.2) approaches the fixed point *zero*, i.e., computing the first and second order decay asymptotics for solutions with small initial data.

Similar study of the asymptotic behavior of the two-dimensional (2-D) vorticity equation for the Navier–Stokes equation can be found in [5], [9], and [10]. In [9] the asymptotics are calculated by applying invariant manifold techniques to the semiflow governing the vorticity problem. Their approach is to scale the vorticity problem into coordinates which are particularly well suited for studying the large time behavior of the Navier–Stokes equation and then apply the invariant manifold theorem in [6] to construct an invariant manifold in the phase space of the scaled problem and foliate the phase space locally, near the fixed point *zero*. The manifolds constructed give insight into the behavior of solutions near the fixed point, and, among other results, the authors calculate the asymptotics through the interaction and properties of these manifolds.

The close relation of the Navier–Stokes equations and the VCHE gives hope that a similar program may be carried out for the 2-D VCHE, especially when comparing the vorticity equations. In fact, such attempts are met with resistance from the filter in the VCHE. In a functional setting the filter eases problems by smoothing the solution, but in a dynamical setting such as this the filter adds complication to the problem. In particular, the filter does not scale well with the other parts of the equation, and the resulting nonlinear term has dependence on the scaled time variable not present in the case of the Navier–Stokes equations. This time dependence carries through into the semigroups generated by the scaled equation and complicates the invariant manifold construction. Specifically, the theorem in [6] cannot be applied. It is possible to add another equation to the system to express it in an autonomous form (does not depend explicitly on the scaled time variable), but the semigroup generated is still dependent on time and does not commute ($\Theta(s+t) \neq \Theta(s)\Theta(t)$). It is not immediately clear how to construct invariant manifolds with such a time dependence or even if such a structure exists.

The invariant manifold theorem in [6] is based on solving Lyapunov–Perron-type equations which are generated through recursive application of the semiflow. To work around the time dependence of this system we construct an infinite family of systems by stepping the original system forward in time a fixed length and work with the corresponding semigroups. These semigroups do not commute but can be composed to reconstruct the flow of the solution. Following in spirit [6] the semiflow generated by these systems is decomposed into a linear term, a nonlinear term for which we can find uniform Lipschitz bounds, and a forcing function which decays sufficiently fast. Applying this decomposition recursively we find a discrete Lyapunov–Perron-type system which is solvable in a rapidly decaying space. The existence of a solution to this discrete system implies decay properties of the difference system which in turn allow

us to compute the asymptotics we desire. Although the notion of invariant manifolds is lost we still retain enough structure to complete the asymptotic calculations. In essence we rework part of [6] with weaker hypotheses reflecting our situation. To calculate the decay of solutions we linearize the scaled VCHE around a fixed point sufficiently close to zero and determined by the initial data considered, subtract the linearized equation from the scaled one to get a system which measures distance from the linear solution, and apply the newly constructed decay theorem. This procedure, in theory, can be continued to arbitrary orders of asymptotics by considering sufficiently localized data. We consider only the cases where the governing ODEs are linear.

The work is based in large part on the work in [6] and [9]; this is reflected in the notation we use and the statements of many theorems. Following this introduction we introduce the majority of our notation and a few useful lemmas relating to the VCHE. Section 3 is dedicated to proving the existence and uniqueness of solutions to the vorticity problem in two dimensions. This section is somewhat based on the work in [1], where similar results were proved for the Navier–Stokes equations, but for us the work is simplified thanks to the smoothing properties of the filter in functional settings. In section 4 we introduce the scaled variables, prove bounds for solutions with these variables in weighted spaces, and discuss other properties of the scaled system. In this section we also discuss the linearized system and the action of the linear operator $\mathcal{L}$ (the scaled form of $\triangle$) on the weighted spaces. Section 5 contains the decay theorems based on the invariant manifold structure. Sections 6 and 7 hold the theorems and computations involving the first and second order asymptotic, respectively. The theorems in section 7 are very similar to section 6, and many are stated with little or no proof when they have nearly identical analogues in section 6. To end this introduction we state the main conclusions of this paper; proofs are contained in sections 3, 6, and 7, respectively. The spaces $L^2(m)$ are weighted spaces; see (2.2) below.

THEOREM 1.1. *Given initial conditions $\tilde{v}_0 \in L^1(\mathbb{R}^2)$, there exists a unique global solution $\tilde{v} \in L^\infty([0,\infty); L^1(\mathbb{R}^2))$ to the PDE* (1.2). *This solution satisfies the decay bound*

$$|\tilde{v}(t)|_q \leq Ct^{-(1-\frac{1}{q})}|\tilde{v}(0)|_1.$$

THEOREM 1.2. *For any $\mu \in (0, \frac{1}{2})$, there exists a $r_0 > 0$ so that for any initial data $\tilde{v}_0 \in L^2(2)$ with $\|\tilde{v}_0\|_2 \leq r_0$ the solution of* (1.2) *satisfies*

$$|\tilde{v}(\cdot,t) - a(\Omega(\cdot,t))|_p \leq C(1+t)^{-1-\mu+\frac{1}{p}},$$

where $a = \int_{\mathbb{R}^2} \tilde{v}_0\, dx$ and

$$\Omega(x,t) = \frac{1}{4\pi(1+t)} e^{\frac{-|x|^2}{4(1+t)}}.$$

THEOREM 1.3. *For any $\mu \in (\frac{1}{2}, 1)$, there exists a $r_0 > 0$ so that for any initial data $\tilde{v}_0 \in L^2(3)$ with $\|\tilde{v}_0\|_3 \leq r_0$ the solution of* (1.2) *satisfies*

$$|\tilde{v}(\cdot,t) - a(\Omega(\cdot,t) - \sum_{i=1,2} b_i(\partial_i \Omega(x,t)))|_p \leq C(1+t)^{-1-\mu+\frac{1}{p}},$$

where $b_i = \int x_i \tilde{v}_0\, dx$ and a, Ω are as in the previous theorem.

2. Notation and preliminaries. Throughout this paper we use $\mathbb{N}$ to refer to the natural numbers including 0. Standard Lebesgue spaces will be denoted $L^p(\mathbb{R}^2)$ or (L^p for short) with the norm $|\cdot|_p = (\int |\cdot|^p\,dx)^{1/p}$. Other Banach spaces defined within will use the norm $\|\cdot\|_X$.

To denote the curl of a vector field we use the tilde. For example, the curl of a vector field v is given by $\nabla \times v = \tilde{v}$. For a divergence-free vector field the curl can be undone through convolution with the well-known Biot–Savart kernel $x^\perp/(2\pi|x|^2)$, $x^\perp = (-x_2, x_1)^T$. We denote this convolution as

$$B(\tilde{w}) = \frac{1}{2\pi}\int_{\mathbb{R}^2} \frac{(x-y)^\perp}{|x-y|^2}\tilde{w}(y)\,dy \tag{2.1}$$

so that $B(\tilde{v}) = v$ for divergence-free vector fields.

We define $\langle\xi\rangle = (1+|\xi|^2)^{1/2}$ and make frequent use of the weighted Hilbert spaces $L^2(m)$ defined by

$$L^2(m) = \{f \in L^2(\mathbb{R}^2)|\ \|f\|_m < \infty \text{ and } \nabla\cdot f = 0\}, \tag{2.2}$$

$$\|f\|_m^2 = \int_{\mathbb{R}^2} \langle\xi\rangle^{2m}(\xi)|f(\xi)|^2\,d\xi.$$

The remaining part of this section contains bounds that will be useful later in the paper. To begin we recall a lemma from [9] concerning the Biot–Savart operator B defined by (2.1) and the curl of divergence-free vector fields.

LEMMA 2.1. *Let*

$$v = B(\tilde{v}) = \frac{1}{2\pi}\int_{\mathbb{R}^2} \frac{(x-y)^\perp}{|x-y|^2}\tilde{v}(y)\,dy.$$

(i) *If* $1 < p < 2 < q < \infty$, $\frac{1}{q} = \frac{1}{p} - \frac{1}{2}$, *and* $\tilde{v} \in L^p(\mathbb{R}^2)$, *then there exists a* $C > 0$ *such that* $|v|_q \le C|\tilde{v}|_p$.

(ii) *If* $1 \le p < 2 < q \le \infty$, $\frac{1}{2} = \frac{\alpha}{p} + \frac{(1-\alpha)}{q}$, *where* $\alpha \in (0,1)$, *and* $\tilde{v} \in L^p \cap L^q(\mathbb{R}^2)$, *then there exists a* $C > 0$ *such that* $|v|_\infty \le C|\tilde{v}|_p^\alpha|\tilde{v}|_q^{1-\alpha}$.

(iii) *If* $1 < p < \infty$ *and* $\tilde{v} \in L^p(\mathbb{R}^2)$, *then there exists a* $C > 0$ *such that* $|\nabla v|_p \le C|\tilde{v}|_p$

Proof. This is Lemma 2.1 in [9]. □

The following bounds for the heat kernel will also be useful in proving the existence of solutions in section 3.

LEMMA 2.2. *Let* Φ *be the fundamental solution to the heat equation. Then*

$$|\Phi(t)|_p \le \frac{C(n,p)}{t^{(1-\frac{1}{p})\frac{n}{2}}},$$

$$|\nabla\Phi(t)|_p \le \frac{C(n,p)}{t^{\frac{1}{2}+(1-\frac{1}{p})\frac{n}{2}}}.$$

Proof. By direct calculation

$$|\Phi(t)|_p^p = \frac{1}{(4\pi t)^{\frac{pn}{2}}}\int_{\mathbb{R}^n} e^{\frac{-p|x|^2}{4t}}\,dx = \frac{1}{p^{\frac{n}{2}}(4\pi t)^{\frac{(p-1)n}{2}}}.$$

This proves the first bound. For the second, start by differentiating the heat kernel and then take the L^p norm

$$|\nabla\Phi(t)|_p^p \leq \frac{C(n,p)}{t^{p(1+\frac{n}{2})}} \int_{\mathbb{R}^n} |x|^p e^{\frac{-p|x|^2}{4t}}\, dx.$$

A change of variables and integration now proves the second bound. □

We also recall some facts about the Helmholtz operator H_α.

LEMMA 2.3. *Let $v \in L^r(\mathbb{R}^n)$, $r \in [1,\infty)$. Then there exists a solution $u \in W^{1,r}(\mathbb{R}^n)$ to the Helmholtz equation*

$$H_\alpha^{-1}(u) = u - \alpha^2 \triangle u = v$$

which satisfies the bounds

$$|u|_p \leq |v|_p \ for\ p \in [1,\infty],$$

$$|u|_p \leq \frac{C(n,p,r)}{\alpha^{1+\gamma}} |v|_r \ for\ \gamma = \frac{n}{2}\left(\frac{1}{r} - \frac{1}{p}\right) < 1,$$

$$|\nabla u|_p \leq \frac{C(n,p,r)}{\alpha^{\frac{3}{2}+\gamma}} |v|_r \ for\ \gamma = \frac{n}{2}\left(\frac{1}{r} - \frac{1}{p}\right) < \frac{1}{2}.$$

Proof. This follows by standard elliptic theory. □

The last lemma in this section concerns the inclusion $L^2(m) \hookrightarrow L^1$ and will be useful in proving many estimates later in this paper.

LEMMA 2.4. *If $f \in L^2(m)$ for some $m > 1$, then $f \in L^q$ for $1 \leq q \leq 2$.*

Proof. The bound $|f|_2 \leq \|f\|_m$ is immediate from the definition of $L^2(m)$. Since $\langle\xi\rangle^{-m}$ is integrable in $\mathbb{R}^2$ if $m > 1$ we have $|f|_1 \leq C\|f\|_m$. Interpolation finishes the proof. □

3. Vorticity problem for the VCHE. This section contains proofs for the existence and uniqueness of solutions to the vorticity form of the VCHE (1.2) with data in $\tilde{v} \in L^1(\mathbb{R}^2)$. Decay of these solutions is provided by the optimal smoothing results in [4].

Instead of working directly with (1.2) we prove existence and uniqueness for the mild form of the PDE by solving the integral equation

$$\tilde{v}(t) = e^{\triangle t}\tilde{v}_0 - \int_0^t \nabla\Phi(t-s) * [B(H_\alpha(\tilde{v})) \otimes \tilde{v}](s)\, ds. \tag{3.1}$$

The corresponding result for the Navier–Stokes equation was proved in [1] using a fixed point argument in a subspace of L^1 and then extending the solution operator to L^1; see also [3] and [14]. This approach was necessary because it is difficult to write the Navier–Stokes equations such that a fixed point argument can be applied in L^1. In the case of the VCHE the filter provides enough leverage to apply a fixed point argument directly. We start by proving a bound on the bilinear term and then follow with the existence of a global solution.

LEMMA 3.1. *The bilinear form*

$$b(\tilde{v},\tilde{w}) : L^\infty([0,T],L^1(\mathbb{R}^2)) \times L^\infty([0,T],L^1(\mathbb{R}^2)) \to L^\infty([0,T],L^1(\mathbb{R}^2))$$

defined by

$$b(\tilde{v},\tilde{w}) = \int_0^t \nabla\Phi(t-s) * [B(H_\alpha(\tilde{v})) \otimes \tilde{w}](s)\, ds$$

satisfies the bound

$$\sup_{t\in[0,T]} |b(\tilde{v},\tilde{w})(t)|_1 \le C(T) \left(\sup_{t\in[0,T]} |\tilde{v}(t)|_1 \right) \left(\sup_{t\in[0,T]} |\tilde{w}(t)|_1 \right),$$

where $C(T) \to 0$ *as* $T \to 0$.

Proof. First apply Young's inequality and then Hölder's inequality to the bilinear form

$$|b(\tilde{v},\tilde{w})(t)|_1 \le \int_0^t |\nabla\Phi(t-s)|_1 |B(H_\alpha(\tilde{v}))(s)|_\infty |\tilde{w}(s)|_1\, ds.$$

Lemmas 2.1 and 2.3 give the bound $|B(H_\alpha(\tilde{v}))(s)|_\infty \le C|\tilde{v}(s)|_1$, so

$$|b(\tilde{v},\tilde{w})(t)|_1 \le C \int_0^t |\nabla\Phi(t-s)|_1\, ds \left(\sup_{s\in[0,t]} |\tilde{v}(t)|_1 \right) \left(\sup_{s\in[0,t]} |\tilde{w}(t)|_1 \right).$$

By Lemma 2.2, $|\nabla\Phi(t-s)|_1 \le C/\sqrt{t-s}$ and the right-hand side is integrable over finite intervals. Taking the supremum over $t \in [0,T]$ yields

$$\sup_{t\in[0,T]} |b(\tilde{v},\tilde{w})(t)|_1 \le CT^{\frac{1}{2}} \left(\sup_{t\in[0,T]} |\tilde{v}(t)|_1 \right) \left(\sup_{t\in[0,T]} |\tilde{w}(t)|_1 \right).$$

This estimate concludes the proof. □

THEOREM 3.2. *Given initial data* $\tilde{v}_0 \in L^1(\mathbb{R}^2)$ *there exists a unique global solution* $\tilde{v} \in L^\infty([0,\infty); L^1(\mathbb{R}^2))$ *to the integral equation* (3.1). *For any* $p \in [1,\infty]$, *this solution satisfies the decay bound*

$$|\tilde{v}(t)|_p \le Ct^{-(1-1/p)} |\tilde{v}(0)|_1. \tag{3.2}$$

Proof. Lemma 3.1 with a standard fixed point argument gives the existence of a mild solution in some possibly small time interval. The length of the time interval depends on the L^1 norm of the initial data. After applying the optimal smoothing results in [4] (see also [16]) the L^1 norm of the solution does not increase beyond the L^1 norm of the data which implies the existence of a global solution. Indeed, the vorticity equation for the VCHE is a viscously damped conservation law, so after applying Theorem 1 in [4] we establish (3.2). In particular, $|v(t)|_1 \le C|v(0)|_1$, which establishes global existence.

It remains to establish uniqueness. Let $\tilde{v}$ and $\tilde{w}$ be two solutions of (3.1) corresponding to the same initial data $\tilde{v}_0 \in L^1(\mathbb{R}^2)$. After adding and subtracting cross terms we see that

$$\begin{aligned}(\tilde{v}-\tilde{w})(t) = &- \int_0^t \nabla\Phi(t-s) * [B(H(\tilde{v},\alpha)) \otimes (\tilde{v}-\tilde{w})](s)\, ds \\ &+ \int_0^t \nabla\Phi(t-s) * [B(H((\tilde{v}-\tilde{w}),\alpha)) \otimes \tilde{w}](s)\, ds.\end{aligned}$$

Similar to the proof of Lemma 3.1, apply Lemmas 2.1 and 2.3 with Young's inequality:

$$|\tilde{v}(t)-\tilde{w}(t)|_1 \le C\int_0^t |\nabla\Phi(t-s)|_1(|\tilde{v}(s)|_1+|\tilde{w}(s)|_1)|\tilde{v}(s)-\tilde{w}(s)|_1\,ds.$$

Apply (3.2) to obtain

$$|\tilde{v}(t)-\tilde{w}(t)|_1 \le C|\tilde{v}_0|_1\int_0^t |\nabla\Phi(t-s)|_1|\tilde{v}(s)-\tilde{w}(s)|_1\,ds.$$

From here the Gronwall inequality with Lemma 2.2 is used to establish uniqueness and conclude the proof. □

This establishes Theorem 1.1.

4. The scaled equations. In this section we introduce scaled variables and rewrite the vorticity equation for the VCHE in these variables, preparing it for use with the theorems in section 5. An existence theorem for the scaled VCHE and related filter equations in the weighted spaces $L^2(m)$ is provided, and we discuss the action of the linear operator $\mathcal{L}$ (scaled Laplacian) on the weighted spaces $L^2(m)$.

The scaled variables are defined as

$$\xi = \frac{x}{\sqrt{1+t}}, \qquad \tau = \ln(1+t),$$

$$v(x,t) = \frac{1}{\sqrt{1+t}}w(\xi,\tau), \qquad u(x,t) = \frac{1}{\sqrt{1+t}}\omega(\xi,\tau),$$

$$\tilde{v}(x,t) = \frac{1}{1+t}\tilde{w}(\xi,\tau), \qquad \tilde{u}(x,t) = \frac{1}{1+t}\tilde{\omega}(\xi,\tau).$$

It has been shown in [5], [9], and [10] that these variables are very useful when studying the large time behavior of the Navier–Stokes equation in vorticity form. Under these variables the vorticity form of the VCHE (1.2) becomes

$$\tilde{w}_\tau = \mathcal{L}\tilde{w} - \omega\cdot\nabla_\xi\tilde{w}, \qquad \tilde{w}(0)=\tilde{w}_0, \tag{4.1}$$

$$\omega = B(\mathcal{H}_{\alpha,\tau}(\tilde{w})), \tag{4.2}$$

where B is again convolution with the Biot–Savart kernel as in (2.1), $\mathcal{L}$ is the linear operator $\mathcal{L}=\triangle_\xi+\frac{1}{2}\xi\cdot\nabla_\xi+I$, and $\mathcal{H}_{\alpha,\tau}$ is the operator defined by solving the scaled Helmholtz equation $\tilde{\omega}-\alpha^2e^{-\tau}\triangle_\xi\tilde{\omega}=\tilde{w}$.

The first goal of the section is to show how the filter $\mathcal{H}_{\alpha,\tau}$ acts on the weighted spaces $L^2(m)$; in particular we will show that the Helmholtz equation has a unique solution in these spaces.

LEMMA 4.1. *If $w\in L^2(m)$ and $\omega\in L^2(m)$ are related by the scaled Helmholtz equation $w=\omega-\alpha^2e^{-\tau}\triangle_\xi\omega$, then $\|\omega(\tau)\|_m^2\le C\|w(\tau)\|_m^2$. In the case $4m^2\alpha^2<1$, $C=1$; otherwise $C=2(1+2^{m-1}(4m^2\alpha^2)^m)$. This lemma shows how the operator $\mathcal{H}_{\alpha,\tau}:L^2(m)\to L^2(m)$ is bounded.*

Proof. The case of $m=0$ follows from the well-known linear elliptic theory; the bound is

$$\|w\|_0^2=\|\omega\|_0^2+2\alpha^2e^{-\tau}\|\nabla\omega\|_0^2+\alpha^4e^{-2\tau}\|\triangle\omega\|_0^2.$$

For $m > 0$ we proceed formally, noting that the following calculations can be applied to a dense set of smooth functions. First square the PDE and then multiply by $\langle\xi\rangle^{2m}$; after integration we have

$$\|w\|_m^2 = \|\omega\|_m^2 + \alpha^4 e^{-2\tau}\|\triangle\omega\|_m^2 - 2\alpha^2 e^{-\tau}\int_{\mathbb{R}^2}\langle\xi\rangle^{2m}\omega\triangle\omega\,d\xi.$$

Using integration by parts

$$\int_{\mathbb{R}^2}(1+|\xi|^2)^m\omega\triangle\omega\,d\xi = 2m\int_{\mathbb{R}^2}(\langle\xi\rangle^{2m-2} + (m-1)\langle\xi\rangle^{2m-4}|\xi|^2)\omega^2\,d\xi - \|\nabla\omega\|_m^2$$

leaves

$$\|w\|_m^2 = \|\omega\|_m^2 + \alpha^4 e^{-2\tau}\|\triangle\omega\|_m^2 + 2\alpha^2 e^{-\tau}\|\nabla\omega\|_m^2 - 4m\alpha^2 e^{-\tau}\|\omega\|_{m-1}^2 - 4m(m-1)\alpha^2 e^{-\tau}\||\xi|\omega\|_{m-2}^2.$$

The bound

$$-e^{-\tau}\langle\xi\rangle^{2m-4}(\langle\xi\rangle^2 + (m-1)|\xi|^2) \geq -2m\langle\xi\rangle^{2m}$$

shows that

$$\|w\|_m^2 \geq (1-4m^2\alpha^2)\|\omega\|_m^2 + \alpha^4 e^{-2\tau}\|\triangle\omega\|_m^2 + 2\alpha^2 e^{-\tau}\|\nabla\omega\|_m^2$$

and proves the result if $4m^2\alpha^2 < 1$.

If $4m^2\alpha^2 \geq 1$, set $\beta^2 = 8m^2\alpha^2 - 1$ so that for $|\xi| \geq \beta$ we have $\langle\xi\rangle^{2m} \geq 8m^2\alpha^2\langle\xi\rangle^{2m-2}$. If $B(\beta)$ is the ball with radius β,

$$-4m^2\alpha^2\int_{\mathbb{R}^2}\langle\xi\rangle^{2m-2}\omega^2\,d\xi \geq -4m^2\alpha^2(8m^2\alpha^2)^{m-1}\int_{B(\beta)}\omega^2\,d\xi - \frac{1}{2}\int_{B^C(\beta)}\langle\xi\rangle^{2m}\omega^2\,d\xi.$$

Applying the case $m = 0$ and the bound $\|\tilde{w}\|_2^2 \leq \|\tilde{w}\|_m^2$ allows

$$-\int_{B(\beta)}\omega^2\,d\xi \geq -\int_{\mathbb{R}^2}\omega^2\,d\xi \geq -\int_{\mathbb{R}^2}w^2\,d\xi \geq -\|w\|_m^2.$$

Also,

$$-\frac{1}{2}\int_{B^C(\beta)}\langle\xi\rangle^{2m}\omega^2\,d\xi \geq -\frac{1}{2}\|\omega\|_m^2.$$

Considering all of this leaves, in the case $4m^2\alpha^2 \geq 1$,

$$C\|w\|_m^2 \geq \frac{1}{2}\|\omega\|_m^2 + \alpha^4 e^{-2\tau}\|\triangle\omega\|_m^2 + 2\alpha^2 e^{-\tau}\|\nabla\omega\|_m^2, \tag{4.3}$$

where $C = 1 + 2^{m-1}(4m^2\alpha^2)^m$. □

The above proof is some indication that the filter is not well suited for the weighted spaces. There is still a smoothing effect, but as can be seen from (4.3) this effect decreases as τ becomes large.

THEOREM 4.2. *Given $w \in L^2(m)$, there exists a unique solution $\omega \in L^2(m)$ to the scaled Helmholtz equations $w = \omega - \alpha^2 e^{-\tau} \triangle_\xi \omega$. This proves that the operator $\mathcal{H}_{\alpha,\tau} : L^2(m) \to L^2(m)$ is well defined.*

Proof. The rough estimate $\langle \xi \rangle^{2m} > 1$ implies $L^2(m) \subset L^2(0)$, so if $w \in L^2(m)$, it is well known that there is a unique $\omega \in L^2(0)$ solving the equation. Lemma 4.1 shows that $\omega \in L^2(m)$. ∎

We now turn our attention to the scaled VCHE. Using the strongly continuous semigroup $e^{\tau \mathcal{L}}$ generated by $\mathcal{L}$ in $L^2(m)$ (see the appendix of [9]) we write the mild form of the scaled vorticity problem:

$$\tilde{w}(\tau) = e^{\tau \mathcal{L}} \tilde{w}_0 - \int_0^\tau e^{-\frac{1}{2}(\tau - s)} \nabla \cdot e^{(\tau - s)\mathcal{L}} (\omega(s) \tilde{w}(s))\, ds, \tag{4.4}$$

$$\tilde{\omega} = \mathcal{H}_{\alpha,\tau}(\tilde{w}).$$

The following lemma, based on Lemma 3.1 in [9], provides an estimate on the bilinear form which will be used to prove the existence of a local solution to (4.4).

LEMMA 4.3. *Given $\tilde{w}_1, \tilde{w}_2 \in C^0([0,T]; L^2(m))$, define*

$$R(\tilde{w}_1, \tilde{w}_2)(\tau) = \int_0^\tau e^{-\frac{1}{2}(\tau - s)} \nabla \cdot e^{(\tau - s)\mathcal{L}} (\omega_1(s) \tilde{w}_2(s))\, ds,$$

where $\omega_1 = w_1 - \alpha^2 e^{-\tau} \triangle w_1$ and w_1 is obtained from $\tilde{w}_1$ via the Biot–Savart law. Then $R \in C^0([0,T], L^2(m))$, and there exists $C_0 = C_0(m,T) > 0$ such that

$$\sup_{0 \le \tau \le T} \|R(\tilde{w}_1, \tilde{w}_2)(\tau)\|_m \le C_0 \left(\sup_{0 \le \tau \le T} \|\tilde{w}_1(\tau)\|_m \right) \left(\sup_{0 \le \tau \le T} \|\tilde{w}_2(\tau)\|_m \right).$$

Moreover, the constant C_0 becomes arbitrarily small as T tends to zero.

Proof. We rely on an estimate of the semigroup $e^{\tau \mathcal{L}}$ proved in the appendix of [9]; if $r \in (\frac{2}{m+1}, 2)$, then

$$|\langle \xi \rangle^m \nabla \cdot e^{\tau \mathcal{L}} u|_2 \le \frac{C}{a(\tau)^{(\frac{1}{r} - \frac{1}{2}) + \frac{1}{2}}} |\langle \xi \rangle^m u|_r,$$

where $a(\tau) = 1 - e^{-\tau}$. This allows

$$|\langle \xi \rangle^m R(\tilde{w}_1, \tilde{w}_2)(\tau)|_2 \le C \int_0^\tau \frac{1}{a(\tau - s)^{\frac{1}{r}}} |\langle \xi \rangle^m \omega_1(s) \tilde{w}_2(s)|_r\, ds.$$

Hölder's inequality, Lemma 2.1, and the inclusion $L^2(m) \hookrightarrow L^r(\mathbb{R}^2)$ provide the bound $|\langle \xi \rangle^m \omega_1(s) \tilde{w}_2(s)|_r \le C \|\tilde{w}_2\|_m \|\tilde{\omega}_1\|_m$. Apply Lemma 4.1 and note that $a(\tau - s)^{-1/r}$ is integrable from 0 to τ to finish the proof. ∎

In addition to providing existence and uniqueness the following theorem shows how we can control the $L^2(m)$ norm of a solution to the integral equation (4.4) by controlling the $L^2(m)$ norm of the initial data. Our proof is modeled after the proof of Theorem 3.2 in [9].

THEOREM 4.4. *Given $\tilde{w}_0 \in L^2(m)$ for some $m > 1$, there exists a global solution $\tilde{w} \in C^0([0,\infty), L^2(m))$ to the integral equation (4.4) with $\tilde{w}(0) = \tilde{w}_0$. Moreover, there exists a constant $C_1 = C_1(\|\tilde{w}_0\|_m)$ such that*

$$\|\tilde{w}(\tau)\|_m \le C_1 \tag{4.5}$$

and $C_1 \to 0$ as $\|\tilde{w}_0\|_m \to 0$.

Proof. The previous lemma and a fixed point argument give local in time existence of a unique solution. Moreover, there exists a $T > 0$ such that

$$\sup_{0\le\tau\le T} \|\tilde{w}(\tau)\|_m \le 2\|\tilde{w}_0\|_m. \tag{4.6}$$

By scaling (3.2) and using the fact that $L^2(m) \hookrightarrow L^1$ for $m > 1$ we see that this solution satisfies, for all $p \in [1,\infty]$,

$$|\tilde{w}(\tau)|_p \le \frac{C_p\|\tilde{w}(0)\|_m}{a(\tau)^{1-\frac{1}{p}}}. \tag{4.7}$$

We will now establish (4.5), which will imply global in time existence. Multiplying (4.1) by $\langle\xi\rangle^{2m}\tilde{w}$ and integrating we find that

$$\frac{1}{2}\frac{d}{d\tau}\int_{\mathbb{R}^2}\langle\xi\rangle^{2m}\tilde{w}^2\,d\xi = \int_{\mathbb{R}^2}\langle\xi\rangle^{2m}NL(\tilde{w})\,d\xi,$$

$$NL(\tilde{w}) = \tilde{w}\triangle\tilde{w} + \frac{\tilde{w}}{2}(\xi\cdot\nabla)\tilde{w} + \tilde{w}^2 - \tilde{w}(\omega\cdot\nabla)\tilde{w}.$$

Integration by parts and the bound $|\xi| \le \langle\xi\rangle$ give the following estimates:

$$\int_{\mathbb{R}^2}\langle\xi\rangle^{2m}\tilde{w}\triangle\tilde{w}\,d\xi \le -\int_{\mathbb{R}^2}\langle\xi\rangle^{2m}\nabla\tilde{w}^2\,d\xi + 2m^2\int_{\mathbb{R}^2}\langle\xi\rangle^{2m-2}\tilde{w}^2\,d\xi,$$

$$\frac{1}{2}\int_{\mathbb{R}^2}\langle\xi\rangle^{2m}\tilde{w}(\xi\cdot\nabla\tilde{w}) = -\frac{1}{2}\int_{\mathbb{R}^2}\langle\xi\rangle^{2m}\tilde{w}^2\,d\xi - \frac{1}{2}m\int_{\mathbb{R}^2}\langle\xi\rangle^{2m-2}|\xi|^2\tilde{w}^2\,d\xi$$

$$-\int_{\mathbb{R}^2}\langle\xi\rangle^{2m}\tilde{w}(\omega\cdot\nabla)\tilde{w}\,d\xi = m\int_{\mathbb{R}^2}\langle\xi\rangle^{2m-2}(\xi\cdot\omega)\tilde{w}^2\,d\xi.$$

Furthermore, given $\epsilon > 0$, there exists a $C_\epsilon > 0$ such that

$$\langle\xi\rangle^{2m-2} \le \epsilon\langle\xi\rangle^{2m} + C_\epsilon$$

and we can bound

$$2m^2\int_{\mathbb{R}^2}\langle\xi\rangle^{2m-2}\tilde{w}^2\,d\xi \le \epsilon\int_{\mathbb{R}^2}\langle\xi\rangle^{2m}\tilde{w}^2\,d\xi + C_\epsilon\int_{\mathbb{R}^2}\tilde{w}^2\,d\xi.$$

Similarly,

$$m\int_{\mathbb{R}^2}\langle\xi\rangle^{2m-2}(\xi\cdot\omega)\tilde{w}^2\,d\xi \le \epsilon\int_{\mathbb{R}^2}\langle\xi\rangle^{2m}\tilde{w}^2\,d\xi + C_\epsilon|\omega|_\infty^{2m}\int_{\mathbb{R}^2}\tilde{w}^2\,d\xi.$$

Putting these bounds together yields

$$\frac{1}{2}\frac{d}{d\tau}\int_{\mathbb{R}^2}\langle\xi\rangle^{2m}\tilde{w}^2\,d\xi \le -\left(\frac{1}{2}-2\epsilon\right)\int_{\mathbb{R}^2}\langle\xi\rangle^{2m}\tilde{w}^2\,d\xi + C_\epsilon(1+|\omega|_\infty^{2m})\int_{\mathbb{R}^2}\tilde{w}^2\,d\xi.$$

Write $\delta = \frac{1}{2} - 4\epsilon$; then

$$\frac{d}{d\tau}\left(e^{\delta\tau}\int_{\mathbb{R}^2}\langle\xi\rangle^{2m}\tilde{w}^2\,d\xi\right) \le C_\epsilon e^{\delta\tau}(1+|\omega|_\infty^{2m})\int_{\mathbb{R}^2}\tilde{w}^2\,d\xi.$$

This inequality implies (4.5), provided

$$\sup_{0\le\tau\le\infty}\left((1+|\omega|_\infty^{2m})\int_{\mathbb{R}^2}\tilde{w}^2\,d\xi\right) \le C(\|\tilde{w}_0\|_m), \tag{4.8}$$

where $C(\|\tilde{w}_0\|_m)\to 0$ as $\|\tilde{w}_m\|_m\to 0$. We now establish this estimate.

When T is such that (4.6) holds, estimate (4.7) with $p=\infty$ and then $p=2$ implies (4.8) when the supremum is taken over $\tau\in[T,\infty)$. When $\tau\le T$, Lemmas 2.1 and 2.3 provide the bound $|\omega(\tau)|_\infty \le C(T)|\tilde{w}(\tau)|_2$; with (4.6) this implies $(1+|\omega(\tau)|_\infty)\le C(T)\|\tilde{w}_0\|_m$ for all $\tau\le T$. In addition (4.6) implies $|\tilde{w}(\tau)|_2\le C(\|\tilde{w}_0\|_m)$ when $\tau\le T$. This concludes the proof. ∎

We spend the remaining portion of this section recalling facts about the operator on the spaces $L^2(m)$ useful to our discussion. The operator $\mathcal{L}$ is studied closely in the appendix of [9], and the reader is referred there for proofs of the statements which are not immediate.

The spectrum of $\mathcal{L}$ in the space $L^2(m)$ is

$$\sigma(\mathcal{L}) = \left\{\lambda\in\mathbb{C}|Re(\lambda)\le\frac{1-m}{2}\right\}\cup\left\{-\frac{k}{2}|k\in\mathbb{N}\right\}.$$

The operator $\mathcal{L}$ generates a strongly continuous semigroup $e^{\tau\mathcal{L}}$ on the space $L^2(m)$. The spectrum of $\mathcal{L}$ acting on $L^2(m)$ has $m-1$ isolated eigenvalues $\lambda_j=\frac{-j}{2}$ for $0\le j\le m-2$. This allows a spectral decomposition that we will use throughout the remaining parts.

Definition 4.5. *Let $X_1^m\subset L^2(m)$ be the finite subspace spanned by the eigenvectors associated with the eigenvalues λ_j, $0\le j\le m-2$, and $X_2^m = L^2(m)-X_1$. Define P_1^m as the spectral projection onto X_1^m and P_2^m the projection onto X_2^m.*

Note that P_1^m and P_2^m are guaranteed to exist because X_1^m and X_2^m are closed subspaces of the Hilbert space $L^2(m)$.

Lemma 4.6. *The projections P_1^m and P_2^m defined above satisfy the following bounds for $\tilde{w}\in L^2(m)$:*

$$\|(e^{\mathcal{L}}P_1^m)^{-j}w\|_m \le C_1 e^{\frac{j(m-2)}{2}}\|w\|_m,$$

$$\|(H_{\alpha,\tau}e^{\mathcal{L}}P_2^m)^j w\|_m \le C_2 e^{\frac{-j(m-1)}{2}}\|w\|_m.$$

Proof. This follows from the definitions of the projections. ∎

Of particular interest to us are the first two eigenvalues and the associated eigenvectors.

DEFINITION 4.7. *The first two eigenvalues of $\mathcal{L}$ acting on $L^2(m)$ are $\lambda_0 = 1$ and $\lambda_1 = \frac{-1}{2}$. For the single eigenvector associated with λ_0 we write $G(\xi) = \frac{1}{4\pi} e^{\frac{-|\xi|}{4}}$. For the two eigenvalues associated with λ_2 we write $F_i = \partial_i G(\xi) = -\frac{\xi}{2} G(\xi)$, $i = 1, 2$.*

The eigenvalue G is the scaled "Oseen vortex" and plays an important role in studying the Navier–Stokes system in two dimensions. It is a stationary solution of the scaled PDE, a fact that is easily checked by finding the associated velocity field (v^G) with the Biot–Savart kernel and computing $v^G \cdot \nabla G = 0$. For reference,

$$v^G(\xi) = \frac{1}{2\pi} \frac{1 - e^{\frac{-|\xi|^2}{4}}}{|\xi|^2} |\xi|^{-2} \begin{pmatrix} -\xi_2 \\ \xi_1 \end{pmatrix}.$$

In the scaled VCHE a similar statement is true when we consider G as the filtered and scaled vorticity; this suggests that in describing the behavior of solutions we will also need to consider the following "unfiltered" eigenvectors.

DEFINITION 4.8.

$$\Gamma(\xi, \tau) = G(\xi) - \alpha^2 e^{-\tau} \triangle G(\xi),$$

$$\Lambda_i(\xi, \tau) = F_i(\xi) - \alpha^2 e^{-\tau} \triangle F_i(\xi).$$

Through straightforward calculations one can check that

$$v^G \cdot \nabla \Gamma = 0, \qquad \partial_\tau \Gamma = \mathcal{L}\Gamma = \alpha^2 e^{-\tau} \triangle \Gamma.$$

It is then clear that Γ is a solution of the VCHE. This solution is stationary from the perspective of the filtered flow ω and plays a similar role as G in the Navier–Stokes equations.

By linearizing the PDE (4.1) about the "fixed" point $a\Gamma$ ($a \in \mathbb{R}$) we obtain

$$\tilde{\psi}_\tau = \mathcal{L}\tilde{\psi} - a\eta \cdot \nabla \Gamma - a v^G \cdot \nabla \tilde{\psi}, \qquad \tilde{\psi}(0) = \tilde{\psi}_0, \tag{4.9}$$

$$\eta = B(\mathcal{H}_{\alpha,\tau}(\tilde{\psi})).$$

This linear PDE has strong global solutions, a fact which can be established following the steps in Theorem 4.4; the proof will be omitted here. Note that $a\Gamma$ is a solution to (4.9).

Letting v^{F_i} denote the velocity field associated with F_i, two other important relations are

$$v^{F_i} \cdot \nabla \Gamma + v^G \cdot \nabla \Lambda_i = 0, \qquad \partial_t \Lambda_i - \mathcal{L}\Lambda_i + \frac{1}{2}\Lambda_i = 0.$$

The first is quickly checked by differentiating the relation $v^G \cdot \nabla \Gamma = 0$, and the second can be checked directly. Together, these show that $a\Gamma + b_1 e^{\frac{-\tau}{2}} \Lambda_1 + b_2 e^{\frac{-\tau}{2}} \Lambda_2$ is a solution to the linearized equation (4.9).

Subtracting (4.9) from (4.1) we find that

$$(\tilde{w} - \tilde{\psi})_\tau = \mathcal{L}(\tilde{w} - \tilde{\psi}) - \omega \cdot \nabla(\tilde{w} - \tilde{\psi}) - (\omega - a v^G) \cdot \nabla \tilde{\psi} + a\eta \cdot \nabla \Gamma. \tag{4.10}$$

This system will be studied in the final sections as the foundation for our asymptotic results.

5. Semigroup theorems. This section contains theorems on collections of semigroups which will be used in the following section to prove our decay results. These theorems are based on the Lyapunov–Perron approach of constructing invariant manifolds in [6] but uses relaxed hypotheses. In particular we allow for the treatment of semigroups which may be time dependent and therefore noncommutative. As our hypotheses are relaxed we are not able to fully construct invariant manifolds. This section instead culminates with a decay theorem.

Unless otherwise specified, in this section let X denote a Hilbert space with norm $\|\cdot\|_X$ and $\{\Theta_n(f,\tau)\}$ a collection of semigroups which map X into itself. That is, each $\Theta_n : X \times \mathbb{R}^+ \to X$ is a semigroup. The idea we will keep in the back of our mind is a noncommutative semigroup $\Theta(\cdot,t)$, where $\Theta_n(\cdot,\tau)$ is the action of the semigroup when $t = n + \tau$. The semigroup Θ can then be reconstructed by composing the semigroups Θ_n. Indeed, if $t = n + \tau$, then

$$\Theta(f,t) = \Theta_n(\Theta_{n-1}(\cdots\Theta_2(\Theta_1(f,1),1),1),\tau).$$

Conversely we make a definition for a natural flow through the collection.

DEFINITION 5.1. *If $\{\Theta_n(f,\tau)\}$ is a collection of semigroups which, for each $\tau \in [0,1]$, map X into itself, we call the function $\Theta : X \times \mathbb{R}^+ \to X$ defined by*

$$\Theta(f,t) = \Theta_n(\Theta_{n-1}(\cdots\Theta_2(\Theta_1(f,1),1),1),\tau),$$

where $t = n + \tau$ and $\tau \in [0,1)$, the natural flow *of the collection through f.*

That $\Theta(f,t)$ is well defined for each f follows from the well-defined properties of each Θ_n.

Throughout this section we assume the following uniform conditions on the collection of semigroups; these conditions are relaxations of (H.1)–(H.4) in [6].

(H.1) Define $Lip(\Theta(\cdot,t)) := \sup_n Lip(\Theta_n(\cdot,t))$. Each $\Theta_n(f,t)$ is continuous in $(f,t) \in X \times [0,1]$ and

$$\sup_{0\le t\le 1} Lip(\Theta(\cdot,t)) = D < \infty.$$

(H.2) Each $\Theta_n(f,1)$ can be decomposed as $\Theta_n(\cdot,1) = L + R_n + S_n$, where $L : X \to X$ is a bounded linear operator, each $R_n : X \to X$ is a global Lipschitz map satisfying $R_n(0) = 0$, and $S_n \in X$.

(H.3) There are subspaces X_i, $i = 1,2$, of X and continuous projections P_i, $i = 1,2$, such that $P_1 + P_2 = I$, $X_1 \oplus X_2 = X$. L leaves X_i invariant and commutes with P_i, $i = 1,2$. Denoting by $L_i : X_i \to X_i$ the restriction of L on X_i, L_1 has bounded inverse and there exist constants $\alpha_1 > \alpha_2 \ge 0$, C_1, and C_2 such that

$$|L_1^{-k}P_1| \le C_1\alpha_1^{-k},$$

$$|L_2^{k}P_1| \le C_1\alpha_1^{k}.$$

We will also write L_i for LP_i when there is no confusion.

(H.4) Define $Lip(R) := \sup_n Lip(R_n)$. Let α_1 and α_2 be as in (H.3); there exists $\alpha_1 > \gamma_1 > \gamma_2 > \alpha_2$ such that the R_n satisfy

$$\frac{C_1}{\alpha_1 - \gamma} + \frac{C_2}{\gamma - \alpha_2} < \frac{1}{Lip(R)}$$

for all $\gamma \in (\gamma_2, \gamma_1)$.

(H.5) With α_1 as in (H.3), S_n satisfies

$$\limsup_{n\to\infty} \frac{1}{n} \ln \|S_n\|_X < \ln \alpha_1.$$

DEFINITION 5.2. *We call a sequence* $\{f_n\}_{n\in\mathbb{N}} \subset X$ *a* discrete positive semiorbit *of* $\{\Theta_n\}$ *through* f_0 *if it satisfies the relation* $f_n = Lf_{n-1} + R_{n-1}(f_{n-1}) + S_{n-1}$ *for all* $n \geq 1$.

For a given initial position $f_0 \in X$ we find the discrete positive semiorbit $\{f_n\}_{n\in\mathbb{N}}$ it generates using this definition recursively:

$$f_n = L^n f_0 + \sum_{j=0}^{n-1} L^{n-j-1}(R_j(f_j) + S_j). \tag{5.1}$$

We now prove a lemma based on Lemma 3.3 from [6] which will provide a link between the semiflows $\{\Theta_n\}$ and a discrete Lyapunov–Perron system.

LEMMA 5.3. *We assume conditions* (H.1)–(H.5) *are satisfied. Let* $\{f_n\}_{n\in\mathbb{N}} \subset X$ *satisfy*

$$\limsup_{n\to\infty} \frac{1}{n} \ln \|f_n\|_X < \ln \alpha_1. \tag{5.2}$$

Then the sequence $\{f_n\}_{n\in\mathbb{N}}$ *is a positive semiorbit of* $\{\Theta_n\}$ *if and only if it satisfies, for all* $n \in \mathbb{N}$,

$$f_n = L_2^n f_0 - \sum_{n\leq j} L_1^{(n-j-1)}(R_j(f_j) + S_j) + \sum_{0\leq j<n} L_2^{(n-j-1)}(R_j(f_j) + S_j). \tag{5.3}$$

Proof. Using the iterative relation (5.1) with the projection P_2 shows that

$$P_2 f_n = L_2^n f_0 + \sum_{0\leq j<n} L_2^{(n-j-1)}(R_j(f_j) + S_j).$$

Likewise, if $m > n$, we can write (5.1) in the form

$$f_m = L^{n-m} f_n + \sum_{n\leq j<m} L_1^{(m-j-1)}(R_j(f_j) + S_j)$$

so that

$$P_1 f_n = L_1^{n-m} f_m - \sum_{n\leq j<m} L_1^{(n-j-1)}(R_j(f_j) + S_j).$$

We will show that the right-hand side of the above equation converges as $m \to \infty$; then it remains only to add these projections together to finish the proof. Condition (H.3) and the Lipschitz property of R_n allow

$$\left\| \sum_{n\leq j<m} L_1^{(n-j-1)}(R_j(f_j) + S_j) \right\|_X \leq C_1 \sum_{n\leq j<m} \alpha_1^{n-j-1}(Lip(R)\|f_j\|_X + \|S_j\|_X).$$

The bound (5.2) and assumption (H.5) give convergence of this sum as $m \to \infty$ as well as the following limit:

$$\limsup_{m\to\infty} \|L_1^{n-m} f_m\|_X \le C_1 \alpha_1^n \limsup_{m\to\infty} (\alpha_1^{-m} \|f_m\|_X)$$
$$= 0. \qquad \square$$

The next step is to show that for any initial data f_0 there exists a unique solution to the system (5.3) in a weighted space where all elements satisfy (5.2); then through the previous lemma we can deduce that this solution is a discrete positive semiorbit. Systems such as (5.3) are called Lyapunov–Perron equations and used in the Lyapunov–Perron approach to invariant manifolds. The existence theorem we prove uses a fixed point argument and is similar to Theorem 2.1 in [6].

DEFINITION 5.4. *Fix $\gamma \in (\gamma_2, \gamma_1)$, where γ_1 and γ_2 are as in* (H.4)*, and let E_n^γ be the Banach space equal to X as a vector space but equipped with the norm $\|\cdot\|_{E_n^\mu} = \gamma^{-n}\|\cdot\|_X$. E^γ is the sequence space $f = \{f_n\}_{n\in\mathbb{N}}$, $f_n \in E_n^\gamma$ equipped with the norm $\|f\|_{E^\gamma} = \sup_{n\in\mathbb{N}} \|f_n\|_{E_n^\gamma}$.*

As $\gamma_1 < \alpha_1$ by assumption this definition is compatible with (5.2).

THEOREM 5.5. *We assume conditions* (H.1)–(H.5) *are satisfied. Pick $\gamma \in (\gamma_2, \gamma_1)$. Given initial data $x \in X_2$, there exists a unique solution $\{f_n\} \in E^\mu$ to* (5.3).

Proof. Define $J : E^\gamma \to E^\gamma$ componentwise as

$$J_n(\{f_n\}) = L_2^n x + \sum_{0\le j<n} L_2^{(n-1-j)}(R_j(f_j) + S_j) - \sum_{n\le j} L_1^{(n-1-j)}(R_j(f_j) + S_j).$$

To apply a fixed point theorem we need to check that J is well defined and is a contraction map. In that direction we start with the bound

$$\|J_n(\{f_n\})\|_{E_n^\gamma} \le \|L_2^n x\|_{E_n^\gamma} + \sum_{0\le j<n} \|L_2^{(n-1-j)}(R_j(f_j) + S_j)\|_{E_n^\gamma}$$
$$- \sum_{n\le j} \|L_1^{(n-1-j)}(R_j(f_j) + S_j)\|_{E_n^\gamma}.$$

The first term on the right-hand side is bounded with the help of (H.3), the definition of the space E_n^γ, and the condition $\gamma > \alpha_2$:

$$\|L_2^n x\|_{E_n^\gamma} \le C_2 \alpha_2^n \gamma^{-n} \|x\|_X$$
$$\le C_2 \|x\|_X.$$

Assumption (H.5) and $\gamma < \alpha_1$ imply $\{S_n\} \in E^\gamma$; then the second term is bounded using the same ideas but adding the uniform Lipschitz constant of R_j:

$$\sum_{0\le j<n} \|L_2^{(n-1-j)}(R_j(f_j) + S_j)\|_{E_n^\gamma} \le C_2 \alpha_2^{-1} \sum_{0\le j<n} (\alpha_2/\gamma)^{n-j} \|(R_j(f_j) + S_j)\|_{E_j^\gamma}$$
$$\le \frac{C_2}{\gamma - \alpha_2} \left(Lip(R) \|\{f_n\}\|_{E^\gamma} + \|\{S_n\}\|_{E^\gamma} \right).$$

Similarly,

$$\sum_{n\le j} \|L_1^{(n-1-j)}(R_j(f_j) + S_j)\|_{E_n^\gamma} \le \frac{C_1}{\alpha_1 - \gamma} \left(Lip(R) \|\{f_n\}\|_{E^\gamma} + \|\{S_n\}\|_{E^\gamma} \right).$$

These bounds show that $\|J_n(\{f_n\})\|_{E_n^\gamma}$ is bounded independent of n and therefore J maps E^γ into itself. Given another sequence $g_n \in E^\gamma$, following nearly the same steps one finds the bound

$$\|J_n(\{f_n\}) - J_n(\{g_n\})\|_{E_n^\gamma} \leq \left(\frac{C_1}{\alpha_1 - \gamma} + \frac{C_2}{\gamma - \alpha_2}\right) Lip(R)\|\{f_n - g_n\}\|_{E^\gamma}.$$

Assumption (H.4) guarantees J is a contraction map, and a standard fixed point argument finishes the theorem. ∎

We would like to remark here that the system (5.3) does not "see" the component of the initial data in X_1 but instead picks out the correct component to form a solution which satisfies (5.2). Thus, by solving the system one creates a map $h : X_2 \to X_1$ defined $h(x) = P_1 f_0$ such that the *natural flow* through $x + h(x)$ decays rapidly. This map is important when constructing invariant manifolds and foliations through the Lyapunov–Perron approach (see, for example, [6]). This next corollary makes this remark precise.

COROLLARY 5.6. *Let $\{\Theta_n\}$ be a collection of semigroups satisfying* (H.1)–(H.5). *For each $x \in X_2$ there exists $h(x) \in X_1$ such that the discrete positive orbit through $x + h(x)$ satisfies* (5.2). *This defines a map $h : X_2 \to X_1$.*

Proof. Define h by using x as initial conditions in the previous theorem to find a sequence $\{f_n\}$; then let $h(x) = P_1 f_0$. As the solution is unique so h is well defined, the remaining properties follow quickly from Lemma 5.3. ∎

We are now in a position to prove the theorem, which is the goal of this section and will be the basis for decay estimates in the following sections.

THEOREM 5.7. *Let $\{\Theta_n\}$ be a collection of semigroups satisfying* (H.1)–(H.5) *and h as in the previous corollary. For each $\gamma \in (\gamma_2, \gamma_1)$ and initial condition $x \in X_2$ denote by $\Theta(x + h(x), t)$ the* natural flow *of $\{\Theta_n\}$ through $x + h(x)$. This satisfies*

$$\limsup_{t\to\infty} \frac{1}{t} \ln \|\Theta(x + h(x), t)\|_X < \gamma. \tag{5.4}$$

Proof. $\{\Theta(x+h(x), n)\}$ is a discrete positive orbit of $\{\Theta_n\}$ through $x+h(x)$. Using Theorem 5.5 and its corollary we obtain

$$\limsup_{n\to\infty} \frac{1}{n} \ln \|\Theta(x + h(x), n)\|_X < \gamma.$$

We now apply the Lipschitz property of the semiflows Θ_n given by (H.1). If $n \in \mathbb{N}$ and $0 \leq \sigma < 1$ are such that $t = n + \sigma$,

$$\begin{aligned}\frac{1}{t} \ln \|\Theta(x + h(x), t)\|_X &\leq \frac{1}{t} |\ln(D\|\Theta(x + h(x), n)\|_X) \\ &\leq \frac{1}{n} \ln \|\Theta(x + h(x), n)\|_X + \frac{1}{t} \ln D.\end{aligned}$$

Taking the limit superior first as $n \to \infty$ and then as $t \to \infty$ in the above expression finishes the proof. ∎

6. First order asymptotic behavior. For the first order calculations we work in the space $L^2(2)$ where the operator $\mathcal{L}$ has a single isolated eigenvalue 0 and corresponding eigenvector G (see Definition 4.7). This eigenvector spans the subspace X_1 as in Definition 4.5. The projection onto this subspace is defined as follows: let

$a = \int \tilde{w}_0 \, d\xi$; then $P_1(\tilde{w}_0) = aG$. Before we proceed it is important to remark that $\int \tilde{w} \, d\xi$ is a conserved quantity under the flow of (4.9), so $P_1(\tilde{w}(\tau)) = aG$. Indeed, since $\mathcal{L}\tilde{w} = \nabla \cdot (\nabla \tilde{w} + \frac{\xi}{2})$,

$$\frac{1}{2}\frac{d}{dt}\int \tilde{w} \, d\xi = \int \nabla \cdot \left(\nabla \tilde{w} + \frac{\xi}{2} \cdot \tilde{w} - \omega \tilde{w} \right) d\xi = 0. \tag{6.1}$$

Fix $\tilde{w}_0$ (we keep this fixed throughout the calculations) and consider the system (4.10) with initial data $\tilde{w}_0 - a\Gamma(\cdot, 0)$, where $a = \int \tilde{w}_0 \, d\xi$; later we will require $\|\tilde{w}_0\|_m$ to be sufficiently small. With these initial conditions $\tilde{\psi} = a\Gamma$ is a "stationary" solution of the linear system (4.9). Writing $\tilde{f} = \tilde{w} - a\Gamma$, $\phi = B\mathcal{H}_{\alpha,\tau}(\tilde{f})$, (4.10) becomes

$$\tilde{f}_\tau = \mathcal{L}\tilde{f} - \omega \cdot \nabla \tilde{f} - a\phi \cdot \nabla \Gamma. \tag{6.2}$$

The associated integral equation is

$$\tilde{f} = e^{\tau\mathcal{L}} P_2(\tilde{w}_0) - \int_0^\tau e^{\frac{-1}{2}(\tau-\sigma)\nabla} \cdot e^{(\tau-\sigma)\mathcal{L}} (\omega \tilde{f} + a\phi\Gamma)(\sigma) \, d\sigma. \tag{6.3}$$

Although it is not obvious at a quick glance, the above system changes with τ. The dependence on τ is buried in the ϕ term (in the filter relation) and destroys the commutative property of the generated semiflow. Commutativity is a very useful property when dealing with semiflows, but we are able to proceed in a fashion consistent with the Lyapunov–Perron approach to constructing invariant manifolds using the results of section 5. To make this work we need a collection of semiflows with uniform bounds which can be put together to reconstruct the original flow. The following system takes us in that direction, and we now change our focus to finding properties of $\tilde{f}_n$ which solve the following system, defined for all $n \in \mathbb{N}$:

$$\tilde{f}_n = e^{\tau\mathcal{L}} \tilde{f}_{n,0} - \int_0^\tau e^{\frac{-1}{2}(\tau-\sigma)\nabla} \cdot e^{(\tau-\sigma)\mathcal{L}} (\omega(n+\sigma)\tilde{f}_n(\sigma) + a\phi_m(\sigma)\Gamma(n+\sigma)) \, d\sigma, \tag{6.4}$$

$$\phi_n(\sigma) = B\mathcal{H}_{\alpha,\sigma+n}(\tilde{f}_n)(\sigma).$$

Note that this system has a global solution; existence can be proved analogously to Theorem 4.4. It should also be noted here that the above system, (6.2), (6.3), and therefore the flow Θ_n depend on $\tilde{w}_0$, which we consider fixed when we write down the system (a depends on $\tilde{w}_0$); this relation is suppressed in the notation. Most solutions of this equation will have no meaning, but by choosing $\tilde{f}_{0,0} = P_2(\tilde{w}_0)$ and then $\tilde{f}_{n,0} = \tilde{f}_{n-1}(1)$ we are able to recover information about the semiflow corresponding to $\tilde{w}_0$.

DEFINITION 6.1. *Throughout the remainder of this section let $\Theta_n(\tilde{f}_{n,0}, \tau)$ denote the global solution to the system* (6.4) *with initial data $\tilde{f}_{n,0}$.*

The parameter n allows us to keep track of our progress in time. For example, if $0 \le \sigma < 1$ is such that $\tau = n + \sigma$ and $\Theta(\tilde{f}_0, \tau)$ is the semiflow of (6.3), then

$$\Theta(\tilde{f}_0, \tau) = \Theta_n(\Theta_{n-1}(\Theta_{n-2}(\cdots \Theta_0(\tilde{f}_0, 1), 1), 1), \sigma). \tag{6.5}$$

We will now prove uniform properties of these semiflows.

LEMMA 6.2. *The semiflows $\Theta_n(\tilde{f}_0, \tau)$, $n \in \mathbb{N}$, are all C^1 in $L^2(2) \times \mathbb{R}^+$. There exist a constant $r_0 > 0$ (possibly small) and $D > 0$ such that for all $\|\tilde{w}_0\|_2 < r_0$ and $n \in \mathbb{N}$ the flow Θ_n satisfies the following Lipschitz property:*

$$\sup_{n\in\mathbb{N}} \sup_{0\le\tau<1} Lip(\Theta_n(\cdot, \tau)) = D < \infty. \tag{6.6}$$

This bound holds as $r_0 \to 0$.

Proof. Since the bound obtained in Lemma 4.1 is independent of τ we can prove this lemma for all $n \in \mathbb{N}$ at once. That the semiflow is C^1 is a classical result. Consider the semiflows $\tilde{f}(\tau) = \Theta_n(\tilde{f}_0, \tau)$ and $\tilde{g}(\tau) = \Theta_n(\tilde{g}_0, \tau)$ found from initial data $\tilde{f}_0$ and $\tilde{g}_0$, respectively, and the corresponding filtered flows ϕ and γ. Subtracting we have

$$\|\Theta_n(\tilde{f}, \tau) - \Theta_n(\tilde{g}, \tau)\|_2 \leq \|e^{\tau\mathcal{L}}(\tilde{f}_0 - \tilde{g}_0)\|_2 + I(\tau) + J(\tau),$$

where

$$I(\tau) = \left\| \int_0^\tau e^{-\frac{1}{2}(\tau-\sigma)} \nabla \cdot e^{(\tau-\sigma)\mathcal{L}} (\omega(n+\sigma)(\tilde{f}(\sigma) - \tilde{g}(\sigma)))\, d\sigma \right\|_2,$$

$$J(\tau) = \left\| \int_0^\tau e^{-\frac{1}{2}(\tau-\sigma)} \nabla \cdot e^{(\tau-\sigma)\mathcal{L}} ((\phi(\sigma) - \gamma(s)) a\Gamma(n+\sigma))\, ds \right\|_2.$$

Similar to the steps in Lemma 4.3 we obtain

$$I(\tau) + J(\tau) \leq C(\tau) \left(\sup_{0\leq\sigma<\tau} \|\tilde{f}(\sigma) - \tilde{g}(\sigma)\|_2 \right) \cdot \left(\sup_{0\leq\sigma<\tau} (\|\tilde{w}(n+\sigma)\|_m + a\|\Gamma(n+\sigma)\|_2) \right)$$

with $C(\tau)$ a continuous function and hence bounded on $\tau \in [0,1]$ by a constant C. The bound (4.5) combined with the definition of a ($a \leq \|\tilde{w}_0\|_2$) allows us to pick $r_0 > 0$ small enough so that, for all $\sigma > 0$,

$$\|\tilde{w}(n+\sigma)\|_2 + a\|\Gamma(n+\sigma)\|_2 < \frac{1}{2C}.$$

After taking the supremum over $\tau \in [0,1]$ we have

$$\sup_{0\leq\tau<1} \|\Theta_n(\tilde{f}, \tau) - \Theta_n(\tilde{g}, \tau)\|_2 \leq \sup_{0\leq\tau<1} \|e^{\tau\mathcal{L}}(\tilde{f}_0 - \tilde{g}_0)\|_2 \leq D\|\tilde{f}_0 - \tilde{g}_0\|_2.$$

This is the bound (6.6). □

LEMMA 6.3. *There exists a constant $r_0 > 0$ (possibly small) such that for all $\|\tilde{w}_0\|_2 < r_0$ and $n \in \mathbb{N}$ the flow Θ_n can be decomposed as $\Theta_n(\tilde{f}_0, 1) = e^{\mathcal{L}}\tilde{f}_0 + R_n(\tilde{f}_0)$, where $R_n(\cdot)$ is Lipschitz as a function from $L^2(2)$ to itself. The uniform Lipschitz constant $Lip(R) := \sup_{n\in\mathbb{N}} Lip(R_n(\cdot))$ satisfies the following conditions:*

(i) *For any $\mu \in (0, \frac{1}{2})$, r_0 may be chosen so that, for all $i \in \mathbb{N}$,*

$$\frac{C_1}{1 - e^{-\mu}} + \frac{C_2}{e^{-\mu} - e^{-1/2}} < \frac{1}{Lip(R)}. \tag{6.7}$$

(ii) *This bound holds as $r_0 \to 0$.*

Proof. As in the previous proof, consider the semiflows $\tilde{f}(\tau) = \Theta_n(\tilde{f}_0, \tau)$ and $\tilde{g}(\tau) = \Theta_n(\tilde{g}_0, \tau)$. Define

$$R_n(\tilde{f}_0) = \Theta_n(\tilde{f}_0, 1) - e^{\mathcal{L}}\tilde{f}_0 = -\int_0^1 e^{-1/2(1-\sigma)} \nabla \cdot e^{(1-\sigma)\mathcal{L}} (\omega(n+\sigma)\tilde{f}(\sigma) + a\phi(\sigma)\Gamma(n+\sigma))\, d\sigma.$$

As in Lemmas 4.3 or 6.2, after adding and subtracting cross terms,

$$\|R_n(\tilde{f}_0) - R_n(\tilde{g}_0)\|_2 \leq C \left(\sup_{0\leq\sigma<\tau} \|\tilde{f}(\sigma) - \tilde{g}(\sigma)\|_2 \right) \cdot \left(\sup_{0\leq\sigma<\tau} (\|\tilde{w}(n+\sigma)\|_2 + a\|\Gamma(n+\sigma)\|_2) \right).$$

Appealing to (6.6),

$$\|R_n(\tilde{f}_0) - R_n(\tilde{g}_0)\|_2 \leq CD \left(\sup_{0\leq\sigma<\tau} \|\tilde{f}_0 - \tilde{g}_0\|_2 \right) \cdot \left(\sup_{0\leq\sigma<\tau} (\|\tilde{w}(n+\sigma)\|_2 + a\|\Gamma(n+\sigma)\|_m) \right).$$

This shows that

$$Lip(R_i(\cdot)) \leq CD \left(\sup_{0\leq\sigma<\infty} (\|\tilde{w}(\sigma)\|_2 + a\|\Gamma(\sigma)\|_2) \right)$$

and, with the help of (4.5), satisfies the Lipschitz condition. The bound (6.7) is established by taking r_0 sufficiently small. □

For the remainder of this section we assume that r_0 is small enough so the conclusions of Lemma 6.2 and 6.3 hold.

THEOREM 6.4. *Pick $\mu \in (0, \frac{1}{2})$ and choose $r_0 > 0$ to satisfy the conclusions of Lemmas 6.2 and 6.3. Let P_i^2 and X_i^2, $i = 1, 2$, be as in Definition 4.5. Given $\tilde{w}_0$ such that $\|\tilde{w}_0\|_2 \leq r_0$ and initial data $\tilde{f}_0 \in X_2$, there exists a unique global solution $\Theta(\tilde{f}_0, \tau) \in C^0([0,\infty), L^2(2))$ of (6.3). This solution satisfies $P_1\Theta(\tilde{f}_0, \tau) = 0$ and*

$$\limsup_{\tau\to\infty} \frac{1}{t} \ln \|\Theta(\tilde{f}_0, \tau)\|_2 < -\mu. \tag{6.8}$$

Proof. The existence of a unique global solution can be argued as in Theorem 4.4; again let $\Theta_n(\tilde{f}_{n,0}, \tau)$ denote the global solution with initial data $\tilde{f}_{n,0}$. We now check the assumptions for Theorem 5.7. Lemma 6.2 shows that the collection $\{\Theta_n\}$ satisfies assumption (H.1) from section 6. Assumption (H.3) is satisfied by Lemma 4.6. As the left-hand side of (6.7) is continuous and the inequality is strict we can find a small neighborhood of μ, say $(\gamma_2, \gamma_1) \subset (0, \frac{1}{2})$, such that the inequality is satisfied for all numbers in this neighborhood. Taking $\alpha_1 = \frac{1}{2}$ and $\alpha_2 = 0$ and applying Lemma 6.3 shows that the collection $\{\Theta_n\}$ satisfies assumptions (H.2), (H.4), and (H.5) with $S_n = 0$.

Applying Theorem 5.7 gives the existence of an element $h(\tilde{f}_0) \in X_1$ such that

$$\limsup_{n\to\infty} \frac{1}{\tau} \ln \|\Theta(y, \tau)\|_2 < -\mu,$$

where $y = \tilde{f}_0 + h(\tilde{f}_0)$ and Θ is constructed as in (6.5); in other words it is the solution of (6.3). To finish the proof we need only argue that $h(\tilde{f}_0) = 0$, which can be inferred from the "conservation of mass" property of (6.2) and the decay implied by the above inequality. Indeed, using (6.2) we see that $\int \tilde{f}\, d\xi$ is a conserved property:

$$\frac{1}{2}\frac{d}{dt} \int \tilde{f}\, d\xi = \int \nabla \cdot (\nabla \tilde{f} + \xi \tilde{f} - \omega \tilde{f} - a\phi\Gamma)\, d\xi = 0. \tag{6.9}$$

The orthogonal relation $X_1^2 \perp X_2^2$ allows (again writing $y = \tilde{f}_0 + h(\tilde{f}_0)$)

$$\|P_1^2\Theta(y,\tau)\|_2 \le \|\Theta(y,\tau)\|_2$$

and, as $\|P_1\Theta(y,\tau)\|_2$, is constant for all $\tau > 0$; (6.9) implies that $P_1\Theta(y,\tau) = h(\tilde{f}_0) = 0$. ∎

We now prove the theorem, which is the goal of this section.

THEOREM 6.5. *Pick $\mu \in (0, \frac{1}{2})$ and choose $r_0 > 0$ to satisfy the conclusions of Lemmas* 6.2 *and* 6.3. *Given initial data $\tilde{w}_0$ such that $\|\tilde{w}_0\|_2 \le r_0$, the solution $\tilde{w}(\tau)$ of the scaled VCHE given by Theorem* 4.4 *is subject to the following decay estimate:*

$$\|\tilde{w}(\tau) - a\Gamma(\tau)\|_2 \le Ce^{-\mu\tau},$$

where $a = \int \tilde{w}_0\, d\xi$.

Proof. In the previous theorem take $\tilde{f}_0 = P_2\tilde{w}_0$; then $\tilde{f}(\tau) = \tilde{w}(\tau) - a\Gamma(\tau)$. The decay follows from (6.8). ∎

Finally, as a corollary to this theorem we prove the result listed as Theorem 1.2 in the introduction.

COROLLARY 6.6. *For any $\mu \in (0, \frac{1}{2})$, there exists a $r_0 > 0$ so that, for any initial data $\tilde{v}_0 \in L^2(2)$ such that $\|\tilde{v}_0\|_2 \le r_0$, the solution of* (1.2) *given by Theorem* 3.2 *satisfies*

$$|\tilde{v}(\cdot,t) - a(\Omega(\cdot,t) - \alpha^2\triangle\Omega(\cdot,t))|_p \le C(1+t)^{-1-\mu+\frac{1}{p}},$$

where $a = \int_{\mathbb{R}^2} \tilde{v}_0\, dx$ and

$$\Omega(x,t) = \frac{1}{4\pi(1+t)} e^{\frac{-|x|^2}{4(1+t)}}.$$

Proof. This is the result of the previous theorem in unscaled coordinates. Let

$$\Omega(x,t) = \frac{1}{(1+t)} G\left(\frac{x}{\sqrt{1+t}}\right) = \frac{1}{4\pi(1+t)} e^{\frac{-|x|^2}{4(1+t)}}; \tag{6.10}$$

then

$$\frac{1}{(1+t)}\Gamma\left(\frac{x}{\sqrt{1+t}}, \ln(1+t)\right) = \Omega(x,t) - \alpha^2\triangle\Omega(x,t).$$

Thanks to the above theorem and the inclusion $L^2(2) \hookrightarrow L^p$ for when $1 \le p \le 2$,

$$\begin{aligned}
&|\tilde{v}(\cdot,t) - a(\Omega(\cdot,t) - \alpha^2\triangle\Omega(\cdot,t))|_p \\
&\qquad \le (1+t)^{-1+\frac{1}{p}} |\tilde{w}(\cdot, \ln(1+t)) - a\Gamma(\cdot, \ln(1+t))|_p \\
&\qquad \le C(1+t)^{-1+\frac{1}{p}} \|\tilde{w}(\cdot, \ln(1+t)) - a\Gamma(\cdot, \ln(1+t))\|_2 \\
&\qquad \le C(1+t)^{-1-\mu+\frac{1}{p}}. \qquad ∎
\end{aligned}$$

As $\triangle\Omega$ decays faster than $(1+t)^{-1-\mu+1/p}$ we can include it on the right-hand side; this is how Theorem 1.2 is stated.

7. Second order asymptotic. This section is similar in spirit to the previous section, and many of the proofs are omitted because they are nearly the same as the proofs given in the previous section. We work in the space $L^2(3)$ where the operator $\mathcal{L}$ has two isolated eigenvalues, 0 and $\frac{-1}{2}$, and three corresponding eigenvectors, G and F_i, $i = 1, 2$ (see Definition 4.7). Together these eigenvalues span the subspace X_1^3 given by Definition 4.5. Any $\tilde{w}_0 \in L^2(3)$ can be written as $\tilde{w}_0 = aG + b_1F_1 + b_2F_2 + \tilde{g}$, where $a = \int \tilde{w}_0\, d\xi$, $b_i = \int \xi_i \tilde{w}_0\, d\xi$, and $\tilde{g} \in X_2$.

In addition to the "conservation of mass" property $\frac{d}{dt}\int \tilde{w}\, d\xi = 0$ used in the previous section (proof of Theorem 5.7), solutions to the scaled VCHE (4.1) also satisfy the scaled form of conservation of the first moments, $\frac{d}{dt}\int \xi_i \tilde{w}\, d\xi = -\frac{1}{2}\int \xi_i \tilde{w}\, d\xi$. Indeed, let $i = 1, 2$ and $j \neq i$; then

$$\xi_i \mathcal{L}\tilde{w} + \frac{1}{2}\xi_i \tilde{w} = \partial_j \left(\xi_i \partial_i \tilde{w} + \frac{1}{2}\xi_i^2 \tilde{w} - \tilde{w} \right) + \partial_j \left(\xi_i \partial_j \tilde{w} + \frac{1}{2}\xi_i \xi_j \tilde{w} \right),$$

and it is clear that $\int_{\mathbb{R}^2} \xi_i \mathcal{L}\tilde{w}\, d\xi = -\frac{1}{2}\int_{\mathbb{R}^2} \xi_i \tilde{w}\, d\xi$. For the nonlinear term, note that if $\nabla \cdot \omega = 0$ and $\tilde{\omega} = \partial_1\omega_2 - \partial_2\omega_1$, then

$$\xi_i \omega \cdot \nabla \tilde{\omega} = \partial_i(\xi_i \omega_i \tilde{\omega} - \omega_i \omega_j) + \partial_j \left(\xi_i \omega_j \tilde{\omega} + \frac{1}{2}(\omega_i^2 - \omega_j^2) \right).$$

Similarly,

$$\begin{aligned} \xi_i \omega \cdot \nabla \partial_i^2 \tilde{\omega} &= \partial_i(\xi_i \omega \cdot \nabla \partial_j \tilde{\omega}) - \omega \cdot \nabla \partial_i \tilde{\omega} - \xi_i \partial_i \omega \cdot \nabla \partial_i \tilde{\omega} \\ &= \partial_i(\xi_i \omega \cdot \nabla \partial_j \tilde{\omega}) - \nabla(\omega \partial_i \tilde{\omega}) - \partial_i(\xi_i \partial_i \omega_i \partial_i \tilde{\omega} - \partial_i \omega_i \partial_i \omega_j) \\ &\quad - \partial_j \left(\xi_i \partial_i \omega_j \partial_i \tilde{\omega} + \frac{1}{2}((\partial_i \omega_i)^2 - (\partial_i \omega_j)^2) \right) \end{aligned}$$

and

$$\begin{aligned} \xi_i \omega \cdot \nabla \partial_j^2 \tilde{\omega} &= \partial_j(\xi_i \omega \cdot \nabla \partial_j \tilde{\omega}) - \partial_i(\xi_i \partial_j \omega_i \partial_j \tilde{\omega} - \partial_j \omega_i \partial_j \omega_j) \\ &\quad - \partial_j \left(\xi_i \partial_j \omega_j \partial_j \tilde{\omega} + \frac{1}{2}((\partial_j \omega_i)^2 - (\partial_j \omega_j)^2) \right) \end{aligned}$$

so that $\int_{\mathbb{R}^2} \xi_i \omega \cdot \nabla \tilde{w}\, d\xi = 0$.

Fix $\tilde{w}_0 \in L^2(m)$ and consider the system (4.10) with initial data $P_2^m \tilde{w}_0$. Let $a = \int \tilde{w}_0\, d\xi$, $b_i = \int \xi_i \tilde{w}_0\, d\xi$; then $\tilde{\psi}(\tau) = a\Gamma + e^{\frac{-1}{2}}(b_1\Lambda_1 + b_2\Lambda_2)$ is the solution to the linear equation (4.9). Write $\tilde{f} = \tilde{w} - a\Gamma - e^{\frac{-1}{2}}(b_1\Lambda_1 + b_2\Lambda_2)$, $\phi = B\mathcal{H}_{\alpha,\tau}(\tilde{f})$, and after simplification,

$$\tilde{f}_\tau = \mathcal{L}\tilde{f} - \omega \cdot \nabla \tilde{f} - \phi \cdot \nabla \psi - e^{-\tau}(b_1 v^{F_1} \cdot \nabla \Lambda_1 - b_2 v^{F_2} \cdot \nabla \Lambda_2). \tag{7.1}$$

In mild form $\tilde{f}$ satisfies the integral equation

$$\begin{aligned} \tilde{f} = e^{\tau\mathcal{L}} \tilde{f}_0 - \int_0^\tau & e^{-1/2(\tau-\sigma)} \nabla \\ & \cdot e^{(\tau-\sigma)\mathcal{L}}(\omega \tilde{f} + \phi \cdot \nabla \psi + e^{-\tau}(b_1 v^{F_1} \cdot \nabla \Lambda_1 - b_2 v^{F_2} \cdot \nabla \Lambda_2))(\sigma)\, d\sigma. \end{aligned}$$

The key difference between this system and (6.3) of the previous section is the "forcing term" $e^{-\tau}(b_1 v^{F_1} \cdot \nabla \Lambda_1 - b_2 v^{F_2} \cdot \nabla \Lambda_2)$.

As in the previous section fix $\tilde{w}_0$ and for any $n \in \mathbb{N}$ consider the system

(7.2)
$$\tilde{f}_n = e^{\tau\mathcal{L}}\tilde{f}_{0,n} - \int_0^\tau e^{-1/2(\tau-\sigma)}\nabla\cdot e^{(\tau-\sigma)\mathcal{L}}(\omega(n+\sigma)\tilde{f}(\sigma) + a\phi(\sigma)\psi(n+\sigma))\,d\sigma$$
$$- \int_0^\tau e^{-1/2(\tau-\sigma)}\nabla\cdot e^{(\tau-\sigma)\mathcal{L}}(e^{-(n+\sigma)}(b_1 v^{F_1}\cdot\nabla\Lambda_1 - b_2 v^{F_2}\cdot\nabla\Lambda_2)(n+\sigma))\,d\sigma,$$
$$\phi(\sigma) = B\mathcal{H}_{\alpha,\sigma+\tau}(\tilde{f})(\sigma), \qquad \tilde{f}(0) = \tilde{f}_0 \in X_2.$$

We redefine Θ_n for this section.

DEFINITION 7.1. *Throughout the remainder of this section let $\Theta_n(\tilde{f}_{n,0},\tau)$ denote the global solution to the system* (7.2) *with initial data $\tilde{f}_{n,0}$.*

LEMMA 7.2. *For each $n \in \mathbb{N}$, the semiflows $\Theta_n(\tilde{f}_0,\tau)$ are C^1 in $L^2(2)\times\mathbb{R}^+$. There exist a constant $r_0 > 0$ (possibly small) and $D > 0$ such that for all $\|\tilde{w}_0\|_3 < r_0$ and $n \in \mathbb{N}$ the flow Θ_n satisfies the following Lipschitz property:*

$$\sup_{n\in\mathbb{N}}\sup_{0\le\tau<1} Lip(\Theta_n(\cdot)(\tau)) = D < \infty. \tag{7.3}$$

This bound holds as $r_0 \to 0$.

Proof. The proof is nearly identical to the proof of Lemma 6.2. □

LEMMA 7.3. *There exists a constant $r_0 > 0$ (possibly small) such that for all $\|\tilde{w}_0\|_3 < r_0$ and $n \in \mathbb{N}$ the flow Θ_n can be decomposed as $\Theta_n(\tilde{f}_0,1) = e^{\mathcal{L}}\tilde{f}_0 + R_n(\tilde{f}_0) + S_n$, where $R_n(\cdot)$ is Lipschitz as a function from $L^2(3)$ to itself and $S_n \in L^2(3)$ satisfies*

$$\limsup_{n\to\infty}\frac{1}{n}\ln\|S_n\|_3 < -1. \tag{7.4}$$

The Lipschitz constant $Lip(R) := \sup_{n\in\mathbb{N}} Lip(R_n(\cdot))$ can be made arbitrarily small and satisfies the following conditions:

(i) *For any $\mu \in (\frac{1}{2},1)$, r_0 may be chosen so that, for all $n \in \mathbb{N}$,*

$$\frac{C_1}{e^{-1/2}-e^{-\mu}} + \frac{C_2}{e^{-\mu}-e^{-1}} < \frac{1}{Lip(R)}. \tag{7.5}$$

(ii) *This bound holds as $r_0 \to 0$.*

Proof. Define

$$S_n = -\int_0^\tau e^{-1/2(\tau-\sigma)}\nabla\cdot e^{(\tau-\sigma)\mathcal{L}}(e^{-(n+\sigma)}(b_1^2 v^{F_1}\cdot\nabla\Lambda_1 - b_2^2 v^{F_2}\cdot\nabla\Lambda_2)(n+\sigma))\,d\sigma$$

and $R_n(\tilde{f}_0) = \Theta_n(\tilde{f}_0)(1) - e^{\mathcal{L}}\tilde{f}_0 - S_n$. Proving (7.4) is similar to the proof of Lemma 4.3:

$$\|S_n\|_3 \le e^{-n}(b_1\|v^{F_1}\|_3\|\Lambda_1\|_3 + b_2\|v^{F_2}\|_3\|\Lambda_2\|_3). \tag{7.6}$$

The remaining statements in this proof follow as in the proof of Lemma 6.3. □

THEOREM 7.4. *Pick $\mu \in (\frac{1}{2},1)$ and choose $r_0 > 0$ to satisfy the conclusions of Lemmas* 6.2 *and* 6.3. *Let P_i^3 and X_i^3, $i = 1,2$, be as in Definition* 4.5. *Given $\tilde{w}_0$ such that $\|\tilde{w}_0\|_3 \le r_0$ and initial data $\tilde{f}_0 \in X_2$, there exists a unique global solution $\Theta(\tilde{f}_0,\tau) \in C^0([0,\infty), L^2(2))$ of* (6.3). *This solution satisfies $P_1^3\Theta(\tilde{f}_0,\tau) = 0$ and*

$$\limsup_{\tau\to\infty}\frac{1}{t}\ln\|\Theta(\tilde{f}_0,\tau)\|_2 < -\mu. \tag{7.7}$$

Proof. The proof progresses similarly to the proof of Theorem 7.4. The main difference is in the argument for the statement $P_1^3\Theta(\tilde{f}_0,\tau)=0$. This is accomplished by decomposing $P_1\Theta(\tilde{f}_0,\tau)$ onto the first three eigenvectors of $\mathcal{L}$ using conservation of mass (discussed at the beginning of section 6) and conservation of first moments (discussed at the start of section 7):

$$P_1^3\Theta(\tilde{f}_0,\tau)=aG+e^{-\frac{1}{2}}(b_1F_1+b_2F_2).$$

That a and b_i must be zero follows by comparing this expression to the decay implied by (7.7). ∎

THEOREM 7.5. *Pick $\mu\in(\frac{1}{2},1)$ and choose $r_0>0$ to satisfy the conclusions of Lemmas* 7.2 *and* 7.3. *Given initial data $\tilde{w}_0$ such that $\|\tilde{w}_0\|_3\le r_0$, the solution $\tilde{w}(\tau)$ of the scaled VCHE given by Theorem* 4.4 *is subject to the following decay estimate:*

$$\|\tilde{w}(\tau)-a\Gamma(\tau)-e^{\frac{-\tau}{2}}(b_1\Lambda_1+b_2\Lambda_2)\|_2\le Ce^{-\mu\tau},$$

where $a=\int\tilde{w}_0\,d\xi$ and $b_i=\int\xi_i\tilde{w}_0\,d\xi$.

Proof. In the previous theorem take $\tilde{f}_0=P_2\tilde{w}_0$; then $\tilde{f}(\tau)=\tilde{w}(\tau)-a\Gamma(\tau)-e^{\frac{-\tau}{2}}(b_1\Lambda_1+b_2\Lambda_2)$. The decay then follows from (6.8). ∎

The following corollary proves Theorem 1.3, which was stated after removing the fast decaying term.

COROLLARY 7.6. *For any $\mu\in(\frac{1}{2},1)$, there exists a $r_0>0$ so that for any initial data $\tilde{v}_0\in L^2(3)$ with $\|\tilde{v}_0\|_3\le r_0$ the solution of* (1.2) *given by Theorem* 3.2 *satisfies*

$$|\tilde{v}(\cdot,t)-a(\Omega(\cdot,t)-\alpha^2\triangle\Omega(\cdot,t))-\sum_{i=1,2}b_i(\partial_i\Omega(x,t)-\alpha^2\triangle\partial_i\Omega(x,t))|_p\le C(1+t)^{-1-\mu+\frac{1}{p}},$$

where $a=\int\tilde{w}_0\,dx$ and $b_i=\int x_i\tilde{v}_0\,dx$ and Ω is defined by (6.10).

Proof. This is the result of the previous theorem in unscaled coordinates. Let $\Omega(x,t)=\frac{1}{(1+t)}G(\frac{x}{\sqrt{1+t}})$; then

$$\frac{1}{(1+t)}\Gamma\left(\frac{x}{\sqrt{1+t}},\ln(1+t)\right)=\Omega(x,t)-\alpha^2\triangle\Omega(x,t),$$

$$e^{-\frac{\tau}{2}}\frac{1}{(1+t)}\Lambda_i\left(\frac{x}{\sqrt{1+t}},\ln(1+t)\right)=\partial_i\Omega(x,t)-\alpha^2\triangle\partial_i\Omega(x,t).$$

The rest follows as in Corollary 6.6. ∎

Acknowledgments. The author would like to thank M. E. Schonbek for suggesting the problem and the anonymous referees for insightful suggestions and comments.

REFERENCES

[1] M. BEN-ARTZI, *Global solutions of two-dimensional Navier-Stokes and Euler equations*, Arch. Rational Mech. Anal., 128 (1994), pp. 329–358.

[2] C. BJORLAND AND M. E. SCHONBEK, *On questions of decay and existence for the viscous Camassa-Holm equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, to appear.

[3] H. BREZIS, *Remarks on the preceding paper by M. Ben-Artzi: "Global solutions of two-dimensional Navier-Stokes and Euler equations"* [Arch. Rational Mech. Anal., 128 (1994), pp. 329–358], Arch. Rational Mech. Anal., 128 (1994), pp. 359–360.

[4] E. A. CARLEN AND M. LOSS, *Optimal smoothing and decay estimates for viscously damped conservation laws, with applications to the 2-D Navier-Stokes equation*, Duke Math. J., 81 (1995), pp. 135–157.

[5] A. Carpio, *Asymptotic behavior for the vorticity equations in dimensions two and three*, Comm. Partial Differential Equations, 19 (1994), pp. 827–872.
[6] X. Chen, J. K. Hale, and B. Tan, *Invariant foliations for C^1 semigroups in Banach spaces*, J. Differential Equations, 139 (1997), pp. 283–318.
[7] C. Foias, D. D. Holm, and E. S. Titi, *The Navier-Stokes-alpha model of fluid turbulence*, Phys. D, 152/153 (2001), pp. 505–519.
[8] C. Foias, D. D. Holm, and E. S. Titi, *The three dimensional viscous Camassa-Holm equations, and their relation to the Navier-Stokes equations and turbulence theory*, J. Dynam. Differential Equations, 14 (2002), pp. 1–35.
[9] T. Gallay and C. E. Wayne, *Invariant manifolds and the long-time asymptotics of the Navier-Stokes and vorticity equations on $\mathbb{R}^2$*, Arch. Ration. Mech. Anal., 163 (2002), pp. 209–258.
[10] Y. Giga and T. Kambe, *Large time behavior of the vorticity of two-dimensional viscous flow and its application to vortex formation*, Comm. Math. Phys., 117 (1988), pp. 549–568.
[11] D. D. Holm, J. E. Marsden, and T. S. Ratiu, *The Euler-Poincaré equations and semidirect products with applications to continuum theories*, Adv. Math., 137 (1998), pp. 1–81.
[12] D. D. Holm and E. S. Titi, *Computational models of turbulence: The LANS-α model and the role of global analysis*, SIAM News, 38 (2005).
[13] A. A. Ilyin and E. S. Titi, *Attractors for the two-dimensional Navier-Stokes-α model: An α-dependence study*, J. Dynam. Differential Equations, 15 (2003), pp. 751–778.
[14] T. Kato, *The Navier-Stokes equation for an incompressible fluid in $\mathbf{R}^2$ with a measure as the initial vorticity*, Differential Integral Equations, 7 (1994), pp. 949–966.
[15] J. E. Marsden and S. Shkoller, *Global well-posedness for the Lagrangian averaged Navier-Stokes (LANS-α) equations on bounded domains*, R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci., 359 (2001), pp. 1449–1468.
[16] H. Osada, *Diffusion processes with generators of generalized divergence form*, J. Math. Kyoto Univ., 27 (1987), pp. 597–619.

SIAM J. MATH. ANAL.
Vol. 40, No. 2, pp. 540–565

RIGOROUS DERIVATION OF INCOMPRESSIBLE e-MHD EQUATIONS FROM COMPRESSIBLE EULER–MAXWELL EQUATIONS*

YUE-JUN PENG† AND SHU WANG‡

Abstract. We derive incompressible e-MHD equations from compressible Euler–Maxwell equations via the quasi-neutral regime. Under the assumption that the initial data are well prepared for the electric density, electric velocity, and magnetic field (but not necessarily for the electric field), the convergence of the solutions of the compressible Euler–Maxwell equations in a torus to the solutions of the incompressible e-MHD equations is justified rigorously by studies on a weighted energy. One of the main ingredients for establishing uniform a priori estimates is to use the curl-div decomposition of the gradient and the wave-type equations of the Maxwell equations.

Key words. Euler–Maxwell equations, incompressible electron magnetohydrodynamics equations, quasi-neutral limit, weighted energy

AMS subject classifications. 35B40, 35C20, 35L60, 35Q35

DOI. 10.1137/070686056

1. Introduction.

Let n and u be the density and velocity vector of the electric particles in a plasma and E and B be, respectively, the electric field and magnetic field. They are vector functions of a three-dimensional position vector $x \in \mathbb{T}$ and of the time $t > 0$, where $\mathbb{T} = (\mathbb{R}/2\pi\mathbb{Z})^3$ is the torus. The fields E and B are coupled to the electron density through the Maxwell equations and act on the electrons via the Lorentz force. We assume that in the plasma the ions are nonmoving and become a uniform background with a fixed unit density. This implies that the density of ions is equal to 1 and the velocity of ions vanishes. Under these assumptions, the dynamics of the compressible electrons obey the (scaled) one-fluid Euler–Maxwell system:

$$\partial_t n + \operatorname{div}(nu) = 0, \tag{1.1}$$

$$\partial_t (nu) + \operatorname{div}(nu \otimes u) + \nabla p(n) = -n(E + \gamma u \times B), \tag{1.2}$$

$$\gamma\epsilon\partial_t E - \nabla \times B = \gamma n u, \quad \gamma\partial_t B + \nabla \times E = 0, \tag{1.3}$$

$$\epsilon \operatorname{div} E = 1 - n, \quad \operatorname{div} B = 0 \tag{1.4}$$

for $x \in \mathbb{T}$ and $t > 0$ subject to initial conditions

$$(n, u, E, B)(t = 0) = (n_0^\epsilon, u_0^\epsilon, E_0^\epsilon, B_0^\epsilon) \tag{1.5}$$

for $x \in \mathbb{T}$. In the above equations, $p = p(n)$ is the pressure, assumed to be smooth and strictly increasing for $n > 0$, $j = nu$ is the current density, and $E + \gamma u \times B$ represents

*Received by the editors March 22, 2007; accepted for publication (in revised form) February 4, 2008; published electronically May 28, 2008.
http://www.siam.org/journals/sima/40-2/68605.html
†Laboratoire de Mathématiques, CNRS UMR 6620, Université Blaise Pascal (Clermont-Ferrand 2), 63177 Aubière cedex, France (peng@math.univ-bpclermont.fr).
‡College of Applied Sciences, Beijing University of Technology, PingLeYuan100, Chaoyang District, Beijing 100022, People's Republic of China (wangshu@bjut.edu.cn). This author's research was partially supported by the NSFC (grant 10771009) and BSFC (grant 1082001) of China, the Educational Ministry of China (grant NCET-04-0203), and the Personal Ministry of China.

the Lorentz force. Equations (1.1)–(1.2) are the mass and momentum balance laws for the electrons, respectively, while (1.3)–(1.4) are the Maxwell equations. It is easy to see that equations (1.4) are redundant with equations (1.3) as soon as they are satisfied by the initial data. However, we keep them in the system because this redundancy may be lost in the asymptotic limit.

Equations (1.1)–(1.4) can be viewed as a fluid version of the Vlasov–Maxwell system for a plasma describing the evolution of the electron phase density [3]. Their form is similar to the two-fluid Euler–Maxwell system (see [1, 5, 19, 21]). However, the natures of the one-fluid and the two-fluid Euler–Maxwell systems are different. In the two-fluid model, the vanishing velocity of ions implies the vanishing electric field and the quasi neutrality of the plasma; i.e., the density of ions is equal to the density of electrons. Thus, the one-fluid equations (1.1)–(1.4) cannot be derived from the two-fluid equations.

The dimensionless parameters ϵ and γ can be chosen independently of each other, according to the desired scaling. Physically, ϵ and γ can be chosen to be proportional to the Debye length and $\frac{1}{c}$, where $c = (\epsilon_0\nu_0)^{-\frac{1}{2}}$ is the speed of light, with ϵ_0 and ν_0 being the vacuum permittivity and permeability (see [5, pp. 349–351]). The limit $\epsilon \to 0$ leads to $n = 1$, which is called the quasi neutrality of the plasma (here the constant 1 denotes the unit density of nonmoving ions). Thus, the limit $\epsilon \to 0$ is called the quasi-neutral limit. Also, the limit $\gamma \to 0$ is physically called the nonrelativistic limit. For the other physical meaning of the dimensionless parameters ϵ and γ, see [3, 10].

In the present paper, we concentrate on the so-called quasi-neutral regime. Hence here we consider only the following quasi-neutral scaling: $\gamma = O(1)$ and $\epsilon \ll 1$.

Now, setting formally $\epsilon = 0$ in the system (1.1)–(1.4), one can arrive at the so-called electron magnetohydrodynamics (e-MHD) equations as follows [3]:

$$\partial_t u + u \cdot \nabla u + E = -\gamma u \times B, \tag{1.6}$$

$$-\nabla \times B = \gamma u, \quad \mathrm{div} B = 0, \tag{1.7}$$

$$\gamma \partial_t B + \nabla \times E = 0, \quad n = 1. \tag{1.8}$$

The e-MHD system (1.6)–(1.8) is incompressible, i.e., $\mathrm{div} u = 0$, which is precisely the limit equation obtained from (1.1) by using the fact that $n = 1$ in (1.8). However, this incompressible condition need not be written separately since it can be obtained by the first equation in (1.7).

The main purpose of this paper is to prove rigorously the above formal limit for smooth solutions of the Euler–Maxwell system (1.1)–(1.4) on time intervals, on which a smooth solution of the incompressible e-MHD equations exists. The precise statement is given in section 2.

From the view of the singular perturbation theory, the quasi-neutral limit $\epsilon \to 0$ in the Euler–Maxwell system is a problem of singular perturbation for hyperbolic systems (see [8, 11, 12, 22]). However, it is very different from the theory of singular low Mach number limit for symmetrizable hyperbolic systems by Klainerman and Majda in [11, 12]. In the latter case, the essential singularity can be cancelled using a symmetrizer of hyperbolic systems. For the quasi-neutral limit in the Euler–Maxwell system, besides the singularities in the Maxwell equations, there exists an extra singularity caused by the coupling electromagnetic field (source term) in the Euler equations, which cannot be overcome by using the symmetric technique of hyperbolic systems. Hence the singular limit theory for symmetrizable hyperbolic systems developed by Klainerman

and Majda [11, 12] or extended further by Schochet [22] cannot be applied here to obtain the uniform a priori estimates of the solution with respect to ϵ.

In this paper, we control these singularities and then derive rigorously the e-MHD system from the Euler–Maxwell system by an elaborate energy method based on studies on an ϵ-weighted energy. Let $(n^\epsilon, u^\epsilon, E^\epsilon, B^\epsilon)$ be the solution to the Euler–Maxwell problem and (u^0, E^0, B^0) be the solution to the incompressible e-MHD equations. Our basic idea to establish the uniform a priori estimates is the need to obtain an estimate not just for $n^\epsilon - 1 - \epsilon \mathrm{div} E^0$ but also for $\frac{n^\epsilon - 1 - \epsilon \mathrm{div} E^0}{\sqrt{\epsilon}}$ and then for $\sqrt{\epsilon}(E^\epsilon - E^0)$. In particular, the order of derivatives that we need to estimate for the latter is one less than for the other main quantities $(n^\epsilon - 1 - \epsilon \mathrm{div} E^0, u^\epsilon - u^0, B^\epsilon - B^0)$. These estimates cannot be obtained by straightforward Sobolev energy estimates for which the same order of derivatives for both is needed. They are achieved due to the dissipation structure of the equation for $\frac{n^\epsilon - 1 - \epsilon \mathrm{div} E^0}{\sqrt{\epsilon}}$ (see (4.21)). Finally, the desired estimates are obtained through a priori estimates of vorticity and divergence (see Lemmas 4.2–4.3).

There have been a lot of studies on the Euler–Poisson equations and their asymptotic analysis contrarily to the study on the Euler–Maxwell equations. See [2, 4, 7, 8, 20, 23, 24, 25] and the references therein. The first mathematical study of the Euler–Maxwell equations with an extra relaxation term is due to Chen, Jerome, and Wang [6], where a global existence result to weak solutions in the one-dimensional case is established by the fractional step Godunov scheme together with a compensated compactness argument. Paper [6] also exhibits some applications of the model (1.1)–(1.4) in the semiconductor theory. Since then little progress has been made on the Euler–Maxwell equations. Recently the convergence of the Euler–Maxwell system to the compressible Euler–Poisson system has been proved in [17] via the nonrelativistic limit, in which general initial data are allowed by performing an initial layer analysis. This limit corresponds to $\gamma \to 0$ and $\epsilon > 0$ being fixed. The convergence of the compressible Euler–Maxwell equations to the incompressible Euler equations is justified as $\gamma = \epsilon \to 0$ (see [18]). Finally, we mention that a related asymptotic limit problem on the Vlasov–Maxwell system is discussed in [2, 3].

We stress that, in this paper, the convergence of the electron density, the current velocity vector, and the magnetic field vector of the Euler–Maxwell systems is strong, whereas the convergence of the electric field vector is only weak. That is why we do not need the assumption that the initial electric field of Euler–Maxwell systems tends to the initial value of the electric field of the limit system. The latter is not arbitrarily given but is determined by the initial data of the limit system (2.6)–(2.9) (see Proposition 2.1). This is different from the situation in [18], where the initial data are prepared. Also, by checking the proof in section 4, we see that it is possible to extend the results of this paper ($\gamma > 0$ being fixed and $\epsilon \to 0$) to the case $\gamma \to 0$ and $\epsilon \to 0$ without any relation between γ and ϵ. This limit is still governed by the incompressible Euler equations. To see this more clearly, the parameter γ is kept in the estimates.

Notation and preliminary results.

(1) Throughout this paper, $\nabla = \nabla_x$ is the gradient, $\nabla\cdot$ is the divergence operator, and $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ and β, etc., are multi-indices. We denote by $H^s(\mathbb{T})$ the standard Sobolev space in torus $\mathbb{T}$, which is defined by Fourier transform, namely, $f \in H^s(\mathbb{T})$ if and only if

$$\|f\|_s^2 = (2\pi)^3 \sum_{k \in \mathbb{Z}^3} (1 + |k|^2)^s |(\mathcal{F}f)(k)|^2 < +\infty,$$

where $(\mathcal{F}f)(k) = \int_{\mathbb{T}} f(x)e^{-ikx}dx$ is the Fourier transform of $f \in H^s(\mathbb{T})$.

(2) Recall the following basic Moser-type calculus inequalities [11, 12]: for $f, g, v \in H^s$ and any nonnegative multi index $\alpha, |\alpha| \le s$,

(i) $\|D_x^\alpha(fg)\|_{L^2} \le C_s(\|f\|_{L^\infty}\|D_x^s g\|_{L^2} + \|g\|_{L^\infty}\|D_x^s f\|_{L^2}), \quad s \ge 0,$

(ii) $\|D_x^\alpha(fg) - fD_x^\alpha g\|_{L^2} \le C_s(\|D_x f\|_{L^\infty}\|D_x^{s-1} g\|_{L^2} + \|g\|_{L^\infty}\|D_x^s f\|_{L^2}), \quad s \ge 1.$

(3) The following vector analysis formulas will be repeatedly used (see [5]):

$$\mathrm{div}(f \times g) = \nabla \times f \cdot g - \nabla \times g \cdot f, \tag{1.9}$$

$$f \cdot \nabla g = (\nabla \times g) \times f + \nabla(f \cdot g) - \nabla f \cdot g, \tag{1.10}$$

$$f \cdot \nabla f = (\nabla \times f) \times f + \nabla\left(\frac{|f|^2}{2}\right), \tag{1.11}$$

$$\nabla \times (f \times g) = f\mathrm{div}g - g\mathrm{div}f + (g \cdot \nabla)f - (f \cdot \nabla)g. \tag{1.12}$$

2. Well-posedness of the e-MHD and main results. For smooth solutions of the Euler–Maxwell system (1.1)–(1.5) with $n > 0$, (1.2) is equivalent to

$$\partial_t u + (u \cdot \nabla)u + \nabla h(n) = -(E + \gamma u \times B),$$

where the enthalpy $h(n)$ is defined by

$$h(n) = \int_1^n \frac{p'(s)}{s} ds.$$

Thus, regarding ϵ as a singular perturbation parameter, we can rewrite the problem (1.1)–(1.5) as

$$\partial_t n^\epsilon + \mathrm{div}(n^\epsilon u^\epsilon) = 0, \tag{2.1}$$

$$\partial_t u^\epsilon + (u^\epsilon \cdot \nabla)u^\epsilon + \nabla h(n^\epsilon) = -(E^\epsilon + \gamma u^\epsilon \times B^\epsilon), \tag{2.2}$$

$$\gamma\epsilon\partial_t E^\epsilon - \nabla \times B^\epsilon = \gamma n^\epsilon u^\epsilon, \quad \gamma\partial_t B^\epsilon + \nabla \times E^\epsilon = 0, \tag{2.3}$$

$$\epsilon\mathrm{div}E^\epsilon = 1 - n^\epsilon, \quad \mathrm{div}B^\epsilon = 0, \tag{2.4}$$

$$(n^\epsilon, u^\epsilon, E^\epsilon, B^\epsilon)(t = 0) = (n_0^\epsilon, u_0^\epsilon, E_0^\epsilon, B_0^\epsilon), \tag{2.5}$$

where γ is a positive constant of order one. This means that the magnetic field does not vanish in the limiting process.

We also rewrite the limit system (1.6)–(1.8) as

$$\partial_t u^0 + u^0 \cdot \nabla u^0 + E^0 = -\gamma u^0 \times B^0, \tag{2.6}$$

$$-\nabla \times B^0 = \gamma u^0, \quad \mathrm{div}B^0 = 0, \tag{2.7}$$

$$\gamma\partial_t B^0 + \nabla \times E^0 = 0, \quad n^0 = 1. \tag{2.8}$$

Following the idea of [3], introduce the general vorticity

$$\omega^0 = \nabla \times (u^0 - \gamma A^0),$$

where A^0 is the magnetic potential such that

$$\nabla \times A^0 = B^0 \quad \text{and} \quad \mathrm{div} A^0 = 0.$$

Noting $\nabla \cdot u^0 = 0$ and using the identity (1.11), we can write the e-MHD equations (2.6)–(2.8) in a different way:

$$\partial_t \omega^0 + u^0 \cdot \nabla \omega^0 - \omega^0 \cdot \nabla u^0 = 0, \quad -\Delta u^0 + \gamma^2 u^0 = \nabla \times \omega^0.$$

Therefore the existence results are the same as for the incompressible Euler equations [14]. In particular, we have local smooth solutions of the e-MHD equations for the smooth initial data given by

$$u^0(t=0) = u_0^0, \quad B^0(t=0) = B_0^0. \tag{2.9}$$

PROPOSITION 2.1 (see [3]). *Assume that* $u_0^0, B_0^0 \in C^\infty(\mathbb{T})$ *satisfy*

$$-\nabla \times B_0^0 = \gamma u_0^0, \quad \mathrm{div} B_0^0 = 0. \tag{2.10}$$

Then there exist $0 < T_* \le \infty$ *(if* $d = 2$, $T_* = \infty$*), the maximal existence time, and a unique smooth solution* $(u^0, B^0, E^0) \in C^\infty(\mathbb{T} \times [0, T_*))$ *of the incompressible e-MHD equations* (2.6)–(2.9) *defined on* $[0, T_*)$.

For the convergence of the compressible Euler–Maxwell system (2.1)–(2.5), our main result is stated as follows.

THEOREM 2.1. *Let* $s_0 > \frac{3}{2} + 2$ *and* $\gamma > 0$ *be fixed. Let* $u_0^0, B_0^0 \in C^\infty(\mathbb{T})$ *satisfy* (2.10) *and* $n_0^\epsilon, E_0^\epsilon, B_0^\epsilon \in C^\infty(\mathbb{T})$ *satisfy*

$$\epsilon \mathrm{div} E_0^\epsilon = 1 - n_0^\epsilon, \quad \mathrm{div} B_0^\epsilon = 0.$$

Assume that

$$\|(n_0^\epsilon - 1, u_0^\epsilon - u_0^0, B_0^\epsilon - B_0^0)\|_{H^{s_0}(\mathbb{T})} + \|\sqrt{\epsilon} E_0^\epsilon\|_{H^{s_0}(\mathbb{T})} \le C\sqrt{\epsilon} \tag{2.11}$$

for some positive constant C *independent of* ϵ. *Let* T_*, $0 < T_* \le \infty$ $(d = 2, T_* = \infty)$, *be the maximal existence time of the smooth solution* $(u^0, B^0, E^0) \in C^\infty(\mathbb{T} \times [0, T_*))$ *of the incompressible e-MHD equations* (2.6)–(2.9). *Then, for any* $T_0 < T_*$, *there exist constants* $\epsilon_0(T_0) > 0$ *and* $\tilde{M}(T_0) > 0$, *depending only upon* T_0 *and the initial data, such that the Euler–Maxwell system* (2.1)–(2.5) *has a classical smooth solution* $(n^\epsilon, u^\epsilon, E^\epsilon, B^\epsilon)$, *defined on* $[0, T_0]$, *satisfying*

$$\|(n^\epsilon - 1, u^\epsilon - u^0, B^\epsilon - B^0)(\cdot, t)\|_{H^{s_0}(\mathbb{T})} + \|\sqrt{\epsilon} E^\epsilon(\cdot, t)\|_{H^{s_0 - 1}(\mathbb{T})} \le \tilde{M}(T_0)\sqrt{\epsilon} \tag{2.12}$$

for all $0 < \epsilon \le \epsilon_0$ *and* $0 \le t \le T_0$. *As a result, the sequence* $(n^\epsilon, u^\epsilon, B^\epsilon)_{\epsilon > 0}$ *converges strongly to* $(1, u^0, B^0)$ *in* $L^\infty(0, T_0; H^{s_0}(\mathbb{T}))$. *Furthermore, the sequence* $(E^\epsilon)_{\epsilon > 0}$ *converges to* E^0 *in* $W^{-1,\infty}(0, T_0; H^{s_0 - 1}(\mathbb{T}))$.

Remark 2.1. Condition (2.11) means that the initial data are well prepared for $(n^\epsilon, u^\epsilon, B^\epsilon)$ but not for E^ϵ, which are only bounded. This is sufficient to conclude the convergence of E^ϵ to E^0 in some weak sense. Since all the terms involving E^ϵ in the Euler–Maxwell system (2.1)–(2.5) are linear, we can pass to the limit in the system (in the sense of distributions, for instance). The fact that the electric field $E^\epsilon(\cdot, t)$ is bounded uniformly in ϵ for all time $t \in [0, T_0]$ is not surprising because, generally speaking, this property should be maintained over time when the initial electric field is bounded uniformly in ϵ and the density n^ϵ has a better convergence rate. As in

[8, 13, 22], some new techniques are required to study the strong convergence of the electric field when its initial value is also prepared. This is related to the initial layer analysis and will be discussed in the future.

Remark 2.2. The convergence rate $O(\sqrt{\epsilon})$ in Theorem 2.1 is probably not optimal. It should be possible to improve it by constructing better approximate solutions under the same assumption. Moreover, if the initial error for $(n^\epsilon, u^\epsilon, B^\epsilon)$ is better than $O(\sqrt{\epsilon})$, then so should the convergence rate. On the other hand, the convergence rate should be better in a weaker norm than $\|\cdot\|_{s_0}$. For example, the estimate (2.12) on E^ϵ and the first equation in (2.4) already give

$$\|n^\epsilon(\cdot,t)-1\|_{H^{s_0-2}(\mathbb{T})}=O(\epsilon).$$

Remark 2.3. The results of this paper hold in the whole space $\mathbb{R}^3$. Indeed, the key point of the proof of Theorem 2.1 is to establish the uniform a priori estimates (2.12) based on the study of a weighted energy combined with singular perturbation methods. This is dealt with in the same way as that of the papers [11, 12] by Klainerman and Majda. Here neither the compactness of $\mathbb{T}$ nor a Poincaré-type inequality is used.

3. Derivation of error equations and local existence. Let $(n^\epsilon, u^\epsilon, E^\epsilon, B^\epsilon)$ be the unknown solution to the problem (2.1)–(2.5) and (u^0, E^0, B^0) be the solution to the incompressible e-MHD equations defined on $[0, T_*)$ given by Proposition 2.1. Denote this by

$$(N^\epsilon, U^\epsilon, F^\epsilon, G^\epsilon)=(n^\epsilon-1+\epsilon\mathrm{div}E^0, u^\epsilon-u^0, E^\epsilon-E^0, B^\epsilon-B^0), \tag{3.1}$$

which satisfies the following problem:

(3.2)
$$\begin{cases}
\partial_t N^\epsilon+\mathrm{div}((N^\epsilon+1-\epsilon\mathrm{div}E^0)U^\epsilon+N^\epsilon u^0)=\epsilon(\partial_t\mathrm{div}E^0+\mathrm{div}(u^0\mathrm{div}E^0)),\\
\partial_t U^\epsilon+[(U^\epsilon+u^0)\cdot\nabla]U^\epsilon+(U^\epsilon\cdot\nabla)u^0+F^\epsilon+\nabla(h(N^\epsilon+1-\epsilon\mathrm{div}E^0)\\
\qquad\qquad -h(1-\epsilon\mathrm{div}E^0))\\
\qquad\qquad =-\gamma\big((U^\epsilon+u^0)\times G^\epsilon+U^\epsilon\times B^0\big)+\epsilon h'(1-\epsilon\mathrm{div}E^0)\nabla(\mathrm{div}E^0),\\
\epsilon\gamma\partial_t F^\epsilon-\nabla\times G^\epsilon=\gamma\big((N^\epsilon+1-\epsilon\mathrm{div}E^0)U^\epsilon+N^\epsilon u^0\big)-\epsilon\gamma\big(\partial_t E^0+u^0\mathrm{div}E^0\big),\\
\gamma\partial_t G^\epsilon+\nabla\times F^\epsilon=0,\\
\epsilon\mathrm{div}F^\epsilon=-N^\epsilon,\quad \mathrm{div}G^\epsilon=0,\\
(N^\epsilon, U^\epsilon, F^\epsilon, G^\epsilon)|_{t=0}\\
\qquad\qquad =\big(n_0^\epsilon-1+\epsilon\mathrm{div}E^0(t=0), u_0^\epsilon-u_0^0, E_0^\epsilon-E^0(t=0), B_0^\epsilon-B_0^0\big).
\end{cases}$$

Note that we use $1-\epsilon\mathrm{div}E^0$ as the approximate solution of the density n^ϵ instead of the formal approximate solution 1. This implies that the equation $\epsilon\mathrm{div}F^\epsilon=-N^\epsilon$ is homogeneous. This is a key point in the following convergence analysis. Otherwise, an extra term $\mathrm{div}E^0$ appears in the divergence equation (4.21), which is just bounded but does not converge to 0 as $\epsilon\to 0$. If so, by the techniques used in this paper, the desired convergence rate $O(\sqrt{\epsilon})$ cannot be obtained.

Set

$$W_I^\epsilon = \begin{pmatrix} N^\epsilon \\ U^\epsilon \end{pmatrix}, \quad W_{II}^\epsilon = \begin{pmatrix} F^\epsilon \\ G^\epsilon \end{pmatrix}, \quad W^\epsilon = \begin{pmatrix} W_I^\epsilon \\ W_{II}^\epsilon \end{pmatrix} = \begin{pmatrix} N^\epsilon \\ U^\epsilon \\ F^\epsilon \\ G^\epsilon \end{pmatrix},$$

$$W_0^\epsilon = \begin{pmatrix} N_0^\epsilon \\ U_0^\epsilon \\ F_0^\epsilon \\ G_0^\epsilon \end{pmatrix} = \begin{pmatrix} n_0^\epsilon - 1 + \epsilon \mathrm{div} E^0(t=0) \\ u_0^\epsilon - u_0^0 \\ E_0^\epsilon - E^0(t=0) \\ B_0^\epsilon - B_0^0 \end{pmatrix}, \ D_0^\epsilon = \begin{pmatrix} \mathbf{I}_{4\times 4} & \mathbf{0} \\ \mathbf{0} & \begin{pmatrix} \epsilon\gamma \mathbf{I}_{3\times 3} & \mathbf{0} \\ \mathbf{0} & \gamma \mathbf{I}_{3\times 3} \end{pmatrix} \end{pmatrix},$$

$$A_i(W^\epsilon) = \begin{pmatrix} \begin{pmatrix} (U^\epsilon + u^0)_i & (N^\epsilon + 1 - \epsilon \mathrm{div} E^0) e_i^T \\ h'(N^\epsilon + 1 - \epsilon \mathrm{div} E^0) e_i & (U^\epsilon + u^0)_i \mathbf{I}_{3\times 3} \end{pmatrix} & \mathbf{0} \\ \mathbf{0} & \begin{pmatrix} \mathbf{0} & B_i \\ B_i^T & \mathbf{0} \end{pmatrix} \end{pmatrix},$$

$$H_1(W_I^\epsilon) = \begin{pmatrix} -\epsilon U^\epsilon \cdot \nabla(\mathrm{div} E^0) \\ (U^\epsilon \cdot \nabla) u^0 - \epsilon (h'(N^\epsilon + 1 - \epsilon \mathrm{div} E^0) - h'(1 - \epsilon \mathrm{div} E^0)) \nabla(\mathrm{div} E^0) \\ 0 \\ 0 \end{pmatrix},$$

$$H_2(F^\epsilon) = \begin{pmatrix} 0 \\ F^\epsilon \\ 0 \\ 0 \end{pmatrix}, \quad H_3(W_I^\epsilon, G^\epsilon) = \begin{pmatrix} 0 \\ (U^\epsilon + u^0) \times G^\epsilon + U^\epsilon \times B^0 \\ -\big((N^\epsilon + 1 - \epsilon \mathrm{div} E^0) U^\epsilon + N^\epsilon u^0\big) \\ 0 \end{pmatrix}$$

and

$$R^\epsilon = \begin{pmatrix} R_n^\epsilon \\ R_u^\epsilon \\ R_E^\epsilon \\ R_B^\epsilon \end{pmatrix} = \begin{pmatrix} \partial_t \mathrm{div} E^0 + \mathrm{div}(u^0 \mathrm{div} E^0) \\ h'(1 - \epsilon \mathrm{div} E^0) \nabla(\mathrm{div} E^0) \\ -\gamma(\partial_t E^0 + u^0 \mathrm{div} E^0) \\ 0 \end{pmatrix},$$

where (e_1, e_2, e_3) is the canonical basis of $\mathbb{R}^3$, $\mathbf{I}_{d\times d}$ $(d = 3, 4)$ is a $d \times d$ unit matrix, y_i denotes the ith component of $y \in \mathbb{R}^3$, and

$$B_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad B_3 = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

From $(3.2)_{1,3}$, the redundant equations $\epsilon \mathrm{div} F^\epsilon = -N^\epsilon$ and $\mathrm{div} G^\epsilon = 0$ in system (3.2) hold as soon as they are satisfied by the initial data. Thus the problem (3.2) for the

unknown W^ϵ can be rewritten as

(3.3)
$$\begin{cases} D_0^\epsilon \partial_t W^\epsilon + \sum_{i=1}^{3} A_i(W^\epsilon)\partial_{x_i} W^\epsilon + H_1(W_I^\epsilon) = H_2(F^\epsilon) + \epsilon R^\epsilon - \gamma H_3(W_I^\epsilon, G^\epsilon), \\ W^\epsilon|_{t=0} = W_0^\epsilon, \end{cases}$$

with

$$\epsilon \mathrm{div} F^\epsilon(x,0) = -N^\epsilon(x,0), \quad \mathrm{div} G^\epsilon(x,0) = 0,$$

which can be guaranteed by the assumptions on the initial data.

It is not difficult to see that the equations for W^ϵ in (3.3) are symmetrizable hyperbolic; i.e., if we introduce

$$A_0(W^\epsilon) = \begin{pmatrix} \begin{pmatrix} h'(N^\epsilon + 1 - \epsilon \mathrm{div} E^0) & 0 \\ 0 & (N^\epsilon + 1 - \epsilon \mathrm{div} E^0)\mathbf{I}_{3\times 3} \end{pmatrix} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{6\times 6} \end{pmatrix},$$

which is positively definite when $N^\epsilon + 1 - \epsilon \mathrm{div} E^0 \geq M_0 > 0$ for $\epsilon \ll 1$ and $\|N^\epsilon\|_{L^\infty} \leq \frac{1}{2}$, then $A_0 D_0^\epsilon$ and $\tilde{A}_i(W^\epsilon) = A_0(W^\epsilon)A_i(W^\epsilon)$ are symmetric for all $1 \leq i \leq 3$. Note that, for smooth solutions, the Euler–Maxwell system (2.1)–(2.5) is equivalent to that of (3.2) or (3.3). Thus, by the standard existence theory of local smooth solutions for symmetrizable hyperbolic equations (see [9, 15, 16]), we have the following result.

PROPOSITION 3.1. *Let $M > 0$ and $u_0^0, B_0^0 \in C^\infty$ be given and W_0^ϵ satisfy $W_0^\epsilon \in H^s$, $s > \frac{3}{2} + 2$, and $\|N_0^\epsilon\|_{H^s(\mathbb{T})} \leq \delta$ for some given $\delta > 0$ (to be chosen sufficiently small so that $M\delta C_s \leq \frac{1}{2}$, where C_s is the Sobolev embedding constant). Then, for any fixed ϵ ($\ll 1$), there exist $0 < T_\epsilon(\delta) \leq \infty$, the maximal existence time, and a unique smooth solution $W^\epsilon \in \bigcap_{l=0}^{1} C^l([0,T_\epsilon); H^{s-l}(\mathbb{T}))$ of the system (3.2) or (3.3) on $[0,T_\epsilon)$ satisfying $\sup_{0<t<T_\epsilon} \|N^\epsilon(t)\|_{H^s(\mathbb{T})} \leq M\delta$. Moreover, if $T_\epsilon < \infty$, then, for any fixed ϵ (sufficiently small), we have*

$$\text{either } \lim_{t\to T_\epsilon} \|N^\epsilon(t)\|_{H^s(\mathbb{T})} = M\delta \text{ or } \lim_{t\to T_\epsilon} \|(U^\epsilon, F^\epsilon, G^\epsilon)(\cdot,t)\|_{H^s(\mathbb{T})} = +\infty. \tag{3.4}$$

Note that the error equations (3.2) or (3.3) from the Euler–Maxwell system are not in the form covered by the well-known Klainerman and Majda theory of singular limits of hyperbolic systems because of the extra singularity caused by the coupling source term $H_2(F^\epsilon)$ in the Euler equations. Hence the singular limit theory for symmetrizable hyperbolic systems developed by Klainerman and Majda [11, 12] or extended further by Schochet [22] cannot be applied here to obtain the uniform a priori estimates of the solution W^ϵ with respect to ϵ.

4. Proof of the main results. In this section, we justify rigorously the convergence of the Euler–Maxwell system to the incompressible e-MHD equations; namely, we prove Theorem 2.1 by using the asymptotic expansion of singular perturbations and careful classical energy methods.

4.1. Convergence rate and uniform a priori estimates. We first establish the convergence rate of the error function $(N^\epsilon, U^\epsilon, B^\epsilon)$ by obtaining the a priori estimates uniformly in ϵ. As a consequence, we obtain the existence of exact solutions

$(n^\epsilon, u^\epsilon, E^\epsilon, B^\epsilon)$ to (2.1)–(2.5) in a time interval independent of ϵ, and the convergence of $(n^\epsilon, u^\epsilon, E^\epsilon, B^\epsilon)$ to $(1, u^0, E^0, B^0)$ as $\epsilon \to 0$, where (u^0, E^0, B^0) is the solution of the incompressible e-MHD equations (2.6)–(2.9).

In order to justify rigorously the convergence, it suffices to obtain the uniform estimates of the smooth solutions to (3.2) or (3.3) with respect to the parameter ϵ. This is achieved by the elaborate energy method. Since the detailed estimates are very lengthy and involved, we would like to outline some main ingredients here. First, based on the L^2 energy conservation of the Euler–Maxwell system, an ϵ-weighted L^2 energy estimate is derived by noting that there exists a cancellation or some kind of balance between the Euler part and Maxwell part of the Euler–Maxwell system. Second, we introduce a general vorticity of the velocity and the magnetic field and give the corresponding equations of the vorticity and the divergence. Then high order Sobolev energy estimates on the vorticity and the divergence are established by the vector analysis techniques and an elaborate energy method. Next, based on the curl-div decomposition of the gradient, an ϵ-weighted high order Sobolev energy estimate on the density and velocity of the Euler part is derived. Finally, based on the wave-type equation of the Maxwell equations, an ϵ-weighted high order Sobolev energy estimate on the electric field and magnetic field is established. Combining these estimates, an entropy production integral inequality is derived by introducing an ϵ-weighted energy, which yields the desired estimates.

In the following, $(\cdot, \cdot)$ stands for the L^2 inner product of two scalar or vector functions in $\mathbb{T}$. Also, we denote $\int_{\mathbb{T}}$ by $\int$ and

$$\|\cdot\| = \|\cdot\|_{L^2(\mathbb{T})}, \quad \|\cdot\|_l = \|\cdot\|_{H^l(\mathbb{T})}, \quad l \in \mathbb{N}^*.$$

For convenience, we introduce the ϵ-weighted Sobolev norms

$$\|W^\epsilon(t)\|_{l,*} = \|(N^\epsilon, U^\epsilon, G^\epsilon)(t)\|_l,$$

$$|||W^\epsilon(t)|||_l = \left(\|W^\epsilon(t)\|_{l,*}^2 + \left\| \left(\frac{N^\epsilon}{\sqrt{\epsilon}}, \sqrt{\epsilon}\gamma F^\epsilon, \epsilon\gamma \partial_t F^\epsilon \right)(t) \right\|_{l-1}^2 \right)^{\frac{1}{2}},$$

$$|||W^\epsilon|||_{l,T} = \sup_{0<t<T} |||W^\epsilon(t)|||_l, \quad l \in \mathbb{N}^*.$$

Note that these norms are different from those used in [18] to prove the convergence of the compressible Euler–Maxwell equations to the incompressible Euler equations.

The key estimate of this paper is contained in the following result.

PROPOSITION 4.1. *Let l be an integer such that $l > \frac{3}{2} + 2$. Assume*

$$|||W_0^\epsilon|||_l \le D_1 \sqrt{\epsilon} \tag{4.1}$$

for sufficiently small ϵ and a constant $D_1 > 0$ independent of ϵ. Then, for any $T_0 \in (0, T_*)$, there are constants $D_2 > 0$ and $\epsilon_0 > 0$, depending upon T_0, such that, for all $\epsilon \le \epsilon_0$, it holds that $T_\epsilon \ge T_0$, and the solution $W^\epsilon(t)$ of (3.2), well defined in $[0, T_\epsilon)$, satisfies

$$|||W^\epsilon|||_{l,T_0} \le D_2 \sqrt{\epsilon}. \tag{4.2}$$

The remainder of section 4 is devoted to the proof of this result.

First, according to the assumption (4.1), using the local existence results in Proposition 3.1 and Sobolev's embedding lemma, we know that there exists an $\epsilon_0 > 0$ such

that, for all $\epsilon \le \epsilon_0$, there exists a smooth solution W^ϵ to the system (3.2) or (3.3), defined on $[0, T_\epsilon)$, satisfying

$$\sup_{0<t<T_\epsilon} \|N^\epsilon(t)\|_{L^\infty(\mathbb{T})} \le C_s \sup_{0<t<T_\epsilon} \|N^\epsilon(t)\|_{H^s(\mathbb{T})} \le MC_sD_1\sqrt{\epsilon} \le \frac{1}{2}. \tag{4.3}$$

Here we take $\delta = D_1\sqrt{\epsilon}$ in Proposition 3.1, and M will be chosen to be a sufficiently large constant independent of ϵ. Hence the rest involves establishing the a priori estimates uniformly with respect to ϵ so as to guarantee $T_\epsilon \ge T_0$ for any given $T_0 < T_*$ and sufficiently small ϵ. Of course, if $T_\epsilon = \infty$, it suffices to obtain the a priori estimates uniformly with respect to ϵ.

In the following, assuming the conditions of Proposition 4.1, we establish a priori estimates by the elaborate energy method in several steps.

For any $T_1 < 1$ independent of ϵ, denote by $T = T^\epsilon = \min\{T_1, T_\epsilon\}$ and by $C > 0$ a constant which depends upon T_0, D_1 but does not depend upon M, T_1, T, and ϵ.

4.2. L^2-estimates. Based on the L^2-conservation of solutions to the Euler–Maxwell system, we obtain L^2-estimates of the error function W^ϵ. Our basic idea is to control the electric field F^ϵ using the special structure between the Euler part and Maxwell part in the Euler–Maxwell system by introducing the extra singular term $\|\frac{N^\epsilon}{\sqrt{\epsilon}}\|$.

LEMMA 4.1. *For all $0 < t < T$ and sufficiently small ϵ, it holds that*

$$\begin{aligned}
&\int \left\{ |U^\epsilon|^2 + \int_0^{N^\epsilon} (h(1-\epsilon \mathrm{div}E^0 + s) - h(1-\epsilon \mathrm{div}E^0))ds + \epsilon|F^\epsilon|^2 + |G^\epsilon|^2 \right\}(t)dx \\
\le &\int \left\{ |U^\epsilon|^2 + \int_0^{N^\epsilon} (h(1-\epsilon \mathrm{div}E^0 + s) - h(1-\epsilon \mathrm{div}E^0))ds + \epsilon|F^\epsilon|^2 + |G^\epsilon|^2 \right\}(t=0)dx \\
&+ \int_0^t \left\{ \|\sqrt{\epsilon}F^\epsilon\|^2 + C(\|W^\epsilon\|_{l,*}^2 + \|W^\epsilon\|_{l,*} + 1) \left\| \left(N^\epsilon, U^\epsilon, \nabla U^\epsilon, \gamma G^\epsilon, \frac{N^\epsilon}{\sqrt{\epsilon}} \right) \right\|^2 \right\}(s)ds \\
&+ C\epsilon^2 + C\epsilon.
\end{aligned} \tag{4.4}$$

Proof. Taking the L^2 inner product of the second equation in the error system (3.2) for U^ϵ, by integration by parts, we get

$$\begin{aligned}
&\frac{d}{dt}(U^\epsilon, U^\epsilon) + 2(F^\epsilon, U^\epsilon) \\
&= \big(\mathrm{div}(U^\epsilon + u^0)U^\epsilon, U^\epsilon\big) + 2\big(h(1-\epsilon \mathrm{div}E^0 + N^\epsilon) - h(1-\epsilon \mathrm{div}E^0), \mathrm{div}U^\epsilon\big) \\
&\quad - 2\big(U^\epsilon \nabla u^0 + \gamma(U^\epsilon \times B^0 + (U^\epsilon + u^0) \times G^\epsilon) - \epsilon R_u^\epsilon, U^\epsilon\big).
\end{aligned} \tag{4.5}$$

Now we estimate each term on the right-hand side of (4.5).

For the first and third terms, using the property of the approximate solution (u^0, E^0, B^0), Cauchy–Schwarz's inequality, and Sobolev's lemma, we get

$$\big(\mathrm{div}(U^\epsilon + u^0)U^\epsilon, U^\epsilon\big) \le C(\|W^\epsilon(t)\|_{l,*} + 1)\|U^\epsilon\|^2 \tag{4.6}$$

and

$$-2\big(U^\epsilon\nabla u^0+\gamma(U^\epsilon\times B^0+(U^\epsilon+u^0)\times G^\epsilon)-\epsilon R_u^\epsilon,U^\epsilon\big)$$

$$\le C(\|W^\epsilon(t)\|_{l,*}+1)\|(U^\epsilon,\gamma G^\epsilon)\|^2+C\epsilon^2. \tag{4.7}$$

For the second term, noting that the first equation in (3.2) can be rewritten as

$$\mathrm{div}U^\epsilon=-(\partial_t N^\epsilon+\mathrm{div}(N^\epsilon(U^\epsilon+u^0)))+\epsilon(\mathrm{div}(\mathrm{div}E^0U^\epsilon)+R_n^\epsilon),$$

from (4.3) we have, for sufficiently small ϵ,

$$\begin{aligned}
&(h(1-\epsilon\mathrm{div}E^0+N^\epsilon)-h(1-\epsilon\mathrm{div}E^0),\mathrm{div}U^\epsilon)\\
&=-\int(h(1-\epsilon\mathrm{div}E^0+N^\epsilon)-h(1-\epsilon\mathrm{div}E^0))(\partial_t N^\epsilon+\mathrm{div}(N^\epsilon(U^\epsilon+u^0)))dx\\
&\quad+\epsilon\int(h(1-\epsilon\mathrm{div}E^0+N^\epsilon)-h(1-\epsilon\mathrm{div}E^0))(\mathrm{div}(\mathrm{div}E^0U^\epsilon)+R_n^\epsilon)dx\\
&=-\frac{d}{dt}\iint_0^{N^\epsilon}(h(1-\epsilon\mathrm{div}E^0+s)-h(1-\epsilon\mathrm{div}E^0))dsdx\\
&\quad+\iint_0^{N^\epsilon}(h'(1-\epsilon\mathrm{div}E^0+s)-h'(1-\epsilon\mathrm{div}E^0))\partial_t(1-\epsilon\mathrm{div}E^0)dsdx\\
&\quad-\int(h(1-\epsilon\mathrm{div}E^0+N^\epsilon)-h(1-\epsilon\mathrm{div}E^0))\mathrm{div}(N^\epsilon(U^\epsilon+u^0))dx\\
&\quad+\epsilon\int(h(1-\epsilon\mathrm{div}E^0+N^\epsilon)-h(1-\epsilon\mathrm{div}E^0))(\mathrm{div}(\mathrm{div}E^0U^\epsilon)+R_n^\epsilon)dx\\
&\le-\frac{d}{dt}\iint_0^{N^\epsilon}(h(1-\epsilon\mathrm{div}E^0+s)-h(1-\epsilon\mathrm{div}E^0))dsdx\\
&\quad+C(\|W^\epsilon(t)\|_{l,*}+1)\|(N^\epsilon,U^\epsilon,\nabla U^\epsilon)\|^2+C\epsilon^2.
\end{aligned}\tag{4.8}$$

Combining (4.5) with (4.6)–(4.8), we have

$$\begin{aligned}
&\frac{d}{dt}\left[(U^\epsilon,U^\epsilon)+\iint_0^{N^\epsilon}(h(1-\epsilon\mathrm{div}E^0+s)-h(1-\epsilon\mathrm{div}E^0))dsdx\right]+2(F^\epsilon,U^\epsilon)\\
&\le C(\|W^\epsilon(t)\|_{l,*}+1)\|(N^\epsilon,U^\epsilon,\nabla U^\epsilon,\gamma G^\epsilon)\|^2+C\epsilon^2.
\end{aligned}\tag{4.9}$$

Multiplying the first equation of the Maxwell system in the error system (3.2) by $\frac{1}{\gamma}F^\epsilon$ and the second one by $\frac{1}{\gamma}G^\epsilon$, by integration by parts, we get

$$\begin{aligned}
&\frac{d}{dt}(\epsilon\|F^\epsilon\|^2+\|G^\epsilon\|^2)+\frac{2}{\gamma}\int(\nabla\times F^\epsilon\cdot G^\epsilon-\nabla\times G^\epsilon\cdot F^\epsilon)dx-2(U^\epsilon,F^\epsilon)\\
&=2(N^\epsilon(U^\epsilon+u^0),F^\epsilon)-2(\epsilon\mathrm{div}E^0U^\epsilon,F^\epsilon)+\frac{2}{\gamma}(\epsilon R_E^\epsilon,F^\epsilon).
\end{aligned}\tag{4.10}$$

On one hand, using the vector analysis formulas (1.9), the term $O(\frac{1}{\gamma})$ appearing in Sobolev's energy estimates vanishes, i.e.,

$$\int(\nabla\times F^\epsilon\cdot G^\epsilon-\nabla\times G^\epsilon\cdot F^\epsilon)dx=\int \mathrm{div}(F^\epsilon\times G^\epsilon)dx=0. \tag{4.11}$$

On the other hand, using Young's inequality, we have

$$\begin{aligned}(N^\epsilon(U^\epsilon+u^0),F^\epsilon)&\le\epsilon\|F^\epsilon\|^2+\frac{1}{4\epsilon}\|N^\epsilon(U^\epsilon+u^0)\|^2\\&\le\epsilon\|F^\epsilon\|^2+C(\|U^\epsilon(t)\|_l^2+1)\left\|\frac{N^\epsilon}{\sqrt{\epsilon}}\right\|^2,\end{aligned} \tag{4.12}$$

where a singular term $\|\frac{N^\epsilon}{\sqrt{\epsilon}}\|$ appears, and

$$-2(\epsilon\mathrm{div}E^0U^\epsilon,F^\epsilon)+\frac{2}{\gamma}(\epsilon R_E^\epsilon,F^\epsilon)\le\epsilon\|F^\epsilon\|^2+C\epsilon\|U^\epsilon\|^2+C\epsilon. \tag{4.13}$$

Thus, combining (4.10) with (4.11)–(4.13), we get

$$\begin{aligned}&\frac{d}{dt}(\epsilon\|F^\epsilon\|^2+\|G^\epsilon\|^2)-2(U^\epsilon,F^\epsilon)\\&\le\epsilon\|F^\epsilon\|^2+C\epsilon\|U^\epsilon\|^2+C(\|U^\epsilon(t)\|_l^2+1)\left\|\frac{N^\epsilon}{\sqrt{\epsilon}}\right\|^2+C\epsilon.\end{aligned} \tag{4.14}$$

It follows from (4.9) and (4.14) that

$$\begin{aligned}&\frac{d}{dt}\int\left\{|U^\epsilon|^2+\int_0^{N^\epsilon}(h(1-\epsilon\mathrm{div}E^0+s)-h(1-\epsilon\mathrm{div}E^0))ds+\epsilon|F^\epsilon|^2+|G^\epsilon|^2\right\}(t)dx\\&\le\|\sqrt{\epsilon}F^\epsilon\|^2+C(\|W^\epsilon(t)\|_{l,*}^2+\|W^\epsilon(t)\|_{l,*}+1)\left\|\left(N^\epsilon,U^\epsilon,\nabla U^\epsilon,\gamma G^\epsilon,\frac{N^\epsilon}{\sqrt{\epsilon}}\right)\right\|^2\\&\quad+C\epsilon^2+C\epsilon.\end{aligned}$$

This completes the proof of Lemma 4.1. □

Note that it follows from the estimate (4.4) that our problem is the need to obtain an estimate not just for N^ϵ but also for $\frac{N^\epsilon}{\sqrt{\epsilon}}$. In particular, the order of derivatives that we need to estimate for the latter is one less than for the other main quantities $(N^\epsilon,U^\epsilon,G^\epsilon)$. However, straightforward Sobolev energy estimates of higher order would require the same number for both. Hence the above method cannot be generalized directly to the high order Sobolev energy estimates. Here the idea is to establish the uniform estimates for $\|\frac{N^\epsilon}{\sqrt{\epsilon}}(t)\|_{l-1}$ through a priori estimates of vorticity and divergence. To this end, we require the equations of vorticity and divergence.

4.3. Derivation of the vorticity and divergence equations. Taking *curl* on the momentum equation and using $\gamma\partial_tG^\epsilon+\nabla\times F^\epsilon=0$ of the magnetic field in the error system (3.2), we have

$$\begin{aligned}&\partial_t(\nabla\times U^\epsilon)+\nabla\times([(U^\epsilon+u^0)\cdot\nabla]U^\epsilon)+\nabla\times((U^\epsilon\cdot\nabla)u^0)-\gamma\partial_tG^\epsilon\\&=-\gamma\nabla\times\big((U^\epsilon+u^0)\times G^\epsilon+U^\epsilon\times B^0\big).\end{aligned} \tag{4.15}$$

Since $\mathrm{div}G^\epsilon = 0$, there exists a vector function $\mathcal{G}^\epsilon$ such that

$$G^\epsilon = \nabla \times \mathcal{G}^\epsilon. \tag{4.16}$$

Using the vector analysis formulas (1.10) and (1.11), we get

$$\nabla \times ([(U^\epsilon + u^0)\cdot\nabla]U^\epsilon) = \nabla \times \big((\nabla \times U^\epsilon) \times (U^\epsilon + u^0)\big) - \nabla \times \big(\nabla u^0 \cdot U^\epsilon\big). \tag{4.17}$$

Then, putting (4.16) and (4.17) into the equations (4.15), we obtain

$$\partial_t(\nabla \times (U^\epsilon - \gamma\mathcal{G}^\epsilon)) + \nabla \times ((\nabla \times (U^\epsilon - \gamma\mathcal{G}^\epsilon)) \times (U^\epsilon + u^0)) = \mathcal{J}_1^\epsilon, \tag{4.18}$$

where

$$\mathcal{J}_1^\epsilon = \nabla \times (\nabla u^0 \cdot U^\epsilon) - \nabla \times ((U^\epsilon \cdot \nabla)u^0) - \gamma\nabla \times (U^\epsilon \times B^0)$$

satisfies

$$\|\mathcal{J}_1^\epsilon\|_{l-1} \le C\|U^\epsilon\|_l. \tag{4.19}$$

Here we require $u^0 \in H^{l+1}$.

Next, introduce the general vorticity

$$\omega^\epsilon = \nabla \times (U^\epsilon - \gamma\mathcal{G}^\epsilon);$$

then it follows from (4.18) that ω^ϵ satisfies the following vorticity equation:

$$\partial_t\omega^\epsilon + (U^\epsilon + u^0)\cdot\nabla\omega^\epsilon - \omega^\epsilon \cdot \nabla(U^\epsilon + u^0) + \omega^\epsilon \mathrm{div}(U^\epsilon + u^0) = \mathcal{J}_1^\epsilon. \tag{4.20}$$

Here we have again used the vector formulas (1.12).

Taking *div* on the momentum equations in the error system (3.2) and using the equation $\epsilon\mathrm{div}F^\epsilon = -N^\epsilon$, we can obtain the following divergence equation:

$$\begin{aligned}&\partial_t \mathrm{div}U^\epsilon + \mathrm{div}([(U^\epsilon + u^0)\cdot\nabla]U^\epsilon) - \frac{N^\epsilon}{\epsilon} + \Delta(h(N^\epsilon + 1 - \epsilon\mathrm{div}E^0) - h(1 - \epsilon\mathrm{div}E^0))\\ &= -\gamma\mathrm{div}\big((U^\epsilon + u^0) \times G^\epsilon\big) + \mathcal{J}_2^\epsilon,\end{aligned} \tag{4.21}$$

where

$$\mathcal{J}_2^\epsilon = -\epsilon\mathrm{div}(R_u^\epsilon) - \mathrm{div}((U^\epsilon \cdot \nabla)u^0) - \gamma\mathrm{div}(U^\epsilon \times B^0)$$

satisfies

$$\|\mathcal{J}_2^\epsilon\|_{l-1} \le C\|U^\epsilon\|_l + C\epsilon. \tag{4.22}$$

4.4. Estimates of the vorticity and divergence. Now we control $\|\nabla \times U^\epsilon\|_{l-1}$ using the vorticity equations (4.20).

LEMMA 4.2. *For any $0 < t < T$ and $l > \frac{3}{2} + 2$, it holds that*

$$\begin{aligned}&\|\nabla \times U^\epsilon(t)\|_{l-1}^2\\ &\le C(\|\nabla \times U^\epsilon(t=0)\|_{l-1}^2 + \gamma^2\|G^\epsilon(t=0)\|_{l-1}^2) + C\gamma^2\|G^\epsilon(t)\|_{l-1}^2\\ &\quad + C\int_0^t (\|W^\epsilon(s)\|_{l,*} + 1)(\|\nabla \times U^\epsilon(s)\|_{l-1}^2 + \gamma^2\|G^\epsilon(s)\|_{l-1}^2)ds.\end{aligned} \tag{4.23}$$

Proof. Let $\alpha \in \mathbb{N}^3$ with $|\alpha| \le l-1$. Taking ∂_x^α on (4.20) and multiplying the resulting equation by $\partial_x^\alpha \omega^\epsilon$, by integration by parts, we have the basic Friedrich energy equation

$$\begin{aligned}\frac{d}{dt}\|\partial_x^\alpha \omega^\epsilon\|^2 &= \big(\mathrm{div}(U^\epsilon + u^0)\partial_x^\alpha \omega^\epsilon, \partial_x^\alpha \omega\big) + 2(H_\alpha^{(1)}, \partial_x^\alpha \omega)\\ &\quad + 2\Big(\partial_x^\alpha\big(\omega^\epsilon \cdot \nabla(U^\epsilon + u^0) - \omega^\epsilon \mathrm{div}(U^\epsilon + u^0) + \mathcal{J}_1^\epsilon\big), \partial_x^\alpha \omega^\epsilon\Big), \end{aligned} \tag{4.24}$$

where the commutator $H_\alpha^{(1)}$ is defined by

$$H_\alpha^{(1)} = -[\partial_x^\alpha((U^\epsilon + u^0)\cdot\nabla\omega^\epsilon) - (U^\epsilon + u^0)\cdot\nabla\partial_x^\alpha\omega^\epsilon],$$

which can be estimated as follows:

$$\begin{aligned}\|H_\alpha^{(1)}\| &\le C(\|\nabla(U^\epsilon + u^0)\|_{L^\infty}\|\partial_x^{l-2}\nabla\omega^\epsilon\| + \|\nabla\omega^\epsilon\|_{L^\infty}\|\partial_x^{l-1}(U^\epsilon + u^0)\|)\\ &\le C(\|\nabla(U^\epsilon + u^0)\|_{l-1}\|\partial_x^{l-2}\nabla\omega^\epsilon\| + \|\nabla\omega^\epsilon\|_{l-2}\|\partial_x^{s-1}(U^\epsilon + u^0)\|)\\ &\le C(\|W^\epsilon(t)\|_{l,*} + 1)\|\omega^\epsilon\|_{l-1}. \end{aligned} \tag{4.25}$$

Here we have used Sobolev's lemma and $l > \frac{3}{2} + 2$.

Using the estimates (4.25) for $H_\alpha^{(1)}$ and (4.19) for $\mathcal{J}_1^\epsilon$, we have, with the aid of Cauchy–Schwarz's inequality and Sobolev's lemma,

$$(\mathrm{div}(U^\epsilon + u^0)\partial_x^\alpha\omega^\epsilon, \partial_x^\alpha\omega) \le C(\|W^\epsilon(t)\|_{l,*} + 1)\|\omega^\epsilon\|_{l-1}^2, \tag{4.26}$$

$$2(H_\alpha^{(1)}, \partial_x^\alpha\omega) \le C(\|W^\epsilon(t)\|_{l,*} + 1)\|\omega^\epsilon\|_{l-1}^2 \tag{4.27}$$

and

$$\begin{aligned}&\Big(\partial_x^\alpha\big(\omega^\epsilon\cdot\nabla(U^\epsilon + u^0) - \omega^\epsilon\mathrm{div}(U^\epsilon + u^0) + \mathcal{J}_1^\epsilon\big), \partial_x^\alpha\omega^\epsilon\Big)\\ &\le C(\|W^\epsilon(t)\|_{l,*} + 1)\|\omega^\epsilon\|_{l-1}^2. \end{aligned} \tag{4.28}$$

Combining (4.24) with (4.26), (4.27), and (4.28), we get

$$\frac{d}{dt}\|\omega^\epsilon\|_{l-1}^2 \le C(\|W^\epsilon(t)\|_{l,*} + 1)\|\omega^\epsilon\|_{l-1}^2,$$

which yields, for any $0 < t < T$,

$$\|\omega^\epsilon(t)\|_{l-1}^2 \le C\|\omega^\epsilon(t=0)\|_{l-1}^2 + C\int_0^t (\|W^\epsilon(s)\|_{l,*} + 1)\|\omega^\epsilon(s)\|_{l-1}^2 ds. \tag{4.29}$$

Using the definition of ω^ϵ, we get

$$\|\nabla\times U^\epsilon\|_{l-1}^2 \le 2\|\omega^\epsilon\|_{l-1}^2 + 2\gamma^2\|G^\epsilon\|_{l-1}^2 \tag{4.30}$$

and

$$\|\omega^\epsilon(t)\|_{l-1}^2 \le 2\|\nabla\times U^\epsilon(t)\|_{l-1}^2 + 2\gamma^2\|G^\epsilon\|_{l-1}^2. \tag{4.31}$$

Then (4.29)–(4.31) give the estimate (4.23).

The proof of Lemma 4.2 is complete. ∎

Next, we estimate $\|\mathrm{div}U^\epsilon\|_{l-1}$ using the divergence equation (4.21). The estimate is contained in the following lemma, whose proof is long and is postponed to the appendix.

LEMMA 4.3. *Let $\alpha \in \mathbb{N}^3$ with $|\alpha| \le l-1$ and $l > \frac{3}{2}+2$. Then, for any $0 < t < T$, we have*

$$\begin{aligned}
&\left[\|\partial_x^\alpha \mathrm{div}U^\epsilon\|^2 + \frac{1}{\epsilon}\left(\frac{1}{1-\epsilon\mathrm{div}E^0}\partial_x^\alpha N^\epsilon, \partial_x^\alpha N^\epsilon\right)\right.\\
&\qquad \left. + \left(\frac{h'(N^\epsilon+1-\epsilon\mathrm{div}E^0)}{N^\epsilon+1-\epsilon\mathrm{div}E^0}\partial_x^\alpha \nabla N^\epsilon, \partial_x^\alpha \nabla N^\epsilon\right)\right](t)\\
&\le C\left[\|\partial_x^\alpha \mathrm{div}U^\epsilon\|^2 + \frac{1}{\epsilon}\left(\frac{1}{1-\epsilon\mathrm{div}E^0}\partial_x^\alpha N^\epsilon, \partial_x^\alpha N^\epsilon\right)\right.\\
&\qquad \left. + \left(\frac{h'(N^\epsilon+1-\epsilon\mathrm{div}E^0)}{N^\epsilon+1-\epsilon\mathrm{div}E^0}\partial_x^\alpha \nabla N^\epsilon, \partial_x^\alpha \nabla N^\epsilon\right)\right](t=0)\\
&\quad + C\int_0^t (\|W^\epsilon(s)\|_{l,*}+1)\|W^\epsilon(s)\|_{l,*}^2 ds\\
&\quad + C\gamma^2\int_0^t (\|U^\epsilon(s)\|_l^2+1)\|G^\epsilon(s)\|_l^2 ds + C\epsilon^2 + C\epsilon. \qquad (4.32)
\end{aligned}$$

4.5. High order energy estimates on the electromagnetic field. Finally, we derive the high order energy estimates on the electromagnetic field. For the electric field, we establish its estimates by using the wave formulas of the Maxwell equations as follows.

LEMMA 4.4. *For any $0 < t \le T$ and $l > \frac{3}{2}+2$, it holds that*

$$\begin{aligned}
&(\epsilon^2\gamma^2\|\partial_t F^\epsilon\|_{l-1}^2 + \epsilon\gamma^2\|F^\epsilon\|_{l-1}^2 + \epsilon\|\nabla\times F^\epsilon\|_{l-1}^2)(t)\\
&\le (\epsilon^2\gamma^2\|\partial_t F^\epsilon\|_{l-1}^2 + \epsilon\gamma^2\|F^\epsilon\|_{l-1}^2 + \epsilon\|\nabla\times F^\epsilon\|_{l-1}^2)(t=0)\\
&\quad + C(\gamma^2+1)\int_0^t (|||W^\epsilon(t)|||_l^4 + |||W^\epsilon(t)|||_l^2+1)|||W^\epsilon(t)|||_l^2(s)ds + C\gamma^2\epsilon^2. \qquad (4.33)
\end{aligned}$$

Proof. It follows from the Maxwell equation in the error equation (3.2) that F^ϵ satisfies the following wave-type equation:

$$\epsilon\gamma\partial_{tt}F^\epsilon - \frac{1}{\gamma}\Delta F^\epsilon + \frac{1}{\gamma}\nabla\mathrm{div}F^\epsilon + \gamma F^\epsilon = -\gamma N^\epsilon F^\epsilon + \gamma\epsilon\mathrm{div}E^0 F^\epsilon + \mathcal{J}_3^\epsilon, \qquad (4.34)$$

where

$$\begin{aligned}
\mathcal{J}_3 &= \gamma(N^\epsilon+1-\epsilon\mathrm{div}E^0)\Big(-[(U^\epsilon+u^0)\cdot\nabla]U^\epsilon - (U^\epsilon\cdot\nabla)u^0\\
&\quad - \nabla(h(N^\epsilon+1-\epsilon\mathrm{div}E^0) - h(1-\epsilon\mathrm{div}E^0)) - \gamma\big((U^\epsilon+u^0)\times G^\epsilon + U^\epsilon\times B^0\big)\\
&\quad + \epsilon h'(1-\epsilon\mathrm{div}E^0)\nabla(\mathrm{div}E^0)\Big) + \gamma\partial_t(N^\epsilon+1-\epsilon\mathrm{div}E^0)U^\epsilon\\
&\quad + \gamma u^0\partial_t N^\epsilon + \gamma\partial_t u^0 N^\epsilon + \epsilon\partial_t(\partial_t E^0 + u^0\mathrm{div}E^0) \qquad (4.35)
\end{aligned}$$

satisfies

$$\|\mathcal{J}_3\|_{l-1} \le C\gamma(|||W^\epsilon(t)|||_l^2 + |||W^\epsilon(t)|||_l + 1)|||W^\epsilon(t)|||_l + C\gamma\epsilon. \tag{4.36}$$

Let $\alpha \in \mathbb{N}^3$ with $|\alpha| \le l-1$. Taking ∂_x^α on (4.34), multiplying the resulting equation by $\partial_t \partial_x^\alpha F^\epsilon$, by integration by parts, we get

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int \left(\epsilon\gamma|\partial_t\partial_x^\alpha F^\epsilon|^2 + \gamma|\partial_x^\alpha F^\epsilon|^2 + \frac{1}{\gamma}|\partial_x^\alpha(\nabla\times F^\epsilon)|^2\right)dx \\
&= -\gamma\int \partial_x^\alpha(N^\epsilon F^\epsilon)\partial_t\partial_x^\alpha F^\epsilon dx + \gamma\epsilon\int \partial_x^\alpha(\mathrm{div}E^0F^\epsilon)\partial_t\partial_x^\alpha F^\epsilon dx \\
&\quad + \int \partial_x^\alpha \mathcal{J}_3\partial_t\partial_x^\alpha F^\epsilon dx.
\end{aligned} \tag{4.37}$$

In the following we control the right-hand side of (4.37) by $\gamma\|F^\epsilon\|_{l-1}^2$ and $\gamma\epsilon\|\partial_t F^\epsilon\|_{l-1}^2$. First, it holds that

$$\begin{aligned}
&-\gamma\int \partial_x^\alpha(N^\epsilon F^\epsilon)\partial_t\partial_x^\alpha F^\epsilon dx \\
&= -\gamma\int N^\epsilon\partial_x^\alpha F^\epsilon\partial_t\partial_x^\alpha F^\epsilon dx - \gamma\int \mathcal{H}_\alpha^{(10)}\partial_t\partial_x^\alpha F^\epsilon dx \\
&\le \frac{\gamma}{6}\int|\partial_x^\alpha F^\epsilon|^2dx + 6\gamma\|N^\epsilon\|_{L^\infty}^2\int|\partial_t\partial_x^\alpha F^\epsilon|^2dx + \gamma\|\mathcal{H}_\alpha^{(10)}\|\|\partial_t\partial_x^\alpha F^\epsilon\| \\
&\le \frac{\gamma}{6}\int|\partial_x^\alpha F^\epsilon|^2dx + C\gamma\|N^\epsilon\|_{l-1}^2\int|\partial_t\partial_x^\alpha F^\epsilon|^2dx + C\gamma\|N^\epsilon\|_{l-1}\|F^\epsilon\|_{l-1}\|\partial_t\partial_x^\alpha F^\epsilon\| \\
&\le \frac{\gamma}{3}\|F^\epsilon\|_{l-1}^2 + C\gamma\|N^\epsilon\|_{l-1}^2\|\partial_t F^\epsilon\|_{l-1}^2 \\
&= \frac{\gamma}{3}\|F^\epsilon\|_{l-1}^2 + C\left\|\frac{N^\epsilon}{\sqrt{\epsilon}}\right\|_{l-1}^2\epsilon\gamma\|\partial_t F^\epsilon\|_{l-1}^2,
\end{aligned} \tag{4.38}$$

where the commutator

$$\mathcal{H}_\alpha^{(10)} = \partial_x^\alpha(N^\epsilon F^\epsilon) - N^\epsilon\partial_x^\alpha F^\epsilon$$

can be estimated by

$$\begin{aligned}
\|\mathcal{H}_\alpha^{(10)}\| &\le C\|\nabla N^\epsilon\|_{L^\infty}\|\partial_x^{l-2}F^\epsilon\| + C\|F^\epsilon\|_{L^\infty}\|\partial_x^{l-1}N^\epsilon\| \\
&\le C\|\nabla N^\epsilon\|_{l-2}\|\partial_x^{l-2}F^\epsilon\| + C\|F^\epsilon\|_{l-1}\|\partial_x^{l-1}N^\epsilon\| \\
&\le C\|N^\epsilon\|_{l-1}\|F^\epsilon\|_{l-1}.
\end{aligned}$$

Then, as in the establishment of (4.38), we can get

$$\gamma\epsilon\int \partial_x^\alpha(\mathrm{div}E^0F^\epsilon)\partial_t\partial_x^\alpha F^\epsilon dx \le C\epsilon\gamma(\|F^\epsilon\|_{l-1}^2 + \|\partial_t F^\epsilon\|_{l-1}^2). \tag{4.39}$$

Finally, by Cauchy–Schwarz's inequality, we have

$$\int \partial_x^\alpha\mathcal{J}_3\partial_t\partial_x^\alpha F^\epsilon dx \le \frac{C}{\epsilon\gamma}\|\partial_x^\alpha\mathcal{J}_3\|^2 + C\epsilon\gamma\|\partial_x^\alpha\partial_t F^\epsilon\|^2. \tag{4.40}$$

Combining (4.37) with (4.38), (4.39), and (4.40), we get

$$\frac{1}{2}\frac{d}{dt}\int\left(\epsilon\gamma|\partial_t\partial_x^\alpha F^\epsilon|^2+\gamma|\partial_x^\alpha F^\epsilon|^2+\frac{1}{\gamma}|\partial_x^\alpha(\nabla\times F^\epsilon)|^2\right)dx$$

$$\le\frac{\gamma}{3}\|F^\epsilon\|_{l-1}^2+C\left\|\frac{N^\epsilon}{\sqrt{\epsilon}}\right\|_{l-1}^2\epsilon\gamma\|\partial_t F^\epsilon\|_{l-1}^2+C\epsilon\gamma(\|F^\epsilon\|_{l-1}^2+\|\partial_t F^\epsilon\|_{l-1}^2)$$

$$+\frac{C}{\epsilon\gamma}\|\partial_x^\alpha\mathcal{J}_3\|^2+C\epsilon\gamma\|\partial_x^\alpha\partial_t F^\epsilon\|^2, \tag{4.41}$$

from which we can easily obtain (4.33).

The proof of Lemma 4.4 is complete. □

We estimate the magnetic field by using the estimates on the electric field and the curl-div decomposition technique of the gradient for the magnetic field.

LEMMA 4.5. *For any* $0<t\le T$ *and* $l>\frac{3}{2}+2$, *it holds that*

$$\|\nabla G^\epsilon\|_{l-1}\le\|\epsilon\gamma\partial_t F^\epsilon\|_{l-1}+C\gamma(\|N^\epsilon\|_{l-1}+1)\|U^\epsilon\|_{l-1}+C\gamma\|N^\epsilon\|_{l-1}+C\gamma\epsilon. \tag{4.42}$$

Proof. Because $\mathrm{div}G^\epsilon=0$, by the curl-div decomposition formulas of the gradient for the magnetic field

$$\|\nabla G^\epsilon\|_{l-1}\le\|\nabla\times G^\epsilon\|_{l-1}+\|\mathrm{div}G^\epsilon\|_{l-1}, \tag{4.43}$$

it suffices to control $\|\nabla\times G^\epsilon\|_{l-1}$.

Using the third equation in the error system (3.2), we get

$$\|\nabla\times G^\epsilon\|_{l-1}\le\|\epsilon\gamma\partial_t F^\epsilon\|_{l-1}+\gamma\|(N^\epsilon+1-\epsilon\mathrm{div}E^0)U^\epsilon+N^\epsilon u^0\|_{l-1}+C\gamma\epsilon$$

$$\le\|\epsilon\gamma\partial_t F^\epsilon\|_{l-1}+C\gamma(\|N^\epsilon\|_{l-1}+1)\|U^\epsilon\|_{l-1}+C\gamma\|N^\epsilon\|_{l-1}+C\gamma\epsilon. \tag{4.44}$$

Combining (4.43) and (4.44), we obtain (4.42).

The proof of Lemma 4.5 is complete. □

4.6. The end of proof of Proposition 4.1. Introduce an ϵ-weighted Sobolev-type energy function

$$\Gamma^\epsilon(t)=\||W^\epsilon(t)\||_l^2.$$

Then it follows from (4.4), (4.23), (4.32), (4.33), and (4.42) that there exists an $\epsilon_0>0$, depending only upon T_0, such that, for any $0<\epsilon\le\epsilon_0$ and any $0<t<T$,

$$\Gamma^\epsilon(t)\le C\Gamma^\epsilon(t=0)+C\int_0^t((\Gamma^\epsilon+\sqrt{\Gamma^\epsilon}+1)\Gamma^\epsilon)(s)ds+C\epsilon. \tag{4.45}$$

Then, applying Gronwall's lemma and using Proposition 3.1 and (3.4), it follows from (4.45) and $\Gamma^\epsilon(t=0)\le C\epsilon$ that there exist a $0<T_1<1$ and an $\epsilon_0>0$ such that $T_\epsilon\ge T_1$ for all $0<\epsilon\le\epsilon_0$ and that there exists an ϵ_0 sufficiently small such that, for any $\epsilon\le\epsilon_0$ and $0<t<T$,

$$\Gamma^\epsilon(t)\le C\epsilon,$$

which gives the desired a priori estimate (4.2). Moreover, by the standard continuous induction method, we can extend $T_\epsilon\ge T_0$ for any $T_0<T_*$.

The proof of Proposition 4.1 is complete. □

4.7. Proof of Theorem 2.1. According to the definitions of the error functions $N^\epsilon, U^\epsilon, F^\epsilon, G^\epsilon$, the regularities of u^0, E^0, B^0, and the error system (3.2), it follows from the assumption (2.11) in Theorem 2.1 that

$$\|\sqrt{\epsilon}F^\epsilon(t=0)\|_{s_0-1} \le C\sqrt{\epsilon},$$

$$\left\|\frac{N^\epsilon}{\sqrt{\epsilon}}(t=0)\right\|_{s_0-1} = \sqrt{\epsilon}\|\mathrm{div}F^\epsilon(t=0)\|_{s_0-1} \le \sqrt{\epsilon}\|F^\epsilon(t=0)\|_{s_0} \le C\sqrt{\epsilon}.$$

Hence, the assumption (4.1) in Proposition 4.1 holds. Thus, the results of Proposition 4.1 imply (2.12). As a consequence, $(n^\epsilon, u^\epsilon, B^\epsilon)$ converges strongly to $(1, u^0, B^0)$ in $L^\infty(0,T_0;H^{s_0}(\mathbb{T}))$. It follows from (2.2), (2.6), and the uniqueness of solutions to the limit problem (2.6)–(2.9) that E^ϵ converges to E^0 in $W^{-1,\infty}(0,T_0;H^{s_0-1}(\mathbb{T}))$.

The proof of Theorem 2.1 is complete. ∎

Appendix A. Proof of Lemma 4.3. Taking ∂_x^α on (4.21) and taking the L^2 inner product of the resulting equation with $\partial_x^\alpha \mathrm{div}U^\epsilon$, by integration by parts, we have the following energy equation:

$$\begin{aligned}
(\mathrm{A}.1)\ & \frac{d}{dt}\|\partial_x^\alpha \mathrm{div}U^\epsilon\|^2 \\
&= (\mathrm{div}(U^\epsilon+u^0)\partial_x^\alpha \mathrm{div}U^\epsilon, \partial_x^\alpha \mathrm{div}U^\epsilon) + 2(\mathcal{H}_\alpha^{(2)}, \partial_x^\alpha \mathrm{div}U^\epsilon) \\
&\quad + \frac{2}{\epsilon}(\partial_x^\alpha N^\epsilon, \partial_x^\alpha \mathrm{div}U^\epsilon) - 2\big(\partial_x^\alpha \Delta(h(N^\epsilon+1-\epsilon \mathrm{div}E^0) - h(1-\epsilon \mathrm{div}E^0)), \partial_x^\alpha \mathrm{div}U^\epsilon\big) \\
&\quad - 2\gamma\Big(\partial_x^\alpha \mathrm{div}\big((U^\epsilon+u^0)\times G^\epsilon\big), \partial_x^\alpha \mathrm{div}U^\epsilon\Big) + 2(\partial_x^\alpha \mathcal{J}_2^\epsilon, \partial_x^\alpha \mathrm{div}U^\epsilon),
\end{aligned}$$

where the commutator is defined by

$$\mathcal{H}_\alpha^{(2)} = -\{\partial_x^\alpha \mathrm{div}([(U^\epsilon+u^0)\cdot\nabla]U^\epsilon) - [(U^\epsilon+u^0)\cdot\nabla]\partial_x^\alpha \mathrm{div}U^\epsilon\},$$

which can be estimated as follows:

$$\begin{aligned}
\|\mathcal{H}_\alpha^{(2)}\| &\le C\|\nabla(U^\epsilon+u^0)\|_{L^\infty}\|\partial_x^{l-1}\nabla U^\epsilon\| + C\|\nabla U^\epsilon\|_{L^\infty}\|\partial_x^l(U^\epsilon+u^0)\| \\
&\le C\|\nabla(U^\epsilon+u^0)\|_{l-1}\|\partial_x^{l-1}\nabla U^\epsilon\| + C\|\nabla U^\epsilon\|_{l-1}\|\partial_x^l(U^\epsilon+u^0)\| \\
&\le C(\|U^\epsilon\|_l+1)\|U^\epsilon\|_l.
\end{aligned}$$

Hence, by Cauchy–Schwarz's inequality, Sobolev's lemma, and using the estimate (4.22) for $\mathcal{J}_2^\epsilon$, we obtain

$$\begin{aligned}
&(\mathrm{div}(U^\epsilon+u^0)\partial_x^\alpha \mathrm{div}U^\epsilon, \partial_x^\alpha \mathrm{div}U^\epsilon) + 2(\mathcal{H}_\alpha^{(2)}, \partial_x^\alpha \mathrm{div}U^\epsilon) \\
&\le C(\|U^\epsilon\|_l+1)\|U^\epsilon\|_l^2 + C\|\mathcal{H}_\alpha^{(2)}\|\|\partial_x^\alpha \mathrm{div}U^\epsilon\| \\
&\le C(\|W^\epsilon(t)\|_{l,*}+1)\|U^\epsilon\|_l^2 \qquad (\mathrm{A}.2)
\end{aligned}$$

and

$$\begin{aligned}
&-2\gamma\Big(\partial_x^\alpha \mathrm{div}\big((U^\epsilon+u^0)\times G^\epsilon\big), \partial_x^\alpha \mathrm{div}U^\epsilon\Big) + 2(\partial_x^\alpha \mathcal{J}_2^\epsilon, \partial_x^\alpha \mathrm{div}U^\epsilon) \\
&\le C\gamma^2(\|U^\epsilon\|_l^2+1)\|G^\epsilon\|_l^2 + \|W^\epsilon(t)\|_{l,*}^2 + C\epsilon^2. \qquad (\mathrm{A}.3)
\end{aligned}$$

The rest involves dealing with the other two terms on the right-hand side of (A.1), which are more difficult to control. To do this, we rewrite the density equation in (3.2) into the following two formulations:

$$\mathrm{div}U^\epsilon = -\frac{\partial_t N^\epsilon + \mathrm{div}(N^\epsilon(U^\epsilon+u^0)) - \epsilon U^\epsilon\cdot\nabla(\mathrm{div}E^0) - \epsilon R_n^\epsilon}{1-\epsilon\mathrm{div}E^0} \tag{A.4}$$

and

$$\mathrm{div}U^\epsilon = -\frac{\partial_t N^\epsilon + (U^\epsilon+u^0)\cdot\nabla N^\epsilon + N^\epsilon\mathrm{div}u^0 - \epsilon U^\epsilon\cdot\nabla(\mathrm{div}E^0) - \epsilon R_n^\epsilon}{N^\epsilon+1-\epsilon\mathrm{div}E^0}. \tag{A.5}$$

In the following, we use the first formulation (A.4) to estimate the electric field term and the second one (A.5) to estimate the nonlinear pressure term integral in order to avoid the presence of the $(l+1)$th order derivative of the velocity because we are now in H^l energy estimates.

First, we control the singular term $O(\frac{1}{\epsilon})$ in (A.1). We have

$$\begin{aligned}
I_1 &= \frac{2}{\epsilon}(\partial_x^\alpha N^\epsilon, \partial_x^\alpha \mathrm{div}U^\epsilon)\\
&= -\frac{2}{\epsilon}\left(\partial_x^\alpha N^\epsilon, \partial_x^\alpha\left(\frac{\partial_t N^\epsilon + \mathrm{div}(N^\epsilon(U^\epsilon+u^0)) - \epsilon U^\epsilon\cdot\nabla(\mathrm{div}E^0) - \epsilon R_n^\epsilon}{1-\epsilon\mathrm{div}E^0}\right)\right)\\
&= -\frac{2}{\epsilon}\left(\partial_x^\alpha N^\epsilon, \partial_x^\alpha\left(\frac{\partial_t N^\epsilon}{1-\epsilon\mathrm{div}E^0}\right)\right) - \frac{2}{\epsilon}\left(\partial_x^\alpha N^\epsilon, \partial_x^\alpha\left(\frac{\mathrm{div}(N^\epsilon(U^\epsilon+u^0))}{1-\epsilon\mathrm{div}E^0}\right)\right)\\
&\quad + 2\left(\partial_x^\alpha N^\epsilon, \partial_x^\alpha\left(\frac{U^\epsilon\cdot\nabla(\mathrm{div}E^0)}{1-\epsilon\mathrm{div}E^0}\right)\right) + 2\left(\partial_x^\alpha N^\epsilon, \partial_x^\alpha\left(\frac{R_n^\epsilon}{1-\epsilon\mathrm{div}E^0}\right)\right)\\
&= \sum_{i=1}^4 I_1^{(i)},
\end{aligned} \tag{A.6}$$

in which each term can be estimated as follows.

For $I_1^{(1)}$, by Cauchy–Schwarz's inequality and using the fact that

$$\left\|\partial_t\left(\frac{1}{1-\epsilon\mathrm{div}E^0}\right)\right\|_{L^\infty} \le C\epsilon,$$

we have

$$\begin{aligned}
I_1^{(1)} &= -\frac{2}{\epsilon}\left(\partial_x^\alpha N^\epsilon, \partial_x^\alpha\left(\frac{\partial_t N^\epsilon}{1-\epsilon\mathrm{div}E^0}\right)\right)\\
&= -\frac{2}{\epsilon}\left(\frac{1}{1-\epsilon\mathrm{div}E^0}\partial_x^\alpha N^\epsilon, \partial_x^\alpha\partial_t N^\epsilon\right) - \frac{2}{\epsilon}(\mathcal{H}_\alpha^{(2)}, \partial_x^\alpha N^\epsilon)\\
&= -\frac{1}{\epsilon}\frac{d}{dt}\left(\frac{1}{1-\epsilon\mathrm{div}E^0}\partial_x^\alpha N^\epsilon, \partial_x^\alpha N^\epsilon\right) + \frac{1}{\epsilon}\left(\partial_t\left(\frac{1}{1-\epsilon\mathrm{div}E^0}\right)\partial_x^\alpha N^\epsilon, \partial_x^\alpha N^\epsilon\right)\\
&\quad - \frac{2}{\epsilon}(\mathcal{H}_\alpha^{(2)}, \partial_x^\alpha N^\epsilon)\\
&\le -\frac{1}{\epsilon}\frac{d}{dt}\left(\frac{1}{1-\epsilon\mathrm{div}E^0}\partial_x^\alpha N^\epsilon, \partial_x^\alpha N^\epsilon\right) + C\|\partial_x^\alpha N^\epsilon\|^2 + \frac{C}{\epsilon}\|\mathcal{H}_\alpha^{(3)}\|\|\partial_x^\alpha N^\epsilon\|,
\end{aligned} \tag{A.7}$$

where the commutator is defined by

$$\mathcal{H}_\alpha^{(3)} = \partial_x^\alpha\left(\frac{\partial_t N^\epsilon}{1-\epsilon \mathrm{div} E^0}\right) - \frac{1}{1-\epsilon \mathrm{div} E^0}\partial_x^\alpha \partial_t N^\epsilon,$$

which can be controlled as follows:

$$\begin{aligned}\|\mathcal{H}_\alpha^{(3)}\| &\le C\left\|\nabla\left(\frac{1}{1-\epsilon \mathrm{div} E^0}\right)\right\|_{L^\infty}\|\partial_x^{l-2}\partial_t N^\epsilon\| + C\|\partial_t N^\epsilon\|_{L^\infty}\left\|\partial_x^{l-1}\left(\frac{1}{1-\epsilon \mathrm{div} E^0}\right)\right\| \\ &\le C\epsilon\|\partial_t N^\epsilon\|_{l-2} \\ &\le C\epsilon\|(N^\epsilon, U^\epsilon)\|_{l-1} + C\epsilon(\|U^\epsilon\|_{l-1}+1)\|N^\epsilon\|_{l-1} + C\epsilon^2. \qquad \text{(A.8)}\end{aligned}$$

Here we have used $\|\nabla(\frac{1}{1-\epsilon \mathrm{div} E^0})\|_{L^\infty} \le C\epsilon$ and $l > \frac{3}{2} + 2$. Combining (A.7) and (A.8), we obtain

$$I_1^{(1)} \le -\frac{1}{\epsilon}\frac{d}{dt}\left(\frac{1}{1-\epsilon \mathrm{div} E^0}\partial_x^\alpha N^\epsilon, \partial_x^\alpha N^\epsilon\right) + C(\|U^\epsilon\|_{l-1}+1)\|W^\epsilon(t)\|_{l,*}^2 + C\epsilon^2. \qquad \text{(A.9)}$$

For $I_1^{(2)}$, a direct calculation yields

$$\begin{aligned} I_1^{(2)} &= -\frac{2}{\epsilon}\left(\partial_x^\alpha N^\epsilon, \partial_x^\alpha\left(\frac{\mathrm{div}(N^\epsilon(U^\epsilon+u^0))}{1-\epsilon \mathrm{div} E^0}\right)\right) \\ &= -\frac{2}{\epsilon}\left(\partial_x^\alpha N^\epsilon, \frac{1}{1-\epsilon \mathrm{div} E^0}\partial_x^\alpha\big(\mathrm{div}(N^\epsilon(U^\epsilon+u^0))\big)\right) - \frac{2}{\epsilon}(\partial_x^\alpha N^\epsilon, \mathcal{H}_\alpha^{(4)}) \\ &= -\frac{2}{\epsilon}\left(\partial_x^\alpha N^\epsilon, \frac{1}{1-\epsilon \mathrm{div} E^0}\partial_x^\alpha\big((U^\epsilon+u^0)\cdot\nabla N^\epsilon + N^\epsilon \mathrm{div}(U^\epsilon+u^0)\big)\right) \\ &\quad - \frac{2}{\epsilon}(\partial_x^\alpha N^\epsilon \mathcal{H}_\alpha^{(4)}) \\ &= \frac{1}{\epsilon}\left(\partial_x^\alpha N^\epsilon, \mathrm{div}\left(\frac{1}{1-\epsilon \mathrm{div} E^0}(U^\epsilon+u^0)\right)\partial_x^\alpha N^\epsilon\right) \\ &\quad - \frac{2}{\epsilon}\left(\partial_x^\alpha N^\epsilon, \frac{1}{1-\epsilon \mathrm{div} E^0}\partial_x^\alpha N^\epsilon \mathrm{div}(U^\epsilon+u^0)\right) \\ &\quad - \frac{2}{\epsilon}(\partial_x^\alpha N^\epsilon, \mathcal{H}_\alpha^{(4)}) - \frac{2}{\epsilon}\sum_{i=5}^{i=6}\left(\frac{1}{1-\epsilon \mathrm{div} E^0}\partial_x^\alpha N^\epsilon, \mathcal{H}_\alpha^{(i)}\right), \end{aligned} \qquad \text{(A.10)}$$

where the commutators $\mathcal{H}_\alpha^{(i)}, i = 4, 5, 6$, are defined by

$$\begin{aligned}\mathcal{H}_\alpha^{(4)} &= \partial_x^\alpha\left(\frac{\mathrm{div}(N^\epsilon(U^\epsilon+u^0))}{1-\epsilon \mathrm{div} E^0}\right) - \frac{1}{1-\epsilon \mathrm{div} E^0}\partial_x^\alpha\big(\mathrm{div}(N^\epsilon(U^\epsilon+u^0))\big), \\ \mathcal{H}_\alpha^{(5)} &= \partial_x^\alpha\big((U^\epsilon+u^0)\cdot\nabla N^\epsilon\big) - (U^\epsilon+u^0)\cdot\partial_x^\alpha\nabla N^\epsilon, \\ \mathcal{H}_\alpha^{(6)} &= \partial_x^\alpha(N^\epsilon \mathrm{div}(U^\epsilon+u^0)) - \partial_x^\alpha N^\epsilon \mathrm{div}(U^\epsilon+u^0),\end{aligned}$$

which, with the aid of Sobolev's lemma, can be estimated, respectively, as follows:

$$\begin{aligned}\|\mathcal{H}_\alpha^{(4)}\| &\le C\left\|\nabla\left(\frac{1}{1-\epsilon\mathrm{div}E^0}\right)\right\|_{L^\infty}\|\partial_x^{l-2}(\mathrm{div}(N^\epsilon(U^\epsilon+u^0)))\|\\ &\quad + C\|\mathrm{div}(N^\epsilon(U^\epsilon+u^0))\|_{L^\infty}\left\|\partial_x^{l-1}\left(\frac{1}{1-\epsilon\mathrm{div}E^0}\right)\right\|\\ &\le C\left\|\nabla\left(\frac{1}{1-\epsilon\mathrm{div}E^0}\right)\right\|_{L^\infty}\|\partial_x^{l-2}(\mathrm{div}(N^\epsilon(U^\epsilon+u^0)))\|\\ &\quad + C\|\mathrm{div}(N^\epsilon(U^\epsilon+u^0))\|_{l-2}\left\|\partial_x^{l-1}\left(\frac{1}{1-\epsilon\mathrm{div}E^0}\right)\right\|\\ &\le C\epsilon(\|U^\epsilon\|_l+1)\|N^\epsilon\|_{l-1}, \qquad \text{(A.11)}\end{aligned}$$

$$\begin{aligned}\|\mathcal{H}_\alpha^{(5)}\| &\le C\|\nabla(U^\epsilon+u^0)\|_{L^\infty}\|\partial_x^{l-2}\nabla N^\epsilon\| + C\|\nabla N^\epsilon\|_{L^\infty}\|\partial_x^{l-1}(U^\epsilon+u^0)\|\\ &\le C\|\nabla(U^\epsilon+u^0)\|_{l-1}\|\partial_x^{l-2}\nabla N^\epsilon\| + C\|\nabla N^\epsilon\|_{l-2}\|\partial_x^{l-1}(U^\epsilon+u^0)\|\\ &\le C(\|U^\epsilon\|_l+1)\|N^\epsilon\|_{l-1}, \qquad \text{(A.12)}\end{aligned}$$

$$\begin{aligned}\|\mathcal{H}_\alpha^{(6)}\| &\le C\|\nabla\mathrm{div}(U^\epsilon+u^0)\|_{L^\infty}\|\partial_x^{l-2}N^\epsilon\| + C\|N^\epsilon\|_{L^\infty}\|\partial_x^{l-1}\mathrm{div}(U^\epsilon+u^0)\|\\ &\le C\|\nabla\mathrm{div}(U^\epsilon+u^0)\|_{l-2}\|\partial_x^{l-2}N^\epsilon\| + C\|N^\epsilon\|_{l-1}\|\partial_x^{l-1}\mathrm{div}(U^\epsilon+u^0)\|\\ &\le C(\|U^\epsilon\|_l+1)\|N^\epsilon\|_{l-1}. \qquad \text{(A.13)}\end{aligned}$$

Here we have again used $\|\nabla(\frac{1}{1-\epsilon\mathrm{div}E^0})\|_{L^\infty}\le C\epsilon$ and $l>\frac{3}{2}+2$.

Thus, combining (A.10) together with (A.11)–(A.13), we get, with the aid of Cauchy–Schwarz's inequality, that

$$I_1^{(2)} \le \frac{C}{\epsilon}(\|U^\epsilon\|_l+1)\|N^\epsilon\|_{l-1}^2. \qquad \text{(A.14)}$$

Also, by Cauchy–Schwarz's inequality, $I_1^{(3)}$ can be estimated as follows:

$$I_1^{(3)} = 2\left(\partial_x^\alpha N^\epsilon, \partial_x^\alpha\left(\frac{U^\epsilon\cdot\nabla(\mathrm{div}E^0)}{1-\epsilon\mathrm{div}E^0}\right)\right) \le C\|W^\epsilon(t)\|_{l,*}^2. \qquad \text{(A.15)}$$

Finally, we estimate $I_1^{(4)}$. Since R_n^ϵ is not small with respect to ϵ (only uniformly bounded), we must use $\|\frac{N^\epsilon}{\sqrt{\epsilon}}\|_{l-1}$ to control $I_1^{(4)}$. Thus, we rewrite $I_1^{(4)}$ as

$$I_1^{(4)} = 2\left(\partial_x^\alpha N^\epsilon, \partial_x^\alpha\left(\frac{R_n^\epsilon}{1-\epsilon\mathrm{div}E^0}\right)\right) = 2\left(\frac{\partial_x^\alpha N^\epsilon}{\sqrt{\epsilon}}, \sqrt{\epsilon}\partial_x^\alpha\left(\frac{R_n^\epsilon}{1-\epsilon\mathrm{div}E^0}\right)\right).$$

By Cauchy–Schwarz's inequality and the uniform bound on R_n^ϵ, we have

$$I_1^{(4)} \le \frac{C}{\epsilon}\|N^\epsilon\|_{l-1}^2 + C\epsilon. \qquad \text{(A.16)}$$

Combining (A.6) with (A.9), (A.14), (A.15), and (A.16), we obtain

$$I_1 \le -\frac{1}{\epsilon}\frac{d}{dt}\left(\frac{1}{1-\epsilon\mathrm{div}E^0}\partial_x^\alpha N^\epsilon, \partial_x^\alpha N^\epsilon\right) + C(\|U^\epsilon\|_l+1)|||W^\epsilon(t)|||_l^2 + C\epsilon. \qquad \text{(A.17)}$$

Next, we control the pressure term I_2. It holds that

$$\begin{aligned}
I_2 &= -2(\partial_x^\alpha \Delta(h(N^\epsilon+1-\epsilon \mathrm{div}E^0)-h(1-\epsilon \mathrm{div}E^0)), \partial_x^\alpha \mathrm{div}U^\epsilon)\\
&= -2\Big(\partial_x^\alpha \mathrm{div}\big(h'(N^\epsilon+1-\epsilon \mathrm{div}E^0)\nabla N^\epsilon\\
&\qquad -\epsilon(h'(N^\epsilon+1-\epsilon \mathrm{div}E^0)-h'(1-\epsilon \mathrm{div}E^0))\nabla(\mathrm{div}E^0)\big), \partial_x^\alpha \mathrm{div}U^\epsilon\Big)\\
&= -2\Big(\mathrm{div}\big(h'(N^\epsilon+1-\epsilon \mathrm{div}E^0)\partial_x^\alpha \nabla N^\epsilon\big), \partial_x^\alpha \mathrm{div}U^\epsilon\Big) - 2(\mathcal{H}_\alpha^{(7)}, \partial_x^\alpha \mathrm{div}U^\epsilon)\\
&\quad + 2\epsilon\Big(\partial_x^\alpha \mathrm{div}\big((h'(N^\epsilon+1-\epsilon \mathrm{div}E^0)-h'(1-\epsilon \mathrm{div}E^0))\nabla(\mathrm{div}E^0)\big), \partial_x^\alpha \mathrm{div}U^\epsilon\Big)
\end{aligned}$$

$$(\text{A.18}) = I_2^{(1)} + I_2^{(2)} + I_2^{(3)},$$

where

$$\begin{aligned}
I_2^{(1)} &= -2\Big(\mathrm{div}\big(h'(N^\epsilon+1-\epsilon \mathrm{div}E^0)\nabla \partial_x^\alpha N^\epsilon\big), \partial_x^\alpha \mathrm{div}U^\epsilon\Big),\\
I_2^{(2)} &= -2(\mathcal{H}_\alpha^{(7)}, \partial_x^\alpha \mathrm{div}U^\epsilon),\\
I_2^{(3)} &= 2\epsilon\Big(\partial_x^\alpha \mathrm{div}\big((h'(N^\epsilon+1-\epsilon \mathrm{div}E^0)-h'(1-\epsilon \mathrm{div}E^0))\nabla(\mathrm{div}E^0)\big), \partial_x^\alpha \mathrm{div}U^\epsilon\Big)
\end{aligned}$$

and the commutator $\mathcal{H}_\alpha^{(7)}$ is defined by

$$\begin{aligned}
\mathcal{H}_\alpha^{(7)} &= \partial_x^\alpha \mathrm{div}\big(h'(N^\epsilon+1-\epsilon \mathrm{div}E^0)\nabla N^\epsilon\big) - \mathrm{div}\big(h'(N^\epsilon+1-\epsilon \mathrm{div}E^0)\partial_x^\alpha \nabla N^\epsilon\big)\\
&= \partial_x^\alpha\big(h'(N^\epsilon+1-\epsilon \mathrm{div}E^0)\mathrm{div}\nabla N^\epsilon\big) - h'(N^\epsilon+1-\epsilon \mathrm{div}E^0)\partial_x^\alpha \mathrm{div}\nabla N^\epsilon\\
&\quad + \partial_x^\alpha\big(\nabla h'(N^\epsilon+1-\epsilon \mathrm{div}E^0)\nabla N^\epsilon\big) - \nabla h'(N^\epsilon+1-\epsilon \mathrm{div}E^0)\partial_x^\alpha \nabla N^\epsilon.
\end{aligned}$$

Noting that $I_2^{(3)}$ does contain only the lth order derivatives of the error function W_I^ϵ, it can be easily estimated by

$$I_2^{(3)} \le \|U^\epsilon\|_l^2 + C\|N^\epsilon\|_l^2. \tag{A.19}$$

Here we have used the regularity of the limit system, i.e., $\|E^0\|_{l+2} \le C$. The estimate techniques of the commutator yield

$$\begin{aligned}
\|\mathcal{H}_\alpha^{(7)}\| &\le C\|\nabla h'(N^\epsilon+1-\epsilon \mathrm{div}E^0)\|_{L^\infty}\|\partial_x^{l-2}\mathrm{div}\nabla N^\epsilon\|\\
&\quad + C\|\mathrm{div}\nabla N^\epsilon\|_{L^\infty}\|\partial_x^{l-1}h'(N^\epsilon+1-\epsilon \mathrm{div}E^0)\|\\
&\quad + C\|\nabla^2 h'(N^\epsilon+1-\epsilon \mathrm{div}E^0)\|_{L^\infty}\|\partial_x^{l-2}\nabla N^\epsilon\|\\
&\quad + C\|\nabla N^\epsilon\|_{L^\infty}\|\partial_x^{l-1}\nabla h'(N^\epsilon+1-\epsilon \mathrm{div}E^0)\|\\
&\le C(\|N^\epsilon\|_l + 1)\|N^\epsilon\|_l,
\end{aligned}$$

which implies by Cauchy–Schwarz's inequality that

$$I_2^{(2)} \le C(\|N^\epsilon\|_l + 1)\|W^\epsilon(t)\|_{l,*}^2. \tag{A.20}$$

In the following, we estimate $I_2^{(1)}$. By the relation (A.5) between the density and the divergence, we have

$$
\begin{aligned}
I_2^{(1)} &= 2\bigg(\mathrm{div}\big(h'(N^\epsilon+1-\epsilon\mathrm{div}E^0)\partial_x^\alpha\nabla N^\epsilon\big),\\
&\qquad \partial_x^\alpha\bigg(\frac{\partial_t N^\epsilon+(U^\epsilon+u^0)\cdot\nabla N^\epsilon-\epsilon U^\epsilon\cdot\nabla(\mathrm{div}E^0)+N^\epsilon\mathrm{div}u^0-\epsilon R_n^\epsilon}{N^\epsilon+1-\epsilon\mathrm{div}E^0}\bigg)\bigg)\\
&= -2\bigg(h'(N^\epsilon+1-\epsilon\mathrm{div}E^0)\partial_x^\alpha\nabla N^\epsilon,\nabla\partial_x^\alpha\bigg(\frac{\partial_t N^\epsilon}{N^\epsilon+1-\epsilon\mathrm{div}E^0}\bigg)\bigg)\\
&\quad -2\bigg(h'(N^\epsilon+1-\epsilon\mathrm{div}E^0)\partial_x^\alpha\nabla N^\epsilon,\nabla\partial_x^\alpha\bigg(\frac{(U^\epsilon+u^0)\cdot\nabla N^\epsilon}{N^\epsilon+1-\epsilon\mathrm{div}E^0}\bigg)\bigg)\\
&\quad -2\bigg(h'(N^\epsilon+1-\epsilon\mathrm{div}E^0)\partial_x^\alpha\nabla N^\epsilon,\\
&\qquad \nabla\partial_x^\alpha\bigg(\frac{-\epsilon U^\epsilon\cdot\nabla(\mathrm{div}E^0)+N^\epsilon\mathrm{div}u^0-\epsilon R_n^\epsilon}{N^\epsilon+1-\epsilon\mathrm{div}E^0}\bigg)\bigg)\\
&= \sum_{i=1}^{3} I_{21}^{(i)}. \qquad \text{(A.21)}
\end{aligned}
$$

Now we estimate each term of $I_2^{(1)}$.

For $I_{21}^{(1)}$ and $I_{21}^{(2)}$, we have

$$
\begin{aligned}
I_{21}^{(1)} &= -2\bigg(h'(N^\epsilon+1-\epsilon\mathrm{div}E^0)\partial_x^\alpha\nabla N^\epsilon,\nabla\partial_x^\alpha\bigg(\frac{\partial_t N^\epsilon}{N^\epsilon+1-\epsilon\mathrm{div}E^0}\bigg)\bigg)\\
&= -\frac{d}{dt}\bigg(\frac{h'(N^\epsilon+1-\epsilon\mathrm{div}E^0)}{(N^\epsilon+1-\epsilon\mathrm{div}E^0)}\partial_x^\alpha\nabla N^\epsilon,\partial_x^\alpha\nabla N^\epsilon\bigg)\\
&\quad +\bigg(\partial_t\bigg(\frac{h'(N^\epsilon+1-\epsilon\mathrm{div}E^0)}{(N^\epsilon+1-\epsilon\mathrm{div}E^0)}\bigg)\partial_x^\alpha\nabla N^\epsilon,\partial_x^\alpha\nabla N^\epsilon\bigg)\\
&\quad -2\bigg(h'(N^\epsilon+1-\epsilon\mathrm{div}E^0)\partial_x^\alpha\nabla N^\epsilon,\mathcal{H}_\alpha^{(8)}\bigg)\\
&\le -\frac{d}{dt}\bigg(\frac{h'(N^\epsilon+1-\epsilon\mathrm{div}E^0)}{2(N^\epsilon+1-\epsilon\mathrm{div}E^0)}\partial_x^\alpha\nabla N^\epsilon,\partial_x^\alpha\nabla N^\epsilon\bigg)\\
&\quad +C(\|W^\epsilon(t)\|_{l,*}+1)\|W^\epsilon(t)\|_{l,*}^2 \qquad \text{(A.22)}
\end{aligned}
$$

and

$$I_{21}^{(2)} = -2\Big(h'(N^\epsilon+1-\epsilon\mathrm{div}E^0)\partial_x^\alpha\nabla N^\epsilon, \nabla\partial_x^\alpha\Big(\frac{(U^\epsilon+u^0)\cdot\nabla N^\epsilon}{N^\epsilon+1-\epsilon\mathrm{div}E^0}\Big)\Big)$$

$$= \Big(\mathrm{div}\Big(\frac{h'(N^\epsilon+1-\epsilon\mathrm{div}E^0)(U^\epsilon+u^0)}{(N^\epsilon+1-\epsilon\mathrm{div}E^0)}\Big)\partial_x^\alpha\nabla N^\epsilon, \partial_x^\alpha\nabla N^\epsilon\Big)$$

$$-2\Big(h'(N^\epsilon+1-\epsilon\mathrm{div}E^0)\partial_x^\alpha\nabla N^\epsilon, \mathcal{H}_\alpha^{(9)}\Big)$$

$$\le C(\|W^\epsilon(t)\|_{l,*}+1)\|W^\epsilon(t)\|_{l,*}^2 \tag{A.23}$$

because the commutators

$$\mathcal{H}_\alpha^{(8)} = \nabla\partial_x^\alpha\Big(\frac{\partial_t N^\epsilon}{N^\epsilon+1-\epsilon\mathrm{div}E^0}\Big) - \frac{\partial_t\nabla\partial_x^\alpha N^\epsilon}{N^\epsilon+1-\epsilon\mathrm{div}E^0}$$

and

$$\mathcal{H}_\alpha^{(9)} = \nabla\partial_x^\alpha\Big(\frac{(U^\epsilon+u^0)\cdot\nabla N^\epsilon}{N^\epsilon+1-\epsilon\mathrm{div}E^0}\Big) - \frac{(U^\epsilon+u^0)\cdot\nabla(\partial_x^\alpha\nabla N^\epsilon)}{N^\epsilon+1-\epsilon\mathrm{div}E^0}$$

can be estimated, respectively, by

$$\|\mathcal{H}_\alpha^{(8)}\| \le C\Big\|\nabla\Big(\frac{1}{N^\epsilon+1-\epsilon\mathrm{div}E^0}\Big)\Big\|_{L^\infty}\|\partial_x^{l-1}\partial_t N^\epsilon\|$$

$$+C\|\partial_t N^\epsilon\|_{L^\infty}\Big\|\partial_x^l\Big(\frac{1}{N^\epsilon+1-\epsilon\mathrm{div}E^0}\Big)\Big\|$$

$$\le C(\|W^\epsilon(t)\|_{l,*}+1)\|W^\epsilon(t)\|_{l,*}+C\epsilon^2$$

and

$$\|\mathcal{H}_\alpha^{(9)}\| \le C\Big\|\nabla\Big(\frac{U^\epsilon+u^0}{N^\epsilon+1-\epsilon\mathrm{div}E^0}\Big)\Big\|_{L^\infty}\|\partial_x^{l-1}\nabla N^\epsilon\|$$

$$+C\|\nabla N^\epsilon\|_{L^\infty}\Big\|\partial_x^l\Big(\frac{U^\epsilon+u^0}{N^\epsilon+1-\epsilon\mathrm{div}E^0}\Big)\Big\|$$

$$\le C(\|W^\epsilon(t)\|_{l,*}+1)\|W^\epsilon(t)\|_{l,*}+C\epsilon^2.$$

For $I_{21}^{(3)}$, by Cauchy–Schwarz's inequality and the regularities of u^0, E^0, we have

$$I_{21}^{(3)} = -2\Big(h'(N^\epsilon+1-\epsilon\mathrm{div}E^0)\partial_x^\alpha\nabla N^\epsilon,$$

$$\nabla\partial_x^\alpha\Big(\frac{-\epsilon U^\epsilon\cdot\nabla(\mathrm{div}E^0)+N^\epsilon\mathrm{div}u^0-\epsilon R_n^\epsilon}{N^\epsilon+1-\epsilon\mathrm{div}E^0}\Big)\Big)$$

$$\le C\|W^\epsilon(t)\|_{l,*}^3 + C\|W^\epsilon(t)\|_{l,*}^2 + C\epsilon^2. \tag{A.24}$$

Combining (A.21) with (A.22)–(A.24), we get

$$I_2^{(1)} \le -\frac{d}{dt}\left(\frac{h'(N^\epsilon+1-\epsilon \mathrm{div}E^0)}{(N^\epsilon+1-\epsilon \mathrm{div}E^0)}, \partial_x^\alpha \nabla N^\epsilon, \partial_x^\alpha \nabla N^\epsilon\right)$$

$$+ C(\|W^\epsilon(t)\|_{l,*}+1)\|W^\epsilon(t)\|_{l,*}^2 + C\epsilon^2. \tag{A.25}$$

Similarly, combining (A.18) with (A.19), (A.20), and (A.25), we get

$$I_2 \le -\frac{d}{dt}\left(\frac{h'(N^\epsilon+1-\epsilon \mathrm{div}E^0)}{(N^\epsilon+1-\epsilon \mathrm{div}E^0)}\partial_x^\alpha \nabla N^\epsilon, \partial_x^\alpha \nabla N^\epsilon\right)$$

$$+ C(\|W^\epsilon(t)\|_{l,*}+1)\|W^\epsilon(t)\|_{l,*}^2 + C\epsilon^2, \tag{A.26}$$

and combining (A.1) with (A.2), (A.3), (A.17), and (A.26), we get

$$\frac{d}{dt}\left[\|\partial_x^\alpha \mathrm{div}U^\epsilon\|^2 + \frac{1}{\epsilon}\left(\frac{1}{1-\epsilon \mathrm{div}E^0}\partial_x^\alpha N^\epsilon, \partial_x^\alpha N^\epsilon\right)\right.$$

$$\left. + \left(\frac{h'(N^\epsilon+1-\epsilon \mathrm{div}E^0)}{N^\epsilon+1-\epsilon \mathrm{div}E^0}\partial_x^\alpha \nabla N^\epsilon, \partial_x^\alpha \nabla N^\epsilon\right)\right]$$

$$\le C(\|W^\epsilon\|_{l,*}+1)\|W^\epsilon(t)\|_{l,*}^2 + C\gamma^2(\|U^\epsilon\|_l^2+1)\|G^\epsilon\|_l^2 + C\epsilon^2 + C\epsilon,$$

which yields (4.32).

This ends the proof of Lemma 4.3. □

Acknowledgments. The authors are very grateful to both referees for their constructive comments and helpful suggestions, which considerably improved the presentation of the paper. The second author would like to express his gratitude for the hospitality of the Laboratory of Mathematics at Blaise Pascal University, France, when he visited there in May through July 2006.

REFERENCES

[1] C. Besse, P. Degond, F. Deluzet, J. Claudel, G. Gallice, and C. Tessieras, *A model hierarchy for ionospheric plasma modeling*, Math. Models Methods Appl. Sci., 14 (2004), pp. 393–415.

[2] Y. Brenier, *Convergence of the Vlasov-Poisson system to the incompressible Euler equations*, Comm. Partial Differential Equations, 25 (2000), pp. 737–754.

[3] Y. Brenier, N. J. Mauser, and M. Puel, *Incompressible Euler and e-MHD as scaling limits of the Vlasov-Maxwell system*, Commun. Math. Sci., 1 (2003), pp. 437–447.

[4] H. Brézis, F. Golse, and R. Sentis, *Analyse asymptotique de l'équation de Poisson couplée à la relation de Boltzmann. Quasi-neutralité des plasmas*, C. R. Acad. Sci. Paris Sér. I Math., 321 (1995), pp. 953–959.

[5] F. Chen, *Introduction to Plasma Physics and Controlled Fusion*, Vol. 1, Plenum Press, New York, 1984.

[6] G. Q. Chen, J. W. Jerome, and D. H. Wang, *Compressible Euler-Maxwell equations*, Transport Theory Statist. Phys., 29 (2000), pp. 311–331.

[7] S. Cordier and E. Grenier, *Quasineutral limit of an Euler-Poisson system arising from plasma physics*, Comm. Partial Differential Equations, 25 (2000), pp. 1099–1113.

[8] E. Grenier, *Pseudo-differential energy estimates of singular perturbations*, Comm. Pure Appl. Math., 50 (1997), pp. 821–865.

[9] T. Kato, *Nonstationary flow of viscous and ideal fluids in* $\mathbb{R}^3$, J. Funct. Anal., 9 (1972), pp. 296–305.

[10] A. Kingsep, K. Chukbar, and V. Yankov, *Electron magnetohydrodynamics*, in Review of Plasma Physics 16, B. B. Kadomtsev, ed., Consultants Bureau, New York, 1990, pp. 243–291.

[11] S. Klainerman and A. Majda, *Singular limits of quasilinear hyperbolic systems with large parameters and the incompressible limit of compressible fluids*, Comm. Pure Appl. Math., 34 (1981), pp. 481–524.

[12] S. Klainerman and A. Majda, *Compressible and incompressible fluids*, Comm. Pure Appl. Math., 35 (1982), pp. 629–651.

[13] H. O. Kreiss, J. Lorenz, and M. J. Naughton, *Convergence of the solutions of the compressible to the solutions of the incompressible Navier-Stokes equations*, Adv. in Appl. Math., 12 (1991), pp. 187–214.

[14] P. L. Lions, *Mathematical Topics in Fluid Mechanics, Vol.* 1*: Incompressible Models*, Oxford Lecture Ser. Math. Appl. 3, Oxford University Press, New York, 1996.

[15] A. Majda, *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*, Springer-Verlag, New York, 1984.

[16] F. J. McGrath, *Nonstationary plane flow of viscous and ideal fluids*, Arch. Rational Mech. Anal., 27 (1968), pp. 229–348.

[17] Y. J. Peng and S. Wang, *Convergence of compressible Euler-Maxwell equations to compressible Euler-Poisson equations*, Chinese Ann. Math. Ser. B, 28 (2007), pp. 583–602.

[18] Y. J. Peng and S. Wang, *Convergence of compressible Euler-Maxwell equations to incompressible Euler equations*, Comm. Partial Differential Equations, 33 (2008), pp. 349–376.

[19] Y. J. Peng and S. Wang, *Asymptotic expansions in two-fluid compressible Euler-Maxwell equations with small parameters*, Discrete Contin. Dyn. Syst., to appear.

[20] Y. J. Peng and Y. G. Wang, *Convergence of compressible Euler-Poisson equations to incompressible type Euler equations*, Asymptot. Anal., 41 (2005), pp. 141–160.

[21] H. Rishbeth and O. K. Garriott, *Introduction to Ionospheric Physics*, Academic Press, New York, 1969.

[22] S. Schochet, *Fast singular limits of hyperbolic PDEs*, J. Differential Equations, 114 (1994), pp. 476–512.

[23] M. Slemrod and N. Sternberg, *Quasi-neutral limit for the Euler-Poisson system*, J. Nonlinear Sci., 11 (2001), pp. 193–209.

[24] S. Wang, *Quasineutral limit of Euler-Poisson system with and without viscosity*, Comm. Partial Differential Equations, 29 (2004), pp. 419–456.

[25] S. Wang and S. Jiang, *The convergence of the Navier-Stokes-Poisson system to the incompressible Euler equations*, Comm. Partial Differential Equations, 31 (2006), pp. 571–591.

SIAM J. MATH. ANAL.
Vol. 40, No. 2, pp. 566–604

SYMMETRY-BREAKING BIFURCATION IN NONLINEAR SCHRÖDINGER/GROSS–PITAEVSKII EQUATIONS*

E. W. KIRR†, P. G. KEVREKIDIS‡, E. SHLIZERMAN§, AND M. I. WEINSTEIN¶

Abstract. We consider a class of nonlinear Schrödinger/Gross–Pitaeveskii (NLS-GP) equations, i.e., NLS with a linear potential. NLS-GP plays an important role in the mathematical modeling of nonlinear optical as well as macroscopic quantum phenomena (BEC). We obtain conditions for a symmetry-breaking bifurcation in a symmetric family of states as $\mathcal{N}$, the squared L^2 norm (particle number, optical power), is increased. The bifurcating asymmetric state is a "mixed mode" which, near the bifurcation point, is approximately a superposition of symmetric and antisymmetric modes. In the special case where the linear potential is a double well with well-separation L, we estimate $\mathcal{N}_{cr}(L)$, the symmetry breaking threshold. Along the "lowest energy" symmetric branch, there is an exchange of stability from the symmetric to the asymmetric branch as $\mathcal{N}$ is increased beyond $\mathcal{N}_{cr}$.

Key words. nonlinear Schrödinger, Gross–Pitaevskii, soliton, bound state

AMS subject classifications. 35Q55, 37K45, 37K50

DOI. 10.1137/060678427

1. Introduction.

Symmetry breaking is a ubiquitous and important phenomenon which arises in a wide range of physical systems. In this paper, we consider a class of partial differential equations (PDEs), which are invariant under a symmetry group. For sufficiently small values of a parameter, $\mathcal{N}$, the preferred (dynamically stable) stationary state of the system is invariant under this symmetry group. However, above a critical parameter, $\mathcal{N}_{cr}$, although the group-invariant state persists, the new preferred state of the system is a state which (i) exists only for $\mathcal{N} > \mathcal{N}_{cr}$ and (ii) is no longer invariant. That is, symmetry is broken and there is an exchange of stability.

Physical examples of symmetry breaking include liquid crystals [31], quantum dots [34], semiconductor lasers [13], and pattern dynamics [28]. This article focuses on spontaneous symmetry breaking as a phenomenon in nonlinear optics [4, 20, 18], as well as in the macroscopic quantum setting of Bose–Einstein condensation (BEC) [1]. Here, the governing equations are PDEs of nonlinear Schrödinger/Gross–Pitaevskii (NLS-GP) type. Symmetry breaking has been observed experimentally in optics for two-component spatial optical vector solitons (i.e., for self-guided laser beams in Kerr media and focusing cubic nonlinearities) in [4], as well as for the electric field distribution between multiple wells of a photorefractive crystal in [20, 18]. In BECs,

*Received by the editors December 22, 2006; accepted for publication (in revised form) July 25, 2007; published electronically June 6, 2008.
http://www.siam.org/journals/sima/40-2/67842.html

†Department of Mathematics, University of Ilinois, Urbana-Champaign, Urbana, IL 61801 (ekirr@math.uiuc.edu). The work of this author was partially supported by grants DMS-0405921 and DMS-060372.

‡Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003 (kevrekid@math.umass.edu). The work of this author was partially supported by DMS-0204585, NSF-CAREER, and DMS-0505663.

§Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel (eli.shlizerman@weizmann.ac.il). Part of this research was done while this author was a visiting graduate student in the Department of Applied Physics and Applied Mathematics at Columbia University.

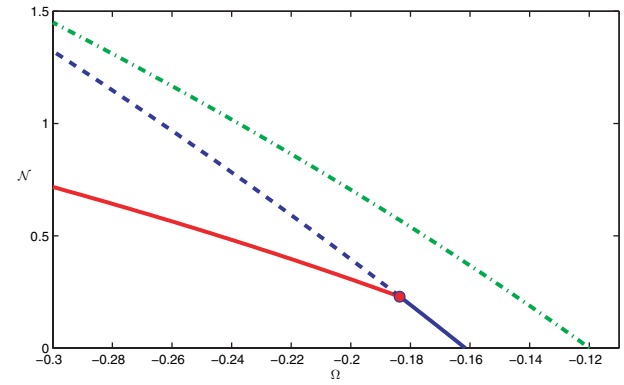
¶Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027 (miw2103@columbia.edu). The work of this author was partially supported by DMS-0412305 and DMS-0530853.

an effective double well formed by a combined (parabolic) magnetic trapping and a (periodic) optical trapping of the atoms may have similar effects [1] and lead to "macroscopic quantum self-trapping."

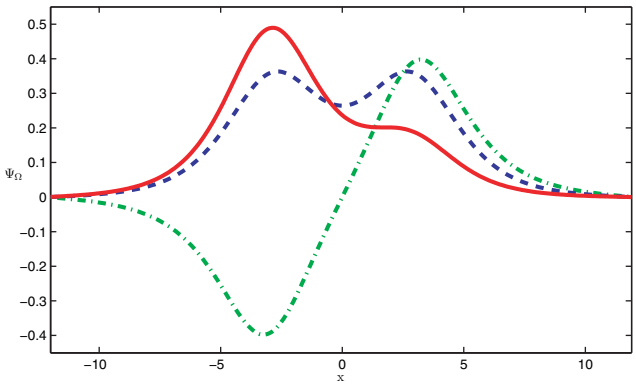
Symmetry breaking in ground states of the three-dimensional NLS-GP equation, with an attractive nonlinearity of Hartree type and a symmetric double-well linear potential, was considered in Aschbacher et. al. [3]; see also Remark 2.2. An example of a double-well potential is the function $V_L(x)$, plotted for the one-dimensional case, in Figure 2b of Example 2.1. The spectral properties of $H_L = -\Delta + V_L(x)$ are discussed in the appendix in section 8. For L sufficiently large, H_L has (at least) two eigenpairs, (Ω_0, ψ_0) and (Ω_1, ψ_1). The normalized ground state eigenfunction ψ_0 is of even parity in x ("symmetric") and the normalized excited state ψ_1 is of odd parity in x ("antisymmetric"). Moreover, $\psi_0(x) \sim 2^{-\frac{1}{2}}(\psi_\omega(x+L) + \psi_\omega(x-L))$ and $\psi_1(x) \sim 2^{-\frac{1}{2}}(\psi_\omega(x+L) - \psi_\omega(x-L))$.

Ground states of NLS-GP are positive nonlinear bound states, arising as *minimizers* of $\mathcal{H}$, the NLS-GP Hamiltonian energy subject to fixed $\mathcal{N}$, the squared L^2 norm. For the class of equations considered in [3], ground states exist for any $\mathcal{N} > 0$. It is proved that for sufficiently large $\mathcal{N}$, any ground state is concentrated in only one of the wells, i.e., symmetry is broken. The analysis in [3] is an asymptotic study of a variational problem (see Remark 2.2) showing that if $\mathcal{N}$ is sufficiently large, then it is energetically preferable for the ground state to localize in a single well. In contrast, for small L^2 norm the ground state is even and bimodal having the symmetries of the ground state of the *linear* Schrödinger operator with symmetric double-well potential. For macroscopic quantum systems, the squared L^2 norm, denoted by $\mathcal{N}$, is the particle number, while in optics it is the optical power. An attractive nonlinearity corresponds to the case of negative scattering length in BEC and focusing Kerr nonlinearity in optics.

An alternative approach to symmetry breaking in NLS-GP is via bifurcation theory. It follows from [26, 25] that a family of "nonlinear ground states" bifurcates from the zero solution ($\mathcal{N} = 0$) at the ground state energy of the Schrödinger operator with a linear double-well potential. This nonlinear ground state branch consists of states having the same bimodal symmetry of the linear ground state. In this article we prove, under suitable conditions, that there is a secondary bifurcation to an asymmetric state at a critical $\mathcal{N} = \mathcal{N}_{cr} > 0$. The bifurcating asymmetric state is a "mixed mode" (see, for example, [8]) which, near the bifurcation point, is approximately a superposition of symmetric and antisymmetric modes. As $\mathcal{N}$ is increased beyond $\mathcal{N}_{cr}$, the asymmetric state tends to concentrate in only one of the two wells. Since the double-well potential is symmetric, the bifurcating state is doubly degenerate; there are two families of asymmetric states associated with energy concentration in each of the wells. Figure 1 shows a typical bifurcation diagram demonstrating symmetry breaking for the NLS-GP system with a double-well potential. At the bifurcation point $\mathcal{N}_{cr}$ (marked by a circle in Figure 1(a)), the symmetric (even with respect to coordinate x_1) ground state becomes unstable and a stable asymmetric (neither even nor odd parity) "mixed (degenerate) state" emanates from it. Further, we show that there is a transfer or exchange of stability which takes place at $\mathcal{N}_{cr}$; for $\mathcal{N} < \mathcal{N}_{cr}$ the symmetric state is stable, while for $\mathcal{N} > \mathcal{N}_{cr}$ the asymmetric state is stable. Since our method is based on local bifurcation analysis we do not require that the states we consider satisfy a minimization principle, as in [3]. Thus, quite generally, symmetry breaking occurs as a consequence of the (finite-dimensional) *normal form* arising in systems with certain symmetry properties. Although we can treat a large class of problems for which there is no minimization principle, our analysis, at present, is



(a)



(b)

FIG. 1. (a) *Bifurcation diagram for bound state solutions* $\psi(x,t) = e^{-i\Omega t}\Psi_\Omega(x)$ *of NLS-GP equation* (2.1), *with double-well potential* (6.1) *and cubic nonlinearity. Double-well parameters are* $s = 1, L = 6$. Ω *denotes the (nonlinear) frequency and* $\mathcal{N} = \mathcal{N}[\Psi_\Omega]$, *the squared* L^2 *norm of* Ψ_Ω *(optical power or particle number). The first bifurcation is from the zero state at the* ground state energy *of the double well. This state is an even function of* x *(symmetric). A secondary bifurcation, to an asymmetric state, at* $\mathcal{N} = \mathcal{N}_{cr}$, *is marked by a (red) circle. For* $\mathcal{N} < \mathcal{N}_{cr}$ *the symmetric state ((blue) solid line) is nonlinearly dynamically stable. For* $\mathcal{N} > \mathcal{N}_{cr}$ *the symmetric state is unstable ((blue) dashed line). The stable asymmetric state, appearing for* $\mathcal{N} > \mathcal{N}_{cr}$, *is marked by a (red) solid line. The (unstable) odd in* x *(antisymmetric) state is marked by a (green) dashed-dotted line.* (b) *Bound state solutions* $\Psi_\Omega(x)$ *plotted for a level set of* $\mathcal{N}[\Psi_\Omega] = 0.5$.

restricted to small norm. As we shall see, a bifurcation occurring at small norm can be ensured, for example, by taking the distance between wells in the double well to be sufficiently large.

In [14] the precise transition point to symmetry breaking, $\mathcal{N}_{cr}$, of the ground state and the transfer of its stability to an asymmetric ground state was considered (by geometric dynamical systems methods) in the exactly solvable NLS-GP, with a double-well potential consisting of two Dirac delta functions separated by a distance L. Additionally, the behavior of the function $\mathcal{N}_{cr}(L)$ was considered. Another solvable model was examined by numerical means in [23]. A study of dynamics for nonlinear double wells appeared in [27].

We study $\mathcal{N}_{cr}(L)$ in general. The value at which symmetry breaking occurs, $\mathcal{N}_{cr}(L)$, is closely related to the spectral properties of the linearization of NLS-GP

about the symmetric branch. Indeed, so long as the linearization of NLS-GP at the symmetric state has no nonsymmetric null space, the symmetric state is locally unique by the implicit function theorem [24]. The mechanism for symmetry breaking is the first appearance of an antisymmetric element in the null space of the linearization for some $\mathcal{N} = \mathcal{N}_{cr}$. This is demonstrated for a finite-dimensional Galerkin approximation of NLS-GP in [20, 17]. The present work extends and generalizes this analysis to the full infinite-dimensional problem using the Lyapunov–Schmidt method [24]. Control of the corrections to the finite-dimensional approximation requires small norm of the states considered. Since, as anticipated by the Galerkin approximation, $\mathcal{N}_{cr}$ is proportional to the distance between the lowest eigenvalues of the double well, which is exponentially small in L, our results apply to double wells with separation L, for L sufficiently large.

Our bifurcation results can be viewed as a very detailed study of a class of $O(2) \times \mathbb{Z}_2$ symmetric dynamical systems with four degrees of freedom in which there is a bifurcation of mixed states. We know of no classification of such systems. We show, under the nondegeneracy condition (4.3), that they are reducible to systems with two degrees of freedom and $\mathbb{Z}_2 \times \mathbb{Z}_2$ symmetry which have been classified in, for example, [8, 11]. In addition to being a self-contained treatment, our study of the nature of the solution set on the manifold of constant $\mathcal{N}$, as $\mathcal{N}$ is varied, shows the key role of $\mathcal{N}$, a physical NLS-GP time-invariant, in determining the bifurcation diagram, which encodes the dynamic stability properties.

The article is organized as follows. In section 2 we introduce the NLS-GP model and give a technical formulation of the bifurcation problem. In section 3 we study a finite-dimensional truncation of the bifurcation problem, identifying a relevant bifurcation point. In section 4, we prove the persistence of this symmetry-breaking bifurcation in the full NLS-GP problem for $\mathcal{N} \geq \mathcal{N}_{cr}$. Moreover, in section 5 we show that the lowest energy symmetric state becomes dynamically unstable at $\mathcal{N}_{cr}$ and the bifurcating asymmetric state is the dynamically stable ground state for $\mathcal{N} > \mathcal{N}_{cr}$.

The main results are stated in Theorem 4.1, Corollary 4.1, and Theorem 5.1. In particular, we obtain an asymptotic formula for the critical particle number (optical power) for symmetry breaking in NLS-GP,

$$\mathcal{N}_{cr} = \frac{\Omega_1 - \Omega_0}{\Xi[\psi_0, \psi_1]} \; + \; \mathcal{O}\left(\frac{(\Omega_1 - \Omega_0)^2}{\Xi[\psi_0, \psi_1]^3} \right). \tag{1.1}$$

Here, (Ω_0, ψ_0) and (Ω_1, ψ_1) are eigenvalue-eigenfunction pairs of the *linear* Schrödinger operator $H = -\Delta + V$, where Ω_0 and Ω_1 are separated from other spectrum and Ξ is a positive constant, given by (4.2), depending on ψ_0 and ψ_1. (Section 8 is an appendix where we discuss the basic spectral properties of $-\Delta + V$, where V is a double-well potential.) *The most important case is where* $\Omega_0 < \Omega_1$ *are the* lowest *two energies (linear ground and first excited states).* For double wells with separation L, we have $\mathcal{N}_{cr} = \mathcal{N}_{cr}(L)$, depending on the eigenvalue spacing $\Omega_0(L) - \Omega_1(L)$, which is exponentially small if L is large and Ξ is of order one. Thus, for large L, the bifurcation occurs at small L^2 norm. This is the weakly nonlinear regime in which the corrections to the finite-dimensional model can be controlled perturbatively. A local bifurcation diagram of this type will occur for *any* simple even-odd symmetric pair of simple eigenvalues of H in the weakly nonlinear regime, so long as the eigenfrequencies are separated from the rest of the spectrum of H; see Proposition 4.1 and the gap condition (4.13). Therefore, a similar phenomenon occurs for higher order, nearly degenerate pairs of eigenstates of the double wells, arising from isolated single

wells with multiple eigenstates. Section 6 contains numerical results validating our theoretical analysis.

2. Technical formulation. Consider the initial value problem for the time-dependent NLS-GP equation

$$i\partial_t\psi = H\psi \ + \ g(x)K[\psi\bar{\psi}]\ \psi, \quad \psi(x,0) \text{ given}, \tag{2.1}$$

$$H = -\Delta + V(x). \tag{2.2}$$

We assume the following.

(H1) The initial value problem for NLS-GP is well-posed in the space $C^0([0,\infty); H^1(\mathbb{R}^n))$.

(H2) The potential $V(x)$ is assumed to be real valued, smooth, and rapidly decaying as $|x| \to \infty$. We also assume symmetry with respect to a hyperplane which, without loss of generality, can be taken to be $\{x_1 = 0\}$:

$$V(x_1, x_2, \ldots, x_n) = V(-x_1, x_2, \ldots, x_n). \tag{2.3}$$

We assume the nonlinear term, $K[\psi\overline{\psi}]$, to be attractive, cubic (local or nonlocal), and symmetric with respect to the same hyperplane. Specifically, we assume the following hypotheses on the nonlinear term:

(H3) (a) $g(x_1, x_2, \ldots, x_n) = g(-x_1, x_2, \ldots, x_n)$ (symmetry).
(b) $g(x) < 0$ (attractive/focusing).
(c) $K[h] = \int K(x-y)h(y)dy$,
$K(x_1, x_2, \ldots, x_n) = K(-x_1, x_2, \ldots, x_n)$, $K > 0$.
(d) Consider the map $N : H^2 \times H^2 \times H^2 \mapsto L^2$ defined by

$$N(\phi_1, \phi_2, \phi_3) = gK[\phi_1\bar{\phi_2}]\phi_3. \tag{2.4}$$

We also write $N(u) = N(u,u,u)$ and note that $\partial_u N(u) = N(\cdot, u, u) + N(u, \cdot, u) + N(u, u, \cdot)$. We assume that there exists a constant $k > 0$ such that

$$\|N(\phi_0, \phi_1, \phi_2)\|_{L^2} \ \leq \ k\ \|\phi_1\|_{H^2}\|\phi_2\|_{H^2}\|\phi_3\|_{H^2}. \tag{2.5}$$

Remark 2.1. The symmetry restrictions in (H3) insure that the nonlinear term is equivariant with respect to the action of $O(2) \times \mathbb{Z}_2$, namely,

$$\begin{aligned} N(e^{i\theta}u) &= e^{i\theta}N(u), \qquad 0 \leq \theta < 2\pi, \\ N(\bar{u}) &= \overline{N(u)}, \\ N(Ru) &= RN(u), \end{aligned}$$

where R is the reflection with respect to the $x_1 = 0$ hyperplane. Although, the nonlinearity taken in (H3) is cubic, our analysis can be easily extended to more general cases satisfying the $O(2) \times \mathbb{Z}_2$ equivariance.

We now give several illustrative and important examples of NLS-GP.

Example 1. *GP equation for BECs with negative scattering* length $g(x) \equiv -1$, $K(x) = \delta(x)$.

Example 2. *NLS equation for optical media with a nonlocal kernel* $g(x) \equiv \pm 1$, $K(x) = A\exp(-x^2/\sigma^2)$ [22]. See also [6] for similar considerations in BECs.

Example 3. *Photorefractive (saturable) nonlinearities and optically induced potentials* [7]. The relevant symmetry breaking phenomenology is experimentally observable, as shown in [20]. Also contained therein is the finite-dimensional Galerkin approach of section 3.

Nonlinear bound states. Nonlinear bound states are solutions of NLS-GP of the form

$$\psi(x,t) = e^{-i\Omega t}\Psi_\Omega(x), \tag{2.6}$$

where $\Psi_\Omega \in H^1(\mathbb{R}^n)$ solves

$$H\ \Psi_\Omega\ +\ g(x)\ K[|\Psi_\Omega|^2]\ \Psi_\Omega\ -\ \Omega\Psi_\Omega = 0,\quad \Psi_\Omega \in H^1. \tag{2.7}$$

If the potential $V(x)$ is such that the operator $H = -\Delta + V(x)$ has a discrete eigenvalue, Ω_*, and a correspsonding eigenstate ψ_*, then for energies Ω near Ω_*, one expects small amplitude *nonlinear* bound states, which are to leading order small multiples of ψ_*. This is the standard setting of bifurcation from a simple eigenvalue [24], which follows from the implicit function theorem.

THEOREM 2.1 (see [21, 25, 26]). *Let $(\Psi,\Omega) = (\psi_*,\Omega_*)$ be a simple eigenpair of the eigenvalue problem $H\Psi = \Omega\Psi$, i.e., $dim\{\rho : (H-\Omega_*)\rho = 0\} = 1$. Then there exists a unique smooth curve of nontrivial solutions $\alpha \mapsto (\ \Psi(\cdot;\alpha),\Omega(\alpha)\ )$, defined in a neighborhood of $\alpha = 0$, such that*

$$\Psi_\Omega = \alpha\left(\ \psi_*\ +\ \mathcal{O}(|\alpha|^2)\ \right),\quad \Omega = \Omega_*\ +\ \mathcal{O}(|\alpha|^2),\quad \alpha \to 0. \tag{2.8}$$

Remark 2.2. For a large class of problems, a nonlinear ground state can be characterized variationally as a constrained minimum of the NLS-GP energy subject to fixed squared L^2 norm. Define the NLS-GP Hamiltonian energy functional

$$\mathcal{H}_{NLS-GP}[\Phi]\ \equiv\ \int |\nabla\Phi|^2\ +\ V|\Phi|^2\ dy\ +\ \frac{1}{2}\int\ g(y)|\Phi(y)|^2K[|\Phi|^2]\ dy \tag{2.9}$$

and the particle number (optical power)

$$\mathcal{N}[\Phi] = \int\ |\Phi|^2\ dy. \tag{2.10}$$

In particular, the following can be proved.

THEOREM 2.2. *Let $I_\lambda = \inf_{\mathcal{N}[f]=\lambda}\ \mathcal{H}[f]$. If $-\infty < I_\lambda < 0$, then the minimum is attained at a positive solution of* (2.7). *Here, $\Omega = \Omega(\lambda)$ is a Lagrange multiplier for the constrained variational problem.*

In [3] the nonlinear Hartree equation is studied; $K[h] = |y|^{-1} \star h$, $g \equiv -1$. It is proved that if $V(x)$ is a double-well potential, then for λ sufficiently large, the minimizer does not have the same symmetry as the linear ground state. By uniqueness, ensured by the implicit function theorem, for small $\mathcal{N}$, the minimizer has the same symmetry as that as the linear ground state and has the expansion (2.8); see [3] and section 4.

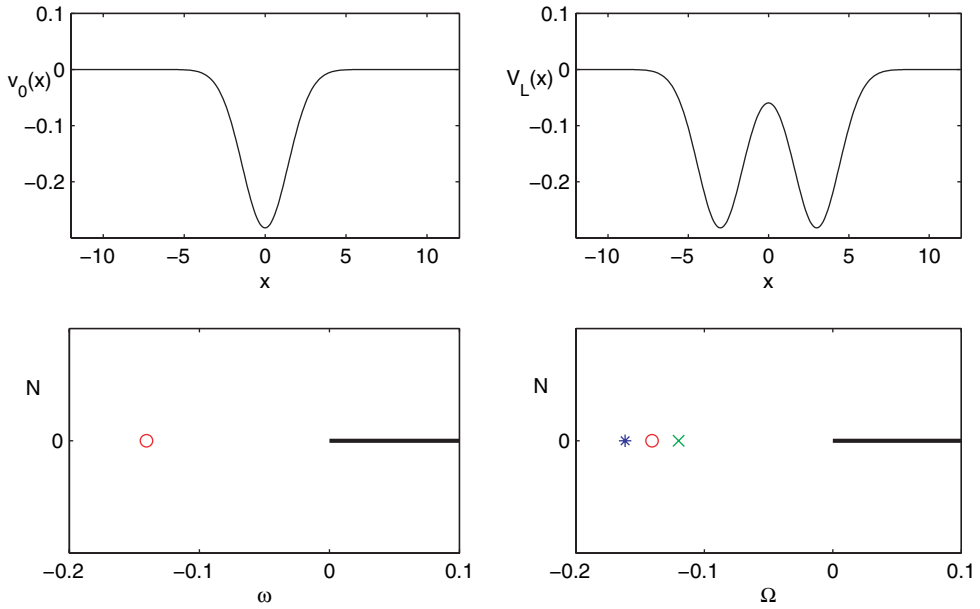
We make the following assumptions.

Spectral assumptions on H.

(H4) H has a pair of simple eigenvalues Ω_0 and Ω_1. ψ_0 and ψ_1, the corresponding (real-valued) eigenfunctions are, respectively, even (symmetric) and odd (antisymmetric) in x_1.

Example 2.1 (the basic example: *double-well potentials*). Our basic example of $V(x)$ is a double-well potential consisting of two identical potential wells and separated by a distance L. Thus, assume symmetry with respect to the hyperplane which, without loss of generality, can be taken to be $\{x_1 = 0\}$:

$$V(x_1,x_2,\ldots,x_n) = V(-x_1,x_2,\ldots,x_n). \tag{2.11}$$



(a) (b)

FIG. 2. *This figure demonstrates a single- and a double-well potential and the spectrum of H and H_L, respectively.* (a) *shows a single-well potential and under it the spectrum of H, with an eigenvalue marked by a (red) mark "o" at ω and continuous spectrum marked by a (black) line for energies $\omega \geq 0$.* (b) *shows the double well centered at $\pm L$ and the spectrum of H_L underneath. The eigenvalues Ω_0 and Ω_1 are each marked by a (blue) mark "*" and a (green) mark "x," respectively, on either side of the location ω—(red) mark "o." The continuous spectrum is marked by a (black) line for energies $\Omega \geq 0$.*

In the appendix, we discuss how spectral hypothesis (H4) is shown for a double wells with well-separation parameter L. The simplest example, in one space dimension, is obtained as follows. Start with a single potential well (rapidly decaying as $|x| \to \infty$), $v_0(x)$, having exactly one eigenvalue, ω, $H_0\psi_\omega = (-\Delta + v_0(x))\psi_\omega = \omega\psi_\omega$; see Figure 2(a). Center this well at $x = -L$ and place an identical well centered at $x = L$. Denote by $V_L(x)$ the resulting *double-well potential* and by H_L the Schrödinger operator:

$$H_L = -\Delta + V_L(x). \tag{2.12}$$

There exists $L > L_0$ such that for $L > L_0$, H_L has a pair of eigenvalues $\Omega_0 = \Omega_0(L)$ and $\Omega_1 = \Omega_1(L)$, $\Omega_0 < \Omega_1$, and corresponding eigenfunctions ψ_0 and ψ_1; see Figure 2(b). ψ_0 is symmetric with respect to $x = 0$ and ψ_1 is antisymmetric with respect to $x = 0$. Moreover, for L sufficiently large, $|\Omega_0 - \Omega_1| = \mathcal{O}(e^{-\kappa L})$, $\kappa > 0$; see [12] and see also the appendix.

The construction can be generalized. If $-\Delta + v_0(x)$ has m bound states, then forming a double well V_L with L sufficiently large, $H_L = -\Delta + V_L$ will have m pairs of eigenvalues: $(\Omega_{2j}, \Omega_{2j+1})$, $j = 0, \ldots, m-1$, eigenfunctions ψ_{2j} (symmetric), and ψ_{2j+1} antisymmetric.

By Theorem 2.1, for small $\mathcal{N}$, there exists a unique nontrivial nonlinear bound state bifurcating from the zero solution at the ground state energy, Ω_0, of H. By uniqueness, ensured by the implicit function theorem, these small amplitude nonlinear bound states have the same symmetries as the double well; they are bimodal. We also

know from [3] that for sufficiently large $\mathcal{N}$ the ground state has broken symmetry. We now seek to elucidate the transition from the regime of $\mathcal{N}$ small to $\mathcal{N}$ large.

We work in the general setting of hypotheses (H1)–(H4). Define spectral projections onto the bound and continuous spectral parts of H:

$$P_0 = (\psi_0, \cdot)\,\psi_0,\ P_1 = (\psi_1, \cdot)\,\psi_1,\quad \tilde{P} = I\ -\ P_0\ -\ P_1. \tag{2.13}$$

Here,

$$(f, g) = \int\ \bar{f} g\ dx. \tag{2.14}$$

We decompose the solutions of (2.7) according to

$$\Psi_\Omega = c_0\psi_0\ +\ c_1\psi_1\ +\ \eta,\quad \eta = \tilde{P}\eta. \tag{2.15}$$

We next substitute the expression (2.15) into (2.7) and then act with projections P_0, P_1, and $\tilde{P}$ on the resulting equation. Using the symmetry and antisymmetry properties of the eigenstates, we obtain three equations which are equivalent to the PDE (2.7):

$$\begin{aligned}(\Omega_0 - \Omega)\, c_0 + a_{0000}|c_0|^2 c_0 + (a_{0110} + a_{0011})\,|c_1|^2 c_0,\\ + a_{0011} c_1^2 \bar{c}_0 + (\psi_0 g, \mathcal{R}(c_0, c_1, \eta)) = 0,\end{aligned} \tag{2.16}$$

$$\begin{aligned}(\Omega_1 - \Omega)\, c_1 + a_{1111}|c_1|^2 c_1 + (a_{1010} + a_{1001})\,|c_0|^2 c_1\\ + a_{1010} c_0^2 \bar{c}_1 + (\psi_1 g, \mathcal{R}(c_0, c_1, \eta)) = 0,\end{aligned} \tag{2.17}$$

$$(H - \Omega)\ \eta = -\tilde{P}\ g\left[F(\cdot; c_0, c_1)\ +\ \mathcal{R}(c_0, c_1, \eta)\right], \tag{2.18}$$

where $F(\cdot, c_0, c_1)$ is independent of η and $\mathcal{R}(c_0, c_1, \eta)$ contains linear, quadratic, and cubic terms in η. The coefficients a_{klmn} are defined by

$$a_{klmn} = \left(\psi_k, gK[\psi_l \bar{\psi_m}]\psi_n\right). \tag{2.19}$$

We shall study the character of the set of solutions of the system (2.16)–(2.18) restricted to the level set

$$\int |\Psi_\Omega|^2 dx = \mathcal{N} \quad\Longleftrightarrow\quad |c_0|^2\ +\ |c_1|^2\ +\ \int |\eta|^2 dx = \mathcal{N} \tag{2.20}$$

as $\mathcal{N}$ varies.

Let Ω_0 and Ω_1 denote the two lowest eigenvalues of H_L. We prove (Theorem 4.1, Corollary 4.1, Theorem 5.1) the following.

- There exist two solution branches, parametrized by $\mathcal{N}$, which bifurcate from the zero solution at the eigenvalues Ω_0 and Ω_1.
- Along the branch (Ω, Ψ_Ω), emanating from the solution $(\Omega = \Omega_0, \Psi = 0)$, there is a symmetry-breaking bifurcation at $\mathcal{N} = \mathcal{N}_{crit} > 0$. In particular, let u_{crit} denote the solution of (2.7) corresponding to the value $\mathcal{N} = \mathcal{N}_{crit}$. Then, in a neighborhood u_{crit}, for $\mathcal{N} < \mathcal{N}_{crit}$ there is only one solution of (2.7), the symmetric ground state, while for $\mathcal{N} > \mathcal{N}_{crit}$ there are two solutions one symmetric and a second asymmetric.
- *Exchange of stability* at the bifurcation point: For $\mathcal{N} < \mathcal{N}_{crit}$ the symmetric state is dynamically stable, while for $\mathcal{N} > \mathcal{N}_{crit}$ the asymmetric state is stable and the symmetric state is exponentially unstable.

3. Bifurcations in a finite-dimensional approximation. It is illustrative to consider the finite-dimensional approximation to the system (2.16)–(2.18), obtained by neglecting the continuous spectral part, η. Let us first set $\eta = 0$, and therefore $\mathcal{R}(c_0, c_1, 0) = 0$. Under this assumption of no coupling to the continuous spectral part of H, we obtain the finite-dimensional system

$$(\Omega_0 - \Omega)\, c_0 + a_{0000}|c_0|^2 c_0 + (a_{0110} + a_{0011})\, |c_1|^2 c_0 + a_{0011} c_1^2 \bar{c}_0 = 0, \tag{3.1}$$

$$(\Omega_1 - \Omega)\, c_1 + a_{1111}|c_1|^2 c_1 + (a_{1010} + a_{1001})\, |c_0|^2 c_1 + a_{1010} c_0^2 \bar{c}_1 = 0. \tag{3.2}$$

Note that in this approximation the physically meaningful parameter $\mathcal{N}$ is given by

$$|c_0|^2 \ + \ |c_1|^2 = \mathcal{N}. \tag{3.3}$$

Our strategy is to first analyze the bifurcation problem for this approximate finite-dimensional system of *algebraic equations*. We then treat the corrections, coming from coupling to the continuous spectral part of H, η, perturbatively. In fact, the analysis in the next section indirectly shows that (3.1)-(3.2) is the universal unfolding of bifurcation problems with four degrees of freedom and $O(2) \times \mathbb{Z}_2$ symmetry; see [8] for related notions.[1]

Our point of view, however, is not the *general* theory of (3.1)-(3.2), its bifurcations and its structural stability, but rather what it implies for solutions of the NLS-GP system for given "data", e.g., spectral properties of the potential $V(x)$, the nonlinearity term defined in terms of $K[\cdot]$ and g. This data, together with structural properties of $NLS\text{-}GP$, such as its Hamiltonian structure and the second-order elliptic character of the nonlinear bound state equations, imply natural assumptions as well as constraints on the coefficients a_{jklm}. In this section, where we study the algebraic problem (3.1)-(3.2), we posit and comment briefly on these properties. The first are

$$a_{0000} - a_{1001} - 2a_{1010} > 0, \tag{3.4}$$

$$a_{1010} \neq 0. \tag{3.5}$$

Condition (3.4) will guarantee a symmetry-breaking bifurcation at small amplitude, while (3.5) is a nondegeneracy condition, easily seen to hold for generic assumptions on $V, K[\cdot]$ and g in NLS-GP.

Additional conditions are

$$a_{1001} - a_{0000} > 0 \tag{3.6}$$

and

$$a_{1111} - a_{0110} - 2a_{0011} > 0, \tag{3.7}$$

$$a_{1111} - a_{0110} > 0, \tag{3.8}$$

which we shall see are related to secondary bifurcations and the exchange of stability at the bifurcation point.

Note also that

$$a_{0000} < 0, \quad a_{1111} < 0 \tag{3.9}$$

since $K[\cdot]$ preserves positivity and $g < 0$; see (2.19) and hypothesis (H3).

[1] Since the full problem, with $\eta \neq 0$, adds higher order terms in (3.1)-(3.2), from singularity theory we expect that they will not qualitatively change the bifurcation diagram. We do not make use of results in singularity theory instead we prefer a self-contained treatment yielding a detailed understanding of the effect of the physically important unfolding parameter $\Omega_1 - \Omega_0$.

One can easily construct, using (3.9), the following two *pure mode solutions* of (3.1)-(3.2):

(1) *Approximate nonlinear symmetric branch:*
$c_1 = 0,\ c_0 = \sqrt{\frac{\Omega-\Omega_0}{a_{0000}}}e^{i\theta},\ 0 \le \theta < 2\pi,\ \Omega \le \Omega_0.$
In terms of the physical parameter $\mathcal{N}$: $c_1 = 0,\ c_0 = \sqrt{\mathcal{N}}e^{i\theta},\ \Omega = \Omega_0 + a_{0000}\mathcal{N}$, where $\mathcal{N} \ge 0,\ 0 \le \theta < 2\pi$.

(2) *Approximate nonlinear antisymmetric branch:*
$c_0 = 0,\ c_1 = \sqrt{\frac{\Omega-\Omega_1}{a_{1111}}}e^{i\theta},\ 0 \le \theta < 2\pi,\ \Omega \le \Omega_1.$
In terms of the physical parameter $\mathcal{N}$: $c_0 = 0,\ c_1 = \sqrt{\mathcal{N}}e^{i\theta},\ \Omega = \Omega_1 + a_{1111}\mathcal{N}$, where $\mathcal{N} \ge 0,\ 0 \le \theta < 2\pi$.

The above two solutions correspond to those of NLS-GP given by Theorem 2.1.

Under assumptions (3.4)–(3.8) the system (3.1)-(3.2) has exactly one more (secondary) branch of solutions which bifurcates from the symmetric branch at

$$\Omega^* = \Omega_0 + \frac{a_{0000}(\Omega_1 - \Omega_0)}{a_{0000} - a_{1001} - 2a_{1010}} < \Omega_0$$

or, equivalently, at

$$\mathcal{N}_{cr} = \frac{\Omega_1 - \Omega_0}{a_{0000} - a_{1001} - 2a_{1010}} > 0. \tag{3.10}$$

In terms of the physical parameter $\mathcal{N}$, these *mixed mode solutions of the approximate system* are given by

(3) *Approximate nonlinear asymmetric branch:* For $\mathcal{N} \ge \mathcal{N}_{cr},\ 0 \le \theta_0,\ \theta_1 < 2\pi,\ \theta_1 - \theta_0 \in \{0, \pi\}$,

$$\begin{aligned}
c_0 &= \sqrt{\frac{(a_{1111} - a_{0110} - 2a_{0011})\mathcal{N} + (\Omega_1 - \Omega_0)}{a_{1111} - a_{0110} - 2a_{0011} + a_{0000} - a_{1001} - 2a_{1010}}}e^{i\theta_0},\\
c_1 &= \sqrt{\frac{(a_{0000} - a_{1001} - 2a_{1010})\mathcal{N} - (\Omega_1 - \Omega_0)}{a_{1111} - a_{0110} - 2a_{0011} + a_{0000} - a_{1001} - 2a_{1010}}}e^{i\theta_1},\\
\Omega &= \Omega_0 + \\
&\frac{(a_{0000}a_{1111} - (a_{1001} + 2a_{1010})(a_{0110} + 2a_{0011}))\mathcal{N}}{a_{1111} - a_{0110} - 2a_{0011} + a_{0000} - a_{1001} - 2a_{1010}}\\
&\frac{- (a_{0000} - a_{1001} - 2a_{1010})(\Omega_1 - \Omega_0)}{a_{1111} - a_{0110} - 2a_{0011} + a_{0000} - a_{1001} - 2a_{1010}}.
\end{aligned} \tag{3.11}$$

To derive the above solutions, first introduce polar coordinates: $c_j = \rho_j e^{i\theta_j},\ j = 1, 2,\ \Delta\theta = \theta_1 - \theta_0$. Then, (3.1)-(3.2) becomes

$$\rho_0\ \left[\Omega_0 - \Omega\ +\ a_{0000}\rho_0^2\ +\ \left(a_{0110} + a_{0011} + a_{0011}e^{i2\Delta\theta}\right)\rho_1^2\right] = 0, \tag{3.12}$$

$$\rho_1\ \left[\Omega_1 - \Omega\ +\ a_{1111}\rho_1^2\ +\ \left(a_{1001} + a_{1010} + a_{1010}e^{-i2\Delta\theta}\right)\rho_0^2\right] = 0. \tag{3.13}$$

Taking the imaginary parts of both equations and using the fact that all coefficients are real, see (2.19), we get $a_{0011}\sin(2\Delta\theta) = 0 = a_{1010}\sin(2\Delta\theta)$. Because of (3.5), the latter are equivalent with $\Delta\theta \in \{0, \pi/2, \pi, 3\pi/2\}$ modulo 2π, note that $a_{0011} = a_{1010}$; see (2.19). The real parts of (3.12)-(3.13) become

$$\rho_0\ \left[\Omega_0 - \Omega\ +\ a_{0000}\rho_0^2\ +\ (a_{0110} + 2a_{0011})\,\rho_1^2\right] = 0, \tag{3.14}$$

$$\rho_1\ \left[\Omega_1 - \Omega\ +\ a_{1111}\rho_1^2\ +\ (a_{1001} + 2a_{1010})\,\rho_0^2\right] = 0 \tag{3.15}$$

for $\Delta\theta \in \{0, \pi\}$, while for $\Delta\theta \in \{\pi/2, 3\pi/2\}$,

$$\rho_0 \left[\Omega_0 - \Omega + a_{0000}\rho_0^2 + a_{0110}\rho_1^2\right] = 0, \tag{3.16}$$

$$\rho_1 \left[\Omega_1 - \Omega + a_{1111}\rho_1^2 + a_{1001}\rho_0^2\right] = 0. \tag{3.17}$$

Both systems (3.14)-(3.15) and (3.16)-(3.17), respectively, can be easily analyzed and even reduced to a linear system by the change of variables: $\mathcal{P}_0 = \rho_0^2$, $\mathcal{P}_1 = \rho_1^2$. The first one, under assumption (3.4), gives the mixed mode solution (3) in addition to the pure mode solutions (1) and (2), and no other solutions provided (3.7) holds. The second system gives no other solutions provided (3.6), (3.8) hold. Thus we have that the finite-dimensional Galerkin approximation yields the qualitative structure of the NLS-GP bifurcation diagram of Figure 3. Dynamic stability can also be addressed in the context of the Galerkin-truncated time-dependent Hamiltonian dynamics. We defer our study of stability to our considerations of the full PDE problem in the coming sections.

The study of the systems (3.14)-(3.15) and (3.16)-(3.17) can be put in the context of singularity theory and bifurcations in systems with symmetry [8]. These systems exhibit a $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ symmetry. Actually the argument in the next section shows that under the nondegeneracy condition (3.5), bifurcation problems with four degrees of freedom and $O(2) \times \mathbb{Z}_2$ symmetry are reducible to two distinct systems each with two degrees of freedom and $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ symmetry. The latter have been classified, for example, in [8, Chapter X]. Under assumptions (3.4), (3.7), our first system (3.14)-(3.15) falls in case A subcase (1) [8, Chapter X] while under (3.6), (3.8) our second system falls in their case A subcases (2) or (3). Note that the above-cited book uses $\lambda = \Omega_0 - \Omega$ as bifurcation parameter and $\sigma = \Omega_0 - \Omega_1$ as the unfolding parameter which explains the differences between our bifurcation diagram in Figure 1 and their Figure 4.3 in Chapter X.

Finally, the posited properties of a_{klmn} can be directly verified for general double-well problems with L sufficiently large; see the proof of Corollary 4.1.

4. Bifurcation/symmetry breaking analysis of the PDE. In this section we prove the following theorem.

THEOREM 4.1 (symmetry breaking for NLS-GP). *Consider NLS-GP with hypotheses* (H2)–(H4). *Let* a_{klmn} *be given by* (2.19) *and let*

$$\Xi[\psi_0, \psi_1, g] \equiv a_{0000} - a_{1001} - 2a_{1010} \tag{4.1}$$

$$= \left(\psi_0^2, gK[\psi_0^2]\right) - \left(\psi_1^2, gK[\psi_0^2]\right) - 2\left(\psi_0\psi_1, gK[\psi_0\psi_1]\right) > 0, \tag{4.2}$$

$$a_{1010} = \left(\psi_0\psi_1, gK[\psi_0\psi_1]\right) \neq 0. \tag{4.3}$$

Assume

$$\frac{\Omega_1 - \Omega_0}{\Xi[\psi_0, \psi_1, g]^2} \quad \textit{is sufficiently small.} \tag{4.4}$$

Then, there exists $\mathcal{N}_{cr} > 0$ *such that*

(i) *for any* $\mathcal{N} \leq \mathcal{N}_{cr}$, *there is (up to the symmetry* $u \mapsto u\, e^{i\gamma}$*) a locally unique symmetric state,* $u_{\mathcal{N}}^{sym}$.

(ii) $\mathcal{N} = \mathcal{N}_{cr}$, $u_{\mathcal{N}_{cr}}^{sym}$ *is a bifurcation point. For* $\mathcal{N} > \mathcal{N}_{cr}$, *there are, in a neighborhood of* $\mathcal{N} = \mathcal{N}_{cr}$, $u_{\mathcal{N}_{cr}}^{sym}$, *two branches of solutions:* (a) *a continuation of the symmetric branch and* (b) *a new asymmetric branch.*

(iii) *The critical* $\mathcal{N}$ *value for bifurcation is given approximately by*

$$\mathcal{N}_{cr} = \frac{\Omega_1 - \Omega_0}{\Xi[\psi_0, \psi_1, g]} \left[1 + \mathcal{O}\left(\frac{\Omega_1 - \Omega_0}{\Xi[\psi_0, \psi_1, g]^2} \right)\right].$$

(iv) *The bifurcation from the zero state at* $\Omega = \Omega_0$ *and the bifurcation described in* (i)–(iii) *are the only ones along this branch for* $\mathcal{N}$ *small.*

COROLLARY 4.1 (application to double wells, V_L, with separation L). *Fix a pair of eigenvalues,* (Ω_{2j}, ψ_{2j}), $(\Omega_{2j+1}, \psi_{2j+1})$, *of the linear double-well potential,* $V_L(x)$; *see Example* 2.1. *Assume that* $K(x) \to 0$ *as* $|x| \to 0$ *and* $\lim_{L\to\pm\infty} g(x_1 + L, x_2, \dots) = g_\pm(x_2, \dots) < 0$. *Then, for the NLS-GP with double-well potential of well-separation* L, *there exists* $\tilde{L} > 0$ *such that for all* $L \geq \tilde{L}$, *there is a symmetry-breaking bifurcation, as described in Theorem* 4.1, *with* $\mathcal{N}_{cr} = \mathcal{N}_{cr}(L; j)$.

Proof of Corollary 4.1. The crux is to verify the spectral properties (H4), the nondegeneracy condition (4.3), and the smallness condition (4.4). The spectral properties of $H_L = -\Delta + V_L$ are handled in Proposition 8.1 of the appendix. In what follows we show that

$$\lim_{L\to\infty} a_{0000} - a_{1001} = 0, \tag{4.5}$$

$$\lim_{L\to\infty} a_{1010} < 0. \tag{4.6}$$

These imply both (4.3) and (4.4) since $\Omega_0 - \Omega_1 \to 0$ as $L \to \infty$; see Proposition 8.1.

For (4.5), (4.6) we use the large L asymptotic behavior of the eigenfunctions of H_L, given in Proposition 8.1, together with the continuity of the nonlinearity (2.5), to rewrite them as

$$\lim_{L\to\infty} a_{0000} - a_{1001} = \lim_{L\to\infty} \frac{1}{2} \left(T_L\psi_\omega, gK[(T_L\psi_\omega + RT_L\psi_\omega)^2]RT_L\psi_\omega\right), \tag{4.7}$$

$$\lim_{L\to\infty} a_{1010} = \lim_{L\to\infty} \frac{1}{4} \left(T_L\psi_\omega^2 - RT_L\psi_\omega^2, gK[T_L\psi_\omega^2 - RT_L\psi_\omega^2]\right). \tag{4.8}$$

To obtain (4.5) approximate $\psi_\omega \in H^2$ by smooth functions with compact support ψ_ω^0. For sufficiently large L, $T_L\psi_\omega^0$ and $RT_L\psi_\omega^0$ have no common support and the scalar product in (4.7) is zero. The continuity of the nonlinearity (2.5) implies (4.5).

Now, the right-hand side in (4.8) has two types of similar terms:

$$\lim_{L\to\infty} \left(T_L\psi_\omega^2, gK[RT_L\psi_\omega^2]\right) = \lim_{L\to\infty} \left(\psi_\omega^2, (T_{-L}g)T_{-2L}K[R\psi_\omega^2]\right) = 0,$$

$$\lim_{L\to\infty} \left(T_L\psi_\omega^2, gK[T_L\psi_\omega^2]\right) = \lim_{L\to\infty} \left(\psi_\omega^2, (T_{-L}g)K[\psi_\omega^2]\right) = \left(\psi_\omega^2, g_-K[\psi_\omega^2]\right) < 0,$$

where we used the symmetries of g and K and their behavior as $x_1 \to \pm\infty$. This completes the proof of the corollary. □

To prove Theorem 4.1 we will establish that, under hypotheses (4.2)–(4.4), the character of the solution set (symmetry-breaking bifurcation) of the finite-dimensional

approximation (3.1)-(3.2) persists for the full (infinite-dimensional) problem:

$$(\Omega_0 - \Omega)\, c_0 + a_{0000}|c_0|^2 c_0 + (a_{0110} + a_{0011})\,|c_1|^2 c_0 + a_{0011} c_1^2 \bar{c}_0 + (\psi_0 g, \mathcal{R}(c_0, c_1, \eta)) = 0, \tag{4.9}$$

$$(\Omega_1 - \Omega)\, c_1 + a_{1111}|c_1|^2 c_1 + (a_{1010} + a_{1001})\,|c_0|^2 c_1 + a_{1010} c_0^2 \bar{c}_1 + (\psi_1 g, \mathcal{R}(c_0, c_1, \eta)) = 0, \tag{4.10}$$

$$(H - \Omega)\ \eta = -\tilde{P}\ g\,[F(\cdot; c_0, c_1)\ +\ \mathcal{R}(c_0, c_1, \eta)\ ]\,, \qquad \eta = \tilde{P}\eta, \tag{4.11}$$

$$|c_0|^2\ +\ |c_1|^2\ +\ \int |\eta|^2 = \mathcal{N}. \tag{4.12}$$

We analyze this system using the Lyapunov–Schmidt-type method. The strategy is to solve (4.11) for η as a functional of c_0, c_1, and Ω. Then, substituting $\eta = \eta[c_0, c_1, \Omega]$ into (4.9), (4.10), and (4.12), we obtain three closed equations, depending on a parameter $\mathcal{N}$, for c_0, c_1, and Ω. This system is a perturbation of the finite-dimensional (truncated) system: (3.1)–(3.3). We then show that under hypotheses (4.2)–(4.4) there is a symmetry-breaking bifurcation. Finally, we show that the terms perturbing the finite-dimensional model have a small and controllable effect on the character of the solution set for a range of $\mathcal{N}$, which includes the bifurcation point. As seen in the proof, the form of the nonlinearity implies analyticity. However, analyticity is not essential for the arguments, at the root of which is the implicit function theorem, and the methods can be adapted to situations where one has a finite degree of smoothness.

We begin with the following proposition, which characterizes $\eta = \eta[c_0, c_1, \Omega]$.

PROPOSITION 4.1. *Consider* (4.11) *for* η. *By* (H4) *we have the following:*

(4.13) Gap condition : $|\Omega_j - \tau|\ \geq\ 2d_*$ for $j = 0, 1$ and for all $\tau \in \sigma(H) \setminus \{\Omega_0, \Omega_1\}$.

Then there exist $n_*,\ r_* > 0$, *depending on* d_*, *such that in the open set*

$$|c_0| + |c_1| < r_*, \qquad \|c_0\psi_0 + c_1\psi_1 + \eta\|_{H^2} < n_*(d_*), \tag{4.14}$$

$$dist(\Omega, \sigma(H) \setminus \{\Omega_0, \Omega_1\}) > d_*, \tag{4.15}$$

the unique solution of (2.18) *is given by the real-analytic mapping*

$$(c_0, c_1, \Omega)\ \mapsto\ \ \eta[c_0, c_1, \Omega], \tag{4.16}$$

defined on the domain given by (4.14)-(4.15). *Moreover, there exists* $C_* > 0$ *such that*

$$\|\eta[c_0, c_1, \Omega]\|_{H^2} \leq C_*(|c_0| + |c_1|)^3. \tag{4.17}$$

Proof. Consider the map

$$N : H^2 \times H^2 \times H^2 \mapsto L^2,$$
$$N(\phi_0, \phi_1, \phi_2) = gK[\phi_1 \bar{\phi}_2]\phi_3.$$

By assumptions on the nonlinearity (see section 2), there exists a constant $k > 0$ such that

$$\|N(\phi_0, \phi_1, \phi_2)\|_{L^2} \leq k\|\phi_1\|_{H^2}\|\phi_2\|_{H^2}\|\phi_3\|_{H^2}. \tag{4.18}$$

Moreover, since the map is real linear in each component, it is real analytic.[2]

[2]The trilinearity follows from the implicit bilinearity of K in formulas (2.16)–(2.18).

Let c_0, c_1, and Ω be restricted according to (4.14)-(4.15). Equation (2.18) can be rewritten in the form

$$\eta + (H-\Omega)^{-1}\tilde{P}N[c_0\psi_0 + c_1\psi_1 + \eta] = 0. \tag{4.19}$$

Since the spectrum of $H\tilde{P}$ is bounded away from Ω by d_*, the resolvent

$$(H-\Omega)^{-1}\tilde{P} : L^2 \mapsto H^2$$

is a (complex) analytic map and bounded uniformly,

$$\|(H-\Omega)^{-1}\tilde{P}\|_{L^2\mapsto H^2} \le p(d_*^{-1}), \tag{4.20}$$

where $p(s) \to \infty$ as $s \to \infty$. Consequently the map $F : \mathbb{C}^2 \times \{\Omega \in \mathbb{C} \ : \ dist(\Omega, \sigma(H) \setminus \{\Omega_0, \Omega_1\}\ \} \ge d_*\} \times H^2 \mapsto H^2$ given by

$$F(c_0, c_1, \Omega, \eta) = \eta + (H-\Omega)^{-1}\tilde{P}N[c_0\psi_0 + c_1\psi_1 + \eta] \tag{4.21}$$

is real analytic. Moreover,

$$F(0,0,\Omega,0) = 0, \quad D_\eta F(0,0,\Omega,0) = I.$$

Applying the implicit function theorem to (4.19), we have that there exists $n_*(\Omega), r_*(\Omega)$ such that whenever $|c_0| + |c_1| < r_*$ and $\|c_0\psi_0 + c_1\psi_1 + \eta\|_{H^2} < n_*$, (4.19) has a unique solution:

$$\eta = \eta(c_0, c_1, \Omega) \in H^2$$

which depends analytically on the parameters $c_0,\ c_1,\ \Omega$. By applying the projection operator $\tilde{P}$ to (4.19), which commutes with $(H-\Omega)^{-1}$, we immediately obtain $\tilde{P}\eta = \eta$, i.e., $\eta \in \tilde{P}L^2$.

We now show that $n_*,\ r_*$ can be chosen independently of Ω, satisfying (4.15). The implicit function theorem can be applied in an open set for which

$$D_\eta F(c_0, c_1, \Omega, \eta) = I + (H-\Omega)^{-1}\tilde{P}D_\eta N[c_0\psi_0 + c_1\psi_1 + \eta]$$

is invertible. For this it suffices to have

$$\|(H-\Omega)^{-1}\tilde{P}D_\eta N[c_0\psi_0 + c_1\psi_1 + \eta]\|_{H^2} \le Lip < 1.$$

A direct application of (2.5) and (4.20) shows that

$$\|(H-\Omega)^{-1}\tilde{P}D_\eta N[c_0\psi_0 + c_1\psi_1 + \eta]\|_{H^2} \le 3k\ p(d_*^{-1})\|c_0\psi_0 + c_1\psi_1 + \eta\|_{H^2}^2. \tag{4.22}$$

Fix $Lip = 3/4$. Then, a sufficient condition for invertibility is

$$3k\ p(d_*^{-1})\ \|c_0\psi_0 + c_1\psi_1 + \eta\|_{H^2}^2 \le Lip = 3/4, \tag{4.23}$$

which allows us to choose $n_* = \frac{1}{2}\sqrt{\frac{1}{kp(d_*^{-1})}}$ independently of Ω.

But, if (4.23) holds, then, from (4.22), the H^2 operator

$$(H-\Omega)^{-1}\tilde{P}N[c_0\psi_0 + c_1\psi_1 + \cdot]$$

is Lipschitz with Lipschitz constant less than or equal to $Lip = 3/4$. The standard contraction principle estimate applied to (4.19) gives

$$\begin{aligned}\|\eta\|_{H^2} &\leq \frac{1}{1-Lip}\|(H-\Omega)^{-1}\tilde{P}N[c_0\psi_0+c_1\psi_1]\|_{H^2}\\ &\leq 4p(d_*^{-1})\ k\|c_0\psi_0+c_1\psi_1\|_{H^2}^3. \end{aligned} \tag{4.24}$$

Plugging the above estimate into (4.23) gives

$$\|c_0\psi_0+c_1\psi_1\|_{H^2}+4p(d_*^{-1})\ k\|c_0\psi_0+c_1\psi_1\|_{H^2}^3 \leq \frac{1}{2\sqrt{p(d_*^{-1})k}}.$$

Since the left-hand side is continuous in $(c_0,c_1)\in\mathbb{C}^2$ and zero for $c_0=c_1=0$, one can construct $r_*>0$ depending only on $d_*,\ k$ such that the above inequality, hence (4.23) and (4.24), all hold whenever $|c_0|+|c_1|\leq r_*$. Finally, (4.17) now follows from (4.24). ∎

In particular, for the double-well potential we have the following proposition.

PROPOSITION 4.2. *Let $V=V_L$ denote the double-well potential with well-separation L. There exist $L_*>0, \varepsilon(L_*)>0$, and $d_*(L_*)>0$ such that for $L>L_*$, we have that for (c_0,c_1,Ω) satisfying dist$(\Omega,\sigma(H)\setminus\{\Omega_0,\Omega_1\})\geq d_*(L_*)$ and $|c_0|+|c_1|<\varepsilon(L_*)$, $\eta[c_0,c_1,\Omega]$ is defined and analytic and satisfies the bound* (4.17) *for some $C_*>0$.*

Proof. Since Ω_0,Ω_1,ψ_0, and ψ_1 can be controlled uniformly in L large, both d_* and r_* in the previous proposition can be controlled uniformly in L large. ∎

Next we study the symmetries of $\eta(c_0,c_1,\Omega)$ and the properties of $\mathcal{R}(c_0,c_1,\eta)$ which we will use in analyzing (2.16)-(2.17). The following result is a direct consequence of the symmetries of (2.18) and Proposition 4.1.

PROPOSITION 4.3. *We have*

$$\eta(e^{i\theta}c_0,e^{i\theta}c_1,\Omega)=e^{i\theta}\eta(c_0,c_1,\Omega) \qquad \text{for } 0\leq\theta<2\pi, \tag{4.25}$$

$$\overline{\eta(c_0,c_1,\Omega)}=\eta(\overline{c_0},\overline{c_1},\overline{\Omega}), \tag{4.26}$$

in particular

$$\eta(e^{i\theta}c_0,c_1=0,\Omega)=e^{i\theta}\eta(c_0,c_1=0,\Omega), \tag{4.27}$$

$$\eta(c_0=0,e^{i\theta}c_1,\Omega)=e^{i\theta}\eta(c_0=0,c_1,\Omega), \tag{4.28}$$

$\eta(c_0,0,\Omega)$ is even in x_1, $\eta(0,c_1,\Omega)$ is odd in x_1, and if c_0,c_1, and Ω are real-valued, then $\eta(c_0,c_1,\Omega)$ is real-valued.

In addition

$$\langle\psi_0,\mathcal{R}(c_0,c_1,\eta)\rangle=c_0\ f_0\left(c_0,c_1,\Omega\right), \tag{4.29}$$

$$\langle\psi_1,\mathcal{R}(c_0,c_1,\eta)\rangle=c_1\ f_1\left(c_0,c_1,\Omega\right), \tag{4.30}$$

where, for any $0\leq\theta<2\pi$ and $j=0,1$,

$$f_j\left(e^{i\theta}c_0,e^{i\theta}c_1,\Omega\right)=f_j\left(c_0,c_1,\Omega\right), \tag{4.31}$$

$$\overline{f_j(c_0,c_1,\Omega\ )}=f_j\left(\overline{c_0},\overline{c_1},\overline{\Omega}\right), \tag{4.32}$$

$$|f_j\left(c_0,c_1,\Omega\right)|\leq C(|c_0|+|c_1|)^4 \tag{4.33}$$

for some constant $C > 0$. Moreover, both f_0 and f_1 can be written as absolutely convergent power series:

(4.34)
$$f_j(c_0, c_1, \Omega) = \sum_{\substack{k+l+m+n \geq 4,\\ k-l+m-n=0,\\ m+n=\text{even}}} b^j_{klmn}(\Omega) c_0^k\ \overline{c}_0^{\,l}\ c_1^m\ \overline{c}_1^{\,n}, \qquad j = 0, 1,$$

where $b^j_{klmn}(\Omega)$ are real valued when Ω is real valued. In particular, if $c_0,\ c_1$, and Ω are real valued, then $f_j(c_0, c_1, \Omega)$ is real valued and, in polar coordinates, for $c_0, c_1 \neq 0$, we have

$$f_j(|c_0|, |c_1|, \Delta\theta, \Omega) = \sum_{k+m\geq 2,\ p\in\mathbb{Z}} b^j_{kmp}(\Omega) e^{ip2\Delta\theta} |c_0|^{2k} |c_1|^{2m}, \qquad j = 0, 1, \tag{4.35}$$

where $\Delta\theta$ is the phase difference between $c_1 \in \mathbb{C}$ and $c_0 \in \mathbb{C}$.

Proof of Proposition 4.3. We start with (4.25) which clearly implies (4.27)-(4.28). We fix Ω and suppress dependence on it in subsequent notation. From (4.19) we have

$$\begin{aligned}
&\eta(e^{i\theta}c_0, e^{i\theta}c_1)\\
&\quad = -(H-\Omega)^{-1}\tilde{P}\\
&\qquad N(e^{i\theta}c_0\psi_0 + e^{i\theta}c_1\psi_1 + \eta, e^{i\theta}c_0\psi_0 + e^{i\theta}c_1\psi_1 + \eta, e^{i\theta}c_0\psi_0 + e^{i\theta}c_1\psi_1 + \eta)\\
&\quad = -(H-\Omega)^{-1}\tilde{P}\\
&\qquad e^{i\theta}N(c_0\psi_0 + c_1\psi_1 + e^{-i\theta}\eta, c_0\psi_0 + c_1\psi_1 + e^{-i\theta}\eta, c_0\psi_0 + c_1\psi_1 + e^{-i\theta}\eta),
\end{aligned}$$

where we used

$$N(a\phi_1, b\phi_2, c\phi_3) = a\overline{b}cN(\phi_1, \phi_2, \phi_3). \tag{4.36}$$

Consequently,

$$\begin{aligned}
e^{-i\theta}\eta(e^{i\theta}c_0, e^{i\theta}c_1) = -(H-\Omega)^{-1}\tilde{P}\\
N[c_0\psi_0 + c_1\psi_1 + e^{-i\theta}\eta, c_0\psi_0 + c_1\psi_1 + e^{-i\theta}\eta, c_0\psi_0 + c_1\psi_1 + e^{-i\theta}\eta]
\end{aligned}$$

which shows that both $e^{-i\theta}\eta(e^{i\theta}c_0, e^{i\theta}c_1)$ and $\eta(c_0, c_1)$ satisfy the same equation (4.19). From the uniqueness of the solution proved in Proposition 4.1 we have the relation (4.25).

A similar argument (and use of the complex conjugate) leads to (4.26) and to the parities of $\eta(c_0, 0)$ and $\eta(0, c_1)$.

To prove (4.29) and (4.30), recall that

$$\begin{aligned}
\mathcal{R}\left(c_0, c_1, \eta(c_0, c_1, \Omega)\right) &= N(c_0\psi_0 + c_1\psi_1 + \eta, c_0\psi_0 + c_1\psi_1 + \eta, c_0\psi_0 + c_1\psi_1 + \eta)\\
&\quad - N(c_0\psi_0 + c_1\psi_1, c_0\psi_0 + c_1\psi_1, c_0\psi_0 + c_1\psi_1).
\end{aligned} \tag{4.37}$$

Consider first the case $c_1 = \rho_1 \in \mathbb{R}$. Note that

$$\langle \psi_1 g, \mathcal{R}(c_0, \rho_1 = 0, \eta(c_0, 0))\rangle = 0.$$

Indeed, for $\rho_1 = 0$, all the functions in the arguments of $\mathcal{R}$ are even functions (in x_1) making $\mathcal{R}$ an even function. Since ψ_1 is odd we get that the above is the integral over the entire space of an odd function, and therefore equal to zero. Since

$\langle \psi_1, \mathcal{R}(c_0, \rho_1, \eta(c_0, \rho_1))\rangle$ is analytic in $\rho_1 \in \mathbb{R}$ by the composition rule and its Taylor series starts with zero, we get (4.30) for real $c_1 = \rho_1$. To extend the result for complex values c_1 we use the rotational symmetry of $\mathcal{R}$, namely from (4.25), (4.36), and (4.37) we have

$$\mathcal{R}\left(e^{i\theta}c_0, e^{i\theta}c_1, \eta(e^{i\theta}c_0, e^{i\theta}c_1, \Omega)\right) = e^{i\theta}\mathcal{R}\left(c_0, c_1, \eta(c_0, c_1, \Omega)\right), \qquad 0 \le \theta < 2\pi,$$

hence (4.30) holds for $c_1 = |c_1|e^{-i\theta}$ by extending f_1 via the equality (4.31).

A similar argument holds for (4.29). Identity (4.33) follows from the definition of $\mathcal{R}$ and (4.17) while identity (4.32) follows from (4.26).

We now turn to a proof of the expansions for f_j: (4.34) and (4.35). Note first that $\mathcal{R}$ is real analytic because in (4.37) N is real linear in each variable and η is real analytic by Proposition 4.1. Hence, both f_0 and f_1 given by (4.29)-(4.30) are real analytic in $c_0,\ c_1$ and can be written in power series of the type (4.34). Estimate (4.33) implies that $k + l + m + n \ge 4$, while the rotational invariance (4.31) implies $k-l+m-n = 0$. The following parity argument shows why $m+n$ hence $m-n = l-k$ and $k + l$ are all even. Assume $m + n$ is odd. Note that because of (4.29), b^0_{klmn} is the scalar product between an even function (in x_1) ψ_0 and the term in the power series of $\mathcal{R}$ in which ψ_1 is repeated $m+n$ times. The latter is an odd function (in x_1) because ψ_1 is an odd function and it is repeated an odd number of times. The scalar product and hence b^0_{klmn} for $m + n$ odd will be zero. A similar argument holds for $b^1_{klmn},\ m + n$ odd. Finally $b^j_{klmn}(\Omega)$ are real-valued when Ω is real because they are scalar products of real-valued functions.

The form (4.35) of the power series follows directly from (4.34) by expressing c_0 and c_1 in their polar forms: $c_0 = |c_0|e^{i\theta_0}$ and $c_1 = |c_1|e^{i\theta_1},\ \Delta\theta = \theta_1 - \theta_0$, and using that $m + n,\ k + l$, and $m - n = -(k - l)$ are all even. The proof of Proposition 4.3 is now complete. ∎

4.1. Ground state and excited state branches, prebifurcation. In this section we prove part (i) of Theorem 4.1 as well as a corresponding statement about the excited state. In particular, we show that for sufficiently small amplitude, the only nonlinear bound state families are those arising via bifurcation from the zero state at the eigenvalues Ω_0 and Ω_1. This is true for general potentials with two bound states. Here, however, we can determine threshold amplitude, $\mathcal{N}_{cr}$, above which the solution set changes.

A closed system of equations for c_0, c_1, and Ω, parametrized by $\mathcal{N}$, is obtained upon substitution of $\eta[c_0, c_1, \Omega]$ (Proposition 4.1) into (4.9)–(4.12). Furthermore, using the structural properties (4.29)-(4.30) of Proposition 4.3, we obtain

(4.38)
$$(\Omega_0 - \Omega)\, c_0 + a_{0000}|c_0|^2 c_0 + (a_{0110} + a_{0011})\,|c_1|^2 c_0 + a_{0011}c_1^2\bar{c}_0 + c_0 f_0(c_0, c_1, \Omega) = 0,$$

(4.39)
$$(\Omega_1 - \Omega)\, c_1 + a_{1111}|c_1|^2 c_1 + (a_{1010} + a_{1001})\,|c_0|^2 c_1 + a_{1010}c_0^2\bar{c}_1 + c_1 f_1(c_0, c_1, \Omega) = 0,$$

(4.40)
$$|c_0|^2 \ + \ |c_1|^2 \ + \mathcal{O}\left(|c_0|^2 + |c_1|^2\right)^3 = \mathcal{N}.$$

This system of equations is valid for $|c_0| + |c_1| < r_*$, independent of L (the distance between wells).

If we choose $c_1 = 0$, then the second equation in the system, (4.39), is satisfied. In this case, a nontrivial solution requires $c_0 \neq 0$. The first equation, (4.38), after

factoring out c_0 becomes

$$\Omega_0 - \Omega + a_{0000}|c_0|^2 + f_0(|c_0|, 0, \Omega) = 0, \tag{4.41}$$

where we used (4.31) to eliminate the phase of the complex quantity c_0. Since Ω is real, (4.41) becomes one equation with two real parameters $\Omega,\ |c_0|$. Since the right-hand side of (4.41) vanishes for $\Omega = \Omega_0$ and $|c_0| = 0$ and since the partial derivative of this function with respect to Ω, evaluated at this solution, is nonzero, we have by the implicit function theorem that there is a unique solution

$$\Omega = \Omega_g(|c_0|) = \Omega_0 + a_{0000}|c_0|^2 + \mathcal{O}(|c_0|^4). \tag{4.42}$$

By (4.40), for small amplitudes, the mapping from $|c_0|^2 + |c_1|^2$ to $\mathcal{N}$ is invertible. The family of solutions

$$|c_0| \ \mapsto\ \left(|c_0|e^{i\theta}, |c_1| = 0, \Omega = \Omega_g(|c_0|)\right), \quad \theta_0 \in [0, 2\pi),$$

defined for $|c_0|$ sufficiently small, corresponds to a family of symmetric nonlinear bound states, $u_{\mathcal{N}}$ with $\|u_{\mathcal{N}}\|_{L^2}^2 = \mathcal{N}$, bifurcating from the zero solution at the linear eigenvalue Ω_0

$$u_{\mathcal{N}} = (|c_0|\psi_0(x) \ +\ \eta[|c_0|, 0, \Omega_g(|c_0|)](x))\ e^{i\theta_0}, \quad \theta_0 \in [0, 2\pi);$$

see, for example, [21, 25, 26]. Since both ψ_0 and $\eta(|c_0|, 0, \Omega_g)$ are even (in x_1) we infer that $u_{\mathcal{N}}$ is symmetric (even).

Remark 4.1. A similar result holds for the case $c_0 = 0$ leading to the antisymmetric excited state branch.

PROPOSITION 4.4. *For $|c_0|+|c_1|$ sufficiently small, these two branches of solutions are the only nontrivial solutions of* (2.7).

Proof. Indeed, suppose the contrary. By local uniqueness of these branches, ensured by the implicit function theorem, a solution not already lying on one of these branches must have both c_0 and c_1 nonzero. Now, divide the first equation by c_0, the second equation by c_1, and subtract the results. Introducing polar coordinates

$$c_0 = \rho_0 e^{i\theta_0},\ c_1 = \rho_1 e^{i\theta_1},\ \Delta\theta = \theta_1 - \theta_0, \tag{4.43}$$

we obtain from (4.38)

$$\begin{aligned} \Omega_1 - \Omega_0 = a_{0000}\rho_0^2 \ +\ \left(a_{0110} + a_{0011} + a_{0011}e^{i2\Delta\theta}\right)\rho_1^2 \ +\ f_0\left(\rho_0, \rho_1, \Delta\theta, \Omega\right) \\ - a_{1111}\rho_1^2 \ -\ \left(a_{1001} + a_{1010} + a_{1010}e^{-i2\Delta\theta}\right)\rho_0^2 \ -\ f_1\left(\rho_0, \rho_1, \Delta\theta, \Omega\right). \end{aligned} \tag{4.44}$$

The left-hand side is nonzero while the right-hand side is continuous uniformly for Ω satisfying (4.15) and zero for $\rho_0 = 0 = \rho_1$. Equation (4.44) cannot hold for $|\rho_0|+|\rho_1| < \varepsilon$ where $\varepsilon > 0$ is independent of Ω. This completes the proof of Proposition 4.4. ∎

Note, however, that nothing can preclude validity of (4.44) for larger ρ_0 and ρ_1, possibly leading to a third branch of solutions of (2.7). In what follows, we show that this is indeed the case and the third branch bifurcates from the ground state one at a critical value of $\rho_0 = \rho_0^*$.

4.2. Symmetry-breaking bifurcation along the ground state/symmetric branch. A consequence of the previous section is that there are no bifurcations from the ground state branch for sufficiently small amplitude. We now seek a bifurcating

branch of solutions to (2.16), (4.40), along which $c_0 \cdot c_1 \neq 0$. As argued just above, along such a new branch one must have

$$\Omega_0 - \Omega + a_{0000}\rho_0^2 + \left(a_{0110} + a_{0011} + a_{0011}e^{i2\Delta\theta}\right)\rho_1^2 + f_0(\rho_0, \rho_1, \Delta\theta, \Omega) = 0, \tag{4.45}$$

$$\Omega_1 - \Omega + a_{1111}\rho_1^2 + \left(a_{1010} + a_{1001} + a_{1010}e^{-i2\Delta\theta}\right)\rho_0^2 + f_1(\rho_0, \rho_1, \Delta\theta, \Omega) = 0. \tag{4.46}$$

We first derive constraints on $\Delta\theta$. Consider the imaginary parts of the two equations and use the expansions (4.35) and the fact that Ω is real:

$$\begin{aligned} &a_{0011}\sin(2\Delta\theta)\rho_1^2 + \sum_{k,m\geq 1,\ p\in\mathbb{Z}} b^0_{kmp}(\Omega)\sin(p2\Delta\theta)\rho_0^{2k}\rho_1^{2m} \\ &\quad = \sin(2\Delta\theta)\rho_1^2(a_{0011} + \mathcal{O}(\rho_0^2 + \rho_1^2)) = 0, \\ &-a_{1010}\sin(2\Delta\theta)\rho_0^2 + \sum_{k,m\geq 1,\ p\in\mathbb{Z}} b^1_{kmp}(\Omega)\sin(p2\Delta\theta)\rho_0^{2k}\rho_1^{2m} \\ &\quad = \sin(2\Delta\theta)\rho_0^2(-a_{1010} + \mathcal{O}(\rho_0^2 + \rho_1^2)) = 0. \end{aligned}$$

Due to the nondegeneracy assumption (4.3) and $a_{0011} = a_{1010}$, see (2.19), both equations hold if and only if $\sin(2\Delta\theta) = 0$ or, equivalently,

$$\Delta\theta \in \left\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\right\}. \tag{4.47}$$

Case 1. $\Delta\theta \in \{0, \pi\}$: Here, the system (4.45)-(4.46) is equivalent to the same system of two real equations with three real parameters $\rho_0 \geq 0$, $\rho_1 \geq 0$, and Ω:

$$F_0(\rho_0, \rho_1, \Omega) \stackrel{def}{=} \Omega_0 - \Omega + a_{0000}\rho_0^2 + (a_{0110} + 2a_{0011})\rho_1^2 + f_0(\rho_0, \rho_1, \Omega) = 0, \tag{4.48}$$

$$F_1(\rho_0, \rho_1, \Omega) \stackrel{def}{=} \Omega_1 - \Omega + a_{1111}\rho_1^2 + (2a_{1010} + a_{1001})\rho_0^2 + f_1(\rho_0, \rho_1, \Omega) = 0. \tag{4.49}$$

We begin by seeking the point along the ground state branch $(\rho_0^*, 0, \Omega_g(\rho_0^*))$ from which a new family of solutions of (4.48)-(4.49), parametrized by $\rho_1 \geq 0$, bifurcates; see (4.42).

Recall first that for *any* $\rho_0 \geq 0$ sufficiently small, $F_0(\rho_0, 0, \Omega_g(\rho_0)) = 0$. A candidate for a bifurcation point is $\rho_0^* > 0$ for which, in addition,

$$F_1(\rho_0^*, 0, \Omega_g(\rho_0^*)) = 0. \tag{4.50}$$

Using (4.33) one can check that

$$F_1(\rho_0, 0, \Omega_g(\rho_0)) = \Omega_1 - \Omega_0 + \left(a_{1001} + 2a_{1010} - a_{0000} + \mathcal{O}(\rho_0^2)\right)\rho_0^2 = 0 \tag{4.51}$$

has a solution:

$$\rho_0^* = \sqrt{\frac{\Omega_1 - \Omega_0}{|a_{1001} + 2a_{1010} - a_{0000}|}}\ \left[1 + \mathcal{O}\left(\ \frac{\Omega_1 - \Omega_0}{|a_{1001} + 2a_{1010} - a_{0000}|^2}\ \right)\right]. \tag{4.52}$$

We now show that a new family of solutions bifurcates from the symmetric state at $(\rho_0^*, 0, \Omega_g(\rho_0^*))$. This is realized as a unique, one-parameter family of solutions

$$\rho_1 \mapsto (\rho_0(\rho_1), \rho_1, \Omega_{asym}(\rho_1)) \tag{4.53}$$

of the equations

$$F_0(\rho_0, \rho_1, \Omega) = 0, \quad F_1(\rho_0, \rho_1, \Omega) = 0. \tag{4.54}$$

To see this, note that by the preceding discussion we have $F_j(\rho_0^*, 0, \Omega_g(\rho_0^*)) = 0, j = 1, 2$. Moreover, the Jacobian

$$\left| \frac{\partial(F_0, F_1)}{\partial(\rho_0, \Omega)}(\rho_0^*, 0, \Omega_g(\rho_0^*)) \right| = 2\rho_0^*(a_{1001} + 2a_{1010} - a_{0000} + \mathcal{O}(\rho_0^{*2}))$$

is nonzero because $\rho_0^* > 0$ and

$$a_{1001} + 2a_{1010} - a_{0000} + \mathcal{O}(\rho_0^{*2}) < 0 \tag{4.55}$$

since ρ_0^* solves (4.51) and $\Omega_1 - \Omega_0 > 0$. Therefore, by the implicit function theorem, for small $\rho_1 > 0$, there is a unique solution of the system (4.48)-(4.49):

$$\rho_0 = \rho_0(\rho_1) = \rho_0^* + \frac{\rho_1^2}{2\rho_0^*}\left(\frac{a_{0110} + 2a_{0011} - a_{1111}}{a_{1001} + 2a_{1010} - a_{0000}} + \mathcal{O}(\rho_0^{*2})\right) + \mathcal{O}(\rho_1^4), \tag{4.56}$$

$$\Omega = \Omega_{asym}(\rho_1) = \Omega_g(\rho_0^*) + \rho_1^2\left(a_{1111} + (2a_{1010} + a_{1001})\frac{a_{0110} + 2a_{0011} - a_{1111}}{a_{1001} + 2a_{1010} - a_{0000}} + \mathcal{O}(\rho_0^{*2})\right) + \mathcal{O}(\rho_1^4). \tag{4.57}$$

Remark 4.2. (1) Due to the equivalence of $\mathcal{N}$ and $\rho_0^2 + \rho_1^2$ as parameters, for small amplitude, we have that symmetry is broken at

$$\mathcal{N}_{cr} \sim \frac{\Omega_1 - \Omega_0}{|a_{0000} - a_{1001} - 2a_{1010}|}. \tag{4.58}$$

(2) Note also that we have the family of solutions

$$e^{i\theta}\left(\rho_0(\rho_1)\psi_0 \pm \rho_1\psi_1 + \eta(\rho_0(\rho_1), \pm\rho_1, \Omega_{asym}(\rho_1))\right), \qquad 0 \le \theta < 2\pi,\ \rho_1 > 0. \tag{4.59}$$

Here the $\pm$ is present because the phase difference $\Delta\theta$ between c_0 and c_1 can be 0 or π; see (4.47) and immediately below it. Because $\rho_0 \neq 0 \neq \rho_1$ this branch is neither symmetric nor antisymmetric. Thus, symmetry breaking has taken place. In the case of the double well, the $\pm$ sign in (4.59) shows that the bound states on this asymmetric branch tend to localize in one of the two wells but not symmetrically in both; see also [3], [23], [14].

Case 2. $\Delta\theta \in \left\{\frac{\pi}{2}, \frac{3\pi}{2}\right\}$: In both cases the system (4.45)-(4.46) is equivalent to the system of two real equations, depending on three real parameters $\rho_0 \ge 0$, $\rho_1 \ge 0$, and Ω:

$$F_0(\rho_0, \rho_1, \Omega) \stackrel{def}{=} \Omega_0 - \Omega + a_{0000}\rho_0^2 + a_{0110}\rho_1^2 + f_0(\rho_0, \rho_1, \Omega) = 0, \tag{4.60}$$

$$F_1(\rho_0, \rho_1, \Omega) \stackrel{def}{=} \Omega_1 - \Omega + a_{1111}\rho_1^2 + a_{1001}\rho_0^2 + f_1(\rho_0, \rho_1, \Omega) = 0. \tag{4.61}$$

As before, in order to have a second bifurcation of the symmetric branch it is necessary to find a point, $(\rho_0^{**}, 0, \Omega_g(\rho_0^{**}))$, for which

$$F_1(\rho_0^{**}, 0, \Omega_g(\rho_0^{**}) = \Omega_1 - \Omega_0 + (a_{1001} - a_{0000})\,\rho_0^{**2} + \mathcal{O}(\rho_0^{**4}) = 0. \tag{4.62}$$

To preclude the existence of this bifurcation, one must have that there are no real solutions for ρ_0^{**} small. This is ensured by inequality (3.6): $a_{1001} - a_{0000} > 0$.

In fact (3.6) can be *deduced* from the following general argument and thus we need not hypothesize it. Let L_- and L_+ denote the second-order, self-adjoint Schrödinger operators, defined in section 5, and related to the real and imaginary parts of the linearized NLS-GP equation about a real-valued nonlinear bound state. Consider the full *complex* linearization of the nonlinear bound state equation for NLS-GP. A nontrivial solution of the type in question would correspond to a second element of the null space of L_-. Since zero is the lowest eigenvalue of L_- and the lowest eigenvalue of an elliptic second-order operator is nondegenerate, we have a contradiction. It follows that there is no bifurcation for small amplitude and therefore (3.6) holds.

Remark 4.3. A similar argument holds along the antisymmetric branch. We now denote $L_\pm$ as linearized operators associated with the *antisymmetric* (odd parity with respect to x_1) branch. A bifurcation would occur along it if and only if $\dim\ker L_- \geq 2$ (the antisymmetric nonlinear bound state is always in $\ker L_-$ along this branch) or $\dim\ker L_+ \geq 1$. In the regime in which the first two eigenvalues are still separated from the rest of the spectrum, the former means that the lowest eigenvalue of L_-, i.e., the one bifurcating from $\Omega_0 - \Omega_1$, crosses zero and becomes double. This contradicts the nondegeneracy of the ground state of the second-order elliptic operator L_-. Now, $\dim\ker L_+ \geq 1$ is also impossible. This follows from the fact that the nonlinearity is attractive, see (H3), which easily implies that, for the nonlinear antisymmetric bound state Ψ_Ω, we have $(\Psi_\Omega, L_+\Psi_\Omega) < 0$. Since L_+ is self-adjoint and its lowest eigenvalue has a symmetric eigenvector orthogonal on Ψ_Ω, we get that the first two eigenvalues must be strictly negative and none can become zero. A consequence of a lack of bifurcation along the antisymmetric branch are inequalities (3.7), (3.8).

In summary, there are no other bifurcations than those stated in Theorem 4.1 for NLS-GP, where our Lyapunov–Schmidt-type reduction applies, i.e., in the regime covered by Proposition 4.1.

5. Exchange of stability at the bifurcation point. In this section we consider the dynamic stability of the symmetric and asymmetric waves, associated with the branch bifurcating from the zero state at the *ground state frequency*, Ω_0, of the linear Schrödinger operator $-\Delta + V(x)$; see Figure 1.

The notion of stability with which we work is H^1-orbital Lyapunov stability.

DEFINITION 5.1. *The family of nonlinear bound states* $\{\Psi_\Omega\ e^{-i\Omega t} : \theta \in [0, 2\pi)\}$ *is* H^1*-orbitally Lyapunov stable if for every* $\varepsilon > 0$ *there is a* $\delta(\varepsilon) > 0$ *such that if the initial data* u_0 *satisfies*

$$\inf_{\theta\in[0,2\pi)} \|u_0(\cdot)\ -\ \Psi_\Omega(\cdot)e^{i\theta}\|_{H^1}\ <\ \delta,$$

then for all $t \neq 0$*, the solution* $u(x,t)$ *satisfies*

$$\inf_{\theta\in[0,2\pi)} \|u(\cdot,t)\ -\ \Psi_\Omega(\cdot)e^{i\theta}\|_{H^1}\ <\ \varepsilon.$$

In this section we prove the following theorem.

THEOREM 5.1. *Consider the bifurcations elucidated in Theorem* 4.1. *The symmetric branch is* H^1*-orbitally Lyapunov stable for* $0 \leq \rho_0 < \rho_0^*$ *or, equivalently,* $0 < \mathcal{N} < \mathcal{N}_{cr}$*. At the bifurcation point* $\rho_0 = \rho_0^*$ $(\mathcal{N} = \mathcal{N}_{cr})$*, there is an exchange of stability from the symmetric branch to the asymmetric branch. In particular, for* $\mathcal{N} > \mathcal{N}_{cr}$ *the asymmetric state is stable and the symmetric state is unstable.*

We summarize basic results on stability and instability. Introduce L_+ and L_-, real and imaginary parts, respectively, of the linearized operators about a real-valued state Ψ_Ω:

$$\begin{aligned} L_+ = L_+[\Psi_\Omega]\cdot &= (H-\Omega)\cdot\ + \partial_u N(u,u,u)\mid_{\Psi_\Omega} \\ &\equiv (H-\Omega)\cdot + D_u N[\Psi_\Omega](\cdot), \\ L_- = L_-[\Psi_\Omega]\cdot &= (H-\Omega)\cdot + N(\Psi_\Omega,\Psi_\Omega,\cdot). \end{aligned} \tag{5.1}$$

By (2.7) and (2.4), $L_-\Psi_\Omega = 0$.

We state a special case of known results on stability and instability, directly applicable to the symmetric branch which bifurcates from the zero state at the ground state frequency of $-\Delta + V$.

THEOREM 5.2 (see [32, 33, 10, 9, 15]).

(1) Stability: *Suppose L_+ has exactly one negative eigenvalue and L_- is nonnegative. Assume that*

$$\frac{d}{d\Omega}\int |\Psi_\Omega(x)|^2 dx < 0. \tag{5.2}$$

Then, Ψ_Ω is H^1-orbitally stable.

(2) Instability: *Suppose L_- is nonnegative. If $n_-(L_+) \geq 2$, then the linearized dynamics about Ψ_Ω has spatially localized solution which is exponentially growing in time. Moreover, Ψ_Ω is not H^1-orbitally stable.*

First we claim that along the branch of symmetric solutions bifurcating from the zero solution at frequency Ω_0, the hypothesis on L_- holds. To see that the operator $L_-[\Psi_\Omega]$ is always nonnegative, consider $L_-[\Psi_{\Omega_0}] = L_-[0] = -\Delta + V - \Omega_0$. Clearly, $L_-[0]$ is a nonnegative operator because Ω_0 is the lowest eigenvalue of $-\Delta + V$. Since clearly we have $L_-\Psi_\Omega = 0$, $0 \in spec(L_-[\Psi_\Omega])$. Since the lowest eigenvalue is necessarily simple, by continuity there cannot be any negative eigenvalues for Ω sufficiently close to Ω_0. Finally, if for some Ω, L_- has a negative eigenvalue, then by continuity there would be an Ω_* for which $L_-[\Psi_{\Omega_*}]$ would have a double eigenvalue at zero and no negative spectrum. But this contradicts that the ground state is simple. Therefore, it is the quantity $n_-(L_+)$ which controls whether or not Ψ_Ω is stable.

Next we discuss the slope condition (5.2). It is clear from the construction of the branch $\Omega \mapsto \Psi_\Omega$ that (5.2) holds for Ω near Ω_0. Suppose now that $\partial_\Omega \int |\Psi_\Omega|^2 = 0$. Then, $\langle \Psi_\Omega, \partial_\Omega \Psi_\Omega\rangle = 0$. As shown below, L_+ has exactly one negative eigenvalue for Ω sufficiently near Ω_0. It follows that $L_+ \geq 0$ on $\{\Psi_\Omega\}^\perp$ (see [32, 33]). Therefore, we have $(L_+^{\frac{1}{2}}\partial_\Omega\Psi_\Omega, L_+^{\frac{1}{2}}\partial_\Omega\Psi_\Omega) = (L_+\partial_\Omega\Psi_\Omega, \partial_\Omega\Psi_\Omega) = (\Psi_\Omega, \partial_\Omega\Psi_\Omega) = 0$. Therefore, $L_+^{\frac{1}{2}}\partial_\Omega\Psi_\Omega = 0$, implying $\Psi_\Omega = L_+\partial_\Omega\Psi_\Omega = 0$, which is a contradiction. It follows that (5.2) holds so long as $L_+ > 0$ on $\{\Psi_\Omega\}^\perp$ and when (5.2) first fails, it does so due to a nontrivial element of the nullspace of L_+.

Therefore, Ψ_Ω is stable so long as $n_-(L_+)$ does not increase. We shall now show that for $\mathcal{N} < \mathcal{N}_{cr}$, $n_-(L_+[\Psi_\Omega]) = 1$ but that along the symmetric branch for $\mathcal{N} > \mathcal{N}_{cr}$, $n_-(L_+[\Psi_\Omega]) = 2$. Furthermore, we show that along the bifurcating asymmetric branch, the hypotheses of Theorem 5.2 ensuring stability hold.

Remark 5.1. For simplicity we have considered the most important case where there is a transition from dynamical stability to dynamical instability along the symmetric branch, bifurcating from the ground state of H. However, our analysis which actually shows that along any symmetric branch, associated with any of the eigenvalues $\Omega_{2j}, j \geq 0$, of H, there is a critical $\mathcal{N} = \mathcal{N}_{cr}(j)$ such that as $\mathcal{N}$ is increased

through $\mathcal{N}_{cr}(j)$, the number of negative eigenvalues of the linearization about the symmetric state along the jth symmetric branch, $n_-(L_+^{(j)})$, increases by one. By the results in [15, 9, 19], this has implications for the number of unstable modes of higher order ($j \geq 1$) symmetric states.

Consider the spectral problem for $L_+ = L_+[\Psi_\Omega]$:

$$L_+[\Psi_\Omega]\phi = \mu\phi. \tag{5.3}$$

We now formulate a Lyapunov–Schmidt reduction of (5.3) and then relate it to our formulation for nonlinear bound states. We first decompose ϕ relative to the states $\psi_0,\ \psi_1$ and their orthogonal complement:

$$\phi = \alpha_0\psi_0 \ + \ \alpha_1\psi_1 \ + \ \xi, \quad (\psi_j, \xi) = 0,\ j = 0, 1.$$

Projecting (5.3) onto $\psi_0,\ \psi_1$ and onto the range of $\tilde{P}$ we obtain the system

$$\langle \psi_0, L_+[\Psi_\Omega](\alpha_0\psi_0 \ + \ \alpha_1\psi_1 \ + \ \xi)\rangle = \mu\alpha_0, \tag{5.4}$$

$$\langle \psi_1, L_+[\Psi_\Omega](\alpha_0\psi_0 \ + \ \alpha_1\psi_1 \ + \ \xi)\rangle = \mu\alpha_1, \tag{5.5}$$

$$(H - \Omega)\xi + D_u N[\Psi_\Omega](\alpha_0\psi_0 + \alpha_1\psi_1 + \xi) = \mu\xi. \tag{5.6}$$

The last equation can be rewritten in the form

$$\Big[I \ + \ (H - \Omega - \mu)^{-1}\tilde{P}D_u N[\Psi_\Omega]\Big]\,\xi = -(H - \Omega - \mu)^{-1}\tilde{P}D_u N[\Psi_\Omega](\alpha_0\psi_0 + \alpha_1\psi_1). \tag{5.7}$$

The operator on the right-hand side of (5.7) is essentially the Jacobian studied in the proof of Proposition 4.1, evaluated at $\Omega+\mu$. Hence, by the proof of Proposition 4.1, if $\Omega+\mu$ satisfies (4.15) and $\|\Psi_\Omega\|_{H^2} \leq \mathcal{N}_*$, then the operator $I+(H-\Omega-\mu)^{-1}\tilde{P}D_u N[\Psi_\Omega]$ is invertible on H^2 and (5.7) has a unique solution

$$\begin{aligned}
\xi \ &\stackrel{def}{=} \ \xi[\mu, \alpha_0, \alpha_1, \Omega] \\
&\equiv \ Q[\mu, \Psi_\Omega](\alpha_0\psi_0 + \alpha_1\psi_1) \\
&= \ -(I + (H - \Omega - \mu)^{-1}\tilde{P}D_u N[\Psi_\Omega])^{-1}(H - \Omega - \mu)^{-1}\tilde{P}D_u N[\Psi_\Omega](\alpha_0\psi_0 + \alpha_1\psi_1) \\
&= \ \mathcal{O}\left[(|\rho_0| + |\rho_1|)^2\right] \ [\alpha_0\psi_0 + \alpha_1\psi_1]\,.
\end{aligned} \tag{5.8}$$

The last relation follows from $D_u N[\psi]$ being a quadratic form in $\Psi_\Omega = \rho_0\psi_0 + \rho_1\psi_1 + \mathcal{O}((|\rho_0| + |\rho_1|)^3)$.

Substitution of the expression for ξ as a functional of α_j into (5.4) and (5.5) we get a closed system of two real equations:

$$\begin{aligned}
&(\Omega_0 - \Omega)\alpha_0 \ + \ \langle \psi_0, D_u N[\Psi_\Omega]\ (I + Q[\mu, \Psi_\Omega])\ (\alpha_0\psi_0 + \alpha_1\psi_1)\rangle = \mu\ \alpha_0, \\
&(\Omega_1 - \Omega)\alpha_1 \ + \ \langle \psi_1, D_u N[\Psi_\Omega]\ (I + Q[\mu, \Psi_\Omega])\ (\alpha_0\psi_0 + \alpha_1\psi_1)\rangle = \mu\ \alpha_1.
\end{aligned} \tag{5.9}$$

The system (5.9) is the Lyapunov–Schmidt reduction of the linear eigenvalue problem for L_+ with eigenvalue parameter μ. *Our next step will be to write it in a form relating it to the linearization of the Lyapunov–Schmidt reduction of the nonlinear problem.*

Remark 5.2. For $\|\Psi_\Omega\|_{H^2} \leq n_*$, the above system is equivalent to the eigenvalue problem for the operator $L_+[\Psi_\Omega]$ with eigenvalue parameter μ as long as (4.15) holds

with Ω replaced by $\Omega+\mu$. This restriction on the spectral parameter, μ, is in fact very mild and has no impact on the analysis. This is because we are primarily interested in μ near zero as we are interested in detecting the crossing of an eigenvalue of L_+ from positive to negative reals as $\mathcal{N}$ is increased beyond some $\mathcal{N}_{cr}$. The values of μ for which (4.15) does not hold do not play a role in any change of index, $n_-(L_+)$.

First rewrite (5.9) as

$$\begin{aligned}(\Omega_0-\Omega-\mu)\alpha_0 \;+\; &\langle\psi_0, D_uN[\Psi_\Omega]\;(I+Q[0,\Psi_\Omega])\;(\alpha_0\psi_0+\alpha_1\psi_1)\rangle\\ &+\;\langle\psi_0, D_uN[\Psi_\Omega]\;\Delta Q[\mu,\Psi_\Omega]\;(\alpha_0\psi_0+\alpha_1\psi_1)\rangle=0,\end{aligned}\tag{5.10}$$

$$\begin{aligned}(\Omega_1-\Omega-\mu)\alpha_1 \;+\; &\langle\psi_1, D_uN[\Psi_\Omega]\;(I+Q[0,\Psi_\Omega])\;(\alpha_0\psi_0+\alpha_1\psi_1)\rangle\\ &+\;\langle\psi_1, D_uN[\Psi_\Omega]\;\Delta Q[\mu,\Psi_\Omega]\;(\alpha_0\psi_0+\alpha_1\psi_1)\rangle=0.\end{aligned}\tag{5.11}$$

Here,

$$\Delta Q\left[\mu,\Psi_\Omega\right]=Q\left[\mu,\Psi_\Omega\right]\;-\;Q\left[0,\Psi_\Omega\right].\tag{5.12}$$

Note that terms involving ΔQ in (5.10), (5.11) are of size $\mathcal{O}[(\rho_0^2+\rho_1^2)\mu\alpha_j]$.

Proposition 5.1.

$$Q[0,\Psi_\Omega](\alpha_0\psi_0+\alpha_1\psi_1)=\partial_{\rho_0}\eta[\rho_0,\rho_1,\Omega]\;\alpha_0\;+\;\partial_{\rho_1}\eta[\rho_0,\rho_1,\Omega]\;\alpha_1.\tag{5.13}$$

Proof. Recall that η satisfies

$$F(\rho_0,\rho_1,\Omega,\eta)\;\equiv\;\eta+(H-\Omega)^{-1}\tilde{P}N[\rho_0\psi_0+\rho_1\psi_1+\eta]=0.\tag{5.14}$$

Differentiation with respect to $\rho_j,\;j=0,1$, yields

$$\Big(I\;+\;(H-\Omega)^{-1}\tilde{P}D_uN[\Psi_\Omega]\Big)\;\partial_{\rho_j}\eta=-\,(H-\Omega)^{-1}\,\tilde{P}D_uN[\Psi_\Omega]\psi_j,\tag{5.15}$$

where

$$\Psi_\Omega=\rho_0\psi_0+\rho_1\psi_1+\eta[\rho_0,\rho_1,\Omega].$$

Thus,

$$\partial_{\rho_j}\;\eta=Q[0,\Psi_\Omega]\;\psi_j,\tag{5.16}$$

from which Proposition 5.1 follows. ∎

We now use Proposition 5.1 to rewrite the first inner products in (5.10), (5.11). For $k=0,1$,

$$\begin{aligned}&\langle\psi_k, D_uN[\Psi_\Omega]\;(I+Q[0,\Psi_\Omega])\;(\alpha_0\psi_0+\alpha_1\psi_1)\rangle\\ &=\sum_{j=0}^{1}\langle\psi_k, D_uN[\rho_0\psi_0+\rho_1\psi_1+\eta](\psi_j+\partial_{\rho_j}\eta)\rangle\alpha_j\\ &=\;\sum_{j=0}^{1}\;\frac{\partial}{\partial\rho_j}\langle\psi_k, N[\Psi_\Omega]\rangle\;\alpha_j\\ &=\;\sum_{j=0}^{1}\;\partial_{\rho_j}\;\langle\psi_k, N[\rho_0\psi_0+\rho_1\psi_1]\rangle\;\alpha_j\;+\;\partial_{\rho_j}\;\left[\rho_k f_k(\rho_0,\rho_1,\Omega)\right],\end{aligned}\tag{5.17}$$

where $N[\psi_\Omega] = N[\rho_0\psi_0 + \rho_1\psi_1] + \mathcal{R}$; see (2.16)–(2.18) and (4.29), (4.30). Therefore, the Lyapunov–Schmidt reduction of the eigenvalue problem for L_+ becomes

$$(5.18)\quad (\Omega_0 - \Omega - \mu)\alpha_0 \;+\; \sum_{j=0}^{1} \partial_{\rho_j}\, \langle \psi_0, N[\rho_0\psi_0 + \rho_1\psi_1]\rangle\, \alpha_j \;+\; \partial_{\rho_j}\, [\rho_0 f_0(\rho_0, \rho_1, \Omega)] + \langle \psi_0, D_u N[\Psi_\Omega]\, \Delta Q[\mu, \Psi_\Omega]\, (\alpha_0\psi_0 + \alpha_1\psi_1)\rangle = 0,$$

$$(5.19)\quad (\Omega_1 - \Omega - \mu)\alpha_1 \;+\; \sum_{j=0}^{1} \partial_{\rho_j}\, \langle \psi_1, N[\rho_0\psi_0 + \rho_1\psi_1]\rangle\, \alpha_j \;+\; \partial_{\rho_j}\, [\rho_1 f_1(\rho_0, \rho_1, \Omega)] + \langle \psi_1, D_u N[\Psi_\Omega]\, \Delta Q[\mu, \Psi_\Omega]\, (\alpha_0\psi_0 + \alpha_1\psi_1)\rangle = 0.$$

This can be written succinctly in a matrix form as

$$(5.20)\qquad [M \;-\mu\; +\; \mathcal{C}(\mu)] \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

where

(5.21)

$$M = M[\Omega, \rho_0, \rho_1] = \begin{pmatrix} \Omega_0 - \Omega + 3a_{0000}\rho_0^2 + (a_{0110} + 2a_{0011})\rho_1^2 + \partial_{\rho_0}(\rho_0 f_0) & 2(a_{0110} + 2a_{0011})\rho_0\rho_1 + \partial_{\rho_1}(\rho_0 f_0) \\ 2(2a_{1010} + a_{1001})\rho_0\rho_1 + \partial_{\rho_0}(\rho_1 f_1) & (\Omega_1 - \Omega) + 3a_{1111}\rho_1^2 + (2a_{1010} + a_{1001})\rho_0^2 + \partial_{\rho_1}(\rho_1 f_1) \end{pmatrix}$$

and

$$(5.22)\qquad \mathcal{C}(\mu)_{lm} = \langle \psi_l, D_u N[\Psi_\Omega]\Delta Q[\mu, \Psi_\Omega]\psi_m\rangle, \quad l, m = 0, 1.$$

Note that

$$(5.23)\qquad \mathcal{C}(\mu = 0) = 0.$$

Recall that μ is the spectral parameter for the eigenvalue problem L_+, (5.3), and we are interested in $n_-(L_+[\Psi_\Omega])$, the number of negative eigenvalues along a family of nonlinear bound states $\Omega \mapsto \Psi_\Omega$. By Theorem 5.2, $n_-(L_+)$ determines the stability or instability of a particular state. This question has now been mapped to the problem of following the roots of

$$(5.24)\qquad D(\mu, \rho_0, \rho_1) = \det(\mu I - M - \mathcal{C}(\mu)) = 0,$$

where ρ_0 and ρ_1 are parameters along the different branches of nonlinear bounds states. Since $\mathcal{C}(\mu)$, defined in (5.22), is small for small amplitude nonlinear bound states, we expect the roots, μ, to be small perturbations of the eigenvalues of the matrix M. We study these roots along the symmetric ($M = M(\Omega_g(\rho_0), \rho_0, 0)$) and asymmetric branch ($M = M(\Omega_{asym}(\rho_1), \rho_0(\rho_1), \rho_1)$) using the implicit function theorem.

Symmetric branch. Along the symmetric branch:

$$\rho_1 = 0, \quad \rho_0 \geq 0, \quad \Omega = \Omega_g = \Omega_0 + a_{0000}\rho_0^2 + \mathcal{O}(\rho_0^4),$$
$$\Psi_\Omega = \rho_0\psi_0 + \eta(\rho_0, 0, \Omega) = \text{symmetric}.$$

Thus, $D = D(\mu, \rho_0)$. Moreover, the system (5.20) is diagonal. This is because Q and hence ΔQ preserve parity at a symmetric Ψ_Ω; see their definitions (5.8) and (5.12). Therefore $\mathcal{C}_{01} = 0 = \mathcal{C}_{10}$, since each is the scalar product of an even and an odd function. Moreover, from (4.35) we get $\frac{\partial f_j}{\partial \rho_1}(\rho_0, 0, \Omega) = 0,\ j = 0, 1.$

Therefore, the matrix $\mu I - M - \mathcal{C}(\mu)$ is diagonal and μ is an eigenvalue of $L_+[\psi_{\Omega_g(\rho_0)}]$ if and only if μ is a root of either

$$P_0(\mu, \rho_0) \ \equiv \ \mu - M_{00}(\rho_0) - \mathcal{C}_{00}(\mu, \rho_0) = 0 \tag{5.25}$$

or

$$P_1(\mu, \rho_0) \ \equiv \ \mu - M_{11}(\rho_0) - \mathcal{C}_{11}(\mu, \rho_0) = 0. \tag{5.26}$$

Both P_0 and P_1 are analytic in μ and ρ_0 and it is easy to check that

$$P_0(0, 0) = 0, \qquad \partial_\mu P_0(0, 0) = 1$$

and

$$P_1(\Omega_1 - \Omega_0, 0) = 0, \qquad \partial_\mu P_1(\Omega_1 - \Omega_0, 0) = 1.$$

Formally differentiating (5.25) or (5.26) with respect to ρ_0 gives

$$\partial_{\rho_0} \mu_j = \frac{\partial_{\rho_0} M_{jj} + \partial_{\rho_0} C_{jj}}{1 - \partial_\mu C_{jj}}. \tag{5.27}$$

By the implicit function theorem (5.25) and (5.26) define, respectively, μ_0 and μ_1 as smooth functions of ρ provided

$$|\partial_\mu \mathcal{C}_{jj}| < 1, \qquad j = 0, 1. \tag{5.28}$$

A direct calculation using (5.8) and estimates (2.5), (4.20) shows that in the regime of interest: Ω satisfying (4.15), it suffices to have

$$\|\Psi_\Omega\|_{H^2} \leq n_* \left(9 \max(\|\psi_0\|_{H^2}, \|\psi_1\|_{H^2})\right)^{-\frac{1}{4}}, \tag{5.29}$$

where n_* is given by Proposition 4.1. The latter can be reduced to an estimate on ρ_0 via the above definition of Ψ_Ω and (4.24) as in the end of the proof of Proposition 4.1.

Therefore, under conditions (4.15) and (5.29), we have a unique solution μ_0, respectively μ_1, of (5.25), respectively (5.26). Moreover, the two solutions are analytic in ρ_0 and, for small ρ_0, we have the following estimates:

$$\mu_0 = 2a_{0000}\rho_0^2 + \mathcal{O}(\rho_0^4) < 0, \tag{5.30}$$

$$\mu_1 = \Omega_1 - \Omega_0 + \mathcal{O}(\rho_0^2) > 0, \tag{5.31}$$

where we used $a_{0000} \equiv g\langle \psi_0^2, K[\psi_0^2]\rangle < 0$ and $\mu_1(\rho_0 = 0) = \Omega_1 - \Omega_0 > 0$.

We claim that μ_1 changes sign for the first time at $\rho_0 = \rho_0^*$. Indeed, by continuity, the sign can only change when $\mu_1 = 0$, i.e., when (5.26) has a solution of the form $(0, \rho_0)$. Since $\mathcal{C}_{11}(0, \rho_0) = 0$ (see (5.23)) (5.26) becomes

$$0 = M_{11}(\rho_0) = \Omega_1 - \Omega_g(\rho_0) + (2a_{1010} + a_{1001})\rho_0^2 + f_1(\rho_0, 0, \Omega_g) = F_1(\rho_0, 0, \Omega_g(\rho_0));$$

see (5.21) and note that $\rho_1 = 0$. But this equation is the same as (4.50), which determines the bifurcation point ρ_0^*. Thus, as expected, $D(\mu, \rho_0) = 0$ has a root

$\rho_1(\rho_0^*) = 0$ or, equivalently, L_+ has a zero eigenvalue at the bifurcation point. Note that the associated null eigenfunction has odd parity in one space dimension and is, more generally, nonsymmetric and changes sign.

To see that $\mu_1(\rho_0)$ changes sign at $\rho_0 = \rho_0^*$ we differentiate (5.26) with respect to ρ_0 at $\rho_0 = \rho_0^*$ and obtain from (5.27) that

$$\partial_{\rho_0}\mu_1 = \frac{\partial_{\rho_0}M_{11} + \partial_{\rho_0}C_{11}}{1 - \partial_\mu C_{11}} < 0.$$

This follows because the denominator is positive, by (5.28), while direct calculation gives for the numerator

$$\partial_{\rho_0}M_{11}(\rho_0^*) + \partial_{\rho_0}C_{11}(\rho_0^*) = 2\rho_0^*\left(a_{1001} + 2a_{1010} - a_{0000} + \mathcal{O}(\rho_0^{*2})\right) < 0;$$

see (4.55). Therefore μ_1 becomes negative for $\rho_0 > \rho_0^*$ at least when $|\rho_0 - \rho_0^*|$ is small enough.

In conclusion, $L_+[\Omega_g(\rho_0)]$ has exactly one negative eigenvalue for $0 \le \rho_0 < \rho_0^*$ and two negative eigenvalues for $\rho_0 > \rho_0^*$ and $|\rho_0 - \rho_0^*|$ small. Therefore, following the criteria of [32, 33, 10, 15, 9, 16], the symmetric branch is stable for $0 \le \rho_0 < \rho_0^*$ and becomes unstable past the bifurcation point.

Asymmetric branch: Stability for $\mathcal{N} > \mathcal{N}_{cr}$. Finally, we study the behavior of the eigenvalue problem (5.20) on the asymmetric branch:

$$0 \le \rho_1 \ll 1,$$

$$\rho_0 = \rho_0(\rho_1) = \rho_0^* + \frac{\rho_1^2}{2\rho_0^*}\left(\frac{a_{0110} + 2a_{0011} - a_{1111}}{a_{1001} + 2a_{1010} - a_{0000}} + \mathcal{O}(\rho_0^{*2})\right) + \mathcal{O}(\rho_1^4),$$

$$\begin{aligned}\Omega = \Omega_{asym}(\rho_1) = \Omega_g(\rho_0^*) &\\ + \rho_1^2\left(a_{1111} + (2a_{1010} + a_{1001})\frac{a_{0110} + 2a_{0011} - a_{1111}}{a_{1001} + 2a_{1010} - a_{0000}} + \mathcal{O}(\rho_0^{*2})\right)&\\ + \mathcal{O}(\rho_1^4),&\end{aligned} \tag{5.33}$$

$$\Psi_\Omega = \rho_0(\rho_1)\psi_0 + \rho_1\psi_1 + \eta(\rho_0(\rho_1), \rho_1, \Omega_{asym}(\rho_1)).$$

The eigenvalues will be given by the zeros of the real-valued function

$$D(\mu, \rho_1) = \det(\mu I - M(\rho_1) - \mathcal{C}(\mu, \rho_1)), \tag{5.34}$$

which is analytic in μ and ρ_1 for $\Omega + \mu$ satisfying (4.15) and ρ_1 small. Note that at $\rho_1 = 0$ we are still on the symmetric branch at the bifurcation point $\rho_0 = \rho_0^*$. Hence, the matrix is diagonal and

$$D(\mu, 0) = P_0(\mu, \rho_0^*)P_1(\mu, \rho_0^*), \tag{5.35}$$

where $P_j,\ j = 0, 1$, are defined in (5.25)-(5.26). In the previous subsection we showed that each $P_j(\cdot, \rho_0^*)$ has exactly one zero, μ_j, on the interval $-\infty < \mu < d_* - \Omega_g(\rho_0^*) > 0$. The zeros were simple by our implicit function theorem application in which

$$\partial_\mu P_j(\mu_j, \rho_0^*) = 1 - \partial_\mu C_{jj} > 0; \tag{5.36}$$

see (5.28). In addition one can easily deduce that $\lim_{\mu\to-\infty} P_j(\mu, \rho_0^*) = -\infty$ by using the definitions (5.22), (5.12) and the fact that $\|(H - \Omega - \mu)^{-1}\|_{L^2\to H^2} \overset{\mu\to-\infty}{\to} 0$ which implies $\|Q[\mu, \Psi_\Omega]\|_{H^2\to H^2} \overset{\mu\to-\infty}{\to} 0$.

Consequently, $D(\cdot,0)$ has exactly two simple zeros $\mu_0 < 0$ and $\mu_1 = 0$ on the interval $-\infty < \mu \le (-d_* - \Omega_g(\rho_0^*))/2 > 0$, which are both simple and $\lim_{\mu\to-\infty} D(\mu,0) = \infty$. It is well known, and a consequence of continuity arguments and of the implicit function theorem, that the previous statement continues to hold for small perturbations. More precisely, there exists $\varepsilon > 0$ such that whenever $|\rho_1| < \varepsilon$, $D(\cdot,\rho_1)$ has exactly two zeros $\mu_0(\rho_1) < 0$ and $\mu_1(\rho_1)$ on the interval $-\infty < \mu \le (-d_* - \Omega_g(\rho_0^*))/2 > 0$, which are both simple and analytic in ρ_1.

Since we are interested in $n_-(L_+)$, the number of negative eigenvalues of L_+, we still need to determine the sign of $\mu_1(\rho_1)$. In what follows we will show that its derivatives satisfy

$$\partial_{\rho_1}\mu_1(0) = 0, \quad \partial_{\rho_1}^2\mu_1(0) > 0. \tag{5.37}$$

We can then conclude that for $0 < \rho_1 \ll 1$, $\mu_1(\rho_1) > 0$ and L_+ has exactly one (simple) negative eigenvalue, $\mu_0(\rho_1)$. Therefore, the asymmetric branch is stable.

We now prove (5.37). By differentiating

$$D(\mu_1(\rho_1),\rho_1) = 0 \tag{5.38}$$

once with respect to ρ_1 at $\rho_1 = 0$ we get

$$\partial_\mu D(0,0)\partial_{\rho_1}\mu_1(0) + \partial_{\rho_1} D(0,0) = 0.$$

Using (5.35) we obtain

$$\partial_\mu D(0,0) = P_0(0,\rho_0^*)\partial_\mu P_1(\mu_1 = 0,\rho_0^*) > 0, \tag{5.39}$$

where we used (5.36) and $P_0(0,\rho_0^*) = -M_{00}(\rho_0^*) > 0$. Using (5.34) and (5.23) we obtain

$$\partial_{\rho_1} D(0,0) = \frac{\partial \det(M)}{\partial\rho_1}(\rho_1 = 0) = \det 10 + \det 01, \tag{5.40}$$

where $\det ij$ is the determinant evaluated at $\rho_1 = 0$ of the matrix obtained from M by differentiating the first row i times, respectively the second row j times. $\det ij$ can be evaluated using (4.34), (4.50), and (5.33).

Note that the second row of $\det 10$ is zero and therefore $\det 10 = 0$. Furthermore, $\det 01$ is zero because its second column is zero. Therefore, by (5.40) we have $\partial_{\rho_1}\mu_1(0) = 0$.

We now calculate $\partial_{\rho_1}^2\mu_1(\rho_1 = 0)$. Differentiate (5.38) twice with respect to ρ_1 at $\rho_1 = 0$ and use $\partial_{\rho_1}\mu_1(0) = 0$ to obtain

$$\partial_\mu D(0,0)\partial_{\rho_1}^2\mu_1(0) + \partial_{\rho_1}^2 D(0,0) = 0$$

which implies, by (5.39),

$$\operatorname{sign}(\partial_{\rho_1}^2\mu_1(0)) = -\operatorname{sign}(\partial_{\rho_1}^2 D(0,0)).$$

But, as before, (5.34) and (5.23) imply

$$\partial_{\rho_1}^2 D(0,0) = \frac{\partial^2 \det(M)}{\partial\rho_1^2}(0) = \det 20 + 2\det 11 + \det 02 < 0.$$

The last inequality is a consequence of the following argument. First, $\det 20 = 0$, since its second row is zero. A direct calculation using the definition of M and relations (5.32) show

$$\det 11 = -4(a_{0110} + 2a_{0011})(2a_{1010} + a_{1001})\rho_0^{*2} + \mathcal{O}(\rho_0^{*4}),$$
$$\det 02 = 8a_{0000}a_{1111}\rho_0^{*2} + \mathcal{O}(\rho_0^{*4}).$$

By hypothesis (4.2) and by (3.7), shown in Remark 4.3, we have

$$a_{0000} > a_{1001} + 2a_{1010}, \quad a_{1111} > a_{0110} + 2a_{0011}.$$

This implies, for ρ_0^* sufficiently small, that

$$2\det 11 + \det 02 < 0.$$

Therefore, $\partial^2_{\rho_1}\mu_1(0) > 0$ and the proof of Theorem 5.1 is now complete.

Remark 5.3. We remark that, in the special case of $g < 0$ a constant and $V = V_L$ a double-well potential with well-separation parameter L, for L large all coefficients $a_{klmn} = a_{klmn}(L)$ converge to the same value $g\alpha^2 < 0$. This implies

$$2\det 11 + \det 02 = (-64g^2\alpha^4 + \mathcal{O}(e^{-\tau L}))\rho_0^{*2} + \mathcal{O}(\rho_0^{*4}) < 0.$$

6. Numerical study of symmetry breaking.

6.1. Symmetry-breaking bifurcation for fixed well-separation, *L*. In this section we numerically compute the bifurcation diagram for the lowest energy nonlinear bound state branch for NLS-GP equation (2.1) and compare these results to the predictions of the finite-dimensional approximation equations (3.12), (3.13). Specifically, we numerically compute the bifurcation structure of (2.1) for a double-well potential, $V_L(x)$, of the form

$$V(x) = V_0\left[\frac{1}{\sqrt{4\pi s^2}}\exp\left(-\frac{(x-L/2)^2}{4s^2}\right) + \frac{1}{\sqrt{4\pi s^2}}\exp\left(-\frac{(x+L/2)^2}{4s^2}\right)\right]. \tag{6.1}$$

The potential for $V_0 = -1$, $s = 1$, and $L = 6$ has two discrete eigenvalues $\Omega_0 = -0.1616$ and $\Omega_1 = -0.12$ and a continuous spectral part for $\Omega > 0$. The linear eigenstates can also be obtained and used to numerically compute the coefficients of the finite-dimensional decomposition of (3.12), (3.13) as $a_{0000} = -0.09397$, $a_{1111} = -0.10375$, $a_{0011} = a_{1010} = a_{1001} = a_{0110} = -0.08836$ (for $g = -1$). Then, using (3.10), we can compute the approximate threshold in $\mathcal{N}$ for bifurcation of an asymmetric branch (and the destabilization of the symmetric one):

$$\mathcal{N}_{cr} \sim \mathcal{N}_{cr}^{(0)} = 0.24331, \quad \Omega_{cr} \sim \Omega_{cr}^{(0)} \equiv \Omega_0 \ + \ a_{0000}\,\mathcal{N}_{cr}^{(0)} = -0.18447.$$

We expect good agreement because the values of s and L suggest the regime of large L, where our rigorous theory holds.

Using numerical fixed-point iterations (in particular Newton's method), we obtain the branches of the nonlinear eigenvalue problem (2.7). To study the stability of a solution, u_0, of (2.7), consider the evolution of a small perturbation of it:

$$u = e^{-i\Omega t}\left[u_0(x) + \left(p(x)e^{\lambda t} + q(x)e^{\bar{\lambda}t}\right)\right]. \tag{6.2}$$

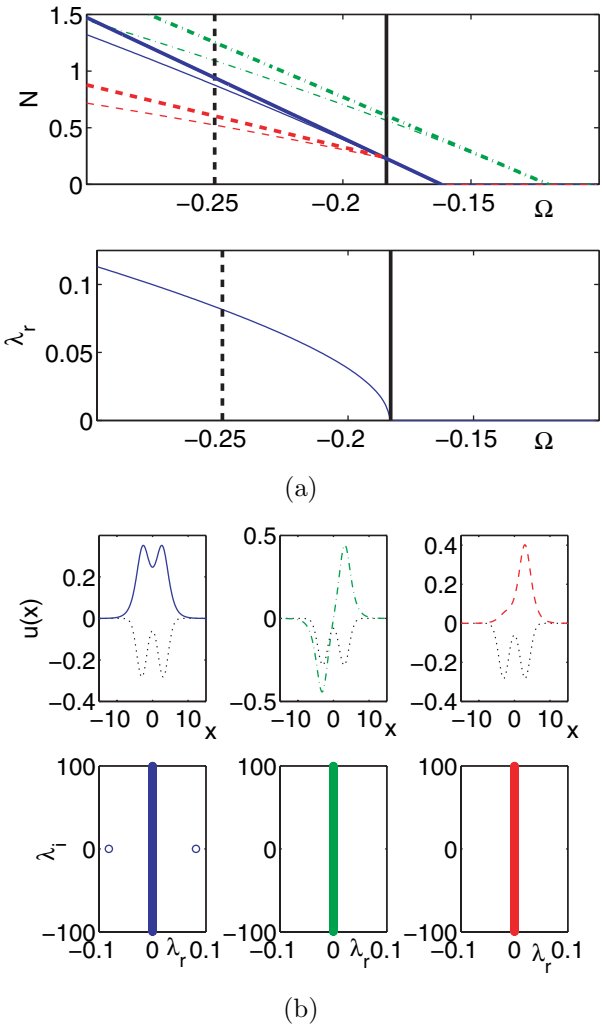


FIG. 3. *The figure shows the typical numerical bifurcation results for the cubic case and their comparison with the finite-dimensional analysis of section* 3. *Panel* (a) *shows the bifurcation diagram in the top subplot and the relevant real eigenvalues in the bottom subplot. In the top, the solid (blue) line represents the symmetric branch, the dashed-dotted (green) line represents the antisymmetric branch, while the dashed (red) line represents the bifurcating asymmetric branch. The thin lines indicate the numerical findings, while the thick ones show the corresponding finite-dimensional, weakly nonlinear predictions. The solid vertical (black) line indicates the critical point (of* $\Omega \approx -0.1835$*) obtained numerically. The dashed vertical (black) line is a guide to the eye for the case with* $\Omega = -0.25$*, whose detailed results are shown in panel* (b). *The bottom subplot of panel* (a) *shows the real eigenvalue (as a function of* Ω*) of the symmetric branch that becomes unstable for* $\Omega < -0.1835$. *Panel* (b) *shows, using the same symbolism as panel* (a)*, the symmetric (left), antisymmetric (middle), and asymmetric (right) branches and their linearization eigenvalues (bottom subplots) for* $\Omega = -0.25$. *The potential is shown by a dotted black line.*

Keeping only linear terms in p, q, we obtain a linear evolution equation, whose normal modes satisfy a linear eigenvalue problem with spectral parameter, which we denote by λ and eigenvector $(p(x), \bar{q}(x))^T$.

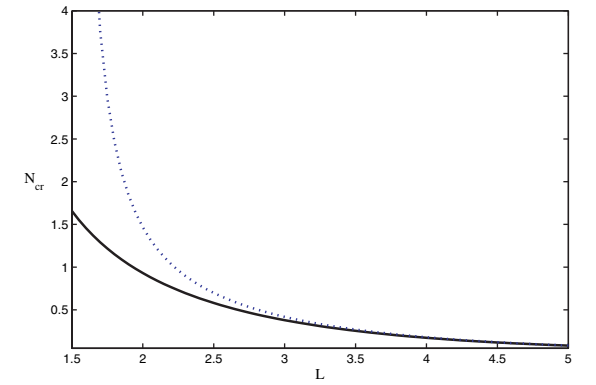
Our computations for the simplest case of the cubic nonlinearity with $K[\psi\bar{\psi}] = \psi\bar{\psi}$ are shown in Figure 3 (for $g(x) = -1$). In particular, the top subplot of panel (a) shows

the full numerical results by thin lines (solid for the symmetric solution, dashed for the bifurcating asymmetric, and dashed-dotted for the antisymmetric one) and compares them with the predictions based on the finite-dimensional truncation of (3.12)-(3.13) shown by the corresponding thick lines. The approximate threshold values $\mathcal{N}_{cr}$ and Ω_{cr} are found numerically to be $\Omega_{cr}^{(0)} \approx -0.1835$ and $\mathcal{N}_{cr}^{(0)} \approx 0.229$. This suggests a relative error in its evaluation by the finite-dimensional reduction of less than 1%. This critical point is indicated by a solid vertical line in panel (a). For $\Omega > \Omega_{cr}^{(0)}$, there exist two branches in the problem, namely the one that bifurcates from the symmetric linear state (this branch exists for $\Omega < \Omega_0$) and the one that bifurcates from the antisymmetric linear state (and, hence, exists for $\Omega < \Omega_1$). For $\Omega < \Omega_{cr}^{(0)}$, the symmetric branch becomes unstable due to a real eigenvalue (see bottom subplot of panel (a)), signalling the emergence of a new branch, namely the asymmetric one. All three branches are shown for $\Omega = -0.25$ (indicated by dashed vertical line in panel (a)) in panel (b) and their corresponding linearization spectrum (λ_r, λ_i) is shown for the eigenvalues $\lambda = \lambda_r + i\lambda_i$.

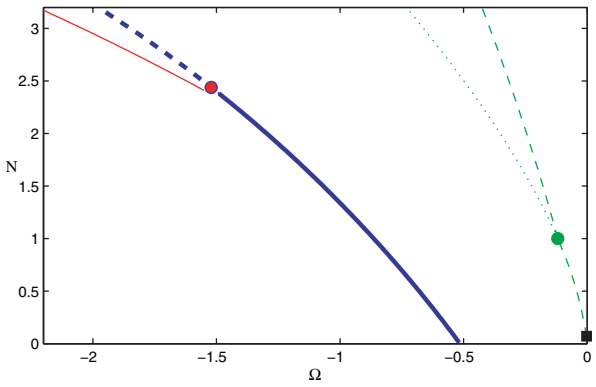
6.2. Symmetry-breaking threshold $\mathcal{N}_{cr}(L)$ as L varies. We now investigate the limits of validity of $\mathcal{N}_{cr}^{(0)}(L)$ as an approximation to $\mathcal{N}_{cr}(L)$ by varying the distance L between the potential wells (6.1). For L large, $\mathcal{N}_{cr}^{(0)}$, given by (3.10), is close to the actual $\mathcal{N}_{cr}(L)$, the exact threshold. In this case the eigenvalues of $-\partial_x^2 + V_L(x)$, $\Omega_0(L)$, and $\Omega_1(L)$ are close to each other; see Example 2.1. Therefore, the bifurcation occurs for small $\mathcal{N}$ and one is in the regime of validity of Theorem 4.1. In Figure 4 we display a comparison between the estimate for $\mathcal{N}_{cr}$ based on the finite-dimensional truncation, $\mathcal{N}_{cr}^{(0)}$, and the actual $\mathcal{N}_{cr}$. For large L the two values are close to each other. As L is decreased the wells approach one another and eventually, at $L = 0$, merge to form a single-well potential. As L is decreased, the eigenvalues of the linear bound states $\Omega_0(L)$ and $\Omega_1(L)$ move farther apart. For some value of L, L_d, the eigenvalue of the excited state, $\Omega_1(L)$, merges at $\Omega = 0$ into the continuous spectrum (and becomes a complex *scattering resonance*). For $L < L_d$ the estimate $\mathcal{N}_{cr}^{(0)}$ is not correct. In fact, $\mathcal{N}_{cr}^{(0)}(L) \to \infty$, while the actual value of $\mathcal{N}_{cr}(L)$ appears to remain finite. In Figure 4(a) we observe that for $L < 2$, $\mathcal{N}^{(0)}$ and $\mathcal{N}_{cr}$ diverge from one another and eventually the approximation $\mathcal{N}_{cr}^{(0)}(L)$ tends to infinity, while the actual $\mathcal{N}_{cr}(L)$ remains finite. In Figure 4(b) we show a bifurcation diagram for small L, in which a second discrete (excited state) eigenvalue of $-\partial_x^2 + V_L$, Ω_1, does not exist. Yet there exists a symmetry-breaking threshold, $\mathcal{N}_{cr}$, along the symmetric branch.

There are interesting observations to make with regard to antisymmetric solutions. Although there is no linear antisymmetric state, from which to bifurcate in the linear (zero amplitude) limit, we do observe a bifurcation of antisymmetric solutions. The black square in Figure 4(b) marks a *strictly positive* threshold value of the squared L^2 norm, $\mathcal{N}_{cr}^{excited} > 0$, at which an antisymmetric nonlinear bound state emerges from zero frequency. The linearization, L_+, about this *excited state branch* has two negative eigenvalues: with corresponding symmetric and antisymmetric eigenstates. The bifurcation along this antisymmetric branch at a larger value of $\mathcal{N}$ to an "asymmetric" state (see Figure 4(b)) is the result of a third eigenvalue of L_+ (with corresponding *even* parity eigenfunction) emerging from the continuous spectrum, as $\mathcal{N}$ is increased, and hitting zero at some critical value $\mathcal{N}_{cr}^{excited,asym} > \mathcal{N}_{cr}^{excited}$.

Finally, note that we expect, at least in the regime of weak nonlinearity, that branches of nonlinear bound states originating from nonground state eigenvalues to be unstable. There can be various mechanisms: linear [10, 15, 9, 16] as well as nonlinear.



(a)



(b)

FIG. 4. *The figure demonstrates the validity of* $\mathcal{N}_{cr}^{(0)}(L)$ *as an approximation to* $\mathcal{N}_{cr}(L)$. *Panel* (a) *compares the linear finite-dimensional estimation for the bifurcation point* $\mathcal{N}_{cr}^{(0)}(L)$ *and the actual numerical bifurcation point* $\mathcal{N}_{cr}$. *The computations are for the double-well potential* (6.1), $V_0 = -1$, *and* $s = 1/4$ *and cubic nonlinearity. The curve* $\mathcal{N}_{cr}(L)$ *is marked by a solid (black) line and the curve* $\mathcal{N}_{cr}^{(0)}(L)$ *is marked by a dotted (blue) line. Panel* (b) *shows a numerical bifurcation diagram for the double-well potential* (6.1), $V_0 = -1$, $s = 1/4$, *and* $L = 1.3$. *The bifurcation point* $\mathcal{N}_{cr}$ *is marked by a (red) circle. For* $\mathcal{N} < \mathcal{N}_{cr}$ *the ground state, marked by a thick (blue) solid line, is stable. For* $\mathcal{N} > \mathcal{N}_{cr}$ *the ground state is unstable and marked by a thick (blue) dashed line. The stable asymmetric state which appears for* $\mathcal{N} > \mathcal{N}_{cr}$ *is marked by a thin (red) solid line. The antisymmetric state* $\Omega_1^{\mathcal{N}}$ *is marked by a thin (light green) dashed line. The point* $\mathcal{N}$ *for which the antisymmetric state appears in the discrete spectrum is marked by a (black) square. Notice that in this bifurcation diagram there is also a bifurcation from the antisymmetric branch. The state which bifurcates from the antisymmetric state is marked by a (dark green) thin dotted line. See the text for remarks on the instability of these latter branches.*

Concerning the latter, if $2\Omega_1 - \Omega_0 > 0$, then it is known that the excited state is unstable due to resonant coupling to the radiation modes; see, for example, [29, 30]. Linearly this is manifested by an exponential instability, computed via perturbation theory of an embedded eigenvalue of the linearization about the excited state at the bifurcation point (zero amplitude) [5]. If $2\Omega_1 - \Omega_0 < 0$, which is the case in the example of Figure 1, the excited state is linearly stable for small amplitude, but it is nonlinearly unstable due to higher nonlinear order coupling to radiation modes [35].

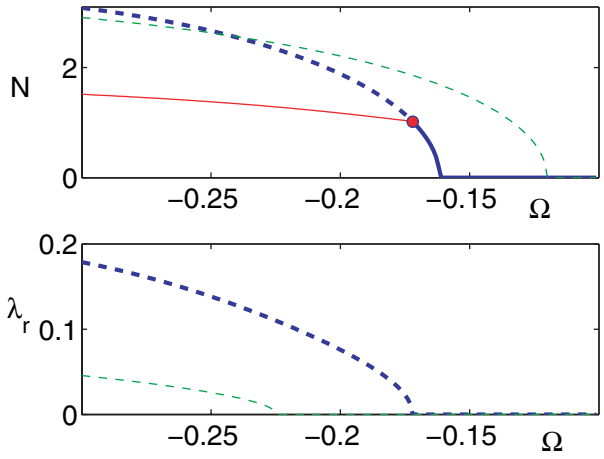


FIG. 5. *Same as panel* (a) *of Figure* 3 *but for the* quintic *nonlinearity. This serves to illustrate the analogies between the bifurcation pictures but also their differences (shifted critical point and also partial instability of the antisymmetric branch).*

6.3. More general nonlinearities. To simplify the analysis in this paper, we assumed a cubic nonlinearity in NLS-GP; see (H3). Our approach is quite general and more general nonlinearities can be treated. That is, a more general finite-dimensional Galerkin approximation can be derived and its normal form/symmetry/bifurcation theory can be developed.

In this subsection we present numerical computations for general power law nonlinearities such as $K[\psi\bar{\psi}] = (\psi\bar{\psi})^p$ and discuss phenomena analogous to the cubic case, $p = 1$.

Our numerical results for the case of $p = 2$, $K[\psi\bar{\psi}]\psi = |\psi|^4\psi$, are presented in Figure 5. The curves are analogous to those of panel (a) of Figure 3. The bifurcation diagram for this higher order nonlinearity is similar to that of the cubic case. However, the critical point for the emergence of the asymmetric branch is now shifted to $\Omega_{cr} \approx -0.1725$, closer to the linear limit. We have also examined the case of $p = 3$, $K[\psi\bar{\psi}]\psi = |\psi|^6\psi$, finding that the relevant critical point is further shifted toward the linear limit, $\Omega_{cr} = -0.168$. Using our methods one could identify, in the double-well case with large separation, $\Omega_{cr}(p; L)$ such that for all $\Omega < \Omega_{cr}(p; L)$, the symmetric branch is unstable. We also note in passing that bifurcation diagrams for higher values of p may also bear additional (to the shift in Ω_{cr}) differences from the cubic case; one such example in Figure 5 is given by the presence of a linear instability (due to a complex eigenvalue quartet emerging for $\Omega < -0.224$) for the antisymmetric branch. The latter was found to be linearly stable in the cubic case of Figure 3.

6.4. Nonlocal nonlinearities. Finally, we consider the case of nonlocal nonlinearities depending on a parameter ϵ, the range of the nonlocal interaction. In particular, consider the case of a nonlocal nonlinearity of the form

$$K[\psi\bar{\psi}] = \int_{-\infty}^{\infty} \mathcal{K}(x-y)\psi(y)\bar{\psi}(y)dy, \tag{6.3}$$

where

$$\mathcal{K}(x-y) = \frac{1}{2\pi\epsilon^2} e^{-\frac{(x-y)^2}{2\epsilon^2}}. \tag{6.4}$$

Here, $\epsilon > 0$ is a parameter controlling the range of the nonlocal interaction. As ϵ tends to 0, $\mathcal{K}(x-y) \to \delta(x-y)$ and we recover the "local" cubic limit. The form of the finite-dimensional reduction, of (3.12), (3.13), does not change; the only modification is that the coefficients a_{klmn} are now functions of the range of the interaction ϵ. The dependence of the coefficients a_{klmn} on ϵ is displayed in panel (a) of Figure 6. The solid (blue) line shows $|a_{0000}|$, the dashed (green) one corresponds to $|a_{1111}|$, the dashed-dotted (red) one corresponds to $|a_{1001}| = |a_{0110}|$ (g is constant, K even), while the thick solid (black) one corresponds to $|a_{0101}| = |a_{0011}| = |a_{1010}|$. Notice in the inset how the coefficients asymptote smoothly to their "local" limit. Additionally, note the expected asymptotic relation $a_{1001} = a_{0011}$. Also note the significant (decaying) dependence of the relevant coefficients on the range of the interaction. The nature of this dependence indicates that while the character of the bifurcation may be the same as in the case of local nonlinearities, its details (such as the location of the critical points) depend sensitively on the range of the nonlocal interaction. This is illustrated in panel (b) for the specific case of $\epsilon = 5$. In this panel (which is analogous to panel (a) of Figure 3, but for the nonlocal case) the critical point for emergence of the asymmetric branch/instability of the symmetric branch is shifted to $\Omega_{cr} = -0.2466$ (and the corresponding $\mathcal{N}_{cr} = 1.4353$) in comparison to the numerically obtained value of $\Omega_{cr} \approx -0.256$; the relative error in the identification of the critical point (by the finite-dimensional reduction) is in this case of the order of 3.7%. This can be attributed to the more nonlinear character (i.e., occurring for higher value of $\mathcal{N}_{cr}$) nature of the bifurcation. However, as the finite-dimensional approximation still yields a reliable estimate for the location of the critical point, in panel (c) we use it to obtain an approximation to the location of the critical point $(\Omega_{cr}, \mathcal{N}_{cr})$ as a function of the nonlocality parameter ϵ.

7. Concluding remarks. We have obtained rigorous results on the spontaneous symmetry-breaking bifurcation for a large class of NLS-GP equations and studied in detail the case of double-well potentials. Our analysis of the symmetry-breaking bifurcation and the exchange of stability is based on an expansion which, to leading order in amplitude, is a superposition of a symmetric-antisymmetric pair of eigenstates of the linear Hamiltonian, H, whose energies are separated (gap condition (4.13)) from all other spectra of H. This gap condition holds for double wells with sufficiently large L, but breaks down as L decreases. Nevertheless, numerical studies show the existence of a finite threshold for symmetry breaking; see the discussion above on the variation of $\mathcal{N}_{cr}(L)$ with L. A theory encompassing this phenomenon is of interest and is currently under investigation.

Finally, we remark that our analysis can be naturally extended to treat cases of general multi-wells, identical or not, since the methods involve a strategy for analysis of the weakly nonlinear regime, given spectral assumptions on the linear limit. An example is the case of a symmetric triple well, studied in [18], where the finite-dimensional Galerkin analysis has been implemented, revealing a rich bifurcation picture, but with no symmetry-breaking bifurcation in the symmetric branch. In such multi-well cases, naturally the dimension of the Galerkin approximation needs to be increased accordingly (e.g., 3-dimensional for 3-wells, 4-dimensional for 4-wells, etc.) introducing greater complexity into the global bifurcation structure. Formally, the derivation may be systematically extended to the case of infinitely many wells, constituting the so-called tight-binding approximation [2], although a rigorous derivation of such lattice equations may pose a considerable challenge.

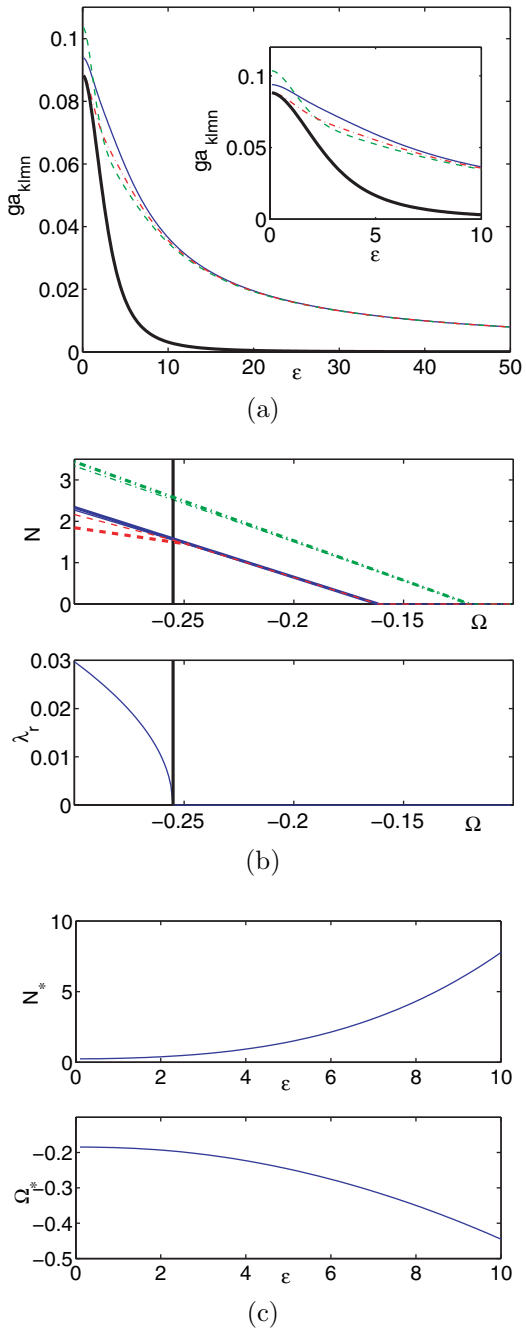


FIG. 6. *This figure shows the nonlocal analogue of Figure* 3. *Panel* (a) *shows the dependence of the (absolute value of the) coefficients of the finite-dimensional approximation on the nonlocality parameter* ϵ *(*$\epsilon = 0$ *denotes the "local" nonlinearity limit). The solid (blue) line denotes* a_{0000}*, the dashed (green) denotes* a_{1111}*, the dashed-dotted (red) denotes* a_{0110}*, while the thick solid (black) one denotes* a_{0101}*. Panel* (b) *is analogous to panel* (a) *of Figure* 3*, but now shown for the nonlocal case, with the nonlocality parameter* $\epsilon = 5$*. Finally, panel* (c) *shows the dependence of the critical point of the finite-dimensional bifurcation* $(N_\star, \Omega_\star)$ *on the nonlocality parameter* ϵ*.*

8. Appendix. Double wells. In this discussion, we are going to follow the analysis of [12]. Consider a (single-well) real-valued potential $v_0(x)$ on $\mathbb{R}^n$ such that $v_0(x) \in L^r + L^\infty_\varepsilon$ for all $1 \le r \le q$ where $q \ge \max(n/2, 2)$ for $n \ne 4$, $q > 2$ for $n = 4$. Then, multiplication by v_0 defines a compact operator from H^2 to L^2 and

$$H_0 = -\Delta + v_0(x)$$

is a self-adjoint operator on L^2 with domain H^2.

Consider now the double-well potential

$$V_L = T_L v_0 T_{-L} + R T_L v_0 T_{-L} R,$$

where T_L and R are the unitary operators

$$T_L g(x_1, x_2, \dots, x_n) = g(x_1 + L, x_2, \dots, x_n),$$
$$Rg(x_1, x_2, \dots, x_n) = g(-x_1, x_2, \dots, x_n),$$

and the self-adjoint operator

$$H_L = -\Delta + V_L(x).$$

PROPOSITION 8.1. *Assume that $\omega < 0$ is a nondegenerate eigenvalue of H_0 separated from the rest of the spectrum of H_0 by a distance greater than $2d_*$. Denote by ψ_ω its corresponding eigenvector, $\|\psi_\omega\|_{L^2} = 1$. Then there exists $L_0 > 0$ such that for $L \ge L_0$ the following are true.*

(i) *H_L has exactly two eigenvalues $\Omega_0(L)$ and $\Omega_1(L)$ nearer to ω than $2d_*$. Moreover, $\lim_{L\to\infty} \Omega_j(L) = \omega,\ j = 0, 1$.*

(ii) *One can choose the normalized eigenvectors $\psi_j(L),\ \|\psi_j(L)\|_{L^2} = 1$, corresponding to the eigenvalues $\Omega_j(L),\ j = 0, 1$, such that they satisfy*

$$\lim_{L\to\infty} \|\psi_j(L) - (T_L\psi_\omega + (-1)^j R T_L \psi_\omega)/\sqrt{2}\|_{H^2} = 0,\ j = 0, 1.$$

(iii) *If P^L_j are the orthogonal projections in L^2 onto $\psi_j(L),\ j = 0, 1$, and $\tilde{P}_L = Id - P^L_0 - P^L_1$, then there exists $d > 0$ independent of L such that*

$$\|(H_L - \Omega)^{-1}\tilde{P}_L\|_{L^2 \mapsto H^2} \ge d \quad \text{for all } L \ge L_0 \text{ and } |\Omega - \omega\| \le d_*.$$

Proof. For (i) we refer the reader to [12]. The L^2 convergence in (ii) has also been proved there. The H^2 convergence follows from the following compactness argument. Let

$$\psi^L_j = n_L \psi_j(L), \qquad j = 0, 1,$$

where n_L is such that $\|\psi^L_j\|_{H^2} = 1,\ j = 0, 1$. From the eigenvector equations: $(H_L - \Omega(L))\psi^L = 0$, where we dropped the index $j = 0, 1$ and the convergence $\Omega(L) \to \omega$, see part (i), we get

$$\lim_{L\to\infty} \|(-\Delta - \omega + V_L)\psi^L\|_{L^2} = 0. \tag{8.1}$$

Denote

$$g_L = (-\Delta - \omega)\psi^L \in L^2. \tag{8.2}$$

Since $-\Delta-\omega : H^2 \mapsto L^2$ is bounded, there exists a constant $C>0$ independent of L such that

$$\|g_L\|_{L^2} \leq C.$$

Since $\omega<0$, $-\Delta-\omega : H^2 \mapsto L^2$ has a continuous inverse, then (8.1) is equivalent to

$$g_L + V_L(-\Delta-\omega)^{-1}g_L \to 0 \text{ in } L^2.$$

By expanding V_L we get

$$g_L + T_L v_0(-\Delta-\omega)^{-1}T_{-L}g_L + RT_L v_0(-\Delta-\omega)^{-1}T_{-L}Rg_L \to 0. \tag{8.3}$$

But $v_0(-\Delta-\omega)^{-1} : L^2 \mapsto L^2$ is compact while the translation and reflection operators are unitary. These and the uniform boundedness of g_L lead to the existence of $\psi \in L^2$ and $\tilde{\psi} \in L^2$ and a subsequence of g_L, which we will redenote by g_L, such that

$$\lim_{L\to\infty} \|v_0(-\Delta-\omega)^{-1}T_{-L}g_L - \psi\|_{L^2} = 0 \text{ and } \lim_{L\to\infty} \|v_0(-\Delta-\omega)^{-1}T_{-L}Rg_L - \tilde{\psi}\|_{L^2} = 0. \tag{8.4}$$

By plugging in (8.3) and multiplying to the left by T_{-L} we get

$$\lim_{L\to\infty} \|T_{-L}g_L + \psi + RT_{2L}\tilde{\psi}\|_{L^2} = 0.$$

But $RT_{2L}\tilde{\psi}$ converges weakly to zero, hence $T_{-L}g_L$ converges weakly to $-\psi$. By plugging now in (8.4) and using compactness we get

$$\psi + v_0(-\Delta-\omega)^{-1}\psi = 0.$$

The latter shows that $(-\Delta-\omega)^{-1}\psi$ is an eigenvector of $-\Delta+v_0$ corresponding to the eigenvalue ω. By nondegeneracy of ω we get

$$\psi = -n(-\Delta-\omega)\psi_\omega, \tag{8.5}$$

where n is a constant. A similar argument shows that

$$\tilde{\psi} = -\tilde{n}(-\Delta-\omega)\psi_\omega, \tag{8.6}$$

where $\tilde{n}$ is a constant.

Combining (8.1)–(8.6) we get

$$\lim_{L\to\infty} \|(-\Delta-\omega)(\psi^L - nT_L\psi_\omega - \tilde{n}RT_L\psi_\omega)\|_{L^2} = 0 \tag{8.7}$$

which by the continuity of $(-\Delta-\omega)^{-1} : L^2 \mapsto H^2$ implies

$$\lim_{L\to\infty} \|\psi^L - nT_L\psi_\omega - \tilde{n}RT_L\psi_\omega\|_{H^2} = 0.$$

Using now that $\|\psi^L\|_{H^2} = 1$ and that the rescaled ψ_j^L, such that it has norm 1 in L^2, converges to $(T_L\psi_\omega + (-1)^j RT_L\psi_\omega)/\sqrt{2}$, we get the conclusion of part (ii) for a subsequence first, then, by uniqueness of the limit, for all $L\to\infty$.

For part (iii), it suffices to show that there are no sequences $(\Omega_L, \psi^L) \in [\omega - d_*, \omega + d_*] \times H^2$ with $\|\psi^L\|_{H^2} = 1$ and $\psi_L \perp \psi_j(L),\ j=0,1$, in L^2 such that

$$\lim_{L\to\infty} \|(H_L - \Omega_L)\psi^L\|_{L^2} = 0. \tag{8.8}$$

The spectral estimate

$$\|(H_L - \Omega_L)\psi^L\|_{L^2} \geq dist(\Omega_L, \sigma(H_L)\backslash\{\Omega_0(L), \Omega_1(L)\})\|\psi^L\|_{L^2} \geq d_*\|\psi^L\|_{L^2},$$

combined with (8.8), implies

$$\lim_{L\to\infty} \|\psi^L\|_{L^2} = 0. \tag{8.9}$$

In principle, we can now employ the compactness argument in part (ii) to get

$$\lim_{L\to\infty} \|\psi^L\|_{H^2} = 0 \tag{8.10}$$

which will contradict $\|\psi^L\|_{H^2} = 1$. More precisely, (8.8), (8.9) imply

$$\lim_{L\to\infty} \|(-\Delta - \omega - d_* + V_L)\psi^L\|_{L^2} = 0$$

which, by repeating the argument after (8.1) with ω replaced by $\omega + d_*$, gives

$$\lim_{L\to\infty} \|\psi^L + T_L\psi_{\omega+d_*} + RT_L\tilde{\psi}_{\omega+d_*}\|_{H^2} = 0,$$

where $\psi_{\omega+d_*}$ and $\tilde{\psi}_{\omega+d_*}$ are eigenvectors of $-\Delta + v_0$ corresponding to eigenvalue $\omega + d_*$. But the latter is not actually an eigenvalue, hence $\psi_{\omega+d_*} = 0$ and $\tilde{\psi}_{\omega+d_*} = 0$. These show (8.10) and finishes the proof of part (iii).

The proposition is now completely proven. ∎

Acknowledgments. The authors acknowledge the support of the US National Science Foundation, Division of Mathematical Sciences (DMS). EK acknowledges valuable discussions with A. Malkin, B. Sandstede, and E. Lerman. PGK acknowledges valuable discussions with T. Kapitula and Z. Chen. Finally, the authors thank the referees for their careful reading and comments on the submitted manuscript.

REFERENCES

[1] M. Albiez, R. Gati, J. Fölling, S. Hunsmann, M. Cristiani, and M. K. Oberthaler, *Direct observation of tunneling and nonlinear self-trapping in a single bosonic Josephson junction*, Phys. Rev. Lett., 95 (2005), p. 010402.

[2] G. L. Alfimov, P. G. Kevrekidis, V. V. Konotop, and M. Salerno, *Wannier functions analysis of the nonlinear Schrödinger equation with a periodic potential*, Phys. Rev. E, 66 (2002), p. 046608.

[3] W. H. Aschbacher, J. Fröhlich, G. M. Graf, K. Schnee, and M. Troyer, *Symmetry breaking regime in the nonlinear Hartree equation*, J. Math. Phys., 43 (2002), pp. 3879–3891.

[4] C. Cambournac, T. Sylvestre, H. Maillotte, B. Vanderlinden, P. Kockaert, Ph. Emplit, and M. Haelterman, *Symmetry-breaking instability of multimode vector solitons*, Phys. Rev. Lett., 89 (2002), p. 083901.

[5] S. Cuccagna, D. Pelinovsky, and V. Vougalter, *Spectra of positive and negative energies in the linearized NLS problem*, Comm. Pure Appl, Math., 58 (2005), pp. 1–29.

[6] B. Deconinck and J. Nathan Kutz, *Singular instability of exact stationary solutions of the nonlocal Gross-Pitaevskii equation*, Phys. Lett. A, 319 (2003), pp. 97–103.

[7] J. Fleischer, G. Bartal, O. Cohen, T. Schwartz, O. Manela, B. Freedman, M. Segev, H. Buljan, and N. Efremidis, *Spatial photonics in nonlinear waveguide arrays*, Optics Express, 13 (2005), pp. 1780–1796.

[8] M. Golubitsky and D. G. Schaeffer, *Singularities and Groups in Bifurcation Theory*, Vol. 1–2, Appl. Math. Sci. 51, Springer, New York, 1985.

[9] M. G. Grillakis, *Linearized instability for nonlinear Schrödinger and Klein-Gordon equations*, Comm. Pure Appl. Math., 41 (1988), pp. 747–774.

[10] M. G. Grillakis, J. Shatah, and W. A. Strauss, *Stability theory of solitary waves in the presence of symmetry* I, J. Funct. Anal., 74 (1987), pp. 160–197.
[11] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcation of Vector Fields*, Appl. Math. Sci. 42, Springer, New York, 1983.
[12] E. M. Harrell, *Double wells*, Comm. Math. Phys., 75 (1980), pp, 239–261.
[13] T. Heil, I. Fischer, and W. Elsässer, *Chaos synchronization and spontaneous symmetry-breaking in symmetrically delay-coupled semiconductor lasers*, Phys. Rev. Lett., 86 (2001), pp. 795–798.
[14] R. K. Jackson and M. I. Weinstein, *Geometric analysis of bifurcation and symmetry breaking in a Gross-Pitaevskii equation*, J. Statist. Phys., 116 (2004), pp. 881–905.
[15] C. K. R. T. Jones, *An instability mechanism for radially symmetric standing waves of a nonlinear Schrödinger equation*, J. Differential Equations, 71 (1988), pp. 34–62.
[16] T. Kapitula, *Stability of waves in perturbed Hamiltonian systems*, Phys. D, 156 (2001), pp. 186–200.
[17] T. Kapitula and P. Kevrekidis, *Bose-Einstein condensates in the presence of a magnetic trap and optical lattice: Two-mode approximation*, Nonlinearity, 18 (2005), pp. 2491–2512.
[18] T. Kapitula, P. G. Kevrekidis, and Z. Chen, *Three is a crowd: Solitary waves in photorefractive media with three potential wells*, SIAM J. Appl. Dyn. Syst., 5 (2006), pp. 598–633.
[19] T. Kapitula, P. G. Kevrekidis, and B. Sandstede, *Counting eigenvalues via the Krein signature in infinite-dimensional Hamiltonian systems*, Phys. D, 195 (2004), pp. 263–282.
[20] P. G. Kevrekidis, Z. Chen, B. A. Malomed, D. J. Frantzeskakis, and M. I. Weinstein, *Spontaneous symmetry breaking in photonic lattices: Theory and experiment*, Phys. Lett. A, 340 (2005), pp. 275–280.
[21] E. Kirr and A. Zarnescu, *On the asymptotic stability of bound states in* $2D$ *cubic Schrödinger equation*, Comm. Math. Phys., 272 (2007), pp. 443–468.
[22] W. Krolikowski, O. Bang, N. I. Nikolov, D. Neshev, J. Wyller, J. J. Rasmussen, and D. Edmundson, *Modulational instability, solitons and beam propagation in spatially nonlocal nonlinear media*, J. Opt. B Quantum Semiclass. Opt., 6 (2004), pp. S288–S294.
[23] K. W. Mahmud, J. N. Kutz, and W. P. Reinhardt, *Bose-Einstein condensates in a one-dimensional double square well: Analytical solutions of nonlinear Schrödinger equation*, Phys. Rev. A, 66 (2002), p. 063607.
[24] L. Nirenberg, *Lectures on Nonlinear Functional Analysis*, Courant Institute, New York, 1974.
[25] C.-A. Pillet and C. E. Wayne, *Invariant manifolds for a class of dispersive, Hamiltonian partial differential equations*, J. Differential Equations, 141 (1997), pp. 310–326.
[26] H. A. Rose and M. I. Weinstein, *On the bound states of the nonlinear Schrödinger equation with a linear potential*, Phys. D, 30 (1988), pp. 207–218.
[27] A. Sacchetti, *Nonlinear time-dependent one-dimensional Schrödinger equation with double-well potential*, SIAM J. Math. Anal., 35 (2003), pp. 1160–1176.
[28] S. Sawai, Y. Maeda, and Y. Sawada, *Spontaneous symmetry breaking turing-type pattern formation in a confined dictyostelium cell mass*, Phys. Rev. Lett., 85 (2000), pp. 2212–2215.
[29] A. Soffer and M. I. Weinstein, *Selection of the ground state for nonlinear Schrödinger equations*, Rev. Math. Phys., 16 (2004), pp. 977–1071.
[30] A. Soffer and M. I. Weinstein, *Theory of nonlinear dispersive waves and selection of the ground state*, Phys. Rev. Lett., 95 (2005), p. 213905.
[31] A. G. Vanakaras, D. J. Fotinos, and E. T. Samulski, *Tilt, polarity, and spontaneous symmetry breaking in liquid crystals*, Phys. Rev. E, 57 (1998), pp. R4875–R4878.
[32] M. I. Weinstein, *Modulational stability of ground states of nonlinear Schrödinger equations*, SIAM J. Math. Anal., 16 (1985), pp. 472–491.
[33] M. I. Weinstein, *Lyapunov stability of ground states of nonlinear dispersive evolution equations*, Comm. Pure Appl. Math., 39 (1986), pp. 51–68.
[34] C. Yannouleas and U. Landman, *Spontaneous symmetry breaking in single and molecular quantum dots*, Phys. Rev. Lett., 82 (1999), pp. 5325–5328.
[35] G. Zhou, *Perturbation expansion and Nth order Fermi golden rule of the nonlinear Schrödinger equations with potential*, J. Math. Phys.., 48 (2007), p. 053509.

SIAM J. MATH. ANAL.
Vol. 40, No. 2, pp. 605–622

ON 2×2 CONSERVATION LAWS AT A JUNCTION*

R. M. COLOMBO†, M. HERTY‡, AND V. SACHERS§

Abstract. This paper deals with 2×2 conservation laws at a junction. For the Cauchy problem, existence, uniqueness, and Lipschitz continuous dependence of the solution from the initial data as well as from the conditions at the junction are proved. The present construction comprehends the case of the p-system used to describe gas flow in networks and hereby unifies different approaches present in the literature. Furthermore, different models for water networks are considered.

Key words. hyperbolic systems of conservation laws, p-system, St. Venant equations

AMS subject classifications. 35L65, 76N10, 34B45

DOI. 10.1137/070690298

1. Introduction.

This paper studies the initial value problem consisting of

$$\partial_t u_l + \partial_{x_l} f(u_l) = 0, \qquad \text{with } l = 1, \dots, n,\ t \in [0, +\infty[\ ,\ x \in [0, +\infty[\ , \tag{1.1}$$

along n pipes together with

$$\Psi\left(u_1(t, 0+), u_2(t, 0+), \dots, u_n(t, 0+)\right) = 0 \tag{1.2}$$

at a junction. In other words, we deal with n initial boundary value problems for systems of 2×2 conservation laws, coupled through nonlinear boundary conditions. In this general setting, extending the results in [10], we derive conditions under which the Cauchy problem for (1.1) has a unique solution. Furthermore, the Lipschitz continuous dependence of the solution on initial data and coupling conditions is proved.

We model a junction connecting n rectilinear pipes by n pairwise distinct vectors $\nu_1, \dots, \nu_n$ parallel to the pipes and such that $\|\nu_l\|$ equals the cross section of the lth pipe. Furthermore, we assign a space coordinate $x_l > 0$ to each pipe. The transport of a specific quantity in the lth pipe is given by (1.1), where u_l is the vector of variables along the lth duct and f is a general nonlinear flux function. Condition (1.2) describes the interaction of the transported quantities at the intersection of the pipes; see [2, 3, 9, 10, 11, 19]. The standard situation of the Cauchy problem on a line is recovered in the case $n = 2$, $\nu_1 + \nu_2 = 0$, and $\Psi(u_1, u_2) = f(u_1) - f(u_2)$; see section 4.1. To simplify the notation, below we denote by x all the coordinates x_l along the various pipes.

Applications of the theoretical results are in the field of fluid flow in networks and in particular in high-pressure gas pipelines in open canals. In recent years, there has been intense research in flow problems on networks; see, e.g., the book on gas networks [23] and the publications of the Pipeline Simulation Interest Group [26]. Most of the proposed models [12, 14, 21, 23, 24, 25, 27, 22, 18] consider each pipe as a

*Received by the editors April 28, 2007; accepted for publication (in revised form) February 14, 2008; published electronically June 6, 2008. This work was supported by DFG Program 1253 and DAAD Vigoni project D/06/19582.
http://www.siam.org/journals/sima/40-2/69029.html
†Department of Mathematics, Brescia University, 25133 Brescia, Italy (rinaldo@ing.unibs.it).
‡Mathematik, RWTH Aachen, 52056 Aachen, Germany (herty@mathc.rwth-aachen.de).
§Department of Mathematics, TU Kaiserslautern, 67653 Kaiserslautern, Germany (sachers@mathematik.uni-kl.de).

one-dimensional domain and use balance laws to describe the dynamics. The validity of one-dimensional models is the subject of intense discussions; we refer the reader to [26] for more details.

In this context, the most challenging and interesting point is the coupling condition at pipe-to-pipe intersections. In an engineering context, this coupling is typically modeled through tables prescribing suitable relations depending on, e.g., the geometry of the pipe, the material, and flow conditions [12, 21]. Among the first mathematical treatments of this situation are [2, 3, 9, 10, 19]. The current presentation considers the *subsonic* case as in [2, 3, 9, 10, 11, 19] and typical physical conditions [14]. Extending the presentations in [4, 16, 20], we consider solutions possibly containing shocks.

Our purpose is to present a general framework for coupling conditions and prove well posedness. Indeed, we unify and extend the approaches [2, 3, 9, 10, 16, 19]. First, the present framework includes the currently used one-dimensional isothermal model for gas flow as well as the shallow-water equations for flows in open channels. In particular, we extend the results in [10] covering not only the isentropic Euler equations but a general 2×2 system. More importantly, the present result allows us to consider any coupling conditions specified through any (possibly nonlinear) function Ψ, the sole constraint being condition (2.2) below. Hence, the present results also extend previous works on open channel flow with gate control or pumps (see [16]), as well as the model for a kink in a pipe introduced in [19]. Finally, within this general setting, we also prove the Lipschitz continuous dependence of the solutions from the condition Ψ at the junction; see (3.2) in Theorem 3.2. As a consequence, in all the cases mentioned above, it is possible to prove the existence of an optimal control.

Numerically, we show that different one-dimensional coupling conditions lead to qualitatively different solutions. The comparison is in the context of the present theory, i.e., for one-dimensional models. Two-dimensional situations are considered, for instance, in [17, 18]. For a comparison with results of the engineering community we refer the reader to the pressure loss tables [12, 21].

The paper is organized as follows. Section 2 is devoted to the Riemann problem and extends [9]. In section 3, the well posedness of the Cauchy problem is stated. Applications of this result to gas flow in pipes as well as flow in open canals are collected in section 4. Section 5 contains the detailed constructions and proofs.

2. The Riemann problem at a junction. Throughout, we refer the reader to [6] for the general theory of hyperbolic systems of conservation laws. Let $\Omega \subseteq \mathbb{R}^2$ be a nonempty open set. Fix a flow $f \in \mathbf{C^4}(\Omega; \mathbb{R}^2)$ satisfying the following assumption:

(F) There exists a $\bar{u} \in \Omega$ such that $Df(\bar{u})$ admits a strictly negative eigenvalue $\lambda_1(\bar{u})$ and a strictly positive one $\lambda_2(\bar{u})$, the corresponding eigenvectors are linearly independent, and each characteristic field is either genuinely nonlinear or linearly degenerate.

Under this condition, (1.1) generates a standard Riemann semigroup; see [6, Chapter 8]. By *Riemann problem at the junction* we mean the problem

$$\begin{cases} \partial_t u_l + \partial_x f(u_l) = 0, & t \in \mathbb{R}^+, \quad l \in \{1, \dots, n\}, \\ u_l(0, x) = \bar{u}_l\,, & x \in \mathbb{R}^+, \quad u_l \in \Omega, \end{cases} \tag{2.1}$$

where $\bar{u}_1, \dots, \bar{u}_n$ are constant states in Ω. For $l = 1, \dots, n$, u_l has two components; i.e., $u_l = (u_{l,1}, u_{l,2})$ denotes the densities of the conserved quantities in the lth tube. Here and in what follows, $\mathbb{R}^+ = [0, +\infty[$.

DEFINITION 2.1. *Fix a map $\Psi \in \mathbf{C^1}(\Omega^n; \mathbb{R}^n)$. A* Ψ-solution *to the Riemann problem* (2.1) *is a function $u\colon \mathbb{R}^+ \times \mathbb{R}^+ \mapsto \Omega^n$ such that the following hold:*

(L) *For $l = 1, \dots, n$, the function $(t, x) \mapsto u_l(t, x)$ is self-similar and coincides with the restriction to $x \in \mathring{\mathbb{R}}^+$ of the Lax solution to the standard Riemann problem*

$$\begin{cases} \partial_t u_l + \partial_x f(u_l) = 0, \\ u_l(0, x) = \begin{cases} \bar{u}_l & \text{if } x > 0, \\ u_l(1, 0+) & \text{if } x < 0\,. \end{cases} \end{cases}$$

(Ψ) *The trace $u(t, 0+)$ of u at the junction satisfies* (1.2) *for a.e. $t > 0$.*

Given an entropy-entropy flux pair (E, F), the Ψ-solution is entropic at the junction if the following holds:

(E) *At the junction, entropy may not decrease; i.e., for a.e. $t > 0$*

$$\sum_{l=1}^{n} \|\nu_l\|\, F\left(u_l(t, 0+)\right) \leq 0\,.$$

For later use, with a slight abuse of notation, we denote

$$\begin{aligned} \|u\| &= \sum_{l=1}^{n} \|u_l\| && \text{for } u \in \Omega^n\,, \\ \|u\|_{\mathbf{L^1}} &= \int_{\mathbb{R}^+} \left\|u(x)\right\| dx && \text{for } u \in \mathbf{L^1}\left(\mathbb{R}^+; \Omega^n\right)\,, \\ \mathrm{TV}\,(u) &= \sum_{l=1}^{n} \mathrm{TV}\,(u_l) && \text{for } u \in \mathbf{BV}\left(\mathbb{R}^+; \Omega^n\right)\,. \end{aligned}$$

The following proposition yields the well posedness of the Riemann problem and the continuous dependence of the solution to the Riemann problem from the initial state and from the function Ψ. These results are used in section 3 to prove well posedness of the Cauchy problem by the wave tracking algorithm. The proofs are deferred to section 5.

PROPOSITION 2.2. *Let $n \in \mathbb{N}$ with $n \geq 2$. Fix the pairwise distinct vectors $\nu_1, \dots, \nu_n$ in $\mathbb{R}^3 \setminus \{0\}$ and an n-tuple of constant states $\hat{u} \in \Omega^n$ giving a stationary solution to the Riemann problem* (2.1) *in the sense of Definition* 2.1. *Assume that for $l = 1, \dots, n$,* **(F)** *holds in $\hat{u}_l$. If $\Psi \in \mathbf{C^1}(\Omega^n; \mathbb{R}^n)$ satisfies*

$$\det \left[\; D_1\Psi(\hat{u}) r_2(\hat{u}_1) \quad D_2\Psi(\hat{u}) r_2(\hat{u}_2) \quad \dots \quad D_n\Psi(\hat{u}) r_2(\hat{u}_n) \;\right] \neq 0, \tag{2.2}$$

where $D_l\Psi = D_{u_l}\Psi$, then there exist positive δ, K such that the following hold:

1. *For all $\bar{u} \in \Omega^n$ satisfying $\|\bar{u} - \hat{u}\| < \delta$, the Riemann problem* (2.1) *admits a unique self-similar solution $(t, x) \mapsto \left(\mathcal{R}^\Psi(\bar{u})\right)(t, x)$ in the sense of Definition* 2.1.
2. *If* (1.1) *admits an entropy-entropy flux pair (E, F), requiring that*

$$\sum_{l=1}^{n} \|\nu_l\|\, F(\hat{u}_l) < 0 \tag{2.3}$$

ensures that the solution $(t, x) \mapsto \left(\mathcal{R}^\Psi(\bar{u})\right)(t, x)$ is also entropic.

3. *If $\bar u, \bar w \in \Omega^n$ both satisfy $\|\bar u - \hat u\| < \delta$ and $\|\bar w - \hat u\| < \delta$, then the traces at the junction of the corresponding solutions to* (2.1) *satisfy*

$$\left\| \left((\mathcal{R}^\Psi(\bar u)) \right)(t, 0+) - \left((\mathcal{R}^\Psi(\bar w)) \right)(t, 0+) \right\| \leq K \cdot \|\bar u - \bar w\| \,. \tag{2.4}$$

4. *For any $\tilde\Psi \in \mathbf{C^1}(\Omega^n; \mathbb{R}^n)$ with $\|\tilde\Psi - \Psi\|_{\mathbf{C^1}} < \delta$, $\tilde\Psi$ also satisfies* (2.2) *and for all $\bar u \in \Omega^n$ satisfying $\|\bar u - \hat u\| < \delta$,*

$$\left\| \left(\mathcal{R}^{\tilde\Psi}(\bar u) \right)(t) - \left(\mathcal{R}^\Psi(\bar u) \right)(t) \right\|_{\mathbf{L^1}} \leq K \cdot \left\| \tilde\Psi - \Psi \right\|_{\mathbf{C^1}} \cdot t \,.$$

In the previous proposition, $r_2(\hat u)$ is the right eigenvector of $Df(\hat u)$ corresponding to the second characteristic field.

We remark that, for subsonic initial data, we obtain here a unique Ψ-solution, without any additional condition, such as **(E)**.

3. The Cauchy problem at an intersection. Next we consider the Cauchy problem at a junction. First, we give a definition of a solution which naturally extends Definition 2.1 to the Cauchy problem. Then we prove the existence of solutions for initial data with small total variation and the continuous dependence on the coupling condition Ψ.

DEFINITION 3.1. *Fix $\hat u \in \Omega^n$ and $T \in]0, +\infty]$. A weak Ψ-solution to*

$$\begin{cases} \partial_t u_l + \partial_x f(u_l) = 0, & t \in \mathbb{R}^+, & l \in \{1, \dots, n\}, \\ u(0, x) = u_o(x), & x \in \mathbb{R}^+, & u_o \in \hat u + \mathbf{L^1}(\mathbb{R}^+; \Omega^n) \end{cases} \tag{3.1}$$

on $[0, T]$ is a map $u \in \mathbf{C^0}\left([0, T]; \hat u + \mathbf{L^1}(\mathbb{R}^+; \Omega^n)\right)$ such that the following hold:

(W) *For all $\varphi \in \mathbf{C^\infty_c}\left(]-\infty, T[\times \mathbb{R}^+; \mathbb{R}\right)$ and for $l = 1, \dots, n$*

$$\int_0^T \int_{\mathbb{R}^+} \left(u_l \, \partial_t \varphi + f(u_l) \, \partial_x \varphi \right) dx \, dt + \int_{\mathbb{R}^+} u_{o,l}(x) \, \varphi(0, x) \, dx = 0 \,.$$

($\mathbf{\Psi}$) *The condition at the junction is met: for a.e. $t \in \mathbb{R}^+$, $\Psi\left(u(t, 0+)\right) = 0$.*

If (1.1) *admits an entropy-entropy flux pair (E, F), then the weak Ψ-solution u is entropic if for all $\varphi \in \mathbf{C^\infty_c}\left(]-\infty, T[\times \mathbb{R}^+; \mathbb{R}^+\right)$*

$$\sum_{l=1}^n \left(\int_0^T \int_{\mathbb{R}^+} \left(E(u_l) \, \partial_t \varphi + F(u_l) \, \partial_x \varphi \right) dx \, dt + \int_{\mathbb{R}^+} E(u_o) \, \varphi(0, x) \, dx \right) \|\nu_l\| \geq 0.$$

We are now ready to state the main result of this paper, namely the well posedness of the Cauchy problem for (3.1) at the junction.

THEOREM 3.2. *Let $n \in \mathbb{N}$, $n \geq 2$. Fix the pairwise distinct vectors $\nu_1, \dots, \nu_n$ in $\mathbb{R}^3 \setminus \{0\}$. Fix an n-tuple of states $\bar u \in \Omega^n$ such that f satisfies* **(F)** *at $\bar u$ and the Riemann problem* (2.1) *with initial datum $\bar u$ admits the stationary solution in the sense of Definition* 2.1. *Let $\Psi \in \mathbf{C^1}(\Omega^n; \mathbb{R}^n)$ satisfy* (2.2). *Then there exist positive δ, L and a map $S\colon [0, +\infty[\times \mathcal{D} \to \mathcal{D}$ such that*

1. *$\mathcal{D} \supseteq \left\{ u \in \bar u + \mathbf{L^1}(\mathbb{R}^+; \Omega^n) \colon \mathrm{TV}(u) \leq \delta \right\}$;*
2. *for $u \in \mathcal{D}$, $S_0 u = u$ and for $s, t \geq 0$, $S_s S_t u = S_{s+t} u$;*
3. *for $u, w \in \mathcal{D}$ and $s, t \geq 0$, $\|S_t u - S_s w\|_{\mathbf{L^1}} \leq L \cdot \left(\|u - w\|_{\mathbf{L^1}} + |t - s| \right)$;*

4. *if $u \in \mathcal{D}$ is piecewise constant, then for $t > 0$ sufficiently small, $S_t u$ coincides with the juxtaposition of the solutions to Riemann problems centered at the points of jumps or at the junction.*

Moreover, for every $u \in \mathcal{D}$, the map $t \mapsto S_t u$ is a Ψ-solution to the Cauchy problem (3.1) *according to Definition* 3.1.

For any $\tilde{\Psi} \in \mathbf{C^1}(\Omega^n; \mathbb{R}^n)$ with $\|\tilde{\Psi} - \Psi\|_{\mathbf{C^1}} < \delta$, $\tilde{\Psi}$ generates a semigroup of solutions on $\mathcal{D}$ and for $u \in \mathcal{D}$,

$$\left\| S_t^{\tilde{\Psi}} \bar{u} - S_t^{\Psi} \bar{u} \right\|_{\mathbf{L^1}} \leq L \cdot \left\| \tilde{\Psi} - \Psi \right\|_{\mathbf{C^1}} \cdot t \,. \tag{3.2}$$

If (1.1) *admits an entropy-entropy flux pair (E, F) and $\bar{u}$ is strictly entropic in the sense of* (2.3)*, then the Ψ-solution $t \mapsto S_t u$ is entropic at the junction.*

The proof is deferred to section 5.

4. Gas networks and open channels. A widely used model for gas flow in pipe networks is the system of isothermal Euler equations; see [23] and the references therein. In this section, we discuss the coupling conditions [2, 10] in the context of the presented theory in the case of a general p-system. By p-system we mean

$$\left\{ \begin{array}{l} \partial_t \rho_l + \partial_x q_l = 0, \\ \partial_t q_l + \partial_x \left(\dfrac{q_l^2}{\rho_l} + p(\rho_l) \right) = 0, \end{array} \right. \qquad \begin{array}{rcl} t & \in & \mathbb{R}^+, \\ x & \in & \mathbb{R}^+, \\ l & \in & \{1, \dots, n\}, \\ (\rho_l, q_l) & \in & \mathring{\mathbb{R}}^+ \times \mathbb{R}, \end{array} \tag{4.1}$$

where $\rho > 0$ is the mass density of a given fluid, q its linear momentum density, and $p = p(\rho)$ the pressure law, which we assume to satisfy the following:

(P) $p \in \mathbf{C^2}\left(\mathbb{R}^+; \mathbb{R}^+\right)$, $p(0) = 0$ and for all $\rho \in \mathbb{R}^+$, $p'(\rho) > 0$, $p''(\rho) \geq 0$.

In the context of gas pipelines, the pressure law typically chosen is $p(\rho) = a^2 \rho$, where the sound speed a depends on the gas type and temperature [23].

As is well known, (4.1) is equipped with the (mathematical) entropy-entropy flux pair

$$\begin{array}{rcll} E(\rho, q) & = & \dfrac{q^2}{2\rho} + \rho \displaystyle\int_{\rho_*}^{\rho} \frac{p(r)}{r^2}\, dr & \text{(total energy)}, \\ F(\rho, q) & = & \dfrac{q}{\rho} \cdot \big(E(\rho, q) + p(\rho) \big) & \text{(flow of the total energy)} \end{array}$$

for a $\rho_* > 0$. Choosing an initial datum $\bar{u} = (\bar{\rho}, \bar{q})$ in the *subsonic* region

$$\Omega = \Big\{ (\rho, q) \in \mathring{\mathbb{R}}^+ \times \mathbb{R} \colon \lambda_1(\rho, q) < 0 < \lambda_2(\rho, q) \Big\}$$

ensures that **(F)** holds at $\bar{u}$. Recall the standard relations

$$\begin{array}{rclcrcl} \lambda_1(\rho, q) & = & (q/\rho) - \sqrt{p'(\rho)}\,, & \qquad & \lambda_2(\rho, q) & = & (q/\rho) + \sqrt{p'(\rho)}, \\ r_1(\rho, q) & = & \begin{bmatrix} -1 \\ -\lambda_1(\rho, q) \end{bmatrix}, & & r_2(\rho, q) & = & \begin{bmatrix} 1 \\ \lambda_2(\rho, q) \end{bmatrix}. \end{array} \tag{4.2}$$

Below we consider different coupling conditions that appeared in the literature. Remark that the geometry of the junction implicitly enters into all the relations below

through the choice of the x_l coordinates. Explicitly, only the area $\|\nu_l\|$ of the cross section appears.

Below we consider the conditions presented in [3, 4, 10, 14, 16, 19, 23]. Any of them prescribes the conservation of mass, so that the first component in (1.2) reads

$$\sum_{l=1}^{n} \|\nu_l\| \, q_l = 0 \,. \tag{4.3}$$

With a slight abuse of notation, we use Ψ to denote the remaining $n-1$ conditions.

(a) We prescribe the equal momentum flow for all connected pipes; see [9] and in the case of open channels [16]:

$$P\left(\rho_i(t,0+), q_i(t,0+)\right) = P\left(\rho_j(t,0+), q_j(t,0+)\right) \qquad \forall\, i \neq j.$$

(b) We prescribe a single pressure at the pipe-to-pipe intersection; see [2, 3] and [14] in the engineering literature. For open canals, see [15]:

$$p\left(\rho_i(t,0+)\right) = p\left(\rho_j(t,0+)\right) \qquad \forall\, i \neq j.$$

(c) In case of only two connected pipes, a further coupling condition is proposed in [19] (see (4.7) for the definition of $k(\theta)$):

$$P\left(\rho_1(t,0+), q_1(t,0+)\right) + k(\theta) = P\left(\rho_2(t,0+), q_2(t,0+)\right).$$

Some remarks are in order. Most of the engineering literature uses (b) modified by so-called *minor loss* factors. These factors are listed in tables and depend on additional information; see [12, 21].

On the other hand, (a) is not so commonly used in the engineering literature but yields the $\mathbf{L^1}$ continuity across stationary transonic shocks (see [10, Example 2.3]), which does not hold when (b) or (c) is adopted. A numerical study of a two-dimensional situation for the p-system can be found in [18] and for the Euler system in [17].

4.1. Equal linear momentum flow for the *p*-system. We consider the setting in [10], i.e., a junction among n pipes, each modeled through the p-system (4.1) with a general pressure law and with the corresponding function Ψ in (1.2) given by (4.3) together with

$$\Psi(\rho, q) = \begin{bmatrix} P(\rho_1, q_1) & - & P(\rho_2, q_2) \\ \vdots & \vdots & \vdots \\ P(\rho_{n-1}, q_{n-1}) & - & P(\rho_n, q_n) \end{bmatrix}, \tag{4.4}$$

where $P(\rho, q) = q^2/\rho + p(\rho)$ is the flow of the linear momentum. The well posedness of the Cauchy problem for (4.1) at a junction was proved in [9, Theorem 3.3].

Remark that this choice of Ψ allows us to consider the problem at the junction as an extension of the standard Cauchy problem. Indeed, setting $n = 2$ and $\nu_1 + \nu_2 = 0$, then (4.1)–(4.4) reduces to the usual situation.

In the general case, the total linear momentum Q varies by

$$Q(t_2) - Q(t_1) = \int_{t_1}^{t_2} \sum_{l=1}^{n} P\left(\rho_l(t,0+), q_l(t,0+)\right) \nu_l dt = \left(\int_{t_1}^{t_2} P_*(t)\, dt \right) \sum_{l=1}^{n} \nu_l$$

(see [9, section 1]), where $P_*(t)$ is the trace of $P\left(\rho_l(t,0+), q_l(t,0+)\right)$, which is independent from l by (1.2)–(4.4). Note that the right-hand side in the equation above depends explicitly on the geometry of the junction.

4.2. Equal pressure for the *p*-system. In [3], the condition at the junction amounts to mass conservation and to pressure equality. Here we extend that approach to the case of n ducts with a general pressure law, so that (1.2) consists of (4.3) with

$$\Psi(\rho, q) = \begin{bmatrix} p(\rho_1) & - & p(\rho_2) \\ \vdots & \vdots & \vdots \\ p(\rho_{n-1}) & - & p(\rho_n) \end{bmatrix}. \tag{4.5}$$

Using [9, Lemma 4.4], the determinant in condition (2.2) evaluates to

$$(-1)^{n+1} \prod_{i=1}^{n} \lambda_2(\hat\rho_i, \hat q_i) \cdot \sum_{i=1}^{n} \prod_{j\neq i} \frac{p'(\hat\rho_j)}{\lambda_2(\hat\rho_j, \hat q_j)}\,. \tag{4.6}$$

Since $(\hat\rho, \hat q) \in \Omega^n$, we have $\lambda_2(\hat\rho_i, \hat q_i) > 0$, while the assumption **(P)** on the pressure law implies $p'(\hat\rho_i) > 0$. Thus, Theorem 3.2 applies, yielding well posedness in the case of n pipes with a general pressure law.

4.3. Two pipes with friction for the *p*-system. The case studied in [19] corresponds to (4.1) with $p(\rho) = \rho$, $n = 2$, $\nu_1 = [-1\ 0]$, and $\nu_2 = [\cos\theta\ \ \sin\theta]$. Here the angular dependence is modeled explicitly and taken into account in the coupling conditions. This situation mimics a kink forming an angle θ in a pipe. Due to the kink, the linear momentum is assumed to vary by a factor k such that

$$k = \sqrt{2(1 - \cos\theta)} \tag{4.7}$$

for $\theta \in [0, \pi/2]$. In the case of pipes possibly with different cross sections $\|\nu_l\|$ and a general pressure law satisfying **(P)**, at the junction we obtain condition (4.3) with

$$\Psi(\rho, q) = \left(\frac{q_1^2}{\rho_1} + p(\rho_1) + fkq_1\right) - \left(\frac{q_2^2}{\rho_2} - p(\rho_2)\right), \tag{4.8}$$

which reduces to the case considered in [19] when $p(\rho) = \rho$ and equal cross sections. The parameter f denotes a nonnegative empirical friction coefficient. The condition (2.2) is

$$\lambda_2(\hat\rho_2, \hat q_2)\, \lambda_2(\hat\rho_1, \hat q_1)\, \big(\lambda_2(\hat\rho_2, \hat q_2) + \lambda_2(\hat\rho_1, \hat q_1) + fk\big) \neq 0\,.$$

Hence, Theorem 3.2 applies, yielding well posedness for general pressure laws.

4.4. Equal momentum flow for open canals. The model presented in [20, formulae (2.3)–(2.7)] for a node among n open canals reads

$$\begin{cases} \partial_t A_l + \partial_x (A_l\, V_l) = 0, \\ \partial_t V_l + \partial_x \left(\dfrac{1}{2} {V_l}^2 + g\, h(A_l)\right) = 0, \end{cases} \qquad \begin{aligned} t &\in \mathbb{R}^+, \\ x &\in \mathbb{R}^+, \\ l &\in \{1, \dots, n\}, \\ (A_l, V_l) &\in \mathring{\mathbb{R}}^+ \times \mathbb{R}, \end{aligned} \tag{4.9}$$

where V_l is the water speed in the lth canal, A_l is the vertical cross section occupied by the water, g is gravity, and h is the water level. Here, differently from [20], we assume

that the canals' beds are all at the same height above sea level, which is acceptable in a neighborhood of the node among the pipes.

Other descriptions for the dynamics of open canals are found in the literature. In particular, [15, section 6.1] presents a different model, based on [13], that fits into our framework of section 4.2.

The coupling condition at the node is given by [20, formulae (2.9)–(2.16)]:

$$\Psi(A,V) = \begin{bmatrix} \displaystyle\sum_{l=1}^{n} A_l\, V_l \\ \displaystyle\frac{1}{2}{V_1}^2 + g\, h(A_1) - \frac{1}{2}{V_2}^2 - g\, h(A_2) \\ \vdots \qquad \vdots \qquad \vdots \\ \displaystyle\frac{1}{2}{V_{n-1}}^2 + g\, h(A_{n-1}) - \frac{1}{2}{V_n}^2 - g\, h(A_n) \end{bmatrix}.$$

The first component in Ψ ensures the conservation of water. For a motivation for the other components, we refer the reader to [20, formulae (2.9)–(2.16)].

This condition is analogous to the equal linear momentum flow condition (4.4) discussed in section 4.1.

Choosing an initial datum $\bar u = (\bar A, \bar V)$ such that $\bar V < \sqrt{\bar A g h'(\bar A)}$ ensures that **(F)** is fulfilled at $\bar u$. In the present case we have

$$\begin{aligned} \lambda_1(A,V) &= V - \sqrt{Agh'(A)}\,, & \lambda_2(\rho,q) &= V + \sqrt{Agh'(A)}\,, \\ r_1(A,V) &= \begin{bmatrix} \sqrt{A} \\ -\sqrt{gh'(A)} \end{bmatrix}, & r_2(A,V) &= \begin{bmatrix} \sqrt{A} \\ \sqrt{gh'(A)} \end{bmatrix}. \end{aligned}$$

The determinant in condition (2.2) evaluates therefore to

$$(-1)^{n+1} \left(\prod_{i=1}^{n} \lambda_2(\bar A_i, \bar V_i) \right) \cdot \sum_{i=1}^{n} \sqrt{\bar A_i} \prod_{j\neq i} \sqrt{gh'(\bar A_j)}\,.$$

Here $A > 0$ by assumption, as well as $\sqrt{gh'(A)} > 0$ by the monotonicity of $h(A)$, which ensures that $\lambda_2(A,V) > 0$. Thus, Theorem 3.2 can be applied, yielding the well posedness for a junction of n open canals.

4.5. Numerical examples for the *p*-system. This section is devoted to comparisons among the different coupling conditions at the junction in the case of the p-system (4.1). Throughout, we use the γ-law: $p(\rho) = p_* \cdot (\rho/\rho_*)^\gamma$ with $\gamma = 1.4$, $\rho_* = 1$, and $p_* = 1$, which clearly satisfies **(P)**. In the case of the coupling conditions (4.8), we set $f \equiv 1$.

In general, stationary solutions for (4.4) fail to be stationary solutions for (4.5) or (4.8). Therefore, we perturb below *static* solutions, which are stationary for all coupling conditions and allow the comparisons.

Below the initial data $u_l(x) = (\rho_l(x), q_l(x))$ along the lth duct attains at most two values, say $u_l^\infty = \lim_{x\to+\infty} u_l(x)$ and $u_l^0 = \lim_{x\to 0+} u_l(x)$. When waves hit the junction, we solve numerically condition (1.2) using Newton's method and obtain the traces u_l^+ of the solution at the junction. The solution to (2.1) is then computed

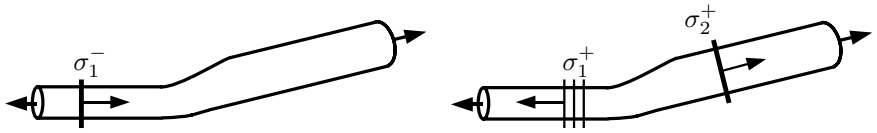


FIG. 1. *Kink with a small angle. Left: a shock approaches a kink with a small angle. Right: a rarefaction is reflected and a shock is refracted.*

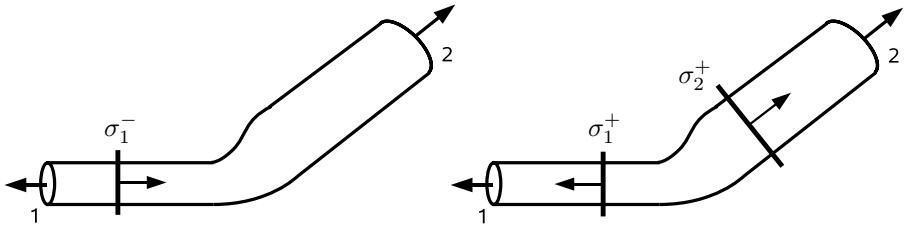


FIG. 2. *Kink with a large angle. Left: a shock approaches a kink. Right: according to condition* (4.8)*, a shock is reflected and a shock is refracted.*

solving a classical Riemann problem between the states u_l^+ (on the left) and u_l^∞ (on the right). Due to the chosen directions of the space variables, waves approaching the junction belong to the first family and those exiting it to the second.

We selected three different examples. In the case of the coupling conditions (4.4) and (4.5), only the cross section of the connected pipes appears explicitly. In the case of (4.8), the angular dependence is taken into account explicitly.

Remark that the numerical values provided in the tables below are expressed in the coordinates x_l adapted to the junction. In particular, the column *Wave speed* is the modulus of the propagation speed of the wave; its x and y components depend on the direction of the pipe. In case of rarefactions, the column *Wave speed* displays the minimal and maximal moduli of the propagation speeds.

4.5.1. Two pipes, different cross sections, possibly different angles. Consider a junction between $n = 2$ horizontal pipes having cross sections $\|\nu_1\| = 1$ and $\|\nu_2\| = 2$. The situation is depicted in Figures 1 and 2. We choose the cases $\theta = 0$, $\theta = \pi/4$, $\theta = \pi/16$, and $\theta = \pi/32$. In each of these cases, a shock with right state $u_1^0 = [1.1000, -0.1253]$ propagating along pipe 1 hits the junction. At first, we compare conditions (4.4) and (4.5). In Table 1, the first column refers to the coupling condition; in the case of (4.8), Table 2 displays the angle θ. The type of the resulting wave is in the fifth column, where S and R refer to 2-(Lax-)shocks and 2-rarefaction waves, respectively.

These numerical integrations suggest that, according to (4.8), there exists an angle θ_* at which no wave is reflected. Clearly, θ_* depends on the initial states, on the ducts' sections, and on the empirical factor f.

4.5.2. Three pipes, different cross sections. Consider $n = 3$ pipes identified by $\nu_1 = [-1 - \sqrt{2} \quad 0]^T$, $\nu_2 = [1 \; 1]^T$, and $\nu_3 = [1 \; 0]^T$ and having cross sections $\|\nu_l\|$. The situation is depicted in Figure 3. Again, we assume that the flow is initially at rest, i.e., $\hat{q}_l = 0$ for $l = 1, 2, 3$. We consider the case of a shock approaching the junction along pipe 1. Due to the choice of the initial data and to the geometry of the junction, the numerical solutions to both coupling conditions (4.4) and (4.5) in fact yield the same results. The final states and the corresponding waves are in Table 3.

TABLE 1
Comparison of results obtained by condition (4.4), (4.8), $\theta = 0$, *and* (4.5) *for two connected pipes with different cross sections.*

Ψ	l	u_l^∞	u_l^+	Wave	Wave speed
(4.4), (4.8), $\theta \equiv 0$	1	$\begin{bmatrix} +1.1000 \\ -0.1253 \end{bmatrix}$	$\begin{bmatrix} +1.0553 \\ -0.1728 \end{bmatrix}$	R	$\begin{bmatrix} +1.0322 \\ +1.0921 \end{bmatrix}$
	2	$\begin{bmatrix} +1.0000 \\ +0.0000 \end{bmatrix}$	$\begin{bmatrix} +1.0701 \\ +0.0864 \end{bmatrix}$	S	$+1.2324$
(4.5)	1	$\begin{bmatrix} +1.1000 \\ -0.1253 \end{bmatrix}$	$\begin{bmatrix} +1.0658 \\ -0.1618 \end{bmatrix}$	R	$\begin{bmatrix} +1.0466 \\ +1.0921 \end{bmatrix}$
	2	$\begin{bmatrix} +1.0000 \\ +0.0000 \end{bmatrix}$	$\begin{bmatrix} +1.0658 \\ +0.0809 \end{bmatrix}$	S	$+1.2294$

TABLE 2
Comparison for the situation of a kink of angle θ. *The case* $\theta = 0$ *can be found in Table* 1.

Ψ	l	u_l^∞	u_l^+	Wave	Wave speed
(4.8)	1	$\begin{bmatrix} +1.1000 \\ -0.1253 \end{bmatrix}$	$\begin{bmatrix} +1.0645 \\ -0.1633 \end{bmatrix}$	R	$\begin{bmatrix} +1.0447 \\ +1.0921 \end{bmatrix}$
$\theta = \frac{\pi}{32}$	2	$\begin{bmatrix} +1.0000 \\ +0.0000 \end{bmatrix}$	$\begin{bmatrix} +1.0664 \\ +0.0816 \end{bmatrix}$	S	$+1.1298$
(4.8)	1	$\begin{bmatrix} +1.1000 \\ -0.1253 \end{bmatrix}$	$\begin{bmatrix} +1.0725 \\ -0.1548 \end{bmatrix}$	R	$\begin{bmatrix} +1.0556 \\ +1.0921 \end{bmatrix}$
$\theta = \frac{\pi}{16}$	2	$\begin{bmatrix} +1.0000 \\ +0.0000 \end{bmatrix}$	$\begin{bmatrix} +1.0631 \\ +0.0774 \end{bmatrix}$	S	$+1.2275$
(4.8)	1	$\begin{bmatrix} +1.1000 \\ -0.1253 \end{bmatrix}$	$\begin{bmatrix} +1.1056 \\ -0.1192 \end{bmatrix}$	S	$+1.0958$
$\theta = \frac{\pi}{4}$	2	$\begin{bmatrix} +1.0000 \\ +0.0000 \end{bmatrix}$	$\begin{bmatrix} +1.0489 \\ +0.0596 \end{bmatrix}$	S	$+1.2177$

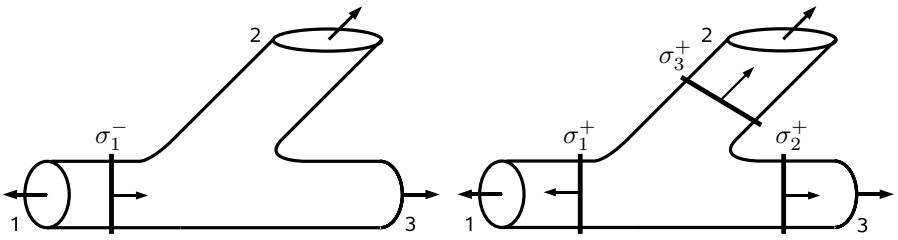


FIG. 3. *A shock hitting a T-junction.*

4.5.3. Four pipes, different cross sections. Finally, we consider a junction with $n = 4$ pipes defined by $\nu_1 = [-1\ 0]$, $\nu_2 = [0\ 1]$, $\nu_3 = [0\ -1]$, and $\nu_4 = [1\ 0]$. The situation is depicted in Figure 4. Initially the gas flow is at rest, i.e., $\hat{q}_l = 0$, and $\hat{\rho}_l = 1$ for $l = 1, \dots, 4$. We let three 1-Lax-shocks of different strengths collide simultaneously at the junction along pipes 1, 3, and 4.

It is remarkable that the two coupling conditions yield qualitatively different results. The wave reflected in tube 1 is a rarefaction according to (4.4) and a shock according to (4.5). However, the propagation speeds in the two cases are close to each other, coherently with the $\mathbf{L^1}$ continuous dependence.

Table 4 displays, for each duct, the type of wave arising after the interaction at the junction, and its speed is reported below for the two different coupling conditions (4.4) and (4.5).

TABLE 3
Numerical results for section 4.5.2.

(4.4) and (4.5)	$l=1$	$l=2$	$l=3$
u_l^∞	$\begin{bmatrix} +1.0000 \\ +0.0000 \end{bmatrix}$	$\begin{bmatrix} +1.0000 \\ +0.0000 \end{bmatrix}$	$\begin{bmatrix} +1.0000 \\ +0.0000 \end{bmatrix}$
u_l^0	$\begin{bmatrix} +1.2000 \\ -0.2642 \end{bmatrix}$	$\begin{bmatrix} +1.0000 \\ +0.0000 \end{bmatrix}$	$\begin{bmatrix} +1.0000 \\ +0.0000 \end{bmatrix}$
u_l^+	$\begin{bmatrix} +1.2002 \\ -0.2640 \end{bmatrix}$	$\begin{bmatrix} +1.2002 \\ +0.2640 \end{bmatrix}$	$\begin{bmatrix} +1.2002 \\ +0.2640 \end{bmatrix}$
Wave type	S	S	S
Wave speed	+1.0071	+1.3206	+1.3206

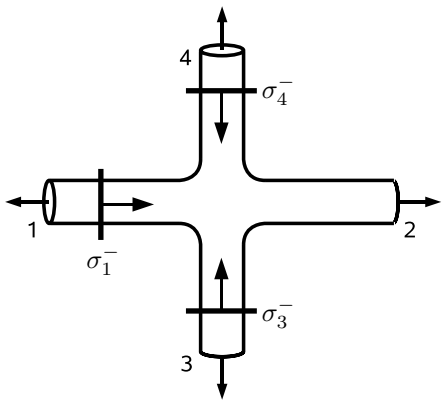


FIG. 4. *Three Lax shocks hit the junction simultaneously.*

TABLE 4
Numerical results for section 4.5.3.

	$l=1$	$l=2$	$l=3$	$l=4$
u_l^∞	$\begin{bmatrix} +1.0000 \\ +0.0000 \end{bmatrix}$	$\begin{bmatrix} +1.0000 \\ +0.0000 \end{bmatrix}$	$\begin{bmatrix} +1.0000 \\ +0.0000 \end{bmatrix}$	$\begin{bmatrix} +1.0000 \\ +0.0000 \end{bmatrix}$
u_l^0	$\begin{bmatrix} +1.1500 \\ -0.1931 \end{bmatrix}$	$\begin{bmatrix} +1.0000 \\ +0.0000 \end{bmatrix}$	$\begin{bmatrix} +1.1000 \\ -0.1253 \end{bmatrix}$	$\begin{bmatrix} +1.0500 \\ -0.0609 \end{bmatrix}$
	Coupling condition (4.4)			
u_l^+	$\begin{bmatrix} +1.1396 \\ -0.2039 \end{bmatrix}$	$\begin{bmatrix} +1.1441 \\ +0.1848 \end{bmatrix}$	$\begin{bmatrix} +1.1625 \\ -0.0545 \end{bmatrix}$	$\begin{bmatrix} +1.1611 \\ +0.0736 \end{bmatrix}$
Wave type	R	S	S	S
Wave speed	$\begin{bmatrix} +1.0357 \\ +1.0489 \end{bmatrix}$	+1.2831	+1.1328	+1.2113
	Coupling condition (4.5)			
u_l^+	$\begin{bmatrix} +1.1520 \\ -0.1909 \end{bmatrix}$	$\begin{bmatrix} +1.1520 \\ +0.1957 \end{bmatrix}$	$\begin{bmatrix} +1.1520 \\ -0.0667 \end{bmatrix}$	$\begin{bmatrix} +1.1520 \\ +0.0620 \end{bmatrix}$
Wave type	S	S	S	S
Wave speed	+1.0501	+1.2884	+1.1260	+1.2053

5. Technical details. As a general reference on the theory of hyperbolic systems of conservation laws, we refer the reader to [6]. Denote by $\sigma \mapsto \mathcal{L}_i(u_o, \sigma)$ the ith Lax curve through u_o for $i = 1, 2$. As usual, $\mathcal{O}(1)$ denotes a sufficiently large constant dependent only on f restricted to a neighborhood of the initial states.

Proof of Proposition 2.2. It is sufficient to show that for all $\bar{u}$ sufficiently near $\hat{u}$ the nonlinear system $\Psi\left(\mathcal{L}_2(\bar{u}_1, \sigma_1), \dots, \mathcal{L}_2(\bar{u}_n, \sigma_n)\right) = 0$ admits a unique n-tuple of solution $\sigma_1, \dots, \sigma_n$. Indeed, condition (2.2) allows us to use the implicit function theorem.

The Lipschitz estimate (2.4) follows immediately from the regularity of the implicit function.

To prove the latter statement, it is sufficient to consider a single pipe, say the first one. Let u, respectively, $\tilde{u}$, be the solutions to (2.1) with condition Ψ, respectively, $\tilde{\Psi}$, at the junction. If u and $\tilde{u}$ contain a single shock, then

$$
\begin{aligned}
&\|u - \tilde{u}\| \\
&= \int_0^{\min\{\Lambda^1, \Lambda^2\}t} \|u(x) - \tilde{u}(x)\| \, dx \\
&\quad + \int_{\min\{\Lambda^1, \Lambda^2\}t}^{\max\{\Lambda^1, \Lambda^2\}t} \|u(x) - \tilde{u}(x)\| \, dx \\
&= \mathcal{O}(1) \cdot \|u(t, 0+) - \tilde{u}(t, 0+)\| \cdot t + \mathcal{O}(1) \cdot \left|\Lambda^2 - \Lambda^1\right| \cdot t \\
&= \mathcal{O}(1) \cdot \left\|\tilde{\Psi} - \Psi\right\|_{\mathbf{C^1}} \cdot t + \mathcal{O}(1) \cdot \|u(t, 0+) - \tilde{u}(t, 0+)\| \cdot t \\
&= \mathcal{O}(1) \cdot \left\|\tilde{\Psi} - \Psi\right\|_{\mathbf{C^1}} \cdot t \, .
\end{aligned}
$$

The case of one or both solutions containing a rarefaction is similar; see also the proof of [5, Corollary 2.5]. □

We now pass to the proof of Theorem 3.2. To do this, we first use wave front tracking to construct approximate solutions to the Cauchy problem (3.1) adapting the wave front tracking technique; see [6, Chapter 7].

Let $\hat{\delta} > 0$ be such that the closed sphere $\overline{B(\hat{u}_l, \hat{\delta})} \subset \Omega$ for $l = 1, \dots, n$, and introduce the compact set $\mathcal{B} = \prod_{l=1}^n \overline{B(\hat{u}_l, \hat{\delta})}$. We omit the proof of the following simple estimate.

LEMMA 5.1. *For all $u \in \overline{B(\hat{u}, \hat{\delta})}$ there exists $C > 0$ such that if $\|u - \hat{u}\| < \hat{\delta}$ and $\|\mathcal{L}_i(u; \sigma) - u\| \leq \hat{\delta}$, then*

$$
\frac{1}{C} \cdot |\sigma| \leq \|\mathcal{L}_i(u; \sigma) - u\| \leq C \cdot |\sigma| \, .
$$

Fix $\varepsilon > 0$. Approximate the initial datum u_o with a sequence $u_{o,\varepsilon}$ of piecewise constant initial data each having a finite number of discontinuities so that $\lim_{\varepsilon \to 0} \left\|u_{o,\varepsilon} - u_o\right\|_{\mathbf{L^1}} = 0$. Then, at the junction and at each point of jump in the approximate initial datum along the pipes, we solve the corresponding Riemann problem according to Definition 2.1. If the total variation of the initial datum is sufficiently small, then Proposition 2.2 ensures the existence and uniqueness of solutions to the Riemann problem. We approximate each rarefaction wave with a rarefaction fan,

i.e., by means of (nonentropic) shock waves traveling at the characteristic speed of the state to the right of the shock and with size at most ε.

This construction can be extended up to the first time $\bar{t}_1$ at which two waves interact in a pipe or a wave hits the junction. Clearly, at time $\bar{t}_1$ the functions so constructed are piecewise constant with a finite number of discontinuities. Hence, at any subsequent interaction or collision with the junction, we repeat the previous construction with the following provisions:

1. no more than two waves interact at the same point or at the junction;
2. a rarefaction fan of the ith family produced by the interaction between an ith rarefaction and any other wave is *not* split any further;
3. when the product of the strengths of two interacting waves falls below a threshold $\check{\varepsilon}$, then we let the waves cross each other, their size being unaltered, and introduce a *nonphysical* wave with speed $\hat{\lambda}$, with $\hat{\lambda} > \sup_u \lambda_2(u)$; see [6, Chapter 7] and the refinement [1].

In the present case, we have to complete the above algorithm stating how the Riemann problem at the junction is to be solved. At time $t = 0$ and whenever a physical wave with size greater than $\check{\varepsilon}$ hits the junction, the accurate solver is used; i.e., the exact solution as in Definition 2.1 is approximated by replacing rarefaction waves with rarefaction fans. When a wave with strength smaller than $\check{\varepsilon}$ hits the junction, then we let it be reflected into a nonphysical wave with speed $\hat{\lambda}$, and no wave in any other pipe is produced.

Repeating recursively this procedure, we construct a wave front tracking sequence of approximate solutions u_ε in the sense of [6, Definition 7.1].

At interactions of waves in a pipe, we have the following classical result.

LEMMA 5.2. *There exists a constant K with the following property.*

1. *If there is an interaction in a pipe between two waves σ_1^- and σ_2^-, respectively, of the first and second family, producing the waves σ_1^+ and σ_2^+ (see Figure 5, left), then*

$$\left|\sigma_1^+ - \sigma_1^-\right| + \left|\sigma_2^+ - \sigma_2^-\right| \leq K \cdot \left|\sigma_1^- \sigma_2^-\right|. \tag{5.1}$$

2. *If there is an interaction in a pipe between two waves σ_i' and σ_i'' of the same ith family producing waves of total size σ_1^+ and σ_2^+ (see Figure 5, right, for the case $i = 2$), then*

$$\left|\sigma_1^+ - (\sigma_1'' + \sigma_1')\right| + \left|\sigma_2^+\right| \leq K \cdot \left|\sigma_1' \sigma_1''\right| \qquad \text{if } i = 1,$$

$$\left|\sigma_1^+\right| + \left|\sigma_2^+ - (\sigma_2'' + \sigma_2')\right| \leq K \cdot \left|\sigma_2' \sigma_2''\right| \qquad \text{if } i = 2.$$

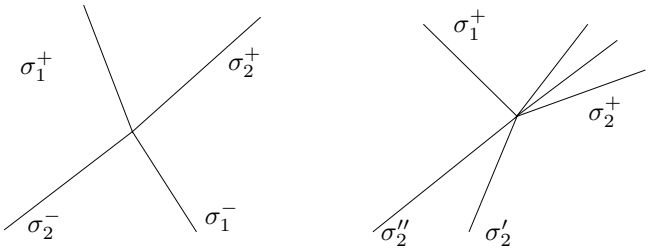


FIG. 5. *Notation for the standard interaction estimates in Lemma* 5.2.

3. *If there is an interaction in a pipe between two physical waves* σ_1^- *and* σ_2^- *producing a nonphysical wave* σ_3^+ *(see Figure* 6*, left), then*

$$\left|\sigma_3^+\right| \leq K \cdot \left|\sigma_1^- \sigma_2^-\right|.$$

4. *If there is an interaction in a pipe between a physical wave* σ *and a nonphysical wave* σ_3^- *producing a physical wave* σ *and a nonphysical wave* σ_3^+ *(see Figure* 6*, right), then*

$$\left|\sigma_3^+\right| - \left|\sigma_3^-\right| \leq K \cdot \left|\sigma\sigma_3^-\right|.$$

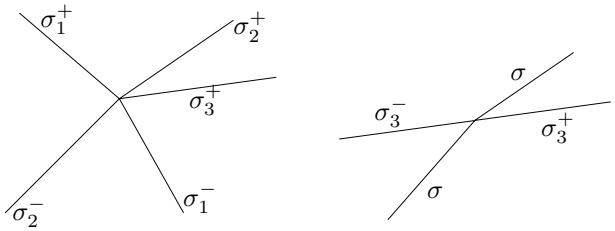


FIG. 6. *Left: a nonphysical wave arises. Right: a nonphysical wave hits a physical one.*

For a proof of this result see [6, Chapter 7]. By construction, nonphysical waves cannot interact with the junction or with other nonphysical waves. In the case of the junction, we have the following result, with notation as depicted in Figure 7.

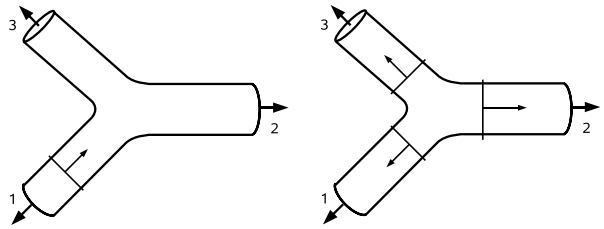


FIG. 7. *Notation for Proposition* 5.3*. Left: before the interaction. Right: after the interaction.*

PROPOSITION 5.3. *There exist* $\delta_J > 0$ *and* $K_J \geq 1$ *with the following property. For any* $\bar{u} \in \mathcal{B}$ *that yields a stationary solution to the Riemann problem* (2.1)*, and for any* 1*-waves* $\sigma_{\bar{l}}^- \in]-\delta_J, \delta_J[$ *hitting the junction and producing the* 2*-waves* σ_l^+*,*

$$\sum_{l=1}^{n} \left|\sigma_l^+\right| \leq K_J \cdot \left|\sigma_{\bar{l}}^-\right|. \tag{5.2}$$

The proof follows immediately from (2.4) and Lemma 5.1.

Define $\bar{K} = 2KK_J + 1$. Fix a wave front tracking approximate solution u_ε. For $t > 0$ and $l \in \{1, \ldots, n\}$, we denote with $\{x_{l,\alpha} : \alpha \in \mathcal{J}_l(u)\}$ the set of the positions of the discontinuities of the approximate solution u in the lth pipe and with $\sigma_{l,1,\alpha}$, $\sigma_{l,2,\alpha}$, $\sigma_{l,3,\alpha}$ the strengths of the waves, respectively, of the first family, of the second family, and of the nonphysical waves at $x_{l,\alpha}$. Introduce the Glimm-type functionals

$$V(t) = \sum_{l=1}^{n} \sum_{\alpha \in \mathcal{J}_l} \left(2\,K_J \cdot \left|\sigma_{l,1,\alpha}\right| + \left|\sigma_{l,2,\alpha}\right| + \left|\sigma_{l,3,\alpha}\right|\right), \tag{5.3}$$

$$Q(t) = \sum_{l=1}^{n} \sum \left\{ \left|\sigma_{l,i,\alpha}\, \sigma_{l,j,\beta}\right| : (\sigma_{l,i,\alpha}, \sigma_{l,j,\beta}) \in \mathcal{A}_l \right\},$$

$$\Upsilon(t) = V(t) + \check{K} \cdot Q(t)\,,$$

where $\mathcal{A}_l$ denotes the set of approaching waves in the lth pipe; see [6, section 7.3].

The functionals above are well defined for every $t > 0$ at which no interaction takes place. Now suppose that at a time $\tau > 0$ there is an interaction between the wave $\sigma_{\bar{l},1}$ of the first family and the junction. In general, this interaction produces n waves $\sigma^+_{l,2}$ of the second family. Thus

$$\Delta V(\tau) \le \left(\sum_{l=1}^{n} \left|\sigma^+_{l,2}\right| \right) - 2\, K_J \left|\sigma_{\bar{l},1}\right| \;\le\; -K_J \cdot \left|\sigma_{\bar{l},1}\right|,$$

$$\Delta Q(\tau) \le \sum_{l=1}^{n} \left(\left|\sigma^+_{l,2}\right| \sum_{i,\alpha} \left|\sigma_{l,i,\alpha}\right| \right) \;\le\; K_J \cdot V(\tau-) \cdot \left|\sigma_{\bar{l},1}\right|,$$

$$\Delta\Upsilon(\tau) \le K_J \cdot \left(\check{K} \cdot V(\tau-) - 1 \right) \cdot \left|\sigma_{\bar{l},1}\right|.$$

Now suppose an interaction between two waves $\sigma_{l,i,\alpha}$, $\sigma'_{l,j,\beta}$ happens in a pipe at time τ. By Lemma 5.2 we deduce that

$$\Delta V(\tau) \le 2 \cdot K \cdot K_J \cdot \left|\sigma_{l,i,\alpha}\sigma'_{l,j,\beta}\right|,$$

$$\Delta Q(\tau) \le K \cdot \left|\sigma_{l,i,\alpha}\sigma'_{l,j,\beta}\right| \cdot V(\tau-) - \left|\sigma_{l,i,\alpha}\sigma'_{l,j,\beta}\right|,$$

$$\Delta\Upsilon(\tau) \le \left|\sigma_{l,i,\alpha}\sigma'_{l,j,\beta}\right| \left(K \cdot (2 \cdot K_J + \check{K} \cdot V(\tau-)) - \check{K} \right).$$

We have thus proved the following basic result.

PROPOSITION 5.4. *Let* $\check{\delta} = \min\{1/(2\check{K}+1), 1/(2K\check{K}+1), \hat{\delta}, \delta_J\}$. *At any interaction time* $\tau > 0$, *if* $V(\tau-) < \check{\delta}$, *then* $\Delta\Upsilon(\tau) < 0$ *with* Υ *defined in* (5.4).

Proof of Theorem 3.2. Let $\check{\delta}$ be as in Proposition 5.4, and define

$$\tilde{\mathcal{D}} = \left\{ u \in \hat{u} + \mathbf{L^1}\left(\mathbb{R}^+; (\mathring{\mathbb{R}}^+ \times \mathbb{R})^n\right) : u \in \mathbf{PC} \text{ and } \Upsilon(u) \le \check{\delta} \right\};$$

here **PC** denotes the set of piecewise constant functions with finitely many jumps. It is immediate to prove that there exists a suitable $C_1 > 0$ such that $\frac{1}{C_1}\,\mathrm{TV}\,(u)(t,\cdot) \le V(t) \le C_1\,\mathrm{TV}\,(u)(t,\cdot)$ for all $u \in \tilde{\mathcal{D}}$. Any initial data in $\tilde{\mathcal{D}}$ yields an approximate solution to (1.1) attaining values in $\tilde{\mathcal{D}}$ by Proposition 5.4.

We now pass to the $\mathbf{L^1}$-Lipschitz continuous dependence of the approximate solutions from the initial datum. Consider two wave front tracking approximate solutions u_1 and u_2. Define the functional

$$\Phi\,(u_1, u_2) = \sum_{l=1}^{n} \sum_{i=1}^{2} \int_0^{+\infty} \left|s_{l,i}(x)\right| W_{l,i}(x)\, dx\,, \tag{5.4}$$

where $s_{l,i}(x)$ measures the strengths of the ith shock wave in the lth pipe at point x (see [6, Chapter 8]) and the weights $W_{l,i}$ are defined by

$$W_{l,1}(x) = \hat{K} \cdot \left(1 + \kappa_1 \, A_{l,i}(x) + \kappa_1 \, \kappa_2 \, \left(\Upsilon(u_1) + \Upsilon(u_2)\right)\right),$$
$$W_{l,2}(x) = 1 + \kappa_1 \, A_{l,i}(x) + \kappa_1 \, \kappa_2 \, \left(\Upsilon(u_1) + \Upsilon(u_2)\right)$$

for suitable positive constants κ_1, κ_2 chosen as in [6, formula (8.7)] and $\hat{K} = 1 + (\max_l W_{l,2}) \, K_j \, \hat{\lambda}/|\inf \lambda_1|$. Here Υ is the functional defined in (5.4), while the $A_{l,i}$ are defined by

$$A_{l,i}(x) = \sum \left\{ \left|\sigma_{l,k_\alpha,\alpha}\right| \colon \begin{array}{l} x_\alpha < x, \, i < k_\alpha \leq 2 \\ x_\alpha > x, \, 1 \leq k_\alpha < i \end{array} \right\}$$
$$+ \begin{cases} \displaystyle\sum \left\{ \left|\sigma_{l,i,\alpha}\right| \colon \begin{array}{l} x_\alpha < x, \, \alpha \in \mathcal{J}_l(u_1) \\ x_\alpha > x, \, \alpha \in \mathcal{J}_l(u_2) \end{array} \right\} & \text{if } s_{l,i}(x) < 0, \\ \displaystyle\sum \left\{ \left|\sigma_{l,i,\alpha}\right| \colon \begin{array}{l} x_\alpha < x, \, \alpha \in \mathcal{J}_l(u_2) \\ x_\alpha > x, \, \alpha \in \mathcal{J}_l(u_1) \end{array} \right\} & \text{if } s_{l,i}(x) \geq 0; \end{cases}$$

see [6, Chapter 8]. We first fix κ_1, κ_2, so that $\check{\delta}$ in the definition of $\tilde{\mathcal{D}}$ can be chosen to make $W_{l,i}(x) \geq 1$ and uniformly bounded for every $l \in \{1, \dots, n\}$, $i \in \{1, 2\}$, and $x \geq 0$. Hence the functional Φ is equivalent to $\mathbf{L^1}$ distance:

$$\frac{1}{C_2} \cdot \|u_1 - u_2\|_{\mathbf{L^1}} \leq \Phi\left(u_1, u_2\right) \leq C_2 \cdot \|u_1 - u_2\|_{\mathbf{L^1}}$$

for a positive constant C_2. The same calculations as in [6, Chapter 8] show that, at any time $t > 0$ when an interaction happens neither in u_1 nor in u_2,

$$\begin{aligned} &\frac{d}{dt} \Phi\left(u_1(t), u_2(t)\right) \\ &\leq C_3 \, \varepsilon + \sum_l \sum_i \left|s_{l,i}(0+)\right| W_{l,i} \, \lambda_{l,i}(0+) \\ &\leq C_3 \, \varepsilon - \left(\sum_l \left|s_{l,1}(0+)\right|\right) \hat{K} \left|\inf_l \lambda_{l,1}\right| + K_J \left(\sum_l \left|s_{l,1}(0+)\right|\right) \max_l W_{l,2} \, \hat{\lambda} \\ &\leq C_3 \, \varepsilon + \sum_l \left|s_{l,1}(0+)\right| \left(K_J \max_l W_{l,2} \, \hat{\lambda} - \hat{K} \left|\inf \lambda_{l,1}\right|\right) \\ &\leq C_3 \, \varepsilon \, , \end{aligned}$$

where C_3 is a suitable positive constant depending only on a bound on the total variation of the initial data. Above we used the analogue of Proposition 5.3 for shock curves, i.e., if $\Psi(S_2(S_1(u, q_{l,1}), q_{l,2})) = 0$, then $\sum_l \left|q_{l,2}\right| \leq K_J \sum_l \left|q_{l,1}\right|$, for a suitable K_J.

If $t > 0$ is an interaction time for u_1 or u_2, then, by Proposition 5.4, $\Delta[\Upsilon(u_1(t)) + (u_2(t))] < 0$ and, choosing κ_2 large enough, we obtain

$$\Delta \Phi\left(u_1(t), u_2(t)\right) < 0 \, .$$

Thus, $\Phi\left(u_1(t), u_2(t)\right) - \Phi\left(u_1(s), u_2(s)\right) \leq C_2\, \varepsilon\, (t-s)$ for every $0 \leq s \leq t$. The proof is now completed using the standard arguments in [6, Chapter 8].

The proof that the semigroup trajectory does indeed yield a solution to (3.1) and, in particular, that **(Ψ)** is satisfied on the traces is exactly as that of [7, Proposition 5.3].

We now pass to the stability estimate (3.2). Its proof is similar to those of [5, Theorem 2.1] or [8, Theorem 3.1] and is based on [6, Theorem 2.9], which we recall for convenience: for every Lipschitz map $w\colon [0,T] \to \mathcal{D}$ and every Lipschitz semigroup $S\colon [0,+\infty[\, \times \mathcal{D} \to \mathcal{D}$, the following estimate holds:

$$\|w(T) - S_T w(0)\|_{\mathbf{L^1}} \leq L \cdot \int_0^T \left(\liminf_{h \to 0+} \frac{1}{h} \|w(t+h) - S_h w(t)\|_{\mathbf{L^1}} \right) dt\,,$$

L being the Lipschitz constant of S. In the present case, we are led to

$$\left\| S^{\tilde\Psi}_t u - S^{\Psi}_t u \right\|_{\mathbf{L^1}} \leq L^{\Psi} \cdot \int_0^t \left(\liminf_{h \to 0+} \frac{1}{h} \left\| S^{\Psi}_h S^{\Psi}_\tau u - S^{\tilde\Psi}_h S^{\Psi}_\tau u \right\| \right) d\tau\,.$$

It remains to estimate $\|S^{\Psi}_h v - S^{\tilde\Psi}_h v\|_{\mathbf{L^1}}$ for $v = S^{\Psi}_\tau u$. We use the wave front tracking approximations $v^{\Psi,\varepsilon}(h,\cdot) = S^{\Psi,\varepsilon}_h u$ and $v^{\tilde\Psi,\varepsilon}(h,\cdot) = S^{\tilde\Psi,\varepsilon}_h u$. For $h > 0$ sufficiently small, we can assume that there is at most a single interaction of the waves of $u \in \mathcal{D}$ with the intersection. Then $S^{\Psi,\varepsilon}_h$ coincides with the Riemann solver of Proposition 2.2, and estimate (2.4) can be applied:

$$\left\| S^{\Psi}_h u - S^{\tilde\Psi}_h u \right\| \leq L^{\Psi} \|\Psi_1 - \Psi_2\|_{\mathbf{C^1}} h\,.$$

Since the right-hand side is independent of ε and since $S^{\Psi,\varepsilon}_h u$ converges in $\mathbf{L^1}$ to $S^{\Psi}_h u$, we obtain

$$\left\| S^{\Psi}_h u - S^{\tilde\Psi}_h u \right\|_{\mathbf{L^1}} \leq \int_0^t \left(\liminf_{h \to 0+} \frac{1}{h} \cdot \left\| \Psi - \tilde\Psi \right\|_{\mathbf{C^1}} \cdot h \right) d\tau \leq L^{\Psi} \cdot \left\| \Psi - \tilde\Psi \right\|_{\mathbf{C^1}} \cdot t,$$

completing the proof. □

Acknowledgments. We thank Graziano Guerra for helpful discussions. This work was started while the first author was visiting TU Kaiserslautern.

REFERENCES

[1] P. Baiti and H. K. Jenssen, *On the front-tracking algorithm*, J. Math. Anal. Appl., 217 (1998), pp. 395–404.

[2] M. K. Banda, M. Herty, and A. Klar, *Coupling conditions for gas networks governed by the isothermal Euler equations*, Netw. Heterog. Media, 1 (2006), pp. 275–294.

[3] M. K. Banda, M. Herty, and A. Klar, *Gas flow in pipeline networks*, Netw. Heterog. Media, 1 (2006), pp. 41–56.

[4] G. Bastin, B. Haut, J.-M. Coron, and B. D'andréa-Novel, *Lyapunov stability analysis of networks of scalar conservation laws*, Netw. Heterog. Media, 2 (2007), pp. 751–759.

[5] S. Bianchini and R. M. Colombo, *On the stability of the standard Riemann semigroup*, Proc. Amer. Math. Soc., 130 (2002), pp. 1961–1973.

[6] A. Bressan, *Hyperbolic Systems of Conservation Laws. The One-Dimensional Cauchy Problem*, Oxford Lecture Ser. Math. Appl. 20, Oxford University Press, Oxford, UK, 2000.

[7] R. M. Colombo and A. Corli, *Sonic hyperbolic phase transitions and Chapman-Jouguet detonations*, J. Differential Equations, 184 (2002), pp. 321–347.

[8] R. M. Colombo and A. Corli, *Stability of the Riemann semigroup with respect to the kinetic condition*, Quart. Appl. Math., 62 (2004), pp. 541–551.

[9] R. M. Colombo and M. Garavello, *A well posed Riemann problem for the p-system at a junction*, Netw. Heterog. Media, 1 (2006), pp. 495–511.
[10] R. M. Colombo and M. Garavello, *On the Cauchy problem for the p-system at a junction*, SIAM J. Math. Anal., 39 (2008), pp. 1456–1471.
[11] R. M. Colombo and C. Mauri, *Euler system for compressible fluids at a junction*, J. Hyperbolic Differ. Equ., to appear.
[12] Crane Valve Group, *Flow of Fluids through Valves, Fittings and Pipes*, Technical report 410, Crane Technical paper, 1998.
[13] A. J. C. B. de Saint-Venant, *Theorie du mouvement non-permanent des eaux avec application aux crues des rivières et à l'introduction des marees dans leur lit*, Comptes Rendus Academie des Sciences, 73 (1871), pp. 148–154, 237–240.
[14] K. Ehrhardt and M. Steinbach, *Nonlinear gas optimization in gas networks*, in Modeling, Simulation and Optimzation of Complex Processes, H. G. Bock, E. Kostina, H. X. Pu, and R. Ranacher, eds., Springer, Berlin, 2005, pp. 139–148.
[15] M. Gugat, *Nodal control of conservation laws on networks*, in Control and Boundary Analysis, Lect. Notes Pure Appl. Math. 240, Chapman and Hall/CRC, Boca Raton, FL, 2005, pp. 201–215.
[16] M. Gugat and G. Leugering, *Global boundary controllability of the de St. Venant equations between steady states*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 20 (2003), pp. 1–11.
[17] M. Herty, *Coupling conditions for networked systems of Euler equations*, SIAM J. Sci. Comput., 30 (2008), pp. 1596–1612.
[18] M. Herty and M. Seaid, *Simulation of transient gas flow at pipe-to-pipe intersections*, Internat. J. Numer. Methods Fluids, 56 (2008), pp. 485–506.
[19] H. Holden and N. H. Risebro, *Riemann problems with a kink*, SIAM J. Math. Anal., 30 (1999), pp. 497–515.
[20] G. Leugering and E. J. P. G. Schmidt, *On the modelling and stabilization of flows in networks of open canals*, SIAM J. Control Optim., 41 (2002), pp. 164–180.
[21] Mapress. *Mapress Pressfitting System*, Technical report, Mapress GmbH & Co. KG, Langenfeld, Germany, http://www.mapress.de (2002).
[22] A. Martin, M. Möller, and S. Moritz, *Mixed integer models for the stationary case of gas network optimization*, Math. Program., 105 (2006), pp. 563–582.
[23] A. J. Osiadacz, *Simulation and Analysis of Gas Networks*, Gulf Publishing, Houston, TX, 1989.
[24] A. J. Osiadacz, *Different transient models—limitations, advantages and disadvantages*, in Proceedings of the 28th Annual Meeting of PSIG (Pipeline Simulation Interest Group), 1996.
[25] A. Osiadacz, *Simulation of transient gas flows in networks*, Internat. J. Numer. Methods Fluids, 4 (1984), pp. 13–24.
[26] Pipeline Simulation Interest Group, http://www.psig.org/.
[27] J. Zhou and M. A. Adewumi, *Simulation of transients in natural gas pipelines using hybrid TVD schemes*, Internat. J. Numer. Methods Fluids, 32 (2000), pp. 407–437.

SIAM J. MATH. ANAL.
Vol. 40, No. 2, pp. 623–652

A VARIATIONAL PRINCIPLE FOR HARDENING ELASTOPLASTICITY*

ULISSE STEFANELLI†

Abstract. We present a variational principle governing the quasi-static evolution of a linearized elastoplastic material. In the case of linear hardening, the novel characterization allows us to recover and partly extend some known results and proves itself to be especially well suited for discussing general approximation and convergence issues. In particular, the variational principle is exploited in order to prove in a novel setting the convergence of time and space-time discretizations as well as to provide some possible a posteriori error control.

Key words. variational principle, elastoplasticity, approximation

AMS subject classifications. 35K55, 49S05, 74C05

DOI. 10.1137/070692571

1. Introduction. The primal initial-boundary value problem of elastoplasticity consists in determining the generalized deformation state of a material subject to external mechanical actions. In particular, starting from some initial state and for a given load and traction, one shall determine the displacement u of the body from the reference configuration, the inelastic (plastic) part p of its strain, and, possibly, a vector of internal hardening variables ξ. In the small deformation regime and within the frame of associative elastoplasticity, the problem is classically formulated in a variational form as that of finding the absolutely continuous trajectory $t \in [0,T] \mapsto y(t) \in Y$ (Y is a Banach space) such that

$$\partial\psi(\dot{y}) + Ay \ni \ell \quad \text{a.e. in } (0,T), \quad y(0) = y_0, \tag{1.1}$$

where $y = (u, p, \xi)$ stands for the vector of unknown fields, $A : Y \to Y^*$ (dual) is linear, continuous, and symmetric, and $\psi : Y \to [0,\infty]$ is the positively 1-homogeneous and convex dissipation potential (∂ is the classical subdifferential in the sense of convex analysis; see below). Moreover, $\ell : [0,T] \to Y^*$ is a given and suitably smooth generalized load (possibly including surface tractions), and y_0 represents the initial state. The reader is referred to section 2 for some brief mechanical motivation as well as to the classical monographs by Duvaut and Lions [7], Han and Reddy [13], Lemaitre and Chaboche [18], and Simo and Hughes [42] for a comprehensive collection of results.

The aim of this paper is that of investigating a global-in-time variational formulation of problem (1.1). In particular, we shall introduce the functional $\mathcal{F} : W^{1,1}(0,T;Y) \to [0,\infty]$ on trajectories as

$$\mathcal{F}(y) = \int_0^T \big(\psi(\dot{y}) + \psi^*(\ell - Ay) - \langle \ell - Ay, \dot{y}\rangle\big),$$

*Received by the editors May 21, 2007; accepted for publication (in revised form) January 17, 2008; published electronically July 3, 2008. This research was performed during a visit to the Seminar for Applied Mathematics at ETH Zürich and the Institute of Mathematics of the University of Zürich under the sponsorship of the STM CNR 2006 program and the Swiss National Science Foundation. The support and hospitality of both institutions is gratefully acknowledged.
http://www.siam.org/journals/sima/40-2/69257.html
†IMATI - CNR, v. Ferrata 1, I-27100, Pavia, Italy (ulisse.stefanelli@imati.cnr.it).

where ψ^* stands for the conjugate $\psi^*(w) = \sup_{v\in Y}(\langle w, v\rangle - \psi(v))$ of ψ and $\langle\cdot,\cdot\rangle$ denotes the duality pairing between Y^* and Y. The starting point of this analysis relies on the fact that solutions of (1.1) and minimizers of $\mathcal{F}$ fulfilling the initial condition and making $\mathcal{F}$ zero coincide, namely (see Theorem 3.1),

$$y \text{ solves (1.1)} \quad \text{iff} \quad \mathcal{F}(y) = \min \mathcal{F} = 0 \ \text{ and } \ y(0) = y_0. \tag{1.2}$$

This variational characterization has a clear mechanical interpretation. Indeed, since ψ is positively 1-homogeneous, its conjugate ψ^* turns out to be the indicator function of the convex set $\partial\psi(0)$. Hence, $\min \mathcal{F}$ is actually a constrained minimization problem, and we have

$$\mathcal{F}(y) = 0 \quad \text{iff} \quad \begin{cases} \ell - Ay \in \partial\psi(0) \quad \text{a.e. in } \ (0,T), \\ \varphi(T, y(T)) + \displaystyle\int_0^T \psi(\dot{y}) = \varphi(0, y(0)) - \int_0^T \langle \dot{\ell}, y\rangle, \end{cases} \tag{1.3}$$

where we have used the notation $(t,y) \mapsto \varphi(t,y) = \frac{1}{2}\langle Ay, y\rangle - \langle \ell(t), y\rangle$. The first relation above expresses the so-called *local stability* [26] of the trajectory, whereas the second is nothing but the energy balance at time T. More precisely, $\varphi(t,y)$ denotes the complementary energy at time t for the state y, $\int_0^T \psi(\dot{y})$ represents the dissipation of the system on $[0,T]$, and $-\int_0^T \langle \dot{\ell}, y\rangle$ is related to external actions on $[0,T]$. Hence, minimizing $\mathcal{F}$ consists in selecting the (only) stable trajectory which conserves the energy. In this regards, the reader is referred to the pioneering papers by Moreau [33, 34, 35].

The interest of variational characterization (1.2) of the differential problem (1.1) relies on the possibility of exploiting the general tools from the calculus of variations. Some care is, however, required. Indeed, although $\mathcal{F}$ is convex and lower semicontinuous with respect to the weak topology of $W^{1,1}(0,T;Y)$, the functional generally fails to be coercive. Moreover, one is not just asked to minimize $\mathcal{F}$ but also to prove that the minimum is 0. These considerations suggest that the direct method is hardly applicable in order to get solutions to (1.1) via the characterization in (1.2).

The first issue of this paper is that of exploiting the variational principle in (1.2) in order to address general approximation procedures. Since solutions and minimizers coincide, a natural tool in order to frame an abstract approach to limiting procedures within (1.1) is that of considering the corresponding minimum problems via Γ-convergence [11]. As the value of the functional is directly quantified to be 0 on the minimizers, what is actually needed here for passing to limits are so-called Γ-lim inf inequalities only, and the latter are generally easily available. We shall specifically focus on the case of linear hardening elastoplasticity and apply the above-mentioned perspective in order to recover in a unified and more transparent frame and partly generalize some convergence results for conformal finite elements (Theorem 5.3), time discretizations (Theorem 6.5), and fully discrete space-time approximations (Theorem 7.1). In particular, for time discretization we develop a discrete version of the variational principle (1.2) in the same spirit of the theory of *variational integrators* [25] (see subsection 6.1). This connection entails also some generalized view at the classical discrete time schemes (see subsection 6.5).

A second novel point of the present variational approach consists in the possibility of exploiting $\mathcal{F}$ in order to estimate a posteriori some approximation error. By letting $\mathcal{F}(y) = 0$, we will check that (Corollary 4.5)

$$\max_{[0,T]} \frac{1}{2}\langle A(y-v), y-v\rangle \le \mathcal{F}(v) \quad \forall v \in W^{1,1}(0,T;Y), \quad v(0) = y_0.$$

If A shows some coercivity (which is precisely the case of linearized hardening; see subsection 2.4), and v is the outcome of some approximation procedure, the estimate above may serve as the basis for some a posteriori estimation procedure, possibly headed to adaptivity (see subsection 6.7). Let us stress that the latter and (1.3) entail that the distance of a (stable) trajectory from the solution to (1.1) can be uniformly estimated by means of its energy production along the path.

The variational characterization in (1.2) is strictly linked to the celebrated principle by Brezis, Ekeland, and Nayroles [3, 4, 38, 39] for the gradient flow of a convex functional $j : Y \to (-\infty, \infty]$ in a Hilbert space Y, namely,

$$\dot{y} + \partial j(y) \ni \ell \quad \text{a.e. in } (0,T). \tag{1.4}$$

In particular, by using the convexity of j (see details in section 3.1), this inclusion may be equivalently rewritten as the scalar relation

$$j(y) + j^*(\ell - \dot{y}) - \langle \ell - \dot{y}, y \rangle = 0 \quad \text{a.e. in } (0,T), \tag{1.5}$$

where j^* is the conjugate of j. As the above left-hand side is a.e. nonnegative for all trajectories y and it is 0 iff y solves (1.4), the latter is equivalent to minimize the global functional

$$\mathcal{I}(y) = \int_0^T \Big(j(y) + j^*(\ell - \dot{y}) - \langle \ell - \dot{y}, y \rangle \Big)$$

and check that $\mathcal{I}(y) = 0$. Since its introduction, the latter principle has continuously attracted attention. In particular, it has been exploited in the direction of proving the existence [40, 2, 41, 10, 9] (note that the above-mentioned obstructions to the application of the direct methods again appear) and the description of long-time dynamics [17]. Moreover, the Brezis–Ekeland–Nayroles approach has been adapted to the case of second order [21, 22] and doubly nonlinear equations [45] and to the study of variable time step discretizations of (1.4) [43] as well.

The variational characterization (1.2) stems basically from the same idea as in (1.5). Namely, the differential inclusion in (1.1) is equivalently rewritten by convexity of ψ as

$$\psi(\dot{y}) + \psi^*(\ell - Ay) - \langle \ell - Ay, \dot{y} \rangle = 0 \quad \text{a.e. in } (0,T),$$

and (1.2) follows by noting that the above left-hand side is a.e. nonnegative for all trajectories y and it is 0 iff y is fulfilling the inclusion in (1.1). On the other hand, apart from this conceptual analogy, one has indeed to mention that the characterization (1.2) has little in common with the original Brezis–Ekeland–Nayroles principle. In particular, the two principles turn out to be different even in the case of a quadratic functional ψ on a Hilbert space.

One has to mention that, of course, (1.2) is not the only possible global-in-time variational characterization of (1.1). Besides minimizing the L^2 space-time norm of the residual (which might be a little interesting since the order of the problem is doubled), one has at least to mention Visintin [48], where generalized solutions are obtained as minimal elements of a certain partial-order relation on the trajectories, and the recent contribution by Mielke and Ortiz [27], where the functional

$$y \mapsto e^{-T/\varepsilon} \varphi(T, y(T)) + \int_0^T e^{-t/\varepsilon} \left(\psi(\dot{y}) + \frac{1}{\varepsilon} \varphi(t, y) \right) \tag{1.6}$$

is minimized among trajectories with $y(0) = y_0$. Under extra-smoothness conditions on ψ (not fulfilled in the current frame), the Euler–Lagrange equations of the latter functional are

$$\begin{aligned} -\varepsilon \mathrm{D}^2\psi(\dot y)\ddot y + \mathrm{D}\psi(\dot y) + Ay = \ell,\\ y(0) = y_0, \qquad \mathrm{D}\psi(\dot y(T)) + Ay(T) = \ell(T). \end{aligned}$$

In particular, minimizing the functional in (1.6) consists in performing a suitable elliptic (in time) regularization of the problem. In the specific case of ψ positively 1-homogeneous, the limit $\varepsilon \to 0$ can be carried out, and the minimizers of the functional in (1.6) are proved to converge to the solution of (1.1). The latter approach is quite different from that of (1.2). On the one hand, it is much more general as it naturally applies to the nonsmooth case as well (no derivatives of ψ and ϕ are involved). The results of [27] have been recently extended in the direction of discretizations and relaxation in [31].

2. Mechanical model. Let us provide the reader with a brief introduction to the mechanical setting under consideration. Our aim is just that of recalling some essential features of the models as well as their variational formulation. In particular, we restrain from reporting here an extensive discussion on associative elastoplasticity as the latter can be easily recovered from the many contributions on the subject. The reader is particularly referred to the mentioned monographs for some comprehensive presentation.

2.1. Preliminaries. We will denote by $\mathbb{R}^{3\times 3}_{\mathrm{sym}}$ the space of symmetric 3×3 tensors endowed with the natural scalar product $a : b := \mathrm{tr}(ab) = a_{ij}b_{ij}$ (summation convention). The space $\mathbb{R}^{3\times 3}_{\mathrm{sym}}$ is orthogonally decomposed as $\mathbb{R}^{3\times 3}_{\mathrm{sym}} = \mathbb{R}^{3\times 3}_{\mathrm{dev}} \oplus \mathbb{R}\,1_2$, where $\mathbb{R}\,1_2$ is the subspace spanned by the identity 2-tensor 1_2 and $\mathbb{R}^{3\times 3}_{\mathrm{dev}}$ is the subspace of deviatoric symmetric 3×3 tensors. In particular, for all $a \in \mathbb{R}^{3\times 3}_{\mathrm{sym}}$, we have that $a = a_{\mathrm{dev}} + \mathrm{tr}(a)1_2/3$.

We shall assume the reference configuration Ω to be a nonempty, bounded, and connected open set in $\mathbb{R}^3$ with a Lipschitz continuous boundary. The space dimension 3 plays essentially no role throughout the analysis, and we would be in the position of reformulating our results in $\mathbb{R}^d$ with no particular intricacy. Our unknown variables are the displacement of the body $u \in \mathbb{R}^3$, the plastic strain $p \in \mathbb{R}^{3\times 3}_{\mathrm{dev}}$, and a vector of internal variables $\xi \in \mathbb{R}^m$ ($m \in \mathbb{N}$) which will describe the hardening of the material. We will denote by $\varepsilon(u)$ the standard symmetric gradient.

2.2. Constitutive relation. Moving within the small-strain regime, we additively decompose the linearized deformation $\varepsilon(u)$ into the elastic strain e and the inelastic (or plastic) strain p as

$$\varepsilon(u) = e + p.$$

Let $\mathbb{C}$ be the elasticity tensor. By regarding the latter as a symmetric positive definite linear map $\mathbb{C} : \mathbb{R}^{3\times 3}_{\mathrm{sym}} \to \mathbb{R}^{3\times 3}_{\mathrm{sym}}$, we shall assume that the orthogonal subspaces $\mathbb{R}^{3\times 3}_{\mathrm{dev}}$ and $\mathbb{R}\,1_2$ are invariant under $\mathbb{C}$. This amounts to saying that indeed

$$\mathbb{C}a = \mathbb{C}_{\mathrm{dev}}a_{\mathrm{dev}} + \kappa\,\mathrm{tr}(a)1_2$$

for a given $\mathbb{C}_{\mathrm{dev}} : \mathbb{R}^{3\times 3}_{\mathrm{dev}} \to \mathbb{R}^{3\times 3}_{\mathrm{dev}}$, a constant κ, and all $a \in \mathbb{R}^{3\times 3}_{\mathrm{sym}}$. The case of isotropic materials is given by $\mathbb{C}_{\mathrm{dev}} = 2G(1_4 - 1_2 \otimes 1_2/3)$, and G and κ are, respectively, the

shear and the bulk moduli. The latter decomposition is not exploited in our analysis, but it is clearly suggested by the mechanical application. Moreover, we shall introduce two linear symmetric positive semidefinite hardening moduli $\mathbb{H}_p : \mathbb{R}^{3\times3}_{\mathrm{dev}} \to \mathbb{R}^{3\times3}_{\mathrm{dev}}$ and $\mathbb{H}_\xi : \mathbb{R}^m \to \mathbb{R}^m$ (to be identified with a fourth order tensor and a matrix, respectively) and define the Helmholtz free energy $W : \mathbb{R}^{3\times3}_{\mathrm{sym}} \times \mathbb{R}^{3\times3}_{\mathrm{dev}} \times \mathbb{R}^m \to [0,\infty)$ of the material as

$$W(\varepsilon(u), p, \xi) := \frac{1}{2}(\varepsilon(u) - p) : \mathbb{C}(\varepsilon(u) - p) + \frac{1}{2}p : \mathbb{H}_p p + \frac{1}{2}\xi^T \cdot \mathbb{H}_\xi \xi.$$

The generalized stresses (σ, η) are conjugate to the above-defined generalized strains (e, ξ) via the energy W. In particular, the material is classically assumed to show elastic response,

$$\sigma = \frac{\partial W}{\partial e} = \mathbb{C}e = \mathbb{C}(\varepsilon(u) - p), \tag{2.1}$$

and the thermodynamic force η driving the evolution of the internal variables ξ is defined as

$$\eta = -\frac{\partial W}{\partial \xi} = -\mathbb{H}_\xi \xi. \tag{2.2}$$

Moreover, by moving within the frame of associative elastoplasticity, we assume the existence of a function $R : \mathbb{R}^{3\times3}_{\mathrm{dev}} \times \mathbb{R}^m \to [0,\infty]$ convex, positively 1-homogeneous, and lower semicontinuous such that

$$\partial R(\dot p, \dot\xi) \ni \begin{pmatrix} \sigma - \mathbb{H}_p p \\ \eta \end{pmatrix}. \tag{2.3}$$

In particular, R is asked to be the support function of a convex set $C^* \in \mathbb{R}^{3\times3}_{\mathrm{dev}} \times \mathbb{R}^m$, i.e., $R(p) = \sup_{q\in C^*} q : p$. We will indicate with R^* its conjugate, namely, the indicator function of C^* given by $R^*(q) = 0$ if $q \in C^*$ and $R^*(q) = \infty$ otherwise. Moreover, we let C be the domain of R, namely, $C = D(R) = \{(p,\xi) \in \mathbb{R}^{3\times3}_{\mathrm{dev}} \times \mathbb{R}^m : R(p,\xi) < \infty\}$.

Finally, the above material relations (2.1)–(2.3) can be condensed as the following constitutive material law:

$$\partial R(\dot p, \dot\xi) + \begin{pmatrix} (\mathbb{C} + \mathbb{H}_p)p \\ \mathbb{H}_\xi \xi \end{pmatrix} \ni \begin{pmatrix} \mathbb{C}\varepsilon(u) \\ 0 \end{pmatrix}, \tag{2.4}$$

which in turn can be rephrased in the form of (1.1) by letting

$$\begin{gathered} y = (p,\xi), \quad Y = \mathbb{R}^{3\times3}_{\mathrm{dev}} \times \mathbb{R}^m, \quad \psi = R, \\ A(p,\xi) = \big((\mathbb{C} + \mathbb{H}_p)p, \mathbb{H}_\xi \xi\big), \quad \ell = (\mathbb{C}\varepsilon(u), 0). \end{gathered} \tag{2.5}$$

Let us close this subsection by explicitly mentioning three classical linear hardening models [13, Ex. 4.8, p. 88]:

Linear kinematic hardening. Choose $\mathbb{H}_p = h_p 1_4$, where $h_p > 0$, and $\mathbb{H}_\xi = 0$. In this case the internal variable ξ is not evolving and shall be removed from the set of unknowns.

Linear isotropic hardening. Choose $\mathbb{H}_p = 0$, $m = 1$, and $\mathbb{H}_\xi = h_\xi > 0$. Moreover, let $D(R) = \{(p,\xi) \in \mathbb{R}^{3\times3}_{\mathrm{dev}} \times \mathbb{R} : |p| \le \xi\}$.

Linear combined kinematic-isotropic hardening. Let $\mathbb{H}_p = h_p 1_4$, $m = 1$, and $\mathbb{H}_\xi = h_\xi$, where $h_p, h_\xi > 0$. Moreover, let $D(R) = \{(p,\xi) \in \mathbb{R}^{3\times3}_{\mathrm{dev}} \times \mathbb{R} : |p| \le \xi\}$.

It is beyond the purpose of this introduction to discuss and justify the above-mentioned material models. The reader should check the cited references for comments on their relevance within applications and some mechanical motivation.

2.3. Variational formulation of the quasi-static evolution. Let us now move to the consideration of the full equilibrium problem. To this aim, we assume that the boundary $\partial\Omega$ is partitioned in two disjoint open sets $\Gamma_{\rm tr}$ and $\Gamma_{\rm Dir}$, with $\partial\Gamma_{\rm tr} = \partial\Gamma_{\rm Dir}$ (in $\partial\Omega$). We ask $\Gamma_{\rm Dir}$ to be such that there exists a positive constant $c_{\rm Korn}$ depending on $\Gamma_{\rm Dir}$ and Ω such that the Korn inequality

$$c_{\rm Korn}\|u\|^2_{H^1(\Omega;\mathbb{R}^3)} \le \|u\|^2_{L^2(\Gamma_{\rm Dir};\mathbb{R}^3)} + \|\varepsilon(u)\|^2_{L^2(\Omega;\mathbb{R}^{3\times 3}_{\rm sym})} \tag{2.6}$$

holds true for all $u \in H^1(\Omega;\mathbb{R}^3)$. It would indeed suffice to impose $\Gamma_{\rm Dir}$ to have a positive surface measure (see, e.g., [7, Thm. 3.1, p. 110]).

For the sake of simplicity, we will prescribe homogeneous Dirichlet boundary conditions on $\Gamma_{\rm Dir}$ (our analysis extends with little notational intricacy to the case of non-homogeneous Dirichlet boundary conditions as well). On $\Gamma_{\rm tr}$ some time-dependent traction will be prescribed instead.

As for the full quasi-static evolution of the material, we shall couple the constitutive relation (2.4) with the equilibrium equation

$$\operatorname{div}\sigma + f = 0 \quad \text{in } \ \Omega. \tag{2.7}$$

Here we assume to be given the body force $f : [0,T] \to L^2(\Omega;\mathbb{R}^3)$ and a surface traction $g : [0,T] \to L^2(\Gamma_{\rm tr};\mathbb{R}^3)$.

Then one can rephrase the problem into the form of (1.1) by choosing

$$y = (u,p,\xi), \tag{2.8}$$

$$Y = \Big\{(u,p,\xi) \in H^1(\Omega;\mathbb{R}^3) \times L^2(\Omega;\mathbb{R}^{3\times 3}_{\rm dev}) \times L^2(\Omega;\mathbb{R}^m) \tag{2.9}$$

$$\text{such that } \ u = 0 \ \text{ on } \ \Gamma_{\rm Dir}\Big\},$$

$$\langle A(u,p,\xi),(v,q,z)\rangle = \int_\Omega \Big((\varepsilon(u)-p) : \mathbb{C}(\varepsilon(v)-q) + p : \mathbb{H}_p q + \xi^T \cdot \mathbb{H}_\xi z\Big)$$

$$\forall (v,q,z) \in Y, \tag{2.10}$$

$$\psi(u,p,\xi) = \int_\Omega R(p,\xi) \tag{2.11}$$

and defining the total load $\ell : [0,T] \to Y^*$ as

$$\langle \ell(t),(u,p,\xi)\rangle = \int_\Omega f\cdot u + \int_{\Gamma_{\rm tr}} g\cdot u \ {\rm d}\mathcal{H}^2 \quad \forall u \in H^1(\Omega;\mathbb{R}^3),\ t\in[0,T],$$

where $\mathcal{H}^2$ is the 2-dimensional Hausdorff measure.

2.4. The coercivity of A. Let us close this introductory discussion by explicitly commenting on the coercivity of the bilinear form induced by A. We shall recall some sufficient conditions on $\mathbb{H}_p$, $\mathbb{H}_\xi$, and R in such a way that there exists a constant $\alpha > 0$ such that

$$\langle Ay, y\rangle \ge \alpha |y|^2 \quad \forall y \in D(\psi), \tag{2.12}$$

where $|\cdot|$ is the norm in Y. This issue is fairly classical [13, sec. 7.3, p. 167], and we discuss it here for the sake of completeness only.

Of course, (2.12) holds (and even for all $y \in Y$) whenever $\mathbb{H}_p$ and $\mathbb{H}_\xi$ are positive definite (this is the case of the above-mentioned linear combined kinematic-isotropic hardening).

As we have already observed, in the case $\mathbb{H}_\xi = 0$, the problem naturally reduces to the pair (u, p) only. Up to this reduction, (2.12) holds (again for all $y \in Y$) when $\mathbb{H}_p$ is positive definite. This is exactly the case of linear kinematic hardening.

On the other hand, in the case $\mathbb{H}_p = 0$, the plastic strain will still evolve, and one has (2.12) if $D(R)$ is bounded in the p-direction for all ξ, namely, if [13, eq. (7.51)]

(2.13)
$$D(R) \subset \{(p,\xi) \in \mathbb{R}^{3\times 3}_{\rm dev} \times \mathbb{R}^m \ : \ \beta|p|^2 \le \xi^T \cdot \mathbb{H}_\xi \xi \ \text{ for some constant } \ \beta > 0\},$$

which is clearly the case for linear isotropic hardening.

Some generalization of the latter condition could in principle be considered for the case when $\mathbb{H}_p$ and $\mathbb{H}_\xi$ are only semidefinite. In particular, (2.12) holds if one assumes (2.13) and

$$\xi \neq 0 \ \text{ and } \ \xi^T \cdot \mathbb{H}_\xi \xi = 0 \quad \Rightarrow \quad R(p,\xi) = \infty \ \ \forall p \in \mathbb{R}^{3\times 3}_{\rm dev}.$$

Let us mention that the most critical case in the class of (2.4) is $\mathbb{H}_p = 0$, $\mathbb{H}_\xi = 0$ where actually no hardening takes place. This is the situation of *perfect plasticity* for which the Sobolev space framework above is not appropriate, and one would consider the space $BD(\Omega)$ of functions of bounded deformations instead [6]. We shall make clear that, even if our variational characterization covers the case of perfect plasticity, the subsequent approximation results apply to the linear hardening situation only.

3. Characterization.

3.1. General assumptions. Let us start by recalling notation and enlisting the basic assumptions for the following. First of all, we will ask that

(3.1) $$Y \quad \text{is a separable and reflexive Banach space.}$$

We will use the symbols $|\cdot|$ for the norm of Y and $\langle\cdot,\cdot\rangle$ for the duality pairing between Y^* (dual) and Y. The norm in Y^* will be denoted by $|\cdot|_*$ instead.

We introduce the functional

(3.2) $$\psi : Y \to [0,\infty] \quad \text{proper, convex, lower semicontinuous, and positively 1-homogeneous.}$$

Equivalently, ψ is required to be the support function of a convex and closed set $C^* \subset Y^*$ containing 0, namely,

(3.3) $$\psi(y) = \sup\{\langle y^*, y\rangle \ : \ y^* \in C^*\}.$$

We shall define $C = D(\psi)$. Hence, the conjugate $\psi^* : Y^* \to [0,\infty]$, which is classically defined as $\psi^*(y^*) = \sup_{y\in Y}(\langle y^*, y\rangle - \psi(y))$, is the indicator function of the convex set C^*, namely, $\psi^*(y^*) = 0$ if $y^* \in C^*$ and $\psi^*(y^*) = \infty$ otherwise. Let us remark that ψ fulfills the triangle inequality $\psi(a) \le \psi(b) + \psi(c)$ whenever $a = b + c$.

We shall use the symbol ∂ in order to denote the usual subdifferential in the sense of convex analysis, namely,

$$y^* \in \partial\psi(y) \quad \text{iff} \quad y \in D(\psi) \ \text{ and } \ \langle y^*, w - y\rangle \le \psi(w) - \psi(y) \quad \forall w \in Y.$$

Similarly, we define

$$\begin{aligned} y \in \partial\psi^*(y^*) \quad &\text{iff} \quad y^* \in D(\psi^*) \ \text{ and } \ \langle w^* - y^*, y\rangle \le \psi^*(w^*) - \psi^*(y^*) \quad \forall w^* \in Y^* \\ &\text{iff} \quad y^* \in C^* \ \text{ and } \ \langle w^* - y^*, y\rangle \le 0 \quad \forall w^* \in C^*. \end{aligned}$$

Finally, we recall Fenchel's inequality

$$\psi(y) + \psi^*(y^*) \geq \langle y^*, y \rangle \quad \forall y \in Y,\ y^* \in Y^*,$$

and remark that equality holds iff $y^* \in \partial\psi(y)$ (or, equivalently, $y \in \partial^*\psi^*(y^*)$).

As for the operator A we require that

$$A : Y \to Y^* \quad \text{linear, continuous, and symmetric} \tag{3.4}$$

and define the function

$$y \to \phi(y) = \frac{1}{2}\langle Ay, y \rangle,$$

so that $A = \mathrm{D}\phi$. Moreover, we will ask ϕ to be coercive on $C = D(\psi)$; namely, we assume that there exists a positive constant α such that

$$\phi(y) \geq \frac{\alpha}{2}|y|^2 \quad \forall y \in C. \tag{3.5}$$

As we have already commented in subsection 2.4, the latter coercivity is fulfilled in the situation of elastoplastic evolution with linear kinematic, isotropic, or combined kinematic-isotropic hardening and will turn out to be sufficient for both of the forthcoming characterization results.

On the other hand, the following uniqueness-type results will be checked under some stronger coercivity frame, and we will ask for

$$\phi(y) \geq \frac{\alpha}{2}|y|^2 \quad \forall y \in C - C. \tag{3.6}$$

Clearly, condition (3.6) is fulfilled when ϕ happens to be coercive on the whole space Y. The latter applies in particular to the case of linear kinematic and combined kinematic-isotropic hardening elastoplasticity. In this case, ϕ defines an equivalent (squared) norm in Y.

We shall make use of the following notation:

$$\chi(y) = \phi(y) + |y|^2 \quad \forall y \in Y.$$

Indeed the latter choice is just motivated by simplicity and could be replaced as well by any other $\chi : Y \to [0, \infty)$ such that $\chi(y) = 0$ iff $y = 0$ and that $y \mapsto \chi(y) - \phi(y)$ is lower semicontinuous.

Finally, we shall fix data such that

$$\ell \in L^\infty(0, T; Y^*),\ y_0 \in C. \tag{3.7}$$

The restriction on the choice of the initial datum in C is motivated by the coercivity assumption on ϕ in (3.5). On the other hand, we shall explicitly mention that the usual choice for y_0 in elastoplasticity is $y_0 = 0$.

In what follows, the above assumptions (3.1)–(3.5) and (3.7) will be tacitly assumed (unless explicitly stated). It should be clear, however, that the above choice is motivated by the sake of simplicity. Indeed, most of the following results still hold under suitably weaker assumptions, as we shall comment.

3.2. The functional. Let the *Lagrangian* $L : (0,T)\times Y\times Y \to [0,\infty]$ be defined as

$$L(t,y,p) = \psi(p) + \psi^*(\ell(t) - Ay) - \langle \ell(t) - Ay, p\rangle \quad \text{for a.e. } t \in (0,T),\ \forall y,\ p \in Y, \tag{3.8}$$

and the functional $F : W^{1,1}(0,T;Y) \to [0,\infty]$ as

$$F(y) = \int_0^T L(t, y(t), \dot y(t))\ \mathrm{d}t + \chi(y(0) - y_0). \tag{3.9}$$

Now, by simply using the chain rule, we obtain

$$F(y) = \int_0^T \Big(\psi(\dot y) + \psi^*(\ell - Ay) - \langle \ell, \dot y\rangle\Big) + \phi(y(T)) - \phi(y(0)) + \chi(y(0) - y_0).$$

A first remark is that, by exploiting the particular form of χ,

$$F(y) = \int_0^T \Big(\psi(\dot y) + \psi^*(\ell - Ay) - \langle \ell, \dot y\rangle\Big) + \phi(y(T)) + \phi(y_0) - \langle Ay(0), y_0\rangle + |y(0) - y_0|^2. \tag{3.10}$$

In particular, F is clearly convex.

3.3. The characterization. Let us state here our variational principle.

THEOREM 3.1 (variational principle). *$y \in W^{1,1}(0,T;Y)$ solves (1.1) iff*

$$F(y) = 0 = \min F.$$

Proof. Owing to Fenchel's inequality we have

$$L(t,y,p) = 0 \quad \text{iff} \quad \ell(t) - Ay \in \partial\psi(p)$$

and, clearly, $\chi(y(0) - y_0) = 0$ iff $y(0) = y_0$. Hence, all solutions y of (1.1) are such that $F(y) = 0$ and vice versa. ∎

Let us remark that the latter variational characterization result holds in much greater generality. The proof made no use of the separability and reflexivity of Y nor of the linearity of A (besides its being single-valued and such that $t \mapsto Ay(t)$ is measurable). Moreover, the positive 1-homogeneity of ψ is unessential [44]. In particular, the variational approach of Theorem 3.1 can be directly extended to a variety of different dissipative systems possibly including viscous evolution as well. We shall address this perspective in a forthcoming contribution.

We have already observed that F is convex. Moreover, F is lower semicontinuous with respect to the weak topology of $W^{1,1}(0,T;Y)$, since all weakly convergent sequences in $W^{1,1}(0,T;Y)$ are pointwise weakly convergent as well. Hence, one could be tempted to use the direct method in order to get the existence of minimizers, i.e., solutions to (1.1). As we commented in the introduction, this seems to be no trivial task.

First of all, the functional F need not be coercive with respect to the weak topology of $W^{1,1}(0,T;Y)$. Indeed, the functional ψ may degenerate and hence not control the norm of its argument. Moreover, even in the case when ψ is nondegenerate, the homogeneity assumption just entails that the sublevels of F are bounded in $W^{1,1}(0,T;Y)$ and no weak compactness follows.

Second, even assuming coercivity in the weak topology of $W^{1,1}(0,T;Y)$, one would still need to prove that the minimum 0 is attained.

3.4. The variational principle for hardening elastoplasticity. By referring to the notations of section 2, let us now present the actual form of the functional F for the case of the constitutive relation for linearized elastoplastic materials with linear hardening (see (2.5)). In this case the functional reads

$$\begin{aligned}F(p,\xi) = &\int_0^T \Big(R(\dot p,\dot\xi) + R^*\big(\mathbb{C}(\varepsilon(u)-p) - \mathbb{H}_p p, -\mathbb{H}_\xi \xi\big)\Big)\\ &- \int_0^T \Big(\big(\mathbb{C}(\varepsilon(u)-p) - \mathbb{H}_p p\big) : \dot p - \xi^T \cdot \mathbb{H}_\xi \dot\xi\Big)\\ &+ \frac{1}{2}(p(0)-p_0) : (\mathbb{C}+\mathbb{H}_p)(p(0)-p_0) + \frac{1}{2}(\xi(0)-\xi_0)^T \cdot \mathbb{H}_\xi(\xi(0)-\xi_0)\\ &+ |(p(0),\xi(0)) - (p_0,\xi_0)|^2\end{aligned}$$

for some given initial datum $(p_0,\xi_0) \in \mathbb{R}^{3\times 3}_{\rm dev} \times \mathbb{R}^m$ and $\varepsilon(u) \in L^\infty(0,T;\mathbb{R}^{3\times 3}_{\rm sym})$.

In the situation of the quasi-static evolution, for some given initial datum $(u_0,p_0,\xi_0) \in Y$, a load $f \in L^\infty(0,T;L^2(\Omega;\mathbb{R}^3))$, and a traction $g \in L^\infty(0,T;L^2(\Gamma_{\rm tr};\mathbb{R}^3))$, the functional reads (see (2.8)–(2.11))

$$\begin{aligned}F(u,p,\xi) = &\int_0^T \int_\Omega \Big(R(\dot p,\dot\xi) + R^*\big(\mathbb{C}(\varepsilon(u)-p) - \mathbb{H}_p p, -\mathbb{H}_\xi \xi\big)\Big)\\ &- \int_0^T \int_\Omega f\cdot\dot u - \int_0^T \int_{\Gamma_{\rm tr}} g\cdot \dot u \,{\rm d}\mathcal{H}^2\\ &+ \int_0^T \int_\Omega \Big((\varepsilon(u)-p) : \mathbb{C}(\varepsilon(\dot u)-\dot p) + p : \mathbb{H}_p\dot p + \xi^T\cdot\mathbb{H}_\xi\dot\xi\Big)\\ &+ \frac{1}{2}\int_\Omega (\varepsilon(u(0)-u_0) - (p(0)-p_0)) : \mathbb{C}(\varepsilon(u(0)-u_0) - (p(0)-p_0))\\ &+ \frac{1}{2}\int_\Omega \Big((p(0)-p_0) : \mathbb{H}_p(p(0)-p_0) + (\xi(0)-\xi_0)^T\cdot\mathbb{H}_\xi(\xi(0)-\xi_0)\Big)\\ &+ \frac{1}{2}\int_\Omega |(u(0),p(0),\xi(0)) - (u_0,p_0,\xi_0)|^2\end{aligned}$$

for all points $(u,p,\xi) \in Y$ such that

$$\begin{aligned}&\int_\Omega (\varepsilon(u)-p) : \mathbb{C}\varepsilon(v) = \int_\Omega f\cdot v - \int_{\Gamma_{\rm tr}} g\cdot v\,{\rm d}\mathcal{H}^2\\ &\forall v \in H^1(\Omega;\mathbb{R}^3), \quad \text{with} \quad v = 0 \quad \text{on} \quad \Gamma_{\rm Dir}, \quad \text{a.e. in} \quad (0,T),\end{aligned}$$

and $F(u,p,\xi) = \infty$ otherwise.

4. Properties of the minimizers. For the sake of illustrating the variational principle of Theorem 3.1, we shall collect here some properties of the trajectories belonging to the domain of the functional F and, in particular, of the minimizers.

4.1. Trajectories are in C.

LEMMA 4.1. *Let $F(y) < \infty$. Then $y(t) \in C$ for all $t \in [0,T]$.*

Proof. Since $F(y) < \infty$ we have $\dot y \in C$ a.e. in $(0,T)$. Hence, for all $t \in [0,T]$, we have $\int_0^t \dot y \in C$ by Jensen's inequality. On the other hand, $y_0 \in C$ and $y(t) = y_0 + \int_0^t \dot y$. The assertion follows by recalling that C is a cone. ∎

4.2. Stability at regular points of ℓ. Assume that $\ell : [0,T] \to Y^*$ is given, and let the set of *stable* states $S(t) \subset Y$ for $t \in [0,T]$ be defined as

$$S(t) = \{y \in Y \ : \ \phi(y) - \langle \ell(t), y\rangle \le \phi(w) - \langle \ell(t), w\rangle + \psi(w-y) \ \ \forall w \in Y\}.$$

In particular, one has $y \in S(t)$ iff it minimizes the convex functional $G : w \mapsto \phi(w) - \langle \ell(t), w\rangle + \psi(w-y)$, namely,

$$y \in S(t) \quad \text{iff} \quad 0 \in \partial G(y) = Ay - \ell(t) + \partial\psi(0).$$

Let us now check that indeed $\partial\psi(0) = C^*$. To this end, owing to (3.3), given $y^* \in C^*$ one has $\langle y^*, w\rangle \le \psi(w)$ for all $w \in Y$. Namely, $C^* \subset \partial\psi(0)$. On the other hand, assume that there exists $\xi \in \partial\psi(0) \setminus C^*$. Hence, by the Hahn–Banach theorem one finds $y \in Y$, $\alpha \in \mathbb{R}$, and $\varepsilon > 0$ such that

$$\psi(y) \overset{\xi\in\partial\psi(0)}{\ge} \langle \xi, y\rangle \ge \alpha + \varepsilon > \alpha - \varepsilon \ge \langle y^*, y\rangle \quad \forall y^* \in C^*,$$

and, by passing to the supremum as $y^* \in C^*$, we obtain a contradiction. We may summarize this discussion as follows:

$$y \in S(t) \quad \text{iff} \quad \ell(t) - Ay \in C^*.$$

LEMMA 4.2 (stability of the minimizers). *Let ℓ be either left- or right-weakly continuous at some point $t \in [0,T]$ and $F(y) < \infty$. Then $y(t) \in S(t)$.*

Proof. Since $F(y) < \infty$, we have $\ell(t) - Ay(t) \in C^*$ for all $t \in (0,T) \setminus N$, where $|N| = 0$. Choose a sequence $t_k \in (0,T) \setminus N$ such that $t_k \to t$ (from the left or from the right) and $\ell(t_k) \to \ell(t)$ weakly in Y^*. Hence, $\ell(t_k) - Ay(t_k) \to \ell(t) - Ay(t)$ weakly in Y^* and $\ell(t) - Ay(t) \in C^*$. ∎

In particular, if ℓ happens to be right-continuous at 0, the functional F will not attain the minimum value 0 unless the initial datum y_0 is stable, namely, $y_0 \in S(0)$.

4.3. Equivalent formulations. Letting now $\ell \in W^{1,1}(0,T;Y^*)$, problem (1.1) admits some alternative equivalent formulations [26, sec. 2.1]. We explicitly mention that $y \in W^{1,1}(0,T;Y)$ is said to be an *energetic solution* if it solves the *energetic formulation* [32] of (1.1), namely,

$$y(t) \in S(t) \quad \forall t \in [0,T], \tag{4.1}$$

$$\phi(y(t)) - \langle \ell(t), y(t)\rangle + \int_0^t \psi(\dot y) = \phi(y(0)) - \langle \ell(0), y(0)\rangle - \int_0^t \langle \dot\ell, y\rangle \tag{4.2}$$

$$\forall t \in [0,T],$$

$$y(0) = y_0. \tag{4.3}$$

Mielke and Theil [32] proved that the latter is equivalent to (1.1) and hence, owing to the characterization of Theorem 3.1, to $F(y) = 0 = \min F$ (note that the analysis in [32] is much more general and is in particular allowing discontinuous in time evolutions by introducing the above notion of energetic solutions in the frame of functions of a bounded variation). For the aim of pointing out some features of our variational approach, we shall present here a direct proof of this fact.

LEMMA 4.3 (equivalence with the energetic formulation). *Let $\ell \in W^{1,1}(0,T;Y^*)$. Then $F(y) = 0 = \min F$ iff y fulfills* (4.1)–(4.3).

Proof. Owing to Lemma 4.2, we readily have that the stability condition (4.1) holds iff $\psi^*(\ell - Ay) = 0$ a.e.

Let y be such that $F(y)=0$. Then (4.1) and (4.3) hold and $L(t,y(t),\dot y(t))=0$ for a.e. $t\in(0,T)$. In particular, for all $t\in[0,T]$,

$$\begin{aligned}0=\int_0^t L(s,y(s),\dot y(s))\,\mathrm{d}s&=\int_0^t\Big(\psi(\dot y)-\langle\ell,\dot y\rangle\Big)+\phi(y)\Big|_0^t\\&=\big(\phi(y)-\langle\ell,y\rangle\big)\Big|_0^t+\int_0^t\psi(\dot y)+\int_0^t\langle\dot\ell,y\rangle,\end{aligned}$$

so that the energy equality (4.2) holds for all $t\in[0,T]$.

On the contrary, let $y\in W^{1,1}(0,T;Y)$ fulfill (4.1)–(4.3). Then $\chi(y(0)-y_0)=0$ and $\psi^*(\ell-Ay)=0$ a.e. (see above). Hence $F(y)=0$ follows from the energy equality (4.2) at time T and an integration by parts. □

Let us mention that the last lemma proves in particular that the energy equality (4.2) could be equivalently enforced *at the final time* T only. Moreover, it proves that, as already commented in the introduction, all *stable trajectories* $t\mapsto y(t)$ (i.e., trajectories such that $y(t)\in S(t)$ for all $t\in[0,T]$) are such that the following energy inequality holds:

$$\phi(y(t))-\langle\ell(t),y(t)\rangle+\int_0^t\psi(\dot y)\ge\phi(y(0))-\langle\ell(0),y(0)\rangle-\int_0^t\langle\dot\ell,y\rangle\quad\forall t\in[0,T].$$

Hence, we have provided a proof to [26, Prop. 5.7] (note that the referred result is, however, more general as it is concerned with the BV situation, the energy is implicitly depending on time, and no linear structure on Y is required).

Before closing this subsection, let us explicitly remark that the above inferred equivalence between formulations has been obtained for the absolutely continuous case $\ell\in W^{1,1}(0,T;Y^*)$ only, whereas the characterization of Theorem 3.1 holds more generally for bounded ℓ.

4.4. The functional controls the uniform distance: Uniqueness. So far, we have simply reformulated known results in a variational fashion. Here we present some novel results instead.

LEMMA 4.4 (uniform distance control via F). *We have*

$$\eta(1-\eta)\max_{t\in[0,T]}\phi\big(u(t)-v(t)\big)\le\eta F(u)+(1-\eta)F(v)\tag{4.4}$$
$$\forall u,\,v\in W^{1,1}(0,T;Y),\ \eta\in[0,1].$$

Proof. The statement follows from the quadratic character of ϕ. Fix $t\in[0,T]$, and define $F^t:W^{1,1}(0,t;Y)\to[0,\infty]$ as

$$\begin{aligned}F^t(y)&=\int_0^t L(s,y(s),\dot y(s))\,\mathrm{d}s+\chi(y(0)-y_0)\\&=\int_0^t\Big(\psi(\dot y)+\psi^*(\ell-Ay)-\langle\ell,\dot y\rangle\Big)\\&\quad+\phi(y(t))+\phi(y_0)-\langle Ay(0),y_0\rangle+|y(0)-y_0|^2.\end{aligned}$$

Then, clearly, $y\mapsto G^t(y)=F^t(y)-\phi(y(t))$ is convex. Hence, by letting $w=\eta u+(1-\eta)v$ we have

$$\begin{aligned}&0\le F^t(w)\\&\le\eta\big(G^t(u)+\phi(u(t))\big)+(1-\eta)\big(G^t(v)+\phi(v(t))\big)-\eta(1-\eta)\phi\big(u(t)-v(t)\big),\end{aligned}$$

whence the assertion follows. □

The latter lemma exploits the quadratic character of ϕ only. In particular, no coercivity for ϕ is assumed. It should, however, be clear that its application (as the title of the lemma indeed suggests) will always be referred to the situation where the stronger (3.6) is required, namely, when the left-hand side of (4.4) controls

$$\eta(1-\eta)\frac{\alpha}{2}\max_{[0,T]}|u-v|^2.$$

We now present two immediate corollaries of Lemma 4.4.

COROLLARY 4.5 (uniform distance from the minimizer). *Let $F(y)=0$. Then*

$$\max_{t\in[0,T]}\phi\big(y(t)-v(t)\big)\le F(v).$$

This corollary encodes an interesting novel feature of our variational approach, for it provides a possible a posteriori error estimator to be used within approximation procedures. It is interesting to remark that the uniform distance of any stable trajectory from the minimizer is controlled by means of its energy production along the path only. Again, although Corollary 4.5 holds under no coercivity assumptions of ϕ, let us mention that its application will be restricted to the frame of (3.6). Finally, we have uniqueness of the minimizers of F attaining the value 0.

COROLLARY 4.6 (uniqueness). *Assume* (3.6). *Then there exists at most one trajectory y such that $F(y)=0$.*

4.5. Lipschitz bound. Throughout the remainder of the paper we shall tacitly assume that

$$\ell\in W^{1,\infty}(0,T;Y^*),\quad y_0\in S(0). \tag{4.5}$$

As already commented after Lemma 4.2, the above restriction on the initial datum is mandatory whenever ℓ admits a weak right limit in 0.

As for ℓ, the extra Lipschitz continuity assumption is motivated by the rate independence of the problem (every absolutely continuous datum can be time-rescaled to a Lipschitz continuous datum) and the following well known result.

LEMMA 4.7 (Lipschitz bound). *Assume* (3.6), *and let $\ell\in W^{1,\infty}(0,T;Y^*)$ and $F(y)=0$. Then*

$$\|\dot y\|_{L^\infty(0,T;Y)}\le\frac{1}{\alpha}\|\dot\ell\|_{L^\infty(0,T;Y^*)}\quad \textit{a.e. in}\ \ (0,T). \tag{4.6}$$

The proof of the lemma is exactly the classical one [32, Thm. 7.5] but formulated by means of our variational arguments. We provide it for the sake of completeness.

Proof. Let $0\le s<t\le T$ be fixed. Since $L(y,\dot y)=0$ a.e. we have

$$\int_s^t\Big(\psi(\dot y)+\langle\dot\ell,y\rangle\Big)+\big(\phi(y)-\langle\ell,y\rangle\big)\Big|_s^t=0.$$

On the other hand, owing to the strong monotonicity of A and the fact that $y(s)\in S(s)$ (see Lemma 4.2), one obtains

$$\begin{aligned}\phi(y(t)-y(s))&\le\phi(y(t))-\langle\ell(s),y(t)\rangle+\psi(y(t)-y(s))-\phi(y(s))-\langle\ell(s),y(s)\rangle\\&=\big(\phi(y)-\langle\ell,u\rangle\big)\Big|_s^t+\int_s^t\langle\dot\ell(r),y(t)\rangle\,\mathrm{d}r+\psi(y(t)-y(s)).\end{aligned}$$

By taking the sum of these two relations and recalling that, by Jensen,

$$\psi(y(t)-y(s)) = \psi\left(\int_s^t \dot{y}\right) \le \int_s^t \psi(\dot{y}),$$

we get

$$\phi(y(t)-y(s)) \le \int_s^t \langle \dot{\ell}(r), y(t)-y(r)\rangle \,\mathrm{d}r.$$

Finally, an application of some extended Gronwall lemma (see [26, Thm. 3.4]) entails that

$$\frac{\alpha}{2}|y(t)-y(s)|^2 \le \frac{1}{2}\|\dot{\ell}\|_{L^\infty(0,T;Y^*)}|y(t)-y(s)|\,(t-s),$$

and the assertion follows. ∎

5. Space approximation and stability under data perturbation. We now apply the characterization results of Theorem 3.1 to the approximation of solutions of (1.1). As already commented in the introduction, we shall proceed via Γ-convergence [11]. The reader is referred to the monographs by Attouch [1] and Dal Maso [5] for some comprehensive discussion on this topic. Indeed, since Theorem 3.1 directly quantifies the value of the minimum to be 0, what is actually needed for passing to limits are Γ-lim inf inequalities only. We shall illustrate this fact by discussing the simple case of stability under data perturbation first.

LEMMA 5.1 (stability under data perturbation). *Assume* (3.6), *and let* $\ell_h \to \ell$ *strongly in* $L^1(0,T;Y^*)$ *being uniformly Lipschitz continuous and* $y_{0,h} \to y_0$. *Moreover, let* $F_h : W^{1,1}(0,T;Y) \to [0,\infty]$ *be defined as*

$$F_h(y) = \int_0^T \Big(\psi(\dot{y}) + \psi^*(\ell_h - Ay) - \langle \ell_h - Ay, \dot{y}\rangle\Big) + \chi(y(0)-y_{0,h}),$$

and let $F_h(y_h) = 0$. *Then* $y_h \to y$ *weakly star in* $W^{1,\infty}(0,T;Y)$ *and* $F(y) = 0$.

Proof. Owing to Lemma 4.7, we find a (not relabeled) subsequence y_h such that $y_h \to y$ weakly star in $W^{1,\infty}(0,T;Y)$. Hence, we have by lower semicontinuity

$$\begin{aligned} 0 \le F(y) &\le \liminf_{h\to 0} \left(\int_0^T \Big(\psi(\dot{y}_h) + \psi^*(\ell_h - Ay_h) - \langle \ell_h, \dot{y}_h\rangle\Big) \right. \\ &\qquad \left. + \phi(y_h(T)) + \phi(y_{0,n}) - \langle Ay_h(0), y_{0,h}\rangle + |y_h(0)-y_{0,h}|^2 \right) \\ &= \liminf_{h\to 0} F_h(y_h) = 0. \end{aligned}$$

Hence, $F(y) = 0$, y is unique, and the assertion follows from the fact that the whole sequence converges. ∎

5.1. Preliminaries on functional convergence. In order to move to more general approximation situations, we are forced to discuss a suitable functional convergence notion. We limit ourselves in introducing the relevant definitions, referring to the mentioned monographs for all of the necessary details.

Recall that Y is a real reflexive Banach space. Letting f_n, $f : Y \to (-\infty, \infty]$ be convex, proper, and lower semicontinuous, we say that $f_n \to f$ in the *Mosco sense in* Y [1, 36] iff, for all $y \in Y$,

$$f(y) \le \liminf_{n\to\infty} f_n(y_n) \quad \forall y_n \to y \ \text{ weakly in } \ Y,$$
$$\exists y_n \to y \ \text{ strongly in } \ Y \ \text{ such that } \ f(y) = \limsup_{n\to\infty} f_n(y_n).$$

In particular, $f_n \to f$ in the Mosco sense iff $f_n \to f$ in the sense of Γ-convergence with respect to both the weak and the strong topology in Y.

We will consider the situation of approximating functionals ψ_h. By [1, Thm. 3.18, p. 295], we have $\psi_h \to \psi$ in the Mosco sense in Y iff $\psi_h^* \to \psi^*$ in the Mosco sense in Y^*. By assuming the functionals ψ_h to be positively 1-homogeneous, it turns out that the Mosco convergence $\psi_h \to \psi$ in Y is equivalent to the Mosco convergence of sets $C_h^* \to C^*$ in Y^*, which reads

$$C_n^* \ni y_n^* \to y^* \quad \text{weakly in } Y^* \quad \Rightarrow \quad y^* \in C^*,$$
$$\forall y^* \in C^*, \quad \exists y_n^* \in C_n^* \ : \ y_n^* \to y^* \quad \text{strongly in } Y^*.$$

Finally, we repeatedly use a lemma from [45] which we report it here, for the sake of completeness.

LEMMA 5.2 ([45, Cor. 4.4]). *Let $p \in [1, \infty]$ and f_h, $f : Y \to (-\infty, \infty]$ be convex, proper, and lower semicontinuous such that*

$$f(y) \le \inf \left\{ \liminf_{h\to 0} f_h(y_h) \ : \ y_h \to y \ \text{ weakly in } \ Y \right\} \quad \forall y \in Y.$$

Moreover, let $y_h \to y$ weakly in $W^{1,p}(0, T; Y)$ (weakly star if $p = \infty$). Then we have

$$\int_0^T f(y(t))\ \mathrm{d}t \le \liminf_{h\to 0} \int_0^T f_h(y_h(t))\ \mathrm{d}t.$$

5.2. Space approximations. We now move to the analysis of some space approximation situation, indeed specifically tailored for the case of conformal finite elements. Let us list here our assumptions for the sake of later referencing.

We assume to be given

$$Y_h \subset Y \ \text{ closed subspaces such that } \ h \mapsto Y_h \ \text{ increases}$$
(5.1) $$\text{and } \ \cup_{h>0} Y_h \ \text{ is dense in } \ Y,$$
(5.2) $$\phi_h(y) = \phi(y) \ \text{ if } \ y \in Y_h \ \text{ and } \ \phi_h(y) = \infty \ \text{ otherwise.}$$
(5.3) $$\psi_h : Y \to (-\infty, \infty] \quad \text{convex, proper, and lower semicontinuous,}$$
(5.4) $$\psi_h \quad \text{positively 1-homogeneous,}$$
(5.5) $$\psi_h \to \psi \quad \text{in the Mosco sense in } Y,$$
(5.6) $$\phi(y) \ge \frac{\alpha}{2}|y|^2 \quad \forall y \in C_h - C_h \quad \text{where } \ C_h = D(\psi_h),$$
(5.7) $$\ell_h \to \ell \quad \text{pointwise strongly in } Y^*,$$
(5.8) $$\ell_h \quad \text{uniformly Lipschitz continuous,}$$
(5.9) $$y_{0,h} \in Y_h, \quad y_{0,h} \to y_0.$$

We shall mention that within the frame of conformal finite elements methods the subspaces Y_h are obviously taken to be finite-dimensional and that the approximating functionals ϕ_h and ψ_h are usually the restrictions of the functionals ϕ and ψ on the subspace Y_h. This is exactly our choice here for ϕ_h. In particular, one shall observe that $\phi_h \to \phi$ in the Mosco sense in Y, $D(\partial\phi_h) = Y_h$, and that

$$A_h y = \partial\phi_h(y) = \partial\phi(y) = D\phi(y) = Ay \quad \forall y \in Y_h. \tag{5.10}$$

As for ψ_h we are allowing some extra freedom (let us remark, however, that (5.6) follows from (3.6) as soon as ψ_h is the restriction of ψ to Y_h since, in this case, $C_h = C \cap Y_h$). On the other hand, we are asking ψ_h to be positively 1-homogeneous; namely, we are considering the case of some rate-independent approximation of (1.1) only. The reader is referred instead to Efendiev and Mielke [8], Mielke, Rossi, and Savaré [29, 28], Toader and Zanini [47], and Zanini [49] for some results in the direction of rate-dependent approximation of rate-independent processes.

Finally, we shall (re)define the approximating functionals as $F_h : W^{1,1}(0,T;Y) \to [0,\infty]$ as

$$F_h(y) = \int_0^T \Big(\psi_h(\dot y) + \psi_h^*(\ell_h - A_h y) - \langle \ell_h - A_h y, \dot y\rangle\Big) + \chi_h(y(0) - y_{0,h}),$$

where $A_h = \partial\phi_h$ and $\chi_h(\cdot) = \phi_h(\cdot) + |\cdot|^2$. We have the following.

THEOREM 5.3 (convergence of space approximations). *Assume* (5.1)–(5.9)*, and let* $F_h(y_h) = 0$*. Then* $y_h \to y$ *weakly star in* $W^{1,\infty}(0,T;Y)$ *and* $F(y) = 0$*.*

Proof. By Lemma 4.7, we find a (not relabeled) subsequence $y_n \to y$ weakly star in $W^{1,\infty}(0,T;Y)$ and weakly pointwise. Since $F_h(y_h) = 0$ we readily check that $y(t) \in Y_h$ for all $t \in [0,T]$. In particular, $A_h y_h = A y_h$ for all $t \in [0,T]$ owing to (5.10). Hence, by lower semicontinuity,

$$\begin{aligned}
0 \le F(y) &\le \liminf_{h\to 0} \bigg(\int_0^T \Big(\psi_h(\dot y_h) + \psi_h^*(\ell_h - A y_h) - \langle \ell_h, \dot y_h\rangle\Big) \\
&\qquad + \phi(y_h(T)) + \phi(y_{0,h}) - \langle A y_h(0), y_{0,h}\rangle + |y_h(0) - y_{0,h}|^2 \bigg) \\
&= \liminf_{h\to 0} \bigg(\int_0^T \Big(\psi_h(\dot y_h) + \psi_h^*(\ell_h - A_h y_h) - \langle \ell_h, \dot y_h\rangle\Big) \\
&\qquad + \phi_h(y_h(T)) - \phi_h(y_h(0)) \bigg) \\
&= \liminf_{h\to 0} F_h(y_h) = 0.
\end{aligned}$$

Note that the integral terms containing ψ and ψ^* pass to the lim inf by means of Lemma 5.2. ∎

By inspecting the proof of Theorem 5.3 (which of course generalizes Lemma 5.1), one realizes that, whenever the weak-star precompactness in $W^{1,\infty}(0,T;Y)$ of the sequence y_h is assumed, the convergence statement holds more generally in the case $F_h(y_h) \to 0$. Namely, by directly asking for the above-mentioned compactness, one could consider the convergence of some approximated solutions y_h such that, possibly, $F_h(y_h) > 0$. We rephrase this fact in the following statement.

LEMMA 5.4 (Γ-lim inf inequality for F_h). *Assume* (5.1)–(5.3), (5.5)–(5.7), *and* (5.9). *Then*

$$F(u) \le \inf\left\{ \liminf_{h\to 0} F_h(y_h) \ : \ y_h \to y \quad \textit{weakly star in} \ \ W^{1,\infty}(0,T;Y) \right\}.$$

Note that the homogeneity of ψ_h, the uniform convexity of ϕ_h, and the Lipschitz continuity of ℓ_h play no role here.

Finally, again by looking carefully to the proof of Theorem 5.3, one could wonder if the requirement on the Mosco convergence of ψ_h could be weakened. Indeed, what we are actually using is only that

$$\psi \le \Gamma\text{-}\liminf_{h\to 0} \psi_h \quad \text{and} \quad \psi^* \le \Gamma\text{-}\liminf_{h\to 0} \psi_h^* \tag{5.11}$$

with respect to the weak topologies of Y and Y^*, respectively. On the other hand, in our specific situation, [45, Lem. 4.1] entails that (5.11) and the fact that $\psi_h \to \psi$ in the Mosco sense in Y are equivalent.

This observation motivates once again the belief that Mosco convergence is the right frame in order to pass to limits within rate-independent problems. For the sake of completeness, let us recall that a first result in the direction of the approximation of the play operator (Y Hilbert and A coercive on Y) under the Hausdorff convergence of the characteristic sets $C_h^* = D(\psi_h^*)$ is contained in [16, Thm. 3.12, p. 34], whereas the extension of this result to the more general situation of Mosco converging sets as well as some application to parabolic PDEs with hysteresis is discussed in [46]. More recently, Mielke, Roubíček, and Stefanelli [30] addressed in full generality the issue of Γ-convergence and relaxation for the energetic solutions of rate-independent processes. An alternative convergence proof in the specific case of convex energies is obtained by means of the Brezis–Ekeland–Nayroles approach in [45].

6. Time discretization. Assume now that we are given the partitions $P_n = \{0 = t_n^0 < t_n^1 < \cdots < t_n^{N_n} = T\}$, and denote by $\tau_n^i = t_n^i - t_n^{i-1}$ the ith time step and by $\tau_n = \max_{1\le i\le N_n} \tau_n^i$ the diameter of the nth partition. No constraints are imposed on the possible choice of the time steps throughout this analysis besides $\tau_n \to 0$ as $n \to \infty$. Moreover, let the parameter $\theta \in [1/2, 1]$ be given.

In the following we will make extensive use of the following notation: Letting $v = (v^0, \dots, v^{N_n})$ be a vector, we will denote by $\widehat{v}_n$ and $\overline{v}_n$ two functions of the time interval $[0,T]$ which interpolate the values of the vector v piecewise linearly and backward constantly on the partition P_n, respectively. Namely,

$$\begin{aligned} \widehat{v}_n(0) = v^0, \ \ \widehat{v}_n(t) = \gamma_n^i(t) v^i + \big(1 - \gamma_n^i(t)\big) v^{i-1}, \\ \overline{v}_n(0) = v^0, \ \ \overline{v}_n(t) = v^i, \quad \text{for} \ \ t \in (t_n^{i-1}, t_n^i], \ \ i = 1, \dots, N_n \end{aligned}$$

where

$$\gamma_n^i(t) = (t - t_n^{i-1})/\tau_n^i \quad \text{for} \ \ t \in (t_n^{i-1}, t_n^i], \ \ i = 1, \dots, N_n.$$

Moreover, we let $\delta v^i = (v^i - v^{i-1})/\tau_n^i$ for $i = 1, \dots, N_n$ (so that $\dot{\widehat{v}}_n = \overline{\delta v_n}$) and denote by v_θ the vector with components $v_\theta^i = \theta v^i + (1-\theta) v^{i-1}$.

Recall that $\ell \in W^{1,\infty}(0,T;Y^*)$ and $y_0 \in S(0)$. We shall be concerned with the so-called θ-scheme for problem (1.1):

$$\partial\psi\left(\frac{y_n^i - y_n^{i-1}}{\tau_n^i}\right) + A(\theta y_n^i + (1-\theta)y_n^{i-1}) \ni \ell(\theta t_n^i + (1-\theta)t_n^{i-1})$$

$$\text{for } \ i = 1, \dots, N_n, \tag{6.1}$$

$$y_n^0 = y_0. \tag{6.2}$$

One usually refers to the latter as the backward or implicit Euler scheme for the choice $\theta = 1$ and as the Crank–Nicolson scheme for $\theta = 1/2$.

Owing to the above-introduced notation, the latter scheme can be equivalently rewritten as

$$\partial\psi\left(y_n^i - y_n^{i-1}\right) + Ay_{n,\theta}^i \ni \ell(t_{n,\theta}^i) \quad \text{for } \ i = 1, \dots, N_n, \quad y_n^0 = y_0. \tag{6.3}$$

Clearly, the θ-scheme (6.3) is rate-independent. Namely, no time step appears in (6.3), and the choice of the partition affects the solution via the values of the load ℓ only. In this concern, our focus on variable time-step partition could be simplified by considering proper rescaled loads ℓ instead. We shall, however, keep up with it, especially in order to underline the possibility of adapting the partition according to some a posteriori analysis (see subsection 6.7).

Before moving on, let us comment that, for all n, the latter scheme is a unique solution. Indeed, given $y_n^{i-1} \in C$, it suffices to (uniquely) solve iteratively the incremental problem

$$y_n^i \in \operatorname{Arg\,min}_{y\in Y}\Big(\theta\phi(y) - \langle \ell(t_{n,\theta}^i) - (1-\theta)Ay_n^{i-1}, y\rangle + \psi(y - y_n^{i-1})\Big). \tag{6.4}$$

Note that, since $y_n^{i-1} \in C$, the functional under minimization turns out to be uniformly convex. Hence, by (3.5), the minimum problem has a unique solution. In particular, exactly as in Lemma 4.1 we have the following.

LEMMA 6.1. *$y_n^i \in C$ for all $i = 0, 1, \dots, N_n$.*

A crucial observation is that, as in the continuous case, the discrete trajectories show some sort of stability as well.

LEMMA 6.2 (stability of the discrete trajectories). *We have*

$$y_n^i \in \operatorname{Arg\,min}_{y\in Y}\Big(\theta\phi(y) - \langle \ell(t_{n,\theta}^i) - (1-\theta)Ay_n^{i-1}, y\rangle + \psi(y - y_n^i)\Big)$$

$$\text{for } \ i = 1, \dots, N_n. \tag{6.5}$$

In particular, if $\theta = 1$, then $y_n^i \in S(t_n^i)$.

Proof. From the incremental formulation (6.4) and the triangle inequality for ψ, we get, for all $y \in Y$,

$$\begin{aligned}
&\theta\phi(y_n^i) - \langle \ell(t_{n,\theta}^i) - (1-\theta)Ay_n^{i-1}, y_n^i\rangle + \psi(y_n^i - y_n^{i-1}) \\
&\quad \le \theta\phi(y) - \langle \ell(t_{n,\theta}^i) - (1-\theta)Ay_n^{i-1}, y\rangle + \psi(y - y_n^{i-1}) \\
&\quad \le \theta\phi(y) - \langle \ell(t_{n,\theta}^i) - (1-\theta)Ay_n^{i-1}, y\rangle + \psi(y - y_n^i) + \psi(y_n^i - y_n^{i-1}),
\end{aligned}$$

whence the assertion follows. ∎

Again as in the continuous case, we readily check that

$$\text{(6.5) holds iff} \quad \ell(t_{n,\theta}^i) - Ay_{n,\theta}^i \in C^*. \tag{6.6}$$

6.1. The discrete variational principle. We shall now present a discrete version of the variational principle of Theorem 3.1.

We define $L_n^{\theta,i}(y,z) : Y \times Y \to [0,\infty]$ as

$$L_n^{\theta,i}(y,z) = \psi\left(\frac{y-z}{\tau_n^i}\right) + \psi^*\left(\ell(t_{n,\theta}^i) - A(\theta y + (1-\theta)z)\right) - \left\langle \ell(t_{n,\theta}^i) - A(\theta y + (1-\theta)z), \frac{y-z}{\tau_n^i}\right\rangle$$

and the functionals $F_n^\theta : Y^{N_n+1} \to [0,\infty]$ as

$$F_n^\theta(y_n^0,\dots,y_n^{N_n}) = \sum_{i=1}^{N_n} \tau_n^i L_n^{\theta,i}(y_n^i, y_n^{i-1}) + \chi(y_n^0 - y_0).$$

LEMMA 6.3 (discrete variational principle). $(y_n^0,\dots,y_n^{N_n})$ *solves* (6.3) *iff* $F_n^\theta(y_n^0, \dots, y_n^{N_n}) = 0 = \min F_n^\theta$.

Proof. Analogously to the continuous case, we have, for all $i = 1,\dots,N_n$,

$$\partial\psi\left(\delta y_n^i\right) + A y_{n,\theta}^i \ni \ell(t_{n,\theta}^i) \quad \text{iff} \quad L_n^{\theta,i}(y_n^i, y_n^{i-1}) = 0,$$

and $y_n^0 = y_0$ iff $\chi(y_n^0 - y_0) = 0$. ∎

Let us observe that the functional F_n^θ is convex and lower semicontinuous. Moreover, by the homogeneity of ψ (see (3.2)), F_n^θ is actually independent of the time steps. In fact, we have

$$F_n^\theta(y_n^0,\dots,y_n^{N_n}) = \sum_{i=1}^{N_n}\left(\psi\left(y_n^i - y_n^{i-1}\right) + \psi^*\left(\ell(t_{n,\theta}^i) - A y_{n,\theta}^i\right) - \langle \ell(t_{n,\theta}^i) - A y_{n,\theta}^i, y_n^i - y_n^{i-1}\rangle\right) + \chi(y_n^0 - y_0).$$

The idea of dealing with time discretizations via a discrete variational principle closely relates our analysis to the theory of so-called *variational integrators*. The latter are numerical schemes stemming from the approximation of the action functional in Lagrangian mechanics. By referring the reader to the monograph [12] and the survey [25], we shall restrain here from giving a detailed presentation of the subject and limit ourselves to some (necessarily sketchy) considerations. By letting $(t,y,p) \in [0,T] \times \mathbb{R}^m \times \mathbb{R}^m \mapsto \mathcal{L}(t,y,p)$ denote the Lagrangian of a (finite-dimensional, for simplicity) system, the Hamilton principle asserts that the actual trajectory $t \mapsto y(t)$ of the system minimizes the action functional

$$y \mapsto \int_0^T \mathcal{L}(t, y(t), \dot y(t))\ \mathrm{d}t$$

among all curves with prescribed end points, thus solving the Lagrange equations

$$\partial_{y_i}\mathcal{L} - \frac{\mathrm{d}}{\mathrm{d}t}\partial_{p_i}\mathcal{L} = 0 \quad \text{for } i = 1,\dots,m. \tag{6.7}$$

Hence, a natural idea is that of deriving numerical schemes for Lagrangian mechanics by applying some quadrature procedure to the action functional, i.e., discretizing

Hamilton's principle. The resulting discrete schemes show comparable performance with respect to other methods but generally enjoy some interesting extra (and often crucial) properties such as the conservation of suitable quantities [20]. Variational integrators have been intensively applied in finite-dimensional contexts and, more recently, to the situation of nonlinear wave equations [24] and nonequilibrium elasticity [19].

The present analysis may bear some resemblance to the above-mentioned theory. Indeed, the formulation of the θ-scheme in the case $\theta = 1/2$ stems exactly from the midpoint quadrature of the functional F as

$$\begin{aligned}
&\int_{t_n^{i-1}}^{t_n^i} L(t, \widehat{y}(t), \dot{\widehat{y}}(t))\, \mathrm{d}t \\
&= \tau_n^i L(t_{n,1/2}^i, \widehat{y}(t_{n,1/2}^i), \dot{\widehat{y}}(t_{n,1/2}^i)) \\
&= \psi\left(y^i - y^{i-1}\right) + \psi^*\left(\ell(t_{n,1/2}^i) - A\left(\frac{y^i + y^{i-1}}{2}\right)\right) \\
&- \left\langle \ell(t_{n,1/2}^i) - A\left(\frac{y^i + y^{i-1}}{2}\right), y_n^i - y_n^{i-1} \right\rangle,
\end{aligned}$$

where $\widehat{y}$ is taken to be piecewise affine on the partition P_n.

On the other hand, our focus here is quite different. First of all, we are not dealing with the Hamilton principle (end points are not fixed) as we are not aimed at solving the Euler–Lagrange equations for F (i.e., solving (6.7)). Second, we are specifically interested at infinite-dimensional situations, namely, PDEs. Finally, the only choice of θ which is directly related with a quadrature of F is $\theta = 1/2$, and we are not considering higher-order schemes.

Before closing this discussion, let us mention that some Γ-convergence techniques have been recently exploited in the (finite-dimensional) frame of variational integrators by Müller and Ortiz [37] (see also [23]).

6.2. Stability of the θ-scheme. It has been known since Han and Reddy [15, 13] that the choice $\theta < 1/2$ in (6.3) leads to an unconditionally unstable scheme and that, on the contrary, for $\theta \in [1/2, 1]$ the θ-scheme is stable in $H^1(0,T;Y)$ when Y is a Hilbert space and the partitions are chosen to be uniform.

Here we shall provide an alternative stability proof by taking into account the Banach-space frame.

LEMMA 6.4 (stability). *Assume* (3.6)*, and let* $\theta \in [1/2, 1]$. *Then the solution to the* θ*-scheme* (6.3) *fulfills*

$$\|\dot{\widehat{y}}_{n,\theta}\|_{L^\infty(0,T;Y)} \le \frac{1}{\alpha}\|\dot{\ell}\|_{L^\infty(0,T;Y^*)} \quad \text{if } \theta = 1 \text{ or } \theta = \frac{1}{2}. \tag{6.8}$$

Moreover, for constant time steps,

$$\|\dot{\widehat{y}}_{n,\theta}\|_{L^\infty(0,T;Y)} \le \frac{1}{\alpha(2\theta - 1)}\|\dot{\ell}\|_{L^\infty(0,T;Y^*)} \quad \text{if } \frac{1}{2} < \theta < 1. \tag{6.9}$$

Our argument coincides with that of [32, Thm. 4.4] in the case of the implicit Euler scheme, i.e., $\theta = 1$, and it is an extension of the latter for the case $1/2 < \theta < 1$. Here we do not play with the variational inequality by choosing suitable tests but use the scalar relations $L_n^{\theta,i}(y_n^i, y_n^{i-1}) = 0$ instead (this, however, makes no substantial

difference since the latter scalar relations are exactly the outcome of the test on the variational inequality in [32, Thm. 4.4]).

The stability proof for the Crank–Nicolson scheme $\theta = 1/2$ is quite different from former arguments and stems as a direct outcome of our variational approach. In both cases $\theta = 1$ and $\theta = 1/2$, the stability constant $1/\alpha$ is sharp (see Lemma 4.7).

We complement this analysis by providing the stability for the θ-scheme for $1/2 < \theta < 1$ in the case of constant time steps (likely with a nonoptimal, although explicit, stability constant).

Note that the stability estimates (6.8)–(6.9) do not hold in the classical parabolic situation (i.e., ψ quadratic on a Hilbert space) unless the very restrictive compatibility assumption $Ay_0 = \ell(0)$ is made (and in this case, the proof below just goes through). Evidence of this failure is available even in the simplest scalar situation ($Y = \mathbb{R}$) of problem $\dot{y}(t) + y(t) = 0$ for $t > 0$ and $y(0) = y_0$. Indeed, the corresponding θ-scheme fulfills (6.8)–(6.9) iff $y_0 = 0$ (and hence $y \equiv 0$). The current rate-independent situation turns out to be better behaved since the restriction $Ay_0 = \ell(0)$ of the parabolic case is replaced by the weaker $y_0 \in S(0)$ (see (4.5)).

Proof. Let us prove the stability of the Crank–Nicolson scheme $\theta = 1/2$ first. For this aim, it suffices to recall that

$$\begin{aligned} 0 &= F_n^{1/2}(y_n^0, \dots, y_n^{N_n}) \\ &= \int_0^T \Big(\psi(\dot{\widehat{y}}_n) + \psi^*(\widehat{\ell}_n - A\widehat{y}_n) - \langle \widehat{\ell}_n - A\widehat{y}_n, \dot{\widehat{y}}_n\rangle\Big) + \chi(\widehat{y}_n(0) - y_0). \end{aligned}$$

Hence $\widehat{y}_n$ minimizes the functional F where ℓ is replaced by $\widehat{\ell}_n$. The stability estimate follows from Lemma 4.7.

Let us now move to the case $1/2 < \theta \le 1$. Relation (6.5) applied at level $i-1$ for some $i = 2, \dots, N_n$ along with the choice $y = y_n^i$ entails that

$$\begin{aligned} &\theta\phi(y_n^i - y_n^{i-1}) + \theta\phi(y_n^{i-1}) - \langle \ell(t_{n,\theta}^{i-1}) - (1-\theta)Ay_n^{i-2}, y_n^{i-1}\rangle \\ &\quad \le \theta\phi(y_n^i) - \langle \ell(t_{n,\theta}^{i-1}) - (1-\theta)Ay_n^{i-2}, y_n^i\rangle + \psi(y_n^i - y_n^{i-1}), \end{aligned}$$

where the extra term $\theta\phi(y_n^i - y_n^{i-1})$ is obtained from the fact ϕ is quadratic. Hence, we have

$$\begin{aligned} &\theta\phi(y_n^i - y_n^{i-1}) + \theta\phi(y_n^{i-1}) - \theta\phi(y_n^i) \\ &\le \langle \ell(t_{n,\theta}^{i-1}), y_n^{i-1} - y_n^i\rangle + (1-\theta)\langle A(y_n^{i-2} - y_n^{i-1}), y_n^i - y_n^{i-1}\rangle + \psi(y_n^i - y_n^{i-1}) \\ &+(1-\theta)\langle Ay_n^{i-1}, y_n^i - y_n^{i-1}\rangle \\ &= \langle \ell(t_{n,\theta}^{i-1}), y_n^{i-1} - y_n^i\rangle + (1-\theta)\langle A(y_n^{i-2} - y_n^{i-1}), y_n^i - y_n^{i-1}\rangle + \psi(y_n^i - y_n^{i-1}) \\ &-(1-\theta)\Big(\phi(y_n^{i-1}) + \phi(y_n^i - y_n^{i-1}) - \phi(y_n^i)\Big), \end{aligned}$$

so that

$$\phi(e_n^i) + \phi(y_n^{i-1}) - \phi(y_n^i) \le -\langle \ell(t_{n,\theta}^{i-1}), e_n^i\rangle + (\theta-1)\langle Ae_n^{i-1}, e_n^i\rangle + \psi(e_n^i), \tag{6.10}$$

where we have used $e_n^i = y_n^i - y_n^{i-1}$ in order to shorten notations.

Next, from $L_n^{\theta,i}(y_n^i, y_n^{i-1}) = 0$ for $i = 1, \dots, N_n$, we obtain

$$\begin{aligned} 0 &= \psi(e_n^i) - \langle \ell(t_{n,\theta}^i) - Ay_{n,\theta}^i, e_n^i\rangle \\ &= \psi(e_n^i) - \langle \ell(t_{n,\theta}^i), e_n^i\rangle + \theta\Big(\phi(y_n^i) + \phi(e_n^i) - \phi(y_n^{i-1})\Big) \\ &\quad - (1-\theta)\Big(\phi(y_n^{i-1}) + \phi(e_n^i) - \phi(y_n^i)\Big). \end{aligned}$$

In particular, we have checked that

$$\psi(e_n^i) + \phi(y_n^i) - \phi(y_n^{i-1}) + (2\theta - 1)\phi(e_n^i) = \langle \ell(t_{n,\theta}^i), e_n^i \rangle. \tag{6.11}$$

We take the sum between the latter and (6.10) and get

$$2\theta\phi(e_n^i) \le \langle \ell(t_{n,\theta}^i) - \ell(t_{n,\theta}^{i-1}), e_n^i \rangle + (\theta - 1)\langle Ae_n^{i-1}, e_n^i \rangle$$

or, equivalently,

$$\langle Ae_{n,\theta}^i, e_n^i \rangle \le \langle \ell(t_{n,\theta}^i) - \ell(t_{n,\theta}^{i-1}), e_n^i \rangle. \tag{6.12}$$

Now, if $\theta = 1$, we conclude that

$$|e_n^i| \le \frac{1}{\alpha} |\ell(t_{n,\theta}^i) - \ell(t_{n,\theta}^{i-1})|_*,$$

and the assertion follows.

In case $1/2 \le \theta < 1$ and for a constant time-step partition, one proceeds from (6.12) by computing

$$\begin{aligned}
&\tau_n \|\dot{\widehat{\ell}}_n\|_{L^\infty(0,T;Y^*)} \sqrt{\frac{2}{\alpha}} \sqrt{\phi(e_n^i)} \\
&\ge \langle \ell(t_{n,\theta}^i) - \ell(t_{n,\theta}^{i-1}), e_n^i \rangle \\
&\ge \theta \langle Ae_n^i, e_n^i \rangle + (1-\theta)\langle Ae_n^{i-1}, e_n^i \rangle \\
&= (2\theta - 1)\langle Ae_n^i, e_n^i \rangle + (1-\theta)\langle A(e_n^i + e_n^{i-1}), e_n^i \rangle \\
&\ge 2(2\theta - 1)\phi(e_n^i) + (1-\theta)\Big(\phi(e_n^i) - \phi(e_n^{i-1})\Big).
\end{aligned}$$

Note that the coefficient $(2\theta - 1)$ is strictly positive as $\theta > 1/2$. By using the fact that $y_n^0 = y_0 \in S(0)$ (recall (4.5)), we readily check that

$$\phi(e_n^1) + \phi(y_n^0) - \langle \ell(0), y_n^0 \rangle \le \phi(y_n^1) - \langle \ell(0), y_n^1 \rangle + \psi(e_n^1), \tag{6.13}$$

and, by adding the latter to (6.11) for $i = 1$, we have

$$2\theta\phi(e_n^1) \le \langle \ell(t_{n,\theta}^1) - \ell(0), e_n^1 \rangle \le \tau_n \|\dot{\widehat{\ell}}_n\|_{L^\infty(0,T;Y^*)} \sqrt{\frac{2}{\alpha}} \sqrt{\phi(e_n^1)}. \tag{6.14}$$

Let us define

$$a_i^2 = \phi\left(\frac{y_n^i - y^{i-1}}{\tau_n}\right) = \phi(e_n^i)/\tau_n^2,$$

$$C_0 = \frac{2(2\theta - 1)}{1 - \theta}, \quad C_1 = \frac{1}{1-\theta}\sqrt{\frac{2}{\alpha}} \|\dot{\widehat{\ell}}_n\|_{L^\infty(0,T;Y^*)}, \quad C_2 = \frac{C_1}{C_0},$$

so that, owing to (6.13) and (6.14) and by using the fact that $2(2\theta - 1) < 2\theta$,

$$\begin{aligned}
&\big(C_0 + 1\big)a_i^2 - a_{i-1}^2 \le C_1 a_i \quad \text{for} \quad i = 2, \dots, N_n, \\
&\qquad a_1 \le \frac{1}{2\theta}\sqrt{\frac{2}{\alpha}} \|\dot{\widehat{\ell}}_n\|_{L^\infty(0,T;Y^*)} \le \frac{1}{2(2\theta - 1)}\sqrt{\frac{2}{\alpha}} \|\dot{\widehat{\ell}}_n\|_{L^\infty(0,T;Y^*)} \\
&\qquad\quad = \frac{1}{1-\theta}\sqrt{\frac{2}{\alpha}} \|\dot{\widehat{\ell}}_n\|_{L^\infty(0,T;Y^*)} \frac{1-\theta}{2(2\theta - 1)} \\
&\qquad\quad = \frac{C_1}{C_0} = C_2.
\end{aligned}$$

Now, since $(C_0+1)C_2^2 - C_1C_2 = C_2^2$, we easily prove by induction that $a_i \le C_2$, and the assertion follows. ∎

6.3. Convergence. We shall prove the weak-star $W^{1,\infty}(0,T;Y)$ convergence for the θ-method. This result has to be compared with that of Han and Reddy [14, Thm. 3.4], where the uniform convergence of the backward constant interpolations is obtained. Our result is weaker than that of [14, Thm. 3.4] since we are not providing strong convergence. On the other hand, we believe our half-page proof to be possibly more transparent than the long argument developed in [14]. Let us moreover mention that, in the Hilbertian case and for A coercive on Y, the strong convergence in $W^{1,p}(0,T;Y)$ for all $p<\infty$ of the Euler method $\theta = 1$ has been proved in [16, Prop. 3.9, p. 33].

THEOREM 6.5 (convergence for the θ-method). *Assume* (3.6)*, and let* $F_n^\theta(y_n^0,\dots,y_n^{N_n}) = 0$*. Then* $\widehat{y}_n \to y$ *weakly star in* $W^{1,\infty}(0,T;Y)$*, where* $F(y) = 0 = \min F$*.*

Proof. Owing to Lemma 6.4, we can extract a (not relabeled) subsequence such that $\widehat{y}_n \to y$ weakly star in $W^{1,\infty}(0,T;Y)$ and hence weakly pointwise in Y. Moreover, we clearly have that both $\overline{y}_n$ and $\overline{y}_{n,\theta}$ converge at the same limit weakly star in $L^\infty(0,T;Y)$. Finally, we directly check that $\overline{\ell}_{n,\theta} \to \ell$ strongly in $L^\infty(0,T;Y^*)$. By observing that, since $\theta \ge 1/2$,

$$\begin{aligned}\tau_n^i\langle A(\theta y_n^i + (1-\theta)y_n^{i-1}), \delta y_n^i\rangle &= \phi(y_n^i) + (2\theta-1)\phi(y_n^i - y_n^{i-1}) - \phi(y_n^{i-1})\\ &\ge \phi(y_n^i) - \phi(y_n^{i-1}),\end{aligned}$$

we compute that

$$\begin{aligned}0 &= F_n^\theta(y_n^0,\dots,y_n^{N_n})\\ &\ge \int_0^T \Big(\psi(\dot{\widehat{y}}_n) + \psi^*(\overline{\ell}_{n,\theta} - A\overline{y}_{n,\theta}) - \langle \overline{\ell}_{n,\theta}, \dot{\widehat{y}}_n\rangle\Big)\\ &\quad + \phi(\widehat{y}_n(T)) - \phi(\widehat{y}_n(0)) + \chi(\widehat{y}_n(0) - y_0)\\ &= \int_0^T \Big(\psi(\dot{\widehat{y}}_n) + \psi^*(\overline{\ell}_{n,\theta} - A\overline{y}_{n,\theta}) - \langle \overline{\ell}_{n,\theta}, \dot{\widehat{y}}_n\rangle\Big)\\ &\quad + \phi(\widehat{y}_n(T)) + \phi(y_0) - \langle A\widehat{y}_n(0), y_0\rangle + |\widehat{y}_n(0) - y_0|^2.\end{aligned}$$

Finally, it suffices to pass to the lim inf above as $n\to\infty$ and exploit lower semicontinuity and the stated convergences in order to obtain $F(y)\le 0$. Hence, by Theorem 3.1 and Corollary 4.6, y is the only solution to (1.1), and the whole sequence $\widehat{y}_n$ converges. ∎

6.4. The functional controls the uniform distance. We shall reproduce at the discrete level the results of subsection 4.6. We begin by showing how to possibly control the uniform distance of two vectors by means of the discrete functional F_n^θ.

LEMMA 6.6 (uniform distance control via F_n^θ). *Let the vectors* $u = (u^0,\dots,u^{N_n})$ *and* $v = (v^0,\dots,v^{N_n}) \in Y^{N_n+1}$ *be given. Then*

$$\begin{aligned}&\eta(1-\eta)\max_{1\le i\le N_n}\phi(u^i - v^i)\\ &\le \eta F_n^\theta(u^0,\dots,u^{N_n}) + (1-\eta)F_n^\theta(v^0,\dots,v^{N_n}) \quad \forall \eta\in[0,1].\end{aligned}$$

Proof. This proof follows the same lines as that of Corollary 4.4. Let $1 \le i \le N_n$ be fixed, and define $F_n^{\theta,i} : Y^{i+1} \to [0,\infty]$ as

$$\begin{aligned} &F_n^{\theta,i}(y^0,\dots,y^i) \\ &\quad= \sum_{j=1}^{i} \tau_n^j L_n^{\theta,j}(y^j, y^{j-1}) + \chi(y^0 - y_0) \\ &\quad= \sum_{j=1}^{i} \psi(y^j - y^{j-1}) + \psi^*(\ell(t_{n,\theta}^j) - Ay_\theta^j) - \langle \ell(t_{n,\theta}^j), y^j - y^{j-1}\rangle \\ &\qquad+ \phi(y^i) + (2\theta - 1)\sum_{j=1}^{i} \phi(y^j - y^{j-1}) + \phi(y_0) - \langle Ay^0, y_0\rangle + |y^0 - y_0|^2. \end{aligned}$$

Then, clearly, $y = (y^0,\dots,y^i) \mapsto G_n^{\theta,i}(y) = F_n^{\theta,i}(y) - \phi(y^i)$ is convex. Hence, by letting $w = \eta u + (1-\eta)v$ for some $\eta \in [0,1]$, we have

$$\begin{aligned} &0 \le F_n^{\theta,i}(w) \\ &\le \eta\big(G_n^{\theta,i}(u) + \phi(u^i)\big) + (1-\eta)\big(G_n^{\theta,i}(v) + \phi(v^i)\big) - \eta(1-\eta)\phi(u^i - v^i), \end{aligned}$$

whence the assertion follows. ∎

Again, note that the latter lemma controls the uniform norm of the distance only if the stronger (3.6) is required. The following corollary of Lemma 6.6 will be the starting point for some possible a posteriori error control procedure (see subsection 6.6).

COROLLARY 6.7 (uniform distance from the minimizer). *Let* $F_n^\theta(y^0,\dots,y^{N_n}) = 0$. *Then*

$$\max_{1\le i\le N_n} \phi(y^i - v^i) \le F_n^\theta(v^0,\dots,v^{N_n}) \quad \forall (v^0,\dots,v^{N_n}) \in Y^{N_n+1}.$$

Moreover, we reobtain a proof of the uniqueness of the solution of the θ-method.

COROLLARY 6.8 (uniqueness of the minimizer). *Assume* (3.6). *Then there exists at most one* $y = (y^0,\dots,y^{N_n})$ *such that* $F_n^\theta(y) = 0$.

6.5. The generalized θ-method. Although minimizers of F_n^θ and solutions of the θ-scheme (6.3) coincide, minimizing sequences of F_n^θ need not solve (6.3). This extra freedom allows the minimization formulation to capture the convergence of some generalized θ-method, where the relations in (6.3) are not solved exactly but rather are approximated. Namely, we shall look for vectors $u_n = (u_n^0,\dots,u_n^{N_n})$ such that

$$F_n^\theta(u_n) \to 0 \quad \text{as} \quad n \to \infty$$

instead of $F_n^\theta(u_n) = 0$ for all $n \in \mathbb{N}$.

From the computational viewpoint, note that the θ-scheme consists in solving N_n nonlinear equations in one unknown each, while checking for stationarity for F_n^θ implies the solution of a tridiagonal system of $N_n + 1$ nonlinear equations with (up to) three unknowns each. This entails in particular that minimizing F_n^θ instead of solving (6.3) could be of a scarce interest if one is merely concerned in reproducing the θ-scheme with no error. On the other hand, the issue of solving up to some tolerance turns out to be particularly relevant whenever one is aimed at implementing an optimization procedure for the solution of (6.3). Indeed, one should be prepared to run the algorithm (some descent method, say) until some given tolerance is reached.

Our starting point for a possible convergence analysis of the generalized θ-method is the following classical error control result.

THEOREM 6.9 (Mielke and Theil [32]). *Assume* (3.6). *Then* $\widehat{y}_n \to y$ *uniformly and* $F(y) = 0$. *In particular,*

$$\max_{t\in[0,T]} |(\widehat{y}_n - y)(t)| \le C_e\sqrt{\tau_n}, \tag{6.15}$$

where C_e *depends only on data and is independent of* n.

More precisely, in [32] solely the case of the Euler scheme $\theta = 1$ is discussed. However, an easy adaptation of the argument entails the result for $\theta \in [1/2, 1)$ as well.

By explicitly comparing the minimizing sequence $u_n = (u_n^0, \dots, u_n^{N_n})$ with the corresponding solution $(y_n^0, \dots, y_n^{N_n})$ of the θ-method, we have the following.

THEOREM 6.10 (convergence for the generalized θ-method). *Assume* (3.6), *and let* $F_n^\theta(u_n^0, \dots, u_n^{N_n}) \to 0$. *Then* $\widehat{u}_n \to y$ *uniformly, where* $F(y) = 0$. *In particular,*

$$\max_{t\in[0,T]} |(\widehat{u}_n - y)(t)| \le C_e\sqrt{\tau_n} + \left(\frac{2}{\alpha}F_n^\theta(u_n^0, \dots, u_n^{N_n})\right)^{1/2}. \tag{6.16}$$

Proof. We have

$$\begin{aligned}\max_{t\in[0,T]} |(y - \widehat{u}_n)(t)| &\le \max_{t\in[0,T]} |(y - \widehat{y}_n)(t)| + \max_{t\in[0,T]} |(\widehat{y}_n - \widehat{u}_n)(t)| \\ &\le C_e\sqrt{\tau_n} + \max_{1\le i\le N_n} |u_n^i - y_n^i| \\ &\le C_e\sqrt{\tau_n} + \left(\frac{2}{\alpha}\max_{1\le i\le N_n} \phi(u_n^i - y_n^i)\right)^{1/2},\end{aligned}$$

and we conclude by applying Corollary 6.7. ∎

6.6. A posteriori error control. Let us now exploit both Corollary 4.5 and Theorem 6.10 in order to provide some possible a posteriori estimates of the approximation error by means of solutions u_n of the generalized θ-method described above.

LEMMA 6.11 (a posteriori error control via F_n^θ). *Assume* (3.6), *and let* $F_n^\theta(u_n^0, \dots, u_n^{N_n}) \sim \tau_n^s$ *for some* $s > 0$ *and* $F(y) = 0$. *Then*

$$\max_{t\in[0,T]} |(\widehat{u}_n - y)(t)| \sim \tau_n^r, \quad \textit{where} \quad 2r = \max\{1, s\}.$$

LEMMA 6.12 (a posteriori error control via F). *Assume* (3.6), *and let* $F(\widehat{u}_n) \sim \tau_n^s$ *for some* $s > 0$ *and* $F(y) = 0$. *Then* $\max_{t\in[0,T]} |(\widehat{u}_n - y)(t)| \sim \tau_n^{s/2}$.

We are also in the position of proving the weak-star convergence of the time derivatives of solutions u_n of the generalized θ-method by comparing them with the corresponding derivatives of the exact solution of the θ-method.

LEMMA 6.13 (improved convergence for the generalized θ-method). *Assume* (3.6), *and let* $F_n^\theta(u_n^0, \dots, u_n^{N_n}) \sim \tau_n^2$. *Then* $\widehat{u}_n$ *is equibounded in* $W^{1,\infty}(0, T; Y)$. *In particular,* $\widehat{u}_n \to y$ *weakly star in* $W^{1,\infty}(0, T; Y)$.

Proof. Let $(y_n^0, \dots, y_n^{N_n})$ be the solution of the θ-scheme. By exploiting Lemma 6.7, we check that

$$\begin{aligned}|u_n^i - u_n^{i-1}| &\le \left|u_n^i - y_n^i\right| + \left|y_n^i - y_n^{i-1}\right| + \left|y_n^{i-1} - u_n^{i-1}\right| \\ &\le \tau_n^i\|\dot{\widehat{y}}_n\|_{L^\infty(0,T;Y)} + 2\left(\frac{2}{\alpha}F_n^\theta(u_n^0, \dots, u_n^{N_n})\right)^{1/2}.\end{aligned}$$

The uniform bound on $\|\widehat{u}_n\|_{W^{1,\infty}(0,T;Y)}$ follows by dividing the latter by τ_n^i, taking the maximum as $1 \le i \le N_n$, and recalling Lemma 6.4. ∎

6.7. Adaptivity. By assuming (3.6), the above-introduced a posteriori error estimators can be exploited in order to develop an adaptive strategy. In particular, the error control in the uniform norm up to a given tolerance $\mathrm{tol} > 0$

$$\max_{t\in[0,T]} |(y - \widehat{y}_n)(t)| \le \mathrm{tol}$$

for some piecewise approximation $\widehat{y}_n$, with $\chi(\widehat{y}_n(0) - y_0) \le \alpha\,\mathrm{tol}^2/4$, can be inferred, for instance, by choosing time steps in such a way that

$$\int_{t_n^{i-1}}^{t_n^i} L(t, \widehat{y}_n(t), \dot{\widehat{y}}_n(t)) \le \frac{\alpha\,\mathrm{tol}^2}{4N_n},$$

namely, by uniformly distributing the error along the partition.

Alternatively, one could develop an adaptive strategy by considering just computed quantities at the discrete level by asking for

$$\tau_n^i L_n^{\theta,i}(y_n^i, y_n^{i-1}) \le \frac{\alpha\,\mathrm{tol}^2}{32N_n} \quad \text{for} \quad \tau_n \le \frac{\mathrm{tol}^2}{16C_e^2}$$

and exploiting Theorem 6.10.

7. Space-time approximations. Let us combine the results of the previous sections (and use the corresponding notation) in order to state and prove a result on the convergence of full space-time approximations. Our results have to be compared with the former convergence analysis by Han and Reddy [13]. Our approach leads to a convergence proof with respect to a weaker topology. However, it is, on the one hand, slightly more general (some assumptions on the spaces and the functionals—see (H1)–(H2) [13, p. 264]—are not required) and, on the other hand, has a much simpler proof.

THEOREM 7.1 (convergence of space-time approximations). *Assuming* (5.1)–(5.9) *and that* $\theta \in [1/2, 1]$, *define* $L_{n,h}^{\theta,i}(y,z) : Y \times Y \to [0,\infty]$ *as*

$$L_{n,h}^{\theta,i}(y,z) = \psi_h\left(\frac{y-z}{\tau_n^i}\right) + \psi_h^*\left(\ell_h(t_{n,\theta}^i) - A_h(\theta y + (1-\theta)z)\right) - \left\langle \ell_h(t_{n,\theta}^i) - A_h(\theta y + (1-\theta)z), \frac{y-z}{\tau_n^i}\right\rangle,$$

where $A_h = \partial\phi_h$, *and let the functionals* $F_{n,h}^\theta : Y^{N_n+1} \to [0,\infty]$ *be defined as*

$$F_{n,h}^\theta(y^0, \dots, y^{N_n}) = \sum_{i=1}^{N_n} \tau_n^i L_{n,h}^{\theta,i}(y^i, y^{i-1}) + \chi_h(y^0 - y_0),$$

where $\chi_h(\cdot) = \phi_h(\cdot) + |\cdot|^2$ *(note that* $D(F_{n,h}^\theta) \subset Y_h^{N_n+1}$*). Finally, let* $F_{n,h}(y_h^0, \dots, y_h^{N_n}) = 0$. *We have the following:*

(a) $\widehat{y}_{n,h} \to y_h$ *weakly star in* $W^{1,\infty}(0,T;Y)$ *as* $(n,h) \to (\infty,h)$ *and* $F_h(y_h) = 0$.
(b) $\widehat{y}_{n,h} \to \widehat{y}_n$ *weakly star in* $W^{1,\infty}(0,T;Y)$ *as* $(n,h) \to (n,0)$ *and* $F_n^\theta(y_n) = 0$.
(c) $\widehat{y}_n \to y$ *weakly star in* $W^{1,\infty}(0,T;Y)$ *as* $(n,0) \to (\infty,0)$ *and* $F(y) = 0$.

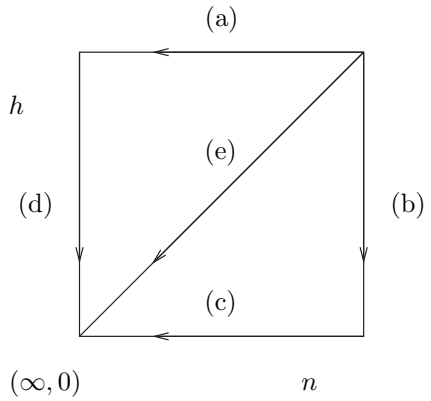


FIG. 1. *Convergences for space-time approximations (see Theorem* 7.1*).*

(d) $y_h \to y$ *weakly star* $W^{1,\infty}(0,T;Y)$ *as* $(\infty,h) \to (\infty,0)$ *and* $F(y)=0$.

(e) $\widehat{y}_{n,h} \to y$ *weakly star in* $W^{1,\infty}(0,T;Y)$ *as* $(n,h) \to (\infty,0)$ *and* $F(y)=0$.

The thesis of the theorem is illustrated in Figure 1. In particular, we aim at showing that the space (or data) and time limit can be taken in any order. Note that limit (c) has been already checked in Theorem 6.5 and that the very same argument yields limit (a) as well (recall that Y_h is closed). Moreover, limit (d) is discussed in Theorem 5.3. So what we are actually left to check are limits (b) and (e) only.

Proof. Limit (b). The assertion follows once we check that, for all $i = 1, \dots, N_n$, if $y^{i-1}_{n,h} \to y^{i-1}_n$ weakly in Y, one has the weak convergence $y^i_{n,h} \to y^i_n$ as well. Recall that

$$y^i_{n,h} \in \operatorname{Arg\,min}_{y\in Y}\Big(\theta\phi_h(y) - \langle \ell_h(t^i_{n,\theta}) - (1-\theta)A_h y^{i-1}_{n,h}, y\rangle + \psi_h(y - y^{i-1}_{n,h})\Big)$$

$$= \operatorname{Arg\,min}_{y\in Y_h}\Big(\theta\phi(y) - \langle \ell_h(t^i_{n,\theta}) - (1-\theta)A y^{i-1}_{n,h}, y\rangle + \psi_h(y - y^{i-1}_{n,h})\Big).$$

Hence, since we have (5.6), the sequence $y^i_{n,h}$ is weakly precompact, and, up to the extraction of a (not relabeled) subsequence, $y^i_{n,h} \to \tilde{y}$ weakly in Y. Let us prove that $\tilde{y}$ solves the incremental problem (6.4). Indeed, we have

$$\begin{aligned} 0 &\le L^{\theta,i}_n(\tilde{y}, y^{i-1}_n) \\ &\le \liminf_{h\to 0}\Big(\psi_h(y^i_{n,h} - y^{i-1}_{n,h}) + \psi^*_h(\ell_h(t^i_{n,\theta}) - A y^i_{n,h,\theta}) \\ &\quad - \langle \ell_h(t^i_{n,\theta}) - A y^i_{n,h,\theta}, y^i_{n,h} - y^{i-1}_{n,h}\rangle\Big) \\ &= \liminf_{h\to 0} L^{\theta,i}_{n,h}(y^i_{n,h}, y^{i-1}_{n,h}) = 0, \end{aligned}$$

where we have used the Mosco convergence in (5.5) and the pointwise convergence of ℓ_h (5.7). Since the only solution of (6.4) is y^i_n, we have $\tilde{y} = y^i_n$, and the whole sequence converges.

Let us mention that, if the functionals ψ_h are uniformly linearly bounded (which is quite common in practice), one could prove the latter convergence to be actually strong: Namely, $y^{i-1}_{n,h} \to y^{i-1}_n$ strongly in Y implies the strong convergence $y^i_{n,h} \to y^i_n$. Indeed, let w_h and $\tilde{w}_h$ be such that $w_h - y^i_{n,h} \to 0$ strongly in Y, $\psi_h(w_h - y^i_{n,h}) \to 0$,

$\tilde w_h \in Y_h$ and $\tilde w_h - w_h \to 0$ strongly in Y. Then

$$\theta\phi(y^i_{n,h}) - \langle \ell(t^i_{n,\theta}) - (1-\theta)A_h y^{i-1}_{n,h}, y^i_{n,h}\rangle + \psi_h(y^i_{n,h} - y^{i-1}_{n,h})$$

$$\le \theta\phi_h(\tilde w_h) - \langle \ell(t^i_{n,\theta}) - (1-\theta)A y^{i-1}_{n,h}, \tilde w_h\rangle + \psi_h(\tilde w_h - w_h) + \psi_h(w_h - y^i_{n,h}).$$

If ψ_h are uniformly linearly bounded above, then $\psi_h(\tilde w_h - w_h) \to 0$, with $h \to 0$. Then, by passing to the lim sup in the latter, we check that $\limsup_{h\to 0} \phi(y^i_{n,h}) \le \phi(y^i_n)$, which together with lower semicontinuity gives $\phi(y^i_{n,h}) \to \phi(y^i_n)$, and the strong convergence follows from the reflexivity of Y.

Limit (e). Lemma 6.4, the uniform Lipschitz continuity of ℓ_h (5.8), and the initial datum convergence (5.9) entail that $\widehat y_{n,h}$ are uniformly Lipschitz continuous as well. Hence, by extracting a (not relabeled) subsequence, $\widehat y_{n,h} \to y$ weakly star in $W^{1,\infty}(0,T;Y)$. In order to check that y solves (1.1), let us remark that, since $\ell^i_{n,h,\theta} = \ell_h(t^i_{n,\theta})$,

$$\overline\ell_{n,h,\theta} \to \ell \quad \text{strongly in} \quad L^1(0,T;Y^*)$$

and that, by [45, Cor. 4.4],

$$\int_0^T \psi(\dot y) \le \liminf_{h\to 0} \int_0^T \psi_h(\dot{\widehat y}_{n,h}),$$

$$\int_0^T \psi^*(\ell - Ay) \le \liminf_{h\to 0} \int_0^T \psi_h^*(\overline\ell_{n,h,\theta} - A\overline y_{n,h,\theta})$$

and compute that

$$\begin{aligned}
0 &\le F(y)\\
&\le \liminf_{h\to 0} \left(\int_0^T \big(\psi_h(\dot{\widehat y}_{n,h}) + \psi_h^*(\overline\ell_{n,h,\theta} - A\overline y_{n,h,\theta}) - \langle \overline\ell_{n,h,\theta}, \dot{\widehat y}_{n,h}\rangle\big)\right.\\
&\quad \left. + \phi(\widehat y_{n,h}(T)) + \phi(y_{0,h}) - \langle A\widehat y_{n,h}(0), y_{0,h}\rangle + |\widehat y_{n,h}(0) - y_{0,h}|^2 \right)\\
&\le \liminf_{h\to 0} F^\theta_{n,h}(y^0_{n,h}, \dots, y^{N_n}_{n,h}) = 0,
\end{aligned}$$

and we have $F(y) = 0$. □

We shall conclude by briefly mentioning some further results which can be obtained by suitably adapting to the current fully discretized situation the arguments developed above for time discretizations. First, in the same spirit of Lemma 6.6, one could consider the possibility of estimating the distance of a vector from the minimizer of $F^\theta_{n,h}$ by means of the functional itself. Second, the use of Corollary 4.5 would entail the possibility of an a posteriori error control, and some adaptive strategy along the lines of subsection 6.7 could be considered. Finally, by relying on the known convergence estimates for full space-time discretized problems [13], one could obtain a convergence and an a posteriori error control result for some generalized space-time approximated problem where $F^\theta_{n,h}$ are not exactly minimized and one considers minimizing sequences instead (see subsection 6.5). We shall develop these considerations elsewhere.

REFERENCES

[1] H. Attouch, *Variational Convergence for Functions and Operators*, Pitman, Boston, 1984.

[2] G. Auchmuty, *Saddle-points and existence-uniqueness for evolution equations*, Differential Integral Equations, 6 (1993), pp. 1161–1171.

[3] H. Brezis and I. Ekeland, *Un principe variationnel associé à certaines équations paraboliques. Le cas dépendant du temps*, C. R. Acad. Sci. Paris Sér. A-B, 282 (1976), pp. Ai, A1197–A1198.

[4] H. Brezis and I. Ekeland, *Un principe variationnel associé à certaines équations paraboliques. Le cas indépendant du temps*, C. R. Acad. Sci. Paris Sér. A-B, 282 (1976), pp. Aii, A971–A974.

[5] G. Dal Maso, *An Introduction to* Γ*-Convergence*, Progr. Nonlinear Differential Equations Appl. 8, Birkhäuser Boston, Cambridge, MA, 1993.

[6] G. Dal Maso, A. DeSimone, and M. G. Mora, *Quasistatic evolution problems for linearly elastic-perfectly plastic materials*, Arch. Ration. Mech. Anal., 180 (2006), pp. 237–291.

[7] G. Duvaut and J.-L. Lions, *Inequalities in Mechanics and Physics*, Springer-Verlag, Berlin, 1976.

[8] M. A. Efendiev and A. Mielke, *On the rate-independent limit of systems with dry friction and small viscosity*, J. Convex Anal., 13 (2006), pp. 151–167.

[9] N. Ghoussoub, *Selfdual Partial Differential Systems and Their Variational Principles*, Universitext, Springer-Verlag, New York, to appear.

[10] N. Ghoussoub and L. Tzou, *A variational principle for gradient flows*, Math. Ann., 330 (2004), pp. 519–549.

[11] E. De Giorgi and T. Franzoni, *Su un tipo di convergenza variazionale*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur., 58 (1975), pp. 842–850.

[12] E. Hairer, Ch. Lubich, and G. Wanner, *Geometric Numerical Integration*, Springer Ser. Comput. Math. 31, 2nd ed., Springer-Verlag, Berlin, 2006.

[13] W. Han and B. D. Reddy, *Plasticity, Mathematical Theory and Numerical Analysis*, Springer-Verlag, New York, 1999.

[14] W. Han and B. D. Reddy, *Convergence of approximations to the primal problem in plasticity under conditions of minimal regularity*, Numer. Math., 87 (2000), pp. 283–315.

[15] W. Han and B. D. Reddy, *Computational plasticity: The variational basis and numerical analysis*, Comput. Mech. Adv., 2 (1995), pp. 283–400.

[16] P. Krejčí, *Hysteresis, Convexity and Dissipation in Hyperbolic Equations*, GAKUTO Internat. Ser. Math. Sci. Appl. 8, Gakkotosho, Tokyo, 1996.

[17] B. Lemaire, *An asymptotical variational principle associated with the steepest descent method for a convex function*, J. Convex Anal., 3 (1996), pp. 63–70.

[18] J. Lemaitre and J.-L. Chaboche, *Mechanics of Solid Materials*, Cambridge University Press, London, 1990.

[19] A. Lew, J. E. Marsden, M. Ortiz, and M. West, *Asynchronous variational integrators*, Arch. Ration. Mech. Anal., 167 (2003), pp. 85–146.

[20] A. Lew, J. E. Marsden, M. Ortiz, and M. West, *Variational time integrators*, Internat. J. Numer. Methods Engrg., 60 (2004), pp. 153–212.

[21] M. Mabrouk, *Un principe variationnel pour une équation non linéaire du second ordre en temps*, C. R. Acad. Sci. Paris Sér. I Math., 332 (2001), pp. 381–386.

[22] M. Mabrouk, *A variational principle for a nonlinear differential equation of second order*, Adv. Appl. Math., 31 (2003), pp. 388–419.

[23] F. Maggi and M. Morini, *A* Γ*-convergence result for variational integrators of Lagrangians with quadratic growth*, ESAIM Control Optim. Calc. Var., 10 (2004), pp. 656–665.

[24] J. E. Marsden, G. W. Patrick, and S. Shkoller, *Multisymplectic geometry, variational integrators, and nonlinear PDEs*, Comm. Math. Phys., 199 (1998), pp. 351–395.

[25] J. E. Marsden and M. West, *Discrete mechanics and variational integrators*, Acta Numer., 10 (2001), pp. 357–514.

[26] A. Mielke, *Evolution of rate-independent systems*, in Handbook of Differential Equations, Evolutionary Equations, Vol. 2, C. Dafermos and E. Feireisl, eds., Elsevier, New York, 2005, pp. 461–559.

[27] A. Mielke and M. Ortiz, *A class of minimum principles for characterizing the trajectories and the relaxation of dissipative systems*, ESAIM Control Optim. Calc. Var., 2008, to appear; also available online from http://www.wias-berlin.de/main/publications/wias-publ/index.cgi.en.

[28] A. Mielke, R. Rossi, and G. Savaré, manuscript, 2008.

[29] A. Mielke, R. Rossi, and G. Savaré, *A metric approach to a class of doubly nonlinear evolution equations and applications*, Ann. Sc. Norm. Super. Pisa Cl. Sci., 5 (2008), pp. 97–169; also available online from http://www.wias-berlin.de/main/publications/wias-publ/index.cgi.en.

[30] A. Mielke, T. Roubíček, and U. Stefanelli, *Γ-limits and relaxations for rate-independent evolutionary problems*, Calc. Var. Partial Differential Equations, 31 (2008), pp. 387–416.

[31] A. Mielke and U. Stefanelli, *A Discrete Variational Principle for Rate-Independent Evolution*, Wias preprint 1295, 2008, also available online from http://www.wias-berlin.de/main/publications/wias-publ/index.cgi.en.

[32] A. Mielke and F. Theil, *On rate-independent hysteresis models*, NoDEA Nonlinear Differential Equations Appl., 11 (2004), pp. 151–189.

[33] J.-J. Moreau, *La notion de sur-potentiel et les liaisons unilatérales en élastostatique*, C. R. Acad. Sci. Paris Sér. A-B, 267 (1968), pp. A954–A957.

[34] J.-J. Moreau, *Sur les lois de frottement, de viscosité et plasticité*, C. R. Acad. Sci. Paris Sér. II Méc. Phys. Chim. Sci. Univers Sci. Terre, 271 (1970), pp. 608–611.

[35] J.-J. Moreau, *Sur l'évolution d'un système élasto-visco-plastique*, C. R. Acad. Sci. Paris Sér. A-B, 273 (1971), pp. A118–A121.

[36] U. Mosco, *Convergence of convex sets and of solutions of variational inequalities*, Adv. Math., 3 (1969), pp. 510–585.

[37] S. Müller and M. Ortiz, *On the Γ-convergence of discrete dynamics and variational integrators*, J. Nonlinear Sci., 14 (2004), pp. 279–296.

[38] B. Nayroles, *Deux théorèmes de minimum pour certains systèmes dissipatifs*, C. R. Acad. Sci. Paris Sér. A-B, 282 (1976), pp. Aiv, A1035–A1038.

[39] B. Nayroles, *Un théorème de minimum pour certains systèmes dissipatifs. Variante hilbertienne*, Travaux Sém. Anal. Convexe, 6 (1976), p. 22.

[40] H. Rios, *Étude de la question d'existence pour certains problèmes d'évolution par minimisation d'une fonctionnelle convexe*, C. R. Acad. Sci. Paris Sér. A-B, 283 (1976), pp. Ai, A83–A86.

[41] T. Roubíček, *Direct method for parabolic problems*, Adv. Math. Sci. Appl., 10 (2000), pp. 57–65.

[42] J. C. Simo and T.J.R. Hughes, *Computational Inelasticity*, Interdiscip. Appl. Math. 7, Springer-Verlag, New York, 1998.

[43] U. Stefanelli, *The Discrete Brezis-Ekeland Principle*, J. Convex. Anal., to appear, also available online from http://www.imati.cnr.it/ulisse/pubbl.html.

[44] U. Stefanelli, *A variational principle in non-smooth mechanics*, in MFO Workshop: Analysis and Numerics for Rate-Independent Processes, Oberwolfach Rep. 4, European Mathematics Society, Zurich, 2007, pp. 622–624.

[45] U. Stefanelli, *The Brezis-Ekeland principle for doubly nonlinear equations*, SIAM J. Control Optim., to appear; also available online from http://www.imati.cnr.it/ulisse/pubbl.html.

[46] U. Stefanelli, *Some remarks on convergence and approximation for a class of hysteresis problems*, Istit. Lombardo Accad. Sci. Lett. Rend. A, to appear; also available online from http://www.imati.cnr.it/ulisse/pubbl.html.

[47] R. Toader and C. Zanini, *An artificial viscosity approach to quasistatic crack growth*, submitted; also available online from http://cvgmt.sns.it/.

[48] A. Visintin, *A new approach to evolution*, C. R. Acad. Sci. Paris Sér. I Math., 332 (2001), pp. 233–238.

[49] C. Zanini, *Singular perturbations of finite dimensional gradient flows*, Discrete Contin. Dyn. Syst., 18 (2007), pp. 657–675.

SIAM J. MATH. ANAL.
Vol. 40, No. 2, pp. 653–674

AVERAGES OVER SPHERES FOR KINETIC TRANSPORT EQUATIONS WITH VELOCITY DERIVATIVES IN THE RIGHT-HAND SIDE*

NIKOLAOS BOURNAVEAS† AND SUSANA GUTIÉRREZ‡

Abstract. We prove estimates in hyperbolic Sobolev spaces $H^{s,\delta}(R^{1+d})$, $d \geq 3$, for velocity averages over spheres of solutions to the kinetic transport equation $\partial_t f + v \cdot \nabla_x f = \Omega_v^{i,j} g$, where $\Omega_v^{i,j} g$ are tangential velocity derivatives of g. Our results extend to all dimensions earlier results of Bournaveas and Perthame in dimension two [*J. Math. Pures Appl.*, 9 (2001), pp. 517–534]. We construct counterexamples to test the optimality of our results.

Key words. velocity-averaging lemmas, kinetic transport equation, hyperbolic Sobolev spaces

AMS subject classifications. 82C70, 35B45, 35F10

DOI. 10.1137/070698415

1. Introduction.

Consider the kinetic transport equation

$$\partial_t f + v \cdot \nabla_x f = g,$$

where $f, g : R_t \times R_x^d \times R_v^d \to R$. It is well known that averaging with respect to the velocity variable v has a smoothing effect which can be measured in classical Sobolev spaces. For example, in all dimensions, the average over the unit ball

$$\rho_{_B}(t,x) = \int_{|v|\leq 1} f(t,x,v)dv$$

is smoother than f by $1/2$ derivatives in L^2, and we have the estimate

$$\|\rho_{_B}\|_{H^{1/2}(R^{1+d})} \leq C\left(\|f\|_{L^2(R^{1+2d})} + \|g\|_{L^2(R^{1+2d})}\right).$$

Results of this type are known as averaging lemmas. They were first discovered in [24, 23] and developed further by many authors [4, 5, 9, 11, 12, 13, 14, 15, 17, 19, 20, 21, 25, 26, 27, 28, 29, 33, 34, 37]. We refer the reader to [3, 35] for a review and an extensive bibliography.

Averages over spheres

$$\rho_{_S}(t,x) = \int_{S^{d-1}} f(t,x,v)\, d\sigma(v)$$

were first studied in [7] and later in [8]. These averages appear in the equation of radiative transfer, a phenomenon that describes the scattering of photons in a hot medium [1, 2, 10, 22, 36]. They also appear in certain kinetic models of chemotaxis when the velocity of the cells is normalized to $|v| = 1$ [16]. It turns out that in

*Received by the editors July 26, 2007; accepted for publication (in revised form) February 26, 2008; published electronically July 3, 2008.
http://www.siam.org/journals/sima/40-2/69841.html
†School of Mathematics, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JZ, UK (n.bournaveas@ed.ac.uk).
‡School of Mathematics, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK (s.gutierrez@bham.ac.uk).

dimensions $d \geq 3$ averages over spheres gain $1/2$ derivatives in L^2 [7, Theorem 1] and we have the estimate

$$\|\rho_S\|_{H^{1/2}(R^{1+d})} \leq C\left(\|f\|_{L^2(R\times R^d\times S^{d-1})} + \|g\|_{L^2(R\times R^d\times S^{d-1})}\right). \tag{1.1}$$

Thus, if $d \geq 3$, ρ_S gains the same amount of regularity as averages over balls, although it is an average over a lower-dimensional set. Moreover, the estimate for spheres implies the estimate for balls (see Remark 1 in [7]).

In dimension $d = 2$ averages over spheres gain only $1/4$ derivatives, but the "missing" regularity can be recovered in the so-called hyperbolic Sobolev spaces $H^{s,\delta}(R^{1+d})$ [6, 30, 31, 32, 38]. More precisely [7, Theorem 2],

$$\|\rho_S\|_{H^{1/4,1/4}(R^{1+2})} \leq C\left(\|f\|_{L^2(R\times R^2\times S^1)} + \|g\|_{L^2(R\times R^2\times S^1)}\right),$$

where

$$\|F\|_{H^{s,\delta}(R^{1+d})} = \left\|w_+(\tau,\xi)^s w_-(\tau,\xi)^\delta \widehat{F}(\tau,\xi)\right\|_{L^2(R_\tau\times R^d_\xi)}, \tag{1.2}$$

$w_\pm(\tau,\xi) = 1 + ||\tau| \pm |\xi||$, and $\widehat{F}$ is the space-time Fourier transform of F.

In [8] these spaces were used to further improve estimate (1.1). We have [8, Theorem 1], when $d \geq 3$, $d \neq 7$,

$$\|\rho_S\|_{H^{s,\delta}(R^{1+d})} \leq C\,\|f\|_{L^2(R\times R^d\times S^{d-1})} + C\,\|g\|_{L^2(R\times R^d\times S^{d-1})}, \tag{1.3a}$$

provided that

$$s + \delta \leq 1/2,\ s \leq \min\left\{\frac{d-1}{4}, 1\right\}. \tag{1.3b}$$

This allows us to take s larger than the classical $1/2$, provided that we compensate by using a negative δ. Moreover, in regions where $||\tau| - |\xi||$ is small (near the cone $|\tau| = |\xi|$) we have $w_+^s w_-^\delta \simeq w_+^s$, and we gain $s \geq 1/2$ derivatives. When $d = 7$ we have the same estimate but with a logarithmic loss:

$$\|\rho_S\|_{H^{s,\delta}_{\log}(R^{1+d})} \leq C\left(\|f\|_{L^2(R\times R^d\times S^{d-1})} + \|g\|_{L^2(R\times R^d\times S^{d-1})}\right), \quad d = 7,$$

where

$$\|F\|_{H^{s,\delta}_{\log}(R^{1+d})} = \left\|\frac{w_+(\tau,\xi)^s w_-(\tau,\xi)^\delta}{\left(1 + \log\frac{w_+(\tau,\xi)}{w_-(\tau,\xi)}\right)^{1/2}}\widehat{F}(\tau,\xi)\right\|_{L^2(R_\tau\times R^d_\xi)}.$$

Now we turn our attention to equations with right-hand sides containing derivatives with respect to the velocity variable:

$$\partial_t f + v\cdot\nabla_x f = \partial_v^m g.$$

This case is of great interest in applications. For example, the Vlasov part of the Vlasov–Maxwell system has this structure with $m = 1$. It is well known (see [24, 23]) that in this case averages of the form $\int_{R^d} f(t,x,v)\phi(v)dv$, where $\phi(v)$ is a smooth cut-off function, gain $\frac{1}{2(m+1)}$ derivatives in L^2. In particular, if $m = 1$, then the gain is $1/4$ derivatives. The proof of these results uses an integration by parts which removes the velocity derivatives from the function g.

Consider now the case of spheres. In order to be able to integrate by parts on spheres we are forced to restrict ourselves to equations with tangential v-derivatives in the right-hand side:

$$\partial_t f + v \cdot \nabla_x f = \Omega_v^{i,j} g, \tag{1.4}$$

where

$$\Omega_v^{i,j} g = v_i \frac{\partial g}{\partial v_j} - v_j \frac{\partial g}{\partial v_i}, \quad 1 \le i, j \le d.$$

In this case, if $d \ge 3$, the optimal gain is $1/4$ derivatives (the same gain as for balls) [7, Theorem 3]. In view of the discussion above, one would expect the gain in $d = 2$ dimensions to be $1/8$ derivatives. However, the tangential derivatives introduce a special structure (see, for example, (2.3) below and the discussion in [8]) reminiscent of the Klainerman null forms structure for the wave equation [30, 31, 32], which allows for better estimates. By using this observation the expected classical gain of $1/8$ derivatives was improved in [7] to $1/7$ derivatives and further improved in [8] to $1/6$ derivatives, and it was also shown that this result is optimal.

If one considers the hyperbolic Sobolev spaces, it is also possible to prove estimates with a total of $s + \delta = 1/4$ derivatives. In this setting, the two-dimensional case has been previously considered in [7, 8]. More precisely, in [7] it was shown that $\rho_{_S} \in H^{1/16,3/16}(R^{1+2})$. This was improved in [8] to $\rho_{_S} \in H^{1/8,1/8}(R^{1+2})$, and it was also shown that this is optimal.

In this paper we study the regularity of velocity averages over spheres for solutions of (1.4) in hyperbolic Sobolev spaces in dimensions $d \ge 3$. Recall first the following result from [7].

Let $d \ge 3$, and let $f \in L^2(R_t \times R_x^d \times \ S_v^{d-1})$ be a solution of (1.4) with $g \in L^2(R_t \times R_x^d \times \ S_v^{d-1})$. Then the velocity average over the unit sphere $\rho_{_S}(t,x) = \int_{S^{d-1}} f(t,x,v) d\sigma(v)$ satisfies

$$\|\rho_{_S}\|_{H^{1/4}(R\times R^d)} \le C \left[\|f\|_{L^2(R\times R^d\times \ S^{d-1})} + C \|g\|_{L^2(R\times R^d\times \ S^{d-1})} \right]. \tag{1.5}$$

We shall improve this result by showing that $\rho_{_S} \in H^{s,\delta}$ for certain choices of (s,δ) depending on the dimension d. More precisely we have the following.

THEOREM 1.1. *Let $d \ge 3$, and let $f \in L^2(R_t \times R_x^d \times \ S_v^{d-1})$ be a solution of* (1.4) *with $g \in L^2(R_t \times R_x^d \times \ S_v^{d-1})$. Define*

$$\begin{aligned} (s,\delta) &= \left(\frac{d-1}{6}, -\frac{2d-5}{12} \right) \quad if\ d \in \{3,4,5\}, \\ (s,\delta) &= \left(\frac{4}{5}, -\frac{11}{20} \right) \quad if\ d = 6, \\ (s,\delta) &= \left(1, -\frac{3}{4} \right), \quad if\ d \ge 7. \end{aligned} \tag{1.6}$$

Then the velocity average $\rho_{_S}(t,x) = \int_{S^{d-1}} f(t,x,v) d\sigma(v)$ satisfies

$$\|\rho_{_S}\|_{H^{s,\delta}(R^{1+d})} \le C \|f\|_{L^2(R\times R^d\times \ S^{d-1})} + C \|g\|_{L^2(R\times R^d\times \ S^{d-1})}, \ d \ne 5, 7, \tag{1.7a}$$

$$\|\rho_{_S}\|_{H^{s,\delta}_{\log}(R^{1+d})} \le C \|f\|_{L^2(R\times R^d\times \ S^{d-1})} + C \|g\|_{L^2(R\times R^d\times \ S^{d-1})}, \ d = 5, 7. \tag{1.7b}$$

Notice that in all cases we have $s+\delta = 1/4$, with $s > 1/4$ and $\delta < 0$. By comparing this result to the classical estimate (1.5) we see that it is possible to improve on the

"good derivative" w_+, provided that we make a sacrifice in the derivative w_-. Observe also that since $\delta = 1/4 - s$ and $s > 1/4$ we have

$$w_+^s w_-^\delta = w_+^{\frac{1}{4}} \left(\frac{w_+}{w_-}\right)^{s-\frac{1}{4}} \geq w_+^{\frac{1}{4}},$$

and therefore estimate (1.7) implies estimate (1.5). Moreover, near the cone $|\tau| = |\xi|$ we have $w_+^s w_-^\delta \simeq w_+^s$, because $w_- \simeq 1$, and therefore our estimate says that we gain $s > 1/4$ derivatives in this region. Notice also that, in contrast to (1.3), we gain even in $d = 3$ dimensions. This is due to the fact that we have not only the usual averaging smoothing effect and the flexibility of the $H^{s,\delta}$-spaces but also the special structure of the tangential derivatives.

It is possible to extend the validity of estimate (1.7) to a whole region of pairs (s, δ). We have the following.

THEOREM 1.2. *Let $d \geq 3$, and let $f \in L^2(R_t \times R_x^d \times S_v^{d-1})$ be a solution of* (1.4) *with $g \in L^2(R_t \times R_x^d \times S_v^{d-1})$. Let $s, \delta \in R$ be such that*

$$s + \delta \leq \frac{1}{4}, \tag{1.8a}$$

$$s \leq \frac{d-1}{6} \quad if\ d \in \{3,4,5\}, \tag{1.8b}$$

$$s \leq \frac{4}{5} \quad if\ d = 6, \tag{1.8c}$$

$$s \leq 1 \quad if\ d \geq 7. \tag{1.8d}$$

Then estimates (1.7) *are true.*

In the region $\{|\tau| > |\xi|\}$ in phase space the symbol of the operator $\partial_t + v \cdot \nabla_x$ satisfies $|\tau + v \cdot \xi| \geq |\tau| - |\xi| > 0$, and therefore it is reasonable to expect better estimates. Indeed, improved estimates in this region are contained in [7, 8]. We shall establish the following improvements to the results of Theorem 1.1. First, estimates (1.7) are valid in $\{|\tau| > |\xi|\}$ with the same s and with $\delta = 0$ (instead of a negative δ).

THEOREM 1.3. *Let $d \geq 3$, and let $f \in L^2(R_t \times R_x^d \times S_v^{d-1})$ be a solution of* (1.4) *with $g \in L^2(R_t \times R_x^d \times S_v^{d-1})$. Define*

$$s = \frac{d-1}{6} \quad if\ d \in \{3,4,5\}, \tag{1.9a}$$

$$s = \frac{4}{5} \quad if\ d = 6, \tag{1.9b}$$

$$s = 1 \quad if\ d \geq 7. \tag{1.9c}$$

Then the velocity average $\rho_s(t,x) = \int_{S^{d-1}} f(t,x,v) d\sigma(v)$ satisfies the following estimates in the classical Sobolev spaces $H^s(\{|\tau| > |\xi|\})$.

If $d \neq 5, 7$,

$$\|w_+(\tau,\xi)^s \widehat{\rho_s}(\tau,\xi)\|_{L^2(\{|\tau|>|\xi|\})} \tag{1.10a}$$

$$\leq C\left(\|f\|_{L^2(R\times R^d\times S^{d-1})} + \|g\|_{L^2(R\times R^d\times S^{d-1})}\right). \tag{1.10b}$$

If $d = 5, 7$,

$$\left\| w_+(\tau,\xi)^s \left[1 + \log \frac{w_+(\tau,\xi)}{w_-(\tau,\xi)}\right]^{-1/2} \widehat{\rho_s}(\tau,\xi) \right\|_{L^2(\{|\tau|>|\xi|\})} \tag{1.10c}$$

$$\leq C\|f\|_{L^2(R\times R^d\times S^{d-1})} + C\|g\|_{L^2(R\times R^d\times S^{d-1})}. \tag{1.10d}$$

Moreover, it is possible to prove $H^{s,\delta}$-estimates in all dimensions $d \geq 3$ with $s+\delta = 1$ and $s, \delta \geq 0$. Precisely, we have the following.

THEOREM 1.4. *Under the same assumptions and notation as in Theorem* 1.3 *and with*

$$(s,\delta) = \left(\frac{d-3}{4}, \frac{7-d}{4}\right) \quad if\ d \in \{3,4,5\}, \tag{1.11a}$$

$$(s,\delta) = \left(\frac{3}{4}, \frac{1}{4}\right) \quad if\ d = 6, \tag{1.11b}$$

$$(s,\delta) = (1,0) \quad if\ d \geq 7, \tag{1.11c}$$

we have

$$\left\| w_+(\tau,\xi)^s w_-(\tau,\xi)^\delta \widehat{\rho_S}(\tau,\xi) \right\|_{L^2(\{|\tau|>|\xi|\})} \tag{1.12a}$$

$$\leq C\|f\|_{L^2(R\times R^d\times\ S^{d-1})} + C\|g\|_{L^2(R\times R^d\times\ S^{d-1})}\ \ if\ d \neq 5,7; \tag{1.12b}$$

$$\left\| w_+(\tau,\xi)^s w_-(\tau,\xi)^\delta \left[1+\log\frac{w_+(\tau,\xi)}{w_-(\tau,\xi)}\right]^{-1/2} \widehat{\rho_S}(\tau,\xi) \right\|_{L^2(\{|\tau|>|\xi|\})} \tag{1.12c}$$

$$\leq C\|f\|_{L^2(R\times R^d\times\ S^{d-1})} + C\|g\|_{L^2(R\times R^d\times\ S^{d-1})} \quad if\ d = 5,7. \tag{1.12d}$$

Remark. Regarding the optimality of the conditions on s and δ we will show in the last section that all of the upper bounds for s and $s+\delta$ are optimal with the possible exception of the bound $s \leq 4/5$ in dimension $d = 6$. In some cases it is possible to have improved estimates if we allow weights of the form $w_+^{s_1} w_-^{\delta_1} + w_+^{s_2} w_-^{\delta_2}$; see page 664 for details. The logarithmic divergence in some of the estimates is due to the logarithmic divergence in Lemmas 3.2 and 3.3 in [8].

Notation. We define the weights $w_\pm$ by $w_\pm(\tau,\xi) = 1 + ||\tau| \pm |\xi||$. We denote the classical Sobolev spaces by H^s and the hyperbolic Sobolev spaces (see (1.2)) by $H^{s,\delta}$. We use $\widehat{\ }$ for the Fourier transform in space-time denoting the dual variables by (τ,ξ). We will always average the solution f over the unit sphere S^{d-1} in R^d and denote the average by ρ_s.

2. Proofs of Theorems 1.1 and 1.2. In this section we prove Theorems 1.1 and 1.2. We need the following result from [8].

PROPOSITION 2.1. *Let $m > -1$, $l > 1/2$, and define*

$$J_l^m(\tau,\xi) = \int_0^\pi \frac{(\sin\theta)^m}{[1+(\tau+|\xi|\cos\theta)^2]^l}\, d\theta, \qquad \tau \in R,\ \xi \in R^d. \tag{2.1}$$

Let $\alpha = \min\left\{\frac{m+1-4l}{2}, 0\right\}$. Then the integrals $J_l^m(\tau,\xi)$ satisfy the following pointwise estimates:

$$J_l^m(\tau,\xi) \leq C\frac{w_-(\tau,\xi)^{2l-1+\alpha}}{w_+(\tau,\xi)^{2l+\alpha}}\ \ if\ m+1 \neq 4l,$$

$$J_l^m(\tau,\xi) \leq C\frac{w_-(\tau,\xi)^{2l-1}}{w_+(\tau,\xi)^{2l}}\left(1+\log\frac{w_+(\tau,\xi)}{w_-(\tau,\xi)}\right)\ \ if\ m+1 = 4l.$$

Proof of Theorem 1.1. By taking the space-time Fourier transform of (1.4) and adding $\lambda\widehat{f}(\tau,\xi,v)$ to both sides, where $\lambda = \lambda(\tau,\xi) > 0$ will be determined later,

we get

$$\widehat{f}(\tau,\xi,v)=\frac{\Omega_v^{i,j}\widehat{g}(\tau,\xi,v)}{\lambda+i\,(\tau+v\cdot\xi)}+\lambda\frac{\widehat{f}(\tau,\xi,v)}{\lambda+i\,(\tau+v\cdot\xi)}.$$

By averaging in v and integrating by parts we obtain

$$\begin{aligned}\widehat{\rho}(\tau,\xi)&=\int_{S^{d-1}}\frac{\Omega_v^{i,j}\widehat{g}(\tau,\xi,v)}{\lambda+i\,(\tau+v\cdot\xi)}\,d\sigma(v)+\ \lambda\int_{S^{d-1}}\frac{\widehat{f}(\tau,\xi,v)}{\lambda+i\,(\tau+v\cdot\xi)}\,d\sigma(v)\\&=i\int_{S^{d-1}}\frac{v_i\xi_j-v_j\xi_i}{\left[\lambda+i\,(\tau+v\cdot\xi)\right]^2}\widehat{g}(\tau,\xi,v)d\sigma(v)+\ \lambda\int_{S^{d-1}}\frac{\widehat{f}(\tau,\xi,v)d\sigma(v)}{\lambda+i\,(\tau+v\cdot\xi)}.\end{aligned}$$

By using the Cauchy–Schwarz inequality we get

$$\begin{aligned}|\widehat{\rho}(\tau,\xi)|\ &\leq\left(\int_{S^{d-1}}|\widehat{g}(\tau,\xi,v)|^2\,d\sigma(v)\right)^{1/2}\cdot K(\tau,\xi)\\&\quad+\left(\int_{S^{d-1}}\left|\widehat{f}(\tau,\xi,v)\right|^2\,d\sigma(v)\right)^{1/2}\cdot L(\tau,\xi),\end{aligned}\tag{2.2}$$

where

$$\begin{aligned}K(\tau,\xi)&=\left(\int_{S^{d-1}}\frac{|v_i\xi_j-v_j\xi_i|^2}{|\lambda+i\,(\tau+v\cdot\xi)|^4}\,d\sigma(v)\right)^{1/2},\\L(\tau,\xi)&=\left(\lambda^2\int_{S^{d-1}}\frac{d\sigma(v)}{|\lambda+i\,(\tau+v\cdot\xi)|^2}\right)^{1/2}.\end{aligned}\tag{2.3}$$

We estimate $K(\tau,\xi)$ and $L(\tau,\xi)$ as follows. Set $\tau'=\frac{\tau}{\lambda},\ \xi'=\frac{\xi}{\lambda}$, and let θ be the angle between ξ and v. Then

$$\begin{aligned}K(\tau,\xi)&=\left(\int_{S^{d-1}}\frac{|v_i\xi_j-v_j\xi_i|^2}{|\lambda+i\,(\tau+v\cdot\xi)|^4}\,d\sigma(v)\right)^{1/2}\\&=\left(\int_{S^{d-1}}\frac{|v_i\xi_j-v_j\xi_i|^2}{\left|\lambda^2+(\tau+v\cdot\xi)^2\right|^2}\,d\sigma(v)\right)^{1/2}\\&=\frac{|\xi|}{\lambda^2}\left(\int_{S^{d-1}}\frac{\left|v_i\frac{\xi_j}{|\xi|}-v_j\frac{\xi_i}{|\xi|}\right|^2}{\left|1+(\tau'+v\cdot\xi')^2\right|^2}\,d\sigma(v)\right)^{1/2}\\&\simeq\frac{|\xi|}{\lambda^2}\left(\int_0^\pi\frac{(\sin\theta)^2\,(\sin\theta)^{d-2}}{\left|1+(\tau'+|\xi'|\cos\theta)^2\right|^2}\,d\theta\right)^{1/2}\\&=\frac{|\xi|}{\lambda^2}J_2^d(\tau',\xi')^{1/2},\end{aligned}\tag{2.4}$$

and

$$
\begin{aligned}
L(\tau,\xi) &= \left(\lambda^2 \int_{S^{d-1}} \frac{d\sigma(v)}{\lambda^2 + (\tau + v\cdot\xi)^2} \right)^{1/2} \\
&= \left(\int_{S^{d-1}} \frac{d\sigma(v)}{1 + (\tau' + v\cdot\xi')^2} \right)^{1/2} \\
&\simeq \left(\int_0^\pi \frac{(\sin\theta)^{d-2}}{1 + (\tau' + |\xi'|\cos\theta)^2} d\theta \right)^{1/2} \\
&= J_1^{d-2}(\tau',\xi')^{1/2}. \qquad (2.5)
\end{aligned}
$$

We start with the case $d \in \{3,4,5\}$. From (2.4) we have $K(\tau,\xi) \le C\frac{|\xi|}{\lambda^2} J_2^d(\tau',\xi')^{1/2}$. Proposition 2.1 with $m = d$, $l = 2$ gives

$$
\begin{aligned}
K(\tau,\xi) &\le C \frac{|\xi|}{\lambda^2} \frac{(1 + ||\tau'| - |\xi'||)^{\frac{d-1}{4}}}{(1 + |\tau'| + |\xi'|)^{\frac{d+1}{4}}} \\
&= C \frac{|\xi|}{\lambda^{\frac{3}{2}}} \frac{(\lambda + ||\tau| - |\xi||)^{\frac{d-1}{4}}}{(\lambda + |\tau| + |\xi|)^{\frac{d+1}{4}}} \\
&\le C \frac{(\lambda + ||\tau| - |\xi||)^{\frac{d-1}{4}}}{\lambda^{\frac{3}{2}} (\lambda + |\tau| + |\xi|)^{\frac{d-3}{4}}} \\
&\le C \frac{(\lambda + w_-)^{\frac{d-1}{4}}}{\lambda^{\frac{3}{2}} (w_+)^{\frac{d-3}{4}}},
\end{aligned}
$$

where we have set $w_\pm = 1 + ||\tau| \pm |\xi||$ and we have used the fact that $\lambda > 1$ (this will be guaranteed by the choice of λ; see (2.9)).

Next we deal with $L(\tau,\xi)$. From estimate (2.5) we have $L(\tau,\xi) \le J_1^{d-2}(\tau',\xi')^{1/2}$, and we now apply Proposition 2.1 with $m = d-2$ and $l = 1$. Notice that $m + 1 = 4l$ when $d = 5$, and hence there is an extra logarithmic term in this case. For $d \in \{3,4\}$ we get

$$
\begin{aligned}
L(\tau,\xi) &\le C \frac{(1 + ||\tau'| - |\xi'||)^{\frac{d-3}{4}}}{(1 + |\tau'| + |\xi'|)^{\frac{d-1}{4}}} \\
&= C \frac{\lambda^{1/2} (\lambda + ||\tau| - |\xi||)^{\frac{d-3}{4}}}{(\lambda + |\tau| + |\xi|)^{\frac{d-1}{4}}} \\
&\le C \frac{\lambda^{1/2} (\lambda + w_-)^{\frac{d-3}{4}}}{(w_+)^{\frac{d-1}{4}}}, \qquad (2.6a)
\end{aligned}
$$

and for $d = 5$ we get

$$
L(\tau,\xi) \le C \frac{\lambda^{1/2} (\lambda + w_-)^{\frac{d-3}{4}}}{(w_+)^{\frac{d-1}{4}}} \left(1 + \log \frac{\lambda + |\tau| + |\xi|}{\lambda + ||\tau| - |\xi||} \right)^{1/2}. \qquad (2.6b)
$$

We choose $\lambda = \lambda(\tau,\xi)$ such that

$$
\frac{(\lambda + w_-)^{\frac{d-1}{4}}}{\lambda^{\frac{3}{2}} (w_+)^{\frac{d-3}{4}}} = \frac{\lambda^{1/2} (\lambda + w_-)^{\frac{d-3}{4}}}{(w_+)^{\frac{d-1}{4}}} \qquad (2.7)
$$

or, equivalently,

$$\lambda^4 - w_+\lambda - w_+w_- = 0. \tag{2.8}$$

We need to know that such a λ exists, and we also need an estimate of λ in terms of the weights $1 + ||\tau| \pm |\xi||$.

Consider the function $f(\lambda) := \lambda^4 - w_+\lambda - w_+w_-$. Clearly, $f(1) < 0$. On the other hand,

$$\begin{aligned} f\left(2w_+^{1/3}w_-^{1/6}\right) &= 16w_+^{4/3}w_-^{2/3} - 2w_+^{4/3}w_-^{1/6} - w_+w_- \\ &= \left(8w_+^{4/3}w_-^{2/3} - 2w_+^{4/3}w_-^{1/6}\right) + \left(8w_+^{4/3}w_-^{2/3} - w_+w_-\right) \\ &= 2w_+^{4/3}w_-^{1/6}\left(4w_-^{1/2} - 1\right) + w_+w_-^{2/3}\left(8w_+^{1/3} - w_-^{1/3}\right) \\ &> 0. \end{aligned}$$

It follows that there exists a λ with

$$1 < \lambda < 2w_+^{1/3}w_-^{1/6} \tag{2.9}$$

such that (2.7) and (2.8) are satisfied. From (2.8) we have $\lambda + w_- = \frac{\lambda^4}{w_+}$; therefore

$$\lambda + w_- < \frac{\left(2w_+^{1/3}w_-^{1/6}\right)^4}{w_+} \leq Cw_+^{1/3}w_-^{2/3}. \tag{2.10}$$

It follows that the right-hand side of (2.7) is bounded by

$$\frac{\left(2w_+^{1/3}w_-^{1/6}\right)^{1/2}\left(w_+^{1/3}w_-^{2/3}\right)^{\frac{d-3}{4}}}{(w_+)^{\frac{d-1}{4}}} \leq Cw_+^{-\frac{d-1}{6}}w_-^{\frac{2d-5}{12}}.$$

Moreover $1 < \lambda \leq Cw_+$, and therefore in (2.6b)

$$1 + \log\frac{\lambda + |\tau| + |\xi|}{\lambda + ||\tau| - |\xi||} \leq C\left(1 + \log\frac{w_+}{w_-}\right).$$

We conclude that $K(\tau,\xi)$ satisfies

$$K(\tau,\xi) \leq Cw_+^{-\frac{d-1}{6}}w_-^{\frac{2d-5}{12}}, \quad d \in \{3,4,5\},$$

while $L(\tau,\xi)$ satisfies

$$L(\tau,\xi) \leq Cw_+^{-\frac{d-1}{6}}w_-^{\frac{2d-5}{12}}, \quad d \in \{3,4\},$$

and

$$L(\tau,\xi) \leq Cw_+^{-\frac{d-1}{6}}w_-^{\frac{2d-5}{12}}\left(1 + \log\frac{w_+}{w_-}\right)^{1/2}, \quad d = 5.$$

Next we deal with all dimensions $d \geq 7$. From estimate (2.4) we have $K(\tau,\xi) \leq C\frac{|\xi|}{\lambda^2}J_2^d(\tau',\xi')^{1/2}$. Proposition 2.1 with $m = d$ and $l = 2$ now gives

$$\alpha = \min\left\{\frac{m+1-4l}{2}, 0\right\} = \min\left\{\frac{d-7}{2}, 0\right\} = 0.$$

Notice, however, that $m+1=4l$ when $d=7$, and therefore there is a logarithmic loss in this case. We get

$$\begin{aligned} K(\tau,\xi) &\le C\frac{|\xi|}{\lambda^2}\left(\frac{\left(1+||\tau'|-|\xi'||\right)^3}{\left(1+|\xi'|+|\tau'|\right)^4}\right)^{1/2} \\ &= C\frac{|\xi|}{\lambda^2}\cdot\frac{\lambda^2\ \left(\lambda+||\tau|-|\xi||\right)^{3/2}}{\lambda^{3/2}\ \left(\lambda+|\tau|+|\xi|\right)^2} \\ &\le C\frac{\left(\lambda+||\tau|-|\xi||\right)^{3/2}}{\lambda^{3/2}\ \left(\lambda+|\tau|+|\xi|\right)} \qquad \text{if } d>7 \end{aligned}$$

and

$$K(\tau,\xi)\le C\frac{\left(\lambda+||\tau|-|\xi||\right)^{3/2}}{\lambda^{3/2}\ \left(\lambda+|\tau|+|\xi|\right)}\left(1+\log\frac{\lambda+|\tau|+|\xi|}{\lambda+||\tau|-|\xi||}\right)^{1/2} \quad \text{if } d=7.$$

From (2.5) we have $L(\tau,\xi)\le CJ_1^{d-2}(\tau',\xi')^{1/2}$. By applying Proposition 2.1, this time with $m=d-2$, $l=1$, and $\alpha=\min\left\{\frac{m+1-4l}{2},0\right\}=0$ (notice that $m+1\neq 4l$, so there is no logarithmic loss for L), we get

$$\begin{aligned} L(\tau,\xi) &\le C\left(\frac{1+||\tau'|-|\xi'||}{\left(1+|\tau'|+|\xi'|\right)^2}\right)^{1/2} \\ &= C\frac{\lambda^{1/2}\ \left(\lambda+||\tau|-|\xi||\right)^{1/2}}{\left(\lambda+|\tau|+|\xi|\right)}. \end{aligned}$$

We need to choose λ such that

$$\frac{\left(\lambda+||\tau|-|\xi||\right)^{3/2}}{\lambda^{3/2}\ \left(\lambda+|\tau|+|\xi|\right)}=\frac{\lambda^{1/2}\ \left(\lambda+||\tau|-|\xi||\right)^{1/2}}{\left(\lambda+|\tau|+|\xi|\right)} \tag{2.11}$$

or, equivalently, $\lambda^2-\lambda-||\tau|-|\xi||=0$. We choose $\lambda=\frac{1+\sqrt{1+4||\tau|-|\xi||}}{2}$. Notice that $\lambda\ge 1$ and that $\lambda\le C(+\sqrt{||\tau|-|\xi||})$, and therefore

$$\lambda+||\tau|-|\xi||\le C\left(1+\sqrt{||\tau|-|\xi||}+||\tau|-|\xi||\right)\simeq 1+||\tau|-|\xi||\,.$$

Therefore the right-hand side of (2.11) is bounded by

$$\frac{\left(1+\sqrt{||\tau|-|\xi||}\right)^{1/2}\left(1+||\tau|-|\xi||\right)^{1/2}}{1+|\tau|+|\xi|}\le C\frac{\left(1+||\tau|-|\xi||\right)^{3/4}}{1+|\tau|+|\xi|}. \tag{2.12}$$

Moreover, since $1<\lambda\le C\left(1+|\tau|+|\xi|\right)$, we have

$$\log\frac{\lambda+|\tau|+|\xi|}{\lambda+||\tau|-|\xi||}\le C\log\frac{1+|\tau|+|\xi|}{1+||\tau|-|\xi||}. \tag{2.13}$$

By putting everything together we conclude that

$$\begin{aligned} &K(\tau,\xi)\le C\frac{\left(1+||\tau|-|\xi||\right)^{3/4}}{1+|\tau|+|\xi|} && \text{if } d>7, \\ &K(\tau,\xi)\le C\frac{\left(1+||\tau|-|\xi||\right)^{3/4}}{1+|\tau|+|\xi|}\left(1+\log\frac{1+|\tau|+|\xi|}{1+||\tau|-|\xi||}\right)^{1/2}, && \text{if } d=7, \end{aligned}$$

and

$$L(\tau,\xi) \leq C\frac{(1+||\tau|-|\xi||)^{3/4}}{1+|\tau|+|\xi|},\ d\geq 7.$$

This completes the proof in the case $d \geq 7$.

It remains to deal with the intermediate case $d = 6$. In this case we have the extra difficulty that K still behaves as in the lower-dimensional case but L behaves as in the higher-dimensional case. Indeed, we have $K(\tau,\xi) \leq C\frac{|\xi|}{\lambda^2}J_2^d(\tau',\xi')^{1/2}$, and Proposition 2.1 gives $\alpha = \min\left\{\frac{d-7}{2},0\right\} = -\frac{1}{2}$; therefore

$$\begin{aligned} K(\tau,\xi) &\leq C\frac{|\xi|}{\lambda^2}\cdot\frac{(1+||\tau'|-|\xi'||)^{5/4}}{(1+|\tau'|+|\xi'|)^{7/4}} \\ &= \frac{|\xi|}{\lambda^2}\cdot\frac{\lambda^{7/4}(\lambda+||\tau|-|\xi||)^{5/4}}{\lambda^{5/4}(\lambda+|\tau|+|\xi|)^{7/4}} \\ &\leq C\frac{(\lambda+||\tau|-|\xi||)^{5/4}}{\lambda^{3/2}(\lambda+|\tau|+|\xi|)^{3/4}}. \end{aligned}$$

For $L(\tau,\xi)$ we have $L(\tau,\xi) \leq CJ_1^{d-2}(\tau',\xi')^{1/2}$, and Proposition 2.1 now gives $\alpha = \min\left\{\frac{d-5}{2},0\right\} = 0$. We get

$$\begin{aligned} L(\tau,\xi) &\leq C\frac{(1+||\tau'|-|\xi'||)^{1/2}}{1+|\tau'|+|\xi'|} \\ &= C\frac{\lambda^{1/2}\,(\lambda+||\tau|-|\xi||)^{1/2}}{\lambda+|\tau|+|\xi|}. \end{aligned}$$

We need to choose λ so that

$$\frac{(\lambda+||\tau|-|\xi||)^{5/4}}{\lambda^{3/2}(\lambda+|\tau|+|\xi|)^{3/4}} = \frac{\lambda^{1/2}\,(\lambda+||\tau|-|\xi||)^{1/2}}{\lambda+|\tau|+|\xi|} \tag{2.14}$$

or, equivalently,

$$\lambda^8 - (\lambda+||\tau|-|\xi||)^3\,(\lambda+|\tau|+|\xi|) = 0. \tag{2.15}$$

To see that such a λ exists and to obtain estimates for it, consider the function

$$f:[1,\infty)\to R,\quad f(\lambda) = \lambda^8 - (\lambda+||\tau|-|\xi||)^3(\lambda+|\tau|+|\xi|).$$

Then, on the one hand, $f(1) \leq 0$ and, on the other hand, with $w_\pm = 1+||\tau|\pm|\xi||$, we have

$$f\left(2w_-^{4/5}w_+^{1/5}\right) = 2^8 w_-^{32/5}w_+^{8/5} \tag{2.16}$$

$$-\left(2w_-^{4/5}w_+^{1/5}+||\tau|-|\xi||\right)^3\left(2w_-^{4/5}w_+^{1/5}+|\tau|+|\xi|\right). \tag{2.17}$$

We have

$$2w_-^{4/5}w_+^{1/5}+|\xi|-|\tau| \leq 2w_-^{4/5}w_+^{1/5}+w_- \leq 3w_-^{4/5}w_+^{1/5}$$

and

$$2w_-^{4/5}w_+^{1/5}+|\xi|+|\tau| \leq 3w_+,$$

and therefore

$$\begin{aligned} f\left(2w_-^{4/5}w_+^{1/5}\right) &\geq 2^8 w_-^{32/5}w_+^{8/5} - \left(3w_-^{4/5}w_+^{1/5}\right)^3 \cdot 3w_+ \\ &= 2^8 w_-^{32/5}w_+^{8/5} - 3^4 w_-^{12/5}w_+^{8/5} \\ &= w_+^{8/5}\left(2^8 w_-^{32/5} - 3^4 w_-^{12/5}\right) \\ &> 0. \end{aligned}$$

It follows that there exists a λ with $1 \leq \lambda \leq 2w_-^{4/5}w_+^{1/5}$ such that both (2.14) and (2.15) are satisfied. Notice that for this λ we have

$$\lambda + ||\tau| - |\xi|| \leq 2w_-^{4/5}w_+^{1/5} + w_- \leq 3w_-^{4/5}w_+^{1/5}. \tag{2.18}$$

Moreover, from (2.15) we get

$$\lambda = \left(\lambda + ||\tau| - |\xi||\right)^{3/8}\left(\lambda + |\tau| + |\xi|\right)^{1/8}. \tag{2.19}$$

By using first (2.19) we find that the right-hand side of (2.14) is bounded by

$$\begin{aligned} &\frac{\left((\lambda + ||\tau| - |\xi||)^{3/8}(\lambda + |\tau| + |\xi|)^{1/8}\right)^{1/2}(\lambda + ||\tau| - |\xi||)^{1/2}}{\lambda + |\tau| + |\xi|} \\ &= \frac{(\lambda + ||\tau| - |\xi||)^{11/16}}{(\lambda + |\tau| + |\xi|)^{15/16}}. \end{aligned}$$

Next we use (2.18) for the numerator and $\lambda \geq 1$ for the denominator to get

$$\cdots \leq C\frac{\left(w_-^{4/5}w_+^{1/5}\right)^{11/16}}{w_+^{15/16}} = C\frac{w_-^{11/20}}{w_+^{4/5}}.$$

We conclude that

$$K(\tau,\xi) \leq C\frac{w_-^{11/20}}{w_+^{4/5}}, \quad L(\tau,\xi) \leq C\frac{w_-^{11/20}}{w_+^{4/5}}.$$

This completes the proof of Theorem 1.1. □

Proof of Theorem 1.2. Let (s_0, δ_0) be defined by

$$(s_0,\delta_0) = \left(\frac{d-1}{6}, -\frac{2d-5}{12}\right) \quad \text{if } d \in \{3,4,5\}, \tag{2.20a}$$

$$(s_0,\delta_0) = \left(\frac{4}{5}, -\frac{11}{20}\right) \quad \text{if } d = 6, \tag{2.20b}$$

$$(s_0,\delta_0) = \left(1, -\frac{3}{4}\right) \quad \text{if } d \geq 7. \tag{2.20c}$$

Notice that in all cases we have $s_0 + \delta_0 = \frac{1}{4}$. Then the conditions (1.8) in Theorem 1.2 amount to $s + \delta \leq \frac{1}{4}$ and $s \leq s_0$. We have

$$w_+^s w_-^\delta \leq w_+^s w_-^{1/4-s} = w_+^{s_0}w_-^{\delta_0}\frac{w_-^{1/4-\delta_0-s}}{w_+^{s_0-s}} = w_+^{s_0}w_-^{\delta_0}\left(\frac{w_-}{w_+}\right)^{s_0-s} \leq w_+^{s_0}w_-^{\delta_0}.$$

The result then follows from Theorem 1.1. □

Remark. It is possible to improve the estimates in this section if we allow weights of the form $w_+^{s_1} w_-^{\delta_1} + w_+^{s_2} w_-^{\delta_2}$. Consider, for example, the case $d \in \{3,4,5\}$. In the proof of Theorem 1.1 we bounded both K and L by

$$B := \frac{\lambda^{\frac{1}{2}} (\lambda + w_-)^{\frac{d-3}{4}}}{w_+^{\frac{d-1}{4}}},$$

with a logarithmic loss if $d = 5$. From (2.8), $\lambda = (\lambda + w_-)^{1/4} w_+^{1/4}$, and hence

$$B \leq C \frac{(\lambda + w_-)^{\frac{2d-5}{8}}}{w_+^{\frac{2d-3}{8}}}.$$

By using $(\lambda + w_-)^{2d-5} \leq C\lambda^{2d-5} + w_-^{2d-5}$ and $\lambda \leq C w_-^{1/6} w_+^{1/3}$ (see (2.9)) we get

$$B \leq C \frac{w_-^{\frac{2d-5}{48}}}{w_+^{\frac{d-1}{6}}} + \frac{w_-^{\frac{2d-5}{8}}}{w_+^{\frac{2d-3}{8}}}.$$

Each of these fractions is smaller than

$$\frac{w_-^{\frac{2d-5}{12}}}{w_+^{\frac{d-1}{6}}}.$$

3. Improved estimates in the region $\{|\tau| > |\xi|\}$. In this section we prove Theorems 1.3 and 1.4.

Proof of Theorems 1.3 and 1.4. By arguing, and using the same notation, as in the proof of Theorem 1.1, we have (see (2.2)–(2.5))

$$\begin{aligned} |\widehat{\rho}(\tau,\xi)| \leq & \left(\int_{S^{d-1}} |\widehat{g}(\tau,\xi,v)|^2 \, d\sigma(v) \right)^{1/2} \cdot K(\tau,\xi) \\ & + \left(\int_{S^{d-1}} \left| \widehat{f}(\tau,\xi,v) \right|^2 \, d\sigma(v) \right)^{1/2} \cdot L(\tau,\xi), \end{aligned} \tag{3.1}$$

where

$$K(\tau,\xi) \simeq \frac{|\xi|}{\lambda^2} J_2^d(\tau',\xi')^{1/2}, \qquad L(\tau,\xi) \simeq J_1^{d-2}(\tau',\xi')^{1/2},$$

with $\tau' = \tau/\lambda$ and $\xi' = \xi/\lambda$.

In what follows we will reduce to proving pointwise estimates for $K(\tau,\xi)$ and $L(\tau,\xi)$. As in the proof of Theorem 1.1, the estimates announced in Theorems 1.3 and 1.4 are a direct consequence of these estimates and (3.1), so we will omit the details here.

Recall from [8] that, when $|\tau| > |\xi|$, the integrals $J_l^m(\tau,\xi)$ defined in (2.1) satisfy

$$J_l^m(\tau,\xi) \simeq \frac{(1 + |\tau| - |\xi|)^\alpha}{(1 + |\tau| + |\xi|)^{2l+\alpha}} \quad \text{if} \quad m + 1 \neq 4l,$$

$$J_l^m(\tau,\xi) \simeq \frac{1}{(1 + |\tau| + |\xi|)^{2l}} \left(1 + \log \frac{1 + |\tau| + |\xi|}{1 + |\tau| - |\xi|} \right) \quad \text{if} \quad m + 1 = 4l,$$

where $\alpha = \min((m+1-4l)/2, 0)$.

We may assume that $|\tau| \geq 1$. First, observe that if $|\tau| \geq 2|\xi|$, then

$$J_l^m(\tau,\xi) \simeq \frac{(1+|\tau|-|\xi|)^\alpha}{(1+|\tau|+|\xi|)^{2l+\alpha}} \simeq \frac{1}{|\tau|^{2l}}$$

and $1+\log\frac{1+|\tau|+|\xi|}{1+|\tau|-|\xi|} \simeq 1$; therefore in all dimensions (use $\lambda = 1$) we have

$$K(\tau,\xi) \simeq |\xi| J_2^d(\tau,\xi)^{1/2} \simeq \frac{|\xi|}{|\tau|^2} \leq C\frac{1}{|\tau|}$$

and

$$L(\tau,\xi) \simeq J_1^{d-2}(\tau,\xi)^{1/2} \leq C\frac{1}{|\tau|}$$

which is stronger than all estimates in (1.10) and (1.12). Thus we may assume that $|\xi| < |\tau| < 2|\xi|$, in which case $1+|\tau|+|\xi| \simeq |\xi|$.

Suppose that $d \in \{3,4,5\}$. Then we have

$$K(\tau,\xi) \simeq \frac{|\xi|}{\lambda^2} J_2^d(\tau',\xi')^{1/2} \leq C\frac{1}{(\lambda+|\tau|-|\xi|)^{\frac{7-d}{4}}|\xi|^{\frac{d-3}{4}}}.$$

Similarly,

$$L(\tau,\xi) \simeq J_1^{d-2}(\tau',\xi')^{1/2} \leq C\frac{\lambda}{(\lambda+|\tau|-|\xi|)^{\frac{5-d}{4}}|\xi|^{\frac{d-1}{4}}} \qquad \text{if} \quad d=3,4,$$

and the same estimate with a logarithmic loss if $d=5$. Choose $\lambda = 1$ to get the estimates in Theorem 1.4. On the other hand, there exists a λ such that

$$\frac{1}{(\lambda+|\tau|-|\xi|)^{\frac{7-d}{4}}|\xi|^{\frac{d-3}{4}}} = \frac{\lambda}{(\lambda+|\tau|-|\xi|)^{\frac{5-d}{4}}|\xi|^{\frac{d-1}{4}}}. \tag{3.2}$$

Indeed, this is equivalent to

$$\lambda\,(\lambda+|\tau|-|\xi|)^{\frac{1}{2}} = |\xi|^{\frac{1}{2}}, \tag{3.3}$$

i.e., $\varphi(\lambda) = 0$, where $\varphi(\lambda) = \lambda^3 + (|\tau|-|\xi|)\lambda^2 - |\xi|$. We have $\varphi(0) = -|\xi| < 0$ and $\varphi(|\xi|^{1/3}) = (|\tau|-|\xi|)|\xi|^{2/3} > 0$, and therefore there exists a $\lambda \in (0, |\xi|^{1/3})$ such that (3.2) is satisfied. For this λ, (3.3) gives $\lambda + |\tau| - |\xi| = \frac{|\xi|}{\lambda^2} \geq |\xi|^{1/3}$, and therefore the left-hand side of (3.2) is no greater than

$$\frac{1}{|\xi|^{\frac{1}{3}\cdot\frac{7-d}{4}}|\xi|^{\frac{d-3}{4}}} = \frac{1}{|\xi|^{\frac{d-1}{6}}}.$$

This gives the estimates in Theorem 1.3.

Consider next $d = 6$. By working as above we get

$$K(\tau,\xi) \le C\frac{1}{(\lambda+|\tau|-|\xi|)^{1/4}|\xi|^{3/4}} \qquad \text{and} \qquad L(\tau,\xi) \le C\frac{\lambda}{|\xi|}.$$

Use $\lambda = 1$ to get the estimates of Theorem 1.4. Alternatively, find $\lambda \in \left(0, |\xi|^{1/5}\right)$ such that

$$\frac{1}{(\lambda+|\tau|-|\xi|)^{1/4}|\xi|^{3/4}} = \frac{\lambda}{|\xi|},$$

and work as above to get the estimates in Theorem 1.3. In the case $d \ge 7$ we are already covered by (1.9c). ∎

Remark. We can obtain more estimates in $H^{s,\delta}$-spaces by interpolating between (1.9) and (1.11). For example, if $d \in \{3,4,5\}$, we have that for all $\alpha \in [0,1]$ estimates (1.10) hold true with $s = \alpha\frac{d-1}{6} + (1-\alpha)\frac{d-3}{4}$ and $\delta = (1-\alpha)\frac{7-d}{4}$, with similar results in all other dimensions.

4. Counterexamples. In this section we discuss the optimality of our results. We use the following notation: We fix $i, j \in \{1,\dots,d\}$, with $i \ne j$. Given a vector $\xi \in R^d$, we denote by $\xi' \in R^{d-2}$ the vector resulting from ξ if the ith and jth coordinates are removed, i.e., $\xi' = (\xi_i,\dots,\xi_{i-1},\xi_{i+1},\dots,\xi_{j-1},\xi_{j+1},\dots,\xi_n)$. We denote by μ the Lebesgue measure in R^d and by σ the corresponding surface measure on S^{d-1}.

PROPOSITION 4.1. *The condition $s+\delta \le 1/4$ in Theorem 1.2 is necessary.*

Proof. Fix $N >> 1$, and define

$$A = \left\{(\tau,\xi) : 5 \le \tau \le 10,\ -2N^{1/2} \le \xi_i \le -N^{1/2},\ N/2 \le \xi_j \le N, |\xi'| \le 1\right\}.$$

Let $\phi : S^{d-1} \to R$ be defined by $\phi(v) = \phi_i(v_i)\phi_j(v_j)$, where $\phi_i, \phi_j : R \to [0,1]$ are smooth cutoff functions such that

$$\begin{aligned}
\phi_i(v_i) = 1 \ \text{ for } \ -\frac{9}{10} \le v_i \le -\frac{3}{5},\\
\phi_i(v_i) = 0 \ \text{ for } \ v_i \ge -\frac{1}{2},\\
\phi_j(v_j) = 1 \ \text{ for } \ \frac{1}{10N^{1/2}} \le v_j \le \frac{3}{20N^{1/2}},\\
\phi_j(v_j) = 0 \ \text{ for } \ v_j \le 0 \quad \text{or} \quad v_j \ge \frac{1}{4N^{1/2}}.
\end{aligned}$$

In particular

$$\operatorname{supp}\phi_i \subseteq [-1,-1/2], \quad \operatorname{supp}\phi_j \subseteq [0,1/4N^{1/2}].$$

For $(\tau,\xi) \in A$ and $u \in \operatorname{supp}\phi$ we have

$$\begin{aligned}
\tau + v\cdot\xi = \tau + v_i\xi_i + v_j\xi_j + v'\cdot\xi' \le 10 + \frac{18}{10N^{1/2}} + \frac{3}{20N^{1/2}}N + 1 \le 11 + 2N^{1/2},\\
\tau + v\cdot\xi = \tau + v_i\xi_i + v_j\xi_j + v'\cdot\xi' \ge 5 + \frac{3N^{1/2}}{5} + 0 - 1 = 4 + \frac{3N^{1/2}}{5},
\end{aligned}$$

and therefore

$$\tau + v\cdot\xi \simeq N^{1/2}. \tag{4.1}$$

Also,

$$v_j\xi_i - v_i\xi_j \le 0 + \frac{9N}{10} \le N,$$
$$v_j\xi_i - v_i\xi_j \ge -\frac{6}{20} + \frac{3N}{10} = \frac{3(N-1)}{10},$$

and therefore

$$v_j\xi_i - v_i\xi_j \simeq N. \tag{4.2}$$

On the other hand,

$$\left|\frac{\Omega_v^{i,j}\phi(v)}{\tau + v\cdot\xi}\right| \le \frac{|v_i|\,|\phi_i|\left|\frac{\partial\phi_j}{\partial v_j}\right| + |v_j|\,|\phi_j|\left|\frac{\partial\phi_i}{\partial v_i}\right|}{|\tau + v\cdot\xi|} \le C\frac{N^{1/2} + N^{-1/2}}{N^{1/2}} \le C. \tag{4.3}$$

We define $f(t,x,v)$ and $g(t,x,v) \in L^2(R\times R^d\times S^{d-1})$ by

$$\widehat{f}(\tau,\xi,v) = \frac{\Omega_v^{i,j}\widehat{g}(\tau,\xi,v)}{i(\tau + v\cdot\xi)} \qquad \text{and} \qquad \widehat{g}(\tau,\xi,v) = \chi_A(\tau,\xi)\phi(v).$$

Then f and g are well defined (recall (4.1)) and $\partial_t f + v\cdot\nabla_x f = \Omega_v^{i,j} g$. The definition of f and integration by parts yield

$$\begin{aligned}\widehat{\rho_S}(\tau,\xi) &= \int_{S^{d-1}} \widehat{f}(\tau,\xi,v)d\sigma v \\ &= -i\chi_A(\tau,\xi)\int_{S^{d-1}} \frac{\Omega_v^{i,j}\phi(v)}{\tau + v\cdot\xi}d\sigma v \\ &= i\chi_A(\tau,\xi)\int_{S^{d-1}} \frac{v_j\xi_i - v_i\xi_j}{(\tau + v\cdot\xi)^2}\phi(v)d\sigma(v).\end{aligned}$$

Since $\phi \ge 0$ and $v_j\xi_i - v_i\xi_j \ge 0$ we have

$$|\widehat{\rho_S}(\tau,\xi)| = \chi_A(\tau,\xi)\int_{S^{d-1}} \frac{v_j\xi_i - v_i\xi_j}{(\tau + v\cdot\xi)^2}\phi(v)d\sigma(v)$$

and, by using (4.1) and (4.2),

$$|\widehat{\rho_S}(\tau,\xi)| \simeq \chi_A(\tau,\xi)\int_{S^{d-1}} \phi(v)d\sigma(v) \ge c\chi_A(\tau,\xi)\frac{1}{N^{1/2}}.$$

For $(\tau,\xi)\in A$ we have $1 + ||\tau| \pm |\xi|| \simeq N$, and therefore

$$\begin{aligned}\|\rho_S\|_{H^{s,\delta}} &= \left\|(1+|\tau|+|\xi|)^s(1+||\tau|-|\xi||)^\delta\widehat{\rho_S}(\tau,\xi)\right\|_{L^2} \\ &\ge cN^{s+\delta-1/2}\mu(A)^{1/2}.\end{aligned} \tag{4.4}$$

On the other hand, by using (4.3),

$$\begin{aligned}\|f\|_{L^2(R\times R^d\times S^{d-1})} &= \left(\int_{R^{1+d}}\int_{S^{d-1}}\left|\frac{\chi_A(\tau,\xi)\Omega_v^{i,j}\phi(v)}{i(\tau + v\cdot\xi)}\right|^2 d\sigma(v)d\tau d\xi\right)^{1/2} \\ &\le C\mu(A)^{1/2}\sigma(\operatorname{supp}\phi)^{1/2} \\ &\le C\mu(A)^{1/2}\frac{1}{N^{1/4}}.\end{aligned} \tag{4.5}$$

Finally,

$$\begin{aligned}\|g\|_{L^2(R\times R^d\times S^{d-1})} &= \left(\int_{R^{1+d}}\int_{S^{d-1}} \chi_A(\tau,\xi)^2\phi(v)^2 d\sigma(v)d\tau d\xi\right)^{1/2}\\ &\leq C\mu(A)^{1/2}\sigma(\operatorname{supp}\phi)^{1/2}\\ &\leq C\mu(A)^{1/2}\frac{1}{N^{1/4}}. \qquad (4.6)\end{aligned}$$

If for some (s,δ) we have $\|\rho_S\|_{H^{s,\delta}} \leq C\left(\|f\|_{L^2}+\|g\|_{L^2}\right)$, then (4.4)–(4.6) imply that $N^{s+\delta-1/2}\leq CN^{-1/4}$, and therefore $s+\delta\leq 1/4$ as required. We get exactly the same result if we have $H^{s,\delta}_{\log}$ instead of $H^{s,\delta}$. □

PROPOSITION 4.2. *The condition $s\leq 1$ in Theorem 1.2 is necessary.*

Proof. Fix $N >> 1$. Define

$$A=\left\{(\tau,\xi): N\leq\xi_i\leq 2N,\ -1\leq\xi_j\leq 0,\ |\xi'|\leq 1,\ |\xi|\leq\tau\leq|\xi|+1\right\}.$$

Let $\phi: R\to R$ be a smooth cutoff function, with $0\leq\phi\leq 1$ and $\phi=1$ on $[1/20,1/10]$ and $\phi=0$ outside $[1/21,1/9]$. Define

$$\widehat{g}(\tau,\xi,v)=\chi_A(\tau,\xi)\phi(v_i)\phi(v_j),\ \widehat{f}(\tau,\xi,v)=\frac{\Omega^{i,j}_v\widehat{g}(\tau,\xi,v)}{i(\tau+v\cdot\xi)}.$$

Notice that for $(\tau,\xi)\in A$ and $v\in S^{d-1}$ such that $v_i,v_j\in\operatorname{supp}\phi$ we have

$$\tau+v\cdot\xi\leq|\tau|+|\xi|\leq CN$$

and, since $v_i\xi_i\geq 0$ and $v_j\xi_j\geq -1$,

$$\tau+v\cdot\xi=\tau+v_i\xi_i+v_j\xi_j+v'\cdot\xi'\geq N+0-1-1\geq cN.$$

Therefore $\tau+v\cdot\xi>0$ and

$$\tau+v\cdot\xi\simeq N.$$

It follows that f is well defined and, moreover,

$$\begin{aligned}\left|\widehat{f}(\tau,\xi,v)\right| &\leq \frac{\chi_A(\tau,\xi)\left[\,|v_i|\phi(v_i)\,|\phi'(v_j)|+|v_j|\,|\phi'(v_i)|\,\phi(v_j)\,\right]}{|\tau+v\cdot\xi|}\\ &\leq C\frac{\chi_A(\tau,\xi)}{N};\end{aligned}$$

therefore

$$\|f\|_{L^2}\leq C\frac{\mu(A)^{1/2}}{N}.$$

Also,

$$\|g\|_{L^2}\leq C\mu(A)^{1/2},$$

and therefore

$$\|f\|_{L^2}+\|g\|_{L^2}\leq C\mu(A)^{1/2}. \qquad (4.7)$$

On the other hand, by using integration by parts we find

$$\widehat{\rho_{_S}}(\tau,\xi) = i\chi_A(\tau,\xi)\int_{S^{d-1}} \frac{v_j\xi_i - v_i\xi_j}{(\tau + v\cdot\xi)^2}\phi(v_i)\phi(v_j)d\sigma(v).$$

Since $v_i \geq 0$ and $\xi_j \leq 0$ we have $v_j\xi_i - v_i\xi_j \geq v_j\xi_i \geq \frac{N}{21} \geq cN$, and therefore

$$\begin{aligned}|\widehat{\rho_{_S}}(\tau,\xi)| &= \chi_A(\tau,\xi)\int_{S^{d-1}} \frac{v_j\xi_i - v_i\xi_j}{(\tau + v\cdot\xi)^2}\phi(v_i)\phi(v_j)d\sigma(v)\\ &\geq c\chi_A(\tau,\xi)\frac{N}{N^2}\int_{S^{d-1}}\phi(v_i)\phi(v_j)d\sigma(v)\\ &\geq c\frac{\chi_A(\tau,\xi)}{N}.\end{aligned}$$

Moreover, for $(\tau,\xi)\in A$, $1+|\tau|+|\xi| \simeq N$ and $1+||\tau|-|\xi|| \simeq 1$; therefore

$$\begin{aligned}&\left\|(1+|\tau|+|\xi|)^s(1+||\tau|-|\xi||)^\delta\,|\widehat{\rho_{_S}}(\tau,\xi)|\right\|_{L^2}\\ &\qquad\geq cN^s\frac{\mu(A)^{1/2}}{N} = N^{s-1}\mu(A)^{1/2}.\end{aligned} \tag{4.8}$$

If $\|\rho_{_S}\|_{H^{s,\delta}} \leq C\left(\|f\|_{L^2} + \|g\|_{L^2}\right)$ is satisfied, then (4.7) and (4.8) imply that $N^{s-1} \leq C$; therefore $s \leq 1$. We get exactly the same result if we have $H^{s,\delta}_{\log}$ instead of $H^{s,\delta}$. □

PROPOSITION 4.3. *The condition $s \leq \frac{d-1}{6}$ in Theorem 1.2 is necessary.*

Proof. For simplicity of notation we use $i = 1$ and $j = d$. Then $\xi' = (\xi_2,\dots,\xi_{d-1})$. Fix $N >> 1$, and define

$$A = \left\{(\tau,\xi) : 1 \leq \tau - |\xi| \leq 2,\ 0 \leq \xi_1 \leq 1,\ -2N \leq \xi_d \leq -N,\ |\xi'| \leq 1\right\}. \tag{4.9}$$

Notice that if $(\tau,\xi)\in A$, then

$$1 + ||\tau| - |\xi|| \simeq 1, \tag{4.10}$$

$$1 + |\tau| + |\xi| \simeq N, \tag{4.11}$$

$$\frac{1}{2} \leq \frac{-\xi_d}{|\xi|} \leq 1.$$

Let $\phi : R \to [0,1]$ be a smooth cutoff function, with $\operatorname{supp}\phi \subseteq [N^{-\frac{1}{3}}, 2N^{-\frac{1}{3}}]$ and $\phi \equiv 1$ on $[\frac{5}{4}N^{-\frac{1}{3}}, \frac{3}{2}N^{-\frac{1}{3}}]$. Let $\psi : R \to [0,1]$ be a smooth cutoff function, with $\operatorname{supp}\psi \subseteq [1-4N^{-\frac{2}{3}}, 1-N^{-\frac{2}{3}}]$ and $\psi \equiv 1$ on $[1-3N^{-\frac{2}{3}}, 1-2N^{-\frac{2}{3}}]$. Define

$$B = \left\{v \in S^{d-1} : v_1 \in \operatorname{supp}\phi,\ v_d \in \operatorname{supp}\psi\right\}. \tag{4.12}$$

For $(\tau,\xi)\in A$ and $v \in B$ we then have $v_d, \xi_1 \geq 0$ and $\xi_d \leq 0$; therefore

$$v_d\xi_1 - v_1\xi_d \geq 0 + v_1\,|\xi_d| \geq N^{\frac{2}{3}}. \tag{4.13}$$

Also

$$\tau + v\cdot\xi = \tau - |\xi| + |\xi|\left(1 + v_d\frac{\xi_d}{|\xi|}\right) + \sum_{i=1}^{d-1} v_i\xi_i. \tag{4.14}$$

We have $\tau - |\xi| \simeq 1$ and $|\sum_{i=1}^{d-1} v_i \xi_i| \leq C1$. Moreover,

$$1 + v_d \frac{\xi_d}{|\xi|} = \left(1 + \frac{\xi_d}{|\xi|}\right) + (1 - v_d)\frac{-\xi_d}{|\xi|} \geq 0 + N^{-\frac{2}{3}} \cdot \frac{1}{2} \geq cN^{-\frac{2}{3}},$$

and therefore

$$|\xi| \left(1 + v_d \frac{\xi_d}{|\xi|}\right) \geq cN^{\frac{1}{3}}.$$

As a consequence,

$$\tau + v \cdot \xi \geq cN^{\frac{1}{3}}.$$

Notice also that

$$|\xi| \left(1 + v_d \frac{\xi_d}{|\xi|}\right) = (|\xi| + \xi_d) + (1 - v_d) \cdot (-\xi_d),$$

with

$$(1 - v_d) \cdot (-\xi_d) \leq 4N^{-\frac{2}{3}} \cdot 2N \leq CN^{\frac{1}{3}}$$

and

$$|\xi| + \xi_d = \frac{|\xi|^2 - \xi_d^2}{|\xi| - \xi_d} = \frac{\sum_{k=1}^{d-1} \xi_k^2}{|\xi| + |\xi_d|} \leq C\frac{1}{N} \leq C.$$

Thus

$$|\xi| \left(1 + v_d \frac{\xi_d}{|\xi|}\right) \leq CN^{\frac{1}{3}}.$$

By using this in (4.14) we get

$$\tau + v \cdot \xi \leq CN^{\frac{1}{3}}.$$

We have shown that, for $(\tau, \xi) \in A$ and $v \in B$,

$$\tau + v \cdot \xi \simeq N^{\frac{1}{3}}. \tag{4.15}$$

Next define f and $g \in L^2(R_t \times R_x^d \times S_v^{d-1})$ by

$$\widehat{f}(\tau, \xi, v) = \frac{\Omega_v^{1,d} \widehat{g}(\tau, \xi, v)}{i(\tau + v \cdot \xi)} \qquad \text{and} \qquad \widehat{g}(\tau, \xi, v) = \chi_A(\tau, \xi)\phi(v_1)\psi(v_d).$$

We have

$$\|g\|_{L^2(R_t \times R_x^d \times S_v^{d-1})} \leq \mu(A)^{1/2}\sigma(B)^{1/2}. \tag{4.16}$$

Also

$$\begin{aligned}
\left|\widehat{f}(\tau, \xi, v)\right| &\leq \frac{\chi_A(\tau, \xi)}{|\tau + v \cdot \xi|} \left[v_1 \phi(v_1) \left|\psi'(v_d)\right| + v_d \left|\phi'(v_1)\right| \psi(v_d)\right] \\
&\leq C\frac{\chi_A(\tau, \xi)\chi_B(v)}{N^{\frac{1}{3}}} \left[N^{-\frac{1}{3}} \cdot 1 \cdot N^{\frac{2}{3}} + 1 \cdot N^{\frac{1}{3}} \cdot 1\right] \\
&\leq C\chi_A(\tau, \xi)\chi_B(v),
\end{aligned}$$

so that

$$\|f\|_{L^2(R_t\times R_x^d\times S_v^{d-1})}\le C\mu(A)^{1/2}\sigma(B)^{1/2}. \tag{4.17}$$

Now observe that B is contained in the set $\{v\in S^{d-1}: 1-4N^{-\frac{2}{3}}\le v_d\le 1\}$, which is a "cap" centered at the North pole of the sphere. It is easy to check that its area is $\le CN^{-\frac{d-1}{3}}$, and therefore $\sigma(B)\le CN^{-\frac{d-1}{3}}$. Then, from (4.16), (4.17), and previous observation, we get

$$\|f\|_{L^2(R_t\times R_x^d\times S_v^{d-1})}+\|g\|_{L^2(R_t\times R_x^d\times S_v^{d-1})}\le C\mu(A)^{1/2}N^{-(d-1)/6}. \tag{4.18}$$

For $\rho_{_S}$ we integrate by parts to get

$$\begin{aligned}\widehat{\rho_{_S}}(\tau,\xi)&=\int_{S^{d-1}}\frac{\Omega_v^{1,d}\widehat{g}(\tau,\xi,v)}{i(\tau+v\cdot\xi)}\,d\sigma(v)\\&=i\chi_A(\tau,\xi)\int_{S^{d-1}}\frac{v_d\xi_1-v_1\xi_d}{(\tau+v\cdot\xi)^2}\phi(v_1)\psi(v_d)\,d\sigma(v).\end{aligned}$$

Since $\phi,\psi\ge 0$ and $v_d\xi_1-v_1\xi_d\ge 0$ we have

$$|\widehat{\rho_{_S}}(\tau,\xi)|=\chi_A(\tau,\xi)\int_{S^{d-1}}\frac{v_d\xi_1-v_1\xi_d}{(\tau+v\cdot\xi)^2}\phi(v_1)\psi(v_d)\,d\sigma(v),$$

and therefore, by using (4.13) and (4.15),

$$|\widehat{\rho_{_S}}(\tau,\xi)|\ge c\chi_A(\tau,\xi)\int_{S^{d-1}}\phi(v_1)\psi(v_d)\,d\sigma(v), \tag{4.19}$$

we claim that

$$\int_{S^{d-1}}\phi(v_1)\psi(v_d)d\sigma(v)\ge cN^{-(d-1)/3}. \tag{4.20}$$

We have (recall that v' denotes the vector $(v_2,\dots,v_{d-1})\in R^{d-2}$)

$$\begin{aligned}&\int_{S^{d-1}}\phi(v_1)\psi(v_d)d\sigma(v)\\&=c\int_{v_1^2+|v'|^2\le 1}\phi(v_1)\psi\left(\sqrt{1-v_1^2-|v'|^2}\right)dv_1dv'\\&\ge c\int_X\phi(v_1)\psi\left(\sqrt{1-v_1^2-|v'|^2}\right)dv_1dv',\end{aligned} \tag{4.21}$$

where $X=\{\frac{5}{4}N^{-1/3}\le v_1\le\frac{3}{2}N^{-1/3},\ \frac{\sqrt{39}}{4}N^{-1/3}\le|v'|\le\frac{3}{2}N^{-1/3}\}$. Clearly, in the domain of integration in (4.21) we have $\phi(v_1)=1$. We claim that we also have $\psi(\sqrt{1-v_1^2-|v'|^2})=1$. Indeed,

$$1-\sqrt{1-v_1^2-|v'|^2}=\frac{v_1^2+|v'|^2}{1+\sqrt{1-v_1^2-|v'|^2}}\le\frac{\frac{9}{4}N^{-2/3}+\frac{9}{4}N^{-2/3}}{\frac{3}{2}}=3N^{-2/3},$$

and therefore

$$1-3N^{-2/3}\le\sqrt{1-v_1^2-|v'|^2}.$$

Moreover,

$$1-\sqrt{1-v_1^2-|v'|^2}=\frac{v_1^2+|v'|^2}{1+\sqrt{1-v_1^2-|v'|^2}}\geq\frac{\frac{25}{16}N^{-2/3}+\frac{39}{16}N^{-2/3}}{2}=2N^{-2/3},$$

and therefore

$$\sqrt{1-v_1^2-|v'|^2}\leq 1-2N^{-2/3}.$$

It follows that

$$\int_{S^{d-1}}\phi(v_1)\psi(v_d)d\sigma(v)\geq c\int_X\ 1\ dv_1dv'\geq cN^{-(d-1)/3}.$$

This proves (4.20). By using (4.20) in (4.19) we get

$$|\widehat{\rho_S}(\tau,\xi)|\geq c\chi_A(\tau,\xi)N^{-\frac{d-1}{3}},$$

and therefore, by using (4.10) and (4.11),

$$\left\|w_+(\tau,\xi)^s w_-(\tau,\xi)^\delta\widehat{\rho_S}(\tau,\xi)\right\|_{L^2(R\times R^d)}\geq\mu(A)^{1/2}N^{s-\frac{d-1}{3}}. \tag{4.22}$$

If the estimate

$$\begin{aligned}&\left\|w_+(\tau,\xi)^s w_-(\tau,\xi)^\delta\widehat{\rho_S}(\tau,\xi)\right\|_{L^2(R\times R^d)}\\&\quad\leq C\left(\|f\|_{L^2(R_t\times R^d_x\times S^{d-1}_v)}+\|g\|_{L^2(R_t\times R^d_x\times S^{d-1}_v)}\right)\end{aligned}$$

is valid, then it follows from (4.22) and (4.18) that

$$N^{s-(d-1)/3}\leq CN^{-(d-1)/6}$$

for $N\gg 1$ which gives

$$s\leq\frac{d-1}{6}.$$

We have exactly the same bound for s if $H^{s,\delta}$ is replaced by $H^{s,\delta}_{\log}$. □

Acknowledgments. The authors acknowledge with pleasure several helpful conversations with Benoît Perthame and Luis Vega.

REFERENCES

[1] C. Bardos, F. Golse, and B. Perthame, *The Rosseland approximation for the radiative transfer equations*, Comm. Pure Appl. Math., 40 (1987), pp. 691–721.

[2] C. Bardos, F. Golse, B. Perthame, and R. Sentis, *The nonaccretive radiative transfer equations: Existence of solutions and Rosseland approximation*, J. Funct. Anal., 77 (1988), pp. 434–460.

[3] F. Bouchut, *Introduction to the mathematical theory of kinetic equations*, in Kinetic Equations and Asymptotic Theories, Ser. Appl. Math. 4, Elsevier, New York, 2000.

[4] F. Bouchut, *Hypoelliptic regularity in kinetic equations*, J. Math. Pures Appl., 81 (2002), pp. 1135–1159.

[5] F. Bouchut and L. Desvillettes, *Averaging lemmas without time Fourier transform and applications to discretized kinetic equations*, Proc. Roy. Soc. Edinburgh Sect. A 129, 1 (1999), pp. 19–36.
[6] J. Bourgain, *Fourier transform restriction phenomena for certain lattice subsets and applications to nonlinear evolution equations. Part* I, Geom. Funct. Anal., 3 (1993), pp. 107–156; *Part* II, Geom. Funct. Anal., 3 (1993), pp. 209–262.
[7] N. Bournaveas and B. Perthame, *Averages over spheres for kinetic transport equations; hyperbolic Sobolev spaces and Strichartz inequalities*, J. Math. Pures Appl., 9 (2001), pp. 517–534.
[8] N. Bournaveas and S. Gutierrez, *On the regularity of averages over spheres for kinetic transport equations in hyperbolic Sobolev spaces*, Rev. Mat. Iberoamericana, 23 (2007), pp. 481–512.
[9] F. Castella and B. Perthame, *Estimations de Strichartz pour les équations de transport cinétique*, C. R. Acad. Sci. Paris Sér. I, 322 (1996), pp. 535–540.
[10] S. Chandrasekhar, *Radiative Transfer*, Clarendon Press, Oxford, 1950.
[11] L. Desvillettes and S. Mischler, *About the splitting algorithm for Boltzmann and B.G.K. equations*, Math. Models Methods Appl. Sci., 6 (1996), pp. 1079–1101.
[12] R. DeVore and G. Petrova, *The averaging lemma*, J. Amer. Math. Soc., 14 (2001), pp. 279–296.
[13] R. J. DiPerna and P. L. Lions, *On the Cauchy problem for the Boltzmann equation: Global existence and weak stability results*, Ann. of Math., 130 (1989), pp. 321–366.
[14] R. J. DiPerna and P. L. Lions, *Global weak solutions of Vlasov - Maxwell systems*, Comm. Pure Appl. Math., 42 (1989), pp. 729–757.
[15] R. J. DiPerna, P. L. Lions, and Y. Meyer, *L^p regularity of velocity averages*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 8 (1991), pp. 271–287.
[16] Y. Dolak and C. Schmeiser, *Kinetic models for chemotaxis: Hydrodynamic limits and spatio-temporal mechanisms*, J. Math. Biol., 51 (2005), pp. 595–615.
[17] M. Escobedo, P. Laurencot, and S. Mischler, *On a kinetic equation for coalescing particles*, Comm. Math. Phys., 246 (2004), pp. 237–267.
[18] D. Foschi and S. Klainerman, *Bilinear space-time estimates for homogeneous wave equations*, Ann. Sci. École Norm. Sup., 33 (2000), pp. 211–274.
[19] P. Gérard, *Regularity of means of solutions of partial differential equations*, Journées Équations aux Dérivées Partielles, 1987, pp. 1–8.
[20] P. Gérard, *Microlocal defect measures*, Comm. Partial Differential Equations, 16 (1991), pp. 1761–1794.
[21] P. Gérard and F. Golse, *Averaging regularity results for PDEs under transversality assumptions*, Comm. Pure Appl. Math., 45 (1992), pp. 1–26.
[22] F. Golse and B. Perthame, *Generalized solutions of the radiative transfer equations in a singular case*, Comm. Math. Phys., 106 (1986), pp. 211–239.
[23] F. Golse, P. L. Lions, B. Perthame, and R. Sentis, *Regularity of the moments of the solution of a transport equation*, J. Funct. Anal., 76 (1988), pp. 110–125.
[24] F. Golse, B. Perthame, and R. Sentis, *Un résultat de compacité pour les équations du transport et application au calcul de la limite de la valeur propre principale d'un opérateur de transport*, C. R. Acad. Sci. Sér. I, 301 (1985), pp. 341–344.
[25] F. Golse and L. Saint-Raymond, *The Navier-Stokes limit of the Boltzmann equation for bounded collision kernels*, Invent. Math., 155 (2004), pp. 81–161.
[26] F. Golse and L. Saint-Raymond, *Velocity averaging in L Sp*1 *for the transport equation*, C. R. Math. Acad. Sci. Paris, 334 (2002), pp. 557–562.
[27] P. E. Jabin and L. Vega, *A real space method for averaging lemmas*, J. Math. Pures Appl., 83 (2004), pp. 1309–1351.
[28] P. E. Jabin and L. Vega, *Averaging lemmas and the x-ray transform*, C. R. Math. Acad. Sci. Paris, 337 (2003), pp. 505–510.
[29] P. E. Jabin and B. Perthame, *Regularity in kinetic formulations via averaging lemmas*, ESAIM Control Optim. Calc. Var., 8 (2002), pp. 761–774.
[30] S. Klainerman and M. Machedon, *Space-time estimates for null forms and the local existence theorem*, Comm. Pure Appl. Math., 46 (1993), pp. 1221–1268.
[31] S. Klainerman and M. Machedon, *Smoothing estimates for null forms and applications*, Duke Math. J., 81 (1995), pp. 99–133.
[32] S. Klainerman and M. Machedon, *Estimates for null forms and the spaces H Sbs, δ*, Int. Math. Res. Not., 17 (1996), pp. 853–865.
[33] P. L. Lions and B. Perthame, *Lemmes de moments, de moyenne et de dispersion*, C. R. Acad. Sci. Paris Sér. I 314, 11 (1992), pp. 801–806.

[34] P. L. Lions, *Régularité optimale des moyennes en vitesses*, C. R. Acad. Sci. Paris Sér. I, 320 (1995), pp. 911–915.
[35] B. Perthame, *Mathematical tools for kinetic equations*, Bull. Amer. Math. Soc. (N.S.), 41 (2004), pp. 205–244.
[36] B. Perthame and J. L. Vázquez, *Bounded speed of propagation for solutions to radiative transfer equations*, Comm. Math. Phys., 130 (1990), pp. 457–469.
[37] B. Perthame and P. E. Souganidis, *A limiting case for velocity averaging*, Ann. Sci. École Norm. Sup. (4), 31 (1998), pp. 591–598.
[38] D. Tataru, *The X Sps Sbθ spaces and unique continuation for solutions to the semilinear wave equation*, Comm. Partial Differential Equations, 21 (1996), pp. 841–887.

SIAM J. MATH. ANAL.
Vol. 40, No. 2, pp. 675–698

STABILITY OF EQUILIBRIA FOR THE STEFAN PROBLEM WITH SURFACE TENSION*

JAN PRÜSS† AND GIERI SIMONETT‡

Abstract. We characterize the equilibrium states for the two-phase Stefan problem with surface tension and with or without kinetic undercooling, and we analyze their stability in terms of dependence on physical and geometric quantities.

Key words. free boundary problem, phase transitions, surface tension, kinetic undercooling, stability, bifurcation

AMS subject classifications. 35R55, 35B55, 35K55, 80A22

DOI. 10.1137/070700632

1. Introduction.

The Stefan problem is a model for phase transitions in solid-liquid systems. In this paper, we consider the two-phase Stefan problem with the modified Gibbs–Thomson law

$$u = \sigma H + \delta V \quad \text{on} \quad \Gamma(t), \quad \sigma > 0, \quad \delta \geq 0, \tag{1.1}$$

and the kinetic condition

$$[d\partial_\nu u] = (\ell - [\kappa]u)V \quad \text{on} \quad \Gamma(t). \tag{1.2}$$

Here $\Gamma(t)$ denotes the unknown moving hypersurface that separates the liquid from the solid phase, u is the temperature, H the mean curvature of $\Gamma(t)$, σ the surface tension coefficient, δ the coefficient of kinetic undercooling, V the normal velocity of $\Gamma(t)$, ℓ the latent heat, $[\kappa]$ the jump of the heat capacities across $\Gamma(t)$, and $[d\partial_\nu u]$ the jump of the heat fluxes across $\Gamma(t)$. Note that in case $\sigma = \delta = 0$, i.e., for the classical Stefan problem, we have $u = 0$ at the interface, and then the kinetic condition becomes the classical Stefan condition.

Under appropriate boundary conditions we will show that spheres (together with constant temperature distributions) are the only equilibrium states for this system, and we will characterize the stability of these equilibria in terms of dependence on physical and geometric quantities.

In order to formulate the Stefan problem we introduce the following notation. Let Ω be a smooth bounded domain in $\mathbb{R}^n$ whose boundary $\partial\Omega$ consists of two disjoint components, an "interior" part J_1 and an "exterior" part J_2. We think of Ω as a homogeneous medium which is occupied by a liquid and a solid phase, say water and ice, that initially occupy the regions Ω_0^1 and Ω_0^2, and that are separated by a sharp interface Γ_0. More precisely, we assume that $\Gamma_0 \subset \Omega$ is a compact closed hypersurface, and that Ω_0^1 and Ω_0^2 are disjoint open sets such that $\overline{\Omega} = \overline{\Omega_0^1} \cup \overline{\Omega_0^2}$, and such that $\partial\Omega_0^i = J_i \cup \Gamma_0$ for $i = 1, 2$. For the sake of definiteness we consider the open set Ω_0^1 as

*Received by the editors August 20, 2007; accepted for publication (in revised form) March 11, 2008; published electronically July 3, 2008.
http://www.siam.org/journals/sima/40-2/70063.html
†Institut für Mathematik, Martin-Luther-Universität Halle-Wittenberg, D-60120 Halle, Germany (jan.pruess@mathematik.uni-halle.de).
‡Department of Mathematics, Vanderbilt University, Nashville, TN 37240 (simonett@math.vanderbilt.edu). The research of this author was partially supported by NSF grant DMS-0600870.

the region occupied by the liquid phase. Consequently, the component J_1 is in contact with the liquid phase, and J_2 is in contact with the solid phase. The boundaries J_1 and J_2, corresponding for instance to the walls of a container, are fixed, whereas Γ_0 will change as time evolves, due to solidification or liquidation of the two different phases.

Given $t \geq 0$, let $\Gamma(t)$ be the position of Γ_0 at time t, and let $V(\cdot,t)$ and $H(\cdot,t)$ be the normal velocity and the mean curvature of $\Gamma(t)$. Moreover, let $\Omega_1(t)$ and $\Omega_2(t)$ be the two regions in Ω separated by $\Gamma(t)$. According to our assumption, $\Omega_1(t)$ is the region occupied by the liquid phase, and $\Gamma(t)$ is a sharp interface which separates the liquid from the solid phase. Let $\nu(\cdot,t)$ be the outer unit normal field on $\Gamma(t)$ with respect to $\Omega_1(t)$. We shall use the convention that the normal velocity is positive if $\Omega_1(t)$ is expanding, and that the mean curvature is positive if the intersection of $\Omega_1(t)$ with a small ball centered at $\Gamma(t)$ is convex. Consequently, the normal velocity is positive if the liquid region is growing, ν points into the solid phase, and H is positive for a water ball surrounded by ice, and negative for an ice ball surrounded by water.

Here we concentrate on the case $J_1 = \emptyset$. Let Γ_0 and $u_0^i : \Omega_0^i \to \mathbb{R}$ be given, where u_0^1 and u_0^2 denote the initial temperatures of the liquid and solid phase, respectively. The strong formulation of the *two-phase Stefan problem with surface tension and kinetic undercooling* consists of finding a family $\Gamma := \{\Gamma(t)\,;\, t \geq 0\}$ of hypersurfaces and functions $u_i : \cup_{t\geq 0}\,(\Omega_i(t) \times \{t\}) \to \mathbb{R}$, satisfying

$$
\left\{
\begin{aligned}
\kappa_i \partial_t u_i - d_i \Delta u_i &= 0 && \text{in} \quad \Omega_i(t),\\
\partial_\nu u_2 &= 0 && \text{on} \quad J_2,\\
u_i &= \sigma H_\Gamma + \delta V && \text{on} \quad \Gamma(t),\\
[d\partial_\nu u] &= (\ell - [\kappa]u)V && \text{on} \quad \Gamma(t),\\
u_i(0) &= u_0^i && \text{in} \quad \Omega_0^i,\\
\Gamma(0) &= \Gamma_0\,, &&
\end{aligned}
\right. \tag{1.3}
$$

where $\kappa_i \geq 0$ is the heat capacity of phase i, d_i is its thermal conductivity coefficient, $\ell > 0$ is the latent heat per unit mass absorbed or released for melting or solidifying, $\sigma > 0$ is the surface tension, and $\delta \geq 0$ is the speed of kinetic undercooling. Moreover,

$$
\begin{aligned}
[\kappa] &:= \kappa_2 - \kappa_1,\\
[d\partial_\nu u] &:= d_2 \partial_\nu u_2 - d_1 \partial_\nu u_1
\end{aligned}
$$

denote the jump of the heat capacities and the heat fluxes, respectively, across the interface $\Gamma(t)$. Note that $[\kappa] = \kappa_2 - \kappa_1 < 0$ is physically reasonable since in the liquid phase there are more degrees of freedom than in the solid phase; hence, the liquid phase can absorb more energy per unit mass. However, we do not assume $[\kappa] < 0$ in what follows. The condition $u_i = \sigma H_\Gamma$ on the free interface is usually called the Gibbs–Thomson law, and $u_i = \sigma H_\Gamma + \delta V$ the modified Gibbs–Thomson law, or the Gibbs–Thomson law with kinetic undercooling; see [2, 3, 16, 17, 19, 21, 24, 25, 32] for more information.

We refer to [12, 13, 14, 22, 23, 28, 29] for existence and regularity results for the Stefan problem with the Gibbs–Thomson law $u_i = \sigma H_\Gamma$ in case $\kappa_1 = \kappa_2$. The Stefan problem with surface tension and kinetic undercooling in case $\kappa_1 = \kappa_2$ has been studied in [5, 28, 29, 31]; see also [20] for the one-phase case.

It will be shown in [26] that the Stefan problem (1.3) has a unique local solution which is analytic in space and time, provided that the well-posedness condition

$$
\ell - \sigma[\kappa] H_{\Gamma_0} > 0 \quad \text{in case} \quad \delta = 0 \tag{1.4}
$$

is satisfied. On the other hand, if $\delta = 0$ and $\kappa_1 > \kappa_2$, problem (1.3) is not well-posed if H_{Γ_0} is too negative, that is, in case the solid region sharply protrudes into the liquid. Associated to the Stefan problem (1.3) is the energy functional

$$E(u(t),\Gamma(t)) := \int_\Omega \kappa u\,dx + \ell\,|\Omega_1(t)| = \int_{\Omega_1(t)} \kappa_1 u_1\,dx + \int_{\Omega_2(t)} \kappa_1 u_2\,dx + \ell\,|\Omega_1(t)|, \tag{1.5}$$

where $|\Omega_1(t)|$ is the volume of the region $\Omega_1(t)$. If (u,Γ) is a sufficiently smooth solution of (1.3), then we obtain

$$\begin{aligned}\frac{d}{dt}E(u(t),\Gamma(t)) &= \int_{\Omega_1(t)} \kappa_1\partial_t u_1\,dx + \int_{\Omega_2(t)} \kappa_1\partial_t u_2\,dx - [\kappa]\int_{\Gamma(t)} uV\,ds + \ell\int_{\Gamma(t)} V\,ds\\ &= \int_{\Gamma(t)} \Big(- [d\partial_\nu u] - [\kappa]uV + \ell V\Big)\,ds = 0,\end{aligned} \tag{1.6}$$

thus showing that energy is conserved.

If $\kappa_1 = \kappa_2 = 0$ and $\delta = 0$, then the resulting problem is the quasi-stationary Stefan problem with surface tension, which has also been termed the Mullins–Sekerka model (or the Hele–Shaw model with surface tension). Existence, uniqueness, regularity, and global existence of solutions for the quasi-stationary approximation have been investigated in [1, 4, 6, 8, 9, 10, 11, 15]. Existence and global existence of classical solutions for the quasi-stationary approximation with $\sigma > 0$ and $\delta > 0$ have been studied in [33, 20].

A major difficulty in the mathematical treatment of the Stefan problem (1.3) is due to fact that the boundary $\Gamma(t)$, and thus the geometry, is unknown and ever changing. A widely used method to overcome this inherent difficulty is to choose a fixed reference surface Σ and then represent the moving surface $\Gamma(t)$ as the graph of a function (which we will denote by $\rho = \rho(s,t)$) in normal direction of Σ. In this way, one obtains a time-dependent (unknown) diffeomorphism from Σ onto $\Gamma(t)$, and in the next step this diffeomorphism is extended to a diffeomorphism of fixed reference regions D^i onto the unknown domains $\Omega_i(t)$. The treatment of the moving boundary problem (1.3) then proceeds by transforming the equations into a new system of equations defined on the fixed domain $D_1 \cup D_2$ from which both the solution and the parameterizing function ρ have to be determined. In the context of the Stefan problem this approach has first been used by Hanzawa [18].

The same approach has also been used in [10, 11] for the quasi-stationary approximation of the Stefan problem with surface tension, and in [26] for the Stefan problem with surface tension. Once the transformed system has been obtained, one can study the mapping properties of the nonlinearities involved, and in particular, one can determine their linearizations; see [26] for more details.

In this paper, we assume that $\Gamma(t)$ does not touch the fixed boundary $J_2 = \partial\Omega$. Under this assumption, we will characterize all of the equilibrium states (u_1, u_2, Σ) of (1.3). In fact, it is easy to see that the equilibria are precisely given by

$$\Sigma = \bigcup_{j=1}^m S_R(x_j), \quad u_1 = u_2 = \sigma/R,$$

where $S_R(x_j)$ denotes disjoint spheres of the same radius R and centers x_j. This can be seen by the following arguments: the equilibria (u_1, u_2, Σ) of the Stefan problem (1.3)

are given by the system of equations

$$\begin{cases} -d_i\Delta u_i = 0 & \text{in} \quad \Omega_i, \\ \partial_\nu u_2 = 0 & \text{on} \quad \partial\Omega, \\ u_i = \sigma H_\Sigma & \text{on} \quad \Sigma, \\ [d\partial_\nu u] = 0 & \text{on} \quad \Sigma. \end{cases} \tag{1.7}$$

Taking the inner product of $(1.7)_1$ with u_i, the divergence theorem and condition $(1.7)_4$ yield

$$\int_{\Omega_i} |\nabla u_i|^2 \, dx = 0;$$

hence u_i is constant on Ω_i. Equation $(1.7)_3$, in turn, shows that $u_1 = u_2$ and also that $H = u/\sigma$ is constant on Σ. But then, since Ω is bounded, Σ must be a sphere $S_R(x_0)$ centered at some point $x_0 \in \Omega$ with radius $R > 0$, if the phases are connected. Otherwise, again due to the boundedness of Ω, Σ is the union of finitely many spheres of the same radius $R > 0$. Here we concentrate on the case of connected phases. Thus there is an $(n+1)$-parameter family of equilibria, the parameters being the n coordinates of the center x_0 and the radius R.

We want to discuss the stability of those equilibria. The linearized problem (associated to the transformed equations) at such an equilibrium state is given by

$$\begin{cases} \kappa\partial_t v - d\Delta v = f & \text{in} \quad (\Omega \setminus \Sigma) \times \mathbb{R}_+, \\ \partial_\nu v = 0 & \text{on} \quad \partial\Omega \times \mathbb{R}_+, \\ v = \sigma\mathcal{A}_\Sigma\rho + \delta\partial_t\rho + g & \text{on} \quad \Sigma \times \mathbb{R}_+, \\ l\partial_t\rho - [d\partial_\nu v] = h & \text{on} \quad \Sigma \times \mathbb{R}_+, \\ v(0) = v_0 & \text{in} \quad \Omega \setminus \Sigma, \\ \rho(0) = \rho_0 & \text{on} \quad \Sigma; \end{cases} \tag{1.8}$$

see [26]. Here, $l = \ell - [\kappa]\sigma/R$, and the operator $\mathcal{A}_\Sigma$ is given by

$$\mathcal{A}_\Sigma = -\frac{1}{n-1}\left(\frac{n-1}{R^2} + \Delta_\Sigma\right),$$

where Δ_Σ denotes the Laplace–Beltrami operator on Σ. This is the linearization of the mean curvature $H'(0)$ at the sphere Σ; cf., e.g., Escher and Simonett [11]. Here we use the notation $v = v_1\chi_{\Omega_1} + v_2\chi_{\Omega_2}$, where χ_G denotes the characteristic function of a set G, and similarly $\kappa = \kappa_1\chi_{\Omega_1} + \kappa_2\chi_{\Omega_2}$ and $d = d_1\chi_{\Omega_1} + d_2\chi_{\Omega_2}$. Associated to the linearization (1.8) is the following eigenvalue problem:

$$\begin{cases} \lambda\kappa v - d\Delta v = 0 & \text{in} \quad \Omega \setminus \Sigma, \\ \partial_\nu v = 0 & \text{on} \quad \partial\Omega, \\ v = \sigma\mathcal{A}_\Sigma\rho + \lambda\delta\rho & \text{on} \quad \Sigma, \\ \lambda l\rho - [d\partial_\nu v] = 0 & \text{on} \quad \Sigma, \end{cases} \tag{1.9}$$

where as before $l = \ell - [\kappa]\sigma/R$. We will now state the main results of this paper. We will formulate our results for a domain in $\mathbb{R}^n$ for $n \in \mathbb{N}$, $n > 1$, although the physically relevant dimensions, naturally, are $n = 2, 3$.

THEOREM 1.1. *Suppose that the phases in the Stefan problem are connected. Then the following assertions hold:*

(a) *The equilibrium states (without boundary contact) for problem* (1.3) *are given by*

$$(u, \Sigma), \quad \textit{where} \quad \Sigma = S_R(x_0) \quad \textit{and} \quad u = \sigma/R,$$

with $S_R(x_0) \subset \Omega$ being the sphere with radius R and center x_0.

(b) *For $l > 0$, the eigenvalue problem* (1.9) *has countably many real eigenvalues of finite algebraic multiplicity.*

(c) *0 is an eigenvalue of* (1.9) *with geometric multiplicity $(n+1)$. The (geometric) eigenspace is spanned by*

$$(-1, Y_0), (0, Y_1), \ldots, (0, Y_n),$$

where $Y_0 = R^2/\sigma$, and where Y_j, $1 \le j \le n$, are the spherical harmonics of degree 1 (normalized by the orthogonality condition $(Y_i|Y_j)_\Sigma = \delta_{ij}$).

(d) *If $\sigma(\kappa|1)_\Omega \le l|\Sigma|R^2$, then* (1.9) *has no positive eigenvalues.*

(e) *If $\sigma(\kappa|1)_\Omega > l|\Sigma|R^2 > 0$, then* (1.9) *has exactly one positive, algebraically simple eigenvalue.*

(f) *If $l < 0$ and $\delta > 0$, then* (1.9) *has at least one positive eigenvalue.*

(g) *If $l < 0$ and $\delta = 0$, then the linearized problem* (1.8) *is not well-posed.*

Proof. The assertion in (a) has been proved above. We refer to Theorem 2.1 for a proof of assertions (b)–(e), and for additional information about the eigenvalue problem (1.9), for the case $l > 0$. The proof of (f) is given at the end of section 5, and (g) follows from [7]. □

Remark 1.2. (a) If $l < 0$, then all equilibrium states are linearly unstable (and the linearized problem (1.8) is not even well-posed in case $\delta = 0$). Therefore we mainly concentrate on the case $l > 0$. Define then

$$\zeta := \frac{\sigma(\kappa|1)_\Omega}{l|\Sigma|R^2}, \quad \text{where} \quad l = \ell - \frac{\sigma[\kappa]}{R}, \quad (\kappa|1)_\Omega := \int_\Omega \kappa \, dx.$$

According to Theorem 1.1.(d)–(e), we know that all eigenvalues of (1.9) are non-positive if $\zeta \le 1$, and that there exists exactly one positive simple eigenvalue if $\zeta > 1$. We will refer to the case $\zeta \le 1$ as a *stability condition.*

Observe that neither the thermal conductivity coefficients d_i nor the kinetic coefficient δ enters this stability condition, as it depends only on the heat capacities κ_i, the latent heat ℓ, the surface tension σ, and on the geometry. In particular, decreasing the size of a ball decreases its stability, as does increasing surface tension; see also Remark 1.5(a). We also mention that the stability condition $\zeta \le 1$ is always valid in the quasi-stationary case $\kappa_i = 0$, i.e., for the Mullins–Sekerka problem.

(b) It will be shown in the forthcoming paper [27] that solutions for the Stefan problem (1.3) that start out close to an equilibrium (u, Σ) exist globally and converge towards an equilibrium state (u', Σ') as time goes to infinity, provided that $l > 0$ and $\zeta < 1$. This gives justice to the wording *stability condition* for the case $\zeta < 1$. We note again that $\zeta = 0$ if the heat capacities κ_i are zero, that is, in the quasi-stationary case. In this case, global existence and convergence to equilibria were obtained in [11, 20] by using a center-manifold analysis; see also [15] for a different approach in the one-phase case.

(c) If the Gibbs–Thomson condition on the free interface $\Gamma(t)$ is replaced by $u_i = 0$, then (1.3) is called the *(classical) Stefan model.* It should be observed that,

in contrast to the problem with surface tension, the classical Stefan problem does not admit nontrivial equilibrium states.

For $l > 0$, the results in Theorem 1.1 suggest that one eigenvalue, λ_*, crosses the imaginary axis at 0 as the quantity ζ increases and exceeds 1. According to part (c) of Theorem 1.1, 0 is always an eigenvalue with geometric multiplicity $(n+1)$. This suggests that as the eigenvalue λ_* crosses through 0, the algebraic multiplicity of 0 raises by one, and then drops again as soon as the eigenvalue has crossed. This is exactly what happens, as will be proved in Theorem 2.1.

Another way to view and understand this situation can be gained from considering the following parameter-dependent eigenvalue problem:

$$\begin{cases} \lambda_* s\kappa v - d\Delta v = 0 & \text{in} \quad \Omega \setminus \Sigma, \\ \partial_\nu v = 0 & \text{on} \quad \partial\Omega, \\ v = \sigma \mathcal{A}_\Sigma \rho + \lambda_* \delta \rho & \text{on} \quad \Sigma, \\ \lambda_* l\rho - [d\partial_\nu v] = 0 & \text{on} \quad \Sigma. \end{cases} \tag{1.10}$$

The following result will be proved in section 6.

THEOREM 1.3. *Let $l > 0$ and set $s_0 := l|\Sigma|R^2/\sigma(\kappa|1)_\Omega$. Then the following hold:*

(a) *The eigenvalue problem* (1.10) *has an analytic curve of solutions*

$$[s \mapsto (\lambda_*(s), v(s), \rho(s))], \quad s \in (s_0 - \varepsilon_0, \infty),$$

such that $\lambda_*(s) > 0$ iff $s > s_0$, where ε_0 is an appropriate positive number.

(b) *$\lambda_*(s)$ crosses the imaginary axis with positive speed at $s = s_0$.*

(c) *$[s \mapsto \lambda_*(s)]$ is strictly increasing.*

(d) *If $\delta > 0$, then $\lambda_*(s)$ is bounded above by $\sigma/\delta R^2$.*

(e) *If $\delta = 0$, then $\lambda_*(s) \to \infty$ as $s \to \infty$.*

Clearly, the eigenvalues of the modified problem (1.10) coincide with the eigenvalues of (1.9) if $s = 1$. In case $\zeta > 1$ we have $s_0 < 1$ and see that $\lambda = \lambda_*(1)$ is a (the only) positive eigenvalue of (1.9).

According to (1.5) an equilibrium state $(\sigma/R, S_R(x_0))$ for the Stefan problem (1.3) has energy

$$\begin{aligned} \phi(R) :&= E\left(\frac{\sigma}{R}, S_R(x_0)\right) = \frac{\sigma}{R}(\kappa|1)_\Omega + \ell\,|\Omega_1| \\ &= \frac{\sigma}{R}\big(\kappa_1|\Omega_1| + \kappa_2|\Omega_2|\big) + \ell\,|\Omega_1|, \end{aligned} \tag{1.11}$$

where $|\Omega_1| = R^n|B|$ and $|\Omega_2| = |\Omega| - R^n|B|$, with $|B|$ the volume of the unit ball. A straightforward computation shows that the function ϕ has a unique minimum. It is attained at the point R_*, where R_* is the unique solution of

$$\frac{\sigma(\kappa|1)_\Omega}{R^2} = \left(\ell - \frac{\sigma[\kappa]}{R}\right)|S_R|, \tag{1.12}$$

with $|S_R| = |S_R(x_0)|$ being the area of the sphere $S_R(x_0)$.

In the following, we denote by R_* the point where ϕ attains its (unique) minimum and by R^* the largest number R such that $\overline{B}_R(x_0) \subset \overline{\Omega}$, and we suppose that $R_* < R^*$. Then we have the following result; see also the stability diagram in Figure 1.

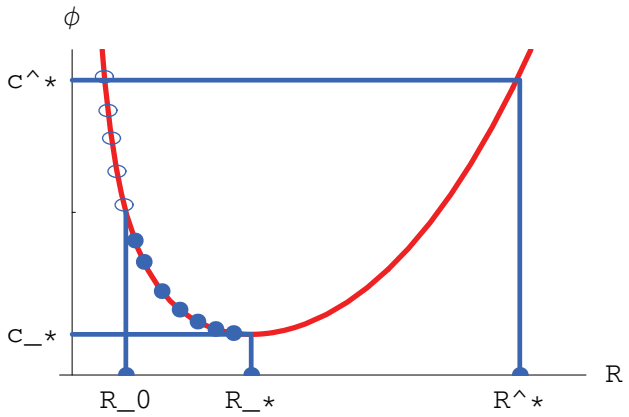


FIG. 1. *Stability diagram for $\kappa_1 < \kappa_2$ and $\delta = 0$; circled: ill-posed, dotted: unstable.*

COROLLARY 1.4. *Let $c_* = \phi(R_*)$ be the minimum value of ϕ, and let $c^* = \phi(R^*)$. Moreover, let $c_0 = E(u_0, \Gamma_0) = \phi(R)$ be a given energy level.*

(a) *If $c_0 < c_*$, then problem* (1.3) *does not admit equilibrium states.*

(b) *If $c_* < c_0 < c^*$, then* (1.3) *admits two branches of equilibrium states. The branch of equilibria $(\sigma/R, S_R(x_0))$ with $0 < R < R_*$ is linearly unstable, whereas the branch with $R_* < R < R^*$ is linearly stable.*

(d) *In case $c_0 = c_*$ or $c_0 > c^*$, the Stefan problem* (1.3) *admits one family of equilibrium states. All equilibria $(\sigma/R, S_R(x_0))$ with $\phi(R) > c^*$ are linearly unstable.*

(e) *If $R_0 := [\kappa]\sigma/\ell > 0$ and $\delta = 0$, then the linearized problem is ill-posed for $R < R_0$.*

Proof. This follows from Remark 1.2(a), (1.12), and the fact that $\phi'(R)$ is negative for $R < R_*$ and positive for $R_* < R < R^*$. □

Remark 1.5. (a) We can show that R_* is increasing with respect to σ (and, for that matter, also with respect to $[\kappa]$). In order to see this, let $R_*(\sigma)$ denote the unique solution of (1.12). Then we have $R_*'(\sigma) > 0$. For this we note that (1.12) can be written as

$$n\ell|B|R_*^{n+1}(\sigma) - \sigma\big(\kappa_2|\Omega| + (n-1)[\kappa]|B|R_*^n(\sigma)\big) = 0. \tag{1.13}$$

Taking the derivative of this equation with respect to σ yields

$$n(n+1)\ell|B|R_*^{n-1}\left(R_* - \frac{(n-1)}{(n+1)}\frac{\sigma[\kappa]}{\ell}\right)R_*'(\sigma)$$
$$= \kappa_2|\Omega| + (n-1)[\kappa]|B|R_*^n = (\ell/\sigma)n|B|R_*^{n+1}(\sigma) > 0.$$

Note that we have used (1.13) for the last equality. It remains to be observed that the parenthetical expression in front of $R_*'(\sigma)$ is always positive. This is clear in case $[\kappa] \le 0$ and follows from the fact that R_* is always greater than $R_0 = \sigma[\kappa]/l$ in case $[\kappa] > 0$. We therefore see that increasing σ increases $R_*(\sigma)$, showing that spheres with a fixed radius R can lose stability as σ increases.

(b) If one considers the case where the domain Ω_1 is occupied by the solid phase, and Ω_2 by the fluid phase, then the third and fourth lines in (1.3) must be replaced

by

$$\begin{aligned} u_i &= -\sigma H_\Gamma - \delta V \qquad \text{on} \quad \Gamma(t), \\ -[d\partial_\nu u] &= (\ell + [\kappa]u)V \qquad \text{on} \quad \Gamma(t), \end{aligned} \tag{1.14}$$

and the energy functional by

$$E(u(t),\Gamma(t)) := \int_\Omega \kappa u\,dx + \ell\,|\Omega_2(t)| = \int_{\Omega_1(t)} \kappa_1 u_1\,dx + \int_{\Omega_2(t)} \kappa_1 u_2\,dx + \ell\,|\Omega_2(t)|,$$

while all other conventions are left unchanged. Thus, one formally has to switch signs in the normal ν and in ℓ and $[\kappa]$. Then all of the results and assertions stated in this paper remain valid for the equilibrium states $(-\sigma/R, S_R(x_0))$.

The plan of this paper is the following. In section 2 we will state a more general and concise version of Theorem 1.1; its proof will be given in sections 3–5. Finally, in section 6 we will prove Theorem 1.3.

2. Main theorem. In this section we will introduce an appropriate functional analytic setting to study the eigenvalue problem (1.9). We always assume $l > 0$ except when proving (f) of Theorem 1.1.

For the case $\delta = 0$ we define the operator L_0 on $E_0 := L_p(\Omega) \times W_p^{2-2/p}(\Sigma)$ by means of

$$\begin{aligned} D(L_0) :=& \big\{(v,\rho) \in W_p^2(\Omega\setminus\Sigma) \times W_p^{4-1/p}(\Sigma) : [d\partial_\nu v] \in W_p^{2-2/p}(\Sigma), \\ &\partial_\nu v = 0 \text{ on } \partial\Omega,\ [v] = 0 \text{ on } \Sigma,\ \ v = \sigma\mathcal{A}_\Sigma\rho \text{ on } \Sigma\big\}, \\ L_0(v,\rho) :=& \big((d/\kappa)\Delta v, [(d/l)\partial_\nu v]\big), \quad (v,\rho) \in D(L_0). \end{aligned}$$

In case $\delta > 0$, we instead set $E_\delta := L_p(\Omega) \times W_p^{1-1/p}(\Sigma)$, and we define L_δ by

$$\begin{aligned} D(L_\delta) :=& \big\{(v,\rho) \in W_p^2(\Omega\setminus\Sigma) \times W_p^{3-1/p}(\Sigma) : \\ &\partial_\nu v = 0 \text{ on } \partial\Omega,\ [v] = 0 \text{ on } \Sigma,\ v - (\delta/l)[d\partial_\nu v] = \sigma\mathcal{A}_\Sigma\rho \text{ on } \Sigma\big\}, \\ L_\delta(v,\rho) :=& \big((d/\kappa)\Delta v, [(d/l)\partial_\nu v]\big), \quad (v,\rho) \in D(L_\delta). \end{aligned}$$

We remark that L_0 and L_δ differ only by their respective domains of definition. It will be shown in [7] that the operators L_δ generate an analytic semigroup on E_δ. This property, in conjunction with the spectral information contained in the next theorem, will be crucial in proving global existence and convergence of solutions for problem (1.3) that start out close to an equilibrium, which will be provided in [27]; see also Remark 1.2(c).

THEOREM 2.1. *Suppose $1 < p < \infty$, and let $l > 0$. For $\delta \geq 0$ let L_δ be defined as above.*

(a) *The spectrum of L_δ consists of countably many real eigenvalues of finite algebraic multiplicity and is independent of p.*

(b) *0 is an eigenvalue of L_δ with geometric multiplicity $(n+1)$. The null space of L_δ is spanned by*

$$(-1, Y_0), (0, Y_1), \ldots, (0, Y_n), \tag{2.1}$$

where $Y_0 = R^2/\sigma$, and where Y_j, $1 \leq j \leq n$, are the spherical harmonics of degree 1 (normalized by the orthogonality condition $(Y_i|Y_j)_\Sigma = \delta_{ij}$).

(c) *Suppose that the degeneracy condition*

$$(\kappa|1)_\Omega := \kappa_1|\Omega_1| + \kappa_2|\Omega_2| = l|\Sigma|R^2/\sigma \tag{2.2}$$

holds. Then the eigenvalue 0 *has algebraic multiplicity* $(n+2)$.

(d) *If the degeneracy condition* (2.2) *does not hold, then* 0 *is semi-simple; that is,* $N(L_\delta^2) = N(L_\delta)$.

(e) *If* $\sigma(\kappa|1)_\Omega \le l|\Sigma|R^2$, *then* L_δ *has no positive eigenvalues.*

(f) *If* $\sigma(\kappa|1)_\Omega > l|\Sigma|R^2$, *then* L_δ *has exactly one positive simple eigenvalue.*

Proof. (a) By the compact embeddings $D(L_\delta) \hookrightarrow E_\delta$, the spectrum of L_δ consists of eigenvalues of finite algebraic multiplicity. The assertion that all eigenvalues are real will be proved in section 4.

Let $1 < p < \infty$ be fixed, and suppose that λ is an eigenvalue of L_δ with a corresponding eigenfunction (v, ρ). Then $v \in W_p^2(\Omega \setminus \Sigma)$, and v solves the elliptic transmission problem

$$\begin{cases} \kappa\lambda v - d\Delta v = 0 & \text{in} \quad \Omega \setminus \Sigma, \\ \partial_\nu v = 0 & \text{on} \quad \partial\Omega, \\ [v] = 0 & \text{on} \quad \Sigma, \\ -[d\partial_\nu v] = -\lambda l\rho & \text{on} \quad \Sigma, \end{cases}$$

with $\rho \in W_p^{4-\mathrm{sign}(\delta)-1/p}(\Sigma)$, where $\mathrm{sign}(\delta) = 1$ if $\delta > 0$, and $\mathrm{sign}(\delta) = 0$ if $\delta = 0$. Due to Sobolev's imbedding theorem we have that $\rho \in W_{p_1}^{1-1/p_1}(\Sigma)$, where $p_1 \in (p, \infty)$ is appropriately chosen. Proposition 5.1 then yields $v \in W_{p_1}^2(\Omega \setminus \Sigma)$. Next, we recall that ρ satisfies

$$\sigma\mathcal{A}_\Sigma\rho = v - (\delta/l)[d\partial_\nu v] =: h \quad \text{on} \quad \Sigma.$$

Since $v \in W_{p_1}^2(\Omega \setminus \Sigma)$ we see that $h \in W_{p_1}^{2-\mathrm{sign}(\delta)-1/p_1}(\Sigma)$, and we obtain from the properties of the elliptic differential operator $\mathcal{A}_\Sigma$ that $\rho \in W_{p_1}^{4-\mathrm{sign}(\delta)-1/p_1}(\Sigma)$. The arguments given above can now be iterated a finite number of times to show that

$$(v, \rho) \in W_q^2(\Omega \setminus \Sigma) \times W_q^{4-\mathrm{sign}(\delta)-1/q}(\Sigma)$$

for any fixed $q > p$. Clearly, this is also true for any $q < p$. We have, thus, shown that the spectrum of L_δ is independent of p. The properties listed in (b)–(d) are proved in section 3, and assertion (e) is shown in section 4 while (f) is established in section 5. ∎

PROPOSITION 2.2. *Let* $1 < p < \infty$. *Suppose that* $(\lambda, v, \rho) \in \mathbb{R} \times W_p^2(\Omega\setminus\Sigma) \times W_p^2(\Sigma)$ *solves the eigenvalue problem* (1.9). *Then the functions* (v, ρ) *are smooth; that is,*

$$v|_{\Omega_i} \in C^\infty(\bar\Omega_i), \quad \rho \in C^\infty(\Sigma).$$

Proof. This follows from a similar bootstrapping argument as in the proof of Theorem 2.1(a), based on regularity properties of the elliptic transmission problems (3.4) and (5.2), and regularity properties of the differential operator $\mathcal{A}_\Sigma$. ∎

Due to Theorem 2.1 and Proposition 2.2 we may restrict our attention to the eigenvalue problem (1.9) in the Hilbert space setting of $L_2(\Omega) \times L_2(\Sigma)$. In the following, we use the notation $(\cdot|\cdot)_\Omega$ and $\|\cdot\|_\Omega$ for the inner product and the norm in $L_2(\Omega)$, respectively, and similarly for $L_2(\Sigma)$.

3. The trivial eigenvalue. Let us first look at the eigenvalue problem (1.9) with $\lambda = 0$. Obviously, here $l \in \mathbb{R}$ can be arbitrary, and also $\delta \in \mathbb{R}$. For this purpose we recall some properties of the operator $\mathcal{A}_\Sigma$.

PROPOSITION 3.1. *Let $\Sigma = S_R(x_0) \subset \mathbb{R}^n$ be a sphere of radius R and center x_0, and let*

$$A_\Sigma = -\frac{1}{n-1}\left(\frac{n-1}{R^2} + \Delta_\Sigma\right)$$

be defined on $L_2(\Sigma)$ with domain $W_2^2(\Sigma)$. Then the following assertions hold:

(a) *A_Σ is self-adjoint. Its spectrum consists of countably many eigenvalues $\lambda_k = \frac{1}{(n-1)R^2}\big(k(k+n-2) - (n-1)\big)$ with $k \geq 0$. The eigenfunctions are given by the spherical harmonics of degree k.*

(b) *The kernel of A_Σ is given by $N(A_\Sigma) = \operatorname{span}\{Y_1, \ldots, Y_n\}$, where Y_j denotes the spherical harmonics of degree 1 on Σ, normalized by $(Y_i|Y_j)_\Sigma = \delta_{ij}$.*

(c) *The range of A_Σ, $R(A_\Sigma)$ is closed, and we have $L_2(\Sigma) = N(A_\Sigma) \oplus R(A_\Sigma)$.*

(d) *There is precisely one negative eigenvalue, namely $-1/R^2$, with eigenfunction 1, which is simple.*

(e) *A_Σ is positive semi-definite on $L_{2,0}(\Sigma) = \{\rho \in L_2(\Sigma) : (\rho|1)_\Sigma = 0\}$ and positive definite on*

$$L_{2,0}(\Sigma) \cap R(A_\Sigma) = \{\rho \in L_2(\Sigma) : (\rho|1)_\Sigma = (\rho|Y_j)_\Sigma = 0,\ j = 1, \ldots, n\}.$$

Proof. We can assume, without loss of generality, that $\Sigma = S_R(0) = R\,\mathbb{S}^{n-1}$, where $\mathbb{S}^{n-1}$ denotes the standard unit sphere in $\mathbb{R}^n$. Let $\Phi : \Sigma \to \mathbb{S}^{n-1}$ be defined by $p \mapsto (1/R)p$. Then Φ is a smooth diffeomorphism of Σ into $\mathbb{S}^{n-1}$, and one readily verifies that

$$(g|h)_{L_2(\Sigma)} = R^{n-1}(\Phi_* g|\Phi_* h)_{L_2(\mathbb{S}^{n-1})}, \quad \Delta_\Sigma = (1/R^2)\,\Phi^* \Delta_{\mathbb{S}^{n-1}}\,\Phi_*, \tag{3.1}$$

where Φ^* and Φ_* are the pull-back and push-forward operators, respectively. We then have

$$(\lambda - A_\Sigma)\rho = 0 \iff \left(\lambda + \frac{1}{(n-1)R^2}\big((n-1) + \Delta_{\mathbb{S}^{n-1}}\big)\right)\Phi_*\rho = 0, \tag{3.2}$$

which shows that λ is an eigenvalue of A_Σ iff

$$\lambda = \frac{1}{(n-1)R^2}\big(\mu - (n-1)\big) \tag{3.3}$$

with μ an eigenvalue of $-\Delta_{\mathbb{S}^{n-1}}$. The assertions in (a)–(b) and (d)–(e) follow now from (3.1)–(3.3) and well-known results for the Laplace–Beltrami operator on $\mathbb{S}^{n-1}$; see, for instance, [30, section 31]. Since A_Σ has compact resolvent we conclude that $R(A_\Sigma)$ is closed, and the fact that A_Σ is self-adjoint then implies the remaining assertion in (c). ∎

Before we proceed we need the following result on the elliptic transmission problem:

$$\begin{cases} -d\Delta v = f & \text{in} \quad \Omega \setminus \Sigma, \\ \partial_\nu v = 0 & \text{on} \quad \partial\Omega, \\ [v] = 0 & \text{on} \quad \Sigma, \\ -[d\partial_\nu v] = g & \text{on} \quad \Sigma. \end{cases} \tag{3.4}$$

PROPOSITION 3.2. *Let $1 < p < \infty$. Then the following hold:*

(a) *The transmission problem* (3.4) *has a solution $v \in W_p^2(\Omega \setminus \Sigma)$ if and only if $(f, g) \in L_p(\Omega) \times W_p^{1-1/p}(\Sigma)$ and the compatibility condition*

$$(f|1)_\Omega + (g|1)_\Sigma = 0$$

is satisfied. The solution is unique with the normalization $(\kappa|v)_\Omega = 0$.

(b) *Let $v = T_0 g$ be the unique solution of* (3.4) *with $f = 0$, $(g|1)_\Sigma = 0$, and $(\kappa|v)_\Omega = 0$. Then T_0 is self-adjoint and positive definite on $L_2(\Sigma)$; that is, there exists a positive constant $c = c(d_i, \Omega_i)$ such that*

$$(T_0 g|g)_{L_2(\Sigma)} \geq c \, \|g\|^2_{L_2(\Sigma)}, \qquad g \in W_2^{1/2}(\Sigma).$$

Proof. (a) This proof follows from known results in elliptic theory since the Lopatinskii–Shapiro conditions are satisfied.

(b) Let $g, h \in W_2^{1/2}(\Sigma)$ be given. Then we have

$$\begin{aligned}(T_0 g|h)_\Sigma = (T_0 g|[-d\partial_\nu T_0 h])_\Sigma &= (d\,\nabla T_0 g|\nabla T_0 h)_\Omega \\ &= (-[d\partial_\nu T_0 g]|T_0 h)_\Sigma = (g|T_0 h)_\Sigma,\end{aligned}$$

thus showing that T_0 is symmetric. For $v := T_0 g$ the computation above yields

$$(T_0 g|g)_\Sigma = (d\,\nabla v|\nabla v)_\Omega.$$

On the other hand, setting $v_i = v|_{\Omega_i}$ we obtain

$$\begin{aligned}\|g\|_{L_2(\Sigma)} &= \|d_1 \partial_\nu v_1 - d_2 \partial_\nu v_2\|_{L_2(\Sigma)} \leq c\big(\|v_1\|_{W_2^2(\Omega_1)} + \|v_2\|_{W_2^2(\Omega_2)}\big) \\ &\leq c\big(\|v_1\|_{L_2(\Omega_1)} + \|\Delta v_1\|_{L_2(\Omega_1)} + \|v_2\|_{L_2(\Omega_2)} + \|\Delta v_2\|_{L_2(\Omega_2)}\big) \\ &= c\|v\|_{L_2(\Omega)} \leq c\|v\|_{W_2^1(\Omega)} \leq c\|\nabla v\|_{L_2(\Omega)} \leq c(T_0 g|g)_\Sigma^{1/2}.\end{aligned}$$

Here we used the fact that

$$v = T_0 g \in W_2^1(\Omega) \cap W_2^2(\Omega \setminus \Sigma).$$

Moreover, we used that $\big(\|\cdot\|_{L_2(\Omega_i)} + \|\Delta \cdot\|_{L_2(\Omega_i)}\big)$ defines an equivalent norm on $W_2^2(\Omega_i)$, and also that $\|\nabla u\|_{L_2(\Omega)}$ defines an equivalent norm on $W_2^1(\Omega)$ for all functions $u \in W_2^1(\Omega)$ with $(\kappa|u)_\Omega = 0$. This completes the proof of Proposition 3.2. □

We are now ready to establish the assertions (b)–(d) of Theorem 2.1.

(b) Suppose that (v, ρ) is a solution of (1.9) with $\lambda = 0$. Then, taking the inner product of $(1.9)_1$ with v, the divergence theorem and $(1.9)_{2,4}$ show that v is constant on $\Omega \setminus \Sigma$; hence, v is constant on Ω and $v = \sigma \mathcal{A}_\Sigma \rho$ due to $(1.9)_3$. A special solution of this problem is $\rho_0 = -R^2 v/\sigma$, and the solutions of the corresponding homogeneous equation are the spherical harmonics Y_j on Σ for $j = 1, \ldots, n$. Thus we obtain an $(n+1)$-dimensional null space spanned by(2.1), which proves Theorem 2.1(b).

This null space is tangent to the (n+1)-dimensional manifold of equilibria, where $(0, Y_j)$ corresponds to the center x_0, and $(-1, Y_0)$ corresponds to the radius R. Note that the null spaces of L_0 and L_δ, $\delta > 0$, coincide.

(c) Suppose that (2.2) holds. Then there exists a pair $(v^*, \rho^*) \in N(L_\delta^2) \setminus N(L_\delta)$. Indeed, this can be seen as follows: we first solve (3.4) with $(f, g) = (-\kappa, lR^2/\sigma)$. According to Proposition 3.2, this problem has a unique solution v_0 with $(\kappa|v_0)_\Omega = 0$ since the necessary compatibility condition is precisely (2.2).

Set $v^* = v_0 + l\sum_{j=1}^n \alpha_j T_0 Y_j$. Then v^* satisfies

$$\begin{cases} -(d/\kappa)\Delta v^* = -1 & \text{in} \quad \Omega\setminus\Sigma, \\ \partial_\nu v^* = 0 & \text{on} \quad \partial\Omega, \\ [v^*] = 0 & \text{on} \quad \Sigma, \\ -[(d/l)\partial_\nu v^*] = R^2/\sigma + \Sigma_{j=1}^n \alpha_j Y_j & \text{on} \quad \Sigma. \end{cases} \tag{3.5}$$

We now want to solve $v^* = \sigma A_\Sigma \rho + (\delta/l)[d\partial_\nu v^*]$ in terms of ρ; that is, we consider the problem

$$\sigma A_\Sigma \rho = v^* - (\delta/l)[d\partial_\nu v^*] =: h.$$

According to Proposition 3.1(b)–(c) this problem has a solution ρ^* iff $(h|Y_i)_\Sigma = 0$ for $i = 1,\dots,n$. The conditions $(h|Y_i)_\Sigma = 0$ will then be employed to determine the coefficients α_j. A short computation yields

$$(h|Y_i)_\Sigma = 0 \quad \Leftrightarrow \quad \delta\alpha_i + \sum_{j=1}^n l(T_0 Y_j|Y_i)_\Sigma\, \alpha_j = -(v_0|Y_i)_\Sigma, \quad i = 1,\dots,n.$$

Since T_0 is self-adjoint and positive definite on $L_2(\Sigma)$, there exists a unique solution of this system, as we shall see in (6.7). Due to $\sigma A_\Sigma(R^2/\sigma + \sum_{j=1}^n \alpha_j Y_j) = -1$ (see Proposition 3.1), we conclude that $(v^*,\rho^*) \in D(L_\delta^2)$. It is then easy to see that $L_\delta(v^*,\rho^*) \neq (0,0)$ and $L_\delta^2(v^*,\rho^*) = (0,0)$. These facts in combination with part (b) show that

$$N(L_\delta^2) = N(L_\delta) \oplus \operatorname{span}\{(v^*,\rho^*)\}, \tag{3.6}$$

in the degenerate case where (2.2) holds.

Next we show, still in the degenerate case (2.2), that $N(L_\delta^3) = N(L_\delta^2)$. In fact, if $(v,\rho) \in N(L_\delta^3)$, then $L_\delta(v,\rho) = (v_N,\rho_N) + \beta(v^*,\rho^*)$ for some $(v_N,\rho_N) \in N(L_\delta)$ and some scalar β. For solvability of this equation, the compatibility condition

$$\big(\kappa(v_N + \beta v^*)|1\big)_\Omega + l\big(\rho_N + \beta\rho^*|1\big)_\Sigma = 0$$

must be valid. Due to the degeneracy condition we have $(\kappa v_N|1)_\Omega + l(\rho_N|1)_\Sigma = 0$, and the compatibility condition is reduced to $\beta\big\{(\kappa v^*)|1)_\Omega + l(\rho^*|1)_\Sigma\big\} = 0$. Using the property that $-(d/\kappa)\Delta v^* = -1$ (see (3.5)), we obtain

$$\begin{aligned} -\big\{(\kappa v^*|1)_\Omega + l(\rho^*|1)_\Sigma\big\} &= -(d\Delta v^*|v^*)_\Omega - l(\rho^*|1)_\Sigma \\ &= \|\sqrt{d}\,\nabla v^*\|_\Omega^2 + ([d\partial_\nu v^*]|v^*)_\Sigma - l(\rho^*|1)_\Sigma \\ &= \|\sqrt{d}\,\nabla v^*\|_\Omega^2 - l\left(R^2/\sigma + \sum_{j=1}^n \alpha_j Y_j|\sigma A_\Sigma\rho^*\right)_\Sigma + (\delta/l)\|[d\partial_\nu v^*]\|_\Sigma^2 - l(\rho^*|1)_\Sigma \\ &= \|\sqrt{d}\,\nabla v^*\|_\Omega^2 + (\delta/l)\|[d\partial_\nu v^*]\|_\Sigma^2 \end{aligned}$$

since A_Σ is self-adjoint on $L_2(\Sigma)$ and $\sigma A_\Sigma(R^2/\sigma + \sum_{j=1}^n \alpha_j Y_j) = -1$; see Proposition 3.1. This implies $\beta = 0$, i.e., $(v,\rho) \in N(L_\delta^2)$, thus establishing Theorem 2.1(c).

(d) Let us examine when the eigenvalue $\lambda_0 = 0$ of L_δ is semi-simple. Assume that $(v,\rho) \in D(L_\delta^2)$ is such that $L_\delta^2(v,\rho) = 0$. Then

$$L_\delta(v,\rho) = \alpha_0(-1,Y_0) + \sum_{j=1}^n \alpha_j(0,Y_j).$$

This implies that

$$\begin{cases} -(d/\kappa)\Delta v = \alpha_0 & \text{in} \quad \Omega \setminus \Sigma, \\ \partial_\nu v = 0 & \text{on} \quad \partial\Omega, \\ [v] = 0 & \text{on} \quad \Sigma, \\ -[(d/l)\partial_\nu v] = -\sum_{j=0}^{n} \alpha_j Y_j & \text{on} \quad \Sigma, \\ v = \sigma \mathcal{A}_\Sigma \rho + (\delta/l)[d\partial_\nu v] & \text{on} \quad \Sigma. \end{cases}$$

According to Proposition 3.2 we necessarily have

$$\alpha_0(\kappa_1|\Omega_1| + \kappa_2|\Omega_2|) = l\sum_{j=0}^{n} \alpha_j (Y_j|1)_\Sigma = l\alpha_0|\Sigma|R^2/\sigma,$$

since the mean value of Y_j over Σ is zero for $j \geq 1$. Assuming the nondegeneracy condition

$$(\kappa|1)_\Omega := \kappa_1|\Omega_1| + \kappa_2|\Omega_2| \neq l|\Sigma|R^2/\sigma, \tag{3.7}$$

we conclude that $\alpha_0 = 0$. But then

$$0 = -\int_\Omega d\Delta v v \, dx = \int_\Omega d|\nabla v|^2 \, dx + \int_\Sigma [d\partial_\nu v] v \, ds,$$

which further yields

$$0 = \|\sqrt{d}\,\nabla v\|_\Omega^2 + l\sigma \sum_{j=1}^{n} \alpha_j (Y_j|\mathcal{A}_\Sigma \rho)_\Sigma + (\delta/l)\|[d\partial_\nu v]\|_\Sigma^2 = \|\sqrt{d}\,\nabla v\|_\Omega^2 + (\delta/l)\|[d\partial_\nu v]\|_\Sigma^2$$

since $\mathcal{A}_\Sigma$ is self-adjoint on $L_2(\Sigma)$ and $A_\Sigma Y_j = 0$ for $j \geq 1$. We conclude that v is constant in Ω and that $0 = [d\partial_\nu v] = l\sum_{j=1}^{n} \alpha_j Y_j$; hence, $\alpha_j = 0$ for all j. This shows that $\lambda_0 = 0$ is a semi-simple eigenvalue of L_δ, that is, $N(L_\delta^2) = N(L_\delta)$ for $\delta \geq 0$, provided the nondegeneracy condition (3.7) is valid, and this proves the assertion of Theorem 2.1(d).

4. Nontrivial eigenvalues. Now we consider the eigenvalue problem (1.9) for $\lambda \in \mathbb{C}$, $\lambda \neq 0$, in case $l > 0$. Suppose that $\lambda \neq 0$ is an eigenvalue with nontrivial eigenfunction (v, ρ). Taking the inner product in $L_2(\Omega)$ of the first equation in (1.9) with v and using the divergence theorem, we get

$$\begin{aligned} \lambda\|\sqrt{\kappa} v\|_\Omega^2 = (d\Delta v|v)_\Omega &= -\|\sqrt{d}\,\nabla v\|_\Omega^2 - ([d\partial_\nu v]|v)_\Sigma \\ &= -\|\sqrt{d}\,\nabla v\|_\Omega^2 - (\delta/l)\|[d\partial_\nu v]\|_\Sigma^2 - \lambda l\sigma(\rho|A_\Sigma \rho)_\Sigma; \end{aligned}$$

hence, we obtain the identity

$$\lambda\Big(\|\sqrt{\kappa} v\|_\Omega^2 + l\sigma(\rho|A_\Sigma\rho)_\Sigma\Big) + \|\sqrt{d}\,\nabla v\|_\Omega^2 + (\delta/l)\|[d\partial_\nu v]\|_\Sigma^2 = 0. \tag{4.1}$$

If $\operatorname{Im}\lambda \neq 0$, then $\|\sqrt{\kappa} v\|_\Omega^2 + l\sigma(\rho|A_\Sigma\rho)_\Sigma = 0$; hence, $\|\sqrt{d}\,\nabla v\|_\Omega^2 + (\delta/l)\|[d\partial_\nu]\|_\Sigma^2 = 0$. We conclude that v is constant and $(1.9)_{1,4}$ now implies that $(v, \rho) = (0, 0)$ since

$\lambda \neq 0$. Therefore the eigenvalues and eigenfunctions are real, thus establishing Theorem 2.1(a).

Using $\lambda l\rho = [d\partial_\nu v]$ and the fact that λ is real, we may rewrite (4.1) as

$$\lambda\Big(\|\sqrt{\kappa}v\|_\Omega^2 + l\sigma(\rho|A_\Sigma\rho)_\Sigma + \lambda l\delta\|\rho\|_\Sigma^2\Big) + \|\sqrt{d}\,\nabla v\|_\Omega^2 = 0. \tag{4.2}$$

Integrating the eigenvalue equation (1.9) we obtain

$$\lambda(v|\kappa)_\Omega = (d\Delta v|1)_\Omega = -([d\partial_\nu v]|1)_\Sigma = -\lambda l(\rho|1)_\Sigma;$$

hence, dividing by λ,

$$(v|\kappa)_\Omega + l(\rho|1)_\Sigma = 0. \tag{4.3}$$

Splitting $\rho = \rho_0 + \bar\rho$ and $v = v_0 + \bar v$, where $(\rho_0|1)_\Sigma = (v_0|\kappa)_\Omega = 0$, from (4.2) and (4.3) we derive an identity equivalent to (4.2), namely,

$$\begin{aligned}\lambda\Big(\|\sqrt{\kappa}v_0\|_\Omega^2 + l\sigma(\rho_0|A_\Sigma\rho_0)_\Sigma + \lambda l\delta\|\rho_0\|_\Sigma^2\Big) + \|\sqrt{d}\,\nabla v\|_\Omega^2 \\ + \lambda\Big(\lambda\delta + \frac{l|\Sigma|}{(\kappa|1)_\Omega} - \frac{\sigma}{R^2}\Big) l|\Sigma|\bar\rho^2 = 0.\end{aligned} \tag{4.4}$$

Since A_Σ is positive semi-definite on $L_{2,0}(\Sigma)$, the L_2-functions with mean zero, we see that in case $\lambda > 0$, (4.4) implies $v =$ constant, and hence $(v,\rho) = (0,0)$, provided that

$$(\kappa|1)_\Omega \leq l|\Sigma|R^2/\sigma. \tag{4.5}$$

Consequently, (1.9) cannot have positive eigenvalues if the stability condition (4.5) is satisfied, thus proving the assertion of Theorem 2.1(e).

5. The unstable eigenvalue. As far as we know, for $l > 0$ and

$$\zeta := \frac{\sigma(\kappa|1)_\Omega}{l|\Sigma|R^2} \leq 1 \tag{5.1}$$

there are no positive eigenvalues; however, the algebraic eigenspace of L_δ rises in dimension by one when ζ becomes 1. This indicates that for $\zeta > 1$ there is exactly one algebraically simple eigenvalue $\lambda_* > 0$. We want to prove that this is indeed the case. In order to do so, we consider the following transmission problem:

$$\begin{cases} \lambda\kappa v - d\Delta v = f & \text{in} \quad \Omega\setminus\Sigma, \\ \partial_\nu v = 0 & \text{on} \quad \partial\Omega, \\ [v] = 0 & \text{on} \quad \Sigma, \\ -[d\partial_\nu v] = g & \text{on} \quad \Sigma. \end{cases} \tag{5.2}$$

Then the following result holds.

PROPOSITION 5.1. *Let $1 < p < \infty$ and $\operatorname{Re}\lambda > 0$. Then the following hold:*

(a) *Problem (5.2) has precisely one solution $v \in W_p^1(\Omega) \cap W_p^2(\Omega\setminus\Sigma)$ iff $(f,g) \in L_p(\Omega) \times W_p^{1-1/p}(\Sigma)$.*

(b) *Let T_λ be the solution operator for (5.2) with $f = 0$. Given any number $\theta \in (0,\pi)$, there exist positive numbers $\lambda_0 = \lambda_0(\theta, d_i, \kappa_i, \Omega_i)$ and $M_0 = M_0(\theta, d_i, \kappa_i, \Omega_i)$ such that*

$$\|T_\lambda g\|_{L_p(\Sigma)} \leq M_0|\lambda|^{-1/2}\|g\|_{L_p(\Sigma)}$$

for $g \in W_p^{1-1/p}(\Sigma)$, whenever $|\lambda| \geq \lambda_0$ and $|\arg\lambda| \leq \theta$.

(c) *For $\lambda > 0$, T_λ is positive definite on $L_2(\Sigma)$; that is, there exists a positive constant $\beta = \beta(d_i, \kappa_i, \Omega_i)$ such that*

$$(T_\lambda g|g)_{L_2(\Sigma)} \geq \beta \frac{\sqrt{\lambda}}{1+\lambda} \|g\|^2_{L_2(\Sigma)}, \qquad g \in W_2^{1/2}(\Sigma).$$

Proof. (a) This proof follows from known results in elliptic theory.

(b) Suppose that $\Omega_1 = \mathbb{R}^n_-$ and $\Omega_2 = \mathbb{R}^n_+$, with $\mathbb{R}^n_\pm = \{(x', x_n) \in \mathbb{R}^n : \pm x_n < 0\}$. Then one readily obtains that

$$T_\lambda g|_{\mathbb{R}^{n-1}\times\{0\}} = \mathcal{F}^{-1}\left(m_\lambda \mathcal{F} g\right),$$

where $\mathcal{F}$ denotes the Fourier transform in the tangential variables, and where

$$m_\lambda(\xi) = \frac{1}{\sqrt{d_1}\sqrt{\kappa_1\lambda + d_1|\xi|^2} + \sqrt{d_2}\sqrt{\kappa_2\lambda + d_2|\xi|^2}}\,.$$

The assertion then follows from Mikhlin's multiplier theorem. The general case can be obtained by the usual procedure of localization.

(c) Let $g, h \in W_2^{1/2}(\Sigma)$ be given. Then we have

$$\begin{aligned}(T_\lambda g|h)_\Sigma = (T_\lambda g|[-d\partial_\nu T_{\bar\lambda} h])_\Sigma &= \lambda(\kappa\, T_\lambda g|T_{\bar\lambda} h)_\Omega + (d\,\nabla T_\lambda g|\nabla T_{\bar\lambda} h)_\Omega \\ &= (-[d\partial_\nu T_\lambda g]|T_{\bar\lambda} h)_\Sigma = (g|T_{\bar\lambda} h)_\Sigma,\end{aligned}$$

showing that $T_\lambda^* = T_{\bar\lambda}$, in particular, that T_λ is symmetric for $\lambda > 0$. For $v := T_\lambda g$, $\lambda > 0$, the computation above yields

$$(T_\lambda g|g)_\Sigma = \lambda(\kappa v|v)_\Omega + (d\,\nabla v|\nabla v)_\Omega. \tag{5.3}$$

Setting $v_i = v|_{\Omega_i}$ we conclude similarly as in the proof of Proposition 3.2 that

$$\begin{aligned}\|g\|_{L_2(\Sigma)} = \|d_1\partial_\nu v_1 - d_2\partial_\nu v_2\|_{L_2(\Sigma)} &\leq c\big(\|d_1 v_1\|_{W_2^2(\Omega_1)} + \|d_2 v_2\|_{W_2^2(\Omega_2)}\big) \\ &\leq c\big(\|d_1 v_1\|_{L_2(\Omega_1)} + \|d_1\Delta v_1\|_{L_2(\Omega_1)} + \|d_2 v_2\|_{L_2(\Omega_2)} + \|d_2\Delta v_2\|_{L_2(\Omega_2)}\big) \\ &= c\big(\|d_1 v_1\|_{L_2(\Omega_1)} + \lambda\|\kappa_1 v_1\|_{L_2(\Omega_1)} + \|d_2 v_2\|_{L_2(\Omega_2)} + \lambda\|\kappa_2 v_2\|_{L_2(\Omega_2)}\big) \\ &\leq c_\lambda\sqrt{\lambda}\|v\|_{L_2(\Omega)} \leq c_\lambda\big(\sqrt{\lambda}\|v\|_{L_2(\Omega)} + \|\nabla v\|_{L_2(\Omega)}\big) \leq c_\lambda\,(T_\lambda g|g)_\Sigma^{1/2},\end{aligned}$$

where $c_\lambda = c(d_i, \kappa_i, \Omega_i)(1+\lambda)/\sqrt{\lambda}$. In the estimates above we have used that $v = T_\lambda g \in W_2^1(\Omega) \cap W_2^2(\Omega\setminus\Sigma)$, and that $\big(\|\cdot\|_{L_2(\Omega_i)} + \|\Delta\cdot\|_{L_2(\Omega_i)}\big)$ defines an equivalent norm on $W_2^2(\Omega_i)$ and, lastly, we employed (5.3). This completes the proof of Proposition 5.1. □

We assume now that $\lambda > 0$ is a fixed number. For given $\rho \in W_2^{1/2}(\Sigma)$, let v be the solution of the transmission problem

$$\begin{cases} \lambda\kappa v - d\Delta v = 0 & \text{in} \quad \Omega\setminus\Sigma, \\ \partial_\nu v = 0 & \text{on} \quad \partial\Omega, \\ [v] = 0 & \text{on} \quad \Sigma, \\ -[d\partial_\nu v] = -\lambda l\rho & \text{on} \quad \Sigma. \end{cases} \tag{5.4}$$

Then $v = T_\lambda(-\lambda l\rho) = -\lambda l T_\lambda\rho$ with T_λ being the solution operator introduced above. By inserting this representation of v into the equation $v = \sigma A_\Sigma\rho + \lambda\delta\rho$, we obtain the problem

$$\lambda\delta\rho + \lambda l T_\lambda\rho + \sigma A_\Sigma\rho = 0, \tag{5.5}$$

which is equivalent to the eigenvalue problem.
Setting $B_\lambda(s) := \lambda\delta I + \lambda l T_\lambda + s\sigma A_\Sigma$ for $s > 0$ and employing Proposition 5.1(c), we obtain the estimate

$$\begin{aligned}(B_\lambda(s)\rho|\rho)_\Sigma &\geq \lambda(\delta+\gamma l)\|\rho\|_\Sigma^2 + s\sigma(A_\Sigma\rho|\rho)_\Sigma \\ &= \lambda(\delta+\gamma l)\|\rho_0\|_\Sigma^2 + s\sigma(A_\Sigma\rho_0|\rho_0)_\Sigma + \big\{\lambda(\delta+\gamma l) - s\sigma/R^2\big\}|\Sigma|\bar\rho^2,\end{aligned}$$

where $\gamma = \beta\sqrt{\lambda}/(1+\lambda)$ and $\rho = \rho_0 + \bar\rho$ with $(\rho_0|1)_\Sigma = 0$. Since $(A_\Sigma\rho_0|\rho_0)_\Sigma \geq 0$ we see that all of the terms in the previous line are nonnegative, provided $\lambda(\delta+\gamma l) \geq s\sigma/R^2$, i.e., for small s. Hence, for small $s > 0$, the operator $B_\lambda(s)$ is positive definite, which means that λ cannot be an eigenvalue of (1.9), where in the third line of (1.9) σ is replaced by $s\sigma$. On the other hand, choosing $\rho = 1$ we have

$$(B_\lambda(s)1|1)_\Sigma = \lambda\delta|\Sigma| + \lambda l(T_\lambda 1|1)_\Sigma - s\sigma|\Sigma|/R^2 < 0$$

if s becomes large. Now we set

$$s_* := s_*(\lambda) := \sup\{s > 0 :\ B_\lambda(s) \text{ is positive definite}\}.$$

Then $B_\lambda(s_*)$ is still semi-definite, but not definite, and hence, by compactness of the resolvent, has a nontrivial kernel. Therefore, for a given $\lambda > 0$ there is an $s_* = s_*(\lambda)$ such that λ is an eigenvalue of (1.9), where σA_Σ is replaced by $s_*\sigma A_\Sigma$ in the third line.

Next, we show that positive eigenvalues are simple. Rewrite (5.5) as

$$\lambda\delta\rho_0 + \lambda l T_\lambda\rho_0 + \sigma A_\Sigma\rho_0 = -\big\{\lambda\delta + \lambda l T_\lambda 1 - \sigma/R^2\big\}\bar\rho.$$

Since B_λ is positive definite on $L_{2,0}(\Sigma)$, this equation has precisely one solution for given $\bar\rho$, which shows that the eigenspace $N(\lambda - L_\delta)$ is at most one-dimensional for any given $\lambda > 0$.

To show that nontrivial eigenvalues are semi-simple, suppose that

$$(\lambda - L_\delta)(v,\rho) = (v_1,\rho_1), \quad (\lambda - L_\delta)(v_1,\rho_1) = 0.$$

Then by Green's formula

$$\begin{aligned}\|\sqrt{\kappa}v_1\|_\Omega^2 &= (\lambda\kappa v - d\Delta v|v_1)_\Omega \\ &= (v|\lambda\kappa v_1 - d\Delta v_1)_\Omega + ([d\partial_\nu v]|v_1)_\Sigma - (v|[d\partial_\nu v_1])_\Sigma \\ &= (\delta/l)([d\partial_\nu v]|[d\partial_\nu v_1])_\Sigma + l\sigma(\lambda\rho - \rho_1|A_\Sigma\rho_1)_\Sigma \\ &\quad - (\delta/l)([d\partial_\nu v]|[d\partial_\nu v_1])_\Sigma - \lambda l\sigma(A_\Sigma\rho|\rho_1)_\Sigma \\ &= -l\sigma(\rho_1|A_\Sigma\rho_1),\end{aligned}$$

which yields

$$\|\sqrt{\kappa}v_1\|_\Omega^2 + l\sigma(\rho_1|A_\Sigma\rho_1).$$

It follows now from (4.1) that v_1 is constant on $\Omega\setminus\Sigma$. Since $\lambda \neq 0$, we then obtain from (1.9) that $(v_1,\rho_1) = (0,0)$. Thus any nontrivial eigenvalue is semi-simple, and, in particular, positive eigenvalues are algebraically simple.

We want to show that for $\zeta > 1$ there is precisely one positive eigenvalue $\lambda_* > 0$. For this purpose we fix the parameters d, l, σ, δ as well as R, but replace κ by $s\kappa$

in the first line of (1.9). Fixing $\mu = \lambda s$ and scaling $\rho \mapsto \rho/s$, we obtain the scaling $\sigma \mapsto s\sigma$. The argument given previously then shows that there is $s_* > 0$ such that μ is a simple eigenvalue of the scaled problem; hence, $\lambda_* = \mu/s_* > 0$ is a simple positive eigenvalue for (1.9) with κ replaced by $s_*\kappa$ in the first line. Since $\lambda_* = \lambda_*(s_*)$ is simple, the eigenvalue problem

$$\begin{cases} \lambda_* s\kappa v - d\Delta v = 0 & \text{in} \quad \Omega \setminus \Sigma, \\ \partial_\nu v = 0 & \text{on} \quad \partial\Omega, \\ v = \sigma \mathcal{A}_\Sigma \rho + \lambda_* \delta\rho & \text{on} \quad \Sigma, \\ \lambda_* l\rho - [d\partial_\nu v] = 0 & \text{on} \quad \Sigma \end{cases} \tag{5.6}$$

has a smooth (analytic) family $[s \mapsto (\lambda_*(s), v(s), \rho(s))]$ of solutions, which exists as long as $\lambda_*(s)$ remains a simple eigenvalue. As $\zeta(s) := s\sigma(\kappa|1)_\Omega/l|\Sigma|R^2$ approaches the value $\zeta = 1$ from above, we must have $\lambda_*(s) \to 0$ from the right. This means that at the value

$$s = s_0 := \frac{l|\Sigma|R^2}{\sigma(\kappa|1)_\Omega} \tag{5.7}$$

the eigenvalue $\lambda_*(s)$ passes through the origin, in accordance with the jump of the algebraic multiplicity by 1 of the eigenvalue 0 for L_δ at $\zeta = 1 = \zeta(s_0)$. This shows that there can be only one positive eigenvalue for (5.6), independently of the values of the parameters, and there is precisely one iff $\zeta > 1$.

If $\zeta = \sigma(\kappa|1)\Omega/l|\Sigma|R^2 > 1$, then we have that $s_0 < 1$. The argument given above shows that the modified eigenvalue problem (5.6) has for each $s > s_0$ exactly one simple eigenvalue. This is, in particular, true for $s = 1$, thus establishing Theorem 2.1(f).

Now we turn our attention to the case $l < 0$. As before, we conclude that the operator L_δ has countably many eigenvalues. We note that the argument given in section 4 also applies to the case $l < 0$ and $\delta = 0$, showing that all eigenvalues of (1.9) are real in this case.

In the following, we assume that $l < 0$ and $\delta > 0$. In order to show Theorem 1.1(f), we consider the operators $B_\lambda := \lambda\delta I + \lambda l T_\lambda + \sigma A_\Sigma$ for $\lambda > 0$. By Proposition 5.1 we have

$$(B_\lambda\rho|\rho)_\Sigma \geq \big(\delta\lambda - |l|M_0\lambda^{1/2} - \sigma/R^2\big)\|\rho\|_\Sigma^2 \geq \|\rho\|_\Sigma^2$$

provided that $\lambda \geq \mu_0$, for some $\mu_0 \geq \lambda_0$. Hence, B_λ is positive definite for large $\lambda > 0$. On the other hand we have

$$(B_\lambda 1|1)_\Sigma = \lambda\delta|\Sigma| - \lambda|l|(T_\lambda 1|1)_\Sigma - \sigma|\Sigma|/R^2 \leq \lambda\delta|\Sigma| - \sigma|\Sigma|/R^2.$$

Thus for λ small we see that B_λ is not positive. Let

$$\lambda_* := \inf\{\lambda > 0 : B_\mu \text{ is positive definite for all } \mu \geq \lambda\}.$$

Then B_{λ_*} is still semi-definite, but not definite, and hence, by compactness of the resolvent, has a nontrivial kernel. This shows that λ_* is an eigenvalue of (1.9), proving Theorem 1.1(f).

Remark 5.2. Suppose $l < 0$ and $\delta > 0$.

(a) While it is still true that all nontrivial eigenvalues of (1.9) are semi-simple, we cannot conclude that positive eigenvalues are simple.

(b) We do not know whether all eigenvalues of (1.9) are real if $l < 0$ and $\delta > 0$. We can, however, prove that every sector $[|\arg\lambda| \le \theta]$ can only contain finitely many eigenvalues for a fixed $\theta \in (0, \pi)$. This can be shown as follows. Let $\theta \in (\pi/2, \pi)$ be fixed, and suppose that $|\arg\lambda| \le \theta$. Moreover, let $\alpha \in (0, \pi/2)$ be an arbitrary fixed number. Then one verifies that

$$|\lambda + \mu| \ge \min\{\sin\alpha, \sin(\pi - \theta)\}|\lambda|, \quad \text{whenever } \mu \in \mathbb{R},\ \alpha \le |\arg\lambda| \le \theta,$$

and this shows that there exists a constant $c > 0$ such that

$$|\lambda\delta + \sigma(A_\Sigma\rho|\rho)_\Sigma| \ge c|\lambda|, \quad \alpha \le |\arg\lambda| \le \theta. \tag{5.8}$$

Using that $\sigma(A_\Sigma\rho|\rho) \ge -\sigma/R^2\|\rho\|_\Sigma^2$, we see that

$$|\lambda\delta + \sigma(A_\Sigma\rho|\rho)_\Sigma| \ge (\delta/2)|\lambda|, \quad \|\rho\|_\Sigma = 1, \quad \operatorname{Re}\lambda \ge 2\sigma/\delta R^2. \tag{5.9}$$

Combining (5.8)–(5.9) yields

$$|\lambda\delta(\rho|\rho)_\Sigma + \sigma(A_\Sigma\rho|\rho)_\Sigma| \ge k|\lambda|, \quad \|\rho\|_\Sigma = 1, \quad |\arg\lambda| \le \theta, \quad |\lambda| \ge \eta, \tag{5.10}$$

where $k = \min\{c, \delta/2\}$ and $\eta = (2\sigma/\delta R^2)(1/\cos\alpha)$. Let λ_0 and M_0 be as in Proposition 5.1. Suppose that $\lambda \in \mathbb{C} \setminus \{0\}$ with $|\arg\lambda| \le \theta$ is an eigenvalue of (1.9) with eigenfunction (v, ρ). Then we have

$$\lambda\delta\rho + \sigma A_\Sigma\rho = \lambda|l|T_\lambda\rho \tag{5.11}$$

and we can assume, without loss of generality, that $\|\rho\|_\Sigma = 1$. If $|\lambda| \ge \max\{\lambda_0, \eta\}$, then we conclude from (5.10)–(5.11) and Proposition 5.1 that

$$k|\lambda| \le M_0|l||\lambda|^{1/2}$$

and so $|\lambda|$ is bounded by $(M_0 l/k)^2$. Clearly, if $|\lambda| \le \max\{\lambda_0, \eta\}$, we have a trivial bound. This shows that all possible eigenvalues in the sector $[|\arg\lambda| \le \theta]$ are bounded. Since eigenvalues cannot accumulate in a bounded set, we see that (1.9) can only have finitely many eigenvalues in the sector $[|\arg\lambda| \le \theta]$.

6. Analysis of the unstable eigenvalue. In this section we analyze the properties of the unstable eigenvalue λ_* of problem (5.6) in case $l > 0$ in more detail. In particular, we study the behavior of $\lambda_*(s)$ and the corresponding eigenfunctions near the critical value 1 of $\zeta(s)$, i.e., for s near s_0; see (5.7).

Proof Theorem 1.3. (a) We will first analyze the behavior of $(\lambda_*(s), v(s), \rho(s))$ for s near s_0. In order to do so we use the following ansatz:

$$\begin{aligned}
\lambda_*(s) &= (s - s_0)\lambda_1(s),\\
v(s) &= -1 + (s - s_0)\lambda_1(s)v_1(s),\\
\rho(s) &= \rho_0 + (s - s_0)\eta + (s - s_0)\lambda_1(s)\big(\rho_1(s) + \vec{\beta}(s)\cdot\vec{y}\,\big),\\
(v_1(s)|\kappa)_\Omega = 0, &\quad (\rho_1(s)|1)_\Sigma = (\rho_1(s)|Y_j) = 0,\ 1 \le j \le n,
\end{aligned} \tag{6.1}$$

with

$$\rho_0 := R^2/\sigma + \vec{\alpha}\cdot\vec{y}\,, \qquad \eta := (\kappa|1)_\Omega/l\Sigma\,, \tag{6.2}$$

where $\vec{\alpha}$, $\vec{\beta}(s) \in \mathbb{R}^n$ and $\vec{y} = (Y_1, \dots, Y_n)$. Setting $r = s - s_0$ and inserting this ansatz into the eigenvalue problem (5.6), we obtain the following system of equations:

$$
\left\{
\begin{aligned}
-d\Delta v_1 &= s_0\kappa + r\kappa(1 - s\lambda_1 v_1) && \text{in} \quad \Omega \setminus \Sigma,\\
\partial_\nu v_1 &= 0 && \text{on} \quad \partial\Omega,\\
[v_1] &= 0 && \text{on} \quad \Sigma,\\
-[d\partial_\nu v_1] &= -l\rho_0 - rl\eta - r\lambda_1 l(\rho_1 + \vec{\beta}\cdot\vec{y}) && \text{on} \quad \Sigma,\\
\sigma A_\Sigma \rho_1 &= \frac{\sigma\eta}{R^2\lambda_1} - \delta\rho_0 + v_1 - r\delta\eta - r\lambda_1\delta(\rho_1 + \vec{\beta}\cdot\vec{y}) && \text{on} \quad \Sigma.
\end{aligned}
\right. \tag{6.3}
$$

We first observe that due to (5.7), $(6.1)_4$, (6.2), and the fact that $(Y_i|1)_\Sigma = 0$, the compatibility condition

$$
\big(1|s_0\kappa + r\kappa(1 - s\lambda_1 v_1)\big)_\Omega - l\big(1|\rho_0 + r\eta + r\lambda_1(\rho_1 + \vec{\beta}\cdot\vec{y})\big)_\Sigma = 0 \tag{6.4}
$$

holds. It is our intention to apply the implicit function theorem to find a smooth (analytic) curve of solutions

$$
[s \mapsto (\lambda_1(s), v_1(s), \rho_1(s), \vec{\beta}(s))]
$$

of (6.3) for s near s_0. The idea is to use the $(n+1)$ orthogonality conditions

$$
(A_\Sigma\rho_1|1) = (A_\Sigma\rho_1|Y_j) = 0, \quad 1 \le j \le n,
$$

to determine the $(n+1)$ scalar functions λ_1 and β_j. In order to do so, we will first derive an expression for $\vec{\alpha}$ and $\vec{\beta}$. Taking the inner product of $(6.3)_5$ with Y_j yields

$$
0 = -\delta\,\vec{\alpha} + (v_1|\,\vec{y}\,)_\Sigma - r\delta\lambda_1\,\vec{\beta}, \tag{6.5}
$$

where $(v_1|\,\vec{y}\,)_\Sigma$ denotes the vector in $\mathbb{R}^n$ with components $(v_1|Y_j)_\Sigma$, $1 \le j \le n$. Due to Proposition 3.2 we have

$$
\begin{aligned}
(v_1|Y_j)_\Sigma &= (v_1| - [d\partial_\nu T_0 Y_j])_\Sigma = \int_\Omega d\,\mathrm{div}(v_1 T_0 Y_j)\,dx\\
&= (d\,\nabla v_1|\nabla T_0 Y_j)_\Omega = (-d\Delta v_1|T_0 Y_j)_\Omega + (-[d\partial_\nu v_1]|T_0 Y_j)_\Sigma\\
&= (s_0\kappa + r\kappa(1 - s\lambda_1 v_1)|T_0 Y_j)_\Omega - l(\rho_0 + r\eta + r\lambda_1(\rho_1 + \vec{\beta}\cdot\vec{y})|T_0 Y_j)_\Sigma.\\
&= -rs\lambda_1(\kappa v_1|T_0 Y_j)_\Omega - l(\rho_0 + r\eta + r\lambda_1(\rho_1 + \vec{\beta}\cdot\vec{y})|T_0 Y_j)_\Sigma
\end{aligned}
$$

for $j = 1, \dots, n$. Setting first $r = 0$ yields an equation for $\vec{\alpha}$, namely,

$$
0 = -\delta\,\vec{\alpha} - (lR^2/\sigma)(1|T_0\,\vec{y}\,)_\Sigma - l(T_0\,\vec{y}\,|\,\vec{y}\,)_\Sigma\,\vec{\alpha},
$$

i.e.,

$$
\vec{\alpha} = -\big(\delta I + l(T_0\,\vec{y}\,|\,\vec{y}\,)_\Sigma\big)^{-1}(lR^2/\sigma)(1|T_0\,\vec{y}\,)_\Sigma, \tag{6.6}
$$

where $(T_0\,\vec{y}\,|\,\vec{y}\,)_\Sigma$ denotes the symmetric matrix with entries $[(T_0 Y_i|Y_j)_\Sigma]_{1\le i,j\le n}$. Here we remind the reader that

$$
\langle (T_0\,\vec{y}\,|\,\vec{y}\,)_\Sigma\xi|\xi\rangle = \sum_{i,j=1}^n \xi_i\xi_j(T_0 Y_i|Y_j)_\Sigma = (T_0(\xi\cdot\vec{y})|(\xi\cdot\vec{y}))_\Sigma \ge c\|\xi\|^2 \tag{6.7}
$$

for $\xi \in \mathbb{R}^n$, where $\langle\cdot|\cdot\rangle$ denotes the Euclidean inner product on $\mathbb{R}^n$. This shows that the matrix $(T_0\,\vec{y}\,|\,\vec{y}\,)_\Sigma$ is positive definite, and hence $\delta I + l(T_0\,\vec{y}\,|\,\vec{y}\,)_\Sigma$ is invertible for any $\delta \geq 0$. Next, we obtain for $\vec{\beta}$ that

$$\vec{\beta} = -\big(\delta I + l(T_0\,\vec{y}\,|\,\vec{y}\,)_\Sigma\big)^{-1}\big\{s\big(\kappa v_1|T_0\,\vec{y}\,\big)_\Omega + l\big(\eta/\lambda_1 + \rho_1|T_0\,\vec{y}\,\big)_\Sigma\big\}. \tag{6.8}$$

Thus we have a function $\vec{\beta} = \vec{\beta}(\lambda_1, v_1, \rho_1, s)$. Finally, we obtain an equation for λ_1 by taking the inner product of $(6.3)_5$ with 1:

$$0 = \sigma\eta|\Sigma|/R^2\lambda_1 - \delta R^2|\Sigma|/\sigma + (v_1|1)_\Sigma - r\delta|\Sigma|\eta. \tag{6.9}$$

Employing the relation $(\kappa|v_1)_\Omega = 0$ and $(6.3)_{1,4}$ as well as (6.5) yields

$$\begin{aligned}
-rs\lambda_1(\kappa v_1|v_1)_\Omega &= (s_0\kappa + r\kappa(1 - s\lambda_1 v_1)|v_1)_\Omega \\
&= (-d\Delta v_1|v_1)_\Omega = \|\sqrt{d}\,\nabla v_1\|_\Omega^2 + ([d\partial_\nu v_1]|v_1)_\Sigma \\
&= \|\sqrt{d}\,\nabla v_1\|_\Omega^2 + (l\rho_0 + rl\eta + r\lambda_1 l(\rho_1 + \vec{\beta}\cdot\vec{y}\,)|v_1)_\Sigma \\
&= \|\sqrt{d}\,\nabla v_1\|_\Omega^2 + \big\{lR^2/\sigma + rl\eta\big\}(v_1|1)_\Sigma + l\delta\big|\,\vec{\alpha} + r\lambda_1\,\vec{\beta}\big|^2 + r\lambda_1 l(\rho_1|v_1)_\Sigma.
\end{aligned}$$

This leads to

$$\begin{aligned}
0 = \|\sqrt{d}\,\nabla v_1\|_\Omega^2 &- \big\{lR^2/\sigma + rl\eta\big\}\big\{(\sigma|\Sigma|\eta/R^2\lambda_1) - \delta R^2|\Sigma|/\sigma - r\delta|\Sigma|\eta\big\} \\
&+ l\delta\big|\,\vec{\alpha} + r\lambda_1\,\vec{\beta}\big|^2 + r\lambda_1\big\{l(\rho_1|v_1)_\Sigma + s(\kappa v_1|v_1)_\Omega\big\},
\end{aligned} \tag{6.10}$$

where we used (6.9).

Suppose now that v_1 solves the first four equations of (6.3). Then one easily verifies that (6.5) is equivalent to (6.6) and (6.8). Moreover, assuming once again that v_1 satisfies the first four equations of (6.3), and that $\vec{\alpha}$ and $\vec{\beta}$ satisfy (6.6) and (6.8), one verifies that

$$(6.9) \iff (6.10). \tag{6.11}$$

For $r = 0$, that is, for $s = s_0$, we obtain from (6.10)

$$\lambda_1(s_0) = l\Sigma\eta/\big\{\|\sqrt{d}\,\nabla v_1(s_0)\|_\Omega^2 + l\delta\big(|\,\vec{\alpha}|^2 + R^4|\Sigma|/\sigma^2\big)\big\}, \tag{6.12}$$

where $v_1(s_0)$ is the unique solution of problem (3.4) with $(f, g) = (s_0\kappa, -l\rho_0)$ and $(\kappa|v_1(s_0))_\Omega = 0$; see Proposition 3.2. This shows that $\lambda_1(s_0)$ is uniquely defined and strictly positive. Moreover, we also know from (6.11) that

$$0 = (\sigma\eta|\Sigma|/R^2\lambda_1(s_0)) - \delta R^2|\Sigma|/\sigma + (v_1(s_0)|1)_\Sigma - r\delta|\Sigma|\eta.$$

We obtain $\rho_1(s_0)$ by solving

$$\sigma A_\Sigma\rho_1 = \sigma\eta/R^2\lambda_1(s_0) - \delta\rho_0 + v_1(s_0)$$

for ρ_1, which is possible since we chose $\lambda_1(s_0)$ and $\vec{\alpha}$ in such a way that the necessary orthogonality conditions of Proposition 3.1(d) hold. Equation (6.8) shows that the mapping

$$\begin{aligned}
&I \times \Big\{v \in W_2^2(\Omega\setminus\Sigma) : [v] = 0 \text{ on } \Sigma\Big\} \times W_2^2(\Sigma) \times \mathbb{R} \to \mathbb{R}^n, \\
&[(\lambda_1, v_1, \rho_1, s) \mapsto \vec{\beta}(\lambda_1, v_1, \rho_1, s)]
\end{aligned} \tag{6.13}$$

is analytic where $I \subset \mathbb{R}$ is an open interval that contains $\lambda_1(s_0)$ but does not contain 0.

We are now in a position to apply the implicit function theorem at the point $(v_1(s_0), \rho_1(s_0), s_0)$ to solve the first four equations in (6.3) and (6.9) for (λ_1, v_1) in terms of (ρ_1, s). We choose the functional analytic setting

$$\begin{aligned}
X_1 &:= \big\{v \in W_2^2(\Omega \setminus \Sigma) :\ \partial_\nu v = 0 \text{ on } \partial\Omega,\ [v] = 0 \text{ on } \Sigma,\ (\kappa|v)_\Omega = 0\big\},\\
X_2 &:= \big\{\rho \in W_2^2(\Sigma) :\ (\rho|1)_\Sigma = (\rho|Y_j)_\Sigma = 0,\ 1 \le j \le n\big\},\\
X &:= \mathbb{R} \times X_1 \times X_2 \times \mathbb{R},\\
Y &:= \mathbb{R} \times \big\{(f,g) \in L_2(\Omega) \times W_2^{1/2}(\Sigma) :\ (f|1)_\Omega + (g|1)_\Sigma = 0\big\}
\end{aligned}$$

and we define $F : V \subset X \to Y$ by means of

$$F(\lambda_1, v_1, \rho_1, s) := \begin{pmatrix} \sigma\eta|\Sigma|/R^2\lambda_1 - \delta R^2|\Sigma|/\sigma + (v_1|1)_\Sigma - r\delta|\Sigma|\eta \\ -\,d\Delta v_1 - s_0\kappa - r\kappa(1 - s\lambda_1 v_1) \\ -\,[d\partial_\nu v_1] + l(\rho_0 + r\eta + r\lambda_1(\rho_1 + \vec{\beta}\cdot\vec{y}\,)) \end{pmatrix},$$

where $V := I \times X_1 \times X_2 \times \mathbb{R}$.

Equation (6.4) implies that F maps V into Y, and (6.13) and the definition of F show that

$$[(\lambda_1, v_1, \rho_1, s) \mapsto F(\lambda_1, v_1, \rho_1, s)] \in C^\omega(V, Y).$$

Clearly, the first four equations of (6.3) together with (6.9) are equivalent to $F(\lambda_1, v_1, \rho_1, s) = (0,0,0)$. Since we already know that

$$F(\lambda_1(s_0), v_1(s_0), \rho_1(s_0), s_0) = (0,0,0),$$

we are left with verifying that the derivative of F at the point $(v_1(s_0), \rho_1(s_0), s_0)$ w.r.t. (ρ_1, v_1) is an isomorphism, i.e.,

$$\mathbb{D}_1 F(\lambda_1(s_0), v_1(s_0), \rho_1(s_0), s_0) \in \mathrm{Isom}(\mathbb{R} \times X_1, Y).$$

It follows from Proposition 3.2(a) that the problem

$$\mathbb{D}_1 F(\lambda_1(s_0), v_1(s_0), \rho_1(s_0), s_0)(\lambda, w) = (\mu, f, g)$$

has for each $(\mu, f, g) \in Y$ a unique solution $(\lambda, w) \in \mathbb{R} \times X_1$, namely,

$$w = R_0(f,g), \quad \lambda = \frac{R^2\lambda_1^2(s_0)}{\sigma\eta|\Sigma|}\big((R_0(f,g)|1)_\Sigma - \mu\big),$$

where $w = R_0(f,g)$ is the unique solution of (3.4) with $(\kappa|w)_\Omega = 0$.

The implicit function theorem then yields a neighborhood U of $(\rho_1(s_0), s_0)$ in $X_2 \times \mathbb{R}$ such that

$$\begin{aligned}
&[(\rho_1, s) \mapsto (\lambda_1(\rho_1, s), v_1(\rho_1, s))] \in C^\omega(U, \mathbb{R} \times X_1)\\
&F(\lambda_1(\rho_1, s) v_1(\rho_1, s), \rho_1, s) = (0,0,0), \quad (\rho_1, s) \in U.
\end{aligned} \tag{6.14}$$

Combining all of the above results, we conclude that

$$[(\rho_1, s) \mapsto (\lambda_1(\rho_1, s), v_1(\rho_1, s), \vec{\beta}(\rho_1, s))] \in C^\omega(U, \mathbb{R} \times X_1 \times \mathbb{R}^n) \tag{6.15}$$

and that the functions $(\lambda_1(\rho_1,s), v_1(\rho_1,s), \vec{\alpha}, \vec{\beta}(\rho_1,s))$ satisfy the first four equations of (6.3) as well as (6.5) and (6.9). We now insert these functions into the equation for ρ_1 which gives an equation of the form

$$G(\rho_1,s) := \sigma A_\Sigma \rho_1 - \sigma\eta/R^2\lambda_1(s_0) + \delta\rho_0 - v_1(s_0) - (s-s_0)R(\rho_1,s) = 0,$$

where $[(\rho_1,s) \mapsto R(\rho_1,s)] \in C^\omega(U, L_2(\Sigma))$. By (6.5) and (6.9) we know that

$$G : X_2 \times \mathbb{R} \to Z := \big\{ g \in L_2(\Sigma) : \ (g|1)_\Sigma = (g|Y_j) = 0, \ 1 \le j \le n \big\}.$$

Moreover, we also know that $G(\rho_1(s_0), s_0) = 0$. The derivative of G with respect to ρ_1 at $(\rho_1(s_0), s_0)$ is σA_Σ, and Proposition 3.1(d) and the implicit function theorem then yield an analytic curve

$$[s \mapsto \rho_1(s)] \in C^\omega((s_0-\varepsilon_0, s_0+\varepsilon_0), X_2)$$

such that $G(\rho_1(s), s) = 0$.

Combining all of the results, we obtain an analytic curve of solutions

$$[s \mapsto (\lambda_*(s), v(s), \rho(s))]$$

of (5.6) for $s \in (s_0-\varepsilon_0, s_0+\varepsilon_0)$. If $s > s_0$, then the statement in (a) follows from the considerations in section 5.

(b) The proof of part (a) shows that the eigenvalue curve $[s \mapsto \lambda_*(s)]$ is analytic near the critical value $s = s_0$ and crosses the imaginary axis at $s = s_0$ with positive speed $\lambda_1(s_0)$; see (6.12).

(c) We show that $\lambda_*(s)$ is strictly increasing. To see this, we differentiate (5.6) w.r.t. s and form the inner product of the resulting equation in Ω with $v = v(s)$. This yields with Green's formula

$$\begin{aligned}
-(s\lambda_*(s))'\|\sqrt{\kappa}v\|_\Omega^2 &= s\lambda_*(\kappa v'|v)_\Omega - (d\Delta v'|v)_\Omega \\
&= (v'|s\lambda_*\kappa v - d\Delta v)_\Omega + ([d\partial_\nu v']|v)_\Sigma - (v'|[d\partial_\nu v])_\Sigma \\
&= (\delta/l)([d\partial_\nu v']|[d\partial_\nu v])_\Sigma + (\lambda_* l\rho' + \lambda_*' l\rho|\sigma A_\Sigma\rho)_\Sigma \\
&\quad - (\delta/l)([d\partial_\nu v]|[d\partial_\nu v'])_\Sigma - (\lambda_* l\rho|\sigma A_\Sigma\rho')_\Sigma \\
&= \lambda_*' l\sigma(A_\Sigma\rho|\rho)_\Sigma;
\end{aligned}$$

hence

$$\lambda_*\|\sqrt{\kappa}v\|_\Omega^2 + \lambda_*'\big\{\|\sqrt{s\kappa}\, v\|_\Omega^2 + l\sigma(A_\Sigma\rho|\rho)_\Sigma\big\} = 0.$$

Employing (4.1) once more (with κ replaced by $s\kappa$), we obtain

$$\lambda_*'(s) = \lambda_*^2(s)\|\sqrt{\kappa}v\|_\Omega^2 / \big\{\|\sqrt{d}\,\nabla v\|_\Omega^2 + (\delta/l)\|[d\partial_\nu v]\|_\Sigma^2\big\},$$

which yields $\lambda_*'(s) > 0$ for $s \ne s_0$. If $s = s_0$, then we have already established in (b) that $\lambda_*'(s_0) > 0$, and this shows the assertion of Theorem 1.3(c).

(d) If the stability condition (5.1) is violated, then we can conclude from the identity (4.4), where κ is now replaced with $s\kappa$ and λ is replaced with λ_*, that

$$\Big(\lambda_*(s) l\delta|\Sigma| + \frac{l^2|\Sigma|^2}{s(\kappa|1)_\Omega} - \frac{l\sigma|\Sigma|}{R^2}\Big) \le 0,$$

which shows that $\lambda_*(s) \le \frac{\sigma}{\delta R^2}\big(1 - 1/\zeta(s)\big)$ provided that $\delta > 0$, i.e., if kinetic undercooling is present.

(e) To show that $\lambda_*(s) \to \infty$ as $s \to \infty$ in case $\delta = 0$, we employ the estimate in Proposition 5.1(b). We first observe that due to the fact that $\lambda_*(s)$ is increasing in s, there exists a number $s_1 > s_0$ such that $s\lambda_*(s) \ge \lambda_0$ for $s \ge s_1$, where λ_0 is the number occurring in Proposition 5.1(b). It then follows from the relation $\lambda_* T_{s\lambda_*}\rho + \sigma A_\Sigma \rho = 0$ that

$$\begin{aligned}\sigma^2\|A_\Sigma\rho\|_\Sigma^2 = \sigma^2\|A_\Sigma\rho_0\|_\Sigma^2 + (\sigma^2|\Sigma|/R^2)\bar\rho^2 &= \lambda_*^2\|T_{s\lambda_*}\rho\|_\Sigma^2 \\ &\le M_1\|\rho\|_\Sigma^2\,\lambda_*(s)/s, \quad s \ge s_1,\end{aligned} \tag{6.16}$$

where we write $\rho = \rho_0 + \bar\rho$ with $(\rho_0|1)_\Sigma = 0$. Multiplying the eigenvalue problem (5.6) with $(s\lambda_* v - (d/\kappa)\Delta v)$ and using the divergence theorem and (6.16), we get

$$\begin{aligned}(s\lambda_*)^2\,\|\sqrt{\kappa}v\|_\Omega^2 + 2s\lambda_*\,\|\sqrt{d}\nabla v\|_\Omega^2 + \|(d/\sqrt{\kappa})\Delta v\|_\Omega^2 + 2s\lambda_*^2 l\sigma(\rho_0|A_\Sigma\rho_0)_\Sigma \\ = \big(2s\lambda_*^2 l\sigma|\Sigma|/R^2\big)\bar\rho^2 \le M_2\|\rho\|_\Sigma^2\,\lambda_*^3, \quad s \ge s_1.\end{aligned}$$

The relation $\lambda_* l\rho = [d\partial_\nu v]$ and the inequality above yield

$$\begin{aligned}\lambda_*^2\|\rho\|_\Sigma^2 = (1/l)^2\|[d\partial_\nu v]\|_\Sigma^2 &\le C\|v\|^2_{W_2^{3/2+\varepsilon}(\Omega\setminus\Sigma)} \le C\|v\|_\Omega^{(1/2-\varepsilon)}\|v\|^{3/2+\varepsilon}_{W_2^2(\Omega\setminus\Sigma)} \\ &\le C\Big\{\|v\|_\Omega^2 + \|v\|_\Omega^{(1/2-\varepsilon)}\|\Delta v\|_\Omega^{3/2+\varepsilon}\Big\} \le C\|\rho\|_\Sigma^2\,\lambda_*^3\left\{\frac{1}{(s\lambda_*)^2} + \frac{1}{(s\lambda_*)^{1/2-\varepsilon}}\right\} \\ &\le C\|\rho\|_\Sigma^2\,\lambda_*^3/(s\lambda_*)^{1/2-\varepsilon} = C\|\rho\|_\Sigma^2\,\lambda_*^{5/2+\varepsilon}/s^{1/2-\varepsilon}\end{aligned}$$

for $s \ge s_1$, where C is a generic constant that may change from line to line. Dividing by λ_*^2 and by $\|\rho\|_\Sigma^2$ implies

$$\lambda_*(s) \ge cs^{(1-2\varepsilon)/(1+2\varepsilon)}, \quad s \ge s_1.$$

Thus we can conclude that $\liminf_{s\to\infty}\lambda_*(s)/s^\theta = \infty$ for each $\theta < 1$. ∎

Acknowledgments. The authors would like to thank Rico Zacher for helpful discussions. We thank the anonymous referees for carefully reading the manuscript.

REFERENCES

[1] B. V. Bazalii, *Stefan problem for the Laplace equation with regard to the curvature of the free boundary*, Ukrainian Math. J., 49 (1997), pp. 1465–1484.
[2] G. Caginalp, *An analysis of a phase field model of a free boundary*, Arch. Rational Mech. Anal., 92 (1986), pp. 205–245.
[3] B. Chalmers, *Principles of Solidification*, Krieger, Huntington, NY, 1977.
[4] X. Chen, *The Hele-Shaw problem and area-preserving curve-shortening motion*, Arch. Rational Mech. Anal., 123 (1993), pp. 117–151.
[5] X. Chen and F. Reitich, *Local existence and uniqueness of solutions of the Stefan problem with surface tension and kinetic undercooling*, J. Math. Anal. Appl., 164 (1992), pp. 350–362.
[6] X. Chen, J. Hong, and F. Yi, *Existence, uniqueness, and regularity of classical solutions of the Mullins-Sekerka problem*, Comm. Partial Differential Equations, 21 (1996), pp. 1705–1727.
[7] R. Denk, J. Prüss, and R. Zacher, *Maximal L_p-regularity of Parabolic Problems with Boundary Conditions of Relaxation Type*, submitted.
[8] J. Escher and G. Simonett, *On Hele-Shaw models with surface tension*, Math. Res. Lett., 3 (1996), pp. 467–474.
[9] J. Escher and G. Simonett, *Classical solutions for the quasi-stationary Stefan problem with surface tension*, in Differential Equations, Asymptotic Analysis, and Mathematical Physics (Potsdam, 1996), Math. Res. 100, Akademie Verlag, Berlin, 1997, pp. 98–104.

[10] J. Escher and G. Simonett, *Classical solutions for Hele-Shaw models with surface tension*, Adv. Differential Equations, 2 (1997), pp. 619–642.
[11] J. Escher and G. Simonett, *A center manifold analysis for the Mullins-Sekerka model*, J. Differential Equations, 143 (1998), pp. 267–292.
[12] J. Escher, J. Prüss, and G. Simonett, *On the Stefan problem with surface tension*, in Elliptic and Parabolic Problems (Rolduc/Gaeta, 2001), World Scientific, River Edge, NJ, 2002, pp. 377–388.
[13] J. Escher, J. Prüss, and G. Simonett, *Analytic solutions for a Stefan problem with Gibbs-Thomson correction*, J. Reine Angew. Math., 563 (2003), pp. 1–52.
[14] A. Friedman and F. Reitich, *The Stefan problem with small surface tension*, Trans. Amer. Math. Soc., 328 (1991), pp. 465–515.
[15] A. Friedman and F. Reitich, *Nonlinear stability of a quasi-static Stefan problem with surface tension: A continuation approach*, Ann. Sc. Norm. Super. Pisa Cl. Sci. (4), 30 (2001), pp. 341–403.
[16] M. E. Gurtin, *On the two-phase problem with interfacial energy and entropy*, Arch. Rational Mech. Anal., 96 (1986), pp. 199–241.
[17] M. E. Gurtin, *Multiphase thermomechanics with interfacial structure*, Arch. Rational Mech. Anal., 104 (1988), pp. 195–221.
[18] E. I. Hanzawa, *Classical solutions of the Stefan problem*, Tôhoku Math. J., 33 (1981), pp. 297–335.
[19] P. Hartman, *Crystal Growth: An Introduction*, North-Holland, Amsterdam, 1973.
[20] C. Kneisel, *Über das Stefan-Problem mit Oberflächenspannung und thermischer Unterkühlung*, Ph.D. thesis, Leibniz Universität, Hannover, Germany, 2007.
[21] J. S. Langer, *Instabilities and pattern formation in crystal growth*, Rev. Mod. Phys., 52 (1980), pp. 1–28.
[22] S. Luckhaus, *Solutions for the two-dimensional Stefan problem with the Gibbs-Thomson law for melting temperature*, European J. Appl. Math., 1 (1990), pp. 101–111.
[23] A. M. Meirmanov, *The Stefan problem with surface tension in the three dimensional case with spherical symmetry: Non-existence of the classical solution*, European J. Appl. Math., 5 (1994), pp. 1–20.
[24] W. W. Mullins, *Thermodynamic equilibrium of a crystal sphere in a fluid*, J. Chem. Phys., 81 (1984), pp. 1436–1442.
[25] W. W. Mullins and R. F. Sekerka, *Stability of a planar interface during solidification of a dilute binary alloy*, J. Appl. Phys., 35 (1964), pp. 444–451.
[26] J. Prüss and G. Simonett, *Smooth Solutions for Two-phase Stefan Problems with Surface Tension*, in preparation.
[27] J. Prüss, G. Simonett, and R. Zacher, *Qualitative Behavior of Stefan Problems with Surface Tension*, in preparation.
[28] E. Radkevitch, *The Gibbs-Thompson correction and conditions for the existence of a classical solution of the modified Stefan problem*, Dokl. Akad. Nauk SSSR, 316 (1991), pp. 1311–1315; translation in Soviet Math. Dokl., 43 (1991), pp. 274–278.
[29] E. Radkevitch, *Conditions for the existence of a classical solution of a modified Stefan problem (the Gibbs-Thomson law)*, Mat. Sb., 183 (1992), pp. 77–101; translation in Russian Acad. Sci. Sb. Math., 75 (1993), pp. 221–246.
[30] H. Triebel, *Höhere Analysis*, Hochschulbücher für Mathematik, Band 76, VEB Deutscher Verlag der Wissenschaften, Berlin, 1972 (in German).
[31] A. Visintin, *Models for supercooling and superheating effects*, in Free Boundary Problems: Applications and Theory, Vol. III, Pitman Res. Notes Math. 120, Pitman, Boston, 1985, pp. 200–207.
[32] A. Visintin, *Models of Phase Transitions*, Progr. Nonlinear Differential Equations Appl. 28, Birkhäuser Boston, Boston, 1996.
[33] W. Yu, *A quasisteady Stefan problem with curvature correction and kinetic undercooling*, J. Partial Differential Equations, 9 (1996), pp. 55–70.

SIAM J. MATH. ANAL.
Vol. 40, No. 2, pp. 699–715

KHOKHLOV–ZABOLOTSKAYA–KUZNETSOV-TYPE EQUATION: NONLINEAR ACOUSTICS IN HETEROGENEOUS MEDIA*

I. KOSTIN† AND G. PANASENKO†

Abstract. The Khokhlov–Zabolotskaya–Kuznetsov model is a PDE describing the wave profile of an acoustic beam in a nonlinear medium. The paper treats the existence and uniqueness questions for this equation in the case of nonconstant coefficients. For the case of rapidly oscillating coefficients the asymptotic behavior of the solution is studied.

Key words. acoustics, nonlinear, homogenization

AMS subject classifications. 35Q35, 35B27

DOI. 10.1137/060674272

1. Introduction.

The so-called Khokhlov–Zabolotskaya–Kuznetsov (KZK) equation [8, 13] belongs to the set of nonlinear acoustics models, such as the well-known Riemann wave equation (or the nonlinear transfer equation), the Burgers equation, the Korteweg–de Vries equation, the Khokhlov–Zabolotskaya equation (see [2, 8, 10, 11, 14]), the Zakharov–Kuznetsov equation [5, 6], and the Rudenko–Sukhorukov equation [7, 9, 12]. These models are derived from the linear or nonlinear wave equation for the acoustic pressure, usually under the hypothesis of small variations of this pressure. More precisely the KZK equation has the form

$$\alpha u_{z\tau} = (f(u_\tau))_\tau + \beta u_{\tau\tau\tau} + \gamma u_\tau + \Delta_x u, \tag{1.1}$$

where $u_\tau = u_\tau(z, x, \tau)$ is the acoustic pressure, $(z, x) \in \mathbb{R} \times \mathbb{R}^d$, $d = 1, 2$, are space variables, and τ is the retarded time.

The nonlinear function f in the KZK equation is quadratic, i.e., $f(s) = \theta s^2$, although for the description of space-limited beams subject to the diffraction and self-action effects it can be taken as cubic: $f(s) = \theta s^3$ (see [13]).

In the real physical setting, f may have a more complicated shape. On the other hand, all of these models are derived under the assumption of small oscillations of the pressure, and so one can always consider f as quadratic for $|s| \le s^*$ and anything different for $|s| > s^*$ with some finite s^*. If $|u_\tau|$ is smaller than s^*, then the two models (with $f(s) = \theta s^2$ $\forall s$ and $f(s) = \theta s^2$ for $|s| \le s^*$) coincide. The advantage of this modified shape of f is that (as it will be proved below) we can get the global existence theorem as soon as f has a bounded derivative.

These arguments motivate us to consider the "KZK-type equation," that is, (1.1) with a nonlinearity f admitting a bounded derivative. Let us emphasize that this shape gives a more convenient physical description than the classical quadratic shape.

Another particularity of the model we consider in the present paper is that the coefficients are rapidly oscillating functions of z. This corresponds to the heterogeneous (stratified in the direction of the axis z) acoustic media. In this case (1.1) takes

*Received by the editors November 7, 2006; accepted for publication (in revised form) December 17, 2007; published electronically July 3, 2008.
http://www.siam.org/journals/sima/40-2/67427.html

†Laboratory of Mathematics of the University of Saint-Etienne (LaMUSE), Université Jean Monnet, 23 rue P. Michelon, 42023 Saint-Etienne, France (kostin@free.fr, grigory.panasenko@univ-st-etienne.fr).

the form

$$u_{zt} = r(z, z/\epsilon)(\phi(u_\tau))_\tau + b(z, z/\epsilon)u_{\tau\tau\tau} + g(z, z/\epsilon)u_\tau + \Delta_x u,$$

where ϵ is a small parameter representing the ratio between the microscale and the macroscale, while the coefficients r, b, g are oscillating with respect to the second variable. In particular, they may be 1-periodic functions of this variable. This feature complicates the problem, although it allows us to apply the homogenization method (see [1]) to obtain the homogenized model. Its solution is close to the one of the initial problem.

It seems that so far there was not any publication on the existence and uniqueness of the solution for the KZK (or KZK-type) equation, although the authors discovered that independently and simultaneously these questions (as well as the derivation of the KZK equation form the Navier–Stokes model) in the case of constant coefficients were studied by Bardos and Rozanova [3, 4]. They consider the KZK equation in the whole space ($x \in \mathbb{R}^d$) in the case of constant coefficients. In the present study we shall consider the varying (and even rapidly oscillating) coefficients of the KZK-type equation set for $x \in \omega$, where $\omega \subset \mathbb{R}^d$ is a bounded domain. The boundedness of f' will ensure the global existence, while in [3, 4] it is proved for small initial data only. In the case of stratified media we homogenize the KZK-type equation and prove the closeness of the solutions of the homogenized and initial models.

The main results of the present paper were announced by the authors in [15].

2. Problem setting. Let ω be a bounded open domain in $\mathbb{R}^d$ with a smooth boundary $\partial\omega$. For a real function $u = u(z, x, \tau)$ of the variables $z \in [0, Z]$, $\tau \in \mathbb{R}$, and $x \in \omega$, consider the following PDE:

$$\alpha u_{z\tau} = (f(u_\tau))_\tau + \beta u_{\tau\tau\tau} + \gamma u_\tau + \Delta u, \tag{2.1}$$

where the Laplace operator Δ (as well as ∇ below) derives with respect to the variable $x \in \omega$. The positive coefficients α, β, and γ are functions of z and x. The nonlinearity f may depend on z and x as well, but neither depends on τ.

In the mathematical setting, the variable z is considered as the evolutionary variable, while x and τ shall be called space variables. Let us impose the following boundary conditions on u:

$$u \text{ is } 2\pi\text{-periodic with respect to } \tau, \tag{2.2}$$

$$\nu \cdot \nabla u = 0 \qquad \text{on } \partial\omega, \tag{2.3}$$

where ν denotes the unit normal vector of ω. It is clear that these conditions are not sufficient to ensure the uniqueness. Indeed, if the coefficients are constant, then any function u depending only on z solves (2.1)–(2.3). This liberty is eliminated by the additional orthogonality condition

$$\int_0^{2\pi} u(z, x, \tau)\, d\tau = 0 \qquad \forall x, \forall z. \tag{2.4}$$

Finally, the evolutionary problem (2.1)–(2.4) requires an initial condition

$$u(0, x, \tau) = u_0(x, \tau). \tag{2.5}$$

To avoid heavy notation, in the estimates below we shall denote by M any large positive constant depending only on ω, f, α, β, γ, and Z. By "large" we mean that the estimate stays true with M replaced by any larger value.

Here is the list of assumptions under which the existence and uniqueness result for problem (2.1)–(2.5) will be established.

(ω) $\omega \subset \mathbb{R}^d$ is an open bounded set. Its boundary $\partial\omega$ is twice continuously differentiable.

($f1$) $f(s) = f(s,z,x)$ is continuously differentiable with respect to s, and its partial derivative $f' \equiv f_s(s,z,x)$ is uniformly bounded on $\mathbb{R} \times \mathbb{R}_+ \times \omega$.

($f2$) $f(0,z,x) = 0$ for all $z \in \mathbb{R}_+$ and $x \in \omega$. As is easy to see, this is not a constraint (otherwise, one can always replace $f(s,z,x)$ by $f(s,z,x) - f(0,z,x)$, leaving the equation intact). Together with the previous assumption this one implies that $|f(s,z,x)| \le M|s|$. Moreover, we assume that f is continuously differentiable with respect to z and its derivative f_z satisfies $|f_z(s,z,x)| \le M|s|$.

($f3$) Denote by F the primitive of f with respect to the first argument such that $F(0,z,x) = 0$. Clearly, $|F(s,z,x)| \le M|s|^2$. We shall assume that F is differentiable with respect to z and $|F_z(v,z,x)| \le M|v|^2$.

($\alpha\beta\gamma$) $\alpha = \alpha(z,x)$, $\beta = \beta(z,x)$, $\gamma = \gamma(z,x)$ are bounded functions defined on $\mathbb{R}_+ \times \omega$. The functions α and β are positive and uniformly bounded away from zero. The derivatives α_z, α_x, β_z, and β_x are assumed to be bounded functions.

These conditions are assumed to hold throughout sections 2–5.

The above list has to be completed by an assumption concerning the regularity of u_0. It will be given along with the definition of a solution.

3. Notation and preliminaries. Let us introduce some functional spaces. First consider the spaces of functions of the variables x and τ. Set $\Omega = \omega \times (0, 2\pi)$. The scalar product and the norm in $L_2(\Omega)$ are denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, respectively.

Let D_Ω be the set of infinitely smooth real functions $v = v(x,\tau)$ defined on $\bar\Omega$ and satisfying

$$\int_0^{2\pi} v(x,\tau)\, d\tau = 0 \quad \forall x \in \bar\omega \qquad \text{(orthogonality)}, \tag{3.1}$$

$$\frac{\partial^k}{\partial\tau^k} v(x,0) = \frac{\partial^k}{\partial\tau^k} v(x,2\pi) \quad \forall x \in \bar\omega, \quad \forall k = 0,1,\dots \qquad \text{(periodicity)}, \tag{3.2}$$

$$\nu(x) \cdot \nabla v(x,\tau) = 0 \quad \forall x \in \partial\omega, \quad \forall \tau \in [0,2\pi] \qquad \text{(Neumann condition)}. \tag{3.3}$$

Denote by X_0 the closure of D_Ω in the $L_2(\Omega)$-norm. Given a function $v \in X_0$, we shall denote by $\tilde v$ its (unique in X_0) primitive with respect to τ, that is,

$$\tilde v_\tau = v \qquad \text{and} \qquad \int_0^{2\pi} \tilde v\, d\tau = 0.$$

The Poincaré inequality yields

$$\|\tilde v\| \le \|v\| \le \|v_\tau\| \le \|v_{\tau\tau}\|$$

for all $v \in D_\Omega$. Note also that the mappings $v \mapsto v_\tau$ and $v \mapsto \tilde v$ are both antisymmetric, that is,

$$\langle v, v_\tau \rangle = 0 \qquad \text{and} \qquad \langle v, \tilde v \rangle = 0$$

for all v in X_0 for which these expressions make sense.

The closures of D_Ω with respect to the norms

$$\|v\|_1^2 = \|v_\tau\|^2 + \|\nabla v\|^2,$$
$$\|v\|_2^2 = \|v_{\tau\tau}\|^2 + \|\Delta \tilde v\|^2$$

are denoted by X_1 and X_2, respectively. Recall that, as well as the Laplace operator, the gradient is assumed to derive only with respect to $x \in \omega$. The continuous embeddings $X_2 \subset X_1 \subset X_0$ take place, the first one following from the estimate

$$2\|\nabla v\|^2 = 2\langle v_\tau, \Delta\tilde{v}\rangle \le \|v_\tau\|^2 + \|\Delta\tilde{v}\|^2 \le \|v\|_2^2,$$

while the second is the classical Poincaré inequality. Moreover, by the Rellich theorem, the embedding $X_1 \subset X_0$ is compact, and therefore the embedding $X_2 \subset X_0$ is also compact.

Now let us consider spaces of functions of three variables. For a given positive Z, set $Q = (0, Z) \times \omega \times (0, 2\pi)$. We shall use the triple bars to denote the $L_2(Q)$-norm:

$$|||u|||^2 = \int_0^Z \|u(z, \cdot\,, \cdot\,)\|^2\, dz.$$

Let D_Q be the set of infinitely smooth real functions $u = u(z, x, \tau)$ defined on $\bar{Q}$ such that $u(z, \cdot\,, \cdot\,) \in D_\Omega$ for all $z \in [0, Z]$. The closure of D_Q in $L_2(Q)$ is denoted U_0.

We shall seek for weak solutions of problem (2.1)–(2.5) in the space U_1, which is defined as the closure of D_Q in the norm

$$|||u|||_1^2 = |||u_{\tau\tau}|||^2 + |||\nabla u_\tau|||^2,$$

while U_2 stands for the space of strong solutions and is defined as the closure of D_Q with respect to the norm

$$|||u|||_2^2 = |||u_{z\tau}|||^2 + |||u_{\tau\tau\tau}|||^2 + |||\Delta u|||^2.$$

The continuous embedding $U_2 \subset U_1$ is proved similarly to the case of $X_2 \subset X_1$ and implies the inequality

$$|||u_{z\tau}|||^2 + |||u_{\tau\tau}|||^2 + |||\nabla u_\tau|||^2 \le 2|||u|||_2^2, \tag{3.4}$$

which we shall need in what follows.

4. Strong solutions: Definition and existence. If $u \in U_2$, then (2.1) writes down for u as an equality of functions in U_0. Note also that the functions of U_2 belong to $C([0, Z], X_0)$, so that the initial condition $u(0, \cdot\,, \cdot\,) = u_0$ can be understood as an identity in X_0. These two observations justify the following definition.

DEFINITION 4.1. *For given $u_0 \in X_2$ and $Z > 0$, a function $u \in U_2$ is called a strong solution to problem* (2.1)–(2.5) *on* $[0, Z]$ *if* (2.1) *holds almost everywhere in Q and $u(0, \cdot\,, \cdot\,) = u_0$ almost everywhere in Ω.*

From now on we shall consider the real functions of $(z, x, \tau) \in Q$ as X_0-valued functions of the variable z and write $u(z)$ instead of $u(z, x, \tau)$.

PROPOSITION 4.2. *For any $u_0 \in X_2$ and any $Z > 0$, problem* (2.1)–(2.5) *admits a strong solution u satisfying*

$$\sup_z \Big[\|u_{\tau\tau}\|^2 + \|u_z\|^2\Big] + |||u_{\tau\tau\tau}|||^2 + |||u_{z\tau}|||^2 + |||\Delta u|||^2 \le M\|u_{0\tau\tau}\|^2 + M\|\Delta\tilde{u}_0\|^2.$$

Proof. Denote by ψ_k the eigenfunctions of the following spectral problem:

$$-\psi_{\tau\tau} - \Delta\psi = \lambda\psi, \qquad \psi \in X_2.$$

The functions ψ_k form an orthogonal basis in X_0 as well as in X_1 and X_2. Let P_m stand for the orthogonal projection onto the subspace of X_0 spanned by $\psi_1, \dots, \psi_m$. The above spectral problem can be solved by the separation of variables, so that P_m commutes not only with $-\partial^2/\partial\tau^2 - \Delta$ but also with $-\partial^2/\partial\tau^2$ and $-\Delta$. The Galerkin approximations for problem (2.1)–(2.5) are defined as continuously differentiable functions $u^m : [0, Z] \to P_m X_0$ satisfying

$$P_m \alpha u^m_{z\tau} = P_m(f(u^m_\tau))_\tau + P_m \beta u^m_{\tau\tau\tau} + P_m \gamma u^m_\tau + \Delta u^m, \tag{4.1}$$

$$u^m(0) = P_m u_0, \tag{4.2}$$

which is equivalent to a Cauchy problem for a system of m ODEs.

First let us prove that the sequence u^m is bounded in U_2. Multiply (4.1) by $u^m_{\tau\tau\tau}$, and observe that $\langle \Delta u^m, u^m_{\tau\tau\tau}\rangle = \langle \nabla u^m_\tau, \nabla u^m_{\tau\tau}\rangle = 0$ to obtain

$$-\langle \alpha u^m_{z\tau\tau}, u^m_{\tau\tau}\rangle = \langle f'(u^m_\tau) u^m_{\tau\tau}, u^m_{\tau\tau\tau}\rangle + \langle \beta u^m_{\tau\tau\tau}, u^m_{\tau\tau\tau}\rangle - \langle \gamma u^m_{\tau\tau}, u^m_{\tau\tau}\rangle.$$

After some obvious transformations using assumptions $(f1)$, $(f2)$, and $(\alpha\beta\gamma)$ as well as the Cauchy–Young inequality, the latter relation becomes

$$\frac{d}{dz}\langle \alpha u^m_{\tau\tau}, u^m_{\tau\tau}\rangle + M^{-1}\|u^m_{\tau\tau\tau}\|^2 \le M\|u^m_{\tau\tau}\|^2,$$

and therefore the Gronwall lemma along with the equivalence of the norms $\langle \alpha u^m_{\tau\tau}, u^m_{\tau\tau}\rangle$ and $\|u^m_{\tau\tau}\|^2$ implies that

$$\sup_z \|u^m_{\tau\tau}\|^2 + \|\!|u^m_{\tau\tau\tau}|\!\|^2 \le M\|u^m_{\tau\tau}(0)\|^2 \le M\|u_{0\tau\tau}\|^2. \tag{4.3}$$

In particular, estimate (4.3) guarantees the existence of the functions u^m satisfying (4.1)–(4.2).

In order to obtain a bound on $\|\!|u^m_{z\tau}|\!\|$, take the primitive of (4.1) with respect to τ, and then derive the result in z. This yields the relation

$$\begin{aligned} &P_m \alpha_z u^m_z + P_m \alpha u^m_{zz} \\ &= P_m(f(u^m_\tau))_z + P_m \beta_z u^m_{\tau\tau} + P_m \beta u^m_{z\tau\tau} + P_m \gamma_z u^m + P_m \gamma u^m_z + \Delta \tilde{u}^m_z + c, \end{aligned} \tag{4.4}$$

where the integration constant c may depend on z and x but not on τ. Let us see now what happens when this identity is multiplied by u^m_z. The last two terms on the right will disappear. On the left, keep the second term and half of the first one:

$$\frac{1}{2}\langle \alpha_z u^m_z, u^m_z\rangle + \langle \alpha u^m_{zz}, u^m_z\rangle = \frac{1}{2}\frac{d}{dz}\langle \alpha u^m_z, u^m_z\rangle.$$

The other half of $\langle \alpha_z u^m_z, u^m_z\rangle$ goes to the right and is estimated by $\|u^m_z\|^2$. The third term on the right goes to the left and, after the integration by parts, admits the following lower bound:

$$-\langle \beta u^m_{z\tau\tau}, u^m_z\rangle = \langle \beta u^m_{z\tau}, u^m_{z\tau}\rangle \ge M^{-1}\|u^m_{z\tau}\|^2.$$

As for the first term on the right of (4.4), use the Cauchy–Young inequality and assumption $(f2)$ to obtain the estimate

$$\begin{aligned} \langle (f(u^m_\tau))_z, u^m_z\rangle &= \langle f'(u^m_\tau) u^m_{z\tau}, u^m_z\rangle + \langle f_z(u^m_\tau), u^m_z\rangle \\ &\le \delta\|u^m_{z\tau}\|^2 + \frac{M}{\delta}\|u^m_z\|^2 + \|u^m_z\|^2 + M\|u^m_\tau\|^2, \end{aligned}$$

which is valid for any positive δ. Quite similarly, for the second term on the right, we have

$$\langle \beta_z u^m_{\tau\tau}, u^m_z\rangle = -\langle \beta_z u^m_\tau, u^m_{z\tau}\rangle \le \delta\|u^m_{z\tau}\|^2 + \frac{M}{\delta}\|u^m_\tau\|^2.$$

The treatment of the remaining terms of (4.4) is evident. By putting the pieces together and choosing δ sufficiently small, we finally arrive at

$$\frac{d}{dz}\langle \alpha u^m_z, u^m_z\rangle + M^{-1}\|u^m_{z\tau}\|^2 \le M\|u^m_z\|^2 + M\|u^m_\tau\|^2.$$

As in the case of estimate (4.3), apply the Gronwall lemma to obtain

$$\sup_z \|u^m_z\|^2 + \|\!|u^m_{z\tau}|\!\|^2 \le M\|u^m_z(0)\|^2 + M\|\!|u^m_\tau|\!\|^2. \tag{4.5}$$

On the other hand, the multiplication of (4.1) by $-\tilde u^m_z$, after the appropriate integration by parts, yields

$$\langle \alpha u^m_z, u^m_z\rangle = \langle f(u^m_\tau), u^m_z\rangle + \langle \beta u^m_{\tau\tau}, u^m_z\rangle + \langle \gamma u^m, u^m_z\rangle + \langle \Delta\tilde u^m, u^m_z\rangle.$$

By applying the Cauchy–Young inequality in each of the terms on the right, we have

$$\|u^m_z\|^2 \le M\|u^m_{\tau\tau}\|^2 + M\|\Delta\tilde u^m\|^2,$$

which allows us to eliminate $\|u^m_z(0)\|^2$ from the right-hand side of (4.5). By taking (4.3) into account, we can rewrite estimate (4.5) as

$$\sup_z \|u^m_z\|^2 + \|\!|u^m_{z\tau}|\!\|^2 \le M\|u_{0\tau\tau}\|^2 + M\|\Delta\tilde u_0\|^2. \tag{4.6}$$

The missing bound on $\|\!|\Delta u^m|\!\|$ can be obtained directly from (4.1) with the help of estimates (4.3) and (4.6):

$$\begin{aligned}\|\!|\Delta u^m|\!\|^2 &\le M\|\!|u^m_{z\tau}|\!\|^2 + M\|\!|u^m_{\tau\tau\tau}|\!\|^2 + M\|\!|u^m_{\tau\tau}|\!\|^2 + M\|\!|u^m_\tau|\!\|^2\\ &\le M\|u_{0\tau\tau}\|^2 + M\|\Delta\tilde u_0\|^2.\end{aligned} \tag{4.7}$$

Note that, according to the convention of section 3, we use the symbol M to denote *any* constant, although it is clear that the values of M in the first and in the second line of (4.7) cannot be the same.

By putting together the estimates (4.3), (4.6), and (4.7) we obtain

$$\begin{aligned}\sup_z \left[\|u^m_{\tau\tau}\|^2 + \|u^m_z\|^2\right] + \|\!|u^m_{\tau\tau\tau}|\!\|^2 + \|\!|u^m_{z\tau}|\!\|^2 + \|\!|\Delta u^m|\!\|^2\\ \le M\|u_{0\tau\tau}\|^2 + M\|\Delta\tilde u_0\|^2.\end{aligned} \tag{4.8}$$

Thus the sequence u^m is bounded in U_2 and therefore contains a subsequence u^{m_j} such that

$$u^{m_j}_{\tau\tau\tau} \xrightarrow{U_0} u_{\tau\tau\tau}, \qquad u^{m_j}_\tau \xrightarrow{U_0} u_\tau, \qquad u^{m_j}_{z\tau} \xrightarrow{U_0} u_{z\tau}, \qquad \Delta u^{m_j} \xrightarrow{U_0} \Delta u,$$

where $u \in U_2$. In order to show the convergence in the nonlinear term, notice that by (3.4) the sequence u^m_τ is bounded in $H^1(Q)$, and hence $u^{m_j}_\tau \to u_\tau$ strongly in U_0.

By the Lipschitzness of f we also have that $f(u_\tau^{m_j}) \to f(u_\tau)$ in $L_2(Q)$. Moreover, the sequence $(f(u_\tau^{m_j}))_\tau = f'(u_\tau^{m_j})u_{\tau\tau}^{m_j}$ is bounded in U_0, and therefore

$$(f(u_\tau^{m_j}))_\tau \xrightarrow{U_0} (f(u_\tau))_\tau.$$

Thus we can pass to the U_0-weak limit in (4.1) and obtain

$$\alpha u_{z\tau} = (f(u_\tau))_\tau + \beta u_{\tau\tau\tau} + \gamma u_\tau + \Delta u$$

as an identity in U_0. The compact embedding $U_2 \subset C([0,Z],X_0)$ implies the convergence $u^{m_j}(z) \to u(z)$ in X_0 for each z. In particular, $u(0) = u_0$. Finally, pass to the limit as $m \to \infty$ in estimate (4.8) to complete the proof. □

5. Strong solutions: Further properties. The next two statements establish further properties of strong solutions and prepare the proof of the existence result for the weak solutions.

PROPOSITION 5.1. *Each strong solution u of problem* (2.1)–(2.5) *satisfies*

$$\sup_z \Big[\|u_\tau\|^2 + \|\nabla u\|^2\Big] + \|\!|u_{\tau\tau}|\!\|^2 + \|\!|\nabla u_\tau|\!\|^2 \le M\Big[\|u_{0\tau}\|^2 + \|\nabla u_0\|^2\Big]. \tag{5.1}$$

Proof. Multiply the equation

$$\alpha u_{z\tau} = (f(u_\tau))_\tau + \beta u_{\tau\tau\tau} + \gamma u_\tau + \Delta u$$

by u_τ, and integrate the result over Ω. The first and the last terms on the right will disappear. After obvious transformations in the remaining terms we obtain for almost all $z \in (0,Z)$ the inequality

$$\frac{d}{dz}\langle \alpha u_\tau, u_\tau\rangle + M^{-1}\|u_{\tau\tau}\|^2 \le M\|u_\tau\|^2. \tag{5.2}$$

The multiplication of the equation by $-u_z$ suppresses the left-hand side and thus yields

$$0 = \langle f(u_\tau), u_{z\tau}\rangle + \langle \beta u_{\tau\tau}, u_{z\tau}\rangle - \langle \gamma u_\tau, u_z\rangle + \frac{1}{2}\frac{d}{dz}\|\nabla u\|^2. \tag{5.3}$$

The first term on the right of (5.3) can be transformed as follows:

$$\langle f(u_\tau), u_{z\tau}\rangle = \frac{d}{dz}\langle F(u_\tau), 1\rangle - \langle F_z(u_\tau), 1\rangle \ge \frac{d}{dz}\langle F(u_\tau), 1\rangle - M\|u_\tau\|^2 \tag{5.4}$$

(here we have used assumption $(f3)$). In order to eliminate the term $\langle \beta u_{\tau\tau}, u_{z\tau}\rangle$ from relation (5.3), multiply the equation by $\alpha^{-1}\beta u_{\tau\tau}$. This time the second and the third terms on the right vanish, and thus we have

$$\begin{aligned} \langle \beta u_{\tau\tau}, u_{z\tau}\rangle &= \langle \alpha^{-1}\beta f'(u_\tau)u_{\tau\tau}, u_{\tau\tau}\rangle + \langle \nabla(\alpha^{-1}\beta u_\tau), \nabla u_\tau\rangle \\ &\ge -M\|u_{\tau\tau}\|^2 + M^{-1}\|\nabla u_\tau\|^2 - M\|\nabla u_\tau\|\|u_\tau\| \\ &\ge -M\|u_{\tau\tau}\|^2 + M^{-1}\|\nabla u_\tau\|^2. \end{aligned} \tag{5.5}$$

Recall that, according to the convention about the use of M, its value may change from line to line.

Similarly, to get rid of $\langle \gamma u_\tau, u_z\rangle$ in (5.3), the proper multiplier is $\alpha^{-1}\gamma u$. It follows that

$$\begin{aligned} -\langle \gamma u_\tau, u_z\rangle &= -\langle \alpha^{-1}\gamma f(u_\tau), u_\tau\rangle - \langle \nabla(\alpha^{-1}\gamma u), \nabla u\rangle \\ &\geq -M\|u_\tau\|^2 - M\|\nabla u\|^2. \end{aligned} \tag{5.6}$$

By summing up relations (5.3)–(5.6), we arrive at

$$\frac{d}{dz}\Big[\|\nabla u\|^2 + 2\langle F(u_\tau), 1\rangle\Big] + M^{-1}\|\nabla u_\tau\|^2 \leq M\|u_{\tau\tau}\|^2 + M\|\nabla u\|^2. \tag{5.7}$$

Choose a positive μ sufficiently large so that (5.2) and (5.7) imply that

$$\begin{aligned} \frac{d}{dz}\Big[\mu\langle \alpha u_\tau, u_\tau\rangle + \|\nabla u\|^2 + 2\langle F(u_\tau), 1\rangle\Big] + M^{-1}\|u_{\tau\tau}\|^2 + M^{-1}\|\nabla u_\tau\|^2 \\ \leq M\|u_\tau\|^2 + M\|\nabla u\|^2. \end{aligned} \tag{5.8}$$

Make μ even larger, if necessary, to ensure that

$$\mu\langle \alpha u_\tau, u_\tau\rangle + 2\langle F(u_\tau), 1\rangle \geq \|u_\tau\|^2.$$

Now the Gronwall lemma applied to (5.8) yields the desired estimate (5.1). □

PROPOSITION 5.2. *Let u^1 and u^2 be two strong solutions of problem* (2.1)–(2.5) *corresponding to the initial conditions $u^1(0) = u_0^1$ and $u^2(0) = u_0^2$, respectively. Then*

$$\sup_z \|u_\tau^1 - u_\tau^2\|^2 + \|\!|u_{\tau\tau}^1 - u_{\tau\tau}^2|\!\|^2 \leq M\|u_{0\tau}^1 - u_{0\tau}^2\|^2. \tag{5.9}$$

In particular, this estimate proves the uniqueness of the strong solution.

Proof. The difference $w = u^1 - u^2$ satisfies the equation

$$\alpha w_{z\tau} = (\bar f w_\tau)_\tau + \beta w_{\tau\tau\tau} + \gamma w_\tau + \Delta w, \tag{5.10}$$

where $\bar f = (f(u_\tau^1) - f(u_\tau^2))/(u_\tau^1 - u_\tau^2)$ is a bounded function. Multiply it by w_τ, and integrate the result over Ω. The last term on the right of (5.10) disappears, and we obtain

$$\begin{aligned} \frac{1}{2}\frac{d}{dz}\langle \alpha w_\tau, w_\tau\rangle + \langle \beta w_{\tau\tau}, w_{\tau\tau}\rangle &= \frac{1}{2}\langle \alpha_z w_\tau, w_\tau\rangle - \langle \bar f w_\tau, w_{\tau\tau}\rangle + \langle \gamma w_\tau, w_\tau\rangle \\ &\leq M\|w_\tau\|\|w_{\tau\tau}\| + M\|w_\tau\|^2 \\ &\leq \frac{1}{2}\langle \beta w_{\tau\tau}, w_{\tau\tau}\rangle + M\|w_\tau\|^2. \end{aligned}$$

Apply the Gronwall lemma to complete the proof. □

Remark. Inequality (5.9) is obtained from (5.10) with the help of an argument similar to the one used in the first step of the proof of Proposition 5.1 (estimate (5.2)). However, the equivalent of (5.1) does not appear to be possible for (5.10) because of the special treatment required for the nonlinear term (see (5.4)).

6. Weak solutions. We have already qualified U_1 as the space of weak solutions. In order to introduce the weak solutions we also need the space of test functions. Let $\hat Y$ be the closure of D_Q with respect to the $H^1(Q)$-norm. Denote by Y the set of all functions $\eta \in \hat Y$ such that $\eta(Z) = 0$. The latter condition makes sense because of the continuous (and even compact) embedding $\hat Y \subset C([0, Z], X_0)$.

DEFINITION 6.1. *For a given $u_0 \in X_1$, a function $u \in U_1$ is called a weak solution of problem* (2.1)–(2.5) *if the identity*

$$\begin{aligned} &-\langle \alpha(0)u_{0\tau}, \eta(0)\rangle - \int_0^Z \langle u_\tau, (\alpha\eta)_z\rangle\, dz \\ &= \int_0^Z \Big[-\langle f(u_\tau), \eta_\tau\rangle - \langle \beta u_{\tau\tau}, \eta_\tau\rangle + \langle \gamma u_\tau, \eta\rangle - \langle \nabla u, \nabla\eta\rangle \Big]\, dz \end{aligned} \tag{6.1}$$

holds for any $\eta \in Y$.

By multiplying (2.1) by η and integrating by parts where necessary, one easily shows that each strong solution is also a weak one.

PROPOSITION 6.2. *For any $u_0 \in X_1$ problem* (2.1)–(2.5) *admits a weak solution u satisfying*

$$\sup_z \Big[\|u_\tau\|^2 + \|\nabla u\|^2 \Big] + \|\!| u_{\tau\tau} |\!\|^2 + \|\!| \nabla u_\tau |\!\|^2 \le M \Big[\|u_{0\tau}\|^2 + \|\nabla u_0\|^2 \Big]. \tag{6.2}$$

Proof. Let u_0^m, $m = 1, 2, \dots$, be a sequence in X_2 such that $u_0^m \to u_0$ in X_1 as $m \to \infty$. For each m denote by u^m the strong solution corresponding to the initial condition $u^m(0) = u_0^m$. It is also a weak solution, that is,

$$\begin{aligned} &-\langle \alpha(0)u_{0\tau}^m, \eta(0)\rangle - \int_0^Z \langle u_\tau^m, (\alpha\eta)_z\rangle\, dz \\ &= \int_0^Z \Big[-\langle f(u_\tau^m), \eta_\tau\rangle - \langle \beta u_{\tau\tau}^m, \eta_\tau\rangle + \langle \gamma u_\tau^m, \eta\rangle - \langle \nabla u^m, \nabla\eta\rangle \Big]\, dz \end{aligned} \tag{6.3}$$

for any $\eta \in Y$. By Proposition 5.2,

$$\|\!| u_{\tau\tau}^m - u_{\tau\tau}^n |\!\|^2 \le M \|u_{0\tau}^m - u_{0\tau}^n\|^2 \underset{m,\, n \to \infty}{\longrightarrow} 0,$$

so that there exists an element $u \in U_0$ such that $u_{\tau\tau} \in U_0$ and $u_{\tau\tau}^m \to u_{\tau\tau}$ in U_0 as $m \to \infty$. This also implies that $u_\tau^m \to u_\tau$ in U_0 and therefore $f(u_\tau^m) \to f(u_\tau)$ in $L_2(Q)$ as well. Finally notice that, by Proposition 5.1,

$$\sup_z \Big[\|u_\tau^m\|^2 + \|\nabla u^m\|^2 \Big] + \|\!| u_{\tau\tau}^m |\!\|^2 + \|\!| \nabla u_\tau^m |\!\|^2 \le M \Big[\|u_{0\tau}^m\|^2 + \|\nabla u_0^m\|^2 \Big], \tag{6.4}$$

so that the sequence ∇u^m is bounded in U_0, and therefore ∇u^m converges to ∇u weakly in U_0. Thus we can pass to the limit in (6.3) to obtain (6.1).

In order to establish estimate (6.2), first observe that the sequence ∇u_τ^m is bounded in U_0 and therefore contains a subsequence weakly convergent to ∇u_τ in U_0. Similarly, the sequences u_τ^m and ∇u^m contain subsequences $*$-weakly convergent, respectively, to u_τ and ∇u in $L_\infty([0, Z], X_0)$. One can therefore justify the limit as $m \to \infty$ in (6.4) and thus obtain (6.2). □

PROPOSITION 6.3. *The weak solution is unique.*

Proof. Let u^1 and u^2 be two weak solutions corresponding to the same initial condition $u^1(0) = u^2(0) = u_0 \in X_1$. Their difference $w = u^1 - u^2$ satisfies the identity

$$\begin{aligned} &-\int_0^Z \langle w_\tau, (\alpha\eta)_z\rangle\, dz \\ &= \int_0^Z \Big[-\langle \bar{f} w_\tau, \eta_\tau\rangle - \langle \beta w_{\tau\tau}, \eta_\tau\rangle + \langle \gamma w_\tau, \eta\rangle - \langle \nabla w, \nabla\eta\rangle \Big]\, dz \end{aligned} \tag{6.5}$$

for any $\eta \in Y$. As well as in the proof of Proposition 5.2, here $\bar{f}$ stands for the bounded function $(f(u_\tau^1) - f(u_\tau^2))/(u_\tau^1 - u_\tau^2)$.

Formally, the estimate we are going to obtain for w is quite simple and resembles much the one of Proposition 5.2. However, the direct approach of the proof of Proposition 5.2 is not valid any more since it requires the manipulation of objects which no longer exist in the case of the weak solutions. In particular, the weak solutions have no derivatives with respect to z.

Let the functions ψ_k and the projections P_m be those defined in the proof of Proposition 4.2. Given a function $\phi \in C_0^\infty([0, Z])$, set $\eta = \alpha^{-1}\phi\tilde{\psi}_k$ in (6.5). After some obvious transformations this yields

$$\left| \int_0^Z \phi_z \langle w, \psi_k \rangle \, dz \right| \le C \int_0^Z |\phi| \, dz \qquad \forall \phi \in C_0^\infty([0, Z]),$$

with some positive C depending on w and k but independent of ϕ. By the Riesz theorem, this proves the existence of a function $\theta \in L_\infty([0, Z])$ such that

$$\int_0^Z \phi_z \langle w, \psi_k \rangle \, dz = -\int_0^Z \phi\theta \, dz \qquad \forall \phi \in C_0^\infty([0, Z]),$$

and thus the function $\langle w, \psi_k \rangle$ admits a derivative $\langle w, \psi_k \rangle_z \in L_\infty([0, Z])$ understood in the sense of distributions. Therefore, for any m, the function $w^m \equiv P_m w$ also admits a bounded derivative with respect to z.

Fix positive numbers $s < Z$, $\delta < s$, and set

$$\sigma(z) = \begin{cases} 1 & \text{if } z \le s - \delta, \\ (s - z)/\delta & \text{if } s - \delta < z \le s, \\ 0 & \text{if } z > s. \end{cases}$$

By taking into account the existence of a bounded derivative w_z^m, it is easy to see that the function $\sigma\alpha^{-1}P_m(\alpha w_\tau^m)$ belongs to the class Y. Inject it in (6.5) instead of η. The integrand at the left becomes

$$\begin{aligned} -\langle w_\tau, (\alpha\eta)_z \rangle &= -\langle w_\tau, P_m(\sigma\alpha w_\tau^m)_z \rangle = -\langle w_\tau^m, (\sigma\alpha w_\tau^m)_z \rangle \\ &= -\frac{1}{2}\sigma_z \langle \alpha w_\tau^m, w_\tau^m \rangle - \frac{\sigma}{2}\langle \alpha_z w_\tau^m, w_\tau^m \rangle - \frac{1}{2}\frac{d}{dz}\big[\sigma\langle \alpha w_\tau^m, w_\tau^m \rangle\big]. \end{aligned}$$

By integrating over $[0, Z]$ and recalling the definition of σ, one obtains for the left-hand side of (6.5) the following lower bound:

$$\begin{aligned} &-\int_0^Z \langle w_\tau, (\alpha\eta)_z \rangle \, dz \\ &\ge \frac{1}{2\delta}\int_{s-\delta}^s \langle \alpha w_\tau^m, w_\tau^m \rangle \, dz - M\int_0^s \|w_\tau^m\|^2 \, dz + \frac{1}{2}\langle \alpha(0) w_\tau^m(0), w_\tau^m(0) \rangle \\ &\ge \frac{1}{2\delta}\int_{s-\delta}^s \langle \alpha w_\tau^m, w_\tau^m \rangle \, dz - M\int_0^s \|w_\tau^m\|^2 \, dz. \end{aligned}$$

As for the right-hand side of (6.5), the first and the third terms of the integrand admit the estimate

$$-\langle \bar{f} w_\tau, \eta_\tau \rangle + \langle \gamma w_\tau, \eta \rangle \le \sigma M \|w_\tau\| \|w_{\tau\tau}^m\|,$$

while the other two terms will be kept intact for the moment. By summing up and passing to the limit as $\delta \to \infty$, we have

$$\frac{1}{2}\langle \alpha(s) w_\tau^m(s), w_\tau^m(s)\rangle \le \int_0^s \Big[M\|w_\tau\|\|w_{\tau\tau}^m\| - \langle \beta w_{\tau\tau}, \alpha^{-1} P_m(\alpha w_{\tau\tau}^m)\rangle - \langle \nabla w, \nabla \alpha^{-1} P_m(\alpha w_\tau^m)\rangle \Big]\, dz.$$

The left-hand side of this relation admits no limit as $m \to \infty$, so first let us integrate it once more with respect to s:

$$\frac{1}{2}\int_0^\xi \langle \alpha(s) w_\tau^m(s), w_\tau^m(s)\rangle\, ds \le \int_0^\xi \int_0^s \Big[M\|w_\tau\|\|w_{\tau\tau}^m\| - \langle \beta w_{\tau\tau}, \alpha^{-1} P_m(\alpha w_{\tau\tau}^m)\rangle - \langle \nabla w, \nabla \alpha^{-1} P_m(\alpha w_\tau^m)\rangle \Big]\, dz\, ds.$$

Recall that $w \in U_1$ and therefore $w^m \to w$ in U_1. Thus one can pass to the limit as $m \to \infty$ in all terms to obtain

$$\frac{1}{2}\int_0^\xi \langle \alpha(s) w_\tau(s), w_\tau(s)\rangle\, ds \le \int_0^\xi \int_0^s \Big[M\|w_\tau\|\|w_{\tau\tau}\| - \langle \beta w_{\tau\tau}, w_{\tau\tau}\rangle \Big]\, dz\, ds.$$

With the help of the Cauchy–Young inequality this can be rewritten as

$$\int_0^\xi \langle \alpha w_\tau, w_\tau\rangle\, dz \le M \int_0^\xi \int_0^s \langle \alpha w_\tau, w_\tau\rangle\, dz\, ds.$$

It remains to apply the Gronwall lemma to obtain $w = 0$. □

7. Stratified media. The present section is devoted to the study of the asymptotic behavior of the solutions to problem (2.1)–(2.5) as its coefficients rapidly oscillate. This study will be undertaken under the following additional assumptions:

($\beta\gamma$) The coefficients α, β, and γ depend only on z. It is clear that in this case the replacement of z by the new variable $\tilde{z} = \int_0^z \alpha(s)\, ds$ will bring the equation to the form with $\alpha = 1$. Thus we assume that $\alpha = 1$ throughout the section. The regularity of β and γ is the same as above; that is, β, β_z, and γ are bounded, while β is also positive and bounded away from zero.

($f4$) The nonlinear function f is of the form $f(s, z, x) = \rho(z)\phi(s)$, where ρ is bounded with a bounded derivative, while ϕ satisfies the same assumptions as f in ($f1$)–($f3$), that is, ϕ admits a bounded continuous derivative and $\phi(0) = 0$. We shall denote by Φ the primitive of ϕ satisfying $\Phi(0) = 0$.

Thus the equation we study takes the form

$$u_{z\tau} = \rho(\phi(u_\tau))_\tau + \beta u_{\tau\tau\tau} + \gamma u_\tau + \Delta u. \tag{7.1}$$

The existence and uniqueness results of sections 4–6 remain valid.

In order to introduce the oscillating character of the coefficients, let us assume that they are of the form

$$\rho(z) = r(z, z/\epsilon), \qquad \beta(z) = b(z, z/\epsilon), \qquad \gamma(z) = g(z, z/\epsilon),$$

where $\epsilon < 1$ is a small parameter. The last assumption we introduce is the following.

($\rho\beta\gamma$) The functions $r(z,\zeta)$, $b(z,\zeta)$, and $g(z,\zeta)$ are bounded as well as the derivatives r_z, r_ζ, b_z, and b_ζ. The function b is positive and bounded away from zero. Finally, the functions $r(z,z/\epsilon)$, $b(z,z/\epsilon)$, and $g(z,z/\epsilon)$ admit limits as $\epsilon \to 0$ in the following weak sense: There exists $q > 1/2$ such that for any $Z > 0$ we have

$$\int_0^Z (r(z,z/\epsilon) - \bar\rho(z))\,dz = O(\epsilon^q),$$
$$\int_0^Z (b(z,z/\epsilon) - \bar\beta(z))\,dz = O(\epsilon^q),$$
$$\int_0^Z (g(z,z/\epsilon) - \bar\gamma(z))\,dz = O(\epsilon^q),$$

where the functions $\bar\rho(z)$, $\bar\beta(z)$, and $\bar\gamma(z)$ as well as the derivatives $\bar\rho_z(z)$, and $\bar\beta_z(z)$ are bounded, while $\bar\rho(z)$ are $\bar\beta(z)$ are also positive and bounded away from zero.

Note that if r, b, and g are Lipschitzian with respect to the first argument and periodic with respect to the second one, then the above convergence conditions hold with $q = 1$ (see [1, 7, 9]).

The asymptotic study of the solutions to (7.1) as $\epsilon \to 0$ require a few estimates. The convention concerning the use of the symbol M introduced in section 2 is still valid with the following remark: the constants M are independent of ϵ.

The first estimate we need is similar to the one of Proposition 5.1.

PROPOSITION 7.1. *Let u be the strong solution of* (7.1) *corresponding to the initial condition $u(0) = u_0 \in X_2$. Then*

$$\sup_z \|\nabla u\|^2 + \frac{1}{\epsilon}\sup_z \|u_\tau\|^2 + M^{-1}\int_0^Z \|u_{\tau\tau}\|^2 dz + M^{-1}\int_0^Z \|\nabla u_\tau\|^2 dz \le M\|\nabla u_0\|^2 + \frac{M}{\epsilon}\|u_{0\tau}\|^2.$$

Proof. Multiply (7.1) by u_τ to obtain

$$\frac{1}{2}\frac{d}{dz}\|u_\tau\|^2 = -\langle \beta u_{\tau\tau}, u_{\tau\tau}\rangle + \langle \gamma u_\tau, u_\tau\rangle \le -M^{-1}\|u_{\tau\tau}\|^2 + M\|u_\tau\|^2. \tag{7.2}$$

The multiplication of (7.1) by u_z yields

$$\begin{aligned}\frac{1}{2}\frac{d}{dz}\|\nabla u\|^2 + \frac{d}{dz}\langle \rho\Phi(u_\tau), 1\rangle &= -\langle \beta u_{\tau\tau}, u_{\tau z}\rangle + \langle \rho_z\Phi(u_\tau), 1\rangle + \langle \gamma u_\tau, u_z\rangle \\ &\le -\langle \beta u_{\tau\tau}, u_{\tau z}\rangle + M\epsilon^{-1}\|u_\tau\|^2 + \langle \gamma u_\tau, u_z\rangle.\end{aligned} \tag{7.3}$$

In order to eliminate the first term on the right-hand side of (7.3), multiply (7.1) by $\beta u_{\tau\tau}$:

$$\begin{aligned}\langle \beta u_{\tau\tau}, u_{\tau z}\rangle &= \langle \rho\beta\phi'(u_\tau)u_{\tau\tau}, u_{\tau\tau}\rangle + \langle \beta\nabla u_\tau, \nabla u_\tau\rangle \\ &\ge -M\|u_{\tau\tau}\|^2 + M^{-1}\|\nabla u_\tau\|^2.\end{aligned} \tag{7.4}$$

Similarly, after the multiplication of (7.1) by γu, we have

$$\langle \gamma u_\tau, u_z\rangle \le M\|u_\tau\|^2 + M\|\nabla u\|^2. \tag{7.5}$$

By combining (7.3)–(7.5), we have

$$\begin{aligned}\frac{1}{2}\frac{d}{dz}\|\nabla u\|^2 + \frac{d}{dz}\langle\rho\Phi(u_\tau),1\rangle + M^{-1}\|\nabla u_\tau\|^2 \\ \le M\|u_{\tau\tau}\|^2 + M\|\nabla u\|^2 + M\epsilon^{-1}\|u_\tau\|^2.\end{aligned} \tag{7.6}$$

For a given positive μ, set

$$E = \frac{1}{2}\|\nabla u\|^2 + \langle\rho\Phi(u_\tau),1\rangle + \frac{\mu}{\epsilon}\|u_\tau\|^2.$$

If the parameter μ is sufficiently large, then

$$E \ge \frac{1}{2}\|\nabla u\|^2 + \frac{\mu}{2\epsilon}\|u_\tau\|^2.$$

Fix μ even larger, if necessary, so that the inequalities (7.2) and (7.6) imply that

$$\frac{dE}{dz} + M^{-1}\|u_{\tau\tau}\|^2 + M^{-1}\|\nabla u_\tau\|^2 \le ME.$$

It remains to apply the Gronwall lemma to obtain the claimed estimate. ∎

The next estimate requires some new notation. The operator $I - \Delta : L_2(\omega) \to L_2(\omega)$ defined on $H^2(\omega)$ with the Neumann boundary conditions is self-adjoint and positive definite. Denote it by A^2. Thus A is its square root, which is also positive definite. Clearly, we have $M^{-1}\|A^{-1}u\| \le \|u\| \le M\|Au\|$. Note also that

$$\|Au\|^2 = \langle A^2u, u\rangle = \langle u - \Delta u, u\rangle = \|u\|^2 + \|\nabla u\|^2.$$

On X_1, the norm $\|Au\|$ is equivalent to $\|\nabla u\|$.

PROPOSITION 7.2. *Let u be the strong solution of* (7.1) *corresponding to the initial condition $u(0) = u_0 \in X_2$. Then*

$$\sup_z \|A^{-1}u_{\tau\tau}\|^2 + \int_0^Z \|A^{-1}u_{z\tau}\|^2\,dz \le M\|\nabla u_0\|^2 + \frac{M}{\epsilon}\|u_{0\tau}\|^2 + \|A^{-1}u_{0\tau\tau}\|^2. \tag{7.7}$$

Proof. The multiplication of (7.1) by $A^{-2}u_{z\tau}$ yields

$$\begin{aligned}\|A^{-1}u_{z\tau}\|^2 + \beta\frac{d}{dz}\|A^{-1}u_{\tau\tau}\|^2 \\ \le M\|u_{\tau\tau}\|\|A^{-2}u_{z\tau}\| + M\|u_\tau\|\|A^{-2}u_{z\tau}\| + M\|A^{-1}\Delta u\|\|A^{-1}u_{z\tau}\| \\ \le M\|u_{\tau\tau}\|\|A^{-1}u_{z\tau}\| + M\|\nabla u\|\|A^{-1}u_{z\tau}\|,\end{aligned}$$

which implies, after the division by β, that

$$\beta^{-1}\|A^{-1}u_{z\tau}\|^2 + \frac{d}{dz}\|A^{-1}u_{\tau\tau}\|^2 \le M\|u_{\tau\tau}\|^2 + M\|\nabla u\|^2.$$

Integrate with respect to z to obtain

$$\sup_z \|A^{-1}u_{\tau\tau}\|^2 + \int_0^Z \|A^{-1}u_{z\tau}\|^2\,dz \le M\int_0^Z \left[\|u_{\tau\tau}\|^2 + \|\nabla u\|^2\right]dz + \|A^{-1}u_{0\tau\tau}\|^2.$$

Finally, make use of Proposition 7.1 to complete the proof. ∎

We can now state and prove the main result of this section.

THEOREM 7.3. *Let u be the strong solution of (7.1) corresponding to the initial condition $u(0) = u_0 \in X_2$. Denote by v the strong solution of the equation*

$$v_{z\tau} = \bar{\rho}(\phi(v_\tau))_\tau + \bar{\beta} v_{\tau\tau\tau} + \bar{\gamma} v_\tau + \Delta v \tag{7.8}$$

satisfying the same initial condition as u. Then

$$\sup_z \|u - v\|^2 \le M\epsilon^{q-1/2}\Big[\|u_{0\tau\tau}\|^2 + \|\Delta\tilde{u}_0\|^2\Big].$$

Proof. It is clear that v is subject to the same type of estimates as u but with the coefficients independent of ϵ. In particular,

$$\begin{aligned}\sup_z \|\nabla v\|^2 + \sup_z \|v_\tau\|^2 + M^{-1}\int_0^Z \|v_{\tau\tau}\|^2 dz + M^{-1}\int_0^Z \|\nabla v_\tau\|^2 dz \\ \le M\|\nabla u_0\|^2 + M\|u_{0\tau}\|^2.\end{aligned} \tag{7.9}$$

The difference $w = u - v$ satisfies the relation

$$\begin{aligned}w_{z\tau} = \rho[\phi(u_\tau) - \phi(v_\tau)]_\tau + \beta w_{\tau\tau\tau} + \gamma w_\tau + \Delta w \\ - (\bar{\rho} - \rho)(\phi(v_\tau))_\tau + (\bar{\beta} - \beta)v_{\tau\tau\tau} + (\bar{\gamma} - \gamma)v_\tau.\end{aligned} \tag{7.10}$$

Multiply it by $-\tilde{w}$ to obtain

$$\begin{aligned}\frac{1}{2}\frac{d}{dz}\|w\|^2 = \langle\rho[\phi(u_\tau) - \phi(v_\tau)], w\rangle - \langle\beta w_\tau, w_\tau\rangle + \langle\gamma w, w\rangle \\ + (\bar{\rho} - \rho)P + (\bar{\beta} - \beta)Q + (\bar{\gamma} - \gamma)R \\ \le M\|w\|^2 + (\bar{\rho} - \rho)P + (\bar{\beta} - \beta)Q + (\bar{\gamma} - \gamma)R,\end{aligned} \tag{7.11}$$

where P, Q, and R stand for the residual terms:

$$P = \langle\phi(v_\tau), u - v\rangle, \qquad Q = \langle v_{\tau\tau}, u - v\rangle, \qquad R = \langle v, u - v\rangle.$$

The integration of (7.11) over $z \in [0, s]$ yields

$$\frac{1}{2}\|w(s)\|^2 \le M\int_0^s \|w\|^2\, dz + \int_0^s \left[(\bar{\alpha} - \alpha)P + (\bar{\beta} - \beta)Q + (\bar{\gamma} - \gamma)R\right] dz.$$

By integrating by parts in the second term on the right and making use of assumption $(\rho\beta\gamma)$, we arrive at

$$\frac{1}{2}\|w(s)\|^2 \le M\int_0^s \|w\|^2\, dz + M\epsilon^q \int_0^s (|P_z| + |Q_z| + |R_z|)\, dz. \tag{7.12}$$

The expressions P_z, Q_z, and R_z need to be estimated separately. The most difficult estimate is required by the term $|\langle v_{\tau\tau}, u_z\rangle|$ appearing in $|Q_z|$. We have

$$\begin{aligned}|\langle v_{\tau\tau}, u_z\rangle| = |\langle(I - \Delta)v_\tau, A^{-2}u_{z\tau}\rangle| \\ \le |\langle v_\tau, A^{-2}u_{z\tau}\rangle| + |\langle\nabla v_\tau, \nabla A^{-2}u_{z\tau}\rangle| \\ \le \|A^{-1}v_\tau\|\|A^{-1}u_{z\tau}\| + \|\nabla v_\tau\|\|\nabla A^{-2}u_{z\tau}\|.\end{aligned} \tag{7.13}$$

On the other hand,

$$\begin{aligned}\|\nabla A^{-2}u_{z\tau}\|^2 = -\langle\Delta A^{-2}u_{z\tau}, A^{-2}u_{z\tau}\rangle = \langle(A^2 - I)A^{-2}u_{z\tau}, A^{-2}u_{z\tau}\rangle \\ = \langle A^{-1}u_{z\tau}, A^{-1}u_{z\tau}\rangle - \langle A^{-2}u_{z\tau}, A^{-2}u_{z\tau}\rangle \le \|A^{-1}u_{z\tau}\|^2.\end{aligned}$$

Thus (7.13) becomes

$$|\langle v_{\tau\tau}, u_z\rangle| \le (\|v_\tau\| + \|\nabla v_\tau\|)\|A^{-1}u_{z\tau}\|.$$

Now with the help of (7.7) and (7.9) we obtain

$$\begin{aligned}\int_0^s |\langle v_{\tau\tau}, u_z\rangle|\,dz &\le M\left[\int_0^s (\|v_\tau\|^2 + \|\nabla v_\tau\|^2)\,dz\right]^{1/2}\left[\int_0^s \|A^{-1}u_{z\tau}\|^2\,dz\right]^{1/2}\\ &\le M\Big[\|\nabla u_0\|^2 + \|u_{0\tau}\|^2\Big]^{1/2}\left[\|\nabla u_0\|^2 + \frac{1}{\epsilon}\|u_{0\tau}\|^2 + \|A^{-1}u_{0\tau\tau}\|^2\right]^{1/2}\\ &\le M\epsilon^{-1/2}\Big[\|u_{0\tau\tau}\|^2 + \|\Delta\tilde{u}_0\|^2\Big].\end{aligned} \tag{7.14}$$

The latter inequality follows from the boundedness of the operator A^{-1} and the fact that $\|\nabla u_0\|^2 = \langle \nabla u_0, \nabla u_0\rangle = -\langle \nabla\tilde{u}_0, \nabla u_{0\tau}\rangle = \langle \Delta\tilde{u}_0, u_{0\tau}\rangle \le \frac{1}{2}(\|\Delta\tilde{u}_0\|^2 + \|u_{0\tau}\|^2)$. Quite similarly to (7.14) one establishes the same upper bound on the term $|\langle v_{\tau\tau}, v_z\rangle|$, and therefore

$$\int_0^s |\langle v_{\tau\tau}, u_z - v_z\rangle|\,dz \le M\epsilon^{-1/2}\Big[\|u_{0\tau\tau}\|^2 + \|\Delta\tilde{u}_0\|^2\Big].$$

Now let us note that

$$\begin{aligned}\int_0^s |Q_z|\,dz &\le \int_0^s |\langle v_{\tau\tau}, u - v\rangle_z|\,dz\\ &\le \int_0^s |\langle v_{\tau\tau}, u_z - v_z\rangle|\,dz + \int_0^s |\langle v_z, u_{\tau\tau}\rangle|\,dz + \int_0^s |\langle v_z, v_{\tau\tau}\rangle|\,dz.\end{aligned} \tag{7.15}$$

The term involving u_z having already been estimated, let us treat the second term on the right of (7.15). Recall that v is a strong solution to (7.8) whose coefficients are independent of ϵ. Therefore, by Proposition 4.2,

$$\sup_z \Big[\|v_{\tau\tau}\|^2 + \|v_z\|^2\Big] + \|\!|v_{\tau\tau\tau}|\!\|^2 + \|\!|v_{z\tau}|\!\|^2 + \|\!|\Delta v|\!\|^2 \le M\|u_{0\tau\tau}\|^2 + M\|\Delta\tilde{u}_0\|^2. \tag{7.16}$$

Along with Proposition 7.1 the latter inequality implies that

$$\begin{aligned}\int_0^s |\langle v_z, u_{\tau\tau}\rangle|\,dz &\le \left[\int_0^s \|v_z\|^2 dz\right]^{1/2}\left[\int_0^s \|u_{\tau\tau}\|^2 dz\right]^{1/2}\\ &\le M\big[\|u_{0\tau\tau}\|^2 + \|\Delta\tilde{u}_0\|^2\big]^{1/2} \times \epsilon^{-1/2}\big[\|\nabla u_0\|^2 + \|u_{0\tau}\|^2\big]^{1/2}.\end{aligned}$$

As has been mentioned above, $\|\nabla u_0\|^2 \le \frac{1}{2}(\|\Delta\tilde{u}_0\|^2 + \|u_{0\tau}\|^2)$ and $\|u_{0\tau}\|^2 \le \|u_{0\tau\tau}\|^2$. So we obtain

$$\int_0^s |\langle v_z, u_{\tau\tau}\rangle|\,dz \le M\epsilon^{-1/2}(\|u_{0\tau\tau}\|^2 + \|\Delta\tilde{u}_0\|^2).$$

The last term on the right of (7.15) admits the same estimate, and we finally obtain

$$\int_0^s |Q_z|\,dz \le M\epsilon^{-1/2}(\|u_{0\tau\tau}\|^2 + \|\Delta\tilde{u}_0\|^2).$$

The estimate on $|P_z|$ is simpler. We have

$$P_z = \langle \phi'(v_\tau) v_{z\tau}, u - v \rangle + \langle \phi(v_\tau), u_z - v_z \rangle.$$

Estimate (7.16) together with Proposition 7.1 yields

$$\int_0^s |P_z|\, dz \le M\epsilon^{-1/2} \Big[\|u_{0\tau\tau}\|^2 + \|\Delta \tilde{u}_0\|^2 \Big] + \int_0^s |\langle \phi(v_\tau), u_z \rangle|\, dz. \tag{7.17}$$

The last term on the right should be treated with the help of Proposition 7.2. First note that

$$\begin{aligned} |\langle \phi(v_\tau), u_z \rangle| &= |\langle A\phi(v_\tau), A^{-1} u_z \rangle| \le \|A\phi(v_\tau)\| \|A^{-1} u_z\| \\ &\le M \|\nabla \phi(v_\tau)\| \|A^{-1} u_z\| = M \|\phi'(v_\tau) \nabla v_\tau\| \|A^{-1} u_z\| \\ &\le M \|\nabla v_\tau\| \|A^{-1} u_z\| \le M \Big[\|\Delta v\|^2 + \|v_{\tau\tau}\|^2 \Big]^{1/2} \|A^{-1} u_z\|. \end{aligned}$$

Now the integration over $z \in [0, s]$ and the use of Proposition 7.2 along with estimate (7.16) yield

$$\int_0^s |\langle \phi(v_\tau), u_z \rangle|\, dz \le M\epsilon^{-1/2} \Big[\|u_{0\tau\tau}\|^2 + \|\Delta \tilde{u}_0\|^2 \Big].$$

We omit the details concerning the treatment of the term involving R_z in the second integral on the right of (7.12). It is quite similar to the case of Q_z since $v = \tilde{\tilde{v}}_{\tau\tau}$. Thus we finally obtain

$$\frac{1}{2} \|w(s)\|^2 \le M \int_0^s \|w\|^2\, dz + M\epsilon^{q-1/2} \Big[\|u_{0\tau\tau}\|^2 + \|\Delta \tilde{u}_0\|^2 \Big]. \tag{7.18}$$

By the Gronwall lemma we have

$$\int_0^s \|w\|^2\, dz \le M\epsilon^{q-1/2} \Big[\|u_{0\tau\tau}\|^2 + \|\Delta \tilde{u}_0\|^2 \Big].$$

Reinject the latter inequality into (7.18) to complete the proof. □

8. Concluding remarks. The obtained global existence and uniqueness results for (2.1) for the case of nonlinearity f with bounded derivative show that this formulation may have some mathematical advantages in comparison with the classical KZK equation (see [3, 4]), although for small variations of the acoustic pressure (which is one of the assumptions of the physical model used to derive the equation) these two models coincide.

The introduction of the varying coefficients allows us to model the acoustic beam propagation in a stratified heterogeneous media (for example, in the atmosphere), while the asymptotic analysis by means of the homogenization techniques simplifies reducing it to the case of constant coefficients, when in some situations an analytical solution is possible.

REFERENCES

[1] N. S. Bakhvalov and G. P. Panasenko, *Homogenization: Averaging Processes in Periodic Media*, Nauka, Moscow, 1984 (in Russian); Kluwer, Dordrecht, 1989, (in English).
[2] N. S. Bakhvalov, Ya. M. Zhileikin, and E. A. Zabolotskaya, *Nonlinear Theory of Sound Beams*, Nauka, Moscow (in Russian); American Institute of Physics, New York, 1987 (in English).
[3] C. Bardos and A. Rozanova, *KZK equation*, in Proceedings of the International Crimean Autumn Mathematical School-Simposium, Laspi-Batiliman, Ukraine, 2004.
[4] C. Bardos and A. Rozanova, *The Khokhlov–Zabolotskaya–Kuznetsov equation*, C. R. Acad. Sci. Paris Sér. I, 344 (2007), pp. 337–342.
[5] A. V. Faminsky, *Cauchy problem for Zakharov–Kuznetsov equation*, Differ. Equ., 31 (1995), pp. 1070–1081 (in Russian).
[6] A. V. Faminsky, *Cauchy problem for quasi-linear equations of odd order*, Math USSR Sbornik, 180 (1989), pp. 1183–1210 (in Russian).
[7] E. A. Lapshin and G. P. Panasenko, *Homogenization of the equations of high frequency nonlinear acoustics*, C. R. Acad. Sci. Paris Sér. I, 325 (1997), pp. 931–936.
[8] B. K. Novikov, O. V. Rudenko, and V. I. Timoshenko, *Nonlinear Underwater Acoustics*, American Institute of Physics, New York, 1987.
[9] E. A. Lapshin and G. P. Panasenko, *Homogenization of high frequency nonlinear acoustics equations*, Appl. Anal., 74 (2000), pp. 311–331.
[10] O. V. Rudenko, *Nonlinear sawtooth-shaped waves*, Physics–Uspekhi, 38 (1995), pp. 965–989.
[11] O. V. Rudenko and S. I. Soluyan, *Theoretical Foundations of Nonlinear Acoustics*, Plenum, New York, 1977.
[12] O. V. Rudenko, A. K. Sukhorukova, and A. P. Sukhorukov, *Equations of high frequency nonlinear acoustics of heterogeneous media*, Acoustic J., 40 (1994), pp. 290–294 (in Russian).
[13] A. P. Sarvazyan, O. V. Rudenko, S. D. Swanson, J. B. Fowlkes, and S. Y. Emelianov, *Shear wave elasticity imaging: A new ultrasonic technology of medical diagnostics*, Ultrasound in Med. and Biol., 24 (1999), pp. 1419–1435.
[14] O. A. Vasil'eva, A. A. Karabutov, E. A. Lapshin, and O. V. Rudenko, *Interaction of One-Dimensional Waves in Dispersion-Free Media*, Moscow University, Moscow, 1983, (in Russian).
[15] I. Kostin and G. Panasenko, *Khokhlov-Zabolotskaya-Kuznetsov type equation: Nonlinear acoustic in heterogeneous media*, C. R. Mécanique, 334 (2006), pp. 220–224.

SIAM J. MATH. ANAL.
Vol. 40, No. 2, pp. 716–737

EXISTENCE OF WEAK SOLUTIONS FOR THE UNSTEADY INTERACTION OF A VISCOUS FLUID WITH AN ELASTIC PLATE*

CÉLINE GRANDMONT†

Abstract. We consider a three-dimensional viscous incompressible fluid governed by the Navier–Stokes equations, interacting with an elastic plate located on one part of the fluid boundary. We do not neglect the deformation of the fluid domain which consequently depends on the displacement of the structure. The purpose of this work is to study the solutions of this unsteady fluid-structure interaction problem, as the coefficient modeling the viscoelasticity (resp., the rotatory inertia) of the plate tends to zero. As a consequence, we obtain the existence of at least one weak solution for the limit problem (Navier–Stokes equation coupled with a plate in flexion) as long as the structure does not touch the bottom of the fluid cavity.

Key words. fluid-structure interaction, weak solutions, plate equations

AMS subject classifications. 74F10, 35Q30, 73K70, 76D03, 35Q35, 35D05

DOI. 10.1137/070699196

1. Introduction. Many physical phenomena deal with a fluid interacting with a moving or deformable structure. These kinds of problems have a lot of important applications, for instance, in areolasticity, biomechanics, hydroelasticity, sedimentation, etc. From the mathematical point of view they have been studied extensivelly over the past few years. Here we consider a viscous incompressible three-dimensional (3D) fluid described by the Navier–Stokes equations interacting with a two-dimensional elastic plate in flexion.

One already knows that if a viscous term or the rotatory inertia are taken into account in the plate equations, there exists at least one weak solution to this problem (see [4]). From the mechanical point of view, adding a viscous term is a way to introduce dissipation in the plate model, and, from the mathematical point of view, this is a way to regularize the structure velocity. Here the dissipation coming from the fluid enables us to control the space high frequencies of the structure velocity and to pass to the limit in the coupled system as the additional viscous plate coefficient tends to zero and, thus, obtain the existence of weak solutions of the limit problem. This limit system can also be obtained as the limit of the plate-Navier–Stokes system (with a regularized initial plate velocity) as the coefficient of the rotatory inertia tends to zero. In most of the previous studies the structure velocity is quite regular because of the model or because of the presence of a regularization operator in the equations. The existing results are concerned mainly with rigid body motions [5], [10], [11], [13], [16], [17], [18], [21], [22], [24], [25] or with the motion of a structure described by a finite number of modal functions [12] or a structure with additional "viscous" terms [2], [4], [8]. Recently, a significant breakthrough has been made by Coutand and Shkoller. In [6], [7] they prove the existence, locally in time, of a unique regular solution (assuming that the data are smooth enough and satisfy suitable compatibility conditions) for the Navier–Stokes equations coupled with linearized elasticity or quasi-linear elasticity. These are the only existence results where the full 3D elasticity is

*Received by the editors August 3, 2007; accepted for publication (in revised form) March 26, 2008; published electronically July 18, 2008.
http://www.siam.org/journals/sima/40-2/69919.html
†REO project, INRIA Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex, France (celine.grandmont@inria.fr).

considered and that don't require additional viscous terms. Nevertheless, despite these new important results, the case of fluid-plate or fluid-shell interaction problems is not, as far as we know, solved. Here, by assuming that the in-plane motions can be neglected and taking advantage of the transverse-only motions of the plate and of the fact that the plate equations enable some regularity of the boundary of the fluid domain, we prove the existence of weak solutions for a fluid-plate interaction problem. Note that the same proof does not apply for general plate or shell models. Yet, even if we consider here a rather simple structure model, this is, to our knowledge, the first existence result of weak solutions in this direction. Note, moreover, that no compatibility assumptions are required and the existence result holds as long as the plate does not touch the bottom of the fluid cavity. Finally, the same results hold for a 2D viscous flow interacting with a one-dimensional membrane.

In the first section we give the equations of the fluid-structure problem for which we derive a priori energy estimates. Next, definitions and properties of the energy spaces are detailed, and the weak formulation of the problem is given. In particular, we build suitable extensions of the fluid test functions and liftings of the structure test functions. In the second section, we state our main results, the third section being devoted to the derivation of compactness properties that enable us to pass to the limit in the equations as the "viscous" plate coefficient tends to zero (section 4).

1.1. Presentation of the problem. We assume that the fluid fills a three-dimensional cavity and interacts with a thin elastic structure, located on a part of the boundary of the fluid, the other part being rigid. For the sake of simplicity, we assume that, in the reference state, the elastic part of the fluid boundary is $\omega \times \{1\}$, where ω denotes a Lipschitz domain in $\mathbb{R}^2$. In the initial state the fluid occupies the domain Ω_{η_I}:

$$\Omega_{\eta_I} = \left\{(x,y,z) \in \mathbb{R}^3,\, (x,y) \in \omega, 0 < z < 1 + \eta_I(x,y)\right\},$$

where η_I is a given initial displacement of the elastic part. The rigid part of $\partial\Omega_{\eta_I}$ is denoted by Γ_0. Note that we could also have considered the case of a fluid between two elastic plates or the case where $\omega \times \{1\}$ is a part of a smooth fluid domain boundary and obtained the same kind of results. We model the deformable part of the boundary by the classical linear plate theory for tranverse motions. We take its edge to be clamped. Note that the equations describing the tranverse motion and the in-plane motions of a plate are decoupled. Here we ignore in-plane motions. From a mechanical point of view, this assumption is justified (at least for small enough deformations) since the membrane stiffness of a plate is much larger than its flexion stiffness. We denote by $\eta_\varepsilon(t,x,y)$ the vertical displacement with respect to the rest configuration. The subscript ε underlines the dependence of the solution with respect to the parameter $\varepsilon \geq 0$, which measures the "viscosity" of the plate (or the rotatory inertia). Then the equations describing the evolution of the transversal displacement η_ε ($\eta_\varepsilon = \eta_\varepsilon(t,x,y) \in \mathbb{R}$) are

$$\left\{\begin{array}{l} \partial_{tt}\eta_\varepsilon + \Delta^2\eta_\varepsilon + \varepsilon\Delta^2\partial_t\eta_\varepsilon = g + (T_f^\varepsilon)_3 \text{ in } \omega, \\ \eta_\varepsilon = \dfrac{\partial\eta_\varepsilon}{\partial n} = 0 \text{ on } \partial\omega, \\ \eta_\varepsilon(0) = \eta_I,\ \partial_t\eta_\varepsilon(0) = \dot\eta_I, \end{array}\right. \tag{1}$$

where g denotes the given body force on the plate and T_f^ε the surface force exerted by the fluid on the structure. The definition of T_f^ε will be made precise later on. Instead

of the additional viscosity term, we could have added $-\varepsilon\Delta\partial_{tt}\eta_\varepsilon$, which models the inertia of rotation. The domain occupied by the fluid at time t is denoted by $\Omega_{\eta_\varepsilon}(t)$:

$$\Omega_{\eta_\varepsilon}(t) = \left\{(x,y,z) \in \mathbb{R}^3\ ,\ (x,y) \in \omega,\ 0 < z < 1 + \eta_\varepsilon(t,x,y)\right\}. \tag{2}$$

The classical forms of the governing equations for the fluid are

$$\begin{cases} \partial_t \mathbf{u}_\varepsilon + (\mathbf{u}_\varepsilon \cdot \nabla)\mathbf{u}_\varepsilon - \nu\Delta\mathbf{u}_\varepsilon + \nabla p_\varepsilon = \mathbf{f} & \text{in } \Omega_{\eta_\varepsilon}(t), \\ \operatorname{div} \mathbf{u}_\varepsilon = 0 & \text{in } \Omega_{\eta_\varepsilon}(t), \\ \mathbf{u}_\varepsilon(t,\cdot) = \mathbf{0} & \text{on } \Gamma_0, \\ \mathbf{u}_\varepsilon(0,\cdot) = \mathbf{u}_I & \text{in } \Omega_{\eta_I}, \end{cases} \tag{3}$$

where $\mathbf{u}_\varepsilon$ denotes the fluid velocity and p_ε the pressure field. The body force $\mathbf{f}$ and the initial velocity $\mathbf{u}_I$ are given.

Since the fluid is viscous, it adheres to the plate, and thus the velocities coincide (in a sense to be defined) at the interface. This is written, since we assume that the plate motion is vertical:

$$\mathbf{u}_\varepsilon(t,x,y,1+\eta_\varepsilon(t,x,y)) = (0,0,\partial_t\eta_\varepsilon(t,x,y))^T, \quad (x,y) \in \omega. \tag{4}$$

This condition, together with the incompressibility of the fluid, leads to

$$\int_\omega \partial_t \eta_\varepsilon = 0. \tag{5}$$

Condition (5) states that the global volume of the cavity is preserved. The surface force T_f exerted by the fluid on the plate can be defined by

$$\int_\omega T_f^\varepsilon \cdot \bar{\mathbf{v}} = \int_{\partial\Omega_{\eta_\varepsilon}(t)\setminus\Gamma_0} (-2\nu D(\mathbf{u}_\varepsilon)\cdot\mathbf{n}_t^\varepsilon + p_\varepsilon \mathbf{n}_t^\varepsilon)\cdot\mathbf{v} \qquad \forall \mathbf{v}, \tag{6}$$

where $D(\mathbf{u}_\varepsilon) = (\nabla\mathbf{u}_\varepsilon + (\nabla\mathbf{u}_\varepsilon)^T)/2$ is the strain tensor, $\mathbf{n}_t^\varepsilon$ denotes the outer unit normal at $\partial\Omega_{\eta_\varepsilon}(t)\setminus\Gamma_0$ ($\mathbf{n}_t^\varepsilon = \frac{1}{\sqrt{1+(\partial_x\eta_\varepsilon)^2+(\partial_y\eta_\varepsilon)^2}}(-\partial_x\eta_\varepsilon, -\partial_y\eta_\varepsilon, 1)^T$), and $\bar{\mathbf{v}}(t,x,y) = \mathbf{v}(t,x,y,1+\eta_\varepsilon(t,x,y))\ \forall (x,y)\in\omega$. Note here that the pressure p_ε is not defined up to a constant but is uniquely defined. Its average value ensures the global volume conservation of the fluid cavity. This average is in fact the Lagrange multiplier associated with the compatibility condition (5). Note that if Neumann boundary conditions had been imposed on Γ_0, then the plate displacement should not verify an additional "volume-preserving" constraint. As noted in [4], the third component of T_f^ε can be rewritten thanks to

$$2(D(\mathbf{u}_\varepsilon)\cdot\mathbf{n}_t^\varepsilon)_3 = (\nabla\mathbf{u}_\varepsilon\cdot\mathbf{n}_t^\varepsilon)_3.$$

This simplification comes from the fact that the displacement at the fluid-structure interface is only transverse and from the incompressibility of the fluid.

Thus $\forall \mathbf{v},$ such that $v_i(t,x,1+\eta_\varepsilon(t,x,y)) = 0,\ i = 1,2,\ (t,x,y) \in (0,T)\times\omega$, we have

$$\int_\omega T_f^\varepsilon \cdot \hat{\mathbf{v}} = \int_{\Gamma_{\eta_\varepsilon}(t)} (-\nu\nabla\mathbf{u}_\varepsilon\cdot\mathbf{n}_t^\varepsilon + p_\varepsilon\mathbf{n}_t^\varepsilon)_3\ v_3. \tag{7}$$

Note moreover that the two first components of T_f^ε correspond to the Lagrange multiplier associated to the vertical velocity constrain that the fluid has to verify at the interface.

1.2. A priori estimates. In this subsection we recall the a priori estimates satisfied by any solution, assuming that it is smooth enough. We multiply the Navier–Stokes equations by $\mathbf{u}_\varepsilon$ and integrate over $\Omega_{\eta_\varepsilon}(t)$, and we multiply the plate equations by $\partial_t\eta_\varepsilon$, integrate over ω, and add these two contributions. After integrations by parts and taking into account the coupling conditions (equality of the velocities (4) and the definition of T_f^ε (7)), we obtain the following energy equality:

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int_{\Omega_{\eta_\varepsilon}(t)}|\mathbf{u}_\varepsilon|^2+2\nu\int_{\Omega_{\eta_\varepsilon}(t)}|\nabla(\mathbf{u}_\varepsilon)|^2\\&+\frac{1}{2}\frac{d}{dt}\int_\omega(\partial_t\eta_\varepsilon)^2+\frac{1}{2}\frac{d}{dt}\int_\omega(\Delta\eta_\varepsilon)^2+\varepsilon\int_\omega(\Delta\partial_t\eta_\varepsilon)^2\\&=\int_{\Omega_{\eta_\varepsilon}(t)}\mathbf{f}\cdot\mathbf{u}_\varepsilon+\int_\omega g\,\partial_t\eta_\varepsilon.\end{aligned}\tag{8}$$

Hence, using Cauchy–Schwarz and Young's inequalities and Gronwall's lemma:

$$\begin{aligned}&\frac{1}{2}\|\mathbf{u}_\varepsilon(t,\cdot)\|^2_{L^2(\Omega_{\eta_\varepsilon}(t))}+2\nu\int_0^t\|\nabla(\mathbf{u}_\varepsilon)(s,\cdot)\|^2_{L^2(\Omega_{\eta_\varepsilon}(s))}ds\\&+\frac{1}{2}\|\partial_t\eta_\varepsilon(t,\cdot)\|^2_{L^2(\omega)}+\frac{1}{2}\|\Delta\eta_\varepsilon(t,\cdot)\|^2_{L^2(\omega)}+\varepsilon\int_0^t\|\Delta\partial_t\eta_\varepsilon(s,\cdot)\|^2_{L^2(\omega)}ds\\&\le e^t\left(\frac{1}{2}\|\mathbf{u}_I\|^2_{L^2(\Omega_{\eta_I})}+\frac{1}{2}\|\dot\eta_I\|^2_{L^2(\omega)}+\frac{1}{2}\|\Delta\eta_I\|^2_{L^2(\omega)}\right)\\&+\frac{1}{2}\int_0^t\exp(t-s)\left(\|\mathbf{f}(s,\cdot)\|^2_{L^2(\Omega_{\eta_\varepsilon}(s))}+\|g(s,\cdot)\|^2_{L^2(\omega)}\right)ds.\end{aligned}\tag{9}$$

Thus, assuming that $\mathbf{f}\in L^2(0,T;L^2(\mathbb{R}^3))$, $g\in L^2(0,T;L^2(\omega))$, $\mathbf{u}_I\in L^2(\Omega_{\eta_I})$, $\eta_I\in H_0^2(\omega)$, and $\dot\eta_I\in L^2(\omega)$,

$$\mathbf{u}_\varepsilon \text{ is bounded, uniformly in } \varepsilon, \text{ in } L^\infty(0,T;L^2(\Omega_{\eta_\varepsilon}(t))),\tag{10}$$

$$\nabla\mathbf{u}_\varepsilon \text{ is bounded, uniformly in } \varepsilon, \text{ in } L^2(0,T;L^2(\Omega_{\eta_\varepsilon}(t))),\tag{11}$$

and

$$\eta_\varepsilon \text{ is bounded, uniformly in } \varepsilon, \text{ in } W^{1,\infty}(0,T;L^2(\omega))\cap L^\infty(0,T;H_0^2.(\omega)).\tag{12}$$

Moreover, if $\varepsilon>0$,

$$\partial_t\eta_\varepsilon\in L^2(0,T;H_0^2(\omega)).$$

Consequently the spaces $L^p(0,T;L^2(\Omega_\gamma(t)))$, $L^2(0,T;H^1(\Omega_\gamma(t)))$ need to be defined for γ belonging to $W^{1,\infty}(0,T;L^2(\omega))\cap L^\infty(0,T;H_0^2(\omega))$. Note that the following continuous injection holds:

$$W^{1,\infty}(0,T;L^2(\omega))\cap L^\infty(0,T;H^2(\omega))\hookrightarrow C^{0,1-\theta}([0,T];H^{2\theta}(\omega))$$

for all $0<\theta<1$. In particular,

$$W^{1,\infty}(0,T;L^2(\omega))\cap L^\infty(0,T;H^2(\omega))\hookrightarrow C^{0,\mu}([0,T];C^{0,1/2-\mu}(\overline{\omega}))\tag{13}$$

for all $0 < \mu < 1/2$. The proof of the first injection relies on standard Hilbertian interpolation inequalities (see [20]). The other is deduced from the first one and from Sobolev injections in dimension two (see [3]). Consequently, this displacement regularity does not imply that the fluid domain boundary is Lipschitz, and we have to pay a special attention to the definitions of the functional spaces. We have also to give a sense to the equality of the velocities. Thus we are going to give some technical lemmas, definitions, and properties, most of which can be found in [4].

1.3. Preliminary definitions and properties. We now turn to the definition of some functional spaces. These definitions can be found in [4], but for the sake of completness we recall them here. Let $T > 0$ and δ belong to $C^0([0,T] \times \overline{\omega})$ such that, for some positive M and α, $M \geq 1 + \delta(t,x,y) \geq \alpha > 0$ for all $(t,x,y) \in [0,T] \times \overline{\omega}$ and such that $\delta = 0$ on $\partial\omega$. The set $\Omega_\delta(t)$ defined by

$$\Omega_\delta(t) = \left\{(x,y,z) \in \mathbb{R}^3,\ (x,y) \in \omega,\ 0 < z < 1 + \delta(t,x,y)\right\}$$

is an open subset of $\mathbb{R}^3$ for every $t \in [0,T]$ which is included in $\mathcal{C}_M = \omega \times (0,M)$. Let $\widehat{\Omega}_\delta$ be the open domain of $\mathbb{R}^4$ defined by

$$\widehat{\Omega}_\delta = \bigcup_{t \in (0,T)} \{t\} \times \Omega_\delta(t).$$

We set $\widehat{\mathcal{C}_M} = (0,T) \times \mathcal{C}_M$. One can define in a standard way the spaces $L^p(\Omega_\delta(t))$, $H^1(\Omega_\delta(t))$, $H^1_0(\Omega_\delta(t))$, for every t, and $L^p(\widehat{\Omega}_\delta)$, $H^1(\widehat{\Omega}_\delta)$, $L^p(\widehat{\mathcal{C}_M})$, and $H^1(\widehat{\mathcal{C}_M})$. The space $H^1_{0,\Gamma_0}(\Omega_\delta(t))$ will denote the subspace of $H^1_{0,\Gamma_0}(\Omega_\delta(t))$ of functions of zero trace on $\Gamma_0 = \omega \times \{0\} \cup \partial\omega \times (0,1)$. We then define:

$$L^2(0,T;H^1(\Omega_\delta(t))) = \left\{v \in L^2(\widehat{\Omega}_\delta),\ \nabla v \in L^2(\widehat{\Omega}_\delta)\right\},$$

$$L^2(0,T;H^1_0(\Omega_\delta(t))) = \overline{\mathcal{D}(\widehat{\Omega}_\delta)}^{L^2(0,T;H^1(\Omega_\delta(t)))},$$

$$\mathcal{V}_\delta = \left\{\mathbf{v} \in C^1\left(\overline{\widehat{\Omega}_\delta}\right),\ \operatorname{div}\mathbf{v} = 0,\ \mathbf{v} = \mathbf{0} \text{ on } (0,T) \times \Gamma_0\right\},$$

$$V_\delta = \overline{\mathcal{V}_\delta}^{L^2(0,T;H^1(\Omega_\delta(t)))},$$

and

$$L^\infty(0,T;L^2(\Omega_\delta(t))) = \left\{v \in L^2(\widehat{\Omega}_\delta),\ \operatorname{sup\,ess}_{t\in(0,T)}\|v\|_{L^2(\Omega_\delta(t))} < +\infty\right\}.$$

Moreover we define

$$V = \left\{\mathbf{v} \in L^2(0,T;H^1(\mathcal{C}_M)),\ \operatorname{div}\mathbf{v} = 0,\ \mathbf{v} = \mathbf{0} \text{ on } (0,T) \times (\Gamma_0 \cup \Gamma_1)\right\},$$

where $\Gamma_1 = \partial\omega \times (1,M)$.

The space V_δ can be characterized as follows:

$$V_\delta = \left\{\mathbf{v} \in L^2(0,T;H^1(\Omega_\delta(t))),\ \operatorname{div}\mathbf{v} = 0,\ \mathbf{v} = \mathbf{0} \text{ on } (0,T) \times \Gamma_0\right\}.$$

In the case of a Lipschitz or a star-shaped domain independent of time, this follows from standard arguments (see [26] or [15]). In our case it can be proved by using the fact that the domain $\Omega_\delta(t)$ is locally a subgraph. This property will be extensively used in all that follows.

Next we recall various lemmas that explain how the trace on $\partial\Omega_\delta(t) \setminus \Gamma_0$ makes sense, define extension and lifting operators, and explore some properties of the spaces defined above. We omit the proofs whenever they can be found in [4]. Note that these results take advantage of the fact that the fluid domain is a subgraph because the displacement of the interface is only transverse. Let us consider the linear mapping $\gamma_{\delta(t)}\colon\ v \mapsto v(x,y,1+\delta(t,x,y))$ defined for $v \in C^0(\overline{\Omega}_\delta(t))$.

LEMMA 1. *For every $t \in [0,T]$, the mapping $\gamma_{\delta(t)}$ from $C^1(\overline{\mathcal{C}_M})$ (resp., $C^1(\overline{\Omega}_\delta(t))$) in $C^0(\overline{\omega})$ can be extended by continuity to a linear mapping from $H^1(\mathcal{C}_M)$ (resp., $H^1(\Omega_\delta(t))$) into $L^2(\omega)$.*

COROLLARY 1. *If $v \in L^2(0,T;H^1(\Omega_\delta(t)))$, then $\gamma_{\delta(t)}(v) \in L^2(0,T;L^2(\omega))$.*

Thus, the trace of $v(x,y,1+\delta(t,x,y))$ on ω makes sense at least in $L^2(\omega)$. The following lemma makes precise the regularity of $\gamma_{\delta(t)}(v)$ when assuming moreover that δ belongs to $L^\infty(0,T;H_0^2(\omega))$. This additional regularity will play a crucial role in our asymptotic study and will enable us to control the space high frequencies of the structure velocity.

LEMMA 2. *Assuming that $\delta \in C^0([0,T];C^0(\overline{\omega})) \cap L^\infty(0,T;H_0^2(\omega))$, then, for any $v \in H^1(\Omega_\delta(t))$, $\gamma_{\delta(t)}(v) \in W^{1-1/p,p}(\omega)$, $\forall 1 < p < 2$ and for $\frac{3}{2} \le p < 2$, $\gamma_{\delta(t)}(v) \in H^{\frac{3p-2}{p}}(\omega)$, for a.e. t. Moreover, if $v \in L^2(0,T;H^1(\Omega_\delta(t)))$, then $\gamma_{\delta(t)}(v) \in L^2(0,T;W^{1-1/p,p}(\omega))$ $\forall 1 < p < 2$ and $\gamma_{\delta(t)}(v) \in L^2(0,T;H^{\frac{3p-2}{p}}(\omega))$ $\forall \frac{3}{2} \le p < 2$.*

Proof. Let us define an auxilary function by

$$w(t,x,y,s) = v(x,y,1+\delta(t,x,y)-s),\ 0 \le s \le \alpha.$$

It is clear that w belongs to $L^2(\mathcal{C}_\alpha)$ $\forall t$. Moreover

$$\nabla w(t,x,y,s) = \begin{pmatrix} \partial_x v(x,y,1+\delta(t,x,y)-s) + \partial_x\delta(t,x,y)\partial_z v(x,y,1+\delta(t,x,y)-s) \\ \partial_y v(x,y,1+\delta(t,x,y)-s) + \partial_y\delta(x,y)\partial_z v(x,y,1+\delta(t,x,y)-s) \\ -\partial_z v(x,y,1+\delta(t,x,y)-s) \end{pmatrix}.$$

Since $H^1(\omega)$ is continuously imbedded in $L^q(\omega)$ $\forall q < \infty$ and $\partial_i v(x,y,1+\delta(t,x,y)-s) \in L^2(\mathcal{C}_\alpha)$ $\forall t$, we deduce that $\nabla w \in L^p(\mathcal{C}_\alpha)$ for all $1 < p < 2$, for a.e. t. Thus $w \in W^{1,p}(\mathcal{C}_\alpha)$, and $\gamma_{\delta(t)}(v) = w|_{s=0} \in W^{1-1/p,p}(\omega)$ $\forall 1 < p < 2$ for a.e. t. Moreover, for a.e. t,

$$\|\gamma_{\delta(t)}(v)\|_{W^{1-1/p,p}(\omega)} \le C(\|\delta\|_{C^0([0,T];C^0(\overline{\omega}))\cap L^\infty(0,T;H_0^2(\omega))}, \|v\|_{H^1(\Omega_\delta(t))}).$$

Furthermore, from Sobolev injections (see [1, Thm. 7.58, p. 218]), we deduce that, for $2 > p \ge \frac{3}{2}$, $\gamma_{\delta(t)}(v) \in H^{\frac{3p-2}{p}}(\omega)$ and

$$\|\gamma_{\delta(t)}(v)\|_{H^{\frac{3p-2}{p}}(\omega)} \le C(\|\delta\|_{C^0([0,T];C^0(\overline{\omega}))\cap L^\infty(0,T;H_0^2(\omega))}, \|v\|_{H^1(\Omega_\delta(t))}). \tag{14}$$

∎

Now we are going to give a characterization of $H_0^1(\Omega_\delta(t))$ with the help of the mapping $\gamma_{\delta(t)}$. An additional assumption on the boundary displacement δ is needed: δ is assumed to belong to $C^0([0,T];H^1(\omega))$ (this is not an optimal assumption). ∎

LEMMA 3. *Assuming that $\delta \in C^0([0,T];C^0(\overline{\omega}) \cap H^1(\omega))$, then*

$$H_0^1(\Omega_\delta(t)) = \left\{ v \in H^1_{0,\Gamma_0}(\Omega_\delta(t)),\ \gamma_{\delta(t)}(v) = 0 \right\}.$$

COROLLARY 2. *If $v \in L^2(0,T;H^1_{0,\Gamma_0}(\Omega_\delta(t)))$ and $\gamma_{\delta(t)}(v) = 0$, for a.e. t, then $v \in L^2(0,T;H_0^1(\Omega_\delta(t)))$, and the converse is also true.*

We now state a lemma that enables us to extend a function $\mathbf{v} \in V_\delta$ such that $\gamma_{\delta(t)}(\mathbf{v}) = (0,0,b)^T,\ b \in L^2(0,T;H^1(\omega))$, the extension belonging to V.

LEMMA 4. *We assume that $\delta \in C^0([0,T];C^0(\overline{\omega}) \cap H_0^1(\omega))$. Let $\mathbf{v} \in V_\delta$ such that, for a.e. $t, \gamma_{\delta(t)}(\mathbf{v}) = (0,0,b)^T, b \in L^2(0,T;H_0^1(\omega))$. The function defined by*

$$\overline{\mathbf{v}} = \left| \begin{array}{l} \mathbf{v} \text{ in } \widehat{\Omega}_\delta \\ (0,0,b)^T \text{ in } \widehat{\mathcal{C}_M} \setminus \widehat{\Omega}_\delta \end{array} \right. \tag{15}$$

belongs to V, and

$$\|\overline{\mathbf{v}}\|_V \leq C(\|\mathbf{v}\|_{V_\delta} + \|b\|_{L^2(0,T;H^1(\omega))}),$$

where C depends only on M.

Remark 1. If $\mathbf{v} \in L^\infty(0,T;L^2(\Omega_\delta(t)))$ and $b \in L^\infty(0,T;L^2(\omega))$, then $\overline{\mathbf{v}} \in L^\infty(0,T;L^2(\mathcal{C}_M))$.

Next we build different lifting operators.

LEMMA 5. *For every $\phi \in H_0^1(\omega)$ there exists $w \in H^1_{0,\Gamma_0}(\Omega_\delta(t))$ such that*

$$\gamma_{\delta(t)}(w) = \phi \quad \text{and} \quad \|w\|_{H^1(\Omega_\delta(t))} \leq C_\alpha \|\phi\|_{H^1(\omega)}.$$

For every $b \in H_0^1(\omega)$ such that $\int_\omega b = 0$ there exists $\mathbf{v}$ such that $\gamma_{\delta(t)}(\mathbf{v}) = (0,0,b)^T$, $\operatorname{div}(\mathbf{v}) = 0$, and

$$\|\mathbf{v}\|_{H^1_{0,\Gamma_0}(\Omega_\delta(t))} \leq C_\alpha \|b\|_{H_0^1(\omega)}.$$

Proof. Indeed

$$w = \left| \begin{array}{l} \phi \text{ in } \Omega_\delta(t) \setminus \mathcal{C}_\alpha, \\ \mathcal{R}(\frac{z}{\alpha}\phi) \text{ in } \mathcal{C}_\alpha\ , \end{array} \right.$$

where $\mathcal{R}$ is a continuous lifting operator from $H^{1/2}(\partial\mathcal{C}_\alpha)$ into $H^1(\mathcal{C}_\alpha)$ and $\mathcal{C}_\alpha = \omega \times (0,\alpha)$ verifies the desired properties. Moreover, if we consider $b \in H_0^1(\omega)$ such that $\int_\omega b = 0$, then $\tilde{\mathcal{R}}$ is a continuous lifting operator from $H^{1/2}(\partial\mathcal{C}_\alpha)$ into $H^1(\mathcal{C}_\alpha)$ such that $\operatorname{div}(\tilde{\mathcal{R}}\mathbf{v}) = 0$,

$$\mathbf{v} = \left| \begin{array}{l} (0,0,b)^T \text{ in } \Omega_\delta(t) \setminus \mathcal{C}_\alpha, \\ \tilde{\mathcal{R}}(0,0,\frac{z}{\alpha}b)^T \text{ in } \mathcal{C}_\alpha\ , \end{array} \right. \tag{16}$$

is divergence-free and belongs to $H^1_{0,\Gamma_0}(\Omega_\delta(t))$. Furthermore, we have for a.e. t

$$\|\mathbf{v}\|_{H^1_{0,\Gamma_0}(\Omega_\delta(t))} \leq C_\alpha \|b\|_{H_0^1(\omega)}.$$

Consequently (16) defines a continuous linear lifting from $\{b \in H_0^1(\omega)$, such that (s.t.) $\int_\omega b = 0\}$ into $\{\mathbf{v} \in H^1_{0,\Gamma_0}(\Omega_\delta(t))$, s.t. $\operatorname{div}(\tilde{\mathcal{R}}\mathbf{v}) = 0\}$. ∎

Remark 2. Thanks to the previous lemma the space

$$\left\{\mathbf{v} \in V_\delta, \gamma_{\delta(t)}(\mathbf{v}) = (0,0,b)^T \text{ for a.e. } t,\ b \in L^2(0,T;H_0^1(\omega))\right\}$$

is equal to the sum of the two following spaces:

$$\overline{\left\{\mathbf{v} \in \mathcal{D}(\widehat{\Omega}_\delta),\ \operatorname{div}\mathbf{v} = 0\right\}}^{L^2(0,T;H^1(\Omega_\delta(t)))}$$

and

$$\left\{\mathbf{v},\ \mathbf{v}=\left|\begin{array}{l}(0,0,b)^T \text{ in } \Omega_\delta(t)\setminus\mathcal{C}_\alpha\\ \tilde{\mathcal{R}}(0,0,\frac{z}{\alpha}b)^T \text{ in } \mathcal{C}_\alpha\end{array}\right. \text{ for a.e. } t,\ b\in L^2(0,T;H^1_0(\omega)),\int_\omega b=0\right\}.$$

We also need to build a "lifting" operator of $(0,0,b)^T$ for any b that belongs only to $H^s(\omega), 0\le s<\frac{1}{2}$ since the structure velocity $\partial_t\eta_\varepsilon$ will be bounded, uniformly in ε, only in $L^2(0,T;H^s(\omega))\ \forall 0\le s<\frac{1}{2}$ and not in $L^2(0,T;H^1_0(\omega))$ (see Lemma 2).

LEMMA 6. *For all $b\in L^2(0,T;H^s(\omega)),\ 0\le s<\frac{1}{2}$ such that $\int_\omega b=0$, there exists a lifting operator $\mathcal{R}_\alpha$ satisfying $\gamma_{\delta(t)}(\mathcal{R}_\alpha(b))=(0,0,b)^T$ and $\operatorname{div}(\mathcal{R}_\alpha(b))=0$ and for a.e. t*

$$\|\mathcal{R}_\alpha(b)\|_{H^s(\mathcal{C}_M)}\le C\|b\|_{H^s(\omega)}\ \forall 0\le s<\frac{1}{2}. \tag{17}$$

Proof. A rather naive construction will satisfy the desired properties. Let us define $\mathcal{R}_\alpha$ by

$$\mathcal{R}_\alpha(b)=\left|\begin{array}{ll}(0,0,b)^T \text{ for } z\ge\alpha & \\ (0,0,\frac{z}{\alpha}b)^T+\mathbf{w}_\alpha \text{ in } \mathcal{C}_\alpha & \text{for a.e. } t,\end{array}\right. \tag{18}$$

where $\mathbf{w}_\alpha$ is such that $\operatorname{div}(\mathbf{w}_\alpha)=b$ and $\mathbf{w}_\alpha\in H^1_0(\mathcal{C}_\alpha)$, $\|\mathbf{w}_\alpha\|_{H^1_0(\mathcal{C}_\alpha)}\le C\|b\|_{L^2(\omega)}$, for a.e. t. Such a function exists (see, for instance, [14]). It is easy to see that $\mathcal{R}_\alpha(b)$ is divergence-free. Moreover, $\mathcal{R}_\alpha$ is linear continuous from $L^2(0,T,L^2(\omega))$ (resp., $L^2(0,T,H^1_0(\omega))$) into $L^2(0,T,L^2(\mathcal{C}_M))$ (resp., $L^2(0,T,H^1_{0,\Gamma_0}(\mathcal{C}_M))$), and thus, by interpolation (see [19]), $\mathcal{R}_\alpha$ is linear continuous from $L^2(0,T,H^s(\omega)))$ into $L^2(0,T,H^s(\mathcal{C}_M))\ \forall 0\le s<\frac{1}{2}$. Consequently, (17) holds true. ∎

Remark 3. The trace $\gamma_{\delta(t)}(\mathcal{R}_\alpha(b))$ makes sense for any $b\in L^2(\omega)$ since $\mathcal{R}_\alpha(b)$ is regular enough with respect to z.

We end this section by mentioning that Korn's and Poincaré's inequalities hold in the considered spaces.

LEMMA 7. *For all $\mathbf{u}$ and $\mathbf{v}$ in*

$$\left\{\mathbf{v}\in V_\delta,\ \exists\, b\in L^2(0,T;H^1_0(\omega)),\ \gamma_{\delta(t)}(\mathbf{v})=(0,0,b)^T \text{ for a.e. } t\right\},$$

we have

$$2\int_{\Omega_\delta(t)}D(\mathbf{u}):D(\mathbf{v})=\int_{\Omega_\delta(t)}\nabla\mathbf{u}:\nabla\mathbf{v}\ \text{ for a.e. } t, \tag{19}$$

and consequently the following Korn's "equality" holds:

$$\sqrt{2}\|D(\mathbf{u})\|_{L^2(\Omega_{\delta(t)})}=\|\nabla\mathbf{u}\|_{L^2(\Omega_{\delta(t)})} \text{ for a.e. } t. \tag{20}$$

LEMMA 8. *Let $v\in H^1_{0,\Gamma_0}(\Omega_\delta(t))$, and then*

$$\|v\|_{L^2(\Omega_\delta(t))}\le M\|\nabla v\|_{L^2(\Omega_\delta(t))}.$$

1.4. Weak formulation. Let $\eta_I\in H^2_0(\omega)$, $(\mathbf{u}_I,\dot\eta_I)\in L^2(\Omega_{\eta_I})\times L^2(\omega)$, such that

$$\begin{aligned}&\min_{\overline{\omega}}(1+\eta_I)>0,\\ &\operatorname{div}\mathbf{u}_I=0 \text{ in } \Omega_{\eta_I},\\ &\mathbf{u}_I\cdot\mathbf{n}=0 \text{ on } \Gamma_0,\\ &\mathbf{u}_I(x,y,1+\eta_I(x,y))\cdot\mathbf{n}_0=(0,0,\dot\eta_I(x,y))^T\cdot\mathbf{n}_0 \text{ on } \omega,\\ &\int_\omega\dot\eta_I(x,y)=0,\end{aligned} \tag{21}$$

where $\mathbf{n}_0$ denotes the unit normal to the initial position of the plate. We refer to [4], where one proves that the normal trace $\mathbf{u}_I(x,y,1+\eta_I(x,y))\cdot\mathbf{n}_0, (x,y)\in\omega$ makes sense for $\mathbf{u}_I\in L^2(\Omega_{\eta_I})$, with $\eta_I\in H_0^2(\omega)$. We shall say that $(\mathbf{u}_\varepsilon,\eta_\varepsilon)$ is a weak solution of the considered model on $(0,T)$ if it satisfies the following problem that will be denoted $(\mathcal{P}_\varepsilon)$:

- $\mathbf{u}_\varepsilon\in V_{\eta_\varepsilon}\cap L^\infty(0,T;L^2(\Omega_{\eta_\varepsilon}(t)))$, $\eta_\varepsilon\in W^{1,\infty}(0,T;L^2(\omega))\cap L^\infty(0,T;H_0^2(\omega))$,
- for $\varepsilon>0$, $\partial_t\eta_\varepsilon\in L^2(0,T;H_0^2(\omega))$,
- $\mathbf{u}_\varepsilon(t,x,y,1+\eta_\varepsilon(t,x,y))=(0,0,\partial_t\eta_\varepsilon(t,x,y))^T$ for a.e. $(t,x,y)\in(0,T)\times\omega$,
- for all $(\ \ ,b)\in(V_{\eta_\varepsilon}\cap H^1(\widehat{\Omega}_{\eta_\varepsilon}))\times(L^2(0,T;H_0^2(\omega))\times H^1(0,T;L^2(\omega)))$ such that $(t,x,y,1+\eta_\varepsilon(t,x,y))=(0,0,b(t,x,y))^T$, for a.e. $(t,x,y)\in(0,T)\times\omega$, we have for a.e. t

$$
\begin{aligned}
(22)\quad &\int_{\Omega_{\eta_\varepsilon}(t)}\mathbf{u}_\varepsilon(t)\cdot\ \ (t)-\int_0^t\int_{\Omega_{\eta_\varepsilon}(s)}\mathbf{u}_\varepsilon\cdot\partial_t\ \ +\nu\int_0^t\int_{\Omega_{\eta_\varepsilon}(s)}\nabla\mathbf{u}_\varepsilon:\nabla\\
&\quad+\int_0^t\int_{\Omega_{\eta_\varepsilon}(s)}(\mathbf{u}_\varepsilon\cdot\nabla)\mathbf{u}_\varepsilon\cdot\ \ -\int_0^t\int_\omega(\partial_t\eta_\varepsilon)^2b+\int_\omega\partial_t\eta_\varepsilon(t)\,b(t)\\
&\quad-\int_0^t\int_\omega\partial_t\eta_\varepsilon\,\partial_tb+\int_0^t\int_\omega\Delta\eta_\varepsilon\,\Delta b+\varepsilon\int_0^t\int_\omega\Delta\partial_t\eta_\varepsilon\,\Delta b\\
&\quad=\int_0^t\int_{\Omega_{\eta_\varepsilon}(t)}\mathbf{f}\cdot\ \ +\int_0^t\int_\omega g\ b+\int_{\Omega_{\eta_I}}\mathbf{u}_I\cdot\ \ (0)+\int_\omega\dot\eta_Ib(0).
\end{aligned}
$$

In what follows, a solution of $(\mathcal{P}_0)$ will be denoted by $(\mathbf{u},\eta)$ (instead of $(\mathbf{u}_0,\eta_0)$).

Remark 4. The test functions depends on the solution and thus, for $\varepsilon>0$, on ε.

Remark 5. The trace at time $t=0$ of $\ \in H^1(\widehat{\Omega}_{\eta_\varepsilon})$ such that $(t,x,y,1+\eta_\varepsilon(t,x,y))=(0,0,b(t,x,y))^T$ for a.e. $(t,x,y)\in(0,T)\times\omega$, with $b\in H^1((0,T)\times\omega)$, is well-defined and makes sense at least in $L^2(\Omega_{\eta_I})$. Indeed we can prove by density arguments ($C^1(\overline{\widehat{\Omega}_{\eta_\varepsilon}})$ is dense in $H^1(\widehat{\Omega}_{\eta_\varepsilon})$ since the domain is a continuous subgraph; see, for instance, [1, Thm. 2, p. 54]) that

$$
-\int_{\Omega_{\eta_\varepsilon(0)}}|\ \ (0)|^2=2\int_0^T\int_{\Omega_{\eta_\varepsilon}(t)}\psi\ \ \partial_t\ \ +\int_0^T\int_{\Omega_{\eta_\varepsilon}(t)}|\ \ |^2\partial_t\psi+\int_0^T\int_\omega|b|^2\psi\partial_t\eta_\varepsilon,
$$

where ψ belongs to $\mathcal{D}([0,T))$ and satisfies $\psi(0)=1$. The right-hand side of this equality makes sense for any $\ \in H^1(\widehat{\Omega}_{\eta_\varepsilon})$ such that $(t,x,y,1+\eta_\varepsilon(t,x,y))=(0,0,b(t,x,y))^T$ for a.e. $(t,x,y)\in(0,T)\times\omega$, with $b\in H^1((0,T)\times\omega)$, since $\partial_t\eta_\varepsilon\in L^\infty(0,T;L^2(\omega))$ and $b\in L^4((0,T)\times\omega)$.

2. Main result. We make the following hypotheses on the data (bulk forces and initial data):

$$
\begin{aligned}
(23)\qquad &\mathbf{f}\in L^2_{loc}((0,+\infty)\times\mathbb{R}^2),\ g\in L^2_{loc}((0,+\infty)\times\omega),\\
&\mathbf{u}_I\in L^2(\Omega_{\eta_I}),\ \dot\eta_I\in L^2(\omega),\ \eta_I\in H_0^2(\omega),
\end{aligned}
$$

and we assume moreover that conditions (21) are satisfied.

First we recall that, for $\varepsilon>0$, there exists at least one weak solution provided that the plate does not touch the bottom of the fluid cavity, in other words, as long as $\min_{(x,y)\in\overline{\omega}}(1+\eta_\varepsilon(t,x,y))>0$. The proof of the following theorem can be found in [4].

THEOREM 1. *Let ε be strictly positive. Under assumptions* (21) *and* (23) *and if* $\min_{(x,y)\in\overline{\omega}}1+\eta_I(x,y)>0$, *there exist* $T_\varepsilon^*\in(0,+\infty]$ *and a weak solution* $(\mathbf{u}_\varepsilon,\eta_\varepsilon)$ *of*

$(\mathcal{P}_\varepsilon)$ on $[0,T]$, $T < T_\varepsilon^*$. This solution satisfies an energy estimate for all $T < T_\varepsilon^*$:

$$\begin{aligned}&\|\mathbf{u}_\varepsilon\|_{L^\infty(0,T;L^2(\Omega_{\eta_\varepsilon}(t)))} + \|\nabla \mathbf{u}_\varepsilon\|_{L^2(0,T;L^2(\Omega_{\eta_\varepsilon}(t)))}\\ &+ \|\partial_t \eta_\varepsilon\|_{L^\infty(0,T;L^2(\omega))} + \|\Delta \eta_\varepsilon\|_{L^\infty(0,T;L^2(\omega))} + \sqrt{\varepsilon}\|\Delta \partial_t \eta_\varepsilon\|_{L^2(0,T;L^2(\omega))}\\ &\quad \le C(T, \|\mathbf{u}_I\|_{L^2(\Omega_{\eta_I})}, \|\mathbf{f}\|_{L^2((0,T)\times\mathbb{R}^2)}, \|g\|_{L^2((0,T)\times\omega)}, \|\eta_I\|_{H_0^2(\omega)}, \|\dot\eta_I\|_{L^2(\omega)}),\end{aligned}$$

where $C > 0$ is nondecreasing with respect to its arguments. Moreover, we have the following alternative:

- *either $T_\varepsilon^* = +\infty$*
- *or $\lim_{t\to T_\varepsilon^*} \min_{\overline{\omega}}(1+\eta_\varepsilon) = 0$.*

Since $\mathbf{u}_\varepsilon$ is bounded in $L^2(0,T;H^1(\Omega_{\eta_\varepsilon}(t)))$ uniformly in ε and thanks to the regularity of the moving elastic boundary, the trace $\gamma_{\eta_\varepsilon(t)}(\mathbf{u}_\varepsilon)$ is bounded in $L^2(0,T;W^{1-1/p,p}(\omega))$ $\forall 1 < p < 2$ and in $L^2(0,T;H^s(\omega))$ $\forall 0 \le s < \frac{1}{2}$ uniformly in ε (see Lemma 2 and (14)). Thus, thanks to the equality of the velocities (4),

$$\partial_t \eta_\varepsilon \text{ is bounded, uniformly in } \varepsilon, \text{ in } L^2(0,T;W^{1-1/p,p}(\omega)) \ \forall 1 < p < 2, \tag{24}$$

and

$$\partial_t \eta_\varepsilon \text{ is bounded, uniformly in } \varepsilon, \text{ in } L^2(0,T;H^s(\omega)) \ \forall 0 \le s < \frac{1}{2}. \tag{25}$$

This will be one of the key arguments for the derivation of the compactness properties of the sequence $(\mathbf{u}_\varepsilon, \eta_\varepsilon)$. In particular thanks to (25) we can control, uniformly in ε, the space high frequencies of $\partial_t \eta_\varepsilon$ in $L^2(0,T;L^2(\omega))$.

In all that follows, the characteristic function of $\widehat{\Omega}_{\eta_\varepsilon}$ will be denoted by ρ_ε, and $\rho_\varepsilon v$ will denote the function equal to v in $\widehat{\Omega}_{\eta_\varepsilon}$ and zero elsewhere. Moreover, we choose M large enough such that, for all $\varepsilon > 0$, $1 + \eta_\varepsilon(t,x,y) \le M$ $\forall (t,x,y) \in [0,T]\times\overline{\omega}$, which is made possible by (12) and (13). We set

$$\overline{\mathbf{u}}_\varepsilon = \begin{cases} \mathbf{u}_\varepsilon & \text{in } \widehat{\Omega}_{\eta_\varepsilon},\\ (0,0,\partial_t\eta_\varepsilon)^T & \text{in } \widehat{\mathcal{C}_M} \setminus \overline{\widehat{\Omega}}_{\eta_\varepsilon}. \end{cases}$$

In all that follows, for any $\mathbf{v}$ in $L^2(0,T;H^1_{0,\Gamma_0}(\Omega_{\eta_\varepsilon}(t)))$ such that $\gamma_{\delta(t)}(\mathbf{v}) = (0,0,b)^T$, $b \in L^2(0,T;H_0^1(\omega))$, $\overline{\mathbf{v}}$ is defined by (15).

The main results of the present paper are the following.

PROPOSITION 1. *The sequence $(T_\varepsilon^*)_{\varepsilon>0}$ is bounded from below away from zero, and the following convergences (up to the extractions of subsequences) hold as ε goes to zero:*

$$\begin{array}{llll} \eta_\varepsilon \to \eta & \textit{strongly in} & C^0([0,T];C^0(\overline{\omega})),\\ \eta_\varepsilon \rightharpoonup \eta & \textit{weakly in} & L^2(0,T;H_0^2(\omega)),\\ \partial_t\eta_\varepsilon \to \partial_t\eta & \textit{strongly in} & L^2(0,T;L^2(\omega)),\\ \rho_\varepsilon \mathbf{u}_\varepsilon \to \rho\mathbf{u} & \textit{strongly in} & L^2(0,T;L^2(\mathcal{C}_M)),\\ \overline{\mathbf{u}}_\varepsilon \to \overline{\mathbf{u}} & \textit{strongly in} & L^2(0,T;L^2(\mathcal{C}_M)),\\ \rho_\varepsilon \nabla\mathbf{u}_\varepsilon \rightharpoonup \rho\nabla\mathbf{u} & \textit{weakly in} & L^2(0,T;L^2(\mathcal{C}_M)), \end{array} \tag{26}$$

where $T > 0$ is a lower bound of T_ε^* independent of ε. Moreover

$$\gamma_{\eta(t)}(\mathbf{u}) = (0,0,\partial_t\eta)^T.$$

This enables us to pass to the limit in (22) as ε tends to zero and thus obtain the following theorem.

THEOREM 2. *Under assumptions* (21) *and* (23) *and if* $\min_{(x,y)\in\overline{\omega}} 1+\eta_I(x,y) > 0$, *there exist* $T^* \in (0,+\infty]$ *and a weak solution* $(\mathbf{u},\eta)$ *of* $(\mathcal{P}_0)$ *on* $[0,T]$, $T < T^*$. *This solution satisfies energy estimates for all* $T < T^*$:

$$\begin{aligned}(27)\qquad &\|\mathbf{u}\|_{L^\infty(0,T;L^2(\Omega_\eta(t)))} + \|\nabla\mathbf{u}\|_{L^2(0,T;L^2(\Omega_\eta(t)))}\\ &\quad + \|\partial_t\eta\|_{L^\infty(0,T;L^2(\omega))} + \|\eta\|_{L^\infty(0,T;H_0^2\omega)}\\ &\le C(T, \|\mathbf{u}_I\|_{L^2(\Omega_{\eta_I})}, \|\mathbf{f}\|_{L^2((0,T)\times\mathbb{R}^2)}, \|g\|_{L^2((0,T)\times\omega)}, \|\eta_I\|_{H_0^2(\omega)}, \|\dot\eta_I\|_{L^2(\omega)}),\end{aligned}$$

where $C > 0$ *is nondecreasing with respect to its arguments. The following alternatives are satisfied:*

- *either* $T^* = +\infty$
- *or* $\lim_{t\to T^*} \min_{\overline{\omega}}(1+\eta) = 0$.

3. Proof of Proposition 1. First we prove that Proposition 1 holds true. We have to verify that T_ε^* is bounded from below independently of ε and obtain compactness properties on $(\mathbf{u}_\varepsilon, \partial_t\eta_\varepsilon)$ in order to prove the desired strong convergences that will enable us to pass to the limit in $(\mathcal{P}_\varepsilon)$ as ε goes to zero.

Lower bound of T_ε^*. For $\varepsilon > 0$ the solution η_ε is bounded uniformly in ε in $L^\infty(0,T;H_0^2(\omega)) \cap W^{1,\infty}(0,T;L^2(\omega))$ for all $T < T_\varepsilon^*$. Thus from (13), η_ε is bounded uniformly in ε in $C^{0,\mu}([0,T];C^0(\overline{\omega}))$, $0 < \mu < \frac{1}{2}$. Consequently

$$1 + \eta_\varepsilon(t,x,y) \ge (1+\eta_I(x,y)) - Ct^\mu \quad \forall (t,x,y) \in [0,T_\varepsilon^*) \times \overline{\omega},$$

where C depends only on the data of the problem. Thus T_ε^* is bounded from below by a time independent of ε. Let T be such that

$$\forall \varepsilon > 0, \min_{(t,x,y)\in[0,T]\times\overline{\omega}} (1+\eta_\varepsilon(t,x,y)) \ge \alpha > 0,$$

where α is chosen such that $\min_{(x,y)\in\overline{\omega}}(1+\eta_I(x,y)) \ge 2\alpha > 0$.

Convergences of the sequence $(\mathbf{u}_\varepsilon, \eta_\varepsilon)$. From (12) and the compact injection (13) we deduce easily the first two convergences announced in Proposition 1. Next we prove strong convergence properties for the fluid and the structure velocities. The solution $(\mathbf{u}_\varepsilon, \eta_\varepsilon)_{\varepsilon>0}$ we build verifies estimate (9) and (25). Furthermore, since $\mathbf{u}_\varepsilon$ is bounded uniformly in ε in $L^2(0,T;H^1(\Omega_{\eta_\varepsilon}(t)))$, it is easy to verify that $\mathbf{w}_\varepsilon$, defined by $\mathbf{w}_\varepsilon(t,x,y,z) = \mathbf{u}_\varepsilon(t,x,y,z(1+\eta_\varepsilon(t,x,y)))$, is bounded uniformly in ε in $L^2(0,T;W^{1,p}(\mathcal{C}_1))$ $\forall 1 < p < 2$. This implies, thanks to Sobolev injections (see [1, Thm. 7.58, p. 218]) that $\mathbf{w}_\varepsilon$ is uniformly bounded in $L^2(0,T;H^\theta(\mathcal{C}_1))$ for any $\theta < 1$. Moreover $\partial_t\eta_\varepsilon$ is uniformly bounded in $L^2(0,T;H^s(\omega))$ $\forall 0 \le s < \frac{1}{2}$. Consequently, $\mathbf{w}_\varepsilon - \mathcal{R}_\alpha(\partial_t\eta_\varepsilon)$ is uniformly bounded in $L^2(0,T;H^s(\mathcal{C}_1))$ for any $0 \le s < \frac{1}{2}$ and its extension by zero for $z \ge 1$ is uniformly bounded in $L^2(0,T;H^s(\mathcal{C}_L))$ for any $s < \frac{1}{2}$, $L \ge 1$ (see [19]). Thus if we extend $\mathbf{w}_\varepsilon$ by $(0,0,\partial_t\eta_\varepsilon)^T$ for $z \ge 1$, this extension is uniformly bounded in $L^2(0,T;H^s(\mathcal{C}_L))$ $\forall 0 \le s < \frac{1}{2}, L \ge 1$. Consequently, since the change of variables $\phi_\varepsilon(t,x,y,z) = (x,y,z(1+\eta_\varepsilon(t,x,y))^T$ is in $L^\infty(0,T;C^{0,\beta}(\mathcal{C}_L))$ $\forall \beta < 1$ as well as its inverse, it is easy to verify that

$$(28)\qquad \overline{\mathbf{u}}_\varepsilon \text{ is bounded, uniformly in } \varepsilon, \text{ in } L^2(0,T;H^{s'}(\mathcal{C}_M))\ \forall 0 \le s' < s, \forall s < \frac{1}{2}.$$

Moreover thanks to the Sobolev injections, $\mathbf{w}_\varepsilon$ is bounded uniformly in ε in $L^2(0,T;L^q(\mathcal{C}_1))$ $\forall q < 6$ and $\partial_t\eta_\varepsilon$ is uniformly bounded in $L^2(0,T;L^r(\omega))$ $\forall r < 4$; thus

$$(29)\qquad \overline{\mathbf{u}}_\varepsilon \text{ is bounded, uniformly in } \varepsilon, \text{ in } L^2(0,T;L^r(\mathcal{C}_M))\ \forall r < 4.$$

Nevertheless, these bounds are not sufficient to obtain the desired strong convergences. We are going to use the following lemma that characterizes the compact sets of $L^p(0,T;X)$, where X is a Banach space (see [23]).

LEMMA 9. *Let X be a Banach space and $F \hookrightarrow L^q(0,T;X)$, with $1 \leq q < \infty$. Then F is a relatively compact set of $L^q(0,T;X)$ if and only if*

(i) $\{\int_{t_1}^{t_2} f(t)dt, f \in F\}$ *is relatively compact in X $\forall < t_1 < t_2 < T$;*

(ii) $\|f(t+h) - f(t)\|_{L^q(0,T;X)} \longrightarrow 0$ *as h goes to zero, uniformly with respect to f in F.*

We are going to apply Lemma 9 to $F = (\overline{\mathbf{u}}_\varepsilon, \partial_t \eta_\varepsilon)_{\varepsilon>0}$, $q = 2$, and $X = L^2(\mathcal{C}_M) \times L^2(\omega)$. The first point (i) is clearly satisfied thanks to (25) and (28), and we have to verify the second point. Given any $h > 0$, we denote that $g^-(t,\cdot) = g(t-h,\cdot)$ and $g^+(t,\cdot) = g(t+h,\cdot)$. The assertion (ii) is a consequence of the following lemma.

LEMMA 10. *Let $T > 0$ such that $\min_{[0,T]\times\overline{\omega}}(1+\eta_\varepsilon) \geq \alpha > 0$. We have $\forall \beta > 0$, $\exists h_0 > 0$, s.t. $\forall \varepsilon > 0$, $\forall h \leq h_0$*

$$\int_0^T \int_{\mathcal{C}_M} \rho_\varepsilon |\overline{\mathbf{u}}_\varepsilon - \overline{\mathbf{u}}_\varepsilon^-|^2 + \int_0^T \int_\omega (\partial_t \eta_\varepsilon - \partial_t \eta_\varepsilon^-)^2 \leq \beta \tag{30}$$

and

$$\int_0^T \int_{\mathcal{C}_M} |\rho_\varepsilon \overline{\mathbf{u}}_\varepsilon - \rho_\varepsilon^- \overline{\mathbf{u}}_\varepsilon^-|^2 \leq \beta, \tag{31}$$

with η_ε extended by η_I for $t < 0$ and $\overline{\mathbf{u}}_\varepsilon$ and $\partial_t \eta_\varepsilon$ extended by 0 for $t < 0$ and where ρ_ε denotes the characteristic function of $\widehat{\Omega}_{\eta_\varepsilon}$.

Proof. We first show that (30) implies (31). Indeed:

$$|\rho_\varepsilon \overline{\mathbf{u}}_\varepsilon - \rho_\varepsilon^- \overline{\mathbf{u}}_\varepsilon^-|^2 \leq C\left(\rho_\varepsilon |\overline{\mathbf{u}}_\varepsilon - \overline{\mathbf{u}}_\varepsilon^-|^2 + |\rho_\varepsilon - \rho_\varepsilon^-||\overline{\mathbf{u}}_\varepsilon^-|^2\right).$$

The estimate of the first contribution comes from (30). For the second contribution we use the fact that $\overline{\mathbf{u}}_\varepsilon$ is bounded uniformly in ε in $L^2(0,T;L^3(\mathcal{C}_M))$ (see (29)):

$$\left|\int_0^T \int_{\mathcal{C}_M} |\rho_\varepsilon - \rho_\varepsilon^-||\overline{\mathbf{u}}_\varepsilon|^2\right| \leq \int_0^T \|\rho_\varepsilon - \rho_\varepsilon^-\|_{L^3(\mathcal{C}_M)} \|\overline{\mathbf{u}}_\varepsilon\|^2_{L^3(\mathcal{C}_M)} \leq C \int_0^T \|\rho_\varepsilon - \rho_\varepsilon^-\|_{L^3(\mathcal{C}_M)}.$$

Remember now that $\partial_t \eta_\varepsilon$ is bounded in $L^\infty(0,T;L^2(\omega))$ uniformly in ε, and thus

$$\begin{aligned}
\int_{\mathcal{C}_M} |\rho_\varepsilon - \rho_\varepsilon^-|^3 &= \int_\omega |\eta_\varepsilon - (\eta_\varepsilon)^-| \\
&= \int_\omega \left|\int_{t-h}^t \partial_t \eta_\varepsilon(s) ds\right| \\
&\leq \int_\omega \int_{t-h}^t |\partial_t \eta_\varepsilon(s)| ds \\
&\leq Ch.
\end{aligned}$$

It leads to

$$\left|\int_0^T \int_{\mathcal{C}_M} (\rho_\varepsilon - \rho_\varepsilon^-)|\overline{\mathbf{u}}_\varepsilon|^2\right| \leq Ch^{\frac{1}{3}}. \tag{32}$$

This shows that (30) implies (31).

To prove (30) we are going to make a suitable choice for the test functions in the weak formulation satisfied by $\mathbf{u}_\varepsilon$ and η_ε:

$$
\begin{aligned}
(33)\quad & \int_{\Omega_{\eta_\varepsilon}(T)} \mathbf{u}_\varepsilon(T)\cdot \quad (T) - \int_0^T \int_{\Omega_{\eta_\varepsilon}(t)} \mathbf{u}_\varepsilon \cdot \partial_t \quad + \nu \int_0^T \int_{\Omega_{\eta_\varepsilon}(t)} \nabla \mathbf{u}_\varepsilon : \nabla \\
& + \int_0^T \int_{\Omega_{\eta_\varepsilon}(t)} (\mathbf{u}_\varepsilon \cdot \nabla)\mathbf{u}_\varepsilon \cdot \quad - \int_0^T \int_\omega (\partial_t \eta_\varepsilon)^2 b + \int_\omega \partial_t \eta_\varepsilon(T)\, b(T) \\
& - \int_0^T \int_\omega \partial_t \eta_\varepsilon\, \partial_t b + \int_0^T \int_\omega \Delta \eta_\varepsilon\, \Delta b + \varepsilon \int_0^T \int_\omega \Delta \partial_t \eta_\varepsilon\, \Delta b \\
& = \int_0^T \int_{\Omega_{\eta_\varepsilon}(t)} \mathbf{f}\cdot \quad + \int_0^T \int_\omega g\ b + \int_{\Omega_{\eta_I}} \mathbf{u}_I \cdot \quad (0) + \int_\omega \dot{\eta}_I b(0), \\
& \text{for all } (\quad, b) \in (V_{\eta_\varepsilon} \cap H^1(\widehat{\Omega}_{\eta_\varepsilon})) \times (L^2(0,T;H_0^2(\omega)) \cap H^1(0,T;L^2(\omega))), \\
& \text{s.t.} \quad (t,x,y,1+\eta_\varepsilon(t,x,y)) = (0,0,b(t,x,y))^T \text{ for a.e. } (t,x,y) \in (0,T)\times\omega.
\end{aligned}
$$

We are going to study separately the low frequencies and the high frequencies of $\partial_t \eta_\varepsilon$ and take advantage of the fact that $\partial_t \eta_\varepsilon$ is bounded in $L^2(0,T;H^s(\omega))\ \forall\ 0 \le s < \frac{1}{2}$ uniformly in ε (see (25)). This implies that we can control, uniformly in ε, the space high frequencies of $\partial_t \eta_\varepsilon$ in $L^2(0,T;L^2(\omega))$.

Definition of admissible test functions. First we introduce a basis of $H_0^2(\omega) \cap L_0^2(\omega)$ by taking eigenfunctions $(\xi_i)_{i\in\mathbb{N}}$ defined by:

$$
\begin{cases}
\displaystyle\int_\omega \Delta\xi_i \Delta b = \lambda_i \int_\omega \xi_i b \quad \forall b \in H_0^2(\omega) \text{ s.t.} \quad \int_\omega b = 0, \\
\xi_i \in H_0^2(\omega), \displaystyle\int_\omega \xi_i = 0,
\end{cases}
$$

with $(\lambda_i)_{i\in\mathbb{N}}$ the sequence of increasing eigenvalues: $\lambda_i > 0,\ \lambda_i \to +\infty$. We choose $(\xi_i)_{i\in\mathbb{N}}$ orthonormal in $L^2(\omega)$. We denote by d^{N_0} the L^2-projection on the finite-dimensional space $\mathrm{span}(\xi_i)_{0\le i\le N_0}$ of any function d and by d^{hf,N_0} the difference $d - d^{N_0}$. Thanks to the choice of the ξ_i, the L^2-projection on the finite-dimensional space $\mathrm{span}(\xi_i)_{0\le i\le N_0}$ is stable in the L^2-norm as well as in the H_0^2-norm. In what follows, we will use the fact that the following property, obtained by Hilbertian interpolation, holds true:

$$
(34)\qquad \forall d \in H^s(\omega), 0 \le s < \frac{1}{2}, \quad \|d^{hf,N_0}\|_{L^2(\omega)} \le \lambda_{N_0}^{-\frac{s}{2}} \|d\|_{H^s(\omega)}.
$$

Next, for $\sigma > 1$ we define $\mathbf{v}_\sigma$ by

$$
\mathbf{v}_\sigma(x,y,z) = (\sigma v_1(x,y,\sigma z), \sigma v_2(x,y,\sigma z), v_3(x,y,\sigma z)).
$$

If $\mathbf{v}$ is divergence-free, $\mathbf{v}_\sigma$ is also divergence-free.

We want now to define admissible test functions. We set

$$
b_\varepsilon = \int_{t-h}^t \partial_t \eta_\varepsilon^{N_0}(s) ds
$$

and

$$
{}_\varepsilon = \int_{t-h}^t \left(\overline{(\mathbf{u}_\varepsilon - \mathcal{R}_\alpha(\partial_t \eta_\varepsilon))^\lambda} \right)_\sigma (s) ds + \int_{t-h}^t \overline{\mathcal{R}_\alpha(\partial_t \eta_\varepsilon^{N_0})}(s) ds,
$$

where the extension $\mathbf{v} \mapsto \overline{\mathbf{v}}$ is defined by (15) and where $\mathcal{R}_\alpha$ is the lifting operator defined at Lemma 6. Moreover a space regularization of $\mathbf{v}_\varepsilon = \mathbf{u}_\varepsilon - \mathcal{R}_\alpha(\partial_t \eta_\varepsilon)$, denoted by $\mathbf{v}_\varepsilon^\lambda$, has been introduced in order to have $\quad_\varepsilon$ uniformly bounded in $H^1(0,T;H^1(\mathcal{C}_M))$. It verifies $\operatorname{div}(\mathbf{v}_\varepsilon^\lambda) = 0$, $\mathbf{v}_\varepsilon^\lambda \in L^2(0,T;H_0^1(\Omega_{\eta_\varepsilon}(t)))$, and

$$\begin{array}{l} \|\mathbf{v}_\varepsilon - \mathbf{v}_\varepsilon^\lambda\|_{L^2(0,T;L^2(\Omega_{\eta_\varepsilon}(t)))} \longrightarrow 0, \text{ uniformly in } \varepsilon, \text{ as } \lambda \text{ goes to zero,} \\ \|\mathbf{v}_\varepsilon^\lambda\|_{L^2(0,T;H^1(\Omega_{\eta_\varepsilon}(t)))} \leq C_\lambda. \end{array} \tag{35}$$

Note that the construction of $\mathbf{v}_\varepsilon^\lambda$ relies on the fact that η_ε converges uniformly to η and that the plate does not touch the bottom of the fluid cavity. Moreover, the uniform convergence of $(\mathbf{v}_\varepsilon^\lambda)_\lambda$ as $\lambda \longrightarrow 0$ in $L^2(0,T;L^2(\Omega_{\eta_\varepsilon}(t)))$ is made possible since $\overline{\mathbf{v}}_\varepsilon$ is uniformly bounded in $L^2(0,T;H^{s'}(\mathcal{C}_M)), 0 < s' < 1/2$ thanks to (17), (25), and (28). Note that, with this choice,

$$\begin{gathered} \|\partial_t \eta_\varepsilon^{N_0}\|_{L^\infty(0,T;L^2(\omega))} \leq C, \quad \|\partial_t \eta_\varepsilon^{N_0}\|_{L^2(0,T;H^s(\omega))} \leq C, \\ \|b_\varepsilon\|_{W^{1,\infty}(0,T;L^2(\omega))} \leq C, \quad \|b_\varepsilon\|_{H^1(0,T;H^s(\omega))} \leq C, \\ \|\overline{\mathbf{v}_\varepsilon^\lambda}\|_{L^\infty(0,T;L^2(\mathcal{C}_M))} \leq C, \|\overline{\mathbf{v}_\varepsilon^\lambda}\|_{L^2(0,T;H^{s'}((\mathcal{C}_M))} \leq C \ \forall s' < s < 1/2, \end{gathered}$$

and

$$\|\partial_t \eta_\varepsilon^{N_0}\|_{L^\infty(0,T;H^2(\omega))} \leq C_{N_0}, \|b_\varepsilon\|_{W^{1,\infty}(0,T;H^2(\omega))} \leq C_{N_0}, \|\overline{\mathbf{v}_\varepsilon^\lambda}\|_{L^2(0,T;H^1(\mathcal{C}_M))} \leq C_\lambda,$$

where C denotes and will denote in all that follows a strictly positive constant that depends only on the data and not on ε and N_0, and C_{N_0} (resp., C_λ) denotes and will denote a strictly positive constant that depends on the data and not on ε but may depend on N_0 (resp., λ). The integer N_0 (resp., the real λ) will be fixed later on and will be large enough (resp., small enough). Then for well chosen σ, $(\ _\varepsilon, b_\varepsilon)$ are admissible tests functions. Indeed, $\ _\varepsilon$ is divergence-free thanks to the definitions of the lifting operator $\mathcal{R}_\alpha$, the extension operator $\mathbf{v} \mapsto \overline{\mathbf{v}}$, the operator $\mathbf{v} \mapsto \mathbf{v}_\sigma$, and the regularization $\mathbf{v} \mapsto \mathbf{v}^\lambda$. Moreover $\ _\varepsilon$ belongs to $H^1(0,T;H^1(\mathcal{C}_M))$. The function b_ε belongs to $H^1(0,T;H_0^2(\omega))$. Both of them are bounded in the previous spaces independently of ε but not of N_0 and λ. Moreover since

$$\|\eta_\varepsilon\|_{L^\infty(0,T;H_0^2(\omega)) \cap W^{1,\infty}(0,T;L^2(\omega))} \leq C,$$

and remembering the imbedding (13), we have

$$\|\eta_\varepsilon - \eta_\varepsilon^-\|_{C^0([0,T]\times\overline{\omega})} \leq C h^\mu, \quad 0 < \mu < \frac{1}{2}.$$

Thus if σ is such that $\sigma \geq 1 + \frac{2C}{\alpha} h^\mu$, we have

$$_\varepsilon(t,x,y,1+\eta_\varepsilon(t,x,y)) = \left(0, \int_{t-h}^t \partial_t \eta_\varepsilon^{N_0}(s,x,y)ds\right)^T \text{ on } \omega.$$

In what follows we choose $\sigma = 1 + \frac{2C}{\alpha} h^\mu$, $0 < \mu < \frac{1}{2}$. Hence, with the choice of test

functions that we made, (33) is written:

$$
\begin{aligned}
(36)\quad & -\int_0^T\int_{\Omega_{\eta_\varepsilon}(t)} \mathbf{u}_\varepsilon\cdot((\overline{\mathbf{v}_\varepsilon^\lambda})_\sigma-(\overline{\mathbf{v}_\varepsilon^\lambda})_\sigma^-)\int_0^T\int_{\Omega_{\eta_\varepsilon}(t)}\mathbf{u}_\varepsilon\cdot(\overline{\mathcal{R}_\alpha(\partial_t\eta_\varepsilon^{N_0})}-\overline{\mathcal{R}_\alpha(\partial_t\eta_\varepsilon^{N_0})}^-)\\
&\qquad+\int_0^T\int_{\Omega_{\eta_\varepsilon}(t)}(\mathbf{u}_\varepsilon\cdot\nabla)\mathbf{u}_\varepsilon\cdot\ {}_\varepsilon+\nu\int_0^T\int_{\Omega_{\eta_\varepsilon}(t)}\nabla\mathbf{u}_\varepsilon:\nabla\ {}_\varepsilon\\
&\qquad+\int_{\Omega_{\eta_\varepsilon}(T)}\mathbf{u}_\varepsilon(T)\cdot\ {}_\varepsilon(T)-\int_0^T\int_\omega\partial_t\eta_\varepsilon\partial_t(\eta_\varepsilon^{N_0}-(\eta_\varepsilon^{N_0})^-)\\
&\quad-\int_0^T\int_\omega(\partial_t\eta_\varepsilon)^2(\eta_\varepsilon^{N_0}-(\eta_\varepsilon^{N_0})^-)+\int_0^T\int_\omega\Delta\eta_\varepsilon\Delta(\eta_\varepsilon^{N_0}-(\eta_\varepsilon^{N_0})^-)\\
&+\varepsilon\int_0^T\int_\omega\Delta\partial_t\eta_\varepsilon\Delta(\eta_\varepsilon^{N_0}-(\eta_\varepsilon^{N_0})^-)+\int_\omega\partial_t\eta_\varepsilon(T)(\eta_\varepsilon^{N_0}(T)-\eta_\varepsilon^{N_0}(T-h))\\
&\qquad=\int_0^T\int_{\Omega_{\eta_\varepsilon}(t)}\mathbf{f}\cdot\ {}_\varepsilon+\int_0^T\int_\omega g(\eta_\varepsilon^{N_0}-(\eta_\varepsilon^{N_0})^-).
\end{aligned}
$$

The two first terms can be written:

(37)

$$
\begin{aligned}
&-\int_0^T\int_{\Omega_{\eta_\varepsilon}(t)}\mathbf{u}_\varepsilon\cdot((\overline{\mathbf{v}_\varepsilon^\lambda})_\sigma-(\overline{\mathbf{v}_\varepsilon^\lambda})_\sigma^-)-\int_0^T\int_{\Omega_{\eta_\varepsilon}(t)}\mathbf{u}_\varepsilon\cdot\Big((\overline{\mathcal{R}_\alpha(\partial_t\eta_\varepsilon^{N_0})}-\overline{\mathcal{R}_\alpha(\partial_t\eta_\varepsilon^{N_0})}^-\Big)\\
&=-\int_0^T\int_{\Omega_{\eta_\varepsilon}(t)}\mathbf{u}_\varepsilon\cdot(\overline{\mathbf{u}}_\varepsilon-\overline{\mathbf{u}}_\varepsilon^-)-\int_0^T\int_{\Omega_{\eta_\varepsilon}(t)}\mathbf{u}_\varepsilon\cdot(\overline{\mathcal{R}_\alpha(\partial_t\eta_\varepsilon^{hf,N_0})}-\overline{\mathcal{R}_\alpha(\partial_t\eta_\varepsilon^{hf,N_0})}^-)\\
&\qquad-\int_0^T\int_{\Omega_{\eta_\varepsilon}(t)}\mathbf{u}_\varepsilon\cdot(((\overline{\mathbf{v}_\varepsilon^\lambda})_\sigma-\overline{\mathbf{v}_\varepsilon^\lambda})-((\overline{\mathbf{v}_\varepsilon^\lambda})_\sigma^--(\overline{\mathbf{v}_\varepsilon^\lambda})^-))\\
&\qquad-\int_0^T\int_{\Omega_{\eta_\varepsilon}(t)}\mathbf{u}_\varepsilon\cdot(((\overline{\mathbf{v}_\varepsilon^\lambda})-\overline{\mathbf{v}}_\varepsilon)-((\overline{\mathbf{v}_\varepsilon^\lambda})^--(\overline{\mathbf{v}}_\varepsilon)^-)).
\end{aligned}
$$

We set $I_1=\int_0^T\int_{\Omega_{\eta_\varepsilon}(t)}\mathbf{u}_\varepsilon\cdot(\overline{\mathbf{u}}_\varepsilon-\overline{\mathbf{u}}_\varepsilon^-)$.

$$
\begin{aligned}
I_1&=-\frac12\int_0^T\int_{\Omega_{\eta_\varepsilon}(t)}|\mathbf{u}_\varepsilon|^2+\frac12\int_0^T\int_{\Omega_{\eta_\varepsilon}(t)}|\overline{\mathbf{u}}_\varepsilon^-|^2-\frac12\int_0^T\int_{\Omega_{\eta_\varepsilon}(t)}|\overline{\mathbf{u}}_\varepsilon-\overline{\mathbf{u}}_\varepsilon^-|^2\\
&=-\frac12\int_0^T\int_{\Omega_{\eta_\varepsilon}(t)}|\mathbf{u}_\varepsilon|^2+\frac12\int_0^{T-h}\int_{\Omega_{\eta_\varepsilon}(t+h)}|\overline{\mathbf{u}}_\varepsilon|^2-\frac12\int_0^T\int_{\Omega_{\eta_\varepsilon}(t)}|\overline{\mathbf{u}}_\varepsilon-\overline{\mathbf{u}}_\varepsilon^-|^2\\
&=\frac12\int_0^{T-h}\int_{\mathcal{C}_M}(\rho_\varepsilon^+-\rho_\varepsilon)|\overline{\mathbf{u}}_\varepsilon|^2-\frac12\int_{T-h}^T\int_{\Omega_{\eta_\varepsilon}(t)}|\mathbf{u}_\varepsilon|^2-\frac12\int_0^T\int_{\Omega_{\eta_\varepsilon}(t)}|\overline{\mathbf{u}}_\varepsilon-\overline{\mathbf{u}}_\varepsilon^-|^2.
\end{aligned}
$$

The same argument we use to prove (32) leads to

$$\int_0^T\int_{\mathcal{C}_M}|\rho_\varepsilon^+-\rho_\varepsilon||\overline{\mathbf{u}}_\varepsilon|^2\le Ch^{\frac13}.$$

This yields

$$I_1\le Ch^{\frac13}-\frac12\int_0^T\int_{\Omega_{\eta_\varepsilon}(t)}|\overline{\mathbf{u}}_\varepsilon-\overline{\mathbf{u}}_\varepsilon^-|^2.\tag{38}$$

For the second term of (37) $\int_0^T \int_{\Omega_{\eta_\varepsilon}(t)} \mathbf{u}_\varepsilon \cdot (\overline{\mathcal{R}_\alpha(\partial_t \eta_\varepsilon^{hf,N_0})} - \overline{\mathcal{R}_\alpha(\partial_t \eta_\varepsilon^{hf,N_0})^-})$ we have, by taking into account the energy estimate (9),

$$\begin{aligned}
&\left| \int_0^T \int_{\Omega_{\eta_\varepsilon}(t)} \mathbf{u}_\varepsilon \cdot (\overline{\mathcal{R}_\alpha(\partial_t \eta_\varepsilon^{hf,N_0})} - \overline{\mathcal{R}_\alpha(\partial_t \eta_\varepsilon^{hf,N_0})^-}) \right| \\
&\qquad \le C \|\mathbf{u}_\varepsilon\|_{L^2(0,T;L^2(\Omega_{\eta_\varepsilon}(t)))} \|\overline{\mathcal{R}_\alpha(\partial_t \eta_\varepsilon^{hf,N_0})}\|_{L^2(0,T;L^2(\mathcal{C}_M))} \\
&\qquad \le C(\|\mathcal{R}_\alpha(\partial_t \eta_\varepsilon^{hf,N_0})\|_{L^2(0,T;L^2(\mathcal{C}_\alpha))} + \|\partial_t \eta_\varepsilon^{hf,N_0}\|_{L^2(0,T;L^2(\omega))}).
\end{aligned}$$

Thanks to the properties satisfied by $\mathcal{R}_\alpha$ and in particular (17) for $s = 0$ we have

$$\|\mathcal{R}_\alpha(\partial_t \eta_\varepsilon^{hf,N_0})\|_{L^2(0,T;L^2(\mathcal{C}_\alpha))} \le C \|\partial_t \eta_\varepsilon^{hf,N_0}\|_{L^2(0,T;L^2(\omega))}, \tag{39}$$

and thus we obtain

$$\left| \int_0^T \int_{\Omega_{\eta_\varepsilon}(t)} \mathbf{u}_\varepsilon \cdot (\overline{\mathcal{R}_\alpha(\partial_t \eta_\varepsilon^{hf,N_0})} - \overline{\mathcal{R}_\alpha(\partial_t \eta_\varepsilon^{hf,N_0})^-}) \right| \le C \|\partial_t \eta_\varepsilon^{hf,N_0}\|_{L^2(0,T;L^2(\omega))}.$$

The definition of $\partial_t \eta_\varepsilon^{hf,N_0}$ and the fact that $\partial_t \eta_\varepsilon$ is bounded in $L^2(0,T;H^s(\omega))$, $s < 1/2$ independently of ε, imply, remembering (34), that

$$\left| \int_0^T \int_{\Omega_{\eta_\varepsilon}(t)} \mathbf{u}_\varepsilon \cdot (\overline{\mathcal{R}_\alpha(\partial_t \eta_\varepsilon^{hf,N_0})} - \overline{\mathcal{R}_\alpha(\partial_t \eta_\varepsilon^{hf,N_0})^-}) \right| \le C \|\partial_t \eta_\varepsilon^{hf,N_0}\|_{L^2(0,T;L^2(\omega))} \le C \lambda_{N_0}^{-\frac{s}{2}}. \tag{40}$$

We have also for the third term of (37):

$$\begin{aligned}
&\left| \int_0^T \int_{\Omega_{\eta_\varepsilon}(t)} \mathbf{u}_\varepsilon \cdot (((\overline{\mathbf{v}_\varepsilon^\lambda})_\sigma - \overline{\mathbf{v}_\varepsilon^\lambda}) - ((\overline{\mathbf{v}_\varepsilon^\lambda})_\sigma^- - (\overline{\mathbf{v}_\varepsilon^\lambda})^-) \right| \\
&\qquad \le 2 \|\mathbf{u}_\varepsilon\|_{L^2(0,T;L^2(\Omega_{\eta_\varepsilon}(t)))} \|(\overline{\mathbf{v}_\varepsilon^\lambda})_\sigma - \overline{\mathbf{v}_\varepsilon^\lambda})\|_{L^2(0,T;L^2(\Omega_{\eta_\varepsilon}(t)))} \\
&\qquad\qquad \le (\sigma - 1) \|\overline{\mathbf{v}_\varepsilon^\lambda}\|_{L^2(0,T;H^1(\Omega_{\eta_\varepsilon}(t)))} \\
&\qquad\qquad\qquad \le C_\lambda (\sigma - 1) \le C_\lambda h^\mu, \quad 0 < \mu < \frac{1}{2}.
\end{aligned} \tag{41}$$

Finally, due to the properties of the space regularization $\mathbf{v} \mapsto \mathbf{v}^\lambda$, the last term of the right-hand side of (37) goes to zero, uniformly in ε, as λ goes to zero. That is written: $\forall \beta > 0$ there exists $\lambda_0 > 0$ such that $\forall \lambda < \lambda_0$

$$\left| \int_0^T \int_{\Omega_{\eta_\varepsilon}(t)} \mathbf{u}_\varepsilon \cdot (((\overline{\mathbf{v}_\varepsilon^\lambda}) - \overline{\mathbf{v}}_\varepsilon) - ((\overline{\mathbf{v}_\varepsilon^\lambda})^- - (\overline{\mathbf{v}}_\varepsilon)^-)) \right| \le C\beta \qquad \forall \varepsilon. \tag{42}$$

Now we take care of the convective term

$$I_2 = \int_0^T \int_{\Omega_{\eta_\varepsilon}(t)} (\mathbf{u}_\varepsilon \cdot \nabla) \mathbf{u}_\varepsilon \cdot \left(\int_{t-h}^t \quad _\varepsilon \right),$$

with $\quad_\varepsilon = (\overline{\mathbf{v}_\varepsilon^\lambda})_\sigma + \overline{\mathcal{R}_\alpha(\partial_t \eta_\varepsilon^{N_0})}$. We have

$$
\begin{aligned}
|I_2| &\le \frac{1}{2}\int_0^T \|\mathbf{u}_\varepsilon\|_{L^4(\Omega_{\eta_\varepsilon}(t))}\|\nabla \mathbf{u}_\varepsilon\|_{L^2(\Omega_{\eta_\varepsilon}(t))}\int_{t-h}^t \|\quad_\varepsilon\|_{L^4(\Omega_{\eta_\varepsilon}(t))}\\
&\le \frac{\sqrt{h}}{2}\int_0^T \|\mathbf{u}_\varepsilon\|_{L^4(\Omega_{\eta_\varepsilon}(t))}\|\nabla \mathbf{u}_\varepsilon\|_{L^2(\Omega_{\eta_\varepsilon}(t))}\left(\int_{t-h}^t \|\quad_\varepsilon\|^2_{L^4(\Omega_{\eta_\varepsilon}(t))}\right)^{\frac{1}{2}} \qquad (43)\\
&\le C_{N_0,\lambda}\sqrt{h}.
\end{aligned}
$$

The next term to consider is $I_3 = \nu\int_0^T \int_{\Omega_{\eta_\varepsilon}(t)} \nabla \mathbf{u}_\varepsilon : \nabla(\int_{t-h}^t \quad_\varepsilon)$.

$$
\begin{aligned}
|I_3| &\le \nu\int_0^T \|\nabla \mathbf{u}_\varepsilon\|_{L^2(\Omega_{\eta_\varepsilon}(t))}\int_{t-h}^t \|\nabla \quad_\varepsilon\|_{L^2(\Omega_{\eta_\varepsilon}(t))}\\
&\le \nu\sqrt{h}\int_0^T \|\nabla \mathbf{u}_\varepsilon\|_{L^2(\Omega_{\eta_\varepsilon}(t))}\left(\int_{t-h}^t \|\nabla \quad_\varepsilon\|^2_{L^2(\Omega_{\eta_\varepsilon}(t))}\right)^{\frac{1}{2}} \qquad (44)\\
&\le C_{N_0,\lambda}\sqrt{h}.
\end{aligned}
$$

The term $I_4 = \int_{\Omega_{\eta_\varepsilon}(T)} \mathbf{u}_\varepsilon(T)\cdot(\int_{T-h}^T \quad_\varepsilon)$ can be estimated as follows:

$$
\begin{aligned}
|I_4| &\le \|\mathbf{u}_\varepsilon(T)\|_{L^2(\Omega_{\eta_\varepsilon}(T))}\int_{T-h}^T \|\quad_\varepsilon\|_{L^2(\Omega_{\eta_\varepsilon}(T))}\\
&\le \sqrt{h}\|\mathbf{u}_\varepsilon(T)\|_{L^2(\Omega_{\eta_\varepsilon}(T))}\left(\int_{T-h}^T \|\quad_\varepsilon\|^2_{L^2(\Omega_{\eta_\varepsilon}(T))}\right)^{\frac{1}{2}} \qquad (45)\\
&\le C\sqrt{h}.
\end{aligned}
$$

We set $I_5 = -\int_0^T \int_\omega \partial_t\eta_\varepsilon \partial_t(\eta_\varepsilon^{N_0} - (\eta_\varepsilon^{N_0})^-)$.

We have thanks to the definition of $\eta_\varepsilon^{N_0}$

$$
\begin{aligned}
I_5 &= -\frac{1}{2}\int_0^T\int_\omega (\partial_t\eta_\varepsilon^{N_0})^2 + \frac{1}{2}\int_0^T\int_\omega ((\partial_t\eta_\varepsilon^{N_0})^-)^2 - \frac{1}{2}\int_0^T\int_\omega (\partial_t\eta_\varepsilon^{N_0} - (\partial_t\eta_\varepsilon^{N_0})^-)^2 \qquad (46)\\
&= -\frac{1}{2}\int_{T-h}^T\int_\omega (\partial_t\eta_\varepsilon^{N_0})^2 - \frac{1}{2}\int_0^T\int_\omega (\partial_t\eta_\varepsilon^{N_0} - (\partial_t\eta_\varepsilon^{N_0})^-)^2\\
&\le -\frac{1}{2}\int_0^T\int_\omega (\partial_t\eta_\varepsilon^{N_0} - (\partial_t\eta_\varepsilon^{N_0})^-)^2.
\end{aligned}
$$

For the next term we have

$$
\begin{aligned}
|I_6| &= \left|\frac{1}{2}\int_0^T\int_\omega (\partial_t\eta_\varepsilon^{N_0})^2(\eta_\varepsilon^{N_0} - (\eta_\varepsilon^{N_0})^-)\right|\\
&\le \frac{1}{2}\int_0^T \|\partial_t\eta_\varepsilon^{N_0}\|^2_{L^3(\omega)}\|\eta_\varepsilon^{N_0} - (\eta_\varepsilon^{N_0})^-\|_{L^3(\omega)},
\end{aligned}
$$

but, taking into account the continuous imbedding of $H^s(\omega), 0 \le s < 1/2$, in $L^3(\omega)$ and the fact that $\partial_t\eta_\varepsilon^{N_0}$ is bounded uniformly in ε in $L^2(0,T,H^s(\omega))\ \forall 0 \le s < 1/2$,

$$
\|\eta_\varepsilon^{N_0} - (\eta_\varepsilon^{N_0})^-\|_{L^3(\omega)} \le \int_{t-h}^t \|\partial_t\eta_\varepsilon^{N_0}\|_{L^3(\omega)} \le C\sqrt{h}.
$$

Consequently

$$|I_6| \leq C\sqrt{h}. \tag{47}$$

The next term to consider is $I_7 = \int_0^T \int_\omega \Delta\eta_\varepsilon^{N_0} \Delta(\eta_\varepsilon^{N_0} - (\eta_\varepsilon^{N_0})^-)$. It can be estimated as follows:

$$\begin{aligned} |I_7| &\leq \int_0^T \|\Delta\eta_\varepsilon^{N_0}\|_{L^2(\omega)} \int_{t-h}^t \|\Delta\partial_t\eta_\varepsilon^{N_0}\|_{L^2(\omega)} \\ &\leq \sqrt{h} \int_0^T \|\Delta\eta_\varepsilon^{N_0}\|_{L^2(\omega)} \left(\int_{t-h}^t \|\Delta\partial_t\eta_\varepsilon^{N_0}\|_{L^2(\omega)}^2 \right)^{\frac{1}{2}} \\ &\leq C_{N_0}\sqrt{h}. \end{aligned} \tag{48}$$

The additional viscous term gives for $\varepsilon \leq 1$:

$$\begin{aligned} |I_8| &= \varepsilon \left| \int_0^T \int_\omega \Delta\partial_t\eta_\varepsilon^{N_0} \Delta(\eta_\varepsilon^{N_0} - (\eta_\varepsilon^{N_0})^-) \right| \\ &\leq \int_0^T \|\Delta\partial_t\eta_\varepsilon^{N_0}\|_{L^2(\omega)} \int_{t-h}^t \|\Delta\partial_t\eta_\varepsilon^{N_0}\|_{L^2(\omega)} \\ &\leq \sqrt{h} \int_0^T \|\Delta\partial_t\eta_\varepsilon^{N_0}\|_{L^2(\omega)} \left(\int_{t-h}^t \|\Delta\partial_t\eta_\varepsilon^{N_0}\|_{L^2(\omega)}^2 \right)^{\frac{1}{2}} \\ &\leq C_{N_0}\sqrt{h}. \end{aligned} \tag{49}$$

We set $I_9 = \int_\omega \partial_t\eta_\varepsilon^{N_0}(T)(\eta_\varepsilon(T) - \eta_\varepsilon^{N_0}(T-h))$.

$$\begin{aligned} |I_9| &= \|\partial_t\eta_\varepsilon(T)\|_{L^2(\omega)} \int_{T-h}^T \|\partial_t\eta_\varepsilon^{N_0}\|_{L^2(\omega)} \\ &\leq Ch. \end{aligned} \tag{50}$$

Next $I_{10} = \int_0^T \int_{\Omega_{\eta_\varepsilon}(t)} \mathbf{f} \cdot (\int_{t-h}^t \quad _\varepsilon)$ and $I_{11} = \int_0^T \int_\omega g(\eta_\varepsilon^{N_0} - (\eta_\varepsilon^{N_0})^-)$ can be estimated respectively by $C\sqrt{h}$ and Ch.

These last two estimates and estimates (38)–(50) yield, for all $\beta > 0$, for λ small enough

$$\int_0^T \int_{\Omega_{\eta_\varepsilon}(t)} |\mathbf{u}_\varepsilon - \overline{\mathbf{u}}_\varepsilon^-|^2 + \int_0^T \int_\omega (\partial_t\eta_\varepsilon^{N_0} - (\partial_t\eta_\varepsilon^{N_0})^-)^2 \leq C_{N_0,\lambda} h^{\frac{1}{3}} + C\lambda_{N_0}^{-\frac{s}{2}} + C\beta \ \forall \varepsilon < \varepsilon_0,$$

with $s < \frac{1}{2}$. Moreover,

$$\int_0^T \int_\omega (\partial_t\eta_\varepsilon^{hf,N_0} - (\partial_t\eta_\varepsilon^{hf,N_0})^-)^2 \leq C\lambda_{N_0}^{-\frac{s}{2}}.$$

Then for all $\beta > 0$, if λ is chosen small enough, we have

$$\int_0^T \int_{\Omega_{\eta_\varepsilon}(t)} |\mathbf{u}_\varepsilon - \overline{\mathbf{u}}_\varepsilon^-|^2 + \int_0^T \int_\omega (\partial_t\eta_\varepsilon - (\partial_t\eta_\varepsilon)^-)^2 \leq C_{N_0,\lambda} h^{\frac{1}{3}} + C\lambda_{N_0}^{-\frac{s}{2}} + C\beta \quad \forall \varepsilon < \varepsilon_0.$$

Thus, by choosing N_0 large enough and λ small enough, we obtain that $\exists h_0 > 0$ such that $\forall h \leq h_0$

$$\int_0^T \int_{\Omega_{\eta_\varepsilon}(t)} |\mathbf{u}_\varepsilon - \overline{\mathbf{u}}_\varepsilon^-|^2 + \int_0^T \int_\omega (\partial_t \eta_\varepsilon - (\partial_t \eta_\varepsilon)^-)^2 \leq C\beta \quad \forall \varepsilon < \varepsilon_0.$$

This proves Lemma 10. □

Thanks to Lemma 9, we obtain that $\partial_t \eta_\varepsilon$ is compact in $L^2(0,T;L^2(\omega))$ and that $\overline{\mathbf{u}}_\varepsilon$ is compact in $L^2(0,T;L^2(\mathcal{C}_M))$.

We now want to verify the convergences announced in Proposition 1 and to verify that the equality between the structure velocity and the fluid velocity at the interface holds in the limit.

Let $T > 0$ such that $\inf_\varepsilon \min_{[0,T]\times\overline{\omega}}(1+\eta_\varepsilon) \geq \alpha > 0$. We will denote any subsequence of $(\eta_\varepsilon, \overline{\mathbf{u}}_\varepsilon)$ by $(\eta_\varepsilon, \overline{\mathbf{u}}_\varepsilon)$. Thanks to the energy estimate and to the compactness properties that have just been derived, we have, denoting by $(\eta, \underline{\mathbf{u}})$ the limit of a subsequence of $(\eta_\varepsilon, \overline{\mathbf{u}}_\varepsilon)$, the following convergences as ε goes to zero:

$$\begin{array}{llll}
\eta_\varepsilon & \to & \eta & \text{uniformly in} \quad C^0([0,T];C^0(\overline{\omega})), \\
\eta_\varepsilon & \rightharpoonup & \eta & \text{weakly in} \quad L^2(0,T;H_0^2(\omega)), \\
\partial_t \eta_\varepsilon & \to & \partial_t \eta & \text{strongly in} \quad L^2(0,T;L^2(\omega)), \\
\partial_t \eta_\varepsilon & \to & \partial_t \eta & \text{weakly in} \quad L^2(0,T;W^{1-1/p,p}(\omega)) \ \forall\ 1 < p < 2, \\
\partial_t \eta_\varepsilon & \to & \partial_t \eta & \text{weakly in} \quad L^2(0,T;H^s(\omega)) \ \forall\ 0 \leq s < 1/2, \\
\overline{\mathbf{u}}_\varepsilon & \to & \underline{\mathbf{u}} & \text{strongly in} \quad L^2(0,T;L^2(\mathcal{C}_M)), \\
\rho_\varepsilon \mathbf{u}_\varepsilon & \to & \rho\underline{\mathbf{u}} & \text{strongly in} \quad L^2(0,T;L^2(\mathcal{C}_M)).
\end{array}$$

Moreover $\rho_\varepsilon \nabla \mathbf{u}_\varepsilon$ tends to some $\mathbf{z}$ weakly in $L^2(0,T;L^2(\mathcal{C}_M))$ as ε goes to zero. It is easy to verify, since $\eta_\varepsilon \to \eta$ in $C^0([0,T]\times\overline{\omega})$, that $\mathbf{z} = 0$ in $\widehat{\mathcal{C}_M}\backslash\widehat{\Omega}_\eta$ and $\mathbf{z}|_{\widehat{\Omega}_\eta} = \nabla(\underline{\mathbf{u}}|_{\widehat{\Omega}_\eta})$. Thus

$$\rho_\varepsilon \nabla \mathbf{u}_\varepsilon \rightharpoonup \rho \nabla(\underline{\mathbf{u}}|_{\widehat{\Omega}_\eta}) \ \text{ in } \ L^2(0,T;L^2(\mathcal{C}_M)). \tag{51}$$

Note also that $\underline{\mathbf{u}} = (0,0,\partial_t\eta)^T$ in $\widehat{\mathcal{C}_M} \setminus \widehat{\Omega}_\eta$, and thus $\underline{\mathbf{u}} = \overline{\mathbf{u}}$ by setting $\mathbf{u} = \underline{\mathbf{u}}|_{\widehat{\Omega}_\eta}$.

Next we take care of the equality

$$\mathbf{u}_\varepsilon(t,x,y,1+\eta_\varepsilon(t,x,y)) = (0,0,\partial_t\eta_\varepsilon(t,x,y))^T$$

on $(0,T)\times\omega$. The right-hand side converges to $(0,0,\partial_t\eta)^T$ strongly in $L^2(0,T;L^2(\omega))$. The left-hand side is the trace of the function $\mathbf{w}_\varepsilon(t,x,y,z) = \mathbf{u}_\varepsilon(t,x,,y,z(1+\eta_\varepsilon(t,x,y))$ on $z=1$, and $\mathbf{w}_\varepsilon$ converges strongly in $L^2(0,T;L^2(\mathcal{C}_1))$ and weakly in $L^2(0,T;W^{1,p}(\mathcal{C}_1))$ $\forall\ 1 < p < 2$ to $\mathbf{u}(t,x,,y,z(1+\eta(t,x,y))$. Hence by the continuity of the trace mapping on $z=1$, we have, for a.e. t,

$$\mathbf{u}(t,x,y,1+\eta(t,x,y)) = (0,0,\partial_t\eta(t,x,y))^T \text{ on } \omega.$$

This ends the proof of Proposition 1.

4. Passage to the limit—Proof of Theorem 2. Next we pass to the limit in the weak formulation:

$$\begin{aligned}\int_{\Omega_{\eta_\varepsilon}(t)} \mathbf{u}_\varepsilon(t)\cdot\ {}_\varepsilon(t) - \int_0^t\int_{\Omega_{\eta_\varepsilon}(s)} \mathbf{u}_\varepsilon\cdot\partial_t\ {}_\varepsilon + \nu\int_0^t\int_{\Omega_{\eta_\varepsilon}(s)} \nabla\mathbf{u}_\varepsilon : \nabla\ {}_\varepsilon \\ + \int_0^t\int_{\Omega_{\eta_\varepsilon}(s)} (\mathbf{u}_\varepsilon\cdot\nabla)\mathbf{u}_\varepsilon\cdot\ {}_\varepsilon - \int_0^t\int_\omega (\partial_t\eta_\varepsilon)^2 b + \int_\omega \partial_t\eta_\varepsilon(t)\, b(t) \\ - \int_0^t\int_\omega \partial_t\eta_\varepsilon\,\partial_t b + \varepsilon\int_0^t\int_\omega \Delta\partial_t\eta_\varepsilon\,\Delta b + \int_0^t\int_\omega \Delta\eta_\varepsilon\,\Delta b \\ = \int_0^t\int_{\Omega_{\eta_\varepsilon}(t)} \mathbf{f}\cdot\ {}_\varepsilon + \int_0^t\int_\omega g\ b + \int_{\Omega(0)} \mathbf{u}_I\cdot\ {}_\varepsilon(0) + \int_\omega \dot\eta_I b(0)\end{aligned}$$

for a.e. t and for all $({}_\varepsilon, b) \in (V_{\eta_\varepsilon}\cap H^1(\widehat{\Omega}_{\eta_\varepsilon}))\times(L^2(0,T;H_0^2(\omega))\times H^1(0,T;L^2(\omega)))$ such that ${}_\varepsilon(t,x,y,1+\eta_\varepsilon(t,x,y)) = (0,0,b(t,x,y))^T$, $(t,x,y)\in[0,T]\times\omega$.

The fluid test functions should depend on ε. However, it is sufficient to consider a dense family of test functions, and it can be chosen independent of ε and admissible for any ε small enough.

First we consider test functions of the form $({}^0, 0)$ such that ^{0} belongs to $\mathcal{D}(\cup_{t\in[0,T]}\{t\}\times\Omega_\eta(t))$ and $\operatorname{div}\ {}^0 = 0$. These test functions satisfy the property that ${}^0(t,\cdot)\in\mathcal{D}(\Omega_\eta(t))\ \forall t$. For ε small enough, since η_ε converges uniformly to η, ${}^0\in\mathcal{D}(\cup_{t\in[0,T]}\{t\}\times\Omega_{\eta_\varepsilon}(t))$.

The second pair of test functions we consider is $({}^1, b)$, where b belongs to $L^2(0,T;H_0^2(\omega))\cap H^1(0,T;L^2(\omega))$, with $\int_\omega b = 0$ and for a.e. t

$${}^1 = \left|\begin{array}{l}(0,0,b)^T \text{ in } \widehat{\mathcal{C}_M}\setminus\overline{\widehat{C}}_\alpha,\\ \mathcal{R}(0,0,\frac{z}{\alpha}b)^T \text{ in } \widehat{C}_\alpha,\end{array}\right.$$

where $\widehat{C}_\alpha = (0,T)\times\mathcal{C}_\alpha$ and $\mathcal{R}$ is a linear lifting operator from $\{\mathbf{w}\in H^{\frac12}(\partial C_\alpha); \int_{\partial C_\alpha}\mathbf{w}\cdot\mathbf{n} = 0\}$ onto $\{\mathbf{v}\in H^1(C_\alpha); \operatorname{div}(\mathbf{v}) = 0\}$. We have easily that ^{1} belongs to $L^2(0,T;H^1(\mathcal{C}_M))$ and $\operatorname{div}({}^1) = 0$. Moreover since $\min_{[0,T]\times\overline{\omega}}(1+\eta_\varepsilon)\ge\alpha$, ${}^1(t,x,y,1+\eta_\varepsilon(t,x,y)) = (0,0,b(t,x,y))^T$ on $(0,T)\times(\omega)\ \forall\varepsilon$. Furthermore we can choose the linear operator $\mathcal{R}$ such that

$$\left\|\mathcal{R}\left(0,0,\frac{zb}{\alpha}\right)^T\right\|_{L^2(\omega\times(0,\alpha))} \le C\|b\|_{L^2(\omega)}. \tag{52}$$

It can be done by solving a Stokes problem in $\omega\times(0,\alpha)$. Indeed this type of inequality can be obtained thanks to a transposition argument and relies on a $H^2\times H^1$ regularity result for the Stokes problem, which is true here since $\omega\times(0,\alpha)$ is a convex set (see [9] for the regularity result for the Stokes problem). Thus if $\mathcal{R}$ is chosen such that (52) holds, we deduce that, for a.e. $t\in I\subset\subset(0,T)$, and for h small enough

$$\left\|\mathcal{R}\left(0,0,\frac{zb}{\alpha}\right)^T(t) - \mathcal{R}\left(0,0,\frac{zb}{\alpha}\right)^T(t+h)\right\|_{L^2(\omega\times(0,\alpha))} \le C\|b(t)-b(t+h)\|_{L^2(\omega)}.$$

Since $\partial_t b\in L^2(0,T;L^2(\omega))$ this implies that $\partial_t\mathcal{R}(0,0,\frac{zb}{\alpha})^T\in L^2(0,T;L^2(\mathcal{C}_\alpha))$ and that $\partial_t\ {}^1\in L^2(0,T;L^2(\mathcal{C}_M))$. Consequently $({}^1, b)$ is a pair of admissible test functions for all ε.

With both types of test functions, it is easy to pass to the limit in the weak formulation as ε goes to zero. Since the considered family of test functions is dense in the set of functions $(\ \ ,b) \in (V_\eta \cap H^1(\widehat{\Omega}_\eta)) \times (L^2(0,T;H_0^2(\omega)) \times H^1(0,T;L^2(\omega)))$ such that $(t,x,y,1+\eta(t,x,y)) = (0,0,b(t,x,y))^T$, $(t,x,y) \in [0,T] \times \omega$, we obtain the existence of one weak solution on $(0,T)$ of (22) that moreover satisfies the energy estimate (27) by passing to the limit as ε tends to zero in (9).

Eventually, we show that we can extend the solution, as long as we have $\min_{[0,T]\times\overline{\omega}}(1+\eta) > 0$. We do exactly as in [4], but for the sake of completness we reproduce the proof here. We build an increasing sequence of times $(T_k)_{k\geq 1}$ as follows. First we choose a time $T_1 > 0$ such that there exists a weak solution up to T_1, with $m_1 = \min_{[0,T_1]\times\overline{\omega}}(1+\eta) > 0$. Possibly changing slightly T_1, we may moreover assume that $\eta(T_1) \in H_0^2(\omega)$, $\partial_t\eta(T_1) \in L^2(\omega)$, and $\mathbf{u}(T_1) \in L^2(\Omega_\eta(T_1))$ (since this is true for almost each time).

Now let $k \geq 1$, and assume that we have built a solution up to some time T_k, with $m_k = \min_{[0,T_k]\times\overline{\omega}}(1+\eta) > 0$. Our construction allows us to build an extension of our solution on some time interval starting from T_k. Thanks to the energy estimate (27) (see also (9)), we have for $s \geq T_k$ for any $0 < \lambda < \frac{3}{4}$

$$1+\eta(s) \geq 1+\eta(T_k) - (s-T_k)^\lambda C(T_k,s) \geq m_k - (s-T_k)^\lambda C(T_k,s)\,, \tag{53}$$

with

$$C(T_k,s) = \tilde{C}\Big(\|\mathbf{u}(T_k)\|_{L^2(\Omega_{\eta_I})}, \|\eta(T_k)\|_{H_0^2(\omega)}, \|\partial_t\eta(T_k)\|_{L_0^2(\omega)}, \int_{T_k}^s \exp(s-u)(\|\mathbf{f}\|_{L^2(\Omega_\eta(u))}(u) + \|g\|_{L^2(\omega)}(u))du\Big),$$

where $\tilde{C}$ is positive and nondecreasing with respect to its arguments, and $C(T_k,s) \leq C(0,s)$. This a priori estimate shows that if we let

$$\tau_k = \min\{1, (m_k/2C(T_k,T_k+1))^{\frac{1}{\lambda}}\},$$

we can build a solution starting from $\mathbf{u}(T_k)$ and $\eta(T_k)$, $\partial_t\eta(T_k)$ up to the time $T_k+\tau_k$ (this corresponds to choosing $\alpha = m_k/2$ in the construction of the solution). The time T_{k+1} is chosen close to $T_k+\tau_k$ (in $[T_k+\tau_k/2, T_k+\tau_k]$), in order to have also $\eta(T_k) \in H_0^2(\omega)$, $\partial_t\eta(T_{k+1}) \in L^2(\omega)$, and $\mathbf{u}(T_{k+1}) \in L^2(\Omega_\eta(T_{k+1}))$.

If the sequence $(T_k)_{k\geq 1}$ is infinite, we let $T^* = \sup_k T_k$. If $T < +\infty$, it must be that $m = \min_{[0,T^*]\times\overline{\omega}}(1+\eta) = 0$. Otherwise, we have $m_k \geq m$ for all k, and hence $\tau_k \geq \min\{1, (m/2C(0,T^*))^{\frac{1}{\lambda}}\} > 0$. But $T_{k+1} - T_k \geq \tau_k/2$ and goes to zero, a contradiction. This achieves the proof of the theorem.

5. Conclusion. We have proved the existence of at least one weak solution for a three-dimensional fluid-plate interaction problem without any (artificial) viscosity of the structure.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Pure Appl. Math. 65, Academic Press, New York, 1975.
[2] M. Boulakia, *Existence of weak solutions for the motion of an elastic structure in an incompressible viscous fluid*, C. R. Math. Acad. Sci. Paris, 336 (2003), pp. 985–990.
[3] H. Brezis, *Analyse Fonctionnelle, Théorie et applications* [*Theory and applications*], Masson, Paris, 1983.

[4] A. Chambolle, B. Desjardins, M. J. Esteban, and C. Grandmont, *Existence of weak solutions for the unsteady interaction of a viscous fluid with an elastic plate*, J. Math. Fluid Mech., 7 (2005), pp. 368–404.
[5] C. Conca, J. H. San Martín, and M. Tucsnak, *Existence of solutions for the equations modelling the motion of a rigid body in a viscous fluid*, Comm. Partial Differential Equations, 25 (2000), pp. 1019–1042.
[6] D. Coutand and S. Shkoller, *Motion of an elastic solid inside an incompressible viscous fluid*, Arch. Ration. Mech. Anal., 176 (2005), pp. 25–102.
[7] D. Coutand and S. Shkoller, *The interaction between quasilinear elastodynamics and the Navier-Stokes equations*, Arch. Ration. Mech. Anal., 179 (2006), pp. 303–352.
[8] H. B. da Veiga, *On the existence of strong solutions to a coupled fluid-structure evolution problem*, J. Math. Fluid Mech., 6 (2004), pp. 21–52.
[9] M. Dauge, *Stationary Stokes and Navier-Stokes systems on two- or three-dimensional domains with corners, Part* I. *Linearized equations*, SIAM J. Math. Anal., 20 (1989), pp. 74–97.
[10] B. Desjardins and M. J. Esteban, *Existence of weak solutions for the motion of rigid bodies in a viscous fluid*, Arch. Ration. Mech. Anal., 146 (1999), pp. 59–71.
[11] B. Desjardins and M. J. Esteban, *On weak solutions for fluid-rigid structure interaction: Compressible and incompressible models*, Comm. Partial Differential Equations, 25 (2000), pp. 1399–1413.
[12] B. Desjardins, M. J. Esteban, C. Grandmont, and P. Le Tallec, *Weak solutions for a fluid–structure interaction model*, Rev. Mat. Complut., 14 (2001), pp. 523–538.
[13] E. Feireisl, *On the motion of rigid bodies in a viscous compressible fluid*, Arch. Ration. Mech. Anal., 167 (2003), pp. 281–308.
[14] G. P. Galdi, *On the motion of a rigid body in a viscous liquid: A mathematical analysis with applications*, in Handbook of Mathematical Fluid Dynamics, Vol. I, North-Holland, Amsterdam, 2002, pp. 653–791.
[15] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations, Theory and Algorithms*, Springer-Verlag, Berlin, 1986.
[16] C. Grandmont and Y. Maday, *Existence for an unsteady fluid-structure interaction problem*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 609–636.
[17] M. D. Gunzburger, H.-C. Lee, and G. A. Seregin, *Global existence of weak solutions for viscous incompressible flows around a moving rigid body in three dimensions*, J. Math. Fluid Mech., 2 (2000), pp. 219–266.
[18] K.-H. Hoffmann and V. N. Starovoitov, *On a motion of a solid body in a viscous fluid. Two-dimensional case*, Adv. Math. Sci. Appl., 9 (1999), pp. 633–648.
[19] J.-L. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications, Vol.* I, Grundlehren Math. Wiss. 181, Springer-Verlag, New York, 1972 (in English).
[20] J.-L. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications. Vol.* I, Grundlehren Math. Wiss. 181, Springer-Verlag, New York, 1972 (in English).
[21] J. A. San Martín, V. Starovoitov, and M. Tucsnak, *Global weak solutions for the two-dimensional motion of several rigid bodies in an incompressible viscous fluid*, Arch. Ration. Mech. Anal., 161 (2002), pp. 113–147.
[22] D. Serre, *Chute libre d'un solide dans un fluide visqueux incompressible, Existence*, Japan J. Appl. Math., 4 (1987), pp. 99–110.
[23] J. Simon, *Compact sets in the space* $L^p(0,T;B)$, Ann. Mat. Pura Appl., 146 (1987), pp. 65–96.
[24] T. Takahashi, *Analysis of strong solutions for the equations modeling the motion of a rigid-fluid system in a bounded domain*, Adv. Differential Equations, 8 (2003), pp. 1499–1532.
[25] T. Takahashi and M. Tucsnak, *Global strong solutions for the two dimensional motion of a rigid body in a viscous fluid*, J. Math. Fluid Mech., 6 (2004), pp. 53–77.
[26] R. Temam, *Navier-Stokes Equations, Theory and Numerical Analysis*, Stud. Math. Appl. 2, North-Holland, Amsterdam, 1977.

SIAM J. MATH. ANAL.
Vol. 40, No. 2, pp. 738–753

TRANSMISSION EIGENVALUES*

LASSI PÄIVÄRINTA† AND JOHN SYLVESTER‡

Abstract. The scattering of a time-harmonic plane wave in an inhomogeneous medium is modeled by the scattering problem for the Helmholtz equation. A transmission eigenvalue is a wavenumber at which the scattering operator has a nontrivial kernel or cokernel. Because many sampling methods for locating scatterers succeed only at wavenumbers that are not transmission eigenvalues, they have been studied for some time. Nevertheless, the existence of transmission eigenvalues has previously been proved only for radial scatterers. In this paper, we prove existence for scatterers without radial symmetry.

Key words. inverse scattering, Helmholtz equation, inverse problems, transmission eigenvalues

AMS subject classifications. 81U40, 35P25

DOI. 10.1137/070697525

1. Introduction.

The scattering of a time-harmonic plane wave in an inhomogeneous medium is modeled by the scattering problem for the Helmholtz equation. The *total wave* u satisfies the perturbed Helmholtz equation

$$\left(\Delta + k^2(1+m)\right) u = 0 \quad \text{in } \mathbb{R}^n. \tag{1.1}$$

The function $m(x)$ denotes the perturbation of the index of refraction from the constant background medium; i.e., $n^2(x) = 1 + m(x)$. We insist that $-1 < m(x)$ be compactly supported and bounded. The relative (far field) scattering operator, s^+, compares the asymptotics of solutions of the free Helmholtz equation to those of (1.1). Both the linear sampling method and the factorization method use the range of this operator to find the support of the scatterer m. These methods are known to succeed at wavenumbers k for which the range of that operator is dense among all far field patterns (i.e., dense in $L^2(S^{n-1})$). If there exists a bounded domain D that contains the support of $m(x)$, and the wavenumber k is not a *transmission eigenvalue* as defined below, then the range of the scattering operator is dense [5].

DEFINITION 1. *A wavenumber k is called a transmission eigenvalue if there exists a nontrivial pair (v, w) solving*

$$\Delta w + k^2 n^2(x) w = 0 \quad \textit{in} \quad D, \tag{1.2}$$

$$\Delta v + k^2 v = 0 \quad \textit{in} \quad D, \tag{1.3}$$

$$w = v, \ \frac{\partial w}{\partial \nu} = \frac{\partial v}{\partial \nu} \quad \textit{on} \quad \partial D. \tag{1.4}$$

If D is not smooth enough, we replace (1.4) with the condition that $u - v \in H_0^2(D)$. Under the conditions that $m > 0$ or $m < 0$ on its support, it has been shown that the set of *transmission eigenvalues* is at most discrete [4], [12], but existence has

*Received by the editors July 17, 2007; accepted for publication (in revised form) April 4, 2008; published electronically July 18, 2008.
http://www.siam.org/journals/sima/40-2/69752.html

†Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland (ljp@rni.helsinki.fi). This author's research was supported by a grant from the Academy of Finland.

‡Department of Mathematics, University of Washington, Seattle, WA 98195 (sylvest@u.washington.edu). This author's research was supported by ONR grant N00014-05-1-0716 and NSF grant DMS-0355455.

been established only for m, which depends only on the radius [6]. Under certain conditions, knowledge of the transmission eigenvalues uniquely determines a radial scatterer [9], [10]. For nonradial scatterers, transmission eigenvalues have also been used to infer simple properties of the scatterer [3].

Under the hypothesis that the infimum of $|m|$ is large enough, we prove existence of transmission eigenvalues, as well as upper and lower bounds on the first transmission eigenvalue. The existence and upper bounds are new; the lower bounds are results from [7] and [3].

In [7], we showed that the following three conditions were equivalent. Notice that (1.5) differs from (1.3) in that the condition below requires that v solve the free Helmholtz equation in all or $\mathbb{R}^n$ rather than just in D. Such v which can be represented as superpositions of plane waves with L^2 densities are called Herglotz wave-functions.

1. There exists a nontrivial pair (v, w) solving

$$\begin{aligned} \Delta w + k^2 n^2(x) w = 0 \quad &\text{in} \quad D, \\ \Delta v + k^2 v = 0 \quad &\text{in} \quad \mathbb{R}^n, \\ w = v,\ \frac{\partial w}{\partial \nu} = \frac{\partial v}{\partial \nu} \quad &\text{on} \quad \partial D. \end{aligned} \tag{1.5}$$

2. There exists a nontrivial $\mu^0 \in \ker\ s^+$.
3. There exists a nontrivial $\mu^0 \in \text{coker}\ \ s^+$.

In the case that v is a Herglotz wave-function, its asymptotic expansion (its far field) belongs to both the kernel and the cokernel of the far field scattering operator. We will show below that, for scatterers supported in a compact set D, the far field scattering operator has a natural extension, and that *transmission eigenvalues* are exactly the wavenumbers for which this natural extension has a kernel or cokernel.

2. The Helmholtz equation and the scattering operator. The scattering operator relates the solutions of (1.1) to solutions of the free Helmholtz equation in all of $\mathbb{R}^n$:

$$\left(\Delta + k^2\right) u^0 = 0 \quad \text{in } \mathbb{R}^n. \tag{2.1}$$

We refer to solutions of (2.1) with finite B^*-norm, defined by

$$||u^0||_{B^*} = \sup_{R>0} \frac{1}{\sqrt{R}} ||u^0||_{L^2(B_R)},$$

as *incident waves* or *free waves*. An *outgoing wave* is a solution to the Helmholtz equation with a compactly supported source f

$$\left(\Delta + k^2\right) v^+ = f \quad \text{in } \mathbb{R}^n \tag{2.2}$$

that satisfies the Sommerfeld radiation condition

$$\lim_{r\to\infty} r^{\frac{n-1}{2}} \left(\frac{\partial v^+}{\partial r} - ikv^+ \right) = 0 \tag{2.3}$$

or, equivalently (for $k > 0$), a *limiting absorption principle*

$$v^+ = \lim_{\varepsilon \downarrow 0} v_\varepsilon^+, \tag{2.4}$$

where v_ε^+ is the unique solution to (2.2) with $k^2 \in \mathbb{R}$ replaced by $k^2 + i\varepsilon$ (see, e.g., section 4 of [1]). We could also define an outgoing wave as a solution to

$$\left(\Delta + k^2(1+m)\right) w^+ = g \tag{2.5}$$

with a compactly supported g, and satisfying (2.3) or (2.4). Because m is compactly supported, the definition based on (2.5) and that based on (2.2) coincide. That is, an outgoing solution v^+ to (2.2) is also an outgoing solution w^+ to (2.5) with $g = f - mv^+$.

Existence and uniqueness of outgoing solutions to (2.5) were proved by Agmon in weighted L^2 spaces[1] [1]. Theorem 2 below is a special case of results in [2], and parts of Theorems 3 and 4 are special cases of results in [2] and section 14 of [8].

Define B to be the completion of $C_0^\infty(\mathbb{R}^n)$ in the B-norm

$$||f||_B = ||f||_{L^2(|x|\in[0,1])} + \sum_{j=1}^{\infty} \frac{1}{\sqrt{2^j}} ||f||_{L^2(|x|\in[2^j,2^{j+1}])}.$$

THEOREM 2. *For every compactly supported g, there exists a unique outgoing solution to* (2.5)*, with*

$$||w^+||_{B^*} \leq C||g||_B,$$

where the constant C depends on m and k.

Because compactly supported functions are dense in B, the correspondence in Theorem 2 defines a bounded map

$$G_m^+ : B \longrightarrow B^*$$

mapping $g \in B$ to w^+. In the rest of the paper, whenever we refer to waves, we mean subspaces of B^*:

1. B^0 is the subspace of incident waves, i.e., solutions to (2.1).
2. B^m is the subspace of total waves, i.e., solutions to (1.1).
3. B^+ is the subspace of outgoing waves, the range of G_0^+.[2]

Both B^0 and B^m are closed in the B^* topology. One way to see this is to note that B^* convergence implies convergence in the sense of tempered distributions so that any u in the closure of B^0 or B^m must satisfy (2.1) or (1.1), respectively, in the sense of distributions. As Schwartz class functions are dense in B, the equations are satisfied in the B^* sense as well. The plane waves, $e^{ik\Theta\cdot x}$, are not in B^0. We shall note in Theorem 4 below that B^0 consists of the Herglotz wave-functions, solutions to (2.1) which have square integrable far fields. The far fields of the plane waves are Dirac deltas. The subspace B^+ is not closed in the B^* topology. In particular, every function in B^+ is in $H^2_{loc}(\mathbb{R}^n)$ and satisfies the radiation condition (2.3), and every compactly supported function in $H^2_{loc}(\mathbb{R}^n)$ belongs to B^+. Because we have defined B^+ as the range of G_0^+, which is injective, B^+ is a Banach space with norm $||G_0^+ f||_{B^+} := ||f||_B$.

A straightforward consequence of Theorem 2 is the correspondence between incident and total waves.

---

[1] $||f||_{L^2_\delta} = ||(1+|x|^2)^{\frac{\delta}{2}} f||_{L^2}$.

[2] This is also the range of G_m^+ for any bounded compactly supported m.

THEOREM 3. *Every total wave has a unique decomposition into an incident wave plus a scattered wave, and every incident wave has a unique decomposition as a total wave minus a scattered wave:*

$$v^m = v^0 + v^+, \tag{2.6}$$

$$u^0 = u^m - u^+. \tag{2.7}$$

Moreover, the scattering map $\mathcal{S}$, defined as $u^0 \mapsto u^m$, is an isomorphism from B^0 onto B^m.

Proof. We prove the second assertion first. Any u_0 that solves (2.1) also solves

$$\left(\Delta + k^2(1+m)\right) u^0 = k^2 m u^0.$$

Let u^+ be the unique outgoing solution to (2.5) with $g = -k^2 m u^0$. Note that

$$||u^+||_{B^*} \leq C_1 ||g||_B \leq k^2 C_1 ||m u^0||_B \leq k^2 C_2 ||u^0||_{B^*},$$

where both constants depend on an upper bound for m and the size of its support.

Defining $u^m = u^0 + u^+$ and noting that it satisfies (1.1) gives decomposition (2.7) and the estimate

$$||u^m||_{B^*} \leq C_3 ||u^0||_{B^*}. \tag{2.8}$$

If $u^0 = w^m - w^+$ is another such decomposition, then w^+ must also satisfy (2.5) with $g = -k^2 m u^0$, but (2.5) has a unique outgoing solution, so $w^+ = u^+$ and $w^m = u^m$.

Similarly, any v^m solving (1.1) is a solution to

$$\left(\Delta + k^2\right) v^m = -k^2 m v^m.$$

Let v^+ be the unique outgoing solution to (2.2) with $f = -k^2 m u^m$, and set $v^0 = v^m - v^+$. Uniqueness follows as in the paragraph above, as does the estimate

$$||u^0||_{B^*} \leq C_3 ||u^m||_{B^*}. \tag{2.9}$$

The existence and uniqueness of the two decompositions (2.6) and (2.7) along with the estimates (2.8) and (2.9) justify the last statement in the theorem—that the scattering map is an isomorphism. □

In order to see the relationship between the *scattering operator* we have defined above and the scattering operator defined on far fields, we need to discuss asymptotics.

THEOREM 4. *Let $u^0 \in B^0$, $u^+ \in B^+$, and $u^m \in B^m$; then, in spherical coordinates $x = r\Theta$ for large r,*

$$u^0 \sim \mu^0(\Theta) \frac{e^{ikr}}{(ikr)^{\frac{n-1}{2}}} + \mu^0(-\Theta) \frac{e^{-ikr}}{(-ikr)^{\frac{n-1}{2}}}, \tag{2.10}$$

$$u^+ \sim \mu^+(\Theta) \frac{e^{ikr}}{(ikr)^{\frac{n-1}{2}}}, \tag{2.11}$$

$$u^m \sim \left(\mu^m(\Theta) + \gamma(\Theta)\right) \frac{e^{ikr}}{(ikr)^{\frac{n-1}{2}}} + \mu^m(-\Theta) \frac{e^{-ikr}}{(-ikr)^{\frac{n-1}{2}}}. \tag{2.12}$$

Moreover, the mappings

$$b^0 : B^0 \longrightarrow L^2(S^{n-1}),$$
$$b^m : B^m \longrightarrow L^2(S^{n-1})$$

defined by $u^0 \mapsto \mu^0$ *and by* $u^m \mapsto \mu^m$ *are isomorphisms. The mapping*

$$b^+ : B^+ \longrightarrow L^2(S^{n-1})$$

defined by $u^+ \mapsto \mu^+$ *is surjective.*[3] *Every compactly supported function in* $H^2(\mathbb{R}^n)$ *belongs to its kernel and any function in its kernel is compactly supported (Rellich's lemma).*

Sketch of proof. The operator $(b^0)^{-1}$is known as the Herglotz operator. We can start with any $\mu_0 \in L^2(S^{n-1})$ and define

$$u^0 = \mathcal{H}\mu^0 = \int_{S^{n-1}} e^{ik\Theta\cdot x}\mu^0(\Theta)dS_\Theta.$$

Noting that any $u^0 \in B^0$ is the inverse Fourier transform of a distribution supported on the sphere $|\xi|^2 - k^2 = 0$ shows that u^0 must have this form, but it requires an estimate [2] to see that $\mu^0 \in L^2$. A stationary phase calculation shows that $\mathcal{H}\mu^0$ has the asymptotics (2.10) when μ^0 is smooth. Again we refer the reader to [2] for the estimate that $||u||_{B^*} \leq C||\mu^0||_{L^2(S^{n-1})}$.

A similar Fourier transform calculation combined with the limiting absorption principle, or a calculation of the asymptotics of the outgoing Green's function, gives (2.11). Alternatively, we may note that $u^+(x,k) - u^+(x,-k)$ belongs to B^0 and deduce (2.11) from (2.10).[4] The surjectivity of b^+ then follows from the surjectivity of b^0.

Because of the decomposition $u^m = u^0 + u^+$ in Theorem 3, (2.12) follows from (2.10) and (2.11). Rellich's lemma and unique continuation imply that b^0 and b^m are injective, as well as the final statement in the theorem. □

We refer to the large r asymptotics as the *far fields* of the corresponding waves; e.g., the far field of u^0 is μ^0, the far field of u^+ is μ^+, and the far field of u^m is μ^m. We use (2.12) to define the far field (relative scattering) operator

$$s^+ : L^2(S^{n-1}) \longrightarrow L^2(S^{n-1})$$

by

$$\mu^m \;\mapsto\; \gamma. \tag{2.13}$$

The wave scattering operator $\mathcal{S}$ and the far field operator s^+ are closely related.

LEMMA 5. *Let* u^0 *and* w^0 *belong to* B^0*, with far fields* μ^0 *and* ω^0*, respectively. Then*

$$\int_{\mathbb{R}^n} \overline{u^0} m\mathcal{S}w^0 = \frac{-2i}{k^n}\int_{S^{n-1}} \overline{\mu^0}s^+\omega^0.$$

[3]The surjectivity of b^+, b^0, and b^m is perhaps the main reason for replacing the L^2_δ spaces in [1] with the Besov spaces, B and B^*, of [2].

[4]Changing the sign of k reverses the sign of ϵ in the limiting absorption principle and changes the sign of the second term in the Sommerfeld radiation condition, thus specifying the unique *incoming*, rather than the *outgoing*, solution.

Proof.

$$\int_{\mathbb{R}^n} \overline{u^0} k^2 m \mathcal{S} w^0 = \int_{\mathbb{R}^n} \overline{u^0} k^2 m w^m,$$

where $w^m = w^0 + w^+$, as in (2.6),

$$\begin{aligned} &= -\int_{\mathbb{R}^n} \overline{u^0} \left(\Delta + k^2\right) w^m \\ &= -\int_{\mathbb{R}^n} \overline{u^0} \left(\Delta + k^2\right) w^+ \\ &= -\lim_{R\to\infty} \int_{|x|<R} \overline{u^0} \left(\Delta + k^2\right) w^+ \\ &= \lim_{R\to\infty} \int_{|x|=R} \frac{\partial \overline{u^0}}{\partial \nu} w^+ - \overline{u^0} \frac{\partial w^+}{\partial \nu}. \end{aligned} \tag{2.14}$$

Making use of the asymptotics in (2.11) and (2.12) gives

$$= \frac{-2i}{k^{n-2}} \int_{S^{n-1}} \overline{\mu^0} \omega^+, \tag{2.15}$$

where ω^+ denotes the far field of w^+,

$$= \frac{-2i}{k^{n-2}} \int_{S^{n-1}} \overline{\mu^0} s^+ \omega^0. \qquad \square$$

A consequence of Lemma 5 is a natural definition of the relative scattering operator, which does not explicitly use asymptotics.

THEOREM 6. *If we define*

$$\mathcal{S}^+ : B^0 \longrightarrow B^{0*}$$

by

$$w^0 \overset{\mathcal{S}^+}{\longmapsto} \frac{-k^n}{2i} \left\langle m \mathcal{S} w^0, \cdot \right\rangle, \tag{2.16}$$

then

$$s^+ \;=\; \mathcal{H}^* \mathcal{S}^+ \mathcal{H}. \tag{2.17}$$

Proof.

$$\begin{aligned} &\langle \mathcal{H}^* \mathcal{S}^+ \mathcal{H} \omega^0, \overline{\mu^0} \rangle \\ &\qquad = \langle \mathcal{S}^+ \mathcal{H} \omega^0, \overline{\mathcal{H} \mu^0} \rangle \\ &\qquad = \frac{-k^n}{2i} \int_{\mathbb{R}^n} \overline{\mathcal{H} \mu^0} m \mathcal{S} \mathcal{H} \omega^0 \\ &\qquad = \frac{-k^n}{2i} \int_{\mathbb{R}^n} \overline{u^0} m \mathcal{S} w^0 \\ &\qquad = \int_{S^{n-1}} \overline{\mu^0} s^+ \omega^0 \\ &\qquad = \langle s^+ \omega^0, \overline{\mu^0} \rangle. \qquad \square \end{aligned}$$

Remark 7. Because $B^0 \subset L^2_{-\delta}$ for any $\delta > \frac{1}{2}$, any $l \in {B^0}^*$ has a (nonunique) extension to an element of $L^2_\delta = {L^2_{-\delta}}^*$, so elements of ${B^0}^*$ can be represented as functions (and called sources).

3. A generalized scattering operator. We describe an incident wave $\mathcal{H}\alpha$ as illuminating the scatterer m. If we use the far field operator s^+, the illumination must always come from the *sphere at infinity*. Many useful sources of illumination are generated by sources outside the scatterer. The waves generated by such sources are never incident waves, although they can be approximated by incident waves on certain compact sets. Solutions to the transmission eigenvalue problem (1.2)–(1.4) are not incident waves, so they do not have a direct interpretation in terms of the far field scattering operator. They do, however, span exactly the kernel of the scattering operator we will define below.

If $m \in L^\infty$ is supported in a bounded domain D, Theorem 2 tells us that we can find a unique $u^+ \in B^+$ solving

$$\left(\Delta + k^2(1+m)\right) u^+ = k^2 m u^0$$

for any $u^0 \in L^2(D)$. It follows that $u^m = u^0 + u^+ \in L^2(D)$. Thus the scattering operator $\mathcal{S}$ has a natural extension,

$$\mathcal{S}_D : B^0_D \longrightarrow B^m_D,$$

where we use the definitions

$$B^0_D = \left\{ w \in L^2(D) \middle| \left(\Delta + k^2\right) w = 0 \text{ in } D \right\},$$
$$B^m_D = \left\{ w \in L^2(D) \middle| \left(\Delta + k^2(1+m)\right) w = 0 \text{ in } D \right\}.$$

The relative scattering operator $\mathcal{S}^+$ has a similar extension:

$$\mathcal{S}^+_D : B^0_D \longrightarrow {B^0_D}^*,$$
$$w^0 \overset{\mathcal{S}^+_D}{\longmapsto} \frac{k^n}{2i} \left\langle m \mathcal{S}_D w^0, \cdot \right\rangle.$$

The scattering and relative scattering operators $\mathcal{S}_D$ and $\mathcal{S}^+_D$ are extensions of $\mathcal{S}$ and $\mathcal{S}^+$ in the sense that, for $u^0, w^0 \in B^0$, then $\mathcal{S}_D u^0$ is the restriction of $\mathcal{S}u^0$ to D and $\langle \mathcal{S}^+_D u^0, w^0 \rangle = \langle \mathcal{S}^+ u^0, w^0 \rangle$.

4. The interior transmission problem. Let D be a bounded domain and $\operatorname{supp} m \subset D$. We will use the notation

$$P^0 = \left(\Delta + k^2\right),$$
$$P^m = \left(\Delta + k^2(1+m)\right),$$
$$H^k(D) = \left\{ u \in L^2(D) \,\middle|\, D^\alpha u \in L^2(D) \quad \forall\ |\alpha| \le k \right\},$$
$$H^k_0(D) = \text{the completion of } C^\infty_0(D) \text{ in } H^k(D).$$

Definition 8. *We say that a wavenumber k is a D-transmission eigenvalue of $m \in L^\infty(D)$ if any of the equivalent conditions in Theorem 9 below are satisfied.*

Theorem 9. *The following are equivalent:*

1. *There exist nontrivial $u^0 \in B^0_D$ and $u^m \in B^m_D$ with $u^0 - u^m \in H^2_0(D)$.*[5]

[5]This is a restatement of (1.2)–(1.4). The condition that $u^0 \in B^0_D$ is (1.3), $u^m \in B^m_D$ is (1.2), and $u^0 - u^m \in H^2_0(D)$ is (1.4).

2. *There exists nontrivial* $u^m \in B_D^m$ *such that the unique outgoing solution* u^+ *to*

$$P^0 u^+ = -k^2 m u^m \tag{4.1}$$

belongs to $H_0^2(D)$.
3. *There exist nontrivial* $u^m \in B_D^m$ *and some* $v \in H_0^2(D)$ *satisfying* (4.1).
4. *There exists nontrivial* $u^0 \in B_D^0$ *such that the unique outgoing solution* u^+ *to*

$$P^m u^+ = -k^2 m u^0 \tag{4.2}$$

belongs to $H_0^2(D)$.
5. *There exist nontrivial* $u^0 \in B_D^0$ *and some* $v \in H_0^2(D)$ *satisfying* (4.2).
6. *There exists nontrivial* $u^0 \in \ker \mathcal{S}_D^+$.
7. *There exists nontrivial* $u^0 \in \operatorname{coker} \mathcal{S}_D^+$.

Proof. We first show that items 1–5 are equivalent.

Condition 2 obviously implies 3, but any $H_0^2(D)$ solution, v, to (4.1) extended to be zero in $\mathbb{R}^n \setminus D$ is outgoing. Since the outgoing solution to (4.1) is unique, $v = u^+$, so 3 implies 2.

Similarly, 4 obviously implies 5, but uniqueness of the outgoing solution to (4.2) implies that any $H_0^2(D)$ solution to (4.2), extended to be zero outside D, must be u^+, so 5 implies 4.

Because Theorem 3 gives a unique decomposition,

$$u^m = u^0 + u^+, \tag{4.3}$$

the unique outgoing solution to (4.1) is also the unique outgoing solution to (4.2). Thus 4 and 2 are equivalent.

The same decomposition shows that $u^+ = u^m - u^0$, so the left-hand side is in $H_0^2(D)$ if and only if the right-hand side is; hence 1 is equivalent to 2.

The equivalence of items 6 and 2 is based on a calculation. Let u^0 and w_0 belong to B_D^0,

$$\begin{aligned} \langle \overline{w^0}, \mathcal{S}_D^+ u^0 \rangle &= -\frac{k^n}{2i} \left\langle \overline{w^0}, m \mathcal{S}_D u^0 \right\rangle \\ &= -\frac{k^n}{2i} \int_D \overline{w^0} m u^m, \end{aligned} \tag{4.4}$$

where u^m in the line above and u^+ in the line below are those uniquely related to u^0 by (4.3) and Theorem 3,

$$= -\frac{k^{n-2}}{2i} \int_D \overline{w^0} P^0 u^+ \tag{4.5}$$

$$= \frac{k^{n-2}}{2i} \int_{\partial D} \frac{\partial \overline{w^0}}{\partial \nu} u^+ - \overline{w^0} \frac{\partial u^+}{\partial \nu}. \tag{4.6}$$

The right-hand side of (4.6) is clearly zero for every w^0 if $u^+ \in H_0^2(D)$, so 2 implies 6. To see the converse, choose $w^0 = \mathcal{H}\omega^0$ (i.e., $w^0 \in B^0$ with asymptotics as in (2.10)). Every $u^+ \in B^+$ has asymptotics as in (2.11), so we may continue the previous calculation,

$$\begin{aligned} \langle \overline{w^0}, \mathcal{S}_D^+ u^0 \rangle &= \frac{k^{n-2}}{2i} \lim_{R \to \infty} \int_{B_R} \frac{\partial \overline{w^0}}{\partial \nu} u^+ - \overline{w^0} \frac{\partial u^+}{\partial \nu} \\ &= \int_{S^{n-1}} \overline{\omega^0} \mu^+. \end{aligned}$$

We conclude that, if the left-hand side vanishes for every w^0, so does the right-hand side for every ω^0, so $\mu^+ \equiv 0$. Now Rellich's lemma and unique continuation tell us that $u^+ \in H_0^2(D)$, so 6 implies 2.

Verifying the equivalence of items 7 and 2 requires a similar computation,

$$\begin{aligned}\frac{2i}{k^{n-2}}\langle u^0, \mathcal{S}_D^+ w^0\rangle &= -\int_D u^0 k^2 m w^m \\ &= \int_D u^0 P^0 w^m \\ &= \int_{\partial D} \frac{\partial u^0}{\partial \nu} w^m - u^0 \frac{\partial w^m}{\partial \nu} \\ &= \int_{\partial D} \frac{\partial u^0}{\partial \nu} w^m - u^0 \frac{\partial w^m}{\partial \nu} - \int_{\partial D} \frac{\partial u^m}{\partial \nu} w^m - u^m \frac{\partial w^m}{\partial \nu},\end{aligned}$$

because the second integral on the right is always zero. Combining the two terms gives

$$\frac{2i}{k^{n-2}}\langle u^0, \mathcal{S}_D^+ w^0\rangle = -\int_{\partial D} \frac{\partial u^+}{\partial \nu} w^m - u^+ \frac{\partial w^m}{\partial \nu}. \tag{4.7}$$

Now 2 implies that the right-hand side of (4.7) is zero, so the left-hand side is zero for every w^0, which implies 7. If we choose $w^0 \in B^0$ and continue the calculation,

$$= \int_{S^{n-1}} \mu^+ \omega^m. \tag{4.8}$$

Item 7 implies that the integral in (4.8) vanishes for every ω^m, so $\mu^+ \equiv 0$. Rellich's lemma and unique continuation then guarantee that $u^+ \in H_0^2(D)$, which implies item 2. □

Note that if $\operatorname{supp} m \subset \tilde{D} \subset D$, then $\mathcal{S}_{\tilde{D}}$ is an extension of $\mathcal{S}_D$. The smaller we make D, the larger we make B_D^0, the domain of the operator $\mathcal{S}_D$. Therefore, if k is a D-transmission eigenvalue of m, then k is also a $\tilde{D}$-transmission eigenvalue.

5. Existence of transmission eigenvalues. In this section we restrict our attention to the case that $D = \operatorname{supp} m$. We assume further that m is bounded away from zero in D. The theorem below was first proved in [12].

THEOREM 10. *If $|m| > \delta > 0$ in D, then k is a D-transmission eigenvalue if and only if there exists $u^+ \in H_0^2(D)$ satisfying*

$$P^m \frac{1}{m} P^0 u^+ = 0. \tag{5.1}$$

Proof. We show that (5.1) is equivalent to item 2 in Theorem 9. If $u^+ \in H_0^2(D)$ satisfies

$$P^0 u^+ = -k^2 m u^m, \tag{5.2}$$

then

$$\frac{1}{m} P^0 u^+ = -k^2 u^m, \tag{5.3}$$

and

$$P^m \frac{1}{m} P^0 u^+ = 0.$$

To see the reverse implication, suppose that $u^+ \in H_0^2(D)$ satisfies (5.1) (recall that any $u^+ \in H_0^2(D)$ is outgoing), and define u^m so that (5.3) holds. It is a consequence of (5.1) that $u^m \in B_D^m$ so that (5.2) implies (4.1). □

Theorem 10 tells us that k is a D-transmission eigenvalue whenever the operator $P^m \frac{1}{m} P^0$ has a kernel in $H_0^2(D)$. We will investigate the existence of this kernel (5.1) by examining the spectrum of the operator as k^2 changes. We will make use of several equivalent formulas for $P^m \frac{1}{m} P^0$ which we list below. We will let $\tau = k^2$.

$$\begin{aligned} P^m \frac{1}{m} P^0 &= P^0 \frac{1}{m} P^m \\ &= \Delta \frac{1}{m} \Delta + \tau \left(\Delta \frac{1}{m} + \left(1 + \frac{1}{m} \right) \Delta \right) + \tau^2 \left(1 + \frac{1}{m} \right) \\ &= (\Delta + \tau) \frac{1}{m} (\Delta + \tau) + \tau (\Delta + \tau) \\ &= (\Delta + \tau(1+m)) \frac{1}{m} (\Delta + \tau(1+m)) - \tau (\Delta + \tau(1+m)) . \end{aligned} \tag{5.4}$$

The following lemma asserts that $P^m \frac{1}{m} P^0$, with the appropriate domain, defines a semibounded self-adjoint operator on $L^2(D)$.

LEMMA 11. *For $\tau \geq 0$, t_τ, defined by*

$$t_\tau(u) = \int_D \frac{1}{m} |(\Delta + \tau) u|^2 - \tau \int_D |\operatorname{grad} u|^2 + \tau^2 \int_D |u|^2 \tag{5.5}$$

with form domain $H_0^2(D)$, is a densely defined, closed, semibounded quadratic form on $L^2(D)$. T_τ, The unique densely defined self-adjoint operator associated to t_τ, T_τ, is equal to $P^m \frac{1}{m} P^0$ on its domain

$$D(T_\tau) = \left\{ u \in H_0^2(D) \quad \middle| \quad \frac{1}{m}(\Delta + \tau) \in H^2(D) \right\}. \tag{5.6}$$

Proof. We state without proof that $H_0^2(D)$ is dense in $L^2(D)$. To see that t_τ is semibounded, we write

$$\begin{aligned} t_\tau(u) &= \int_D \frac{1}{m} |(\Delta + \tau) u|^2 + \tau \int_D \overline{u}((\Delta + \tau) u) \\ &\geq \frac{1}{\sup(m)} ||(\Delta + \tau) u||^2 - \tau ||(\Delta + \tau) u|| ||u|| \\ &\geq \left(\frac{1}{\sup(m)} - \tau\varepsilon \right) ||(\Delta + \tau) u||^2 - \frac{\tau}{\varepsilon} ||u||^2 \\ &\geq \frac{1}{2\sup(m)} ||(\Delta + \tau) u||^2 - 2\sup(m)\tau^2 ||u||^2 \end{aligned} \tag{5.7}$$

after choosing $\varepsilon = \frac{1}{2\tau \sup(m)}$. We record for later use the consequence of (5.7) that

$$||(\Delta + \tau) u||^2 \leq 2\sup(m) t_\tau(u) + (2\sup(m)\tau)^2 ||u||^2. \tag{5.8}$$

Every densely defined semibounded quadratic form defines a unique self-adjoint operator [11, page 278], T_τ, with domain the set of $u \in H_0^2(D)$ such that there is an $f \in L^2(D)$ with

$$t(v, u) = (v, f)$$

for all $v \in H_0^2(D)$, where $t(v,u)$ is the bilinear form

$$\begin{aligned} t(v,u) &= \int_D (\overline{(\Delta+\tau)\, v} \frac{1}{m} (\Delta+\tau)\, u + \tau \overline{v} (\Delta+\tau)\, u \\ &= \int_D \overline{v} \left((\Delta+\tau) \frac{1}{m} (\Delta+\tau) \right) u + \tau \int_D \overline{v}\, (\Delta+\tau)\, u \\ &= (v, T_\tau u)\,, \end{aligned}$$

where the second and third equalities hold for all $u \in D(T_\tau)$ and illustrate that $T_\tau = P^m \frac{1}{m} P^0$ with $D(T_\tau)$, as asserted in (5.6). □

LEMMA 12. *T_τ has discrete spectrum which depends continuously on τ.*

Proof. Because T_τ is semibounded and $H_0^2(D)$ is compactly embedded in $L^2(D)$, T_τ has a compact resolvent and therefore discrete spectrum. For m's that are not smooth, the domains of T_τ may depend on τ. We give a direct proof of the continuity of the eigenvalues. We shall show below that, for all positive real σ and τ,

$$t_\sigma(u) \le (1 + M|\sigma-\tau|)\, t_\tau(u) + M(\tau^2+\sigma+1)|\sigma-\tau|\ ||u||^2, \tag{5.9}$$

where the constant M depends only on m. We recall the min-max characterization of the eigenvalues of a self-adjoint operator defined by a quadratic form [13, p. 71]

$$\lambda^n = \max_{W \in \mathcal{W}_n} \min_{\substack{u \in W \\ ||u||=1}} q(u), \tag{5.10}$$

where $\mathcal{W}_n$ denotes the codimension n subspaces of the form domain of q. An immediate consequence of (5.10) is that inequalities between quadratic forms imply the same inequalities for their ordered eigenvalues so that (5.9) implies

$$\lambda_\sigma^n \le (1 + M|\sigma-\tau|)\, \lambda_\tau^n + M(\tau^2+\sigma+1)|\sigma-\tau|$$

and, consequently,

$$\lambda_\sigma^n - \lambda_\tau^n \le |\sigma-\tau| M \left(\lambda_\tau^n + (\tau^2+\sigma+1) \right). \tag{5.11}$$

Because we may interchange σ and τ,

$$|\lambda_\sigma^n - \lambda_\tau^n| \le |\sigma-\tau| M \left(\max(\lambda_\tau^n, \lambda_\sigma^n) + 2(\tau^2+\sigma^2+1) \right). \tag{5.12}$$

First fix σ in (5.11), and set $\tau = 0$ to conclude that each λ_σ^n varies only over a compact set when σ varies over a compact set. Thus the maximum, $\max(\lambda_\tau^n, \lambda_\sigma^n)$, is bounded for σ and τ on compact sets, and therefore (5.12) proves continuity of the eigenvalues.

It remains only to prove (5.9). We begin by writing

$$\begin{aligned} t_\sigma(u) - t_\tau(u) &= (\sigma-\tau) \int \left(\overline{u} \frac{1}{m} (\Delta+\tau)\, u + \overline{(\Delta+\tau)\, u} \frac{1}{m} u + \tau^2 |u|^2 \right) \\ &\quad + (\sigma-\tau)^2 \int \left(\overline{u}\, (\Delta+\tau)\, u + \frac{1}{m} |u|^2 \right) \\ &\le |\sigma-\tau|(1+|\sigma-\tau|) M \left(||u||\ ||\, (\Delta+\tau)\, u|| + \tau^2 ||u||^2 \right), \end{aligned}$$

where M depends only on $\frac{1}{m}$. For any $\varepsilon > 0$

$$\le |\sigma-\tau|(1+|\sigma-\tau|) M \left(\varepsilon ||\, (\Delta+\tau)\, u||^2 + \left(\tau^2 + \frac{1}{\varepsilon} \right) ||u||^2 \right).$$

We make use of (5.8) to obtain with a different M

$$\leq |\sigma-\tau|(1+|\sigma-\tau|)M\left(\varepsilon t_\tau(u)+\left((\varepsilon+1)\tau^2+\frac{1}{\varepsilon}\right)||u||^2\right)$$
$$\leq |\sigma-\tau|M\left(t_\tau(u)+(\tau^2+\sigma+1)||u||^2\right)$$

after choosing $\frac{1}{\varepsilon}=1+|\sigma-\tau|$. ∎

LEMMA 13. *If*

$$\operatorname{sign}(m)\inf_{u\in H_0^2(D)}\frac{t_\tau(u)}{||u||^2}>0, \tag{5.13}$$

then τ is not a transmission eigenvalue. If there exists $u\in H_0^2(D)$ such that

$$\operatorname{sign}(m)\frac{t_\tau(u)}{||u||^2}\leq 0, \tag{5.14}$$

then there is a transmission eigenvalue $\tau^*\in[0,\tau]$.

Proof. The hypothesis (5.13) implies that the spectrum of T_τ is strictly positive or strictly negative; hence it has no kernel.

The hypothesis (5.14) implies that $\operatorname{sign}(m)T_\tau$ has at least one nonpositive eigenvalue. But $\operatorname{sign}(m)T_0$ is easily seen to be positive definite, so the lowest eigenvalue, which is a continuous function of τ, must have passed through zero for some $\tau^*\in[0,\tau]$. ∎

We will use a simple modification of Lemma 13 to show the existence of more than one transmission eigenvalue. We define the multiplicity of a transmission eigenvalue τ_* to be the multiplicity of 0 as an eigenvalue of T_{τ_*}.

LEMMA 14. *If there exists a $\tau>0$ and a p-dimensional subspace $V^p\in H_0^2(D)$ such that*

$$\operatorname{sign}(m)\frac{t_\tau(u)}{||u||^2}\leq 0$$

for all $u\in V^p$, then there are p-transmission eigenvalues, counting multiplicity, in $[0,\tau]$.

Proof. The hypothesis guarantees that t_τ has p negative eigenvalues, counting multiplicity. The continuity of the spectrum implies that each of those eigenvalues must pass through zero as τ^* decreases from τ to 0. Each time an eigenvalue passes through 0, the dimension of the negative definite subspace, V^p, decreases by the multiplicity of the zero eigenvalue, so the sum of the multiplicities of the transmission eigenvalues between 0 and τ must be at least p. ∎

We will need a few simple inequalities to prove the theorems to follow. We collect them in the lemma below.

LEMMA 15.

$$\lambda_0(D)=\inf_{u\in H_0^1(D)}\frac{\int_D|\operatorname{grad}u|^2}{\int_D|u|^2}=\inf_{u\in H_0^2(D)}\frac{\int_D|\operatorname{grad}u|^2}{\int_D|u|^2}>0, \tag{5.15}$$

$$\mu_0(D)=\inf_{u\in H_0^2(D)}\frac{\int_D|\Delta u|^2}{\int_D|u|^2}\geq\inf_{u\in H_0^1\cap H^2}\frac{\int_D|\Delta u|^2}{\int_D|u|^2}=\lambda_0(D)^2. \tag{5.16}$$

If $u\in H_0^2(D)$,

$$\lambda_0(D)\leq\frac{\int_D|\operatorname{grad}u|^2}{\int_D|u|^2}\leq\frac{\left(\int_D|\Delta u|^2\right)^{\frac{1}{2}}}{\left(\int_D|u|^2\right)^{\frac{1}{2}}}. \tag{5.17}$$

Proof. The first equality in (5.15) is the Rayleigh–Ritz characterization of the first Dirichlet eigenvalue. The second follows because $H_0^2(D)$ is dense in $H_0^1(D)$. The first equality in (5.16) is the Rayleigh–Ritz characterization of the lowest eigenvalue of the biharmonic operator with Dirichlet boundary conditions, the lowest eigenvalue of the *clamped plate*. The inequality holds because $H_0^2(D) \subset H_0^1(D) \bigcap H^2(D)$, so the first infimum must be larger. The second infimum is exactly the Rayleigh–Ritz characterization for the lowest eigenvalue of the Dirichlet Laplacian squared,[6] which is the square of the first Dirichlet eigenvalue, which proves the final equality in (5.16). The first inequality in (5.17) follows from the meaning of infimum and the second from integration by parts and the Cauchy–Schwarz inequality. ∎

THEOREM 16. *Suppose that $m > -1$ is a constant. If*

$$\tau \leq \min\left(1, \frac{1}{m+1}\right) \lambda_0(D), \tag{5.18}$$

then τ is not a transmission eigenvalue. If

$$\frac{(1+\frac{m}{2})^2}{1+m} \geq \frac{\mu_p}{\lambda_0^2} \geq 1, \tag{5.19}$$

where $\mu_p(D)$ is the $(p+1)$st clamped plate eigenvalue, then there are $p+1$ transmission eigenvalues τ with

$$\tau \leq \left(\frac{m+2}{m+1}\right) \frac{\lambda_0(D)}{2}. \tag{5.20}$$

Proof. It follows from (5.4) that

$$\begin{aligned} mt_\tau(u) &= ||\left(\Delta + \tau(1+m)\right) u||^2 - m\tau \int_D \overline{u}\left(\Delta + \tau(1+m)\right) u \\ &> m\tau\left[||\operatorname{grad} u||^2 - (1+m)\tau||u||^2\right] \\ &\geq m\tau\left[\lambda_0(D) - (1+m)\tau\right]||u||^2, \end{aligned}$$

which shows that, for $0 < m$ and $\tau \leq \frac{\lambda_0(D)}{1+m}$, t_τ is positive definite and therefore that τ is not a transmission eigenvalue. If $m < 0$, we express t_τ as in (5.5),

$$\begin{aligned} mt_\tau(u) &= ||\left(\Delta + \tau\right) u||^2 - m\tau||\operatorname{grad} u||^2 + m\tau^2||u||^2 \\ &\geq (-m)\tau||u||^2(\lambda_0(D) - \tau), \end{aligned}$$

which shows that t_τ is positive definite as long as $\tau < \lambda_0(D)$ and finishes the proof of the assertion that τ satisfying (5.18) is not a transmission eigenvalue.

[6] Functions in the domain of the square of the Dirichlet Laplacian must satisfy a second boundary condition, $\Delta u\big|_{\partial D} = 0$. Analogous to the case of the Neumann Laplacian, this is a free boundary condition which does not appear explicitly in the definition of the form domain and therefore does not appear explicitly in (5.16).

To prove the existence of transmission eigenvalues, we will use Lemma 14. Restricting our attention to the sphere, $||u||^2 = 1$, we may write

$$\begin{aligned} mt_\tau(u) &= (m+1)\tau^2 - 2\left(1+\frac{m}{2}\right)||\operatorname{grad} u||^2\tau + ||\Delta u||^2 \\ &\le (m+1)\tau^2 - 2\left(1+\frac{m}{2}\right)\lambda_0\tau + ||\Delta u||^2. \end{aligned}$$

We choose $\tau = \frac{1+\frac{m}{2}}{1+m}\lambda_0$ to obtain

$$\le -\frac{\left(1+\frac{m}{2}\right)^2}{1+m}\lambda_0^2 + ||\Delta u||^2$$

and restrict u to the eigenspace associated with the lowest $p+1$ clamped plate eigenvalues so that

$$mt_\tau(u) \le -\frac{\left(1+\frac{m}{2}\right)^2}{1+m}\lambda_0^2 + \mu_p.$$

Our hypothesis (5.19) is that this quantity is negative, so the conclusion (5.20) follows from Lemma 14. □

THEOREM 17. *Suppose that $m \in L^\infty(D)$. If*

$$\tau \le \min\left(1, \frac{1}{\sup(m)+1}\right)\lambda_0(D),$$

then τ is not a transmission eigenvalue. If $m > 0$ and

$$\inf(m) \ge 4\frac{\mu_p^{\frac{1}{2}}}{\lambda_0} + \frac{\mu_p}{\lambda_0^2}, \tag{5.21}$$

then there are $p+1$ transmission eigenvalues τ with

$$\tau \le \frac{\lambda_0(D)}{2}\left(\frac{\inf(m) - 2\frac{\mu_p^{\frac{1}{2}}}{\lambda_0}}{\inf(m)+1}\right).$$

Proof. For $m > 0$,

$$\begin{aligned} t_\tau(u) &= \int \frac{1}{m}|\left(\Delta + \tau(1+m)\right)u|^2 - \tau\int \overline{u}\left(\Delta + \tau(1+m)\right)u \\ &\ge \tau||\operatorname{grad} u||^2 - \tau^2\int(1+m)|u|^2 \\ &\ge \tau||u||^2\left(\lambda_0 - \tau(1+\sup m)\right), \end{aligned}$$

which shows that t_τ is positive definite if $\tau < \frac{\lambda_0}{1+\inf m}$ and therefore that τ is not a transmission eigenvalue. For $m < 0$

$$\begin{aligned} -t_\tau(u) &= \int \frac{1}{|m|}|\left(\Delta + \tau\right)u|^2 - \tau\int \overline{u}\left(\Delta + \tau\right)u \\ &> \tau||\operatorname{grad} u||^2 - \tau^2||u||^2 \\ &\ge \tau||u||^2\left(\lambda_0 - \tau\right) \end{aligned}$$

so that $-t_\tau$ is positive definite if $\tau < \lambda_0$, completing the proof of the first assertion.

To prove existence, we write

$$t_\tau(u) = \tau^2 \int \left(1 + \frac{1}{m}\right) |u|^2 - \tau \left(|| \operatorname{grad} u||^2 + \int \frac{1}{m} \left(\overline{u}\Delta u + u\Delta\overline{u}\right)\right) + \int \frac{1}{m} |\Delta u|^2.$$

We restrict our attention to functions u with $||u||^2 = 1$ and write $S = \sup(\frac{1}{m})$ to see that

$$t_\tau(u) \leq \tau^2(1+S) - \tau \left(|| \operatorname{grad} u||^2 - 2S||\Delta u|| \right) + S||\Delta u||^2.$$

Restricting to V^p gives

$$t_\tau(u) \leq \tau^2(1+S) - \tau \left(\lambda_0 - 2S\mu_p^{\frac{1}{2}} \right) + S\mu_p.$$

We minimize the sum of the first two terms by choosing $\tau = \frac{\lambda_0 - 2S\mu_p^{\frac{1}{2}}}{2(1+S)}$ to obtain

$$\leq -\frac{\left(\lambda_0 - 2S\mu_p^{\frac{1}{2}}\right)^2}{4(1+S)} + S\mu_p.$$

If we set $A = \frac{\lambda_0}{\mu_p^{\frac{1}{2}}}$, t_τ restricted to V^p is nonpositive if

$$\begin{aligned} (A-2S)^2 - 4S(1+S) &\geq 0, \\ A^2 - 4(A+1)S &\geq 0, \\ \frac{A^2}{4(A+1)} &\geq S, \end{aligned}$$

which is equivalent to (5.21). ∎

6. Conclusions. Under the hypothesis that the perturbation of the index of refraction is large enough, we have shown the existence of D-transmission eigenvalues, given upper and lower bounds for their locations, and identified the corresponding solutions to the transmission eigenvalue problem with the kernel of a scattering operator.

The upper and lower bounds for the transmission eigenvalues depend on the lowest eigenvalues of the Dirichlet Laplacian and the Dirichlet bi-Laplacian (the *clamped plate* operator). These bounds show that the lowest transmission eigenvalue increases as the L^∞-norm of m, or the size of its support, decreases. All of our bounds depend on D only through these eigenvalues and on m only through its infimum or supremum. Thus our estimates for the transmission eigenvalues, k^2, scale with dilations just like the Dirichlet eigenvalues, as the reciprocal of the area of D.

In the Born or weak scattering approximation, there are no transmission eigenvalues [7] when m is strictly positive or strictly negative. Our results are consistent with this, but we do not know if there is a threshold below which there are no transmission eigenvalues, or if the lowest transmission eigenvalue simply goes to infinity as m decreases to zero.[7]

In the radial case, there are infinitely many transmission eigenvalues, so it is reasonable to expect the same result here, but we have no results in this direction.

In summary, there are many questions remaining, some of which may be accessible by a further analysis of the quadratic forms of the operators introduced here. Because

[7] See [3] for an inequality relating the lowest transmission eigenvalue (if it exists) to the supremum of m and the diameter of D.

array imaging techniques are making the scattering operator, and hence its kernel, possible to measure, these questions are becoming increasingly relevant.

REFERENCES

[1] S. Agmon, *Spectral properties of Schrödinger operators and scattering theory*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 2 (1975), pp. 151–218.

[2] S. Agmon and L. Hörmander, *Asymptotic properties of solutions of differential equations with simple characteristics*, J. Anal. Math., 30 (1976), pp. 1–38.

[3] F. Cakoni, D. Colton, and P. Monk, *On the use of transmission eigenvalues to estimate the index of refraction from far field data*, Inverse Problems, 23 (2007), pp. 507–522.

[4] D. Colton, A. Kirsch, and L. Päivärinta, *Far-field patterns for acoustic waves in an inhomogeneous medium*, SIAM J. Math. Anal., 20 (1989), pp. 1472–1483.

[5] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Appl. Math. Sci. 93, Springer-Verlag, Berlin, 1998.

[6] D. Colton and P. Monk, *The inverse scattering problem for acoustic waves in an inhomogeneous medium*, in Inverse Problems in Partial Differential Equations (Arcata, CA, 1989), D. Colton, R. Ewing, and W. Rundell, eds., SIAM, Philadelphia, 1990, pp. 73–84.

[7] D. Colton, L. Päivärinta, and J. Sylvester, *The interior transmission problem*, Inverse Probl. Imaging, 1 (2007), pp. 13–28.

[8] L. Hörmander, *The Analysis of Linear Partial Differential Operators*. II. *Differential Operators with Constant Coefficients*, Classics in Mathematics, Springer-Verlag, Berlin, 2005.

[9] J. R. McLaughlin and P. L. Polyakov, *On the uniqueness of a spherically symmetric speed of sound from transmission eigenvalues*, J. Differential Equations, 107 (1994), pp. 351–382.

[10] J. R. McLaughlin, P. L. Polyakov, and P. E. Sacks, *Reconstruction of a spherically symmetric speed of sound*, SIAM J. Appl. Math., 54 (1994), pp. 1203–1223.

[11] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*. I. *Functional Analysis*, 2nd ed., Academic Press, New York, 1980.

[12] B. P. Rynne and B. D. Sleeman, *The interior transmission problem and inverse scattering from inhomogeneous media*, SIAM J. Math. Anal., 22 (1991), pp. 1755–1762.

[13] B. Simon, *Quantum Mechanics for Hamiltonians Defined as Quadratic Forms*, Princeton Series in Physics, Princeton University Press, Princeton, NJ, 1971.

SIAM J. MATH. ANAL.
Vol. 40, No. 2, pp. 754–775

PRESSURELESS EULER/EULER–POISSON SYSTEMS VIA ADHESION DYNAMICS AND SCALAR CONSERVATION LAWS*

TRUYEN NGUYEN† AND ADRIAN TUDORASCU‡

Abstract. The "sticky particles" model at the discrete level is employed to obtain global solutions for a class of systems of conservation laws among which lie the pressureless Euler and the pressureless attractive/repulsive Euler–Poisson system with zero background charge. We consider the case of finite, nonnegative initial Borel measures with finite second-order moment, along with continuous initial velocities of at most quadratic growth and finite energy. We prove the time regularity of the solution for the pressureless Euler system and obtain that the velocity satisfies the Oleinik entropy condition, which leads to a partial result on uniqueness. Our approach is motivated by earlier work of Brenier and Grenier, who showed that one-dimensional conservation laws with special initial conditions and fluxes are appropriate for studying the pressureless Euler system.

Key words. pressureless Euler, Euler–Poisson system, sticky particles, scalar conservation laws, Wasserstein distance, adhesion dynamics

AMS subject classifications. 35L65, 35L67, 82C40

DOI. 10.1137/070704459

1. Introduction.

Let $\alpha,\ \beta \in \mathbb{R}$ and consider the system

$$\begin{cases} \partial_t \rho + \partial_x(\rho v) = 0, \\ \partial_t(\rho v) + \partial_x(\rho v^2) = \rho(\alpha \partial_x \Phi + \beta) \quad \text{in } \mathbb{R} \times (0,T), \\ \partial_{xx}^2 \Phi = \rho. \end{cases} \tag{1}$$

If $\alpha = \beta = 0$, then (1) describes the pressureless Euler system in spatial dimension one. The most commonly known form of the pressureless, attractive/repulsive Euler–Poisson system with zero background charge is also obtained from (1) by taking $\alpha = \pm 1$ and $\beta = 0$. In this paper, we are concerned with global existence of solutions for the initial value problem. Unlike the Euler with pressure case, the natural environment for the evolution is the space of nonnegative Borel measures on the real line. We consider the case of finite total mass, which we normalize to unity. The pressureless Euler ($\alpha = \beta = 0$) problem was studied using different techniques in [5], [6], [7], [9], [23], [15], [16], [17], [18], [20]. We point out that in these papers, generally, the velocity is taken to be at least bounded on the support of the initial measure. It appears that [23] and [20] are the only references (to our knowledge) which allow for unbounded velocities. Also, [23] is remarkable for dealing with the gravitational term as well ($\alpha = -1,\ \beta = 0$). In spite of that, the serious limitation of [23] is the assumptions that the initial velocity be sublinear growth and the initial mass distribution ρ_0 be either discrete or absolutely continuous with respect to the Lebesgue measure. Our main contribution is proving existence of global solutions for (1) if ρ_0 is just in $\mathcal{P}_2(\mathbb{R})$

*Received by the editors October 3, 2007; accepted for publication (in revised form) April 16, 2008; published electronically August 1, 2008. Both authors were supported in part by the School of Mathematics, Georgia Institute of Technology.

http://www.siam.org/journals/sima/40-2/70445.html

†Department of Theoretical and Applied Mathematics, University of Akron, Akron, OH 44325 (tnguyen@uakron.edu). This author gratefully acknowledges the postdoctoral support provided by NSF grant DMS-03-54729.

‡School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 (adriant@math.gatech.edu).

and v_0 is continuous of at most quadratic growth and finite energy. As a consequence of an important result from [13], we also manage to show that the solution we obtain for Euler pressureless satisfies the Oleinik entropy condition which was conjectured in [7] and [23]. Note that similar constraints to ours on ρ_0 and v_0 were anticipated by Shnirelman [20] for the pressureless Euler system. However, our approach is more general than that of Shnirelman which cannot easily accommodate the α and β terms on the right-hand side of the momentum equation. The solutions constructed by Shnirelman were expressed in the form of a variational problem, which was shown in [3] to be equivalent to the variational principle considered in [23].

During the last decade, significant progress has been achieved in the study of partial differential equations in the context of optimal mass transportation. Much of the work on parabolic, dissipative equations was synthesized and placed in a very general setting in [2]. Much more recent and much less explored is the study of Hamiltonian systems in this context [1], [14]. The connection with the pressureless Euler system in arbitrary dimension was discovered by Benamou and Brenier [4], who showed that this system describes the geodesics in the Wasserstein space $\mathcal{P}_2(\mathbb{R})$. By definition, $\mathcal{P}_p(\mathbb{R})$ is the set of all Borel probabilities on $\mathbb{R}$ with finite p-order moment. The set $\mathcal{P}_2(\mathbb{R})$ is endowed with the quadratic Wasserstein metric defined by

$$W_2^2(\mu,\nu) := \min_{\gamma} \int_{\mathbb{R}^2} |x-y|^2 d\gamma(x,y),$$

where the infimum is taken among all probabilities γ on the product space $\mathbb{R}^2$ with marginals $\mu,\ \nu$. The theory of absolutely continuous curves in $\mathcal{P}_2(\mathbb{R})$ [2] asserts the existence of velocities satisfying the conservation of mass equation in (1), regarded as a continuity equation. The left-hand side of the momentum equation can also be interpreted as the acceleration along the curve. We shall discuss these interesting connections at the end of this paper. A different version of the pressureless Euler–Poisson system was analyzed in [13] in the context of optimal mass transportation. The focus was on the two-point boundary problem, and existence and uniqueness for solutions as action-minimizing paths in $\mathcal{P}_2(\mathbb{R})$ were obtained.

We shall need the following assumptions:

(H1) *The initial distribution of mass* $\rho_0 \in \mathcal{P}_2(\mathbb{R})$.

(H2) *There exists* $0 \le \Lambda < +\infty$ *such that*

$$v_0 \in C(\mathbb{R}) \cap L^2(\rho_0) \text{ and } |v_0(x)| \le \Lambda(1+x^2) \text{ for all } x \in \mathbb{R}.$$

The main objective is the following result.

THEOREM 1.1. *The initial-value problem for* (1) *admits a global weak solution in the sense of distributions if* (H1) *and* (H2) *hold.*

Two independent papers that appeared in 1996 and 1998 used adhesion dynamics to obtain global solutions for (1) in the $\alpha = \beta = 0$ case [7], [23] and in the $\alpha = -1$, $\beta = 0$ case [23]. Not only are we able to deal with the more general (1), but we also establish our results under less restrictive conditions on the initial mass distribution and velocity. In [7] the initial ρ_0 is compactly supported, while the initial velocity v_0 is continuous and bounded. These assumptions are relaxed in [23], e.g., $\sup_{|x|\le R} |v_0(x)|/R \to 0$ as $R \to +\infty$, while $\mathrm{spt}(\rho_0)$ may be unbounded, in which case $\int_0^x y \mathrm{d}\rho_0(y) \to +\infty$ as $|x| \to +\infty$. As opposed to [7], however, [23] makes the extra assumptions that ρ_0 be either discrete or absolutely continuous with respect to the Lebesgue measure, in which case $\rho_0 > 0$ on $\mathrm{spt}(\rho_0)$.

More recent work [16] treats the case $\alpha = \beta = 0$ for nonnegative Radon measures ρ_0 (not necessarily of finite total mass) and velocities $v_0 \in L^\infty(\rho_0)$. This paper is also remarkable in that it gives a necessary and sufficient condition for uniqueness of the Oleinik entropy solution—the initial weak continuity of the energy. As the example showing the necessity of this condition involves infinite mass initial measures, it remains unclear whether that is really needed in the finite mass case. Another paper that deals with possibly discontinuous (but bounded) initial velocities is [9], where the solution is produced constructively.

In [5] we find the concept of duality solutions, which is based on earlier work by the same authors. Existence and uniqueness are obtained under the assumption of atom-free initial density and bounded and continuous initial velocity. Boudin [6] obtains global existence of smooth solutions when the initial data has some higher regularity and is bounded away from zero and infinity (and thus of infinite total mass). Interestingly, the initial velocity does not have to be nondecreasing in order to rule out formation of singularities in finite time. We will consider this issue in section 4.3.

Whereas [7] and [23] use different approaches, they are still closely connected in principle. The fundamental underlying assumption for the discrete dynamics is the "sticky particle" hypothesis. The idea goes back to Zeldovich [24] and can be briefly described as follows. If $m_i,\ i = 1, n$ is a discrete system of masses initially located at $-\infty < x_1 < \cdots < x_n < +\infty$ and moving with initial velocities $v_i,\ i = 1, n$, then one makes the assumption that the velocities remain constant while there is no collision. At the collision of a group of particles, the particles stick together and the initial velocity of the newly formed particle is given by the conservation of momentum. In [23] the authors successfully implemented a version adapted to the case $\alpha = -1,\ \beta = 0$. Instead of the constant speed intercollisional motion, we now assume uniformly accelerated motion between collisions. The acceleration is of gravitational nature and is proportional to the difference between the total mass to the left and the total mass to the right. Thus, in both cases, the trajectory of the ith particle before collision is given by

$$x_i(t) = x_i + tv_i + \frac{t^2}{2} a_n^i,$$

where

$$a_n^i = \begin{cases} 0 & \text{if } \alpha = \beta = 0, \\ \frac{1}{2}\left(\sum_{j<i} m_j - \sum_{j>i} m_j\right) & \text{if } \alpha = -1,\ \beta = 0. \end{cases}$$

(Here we convene that $m_0 = m_{n+1} = 0$.) If the masses $m_j,\ i \le j \le k$, collide at time $t_0 > 0$, then conservation of momentum yields

$$v_i(t_0+) = \frac{\sum_{j=i}^k m_j v_j(t_0-)}{\sum_{j=i}^k m_j}.$$

Of course, only finitely many collisions can occur; therefore, the evolution of the system is completely determined by the above assumptions.

Next we briefly describe the technique employed in [23], whose approach does not distinguish between the discrete and absolutely continuous cases. Here the problem is attacked from a "continuation of characteristics" point of view. When shocks occur, i.e., when the map $\phi_t(y) := y + tv_0(y) + t^2 a_0(y)/2$ is no longer invertible, one needs to

redefine $\phi_t(y)$ in such a way that it remains nondecreasing and $\phi_{t\#}\rho_0 =: \rho_t$ satisfies the equation in a weak sense. This redefinition uses the so-called generalized variational principle which comes from the intuition provided by the discrete case (see [23] for details).

A more elegant approach [7], in our opinion, makes use of standard results on approximations for scalar conservation laws. It is applied to the Euler pressureless system ($\alpha = \beta = 0$) and relies on the fact that the distribution function M of ρ satisfies an autonomous scalar conservation law. Another advantage lies in the fact that the solution for the continuous problem is obtained from the discrete ones via approximation theory for scalar conservation laws. We shall adopt this point of view and prove the more general Theorem 1.1 by an appropriate adaptation of Brenier and Grenier's method.

The plan is as follows: we use (H1) to produce a sequence of discrete probabilities

$$\rho_0^n := \sum_{i=1}^{n} m_i^{(n)} \delta_{x_i^{(n)}} \to \rho_0 \text{ as } n \to \infty$$

in the 2-Wasserstein distance. We shall, in fact, prove that this sequence may be taken such that $\int_{\mathbb{R}} \zeta \mathrm{d}\rho_0^n$ is uniformly bounded for some superquadratic growth function $\zeta : [0,\infty) \to [0,\infty)$. Denote by M_0^n the right continuous distribution function of ρ_0^n and let

$$a_n(m) := \alpha\Big(\sum_{j=1}^{i} m_j - \frac{1}{2} m_i \Big) + \beta \tag{2}$$

whenever $M_0^n\big(x_i^{(n)} - \big) \le m < M_0^n\big(x_i^{(n)}\big)$. We then define flux functions $\tilde{F}_n : [0,T] \times [0,1] \to \mathbb{R}$ as

$$\tilde{F}_n(t,m) := \int_0^m f_n(\omega)\mathrm{d}\omega + t \int_0^m a_n(\omega)\mathrm{d}\omega, \tag{3}$$

where $N_{0\,\#}^n \chi_{(0,1)} = \rho_0^n$ optimally and $f_n := v_0 \circ N_0^n$ (the map N_0^n is taken to be right continuous). By adhesion dynamics we construct the unique entropy solution M_n for the first-order problem

$$\partial_t M + \partial_x\big[\tilde{F}_n(t,M)\big] = 0, \quad M(0,\cdot) = M_0^n. \tag{4}$$

We then use (H2) to show that, for an appropriate choice of approximating initial data, the sequence M_n will converge in some sense to the unique entropy solution for

$$\partial_t M + \partial_x\big[\tilde{F}(t,M)\big] = 0, \quad M(0,\cdot) = M_0, \tag{5}$$

where $\tilde{F} : [0,T] \times [0,1] \to \mathbb{R}$ is given by

$$\tilde{F}(t,m) := \int_0^m f(\omega)\mathrm{d}\omega + t \int_0^m a(\omega)\mathrm{d}\omega = \int_0^m f(\omega)\mathrm{d}\omega + t(\alpha m^2/2 + \beta m) \tag{6}$$

for $f := v_0 \circ N_0$ and $a(m) := \alpha m + \beta$ for $m \in [0,1]$. Here $N_0 := M_0^{-1}$ (generalized inverse) [12] is the right continuous optimal map pushing $\chi_{(0,1)}$ forward to ρ_0. The solution M of (5) will produce the solution $\rho := \partial_x M$ and $v\rho := \partial_x[\tilde{F}(t,M)]$ for (1) via a generalization of a result due to Volpert [22] on BV calculus. The last section is

dedicated to the $\alpha = 0 = \beta$ case. We give an explicit formula for v in terms of M, and, more importantly, we prove that our solution satisfies the Oleinik entropy condition. Some qualitative properties of the solution are discussed, e.g., time regularity and the impact of the initial velocity on the occurrence of spatial singularities. We finish with a partial result on uniqueness; i.e., we show that the energy of our solution for the pressureless Euler system is weakly continuous initially, which, along with the Oleinik entropy condition, leads to uniqueness in the case of bounded initial velocities [16].

2. Elements of one-dimensional BV calculus.

2.1. A BV chain rule in dimension one. To prove a chain rule for BV functions, we need the following lemma.

LEMMA 2.1. *Let μ be a Borel probability measure on $\mathbb{R}$ and M be its right continuous distribution function. Write*

$$\mu = \sum_{j\in J} m_i \delta_{x_j} + \rho,$$

where $\{x_j\}_{j\in J}$ is the set (at most countable) of discontinuities of M and $m_j := \mu(\{x_j\})$. If ρ is nonzero, then we have

$$M_{\#}\rho = \chi_{U^c} \text{ for } U := \bigcup_{j\in J} \big(M(x_j-), M(x_j)\big). \tag{7}$$

Proof. We first observe that the balance of mass is satisfied. Since M is monotone nondecreasing, (7) is equivalent to the fact that M is the optimal map pushing ρ forward to χ_{U^c}. That is what we prove next. It is well known [12] that, since both measures are atom-free, the optimal map is given by $G^{-1} \circ F$, where F, G are the right continuous distribution functions of ρ and χ_{U^c}, respectively, and G^{-1} is the generalized inverse of G, given by $G^{-1}(y) = \inf\{m \in [0,1] : G(m) > y\}$. We shall show that $M(x) = G^{-1} \circ F(x)$ for ρ-a.e. $x \in \mathrm{spt}(\rho)$, i.e., $G^{-1}(F(x)) = \inf\{m \in [0,1] : G(m) > F(x)\} = M(x)$. Note that

$$F(x) = M(x) - \sum_{x_j \le x} m_j \text{ and } G(m) = m - \sum_{M(x_j)\le m} m_j \text{ if } m \in U^c.$$

Thus, if $m \in U^c$,

$$G(m) - F(x) = \begin{cases} m - M(x) - \sum_{M(x)<M(x_j)\le m} m_j & \text{if } m > M(x), \\ 0 & \text{if } m = M(x), \\ m - M(x) + \sum_{m<M(x_j)\le M(x)} m_j & \text{if } m < M(x). \end{cases}$$

Since $G(M(x)) - F(x) = 0$ and G is a right continuous, nondecreasing function, all we need to prove is that there does not exist a nondegenerate interval $[M(x), M(x)+\epsilon]$ such that $G(m) = F(x)$ for all $m \in [M(x), M(x)+\epsilon]$. Suppose such an interval exists. We know M is right continuous at x; therefore, if $x < z < x + \delta(\epsilon)$, we have that $M(x) \le M(z) \le M(x) + \epsilon$. Thus, as $M(z) \in U^c$, we infer that $M(z) - M(x) - \sum_{x<x_j\le z} m_j = 0$, i.e., $\rho([x,z]) = 0$. This means $x \in \partial\mathrm{spt}(\rho)$, and, by taking $[x, z_x]$, the maximal interval for which $\rho([x,z]) = 0$, we see that $(x, z_x) \cap (y, z_y) = \emptyset$ for any $x \neq y$ in $\partial\mathrm{spt}(\rho)$ for which these nondegenerate intervals exist. This means that there are at most countably many such points, and, since ρ is atom-free, it follows that the ρ-measure of this set of points is zero. The lemma is thus proved. ∎

The following theorem is fundamental. The first part of (i) can be trivially obtained from [22]; see, e.g., [5] for the exact formula on the derivative. The novelty of our result appears in (ii) as we go from the Lipschitz to the $W^{1,p}$ case. However, for the reader's convenience we sketch an elementary proof for (i) as well.

THEOREM 2.2. *Let μ be a Borel probability measure on $\mathbb{R}$ and let M be its right continuous distribution function.*

(i) *Assume $f \in W^{1,\infty}(0,1)$. Then, $f \circ M \in BV(\mathbb{R})$ with distributional derivative $g\mu$, where*

$$g(x) = \begin{cases} \overline{f' \circ M}(x) & \text{if } \mu(\{x\}) = 0, \\ \dfrac{f \circ M(x) - f \circ M(x-)}{\mu(\{x\})} & \text{if } \mu(\{x\}) \neq 0, \end{cases} \tag{8}$$

μ-a.e. for some function $\overline{f' \circ M} \in L^\infty(\mathbb{R})$. Furthermore, suppose f' is defined unambiguously on $(0,1)$, and there exists a bounded $C^1[0,1]$ sequence f_n such that $f_n \to f$ uniformly and $f_n' \to f'$ everywhere on $(0,1)$. Then $\overline{f' \circ M} \equiv f' \circ M$ in the μ-a.e. sense, or, equivalently,

$$g(x) = \int_0^1 f'\big((1-s)M(x-) + sM(x)\big)\mathrm{d}s \text{ for } \mu\text{-a.e. } x \in \mathbb{R}. \tag{9}$$

(ii) *Let $1 \le p < +\infty$ and assume $f \in W^{1,p}(0,1) \cap C^1(0,1)$ and g is given by (8), with $\overline{f' \circ M} \equiv f' \circ M$ in the μ-a.e. sense. Then $g \in L^p(\mu)$ with $\|g\|_{L^p(\mu)} \le \|f'\|_{L^p(0,1)}$ and $f \circ M \in BV(\mathbb{R})$ with distributional derivative $g\mu$.*
Now take $f \in W^{1,p}(0,1)$. Suppose f' is defined unambiguously on $(0,1)$, and there exists a $W^{1,p}(0,1) \cap C^1(0,1)$ sequence f_n such that $f_n \to f$ in $W^{1,p}$ and $f_n' \to f'$ everywhere on $(0,1)$. Then we still have the same result with $\overline{f' \circ M} \equiv f' \circ M$ in the μ-a.e. sense.

Proof. W.l.o.g. we may assume that μ is supported in some bounded interval I. Also, we shall first assume $f \in C^1[0,1]$ to show that (8) is valid with $\overline{f' \circ M} \equiv f' \circ M$. Now consider $\varphi \in C_c^\infty(I)$. We need to show that

$$-\int_I \varphi'(x) f \circ M(x)\mathrm{d}x = \int_I \varphi(x) g(x)\mathrm{d}\mu(x). \tag{10}$$

If μ has finitely many atoms, then the validity of (10) can be checked by direct computation. Indeed, since M is piecewise continuous and bounded, we may approximate it uniformly by nondecreasing piecewise $W^{1,1}$ functions M_ϵ (may take M_ϵ piecewise linear and continuous on each continuity interval for M, such that M and M_ϵ agree at the endpoints). Thus, $\mu_\epsilon := M_\epsilon' \rightharpoonup M' = \mu$ weakly as nonnegative, bounded measures. The chain rule for Sobolev functions [8] applies piecewise and yields (10) for μ_ϵ. Then we pass to the limit to obtain the result for μ. Thus, let us assume $D := \{x_1, x_2, \ldots, x_n, \ldots\}$ is the infinite set of all atoms of μ and write

$$\mu = \sum_{i=1}^\infty m_i \delta_{x_i} + \rho, \text{ where } m_i > 0,\ i = 1, 2, \ldots,$$

and ρ is an atom-free nonnegative Borel measure of total mass $1 - \sum_{i=1}^\infty m_i \ge 0$. We shall call the atomic measure the singular part while ρ shall be called the regular part (although it may not be absolutely continuous with respect to the Lebesgue

measure). Consider now the sequence of measures μ_n given by $\mu_n = \sum_{i=1}^{n} m_i \delta_{x_i} + \rho$. Of course, $\mu_n \rightharpoonup \mu$ weakly $\star$ as measures. Since M_n, the right continuous distribution function of μ_n, has only finitely many discontinuities, (8) holds for μ_n, as proved above. It is easy to see that $M_n \to M$ Lebesgue a.e.; thus, the continuity of f along with its boundedness gives the convergence of the left-hand side of (10) by dominated convergence. Therefore, $f \circ M_n \in BV(\mathbb{R})$ with distributional derivative $g_n \mu_n$, where

$$g_n(x) := \begin{cases} f' \circ M_n(x) & \text{if } \mu_n(\{x\}) = 0, \\ \dfrac{f \circ M_n(x) - f \circ M_n(x-)}{\mu_n(\{x\})} & \text{if } \mu_n(\{x\}) \neq 0. \end{cases}$$

Now we write

$$\int_{\mathbb{R}} \varphi(x) g_n(x) \mathrm{d}\mu_n(x) = \sum_{i=1}^{n} \varphi(x_i)[f(M_n(x_i)) - f(M_n(x_i-))] + \int_{\mathbb{R}} \varphi(x) f' \circ M_n(x) \mathrm{d}\rho(x).$$

By the continuity of f' we obtain the convergence of the second term on the right-hand side. Thus, it remains to prove that

$$\sum_{i=1}^{n} \varphi(x_i)[f(M_n(x_i)) - f(M_n(x_i-))] \to \sum_{i=1}^{\infty} \varphi(x_i)[f(M(x_i)) - f(M(x_i-))],$$

which can be obtained after some calculations as a consequence of the convergence of the series $\sum m_i$ if $f \in C^2[0,1]$, and then for $C^1[0,1]$ functions by approximation. If f is $W^{1,\infty}(0,1)$, then we conclude by taking a sequence of functions in $C^1[0,1]$ such that $f_n \to f$ uniformly and f_n' are uniformly bounded. Indeed, one has

$$-\int_{\mathbb{R}} \varphi' f_n \circ M \mathrm{d}x = \int_{\mathbb{R}} \varphi f_n' \circ M \mathrm{d}\rho + \int_{\mathbb{R}} \varphi(x) \frac{f_n(M(x)) - f_n(M(x-))}{M(x) - M(x-)} \mathrm{d}\mu_s(x),$$

where $\mu_s = \sum m_j \delta_{x_j}$ is the singular part of μ. Note that the ratio in the second term of the right-hand side converges uniformly to $[f(M(x)) - f(M(x-))]/[M(x) - M(x-)]$ on the support of μ_s; thus the uniform bound on f_n' ensures the convergence of the integral by dominated convergence. Since the left-hand side is trivially convergent, it follows that

$$\int_{\mathbb{R}} \varphi(x) f_n' \circ M(x) \mathrm{d}\rho(x) \text{ converges as } n \to \infty,$$

which, along with the uniform bound on f_n', yields the convergence of $f_n' \circ M$ in the L^∞ weak $\star$ topology. We also deduce that the limit, denoted by $\overline{f' \circ M}$, is μ-a.e. independent of the chosen sequence f_n. The second statement from (i) easily follows (by dominated convergence) from the fact that $g_n \to g$ everywhere as an L^∞ bounded sequence.

To prove the first part of (ii) we truncate f' by

$$f_n'(x) := \begin{cases} -n & \text{if } f'(x) < -n, \\ f'(x) & \text{if } |f'(x)| \leq n, \\ n & \text{if } f'(x) > n \end{cases} \tag{11}$$

and let f_n be the antiderivative of f_n' vanishing at zero. Note that $|f_n'| \leq |f'|$ on $(0,1)$, which implies that f_n is uniformly bounded with respect to n in $L^p(0,1)$. First,

since $f_n \in C^1[0,1]$, we infer, according to (i), that $f_n \circ M$ is BV with distributional derivative $g_n\mu$, where

$$g_n(x) := \begin{cases} f_n' \circ M(x) & \text{if } \mu(\{x\}) = 0, \\ \dfrac{f_n \circ M(x) - f_n \circ M(x-)}{\mu(\{x\})} & \text{if } \mu(\{x\}) \neq 0. \end{cases} \tag{12}$$

This (see (9) for the equivalent integral expression) together with the fact that $|f_n'| \leq |f'|$ implies

$$|g_n(x)| \leq \int_0^1 \big|f'\big((1-s)M(x-) + sM(x)\big)\big|\mathrm{d}s =: h(x) \text{ for } x \in I.$$

Since $f_n' \to f'$, we infer that $g_n \to g$ pointwise. Assume that $h \in L^p(\mu)$. Then, we may pass to the limit in the right-hand side of

$$-\int_I \varphi'(x) f_n \circ M(x)\mathrm{d}x = \int_I \varphi(x) g_n(x)\mathrm{d}\mu(x).$$

The left-hand side converges to the appropriate quantity because $f_n \to f$ uniformly and f_n are uniformly bounded.

Thus, we are done if we can prove that $h \in L^p(\mu)$. For this let $\{x_j\}_{j\in J}$ be the set of discontinuities of M and $m_j := \mu(\{x_j\})$. Now consider, for $s \in [0,1]$, the sum

$$\sum_{i=1}^n m_i \big|f'\big((1-s)M(x_i-) + sM(x_i)\big)\big|^p = \sum_{i=1}^n m_i \big|f'\big(M(x_i-) + sm_i\big)\big|^p.$$

By monotone convergence

$$\sum_{i=1}^n \int_0^1 m_i \big|f'\big(M(x_i-) + sm_i\big)\big|^p \mathrm{d}s \to \int_0^1 \sum_{i=1}^\infty m_i \big|f'\big(M(x_i-) + sm_i\big)\big|^p \mathrm{d}s,$$

which, after obvious linear changes of variables, is equivalent to

$$\int_{\cup_{i=1}^n [M(x_i-), M(x_i)]} |f'(m)|^p \mathrm{d}m \to \int_0^1 \sum_{i=1}^\infty m_i \big|f'\big(M(x_i-) + sm_i\big)\big|^p \mathrm{d}s.$$

Again, by monotone convergence the left-hand side converges to $\int_U |f'(m)|^p \mathrm{d}m$, where $U := \bigcup_{j\in J} \big(M(x_j-), M(x_j)\big)$. Thus, since $f' \in L^p(0,1)$ and $U \subset [0,1]$, it follows that

$$\int_0^1 \sum_{i=1}^\infty m_i \big|f'\big(M(x_i-) + sm_i\big)\big|^p \mathrm{d}s = \int_U |f'(m)|^p \mathrm{d}m \leq \|f'\|^p_{L^p(0,1)},$$

so we can apply Fubini's theorem to obtain

$$\int_I \int_0^1 \big|f'\big((1-s)M(x-) + sM(x)\big)\big|^p \mathrm{d}s \mathrm{d}\mu_s(x) = \int_U |f'(m)|^p \mathrm{d}m < +\infty, \tag{13}$$

where μ_s denotes, as before, the singular part of μ (unrelated to the integration variable s). If $\mu = \mu_s$, then we are done. Else, by using Lemma 2.1, we obtain

$$\int_I |f' \circ M(x)|^p \mathrm{d}\rho(x) = \int_{U^c} |f'(m)|^p \mathrm{d}m,$$

which, combined with (13), yields

$$\int_I \int_0^1 |f'\big((1-s)M(x-)+sM(x)\big)|^p \mathrm{d}s\mathrm{d}\mu(x) = \|f'\|^p_{L^p(0,1)}. \tag{14}$$

Therefore, $h \in L^p(\mu)$ with $\|h\|_{L^p(\mu)} \leq \|f'\|_{L^p(0,1)}$ (with equality if $p = 1$). Since $|g| \leq h$ pointwise, the proof of the first part of (ii) is concluded.

Consequently,

$$-\int_{\mathbb{R}} \varphi' f_n \circ M \mathrm{d}x = \int_{\mathbb{R}} \varphi g_n \mathrm{d}\mu \text{ for all positive integers } n, \tag{15}$$

where g_n is defined by (12) for the approximating sequence f_n considered in the second part of (ii). We have, just as in deducing (14), that

$$\begin{aligned}\int_I |g_n(x) - g_m(x)|^p \mathrm{d}\mu(x) \leq & \int_I \int_0^1 |(f_n' - f_m')\big((1-s)M(x-)+sM(x)\big)|^p \mathrm{d}s\mathrm{d}\mu(x) \\ = & \|f_n' - f_m'\|^p_{L^p(0,1)}\end{aligned}$$

for all natural $m,\ n \geq 1$. Thus, $\{g_n\}$ is convergent in $L^p(\mu)$, and, due to the hypothesis of *everywhere* convergence of f_n' to f', we obtain $g_n \to g$ in $L^p(\mu)$, where

$$g(x) := \int_0^1 f'\big((1-s)M(x-)+sM(x)\big)\mathrm{d}s.$$

Passing to the limit in (15) concludes our proof. □

The following corollary holds for any $1 \leq p \leq +\infty$.

COROLLARY 2.3. *Let $f \in L^p(0,1)$ be right continuous (thus unambiguously defined everywhere in $(0,1)$), and take F to be its antiderivative vanishing at zero. If M is the cumulative distribution function of some Borel probability measure μ on $\mathbb{R}$, then $F \circ M \in BV(\mathbb{R})$ with distributional derivative $g\mu$, where*

$$g(x) := \begin{cases} f \circ M(x) & \text{if } \mu(\{x\}) = 0, \\ \dfrac{F \circ M(x) - F \circ M(x-)}{\mu(\{x\})} & \text{if } \mu(\{x\}) \neq 0, \end{cases} \tag{16}$$

in the μ-a.e. sense. Furthermore, $g \in L^p(\mu)$ and $\|g\|_{L^p(\mu)} \leq \|f\|_{L^p(0,1)}$.

The proof is an immediate consequence of Theorem 2.2. Indeed, we take (upon extending f by zero outside $(0,1)$) the function $f_n := \eta^n * f$, where η^n is obtained from the standard mollifier supported in $[-1/n, 1/n]$ by shifting it to the left by $1/n$. The classic properties of mollification still hold, e.g., $f_n \to f$ in $L^p(0,1)$ (if $p \neq +\infty$) and $f_n \in C^\infty[0,1]$. However, the interesting feature of these "shifted mollifiers" is that the right continuity of f on $(0,1)$ is enough to easily prove that $f_n \to f$ *everywhere* in $(0,1)$. Thus, if we take $F_n(m) := \int_0^m f_n(\omega)\mathrm{d}\omega$, we are within the hypotheses of Theorem 2.2.

Remark 2.4. To understand the relevance of this result, observe that the functions $f_n,\ f$ defined in the introduction are right continuous on $(0,1)$. Indeed, that comes as a consequence of the continuity of v_0 on $\mathbb{R}$ and the right continuity of $N_0^n,\ N_0$ on $(0,1)$ (as generalized inverses of nondecreasing, right continuous functions [21]).

2.2. A two-dimensional extension. Now let us assume $M : [0,T]\times\mathbb{R} \to [0,1]$ for some $0 < T < +\infty$, such that $M \in BV([0,T]\times\mathbb{R})$ and $M(t,\cdot)$ is a right continuous probability distribution function for Lebesgue a.e. $t \in (0,T)$. The following lemma is easy to check as a consequence of Fubini's theorem.

LEMMA 2.5. *Let ∇M be the vector-valued measure given by the BV gradient $\nabla = (\partial_t, \partial_x)$ of M. Then, its x-component $\mathrm{d}\partial_x M$ admits the decomposition $\mathrm{d}\partial_x M(t,\cdot)\mathrm{d}t$.*

We shall use this to prove the main result of this section, which we state below. We denote by C_r the space of right continuous functions.

THEOREM 2.6. *Consider $M : [0,T]\times\mathbb{R} \to [0,1]$ as above, and let $f \in L^2(0,1)\cap C_r(0,1)$ and F be its antiderivative vanishing at zero. Assume further that*

$$\partial_t M + \partial_x[F\circ M] = 0 \text{ in } \mathcal{D}'([0,T]\times\mathbb{R}). \tag{17}$$

Then $F\circ M \in BV([0,T]\times\mathbb{R})$ and $\nabla[F\circ M] = g\nabla M$ for some $g \in L^1(|\nabla M|)$.

Proof. We may consider w.l.o.g. the case $f \in C(0,1)$. Else, we simply use Corollary 2.3 instead of Theorem 2.2. Let us start by taking the truncations f_n for f as in (11). The corresponding F_n's are in $C^1[0,1]$, $F_n \to F$ uniformly and $f_n \to f$ in $L^2(0,1)$. According to [22], $F_n\circ M \in BV([0,T]\times\mathbb{R})$ and there exist $\bar f_n \in L^\infty$ such that $\nabla[F_n\circ M] = \bar f_n \nabla M$ as vector-valued measures. This means that

$$-\int_0^T\!\!\int_{\mathbb{R}} F_n(M)\nabla\cdot\boldsymbol{\phi}\mathrm{d}x\mathrm{d}t = \int_0^T\!\!\int_{\mathbb{R}} \bar f_n\phi_1\mathrm{d}\partial_t M + \int_0^T\!\!\int_{\mathbb{R}} \bar f_n\phi_2\mathrm{d}\partial_x M \tag{18}$$

for all $\boldsymbol{\phi} := (\phi_1,\phi_2) \in C_c^\infty((0,T)\times\mathbb{R};\mathbb{R}^2)$. We now use Lemma 2.5 and (17) to infer

$$-\int_0^T\!\!\int_{\mathbb{R}} F_n(M)\nabla\cdot\boldsymbol{\phi}\mathrm{d}x\mathrm{d}t = \int_0^T\!\!\int_{\mathbb{R}} \big(-g\phi_1+\phi_2\big)\bar f_n\mathrm{d}\partial_x M(t,\cdot)\mathrm{d}t, \tag{19}$$

where, according to Theorem 2.2 (ii),

$$g(t,\cdot) := \int_0^1 f\big((1-s)M(t,\cdot-) + sM(t,\cdot)\big)\mathrm{d}s \in L^2(\partial_x M(t,\cdot))$$

for Lebesgue a.e. $t \in (0,T)$. Also, Theorem 2.2 (ii) ensures that

$$\|g(t,\cdot)\|_{L^2(\partial_x M(t,\cdot))} \le \|f\|_{L^2(0,1)} \text{ uniformly with respect to } t.$$

Note that (19) implies

$$-\int_0^T\!\!\int_{\mathbb{R}} F_n(M)\partial_x\phi_1\mathrm{d}x\mathrm{d}t = \int_0^T\!\!\int_{\mathbb{R}} \bar f_n\phi_1\mathrm{d}\partial_x M(t,\cdot)\mathrm{d}t \tag{20}$$

for all $\phi_1 \in C_c^\infty((0,T)\times\mathbb{R})$. For Lebesgue a.e. $t \in (0,T)$ one has (Theorem 2.2 (i))

$$\partial_x[F_n\circ M(t,\cdot)] = g_n(t,\cdot)\partial_x M(t,\cdot) \text{ for } g_n(t,x) := \int_0^1 f_n\big((1-s)M(t,x-)+sM(t,x)\big)\mathrm{d}s$$

and

$$\|g_n(t,\cdot)\|_{L^2(\partial_x M(t,\cdot))} \le \|f\|_{L^2(0,1)} \text{ uniformly in } t \text{ and } n.$$

Along with (20), this implies $\bar f_n \equiv g_n$ in the $\mathrm{d}\partial_x M(t,\cdot)\mathrm{d}t$-a.e. sense. Therefore, (19) is equivalent to

$$-\int_0^T\!\!\int_{\mathbb{R}} F_n(M)\nabla\cdot\boldsymbol{\phi}\mathrm{d}x\mathrm{d}t = \int_0^T\!\!\int_{\mathbb{R}} \big(-g\phi_1+\phi_2\big)g_n\mathrm{d}\partial_x M(t,\cdot)\mathrm{d}t. \tag{21}$$

Furthermore, due to the pointwise convergence and the uniform (in n and t) $L^2(\partial_x M)$-bounds we obtain $g_n \to g$ in $L^2(\mathrm{d}\partial_x M(t,\cdot)\mathrm{d}t)$. By the uniform (and bounded) convergence of F_n to F we can also pass to the limit in the left-hand side. Thus,

$$-\int_0^T\int_{\mathbb{R}} F(M)\nabla\cdot\boldsymbol{\phi}\mathrm{d}x\mathrm{d}t = \int_0^T\int_{\mathbb{R}} \big(-g^2\phi_1 + g\phi_2\big)\mathrm{d}\partial_x M(t,\cdot)\mathrm{d}t,$$

which, after using Lemma 2.5 and (17) once more, leads to

$$-\int_0^T\int_{\mathbb{R}} F(M)\nabla\cdot\boldsymbol{\phi}\mathrm{d}x\mathrm{d}t = \int_0^T\int_{\mathbb{R}} g\phi_1\mathrm{d}\partial_t M + \int_0^T\int_{\mathbb{R}} g\phi_2\mathrm{d}\partial_x M.$$

Since $g \in L^2(\partial_x M) \subset L^1(\partial_x M)$ and $g \in L^1(|\partial_t M|)$, we obtain the result. □

We are going to need a slightly different result which we now state as a consequence. It may be proved by retracing the proof of Theorem 2.6.

COROLLARY 2.7. *Assume now that*

$$\partial_t M + \partial_x[\tilde{F}(t,M)] = 0 \text{ in } \mathcal{D}'([0,T]\times\mathbb{R})$$

for some time-linear perturbation $\tilde{F}(t,\cdot) := F + t\Psi$, $\Psi \in C^1[0,1]$. *Then* $F\circ M \in BV([0,T]\times\mathbb{R})$ *and*

$$\nabla[\tilde{F}(t,M)] = (g + t\psi)\nabla M + \mathbf{V}$$

for g *from Theorem* 2.6, *and the vector field* $\mathbf{V} := \big(\Psi\circ M, 0\big)$. *Here,*

$$\psi(t,x) := \int_0^1 \Psi'\big((1-s)M(t,x-) + sM(t,x)\big)\mathrm{d}s.$$

3. Pressureless Euler/Euler–Poisson systems via scalar conservation laws.

3.1. Convergence of measures and their distribution functions. Let $1 \le p < \infty$, $\mu \in \mathcal{P}_p(\mathbb{R})$, and $v \in L^p(\mu)\cap C(\mathbb{R})$. Then by the de la Vallée-Poussin lemma which can be found in [10], there exists a nonnegative, convex, increasing function $\zeta \in C^1([0,+\infty))$ satisfying $\zeta(0) = 0$ and $\frac{\zeta(t)}{t} \uparrow +\infty$ as $t \to +\infty$ such that

$$\int_{\mathbb{R}} \zeta(|x|^p + |v(x)|^p)\mathrm{d}\mu(x) < +\infty.$$

It is also well known that (see [11], for example) there exist a probability space (Ω,Σ,P) and a sequence of independent random variables $\xi_i : \Omega \to \mathbb{R}$ such that

$$\xi_{i\#}P = \mu.$$

Now for each positive integer n and each $\omega \in \Omega$, define

$$\mu^{n,\omega} := \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i(\omega)}.$$

Then by the strong law of large numbers and the separability of $C_c(\mathbb{R})$ we have for P-a.e. $\omega \in \Omega$

$$\int_{\mathbb{R}} f(x)\mathrm{d}\mu^{n,\omega}(x) = \frac{1}{n}\sum_{i=1}^n f(\xi_i(\omega)) \to \mathbb{E}(f\circ\xi_1) = \int_{\mathbb{R}} f(x)\mathrm{d}\mu(x)$$

for all functions $f \in C_c(\mathbb{R})$. Consequently, $\mu^{n,\omega}$ converges narrowly to μ. Thus, by using the strong law of large numbers again, we also have that for P-a.e. $\omega \in \Omega$

$$\int_{\mathbb{R}} \zeta(|x|^p)\mathrm{d}\mu^{n,\omega}(x) \to \int_{\mathbb{R}} \zeta(|x|^p)\mathrm{d}\mu(x), \quad \int_{\mathbb{R}} \zeta(|v|^p)\mathrm{d}\mu^{n,\omega}(x) \to \int_{\mathbb{R}} \zeta(|v|^p)\mathrm{d}\mu(x).$$

The last fact together with the properties of ζ yields for $k > 0$ large enough

$$\frac{\zeta(k^p)}{k^p} \int_{\{x:|x|\geq k\}} |x|^p \mathrm{d}\mu^{n,\omega} \leq \int_{\{x:|x|\geq k\}} \frac{\zeta(|x|^p)}{|x|^p} |x|^p \mathrm{d}\mu^{n,\omega} \leq \int_{\mathbb{R}} \zeta(|x|^p)\mathrm{d}\mu^{n,\omega} \leq C(\omega).$$

Therefore, $\{\mu^{n,\omega}\}$ has uniformly integrable p-moments. Consequently, by Proposition 7.1.5 in [2] we obtain that for P-a.e. $\omega \in \Omega$, $\mu^{n,\omega} \to \mu$ in W_p and

$$\int_{\mathbb{R}} \zeta(|x|^p)\mathrm{d}\mu^{n,\omega}(x) \to \int_{\mathbb{R}} \zeta(|x|^p)\mathrm{d}\mu(x), \quad \int_{\mathbb{R}} \zeta(|v|^p)\mathrm{d}\mu^{n,\omega}(x) \to \int_{\mathbb{R}} \zeta(|v|^p)\mathrm{d}\mu(x).$$

LEMMA 3.1. *Suppose $1 \leq p < \infty$ and $\{\mu_n\}$ is a sequence of measures in $\mathcal{P}_p(\mathbb{R})$ converging to $\mu \in \mathcal{P}_p(\mathbb{R})$ in W_p. When $p > 1$ we assume further that $\int_{\mathbb{R}} \zeta(|x|^p)\mathrm{d}\mu_n(x)$ is uniformly bounded in n for some nonnegative, convex, increasing function $\zeta \in C^1([0,+\infty))$ satisfying $\zeta(0) = 0$ and $\frac{\zeta(t)}{t} \uparrow +\infty$ as $t \to +\infty$. Then we have the following:*

(i) *The L^1 norm of $M_n - M$ on $\mathbb{R}$ goes to zero, where*

$$M(x) := \mu((-\infty, x]).$$

(ii) *For any nondecreasing C^1 function B on $[0,1]$ such that $B(0) = 0, B(1) = 1$,*

$$\partial_x(B(M_n)) \to \partial_x(B(M)) \quad \text{in} \quad W_p.$$

Proof. From Theorem 6.0.2 in [2], we obtain

$$W_1(\mu_n, \mu) = \int_0^1 |M_n^{-1}(s) - M^{-1}(s)|\mathrm{d}s,$$

where the generalized inverse M^{-1} of M is defined by

$$M^{-1}(s) := \inf\{x \in \mathbb{R} : M(x) > s\}, \quad s \in [0,1].$$

Then by using Fubini's theorem we get $W_1(\mu_n, \mu) = \|M_n - M\|_{L^1(\mathbb{R})}$. This together with the fact that $W_1(\mu_n, \mu) \leq W_p(\mu_n, \mu)$ gives (i). By (i) and the Lipschitz condition on B we clearly have $\|B(M_n) - B(M)\|_{L^1(\mathbb{R})} \to 0$. But as

$$W_1(\partial_x(B(M_n)), \partial_x(B(M))) = \|B(M_n) - B(M)\|_{L^1(\mathbb{R})},$$

we infer in particular that $\partial_x(B(M_n)) \to \partial_x(B(M))$ narrowly.

We have from Theorem 2.2

$$\int_{\mathbb{R}} \zeta(|x|^p)\mathrm{d}\partial_x(B(M_n)) = \int_{\mathbb{R}} \zeta(|x|^p) g_n(x)\mathrm{d}\mu_n(x),$$

where $g_n(x) = \int_0^1 B'(sM_n(x) + (1-s)M_n(x-))\mathrm{d}s$. Therefore,

$$\int_{\mathbb{R}} \zeta(|x|^p)\mathrm{d}\partial_x(B(M_n)) \leq \|B'\|_\infty \int_{\mathbb{R}} \zeta(|x|^p)\mathrm{d}\mu_n(x) \leq C\|B'\|_\infty.$$

It follows from this and the properties of the function ζ that for any $k > 0$ large enough,

$$\int_{\{x:|x|\geq k\}} |x|^p \mathrm{d}\partial_x(B(M_n)) \leq \frac{k^p}{\zeta(k^p)} C\|B'\|_\infty.$$

That is, the sequence of probability measures $\partial_x(B(M_n))$ has uniformly integrable p-moments. Consequently, we can conclude from Proposition 7.1.5 in [2] that

$$\partial_x(B(M_n)) \to \partial_x(B(M))$$

in W_p, as desired. ∎

3.2. Convergence of the discrete problem. Following [7], we take a discrete probability measure

$$\rho_0^n := \sum_{j=1}^n m_j \delta_{x_j}, \quad x_1 < x_2 < \cdots < x_n,$$

and define

$$\rho^n(t,x) := \sum_{j=1}^n m_j \delta_{x_j(t)}, \tag{22}$$

where the characteristics are given by

$$x_j(t) = x_j + tv_j + \frac{t^2}{2} a_n\big(M_n(t, x_j(t)-)\big). \tag{23}$$

Here, as in the introduction, we have

$$a_n\big(M_n(t, x_j(t)-)\big) = \left[\alpha\left(\sum_{i=1}^j m_i - \frac{1}{2}m_j\right) + \beta\right].$$

We impose the adhesion dynamics at collision (see the introduction) and consider

$$M_n(t,x) := \sum_{j=1}^n m_j H\big(x - x_j(t)\big), \tag{24}$$

where H is the right continuous Heaviside function. Since M_n is piecewise constant, we need only show that M_n solves (4) by checking the Rankine–Hugoniot jump conditions across the shocks $x = x_j(t)$, i.e.,

$$\dot{x}_j(t) = \frac{\tilde{F}_n(t, M_n(t, x_j(t))) - \tilde{F}_n(t, M_n(t, x_j(t)-))}{M_n(t, x_j(t)) - M_n(t, x_j(t)-)}, \quad j = 1, \ldots, n. \tag{25}$$

Assume the masses m_i for $j_0 \leq i \leq j_1$, including m_j, are all amassed at time t. Then, exactly as in [7], we have

$$v_j(t) = \frac{F_n(M_n(t, x_j(t))) - F_n(M_n(t, x_j(t)-))}{M_n(t, x_j(t)) - M_n(t, x_j(t)-)}, \quad j = 1, \ldots, n,$$

where F_n is defined by

$$F_n(m) := \int_0^m f_n(\omega)\mathrm{d}\omega \quad \text{for } m \in [0,1].$$

Observe that $\tilde{F}_n(t,m) = F_n(m) + t\int_0^m a_n(\omega)\mathrm{d}\omega$ by the definition of $\tilde{F}_n$ in the introduction. The integrand below is constantly a_n^j on each interval of the form $[M_n(t,x_j(t)-), M_n(t,x_j(t)))$; thus,

$$a_n\big(M_n(t,x_j(t)-)\big)\sum_{i=j_0}^{j_1} m_i = \int_{M_n(t,x_j(t)-)}^{M_n(t,x_j(t))} a_n(\omega)\mathrm{d}\omega$$

implies (25). To prove that M_n is an entropy solution for (4), we check the entropy inequality

$$\dot{x}_j(t) \le \frac{\tilde{F}_n(t,X) - \tilde{F}_n(t,M_n(t,x_j(t)-))}{X - M_n(t,x_j(t)-)}, \tag{26}$$

where $X = \sum_{i\le k} m_i$, for some $j_0 \le k \le j_1$. The inequality

$$v_j(t) \le \frac{F_n(X) - F_n(M_n(t,x_j(t)-))}{X - M_n(t,x_j(t)-)}$$

is justified in [7] as a consequence of the *barycentric lemma* (which simply formulates the fact that if two groups of particles collide, then the averaged velocity of the group to the left decreases). Then we see that

$$a_n\big(M_n(t,x_j(t)-)\big)\sum_{i=j_0}^{k} m_i = \int_{M_n(t,x_j(t)-)}^{X} a_n(\omega)\mathrm{d}\omega,$$

which, together with the previous inequality, yields (26). We have just sketched the proof of the following proposition.

PROPOSITION 3.2. *The function M_n given by* (24) *is the entropy solution of the problem* (4).

Next we want to show that M_n converges in some sense to an entropy solution of (5). The following proposition applies if the initial approximating measures are of the form

$$\rho_0^n := \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i^{(n)}}.$$

Note that section 3.1 shows that we may consider such approximations.

PROPOSITION 3.3. *Let $\rho_0 \in \mathcal{P}_2(\mathbb{R})$ and consider a sequence of discrete probabilities ρ_0^n as above such that $W_2\big(\rho_0,\rho_0^n\big) \to 0$ and*

$$\int_{\mathbb{R}} \zeta(v_0^2)\mathrm{d}\rho_0^n \to \int_{\mathbb{R}} \zeta(v_0^2)\mathrm{d}\rho_0 < +\infty, \ \int_{\mathbb{R}} \zeta(x^2)\mathrm{d}\rho_0^n \to \int_{\mathbb{R}} \zeta(x^2)\mathrm{d}\rho_0 < +\infty \tag{27}$$

for some nonnegative, convex, increasing function $\zeta \in C^1([0,+\infty))$ satisfying $\zeta(0) = 0$ and $\frac{\zeta(t)}{t} \uparrow +\infty$ as $t \to +\infty$. Let M_0^n and M_0 be the right continuous cumulative distribution functions of ρ_0^n and ρ_0, respectively. Consider, as above, the entropy

solutions M_n of (4) *for all* n. *Then there exists a Borel function* $M : [0,\infty)\times\mathbb{R} \to [0,1]$ *such that, for any given* $T > 0$,

$$\max_{0\le t\le T} W_2\left(\partial_x M_n(t,\cdot), \partial_x M(t,\cdot)\right) \to 0 \text{ and } \max_{0\le t\le T} \|M_n(t,\cdot) - M(t,\cdot)\|_{L^1(\mathbb{R})} \to 0. \tag{28}$$

Moreover, M *is the entropy solution of the problem* (5).

Proof. First recall that

$$\rho^n(t,x) := \frac{1}{n}\sum_{j=1}^n \delta_{x_j^{(n)}(t)},$$

where

$$x_j^{(n)}(t) = x_j^{(n)} + tv_j^{(n)} + \frac{t^2}{2}\left(\alpha\frac{2j-1}{2n} + \beta\right).$$

Let $\zeta_t(x) := \zeta(\frac{1}{\kappa}x)$, where κ is some positive constant depending only on t which will be determined later. Then, by the properties of ζ, we have

$$\begin{aligned}
&\int_{\mathbb{R}} \zeta_t(|x|^2)\mathrm{d}\rho_n(t,x)\\
&= \frac{1}{n}\sum_{j=1}^n \zeta\left(\frac{1}{\kappa}|x_j^{(n)}(t)|^2\right) \le \frac{1}{n}\sum_{j=1}^n \zeta\left(\frac{3}{\kappa}\left[|x_j^{(n)}|^2 + t^2|v_j^{(n)}|^2 + \frac{(|\alpha|+|\beta|)^2t^4}{4}\right]\right)\\
&\le \frac{1}{3}\zeta\left(\frac{9(|\alpha|+|\beta|)^2t^4}{4\kappa}\right) + \frac{1}{3n}\sum_{j=1}^n \zeta\left(\frac{9}{\kappa}|x_j^{(n)}|^2\right) + \frac{1}{3n}\sum_{j=1}^n \zeta\left(\frac{9t^2}{\kappa}|v_j^{(n)}|^2\right).
\end{aligned}$$

Hence, by choosing $\kappa = \max\{9, 9t^2\}$ and using again the fact that ζ is increasing, we obtain

$$\int_{\mathbb{R}} \zeta_t(|x|^2)\mathrm{d}\rho_n(t,x) \le C_t + \frac{1}{3}\int_{\mathbb{R}} \zeta(|x|^2)\mathrm{d}\rho_0^n(t,x) + \frac{1}{3}\int_{\mathbb{R}} \zeta(|v_0|^2)\mathrm{d}\rho_0^n(t,x) \le C_t$$

uniformly in n (the last inequality is due to (27)). Since ζ_t has superlinear growth (it is just an argument-rescaled version of ζ), there exists $\rho(t,\cdot) \in \mathcal{P}_2(\mathbb{R})$ such that, up to a subsequence that may depend on t, $W_2(\rho^n(t,\cdot), \rho(t,\cdot)) \to 0$ as $n \to \infty$. By a standard diagonal argument we can choose a subsequence independent of t satisfying $W_2(\rho^n(t,\cdot), \rho(t,\cdot)) \to 0$ for all $t \in [0,\infty)\cap\mathbb{Q}$. In order to see that this conclusion also holds for all t in $[0,\infty)$, we are going to show that the paths $\rho^n(t,\cdot)$ are uniformly locally Lipschitz in t. Indeed, let $T > 0$ and $t, s \in [0,T]$ be arbitrary. Then, since

$$|x_j^{(n)}(t) - x_j^{(n)}(s)|^2 \le C|t-s|^2\left(|v_j^{(n)}|^2 + T^2\right),$$

we have

$$\begin{aligned}
W_2^2(\rho^n(t,\cdot), \rho^n(s,\cdot)) &\le \frac{1}{n}\sum_{j=1}^n |x_j^{(n)}(t) - x_j^{(n)}(s)|^2 \le C|t-s|^2\left(\frac{1}{n}\sum_{j=1}^n |v_j^{(n)}|^2 + T^2\right)\\
&= C|t-s|^2\left(\int_{\mathbb{R}} |v_0(x)|^2\mathrm{d}\rho_0^n + T^2\right) \le C|t-s|^2 \quad \text{uniformly in } n.
\end{aligned}$$

Using this uniformly Lipschitz property, we can conclude that, in fact,

$$W_2(\rho^n(t,\cdot),\rho(t,\cdot)) \to 0 \text{ as } n\to\infty, \text{ uniformly in } t\in[0,T], \tag{29}$$

which yields (28) with $M(t,x) := \rho(t,(-\infty,x])$.

Next we show that M is a solution of (5). By the assumptions and Lemma 3.1, we have

$$\int_{\mathbb{R}} |M_0^n - M_0| \mathrm{d}x \to 0$$

and

$$\partial_x(B(M_0^n)) \to \partial_x(B(M_0)) \quad \text{in} \quad W_2$$

for any nondecreasing C^1 function B on $[0,1]$ satisfying $B(0)=0$ and $B(1)=1$. But as v_0 satisfies the assumption (H2), we obtain

$$\int_{\mathbb{R}} v_0(x)\mathrm{d}\partial_x(B(M_0^n)) \to \int_{\mathbb{R}} v_0(x)\mathrm{d}\partial_x(B(M_0)).$$

By the definitions of f_n and f, this means that

$$\int_0^1 f_n(m)B'(m)\mathrm{d}m \to \int_0^1 f(m)B'(m)\mathrm{d}m$$

for all nondecreasing C^1 functions B on $[0,1]$ satisfying $B(0)=0$ and $B(1)=1$. Indeed,

$$\begin{aligned}\int_{\mathbb{R}} v_0(x)\mathrm{d}\partial_x(B(M_0^n))(x) &= \sum_{i=1}^n v_0(x_i^{(n)})\left[B(M_0^n(x_i^{(n)})) - B(M_0^n(x_i^{(n)}-))\right] \\ &= \sum_{i=1}^n \int_{M_0^n(x_i^{(n)}-)}^{M_0^n(x_i^{(n)})} v_0\circ N_0^n(m)B'(m)\mathrm{d}m = \int_0^1 f_n(m)B'(m)\mathrm{d}m.\end{aligned}$$

For the continuous version we use (7) to conclude in a similar way. It then follows that

$$\int_0^1 f_n(m)g(m)\mathrm{d}m \to \int_0^1 f(m)g(m)\mathrm{d}m \quad \text{for all functions} \quad g\in C([0,1]).$$

This together with the fact

$$\int_0^1 |f_n(m)|\mathrm{d}m = \int_{\mathbb{R}} |v_0(x)|\mathrm{d}\rho_0^n \le C \quad \text{uniformly in } n$$

yields $f_n \to f$ weakly in L^1. By using the uniform convergence of $\tilde{F}_n$ to $\tilde{F}$ as continuous and bounded functions, we deduce that M is a solution of the problem (5). Now, for a fixed $t\ge 0$, let U be any C^1 function on $[0,1]$. Define

$$\tilde{F}_{n,U}(t,m) := \int_0^m \left[f_n(\omega) + ta_n(\omega)\right] U'(\omega)\mathrm{d}\omega,$$

$$\tilde{F}_U(t,m) := \int_0^m \left[f(\omega) + ta(\omega)\right] U'(\omega)\mathrm{d}\omega.$$

Then we use the uniform convergence of a_n to $\alpha\mathrm{id}+\beta$ and the facts that $f_n \to f$ weakly in L^1, and

$$\int_0^1 \zeta(|f_n|)\mathrm{d}m = \int_{\mathbb{R}} \zeta(|v_0|)\mathrm{d}\rho_0^n \le C \tag{30}$$

to deduce that $\tilde{F}_{n,U}$ converges to $\tilde{F}_U$ uniformly in $(t,m)\in[0,T]\times[0,1]$. (Notice that we have used the assumption (H2) to derive (30), the equi-integrability of the sequence $\{f_n\}$ in $L^1([0,1])$.) Therefore, as in [7], we conclude that M is, in fact, an entropy solution for (5). ∎

3.3. The existence result. We finally have all the necessary tools to prove Theorem 1.1. Note, however, that we leave out the proof of the fact that the initial conditions are satisfied. We will show that in Proposition 4.9.

Proof of Theorem 1.1. Since M is a solution of the problem (5) by Proposition 3.3, we can use Corollary 2.7 to conclude that $\partial_t M + v\partial_x M = 0$, where $v(t,\cdot) := g(t,\cdot) + t\psi(t,\cdot)$ is well defined $\partial_x M(t,\cdot) =: \rho(t,\cdot)$-a.e. By differentiation in the sense of distributions, we obtain the first equation in (1). Then, Corollary 2.7 also gives

$$\begin{aligned}\partial_t(\rho v) &= \partial_t\big[\partial_x[\tilde{F}(t,M)]\big] = \partial_x\big[\partial_t[\tilde{F}(t,M)]\big]\\ &= \partial_x[v\partial_t M + \Psi(M)] = \partial_x(-v^2\rho) + \psi\rho.\end{aligned}$$

Note that, in our case, $\Psi(m) = \alpha m^2/2 + \beta m$. Thus,

$$\psi(t,x) = \beta + \alpha\int_0^1 [(1-s)M(t,x-) + sM(t,x)]\mathrm{d}s = \beta + \alpha\left[M(t,x) - \frac{1}{2}\rho(t,\{x\})\right].$$

Since $\partial_x M(t,\cdot) = \rho(t,\cdot)$ and $\rho(t,\cdot)$ has at most countably many atoms, we may take $\Phi(t,x) = \int_{-\infty}^x M(t,y)\mathrm{d}y$ to conclude. ∎

Remark 3.4. Note that the term $M(t,x) - \frac{1}{2}\rho(t,\{x\})$ is precisely the *barycentric projection* [2] onto $\rho(t,\cdot)$ of the *optimal coupling* between $\chi_{(0,1)}$ and $\rho(t,\cdot)$. It differs from the projection considered in [13] by an additive factor of 0.5 due to the fact that, instead of $\chi_{(0,1)}$, the reference measure in [13] was $\chi_{(-0.5,0.5)}$.

4. Time regularity, entropy condition, and shocks. In this section we discuss some qualitative properties of the "sticky particle" solution. The family of absolutely continuous curves in $\mathcal{P}_2(\mathbb{R})$ is central to our approach. Thus, recall that $(\mathcal{P}_2(\mathbb{R}), W_2)$ is a Polish space on which we define absolutely continuous curves by saying that $[0,T]\ni t\to\mu_t\in\mathcal{P}_2(\mathbb{R})$ lies in $AC^2(0,T;\mathcal{P}_2(\mathbb{R}))$ provided that there exists $f\in L^2(0,T)$ such that $W_2(\mu_t,\mu_{t+h})\le\int_t^{t+h} f(s)ds$ for all $0<t<t+h<T$.

4.1. The solution path is locally Lipschitz. First we show that our solution path is locally Lipschitz.

PROPOSITION 4.1. *The solution path $t\to\rho(t,\cdot)$ satisfies the following:*

(i) *For any $0<T<+\infty$, we have*

$$W_2(\rho(t,\cdot),\rho(s,\cdot)) \le C_T|t-s| \quad \textit{for all } t,s\in[0,T].$$

(ii) *The energy is nonincreasing, i.e.,*

$$\int_{\mathbb{R}} |v(t,x)|^2\mathrm{d}\rho(t,x) \le \int_{\mathbb{R}} |v_0(x)|^2\mathrm{d}\rho_0(x) \quad \textit{for all } t\ge 0.$$

Proof. Recall that $\rho^n(t,\cdot)$ are uniformly Lipschitz in n and t. Thus, by the triangle inequality,

$$W_2(\rho(t,\cdot),\rho(s,\cdot)) \le W_2(\rho(t,\cdot),\rho^n(t,\cdot)) + C|t-s| + W_2(\rho^n(s,\cdot),\rho(s,\cdot)).$$

Therefore, by letting n go to infinity and using Proposition 3.3, we obtain (i). Also as $v_n(t,\cdot)\rho^n(t,\cdot) \to v(t,\cdot)\rho(t,\cdot)$ weakly for each t, we deduce that

$$\int_{\mathbb{R}} |v(t,x)|^2 \mathrm{d}\rho(t,x) \le \liminf_{n\to\infty} \int_{\mathbb{R}} |v_n(t,x)|^2 \mathrm{d}\rho^n(t,x) \le \liminf_{n\to\infty} \int_{\mathbb{R}} |v_0(x)|^2 \mathrm{d}\rho_0^n(x)$$

$$= \int_{\mathbb{R}} |v_0(x)|^2 \mathrm{d}\rho_0(x). \qquad \square$$

Remark 4.2. As a consequence of Proposition 4.1, we have $\rho \in AC^2(0,T;\mathcal{P}_2(\mathbb{R}))$.

We now recall a result, slightly modified, proved in [13].

PROPOSITION 4.3. *Suppose $\sigma \in AC^2(0,T;\mathcal{P}_2(\mathbb{R}))$. Let v be the velocity of minimal norm associated to σ and $N(t,\cdot):(0,1)\to\mathbb{R}$ be the monotone nondecreasing map such that $N(t,\cdot)_{\#}\chi_{(0,1)} = \sigma(t,\cdot)$. For each t, modifying $N(t,\cdot)$ on a countable subset of $(0,1)$ if necessary, we may assume w.l.o.g. that $N(t,\cdot)$ is right continuous. Then, $N \in H^1(0,T;L^2(0,1))$ and*

$$\dot{N}(t,x) = v(t,N(t,x)) \tag{31}$$

for $\mathcal{L}^2$-a.e. $(t,x)\in(0,T)\times(0,1)$.

Note that the *minimal norm* assumption is, in fact, redundant. Indeed, we prove the following lemma.

LEMMA 4.4. *Consider the path $t\to\mu\in AC^2(0,T;\mathcal{P}_2(\mathbb{R}))$ for some $0<T<+\infty$. Then the velocity defined in* [2] *(called "of minimal norm") is the unique velocity along the curve μ in the following sense: if*

$$\partial_t\mu + \partial_x(\mu v_i) = 0 \text{ in } \mathcal{D}'([0,T]\times\mathbb{R}),\ i=1,2,$$

for some v_i Borel measurable in (t,x) such that $v_i(t,\cdot)\in L^2(\mu(t,\cdot))$ for Lebesgue a.e. $t\in(0,T)$, then for Lebesgue a.e. $t\in(0,T)$ we have $v_1(t,\cdot)\equiv v_2(t,\cdot)$ in the $\mu(t,\cdot)$-a.e. sense.

Proof. By subtraction and by taking test functions $\varphi(t,x)=\xi(t)\zeta(x)$, the equations above readily yield

$$\int_{\mathbb{R}} u(t,x)\zeta'(x)\mathrm{d}\mu(t,x) = 0 \text{ for a.e. } t\in(0,T) \text{ and any } \zeta\in C_c^1(\mathbb{R}),$$

where $u := v_1 - v_2$. Fix $\varepsilon>0$ and $\phi\in C_c(\mathbb{R})$. If $\phi=0$ on $[R,+\infty)$, consider, for each natural number $n>R$, the function

$$\Phi_n(x) := \begin{cases} \int_{-\infty}^x \phi(y)\mathrm{d}y & \text{if } x<n,\\ \omega(x-n) & \text{if } n\le x\le n+1,\\ 0 & \text{if } x>n+1, \end{cases} \tag{32}$$

where $\omega\in C^1[0,1]$ such that $\omega(0)=\int_{-\infty}^R\phi(y)\mathrm{d}y$, $\omega(1)=0$, and $\omega'(0)=0=\omega'(1)$. Clearly, $\Phi_n\in C_c^1(\mathbb{R})$. Thus,

$$\int_{\mathbb{R}} u(t,x)\phi(x)\mathrm{d}\mu(t,x) + \int_n^{n+1} u(x,t)\omega'(x-n)\mathrm{d}\mu(t,x) = 0 \text{ for a.e. } t\in(0,T).$$

We have $|\omega'(x-n)| \le \|\omega'\|_{L^\infty(0,1)} =: C$ for all $n > R$ and all $x \in (n, n+1)$. Since $\mu(t,\cdot)$ is a Borel probability for Lebesgue a.e. $t \in (0,T)$, we conclude that for such t we have

$$\left| \int_{\mathbb{R}} u(t,x)\phi(x) \mathrm{d}\mu(t,x) \right| \le C\|u(t,\cdot)\|_{L^2(\mu(t,\cdot))}\mu(t,[n,n+1])^{1/2} \le \varepsilon$$

if n is sufficiently large. Due to the arbitrariness of ε and ϕ, the proof is concluded. ∎

Remark 4.5. Note that, in fact, we have just proved that $\{\varphi' : \varphi \in C_c^\infty(\mathbb{R})\}$ is dense in $L^2(\mu)$, *even though* μ may not necessarily have finite p-order moment (for any $p > 0$). Also, as a consequence, the tangent space $\mathcal{T}_\mu\mathcal{P}_2(\mathbb{R})$ [2] is the whole $L^2(\mu)$. This property was brought to our attention by W. Gangbo.

4.2. Recovery of the entropy condition. We shall now prove that the solution we obtained for (1) in the $\alpha = 0 = \beta$ case satisfies the *Oleinik entropy condition*.

THEOREM 4.6. *For Lebesgue a.e.* $t \in (0,T)$ *we have*

$$v(t,x_2) - v(t,x_1) \le \frac{1}{t}(x_2 - x_1) \text{ for } \rho(t,\cdot)\text{-a.e. } x_1 \le x_2. \tag{33}$$

Proof. Let us look back at the discrete problem. Note that

$$N_n(t,\omega) := x_j(t) \text{ whenever } M_n(t,x_j(t)-) \le \omega < M_n(t,x_j(t))$$

is the optimal map such that $N_n(t,\cdot)_{\#\chi_{(0,1)}} = \rho^n(t,\cdot)$. It is known [7], [15] that the discrete problem satisfies the Oleinik entropy condition, i.e.,

$$t\big[v_n(t,x_{i_2}(t)) - v_n(t,x_{i_1}(t))\big] \le x_{i_2}(t) - x_{i_1}(t) \text{ whenever } i_1 \le i_2.$$

Since $v_n(t,x_j(t)) = \dot{x}_j(t)$ away from collision times, we infer that the map $t \to \big[x_{i_2}(t) - x_{i_1}(t)\big]/t$ is piecewise nonincreasing. But this map is continuous, so it is globally nonincreasing, and, due to the definition of N_n, it follows that

$$t \to \frac{1}{t}\big[N_n(t,\omega_2) - N_n(t,\omega_1)\big] \tag{34}$$

is nonincreasing in $(0,T)$ for all $\omega_1 \le \omega_2 \in (0,1)$. Now let $\Delta := \{\omega = (\omega_1,\omega_2) \in [0,1]^2 : \omega_1 \le \omega_2\}$ and $S_n(t,\omega) = N_n(t,\omega_2) - N_n(t,\omega_1)$ be defined on $[0,T]\times\Delta$. It is easy to see that $S_n \in H^1(0,T;L^2(\Delta))$ and that (34) implies

$$\int_0^T\int_\Delta \frac{S_n(t,\omega)}{t}\partial_t\varphi(t,\omega)\mathrm{d}\omega\mathrm{d}t \ge 0 \text{ for all nonnegative } \varphi \in C_c^\infty((0,T)\times\Delta). \tag{35}$$

On the other hand, due to (28), we infer that

$$\|N_n - N\|_{L^1((0,T)\times(0,1))} = \int_0^T W_1(\rho^n(t,\cdot),\rho(t,\cdot))\mathrm{d}t \to 0,$$

where $N(t,\cdot)$ is the optimal map such that $N(t,\cdot)_{\#\chi_{(0,1)}} = \rho(t,\cdot)$. In particular, $(S_n/t) \to (S/t)$ in $\mathcal{D}'((0,T)\times\Delta)$, where $S(t,\omega) = N(t,\omega_2) - N(t,\omega_1)$ if $\omega \in \Delta$. Thus, (35) implies

$$\int_0^T\int_\Delta \frac{S(t,\omega)}{t}\partial_t\varphi(t,\omega)\mathrm{d}\omega\mathrm{d}t \ge 0 \text{ for all nonnegative } \varphi \in C_c^\infty((0,T)\times\Delta).$$

Therefore, $\partial_t[S/t] \le 0$ in the distributional sense, which implies $t\dot{S} - S \le 0$ in the $\mathcal{L}^3$-a.e. sense, i.e.,

$$\dot{N}(t,\omega_2) - \dot{N}(t,\omega_1) \le \frac{1}{t}\big[N(t,\omega_2) - N(t,\omega_1)\big] \text{ for Lebesgue a.e. } (t,\omega) \in (0,T)\times\Delta.$$

Consequently, (31) yields

$$v(t,N(t,\omega_2)) - v(t,N(t,\omega_1)) \le \frac{1}{t}\big[N(t,\omega_2) - N(t,\omega_1)\big] \text{ for } \mathcal{L}^3\text{-a.e. } (t,\omega) \in (0,T)\times\Delta,$$

which, due to $N(t,\cdot)_{\#}\chi_{(0,1)} = \rho(t,\cdot)$, implies (33). ∎

4.3. Formation of shocks. Now let us assume that ρ_0 is atom-free and take N_0 to be the optimal map pushing $\chi_{(0,1)}$ forward to ρ_0.

PROPOSITION 4.7. *Let* $T := \sup\{t \in [0,\infty) : \mathrm{id} + tv_0$ *is nondecreasing on* $\mathrm{spt}(\rho_0)\}$. *If* $T = 0$, *then the solution develops atomic singularities instantaneously. If* $0 < T < +\infty$, *then the solution remains nonatomic before* T *and develops atomic singularities instantaneously after* T. *If* $T = +\infty$, *then the solution is atom-free at all times.*

Proof. Let us treat the case $0 < T < +\infty$. Indeed, it will become clear that the other two cases can be handled almost identically. Note that, due to the definition of T and the fact that ρ_0 is nonatomic, $\mathrm{id} + tv_0$ is (strictly) increasing on $\mathrm{spt}(\rho_0)$ for $t \in [0,T)$. Thus, $N_0 + tv_0 \circ N_0$ is increasing on $(0,1)$ for $t \in [0,T)$. It follows that

$$\bar{M}(t,\cdot) := (N_0 + tv_0 \circ N_0)^{-1} = M_0 \circ (\mathrm{id} + tv_0)^{-1}$$

is the entropy solution (given by characteristics) of

$$\partial_t \bar{M} + \partial_x[F(\bar{M})] = 0, \quad \bar{M}(0,\cdot) = M_0,$$

where $F' = v_0 \circ N_0$, $F(0) = 0$. Due to uniqueness of the entropy solution, we get $\bar{M} \equiv M$. Thus, $\rho(t,\cdot) = (\mathrm{id} + tv_0)_{\#}\rho_0$ for $t \in [0,T]$ (this is precisely the geodesic connecting ρ_0 and ρ_T in $\mathcal{P}_2(\mathbb{R})$). Again, since $\mathrm{id}+tv_0$ is (strictly) increasing on $\mathrm{spt}(\rho_0)$ for $t \in [0,T)$, we deduce that $\rho(t,\cdot)$ has no atoms if $t \in [0,T)$. To conclude, we argue by contradiction and suppose that $\rho(t,\cdot)$ is atom-free on $[0,T+\epsilon]$ for some $\epsilon > 0$. Thus, $M(t,\cdot)$ is continuous, and so $v(t,\cdot) = v_0 \circ N_0 \circ M(t,\cdot)$ for all $t \in [0,T+\epsilon]$. Since $M(t,\cdot)$ is now the optimal map pushing $\rho(t,\cdot)$ forward to $\chi_{(0,1)}$, we infer that

$$v(t,N(t,m)) = v_0 \circ N_0 \circ M(t,N(t,m)) = v_0 \circ N_0(m) \text{ a.e. } m \in (0,1),$$

which, in light of (31), yields $\dot{N}(t,m) = v_0 \circ N_0(m)$ for a.e. $(t,m) \in (0,T+\epsilon)\times(0,1)$. Thus, as an $H^1(0,T+\epsilon;L^2(0,1))$ map, $N(t,\cdot) = N_0 + tv_0 \circ N_0$. In particular, we obtain that $\mathrm{id}+tv_0$ is nondecreasing on the support of ρ_0 for all $t \in [0,T+\epsilon]$, which contradicts the definition of T. Therefore, the solution becomes atomic instantaneously after T. ∎

Remark 4.8. What happens at $t = T$ depends on whether or not $\mathrm{id} + Tv_0$ has "flat spots" on $\mathrm{spt}(\rho_0)$. It is easy to construct examples illustrating that each of these situations may occur.

4.4. Continuity of the energy and a remark on uniqueness.

PROPOSITION 4.9. *Suppose* (ρ_0, v_0) *satisfies the conditions* (H1) *and* (H2). *Let* (ρ, v) *be the weak solution to the system* (1) *given in the proof of Theorem* 1.1 *in subsection* 3.3. *Then* (ρ, v) *has the following property:*

$$\lim_{t\to 0^+} \int_{\mathbb{R}} v(t,x)\varphi(x)\mathrm{d}\rho(t,x) = \int_{\mathbb{R}} v_0(x)\varphi(x)\mathrm{d}\rho_0(x) \quad \textit{for all} \quad \varphi \in C_b(\mathbb{R}), \tag{36}$$

which shows that the initial condition for the velocity is satisfied. Moreover, we have

$$\lim_{t\to 0^+}\int_{\mathbb{R}} v^2(t,x)\varphi(x)\mathrm{d}\rho(t,x)=\int_{\mathbb{R}} v_0^2(x)\varphi(x)\mathrm{d}\rho_0(x)\quad \textit{for all}\quad \varphi\in C_b(\mathbb{R}). \tag{37}$$

Proof. Due to $\|M(t,\cdot)-M_0\|_{L^1(\mathbb{R})}=W_1(\rho(t,\cdot),\rho_0)\to 0$ and the BV calculus, we have for any function $\varphi\in C_c^\infty(\mathbb{R})$

$$\begin{aligned}\lim_{t\to 0^+}\int_{\mathbb{R}} v(t,x)\varphi(x)\mathrm{d}\rho(t,x)&=-\lim_{t\to 0^+}\int_{\mathbb{R}}\varphi'(x)F(M(t,x))\mathrm{d}x\\&=-\int_{\mathbb{R}}\varphi'(x)F(M_0(x))\mathrm{d}x=\int_{\mathbb{R}}\varphi(x)g(x)\mathrm{d}\rho_0(x),\end{aligned} \tag{38}$$

where $g(x)$ is given by

$$g(x)=\int_0^1 v_0\circ N_0\left((1-s)M_0(x-)+sM_0(x)\right)\mathrm{d}s,\quad x\in\mathbb{R}.$$

We claim that $N_0\left((1-s)M_0(x-)+sM_0(x)\right)=x$ for ρ_0-a.e. $x\in\mathbb{R}$. Indeed, if a point x satisfies $\rho_0(\{x\})\neq 0$, then M_0 has a jump at x. Therefore, it is clear in this case that

$$N_0\left((1-s)M_0(x-)+sM_0(x)\right)=\inf\left\{z: M_0(z)>(1-s)M_0(x-)+sM_0(x)\right\}=x.$$

As M_0 has at most countably many "flat spots," the claim shall be proved if we can show that $g(x)=v_0(x)$ whenever x satisfies $x\in\mathrm{spt}(\rho_0)$, $\rho_0(\{x\})=0$ and x is not one of the endpoints of some flat spot of M_0. To see this is true, let x be a point with such properties. Then there exists $\epsilon>0$ such that the function M_0 is strictly increasing on $(x-\epsilon,x+\epsilon)$. It follows that

$$N_0\left((1-s)M_0(x-)+sM_0(x)\right)=N_0\left(M_0(x)\right)=x.$$

Thus, we obtain the claim, which in turn yields $g(x)=v_0(x)$ for ρ_0-a.e. $x\in\mathbb{R}$. Hence, by combining this with (38), we get

$$\lim_{t\to 0^+}\int_{\mathbb{R}} v(t,x)\varphi(x)\mathrm{d}\rho(t,x)=\int_{\mathbb{R}} v_0(x)\varphi(x)\mathrm{d}\rho_0(x)\quad\text{for all}\quad\varphi\in C_c^\infty(\mathbb{R}).$$

By a simple approximation and the fact that $\|v_0\|_{L^2(\rho_0)}$ is finite, this gives (36). As a consequence of (36) and the fact that the energy is nonincreasing which was proved in Proposition 4.1, we obtain (see, e.g., Theorem 5.4.4 in [2])

$$\lim_{t\to 0^+}\int_{\mathbb{R}} v^2(t,x)\mathrm{d}\rho(t,x)=\int_{\mathbb{R}} v_0^2(x)\mathrm{d}\rho_0(x).$$

In addition, we have $W_2(\rho(t,\cdot),\rho_0)\le Ct$ from Proposition 4.1. Therefore, by using Proposition 7.1.5 and Theorem 5.4.4 in [2] we can conclude that (37) holds. ∎

We end the paper by the following remark on the uniqueness of the solution.

Remark 4.10. If we assume, in addition, that v_0 is bounded, then the weak solution (ρ,v) constructed above is the unique weak solution satisfying the entropy condition in Theorem 4.6 and the weak convergence of $v^2(t,\cdot)\,\rho(t,\cdot)$ to $v_0^2\rho_0$ (such a weak solution is called an entropy solution for pressureless gases [16]). This fact follows directly from Theorem 4.6, Proposition 4.9, and the uniqueness result in [16]. Thus, if $v_0\in C_b(\mathbb{R})$, we have proved that any entropy solution to the pressureless gases system can be obtained as a weak limit of sticky particles. We note that a similar result was obtained by Bouchut and James in [5] for duality solutions.

Acknowledgments. The authors would like to thank W. Gangbo, R. Pan, and A. Swiech for numerous and fruitful discussions. We are also grateful to both anonymous referees for their helpful comments and suggestions.

REFERENCES

[1] L. Ambrosio and W. Gangbo, *Hamiltonian ODEs in the Wasserstein spaces of probability measures*, Comm. Pure Appl. Math., 61 (2008), pp. 18–53.
[2] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows in Metric Spaces and the Wasserstein Spaces of Probability Measures*, Lectures Math. ETH Zürich, Birkhäuser, Basel, 2005.
[3] A. Andrievskii, S. Gurbatov, and A. Sobolevskii, *Ballistic aggregation in symmetric and nonsymmetric flows*, J. Exp. Theor. Phys., 104 (2007), pp. 887–896.
[4] J. D. Benamou and Y. Brenier, *A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem*, Numer. Math., 84 (2000), pp. 375–393.
[5] F. Bouchut and F. James, *Duality solutions for pressureless gases, monotone scalar conservation laws, and uniqueness*, Comm. Partial Differential Equations, 24 (1999), pp. 2173–2189.
[6] L. Boudin, *A solution with bounded expansion rate to the model of viscous pressureless gases*, SIAM J. Math. Anal., 32 (2000), pp. 172–193.
[7] Y. Brenier and E. Grenier, *Sticky particles and scalar conservation laws*, SIAM J. Numer. Anal., 35 (1998), pp. 2317–2328.
[8] H. Brezis, *Analyse fonctionnelle; théorie et applications*, Collection Mathématiques Appliquées pour la Maitrise, Masson, Paris, 1983.
[9] S. Cheng, J. Li, and T. Zhang, *Explicit construction of measure solutions of Cauchy problem for transportation equations*, Sci. China Ser. A, 40 (1997), pp. 1287–1299.
[10] C. Dellacherie and P. A. Meyer, *Probabilités et potentiel*, Hermann, Paris, 1975.
[11] R. M. Dudley, *Real Analysis and Probability*, Cambridge Stud. Adv. Math. 74, Cambridge University Press, Cambridge, UK, 2002.
[12] M. Fréchet, *Sur la distance de deux lois de probabilité*, C. R. Acad. Sci. Paris, 244 (1957), pp. 689–692.
[13] W. Gangbo, T. Nguyen, and A. Tudorascu, *Euler-Poisson systems as action-minimizing paths in the Wasserstein space*, Arch. Ration. Mech. Anal., to appear.
[14] W. Gangbo and T. Pacini, *Infinite dimensional Hamiltonian systems in terms of the Wasserstein distance*, in progress.
[15] E. Grenier, *Existence globale pour le système des gaz sans pression*, C. R. Acad. Sci. Paris Sér. I Math., 321 (1995), pp. 171–174.
[16] F. Huang and Z. Wang, *Well posedness for pressureless flow*, Comm. Math. Phys., 222 (2001), pp. 117–146.
[17] P. Martin and J. Piasecki, *One-dimensional ballistic aggregation: Rigorous long-time estimates*, J. Statist. Phys., 76 (1994), pp. 447–476.
[18] P. Martin and J. Piasecki, *Aggregation dynamics in a self-gravitating one-dimensional gas*, J. Statist. Phys., 84 (1996), pp. 837–857.
[19] G. Monge, *Mémoire sur la théorie des déblais et de remblais*, Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année, 1781, pp. 666–704.
[20] A. Shnirelman, *On the principle of the shortest way in the dynamics of systems with constraints*, in Global Analysis—Studies and Applications II, Lecture Notes in Math. 1214, Springer-Verlag, Berlin, 1986, pp. 117–130.
[21] C. Villani, *Topics in optimal transportation*, Grad. Stud. Math. 58, AMS, Providence, RI, 2003.
[22] A. Volpert, *Spaces BV and quasilinear equations*, Mat. Sb. (N.S.), 73 (1967), pp. 255–302.
[23] E. Weinan, Y. Rykov, and Y. Sinai, *Generalized variational principles, global weak solutions and behavior with random initial data for systems of conservation laws arising in adhesion particle dynamics*, Comm. Math. Phys., 177 (1996), pp. 349–380.
[24] Ya. B. Zeldovich, *Gravitational instability: An approximate theory for large density perturbations*, Astro. & Astrophys., 5 (1970), pp. 84–89.

SIAM J. MATH. ANAL.
Vol. 40, No. 2, pp. 776–789

SPREADING SPEEDS AND TRAVELING WAVES FOR NONMONOTONE INTEGRODIFFERENCE EQUATIONS*

SZE-BI HSU† AND XIAO-QIANG ZHAO‡

Abstract. The spreading speeds and traveling waves are established for a class of nonmonotone discrete-time integrodifference equation models. It is shown that the spreading speed is linearly determinate and coincides with the minimal wave speed of traveling waves.

Key words. integrodifference equations, nonmonotone integral operators, spreading speeds, linear determinacy, traveling waves

AMS subject classifications. 37L15, 39A11, 92D25

DOI. 10.1137/070703016

1. Introduction.

The invasion speed is a fundamental characteristic of biological invasions, since it describes the speed at which the geographic range of the population expands; see, e.g., [6, 8, 9, 15] and references therein. Aronson and Weinberger [1, 2] first introduced the concept of the asymptotic speed of spread (in short, spreading speed) for reaction-diffusion equations and showed that it coincides with the minimal wave speed for traveling waves under appropriate assumptions. Weinberger [20] and Lui [13] established the theory of spreading speeds and monostable traveling waves for monotone (order-preserving) operators. This theory has been greatly developed recently in [21, 10, 11, 12] to monotone semiflows so that it can be applied to various discrete- and continuous-time evolution equations admitting the comparison principle.

It is known that many discrete- and continuous-time population models with spatial structure are not monotone. For example, scalar discrete-time integrodifference equations with nonmonotone growth functions, and predator-prey type reaction-diffusion systems are among such models. The spreading speeds were obtained for some nonmonotone continuous-time integral equations and time-delayed reaction-diffusion models in [17, 19], and a general result on the nonexistence of traveling waves was also given in [19, Theorem 3.5]. The existence of monostable traveling waves was established for several classes of nonmonotone time-delayed reaction-diffusion equations in [22, 4, 16, 14]. For certain types of nonmonotone discrete-time integrodifference equation models, nonmonotone traveling waves and even traveling cycles were observed in [7] by numerical simulations. In [7, 9, 15, 8], the monotone linear systems, resulting from the linearization of the nonmonotone discrete-time models at zero, were used to estimate spreading speeds. It is worthy to find sufficient conditions under which the spreading speed is linearly determinate for these nonmonotone systems.

*Received by the editors September 16, 2007; accepted for publication (in revised form) April 22, 2008; published electronically August 1, 2008.
http://www.siam.org/journals/sima/40-2/70301.html
†Department of Mathematics, National Tsing Hua University, Hsinchu, Taiwan, Republic of China (sbhsu@math.nthu.edu.tw). Research supported in part by the National Council of Science of Republic of China.
‡Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, NL A1C 5S7, Canada (xzhao@math.mun.ca). Research supported in part by the NSERC of Canada and the MITACS of Canada.

The purpose of our current paper is to study the spreading speeds and traveling waves for nonmonotone discrete-time systems. As a starting point, we consider scalar integrodifference equations with nonmonotone growth functions. The key techniques are to sandwich the given growth function in between two appropriate nondecreasing functions (for spreading speeds) and to construct a closed and convex subset in an appropriate Banach space (for traveling waves). Consequently, we obtain a set of sufficient conditions for the existence of the spreading speed, and the existence and nonexistence of traveling waves. It turns out that the spreading speed is linearly determinate and coincides with the minimal wave speed of traveling waves for this class of nonmonotone discrete-time integrodifference equation population models.

The rest of this paper is organized as follows. In section 2, we first present a general result for monotone integrodifference equations, then we establish the spreading speed c^* by the comparison method and a fluctuation type argument. In section 3, we use the Schauder fixed point theorem to obtain the existence of traveling waves with the wave speed $c > c^*$. The property of the spreading speed is employed to prove the asymptotic property of the wave profile at $+\infty$ and the nonexistence of traveling waves with $c < c^*$. A limiting argument gives the existence of the traveling wave with the wave speed c^*. Section 4 is aimed at the applications of the main results to three types of growth functions arising from population biology.

2. Spreading speeds. Let $\mathcal{C}$ be the space of all bounded and continuous functions from $\mathbb{R}$ to $\mathbb{R}$ equipped with the compact open topology. For a given number $r > 0$, let $\mathcal{C}_r := \{\phi \in \mathcal{C} : 0 \le \phi(x) \le r, \forall x \in \mathbb{R}\}$.

Let $k(x)$ be a nonnegative Lebesgue measurable function on $\mathbb{R}$. Throughout this paper, we assume that the kernel $k(x)$ has the following property:

(K) $\int_{\mathbb{R}} k(y)dy = 1$, $k(-y) = k(y)$, $\forall y \in \mathbb{R}$, and $\int_{\mathbb{R}} e^{-\alpha y}k(y)dy < \infty$, $\forall \alpha \in [0, \Delta)$, where $\Delta > 0$ is the abscissa of convergence and it may be infinity.

We consider a discrete-time integrodifference equation

$$u_{n+1}(x) = \int_{\mathbb{R}} h(u_n(y))k(x-y)dy, \ x \in \mathbb{R}, \ n \ge 0 \tag{2.1}$$

with $u_0 \in \mathcal{C}$. Assume that there exists $\beta > 0$ such that

(H1) $h \in C([0,\beta],[0,\beta])$, $h(0) = 0$, $h'(0) > 1$, $h(\beta) = \beta$, and there is $L_0 > 0$ such that $|h(u_1) - h(u_2)| \le L_0|u_1 - u_2|$, $\forall u_1, u_2 \in [0,\beta]$.

(H2) $u < h(u) \le h'(0)u$, $\forall u \in (0,\beta)$, and $h(u)$ is nondecreasing in $u \in [0,\beta]$.

Let $U(x)$ be a continuous function on $\mathbb{R}$. We say $U(x + cn)$ is a traveling wave solution of (2.1) with the wave speed c if $u_n(x) = U(x + cn)$, $\forall n \ge 0$, satisfies (2.1), and $U(x + cn)$ connects 0 to β if $U(-\infty) = 0$ and $U(+\infty) = \beta$. It is easy to see that $U(x + cn)$ is a traveling wave solution of (2.1) if and only if

$$U(\xi) = \int_{\mathbb{R}} h(U(\xi - c - y))k(y)dy, \quad \forall \xi \in \mathbb{R}.$$

Define

$$c_h^* = \inf_{\mu \in (0,\Delta)} \frac{\ln\left(h'(0)\int_{\mathbb{R}} e^{-\mu y}k(y)dy\right)}{\mu}. \tag{2.2}$$

The following result is essentially due to Weinberger [20], and shows that c_h^* is not only the spreading speed but also the minimal wave speed of monotone traveling waves for system (2.1).

THEOREM 2.1. *Let* (H1) *and* (H2) *hold. Then the following statements are valid:*

(i) *For any $u_0 \in \mathcal{C}_\beta$ with compact support, the solution of* (2.1) *satisfies*

$$\lim_{n\to\infty, |x|\geq cn} u_n(x) = 0, \ \forall c > c_h^*.$$

(ii) *For any $u_0 \in \mathcal{C}_\beta \setminus \{0\}$, the solution of* (2.1) *satisfies*

$$\lim_{n\to\infty, |x|\leq cn} u_n(x) = \beta, \ \forall c \in (0, c_h^*).$$

(iii) *For any $c \geq c_h^*$,* (2.1) *has a traveling wave $U(x + cn)$ connecting 0 to β such that $U(x)$ is nondecreasing in x, and for any $c \in (0, c_h^*)$,* (2.1) *has no traveling wave $U(x + cn)$ connecting 0 to β.*

Proof. The existence of the spreading speed, together with the formula (2.2), and traveling waves is a straightforward consequence of [20, Theorems 6.1–6.6] in the case where $\Delta = +\infty$, and [11, Theorem 2.11 and Theorem 2.15] with $\tau = 0$ in the case where $\Delta < +\infty$. Indeed, the proof of [11, Theorem 3.10] implies that [11, Theorem 3.10] with $\inf_{\mu>0} \Phi(\mu)$ replaced by $\inf_{0<\mu<\Delta} \Phi(\mu)$ is still valid, provided that (C5) holds for all $\mu_1, \mu_2 \in (-\Delta, \Delta)$ and $\Phi(\mu)$ assumes its minimum value at $\mu^* \in (0, \Delta)$. Thus, the formula (2.2) also holds in the case where $\Delta < +\infty$. By [11, Theorem 3.5], it follows that the number $r = r_\sigma$ in [11, Theorem 2.15] can be chosen to be independent of $\sigma > 0$. For any $u_0 \in \mathcal{C}_\beta \setminus \{0\}$, it is easy to show that there exists an integer $n_0 \geq 1$ such that $u_{n_0}(x) > 0$ for x in an interval of length greater $2r$. Taking $u_{n_0}(x)$ as a new initial data, we see that conclusion (ii) holds. The nonexistence of traveling waves is implied by conclusion (ii) (see also [11, Theorem 4.1]). □

Next we consider the discrete-time integrodifference equation

$$u_{n+1}(x) = \int_{\mathbb{R}} f(u_n(y))k(x-y)dy, \quad x \in \mathbb{R}, \ n \geq 0 \tag{2.3}$$

with $u_0 \in \mathcal{C}$. Assume that there exists $b > 0$ such that

(F1) $f \in C([0, b], [0, b])$, $f(0) = 0$, $f'(0) > 1$, and there is $L > 0$ such that $|f(u_1) - f(u_2)| \leq L|u_1 - u_2|$, $\forall u_1, u_2 \in [0, b]$.

(F2) $f(u) \leq f'(0)u$, $\forall u \in [0, b]$, and there is $u^* \in (0, b]$ such that $f(u^*) = u^*$, $f(u) > u$, $\forall u \in (0, u^*)$, and $0 < f(u) < u$, $\forall u \in (u^*, b]$.

Define

$$f^+(u) = \max_{0\leq v\leq u} f(v), \quad f^-(u) = \min_{u\leq v\leq b} f(v), \quad \forall u \in [0, b].$$

It then follows that

$$f^-(u) \leq f(u) \leq f^+(u), \ \forall u \in [0, b],$$

that both f^+ and f^- are nondecreasing and Lipschitz continuous, with the Lipschitz constant L, on $[0, b]$, and that there exists $\delta_0 \in (0, b]$ such that $f^\pm(u) = f(u)$, $\forall u \in [0, \delta_0]$. Let $u_\pm^*$ be such that $f^\pm(u_\pm^*) = u_\pm^*$. Then $0 < u_-^* \leq u^* \leq u_+^* \leq b$.

To obtain the upward convergence as stated in Theorem 2.1 (ii), we need to impose one of the following two additional conditions on f.

(C1) $u^* = b$ and $f(u)$ is nondecreasing in $u \in [b - \epsilon_0, b]$ for some $\epsilon_0 \in (0, b)$.

(C2) $\frac{f(u)}{u}$ is strictly decreasing for $u \in (0, b]$, and $f(u)$ has the property (P) that for any $v, w \in (0, b]$ satisfying $v \leq u^* \leq w$, $v \geq f(w)$ and $w \leq f(v)$, we have $v = w$.

Motivated by the proofs of [17, Theorems 2.9 and 2.12], we have the following observation.

LEMMA 2.1. *Either of the following two conditions is sufficient for the property* (P) *in condition* (C2) *to hold:*

(P1) $uf(u)$ *is strictly increasing for* $u \in (0, b]$.

(P2) $f(u)$ *is nonincreasing for* $u \in [u^*, b]$, *and* $\frac{f^2(u)}{u}$ *is strictly decreasing for* $u \in (0, u^*]$.

Proof. Let $v, w \in (0, b]$ be given such that $v \le u^* \le w$, $v \ge f(w)$, and $w \le f(v)$. In the case where (P1) holds, since $vf(v) \ge f(w)f(v) \ge wf(w)$, it follows that $v \ge w$, and hence $v = w$. In the case where (P2) holds, we have $v \ge f(w) \ge f(f(v)) = f^2(v)$, and hence, $\frac{f^2(v)}{v} \le 1 = \frac{f^2(u^*)}{u^*}$. It then follows that $v \ge u^*$, and hence, $v = u^*$. Since $u^* \le w \le f(v) = f(u^*) = u^*$, we further have $w = u^* = v$. ∎

Now we are in a position to prove the main result of this section.

THEOREM 2.2. *Let* (F1) *and* (F2) *hold and* c_f^* *be defined as in* (2.2) *with* $h = f$. *Then the following statements are valid:*

(1) *For any* $u_0 \in \mathcal{C}_{u_+^*}$ *with compact support, the solution of* (2.3) *satisfies*

$$\lim_{n\to\infty, |x|\ge cn} u_n(x) = 0, \ \forall c > c_f^*.$$

(2) *For any* $u_0 \in \mathcal{C}_{u_+^*} \setminus \{0\}$, *the solution of* (2.3) *satisfies*

$$u_-^* \le \liminf_{n\to\infty, |x|\le cn} u_n(x) \le \limsup_{n\to\infty, |x|\le cn} u_n(x) \le u_+^*, \ \forall c \in (0, c_f^*).$$

(3) *If, in addition, either* (C1) *or* (C2) *holds, then for any* $u_0 \in \mathcal{C}_{u_+^*} \setminus \{0\}$, *the solution of* (2.3) *satisfies*

$$\lim_{n\to\infty, |x|\le cn} u_n(x) = u^*, \ \forall c \in (0, c_f^*).$$

Proof. For convenience, let $c^* = c_f^*$. Define

$$Q(\phi)(x) = \int_{\mathbb{R}} f(\phi(x-y))k(y)dy, \quad Q^{\pm}(\phi)(x) = \int_{\mathbb{R}} f^{\pm}(\phi(x-y))k(y)dy.$$

Clearly, $Q^{\pm}$ is monotone (order preserving) on $\mathcal{C}_b$ and

$$Q^-(\phi) \le Q(\phi) \le Q^+(\phi), \quad \forall \phi \in \mathcal{C}_b.$$

By Theorem 2.1, it follows that c^* is the spreading speed for the discrete-time system $u_{n+1} = Q^{\pm}(u_n)$ on $\mathcal{C}_{u_\pm^*}$.

Case 1. For a given $\phi \in \mathcal{C}_{u_+^*}$ with compact support, let $u_n = Q^n(\phi)$, $u_n^+ = (Q^+)^n(\phi)$, $\forall n \ge 0$. By the comparison principle (see, e.g., [20, Proposition 4.1]), we have

$$0 \le u_n(x) \le u_n^+(x), \quad \forall x \in \mathbb{R}, \ n \ge 0.$$

For any $c > c^*$, Theorem 2.1 (i) implies that $\lim_{n\to\infty, |x|\ge cn} u_n^+(x) = 0$, and hence $\lim_{n\to\infty, |x|\ge cn} u_n(x) = 0$.

Case 2. For a given $\phi \in \mathcal{C}_{u_+^*} \setminus \{0\}$, define $\psi(x) = \min(\phi(x), u_-^*)$. Then $\psi \in \mathcal{C}_{u_-^*} \setminus \{0\}$. Let $u_n^- = (Q^-)^n(\psi)$, $\forall n \ge 0$. Since $\psi \le \phi$, it follows from the comparison principle that

$$0 \le u_n^-(x) \le u_n(x) \le u_n^+(x), \quad \forall x \in \mathbb{R}, \ n \ge 0.$$

For any $c \in (0, c^*)$, Theorem 2.1 (ii) implies that $\lim_{n\to\infty, |x|\le cn} u_n^{\pm}(x) = u_{\pm}^*$. Thus, we have

$$u_-^* \le \liminf_{n\to\infty, |x|\le cn} u_n(x) \le \limsup_{n\to\infty, |x|\le cn} u_n(x) \le u_+^*.$$

Case 3. In the case where (C1) holds, we see that $u^* \le u_+^* \le b = u^*$. Further, it follows from the definition of f^- that $f^-(u) = f(u), \forall u \in [b - \epsilon_0, b]$, and hence, $u_-^* = u^*$. Since $u_-^* = u_+^* = u^*$, the upward convergence in case (3) follows from the conclusion in case (2).

In the case where (C2) holds, we use similar arguments as in the proof of [17, Lemma 3.10] (see also the proof of [19, Theorem 2.5]). For any $(v, w) \in [0, b]^2$, let

$$g(v,w) = \begin{cases} \min\{f(u) : v \le u \le w\}, & \text{if } v \le w, \\ \max\{f(u) : w \le u \le v\}, & \text{if } w \le v. \end{cases} \tag{2.4}$$

Then $g(v, w)$ is nondecreasing in $v \in [0, b]$ and nonincreasing in $w \in [0, b]$. Moreover, $f(u) = g(u, u)$, and $g(v, w)$ is continuous in $(v, w) \in [0, b]^2$ (see [18, section 2]). For a given $u_0 \in \mathcal{C}_{u_+^*} \setminus \{0\}$, let $u_n = Q^n(u_0)$, $\forall n \ge 0$. Then we have

$$u_{n+1}(x) = \int_{\mathbb{R}} g(u_n(x-y), u_n(x-y))k(y)dy, \ \ n \ge 0.$$

For any $\beta \in (0, c^*)$, we define

$$U_*(\beta) := \liminf_{n\to\infty, |x|\le \beta n} u_n(x), \quad U^*(\beta) := \limsup_{n\to\infty, |x|\le \beta n} u_n(x).$$

Let $c \in (0, c^*)$ be given. We fix a number $\gamma \in (c, c^*)$ and define

$$V_*(c,\gamma) = \inf_{c<\beta<\gamma} U_*(\beta), \quad V^*(c,\gamma) = \sup_{c<\beta<\gamma} U^*(\beta).$$

It then follows that

$$V_*(c,\gamma) \le U_*(\beta) \le U^*(\beta) \le V^*(c,\gamma), \quad \forall \beta \in [c,\gamma].$$

By the conclusion in case 2, we have

$$0 < u_-^* \le U_*(\beta) \le U^*(\beta) \le u_+^*, \quad \forall \beta \in (0, c^*),$$

and hence,

$$0 < u_-^* \le V_*(c,\gamma) \le V^*(c,\gamma) \le u_+^* \le b.$$

For any $\beta \in (c, \gamma)$, we choose two sequences $n_j \to \infty$ and $x_j \in \mathbb{R}$ with $|x_j| \le \beta n_j$ such that $\lim_{j\to\infty} u_{n_j}(x_j) = U_*(\beta)$. It is easy to see that

$$U_*(\gamma) \le \liminf_{j\to\infty} u_{n_j-1}(x_j - y) \le \limsup_{j\to\infty} u_{n_j-1}(x_j - y) \le U^*(\gamma), \ \forall y \in \mathbb{R}.$$

Since $u_{n_j}(x_j) = \int_{\mathbb{R}} g(u_{n_j-1}(x_j - y), u_{n_j-1}(x_j - y))k(y)dy$, it follows from Fatou's lemma that

$$U_*(\beta) \ge \int_{\mathbb{R}} \liminf_{j\to\infty} g(u_{n_j-1}(x_j - y), u_{n_j-1}(x_j - y))k(y)dy,$$

and hence

$$\begin{aligned} U_*(\beta) &\geq \int_{\mathbb{R}} g(U_*(\gamma), U^*(\gamma))k(y)dy \\ &= g(U_*(\gamma), U^*(\gamma)) \geq g(V_*(c,\gamma), V^*(c,\gamma)). \end{aligned}$$

Similarly, we have

$$U^*(\beta) \leq g(U^*(\gamma), U_*(\gamma)) \leq g(V^*(c,\gamma), V_*(c,\gamma)).$$

Thus,

$$V_*(c,\gamma) \geq g(V_*(c,\gamma), V^*(c,\gamma)), \quad V^*(c,\gamma) \leq g(V^*(c,\gamma), V_*(c,\gamma)). \tag{2.5}$$

By the definition of function g, we can find $v, w \in [V_*(c,\gamma), V^*(c,\gamma)] \subset (0, b]$ such that

$$g(V^*(c,\gamma), V_*(c,\gamma)) = f(v) \qquad \text{and} \qquad g(V_*(c,\gamma), V^*(c,\gamma)) = f(w).$$

It then follows from (2.5) that

$$f(w) \leq V_*(c,\gamma) \leq v,\ w \leq V^*(c,\gamma) \leq f(v), \tag{2.6}$$

and hence,

$$\frac{f(w)}{w} \leq 1 = \frac{f(u^*)}{u^*} \leq \frac{f(v)}{v}.$$

This, together with the strict monotonicity of $\frac{f(u)}{u}$ on $(0, b]$, implies that $v \leq u^* \leq w$. By (2.6) and the property (P), we obtain $v = w$. It then follows from (2.6) that $V_*(c,\gamma) \geq f(w) = f(v) \geq V^*(c,\gamma)$, and hence, $0 < V_*(c,\gamma) = V^*(c,\gamma) \leq b$. From (2.5), we have

$$V_*(c,\gamma) \geq g(V_*(c,\gamma), V_*(c,\gamma)) \geq V^*(c,\gamma),$$

and hence, $0 < V_*(c,\gamma) = f(V_*(c,\gamma))$. By the uniqueness of the positive fixed point of f in $[0, b]$, it follows that $V_*(c,\gamma) = u^*$. Consequently,

$$u^* = V_*(c,\gamma) \leq U_*(c) \leq U^*(c) \leq V^*(c,\gamma) = u^*,$$

which implies that $\lim_{n\to\infty, |x|\leq cn} u_n(x) = u^*$ for any $c \in (0, c^*)$. □

Remark 2.1. Under the assumption that $\int_{\mathbb{R}^m} k(y)dy = 1$, $k(x) = k(y)$, $\forall x, y \in \mathbb{R}^m$ with $|x| = |y|$, Theorem 2.2 is still valid if we replace $\int_{\mathbb{R}}$ with $\int_{\mathbb{R}^m}$.

3. Traveling waves. In this section, we establish the existence and nonexistence of traveling waves for systems (2.3) by appealing to the Schauder fixed point theorem and the property of the spreading speed.

For a given $\lambda > 0$, let

$$X_\lambda := \{\phi \in C(\mathbb{R}, \mathbb{R}) : \quad \sup_{\xi\in\mathbb{R}} |\phi(\xi)|e^{-\lambda\xi} < +\infty\}$$

and $\|\phi\|_\lambda = \sup_{\xi\in\mathbb{R}} |\phi(\xi)|e^{-\lambda\xi}$. It then follows that $(X_\lambda, \|\cdot\|_\lambda)$ is a Banach space.

Define

$$\Phi(\lambda) = \frac{\ln\left(f'(0)\int_{\mathbb{R}} e^{-\lambda y}k(y)dy\right)}{\lambda}, \quad \forall \lambda \in (0, \Delta),$$

and

$$K(c,\lambda) = f'(0)e^{-c\lambda}\int_{\mathbb{R}} e^{-\lambda y}k(y)dy, \quad \forall c \in \mathbb{R}_+,\ \lambda \in (0,\Delta).$$

By [11, Lemma 3.8], it follows that for any $c > c_f^*$, there exist $0 < \lambda_1 = \lambda_1(c) < \lambda_2 = \lambda_2(c) < \Delta$ such that $\Phi(\lambda_1) = c$ and $\Phi(\lambda) < c$, $\forall \lambda \in (\lambda_1, \lambda_2)$. Thus, we have

$$K(c,\lambda_1) = 1, \quad K(c,\lambda) < 1,\ \forall \lambda \in (\lambda_1, \lambda_2).$$

Note that if $f''(0)$ exists, then $f(u) \geq f'(0)u - au^2$, $\forall u \in [0,\delta]$, for appropriate $a > 0$ and $\delta > 0$. To obtain the existence of traveling waves, we impose the following weaker condition on f (cf. [3]).

(F3) There exist real numbers $\delta^* \in (0, \min(\delta_0, u_+^*)]$, $\sigma > 1$ and $a > 0$ such that $f(u) \geq f'(0)u - au^\sigma$, $\forall u \in [0,\delta^*]$.

THEOREM 3.1. *Let* (F1)–(F3) *hold. Then the following statements are valid:*

(1) *For any* $c \in (0, c_f^*)$, (2.3) *has no traveling wave* $U(x+cn)$ *with* $U \in \mathcal{C}_{u_+^*} \setminus \{0\}$ *and* $U(-\infty) = 0$.

(2) *For any* $c > c_f^*$, (2.3) *has a traveling wave* $U(x+cn)$ *such that* $U \in \mathcal{C}_{u_+^*} \setminus \{0\}$, $U(-\infty) = 0$, *and*

$$u_-^* \leq \liminf_{\xi\to+\infty} U(\xi) \leq \limsup_{\xi\to+\infty} U(\xi) \leq u_+^*.$$

If, in addition, either (C1) *or* (C2) *holds, then* $U(+\infty) = u^*$.

Proof. Case 1. Assume, by contradiction, that for some $c_0 \in (0, c_f^*)$, (2.3) has a traveling wave $u_n(x) := U(x + c_0 n)$ with $U \in \mathcal{C}_{u_+^*} \setminus \{0\}$ and $U(-\infty) = 0$. By Theorem 2.2 (2), there holds

$$\liminf_{n\to\infty, |x|\leq cn} u_n(x) \geq u_-^* > 0, \quad \forall c \in (0, c_f^*).$$

Choose $\tilde{c} \in (c_0, c_f^*)$ and let $x = -\tilde{c}n$. Then $\liminf_{n\to\infty} u_n(-\tilde{c}n) = \liminf_{n\to\infty} U((c_0 - \tilde{c})n) > 0$, but $\lim_{n\to\infty} U((c_0 - \tilde{c})n) = U(-\infty) = 0$, a contradiction.

Case 2. Let $c > c_f^*$ be given. Define a mapping $T : \mathcal{C}_b \to \mathcal{C}_b$ by

$$T(\phi)(\xi) = \int_{\mathbb{R}} f(\phi(\xi - c - y))k(y)dy, \quad \forall \xi \in \mathbb{R},\ \phi \in \mathcal{C}_b. \tag{3.1}$$

Let $T^\pm$ be defined as in (3.1) with f replaced by $f^\pm$. It then follows that $T^\pm$ is nondecreasing with respect to the pointwise ordering on $\mathcal{C}_b$, and that

$$T^-(\phi) \leq T(\phi) \leq T^+(\phi),\ \forall \phi \in \mathcal{C}_b.$$

Following [3], we define

$$\phi^+(\xi) := \min\{u_+^* e^{\lambda_1 \xi}, u_+^*\}, \quad \forall \xi \in \mathbb{R}.$$

Since $f^+(u)$ is nondecreasing in u and $\phi^+(\xi) \leq u_+^*$, $\forall \xi \in \mathbb{R}$, we obtain

$$T^+(\phi^+)(\xi) \leq \int_{\mathbb{R}} f^+(u_+^*)k(y)dy = f^+(u_+^*) = u_+^*, \quad \forall \xi \in \mathbb{R}. \tag{3.2}$$

Note that $f^+(u) \le f'(0)u$, $\forall u \in [0,b]$, and $\phi^+(\xi) \le u_+^* e^{\lambda_1 \xi}$, $\forall \xi \in \mathbb{R}$. Thus, we further have

$$\begin{aligned} T^+(\phi^+)(\xi) &\le \int_{\mathbb{R}} f'(0)\phi^+(\xi - c - y)k(y)dy \\ &\le f'(0) \int_{\mathbb{R}} u_+^* e^{\lambda_1(\xi - c - y)} k(y)dy \\ &= u_+^* e^{\lambda_1 \xi} K(c, \lambda_1) = u_+^* e^{\lambda_1 \xi}, \ \forall \xi \in \mathbb{R}. \end{aligned} \tag{3.3}$$

By (3.2), (3.3), and the definition of ϕ^+, it then follows that $T^+(\phi^+) \le \phi^+$. We first fix a sufficiently small $\epsilon^* \in (0, \lambda_1(\sigma - 1)]$ such that $\lambda_1 + \epsilon^* < \lambda_2$, and hence $K(c, \lambda_1 + \epsilon^*) < 1$. We then choose a sufficiently large number $M \ge 1$ such that

$$\left(1 + \frac{a(u_+^*)^\sigma}{f'(0)\delta^* M}\right) K(c, \lambda_1 + \epsilon^*) < 1. \tag{3.4}$$

Following [19], we define

$$\phi^-(\xi) := \max\{0, \delta^*(1 - Me^{\epsilon^* \xi})e^{\lambda_1 \xi}\}, \quad \forall \xi \in \mathbb{R}.$$

Let $\xi_0 := -\frac{\ln M}{\epsilon^*}$. Then we have

$$\phi^-(\xi) = 0, \ \forall \xi \ge \xi_0, \quad \phi^-(\xi) = \delta^* e^{\lambda_1 \xi} - \delta^* M e^{(\lambda_1 + \epsilon^*)\xi}, \ \forall \xi \le \xi_0.$$

Since $\delta^* \le u_+^*$, $\xi_0 \le 0$, and $0 < \epsilon^* \le \lambda_1(\sigma - 1)$, it is easy to see that

$$0 \le \phi^-(\xi) \le \phi^+(\xi), \quad (\phi^-(\xi))^\sigma \le (u_+^*)^\sigma e^{(\lambda_1 + \epsilon^*)\xi}, \quad \forall \xi \in \mathbb{R}.$$

Clearly, we have

$$T^-(\phi^-)(\xi) \ge 0, \quad \forall \xi \in \mathbb{R}. \tag{3.5}$$

Since $\phi^-(\xi) \ge \delta^* e^{\lambda_1 \xi} - \delta^* M e^{(\lambda_1 + \epsilon^*)\xi}$, $\forall \xi \in \mathbb{R}$, it follows from (3.4) that

$$\begin{aligned} f'(0)\phi^-(\xi) - a(\phi^-(\xi))^\sigma &\ge f'(0)\delta^* e^{\lambda_1 \xi} - f'(0)\delta^* M e^{(\lambda_1 + \epsilon^*)\xi} - a(u_+^*)^\sigma e^{(\lambda_1 + \epsilon^*)\xi} \\ &= f'(0)\delta^* e^{\lambda_1 \xi} - f'(0)\delta^* M e^{(\lambda_1 + \epsilon^*)\xi} \left(1 + \frac{a(u_+^*)^\sigma}{f'(0)\delta^* M}\right) \\ &\ge f'(0)\delta^* e^{\lambda_1 \xi} - \frac{f'(0)\delta^* M}{K(c, \lambda_1 + \epsilon^*)} e^{(\lambda_1 + \epsilon^*)\xi}, \quad \forall \xi \in \mathbb{R}. \end{aligned}$$

In view of (F3) and the fact that $f^-(u) = f(u)$, $\forall u \in [0, \delta_0]$, we then have

$$\begin{aligned} T^-(\phi^-)(\xi) &\ge \int_{\mathbb{R}} \left(f'(0)\phi^-(\xi - c - y) - a(\phi^-(\xi - c - y))^\sigma\right) k(y)dy \\ &\ge \int_{\mathbb{R}} \left(f'(0)\delta^* e^{\lambda_1(\xi - c - y)} - \frac{f'(0)\delta^* M}{K(c, \lambda_1 + \epsilon^*)} e^{(\lambda_1 + \epsilon^*)(\xi - c - y)}\right) k(y)dy \\ &= \delta^* e^{\lambda_1 \xi} K(c, \lambda_1) - \frac{\delta^* M}{K(c, \lambda_1 + \epsilon^*)} e^{(\lambda_1 + \epsilon^*)\xi} K(c, \lambda_1 + \epsilon^*) \\ &= \delta^* e^{\lambda_1 \xi} - \delta^* M e^{(\lambda_1 + \epsilon^*)\xi}, \quad \forall \xi \in \mathbb{R}. \end{aligned} \tag{3.6}$$

By (3.5), (3.6), and the definition of ϕ^-, it follows that $T^-(\phi^-) \ge \phi^-$.

Now we fix a number $\lambda \in (0, \lambda_1)$. It is easy to see that both ϕ^- and ϕ^+ are elements in X_λ. Thus, the set

$$Y := \{\phi \in X_\lambda : \phi^-(\xi) \le \phi(\xi) \le \phi^+(\xi), \forall \xi \in \mathbb{R}\}$$

is a nonempty, closed, and convex subset of X_λ. For any $\phi \in Y$, we have

$$\phi^- \le T^-(\phi^-) \le T^-(\phi) \le T(\phi) \le T^+(\phi) \le T^+(\phi^+) \le \phi^+,$$

and hence $T(Y) \subset Y$. For any $\phi, \psi \in Y$, there holds

$$\begin{aligned}
\|T(\phi) - T(\psi)\|_\lambda &= \sup_{\xi\in\mathbb{R}} |T(\phi)(\xi) - T(\psi)(\xi)| e^{-\lambda\xi} \\
&\le L \cdot \sup_{\xi\in\mathbb{R}} \int_{\mathbb{R}} |\phi(\xi - c - y) - \psi(\xi - c - y)| e^{-\lambda\xi} k(y) dy \\
&\le L\|\phi - \psi\|_\lambda \int_{\mathbb{R}} e^{-\lambda(c+y)} k(y) dy \\
&= \left(L e^{-\lambda c} \int_{\mathbb{R}} e^{-\lambda y} k(y) dy \right) \|\phi - \psi\|_\lambda.
\end{aligned}$$

This implies that $T : Y \to Y$ is continuous. We further show that $T(Y)$ is precompact in X_λ. For any $\phi \in Y$, $\xi_1, \xi_2 \in \mathbb{R}$, we have

$$\begin{aligned}
|T(\phi)(\xi_1) - T(\phi)(\xi_2)| &= \left| \int_{\mathbb{R}} f(\phi(z)) \left(k(\xi_1 - c - z) - k(\xi_2 - c - z) \right) dz \right| \\
&\le b \int_{\mathbb{R}} |k(\xi_1 - c - z) - k(\xi_2 - c - z)| dz \\
&= b \int_{\mathbb{R}} |k(\xi_1 - \xi_2 + y) - k(y)| dy \\
&= b \cdot g(\xi_1 - \xi_2),
\end{aligned}$$

where $g(\xi) = \int_{\mathbb{R}} |k(\xi + y) - k(y)| dy$, $\forall \xi \in \mathbb{R}$. Since $\lim_{\xi\to 0} g(\xi) = 0$, it follows that the family of functions $\{T(\phi)(\xi) : \phi \in Y\}$ is uniformly bounded and equicontinuous in $\xi \in \mathbb{R}$. Thus, for any given sequence $\{\psi_n\}_{n\ge 1}$ in $T(Y)$, there exist $n_k \to \infty$ and $\psi \in C(\mathbb{R}, \mathbb{R})$ such that $\lim_{k\to\infty} \psi_{n_k}(\xi) = \psi(\xi)$ uniformly for ξ in any compact subset of $\mathbb{R}$. Since $\phi^-(\xi) \le \psi_{n_k}(\xi) \le \phi^+(\xi)$, $\forall \xi \in \mathbb{R}$, we have $\phi^-(\xi) \le \psi(\xi) \le \phi^+(\xi)$, $\forall \xi \in \mathbb{R}$, and hence, $\psi \in Y$. Note that

$$\lim_{\xi\to+\infty} (\phi^+(\xi) - \phi^-(\xi)) e^{-\lambda\xi} = 0$$

and

$$\lim_{\xi\to-\infty} (\phi^+(\xi) - \phi^-(\xi)) e^{-\lambda\xi} = 0.$$

Therefore, for any $\epsilon > 0$, there exists $B > 0$ such that

$$0 \le (\phi^+(\xi) - \phi^-(\xi)) e^{-\lambda\xi} < \epsilon, \quad \forall |\xi| \ge B.$$

Since $\lim_{k\to\infty} (\psi_{n_k}(\xi) - \psi(\xi)) e^{-\lambda\xi} = 0$ uniformly for $\xi \in [-B, B]$, there exists an integer $N > 0$ such that

$$|\psi_{n_k}(\xi) - \psi(\xi)| e^{-\lambda\xi} < \epsilon, \quad \forall \xi \in [-B, B],\ k \ge N.$$

It then follows that

$$\|\psi_{n_k} - \psi\|_\lambda = \sup_{\xi\in\mathbb{R}} |\psi_{n_k}(\xi) - \psi(\xi)|e^{-\lambda\xi} \le \epsilon, \quad \forall k \ge N.$$

This implies that $\lim_{k\to\infty} \psi_{n_k} = \psi$ in X_λ. By the Schauder fixed point theorem, there exists $U \in Y$ such that $U = T(U)$, and hence, $U(x + cn)$ is a traveling wave of (2.3). Since $\phi^-(\xi) \le U(\xi) \le \phi^+(\xi)$, $\forall \xi \in \mathbb{R}$, we have $U(-\infty) = 0$ and $U \in \mathcal{C}_{u_+^*} \setminus \{0\}$.

Let $u_n(x) := U(x + cn)$, $\forall n \ge 0$, and fix a number $\bar{c} \in (0, c_f^*)$. By Theorem 2.2 (2), it follows that

$$0 < u_-^* \le \liminf_{n\to\infty, |x|\le \bar{c}n} u_n(x) \le \limsup_{n\to\infty, |x|\le\bar{c}n} u_n(x) \le u_+^*,$$

and hence,

$$u_-^* \le \liminf_{n\to\infty} u_n(-\gamma n) \le \limsup_{n\to\infty} u_n(-\gamma n) \le u_+^*$$

uniformly for $\gamma \in [0, \bar{c}]$. This implies that

$$u_-^* \le \liminf_{n\to\infty} U(sn) \le \limsup_{n\to\infty} U(sn) \le u_+^*$$

uniformly for $s \in [c - \bar{c}, c]$. Let

$$a_n = n(c - \bar{c}), \quad b_n = nc, \quad \forall n \ge 1.$$

Thus, there exists $N_0 > 0$ such that $a_{n+1} - b_n < 0$, $\forall n \ge N_0$, and hence,

$$\cup_{n\ge m}[a_n, b_n] = [a_m, +\infty), \ \forall m \ge N_0.$$

It then follows that

$$u_-^* \le \liminf_{\xi\to+\infty} U(\xi) \le \limsup_{\xi\to+\infty} U(\xi) \le u_+^*.$$

If, in addition, either (C1) or (C2) holds, then Theorem 2.2 (3) implies that

$$\lim_{n\to\infty, |x|\le\bar{c}n} u_n(x) = u^*, \ \forall \bar{c} \in (0, c_f^*).$$

By the same arguments as above, we further have $U(+\infty) = u^*$. □

THEOREM 3.2. *Let* (F1)–(F3) *hold. Then* (2.3) *has a traveling wave* $U(x + c_f^* n)$ *such that* $U \in \mathcal{C}_{u_+^*} \setminus \{0, u^*\}$ *and*

$$u_-^* \le \liminf_{\xi\to+\infty} U(\xi) \le \limsup_{\xi\to+\infty} U(\xi) \le u_+^*.$$

If, in addition, either (C1) *or* (C2) *holds, then* $U(+\infty) = u^*$.

Proof. Choose a sequence $\{c_j\}_{j\ge1} \subset (c_f^*, +\infty)$ such that $\lim_{j\to\infty} c_j = c_f^*$. By Theorem 3.1 (2), it follows that (2.3) has a traveling wave $U_j(x + c_j n)$ such that $U_j \in \mathcal{C}_{u_+^*} \setminus \{0\}$, $U_j(-\infty) = 0$, and

$$u_-^* \le \liminf_{\xi\to+\infty} U_j(\xi) \le \limsup_{\xi\to+\infty} U_j(\xi) \le u_+^*.$$

Without loss of generality, we assume that $U_j(0) = \frac{1}{2}u_-^* > 0$, $\forall j \geq 1$. Note that

$$U_j(\xi) = \int_{\mathbb{R}} f(U_j(\xi - c_j - y))k(y)dy, \ \forall \xi \in \mathbb{R}, \ j \geq 1. \tag{3.7}$$

It follows that

$$|U_j(\xi_1) - U_j(\xi_2)| \leq b \cdot g(\xi_1 - \xi_2), \ \forall \xi_1, \xi_2 \in \mathbb{R}, \ j \geq 1,$$

where $g(\xi)$ is defined as in the proof of Theorem 3.1. Then the family of functions $\{U_j(\xi) : j \geq 1\}$ is uniformly bounded and equicontinuous in $\xi \in \mathbb{R}$. Thus, there exist $j_k \to +\infty$ and $U \in C(\mathbb{R}, \mathbb{R})$ such that $\lim_{k\to\infty} U_{j_k}(\xi) = U(\xi)$ uniformly for ξ in any compact subset of $\mathbb{R}$. Clearly, $U \in \mathcal{C}_{u_+^*}$ and $U(0) = \frac{1}{2}u_-^*$. Letting $j = j_k \to +\infty$ in (3.7) and using the dominated convergence theorem, we obtain

$$U(\xi) = \int_{\mathbb{R}} f(U(\xi - c_f^* - y))k(y)dy, \ \forall \xi \in \mathbb{R},$$

and hence, $u_n(x) := U(x + c_f^* n)$ is a traveling wave of (2.3). As in the proof of Theorem 3.1 (2), we see that Theorem 2.2 (2) and (3) imply the asymptotic behavior of $U(\xi)$ as $\xi \to +\infty$. □

Compared with Theorem 3.1 (2), we expect that $U(-\infty) = 0$ in Theorem 3.2. However, we are not able to prove it at this moment since the limiting function $U(\xi)$ may not be nondecreasing on $\mathbb{R}$.

4. Examples. In this section, we present illustrative examples by choosing three types of growth functions from population biology.

First, we consider the logistic type function $f(u) = ru(1 - \frac{u}{K})$, $r > 0$, $K > 0$. Clearly, $f'(0) = r$, $\max_{u\in[0,K]} f(u) = f(K/2) = \frac{rK}{4}$, $\frac{f(u)}{u} = r(1 - \frac{u}{K})$ is strictly decreasing on $(0, K]$, and $u^* := K(1 - \frac{1}{r})$ is the unique positive fixed point of f on $[0, K]$. Assume that $1 < r < 4$ so that we have $f'(0) > 1$ and $f((0, K]) \subset (0, K]$. It is easy to verify that $f(u)$ is strictly increasing on $[0, u^*]$ if $r \in (0, 2]$. In the case where $r \in (1, 2]$, we choose $b =: u^*$, and hence, $u_-^* = u_+^* = u^*$. In the case $r \in (2, 4)$, we choose $b := \frac{rK}{4}$, and hence, $u_+^* = b$, $u_-^* = f(b) = \frac{r^2K(4-r)}{16}$. Note that

$$\frac{f^2(u)}{u} = \frac{r^2}{K^3}\left(K^2(K - u) - ru(K - u)^2\right).$$

It then follows that $f(u)$ satisfies the property (P2) if $r \in (2, 3]$. By Theorems 2.1 and 2.2 and Theorems 3.1 and 3.2, we have the following result.

EXAMPLE 4.1. *Let $f(u) = ru(1 - \frac{u}{K})$ with $K > 0$ and $r \in (1, 4)$, b, u_+^* and u_-^* be defined as above, and c_f^* be defined as in* (2.2) *with $h = f$. Then the following statements are valid:*

(i) *c_f^* is the spreading speed of* (2.3) *in the sense that both conclusions* (1) *and* (2) *in Theorem* 2.2 *hold. Further, the conclusion* (3) *in Theorem* 2.2 *holds in the case where $r \in (1, 3]$.*

(ii) *For any $c \in (0, c_f^*)$,* (2.3) *has no traveling wave $U(x + cn)$ with $U \in \mathcal{C}_b \setminus \{0\}$ and $U(-\infty) = 0$, and for any $c \geq c_f^*$,* (2.3) *has a traveling wave $U(x + cn)$ with $U \in \mathcal{C}_{u_+^*} \setminus \{0, u^*\}$ and*

$$u_-^* \leq \liminf_{\xi\to+\infty} U(\xi) \leq \limsup_{\xi\to+\infty} U(\xi) \leq u_+^*.$$

Further, $U(+\infty) = u^*$ in the case where $r \in (1, 3]$. If $r \in (1, 2]$, then $U(-\infty) = 0$ and $U(\xi)$ is nondecreasing in ξ for all $c \geq c_f^*$. If $r \in (2, 3]$, then $U(-\infty) = 0$ for all $c > c_f^*$.

In Example 4.1, we can also verify that for any $r > 0$, $uf(u)$ is strictly increasing on $[0, 2K/3]$ and strictly decreasing on $[2K/3, +\infty)$. It then follows that $f(u)$ satisfies the property (P1) if $r \in (2, 8/3]$, but does not satisfy the property (P1) if $r \in (8/3, 4)$. So we chose to use the property (P2) to obtain the upward convergence as stated in Theorem 2.2 (3) for r in a larger interval $(1, 3]$. By taking u_0 as a constant function in integrodifference equation (2.3), we see that the upward convergence implies that u^* is a globally attractive fixed point for the map f on $(0, u_+^*]$. Note that $|f'(u^*)| = |2 - r| > 1$, $\forall r \in (3, 4)$. By [5, Theorem 3.8], it then follows that for any $r \in (3, 4)$, u^* is a unstable fixed point of f. This implies that the upward convergence does not hold for any $r \in (3, 4)$. Thus, the interval $(1, 3]$ for parameter r is optimal for the upward convergence.

Among other things, Kot [7] observed numerically four types of discrete-time traveling waves for the integrodifference equation (2.3) with $f(u) = (1 + r_0)u - r_0u^2$ and $k(x) = 3e^{-6|x|}$: a simple monotone traveling for $r_0 = 0.9$ ([7, Figure 7]); a traveling wave with damped spatial oscillations for $r_0 = 1.9$ ([7, Figure 8]); a traveling two-cycle for $r_0 = 2.2$ ([7, Figures 9a,b]; and a traveling four-cycle for $r_0 = 2.5$ ([7, Figures 10a–d]. Clearly, this system is a special case of Example 4.1 with $r = 1 + r_0$ and $K = (1 + r_0)/r_0$. It is easy to see that our analytic results in Example 4.1 are consistent with these numerical simulations. Note that there is an increasing sequence of parameter values $r_1 = 3 < r_2 \approx 3.449 < r_3 \approx 3.544 < r_4 \approx 3.564 < \dots$ at which the logistic map $f(u) = ru(1 - \frac{u}{K})$ repeatedly undergoes a period-doubling bifurcation (see, e.g., [5, section 3.5]). Further, when $r \approx 3.839$, f has a unique asymptotically stable periodic orbit of minimal period 3. It is a challenging problem to prove the existence of a traveling three-cycle for the integrodifference equation (2.3) associated with the logistic map when $r \approx 3.839$.

Next, we consider the Ricker type function $f(u) = que^{-pu}$, $q > 1$, $p > 0$. Clearly, $f'(0) = q$ and $\frac{f(u)}{u} = qe^{-pu}$ is strictly decreasing on $(0, +\infty)$. It is easy to see that $u^* := \frac{\ln q}{p}$ is the unique positive fixed point of f on $[0, +\infty)$, that $\max_{u \in [0, +\infty)} f(u) = f(1/p) = \frac{q}{pe}$, and that $f(u)$ is strictly increasing on $[0, u^*]$ if $q \in (1, e]$. In the case where $q \in (1, e]$, we choose $b := u^*$, and hence, $u_-^* = u_+^* = u^*$. In the case where $q > e$, we choose $b := \frac{q}{pe}$, and hence, $u_+^* = b$, $u_-^* = f(b) = \frac{q^2}{pe}e^{-q/e}$. Note that

$$\frac{f^2(u)}{u} = q^2 e^{-p(u + que^{-pu})}.$$

An elementary analysis shows that $f(u)$ satisfies the property (P2) if $q \in (e, e^2]$. By Theorems 2.1 and 2.2 and Theorems 3.1 and 3.2, we have the following result.

EXAMPLE 4.2. *Let $f(u) = que^{-pu}$ with $q > 1$ and $p > 0$, b, u_+^* and u_-^* be defined as above, and c_f^* be defined as in* (2.2) *with $h = f$. Then the following statements are valid:*

(i) *c_f^* is the spreading speed of* (2.3) *in the sense that both conclusions* (1) *and* (2) *in Theorem* 2.2 *hold. Further, conclusion* (3) *in Theorem* 2.2 *holds in the case where $q \in (1, e^2]$.*

(ii) *For any $c \in (0, c_f^*)$,* (2.3) *has no traveling wave $U(x + cn)$ with $U \in \mathcal{C}_b \setminus \{0\}$ and $U(-\infty) = 0$, and for any $c \geq c_f^*$,* (2.3) *has a traveling wave $U(x + cn)$*

with $U \in \mathcal{C}_{u_+^*} \setminus \{0, u^*\}$ *and*

$$u_-^* \le \liminf_{\xi\to+\infty} U(\xi) \le \limsup_{\xi\to+\infty} U(\xi) \le u_+^*.$$

Further, $U(+\infty) = u^*$ *in the case where* $q \in (1, e^2]$. *If* $q \in (1, e]$, *then* $U(-\infty) = 0$ *and* $U(\xi)$ *is nondecreasing in* ξ *for all* $c \ge c_f^*$. *If* $q \in (e, e^2]$, *then* $U(-\infty) = 0$ *for all* $c > c_f^*$.

In Example 4.2, for any $q > e^2$, we have $|f'(u^*)| = |1 - \ln q| > 1$, and hence, u^* is an unstable fixed point of f. As discussed in Example 4.1, it follows that the interval $(1, e^2]$ for parameter q is optimal for the upward convergence.

Finally, we consider the generalized Beverton–Holt-type function $f(u) = \frac{pu}{q+u^m}$, $m > 0$, and $p > q > 0$. Clearly, $f'(0) = p/q$ and $\frac{f(u)}{u} = \frac{p}{q+u^m}$ is strictly decreasing on $(0, +\infty)$. It is easy to see that $u^* := (p-q)^{\frac{1}{m}}$ is the unique positive fixed point of f on $[0, +\infty)$, and that $f(u)$ is strictly increasing on $[0, +\infty)$ in the case where $m \in (0, 1]$. In the case where $m > 1$, we have

$$\max_{u\in[0,+\infty)} f(u) = f(\bar{u}) = \frac{p(m-1)\bar{u}}{qm}, \quad \bar{u} := \left(\frac{q}{m-1}\right)^{\frac{1}{m}}.$$

By elementary analysis, it follows that $f(u)$ is strictly increasing on $[0, u^*]$ if $m \in (1, p/(p-q)]$, that $uf(u)$ is strictly increasing on $[0, +\infty)$ if $m \in (0, 2]$, and that $uf(u)$ is strictly increasing on $[0, (2q/(m-2))^{\frac{1}{m}}]$ if $m > 2$. Define $b := u^*$ if $m \in (0, p/(p-q)]$ and $b := f(\bar{u})$ if $m > p/(p-q)$. It then follows that $u_-^* = u_+^* = u^*$ in the case where $m \in (1, p/(p-q)]$, and

$$u_+^* = b, \quad u_-^* = f(b) = \frac{p^2(m-1)\bar{u}}{q^2 m + \frac{p^m(m-1)^{m-1}q}{(qm)^{m-1}}},$$

in the case where $m > p/(p-q)$. Note that $f(u)$ satisfies the property (P1) if either $m \in (0, 2]$, or $m > \max(2, p/(p-q))$ and $f(\bar{u}) \le (2q/(m-2))^{\frac{1}{m}}$. By Theorems 2.1 and 2.2 and Theorems 3.1 and 3.2, we have the following result.

EXAMPLE 4.3. *Let* $f(u) = \frac{pu}{q+u^m}$ *with* $m > 0$ *and* $p > q > 0$, b, u_+^* *and* u_-^* *be defined as above, and* c_f^* *be defined as in* (2.2) *with* $h = f$. *Then the following statements are valid:*

(i) c_f^* *is the spreading speed of* (2.3) *in the sense that both conclusions* (1) *and* (2) *in Theorem* 2.2 *hold. Further, conclusion* (3) *in Theorem* 2.2 *holds in the case where either* $m \in (0, \max(2, p/(p-q))]$, *or* $m > \max(2, p/(p-q))$ *and* $f(\bar{u}) \le (2q/(m-2))^{\frac{1}{m}}$.

(ii) *For any* $c \in (0, c_f^*)$, (2.3) *has no traveling wave* $U(x + cn)$ *with* $U \in \mathcal{C}_b \setminus \{0\}$ *and* $U(-\infty) = 0$, *and for any* $c \ge c_f^*$, (2.3) *has a traveling wave* $U(x + cn)$ *with* $U \in \mathcal{C}_{u_+^*} \setminus \{0, u^*\}$ *and*

$$u_-^* \le \liminf_{\xi\to+\infty} U(\xi) \le \limsup_{\xi\to+\infty} U(\xi) \le u_+^*.$$

Further, $U(+\infty) = u^*$ *in the case where either* $m \in (0, \max(2, p/(p-q))]$, *or* $m > \max(2, p/(p-q))$ *and* $f(\bar{u}) \le (2q/(m-2))^{\frac{1}{m}}$. *If* $m \in (0, p/(p-q)]$, *then* $U(-\infty) = 0$ *and* $U(\xi)$ *is nondecreasing in* ξ *for all* $c \ge c_f^*$. *If either* $p/(p-q) < m \le 2$, *or* $m > \max(2, p/(p-q))$ *and* $f(\bar{u}) \le (2q/(m-2))^{\frac{1}{m}}$, *then* $U(-\infty) = 0$ *for all* $c > c_f^*$.

Acknowledgments. We are grateful to two anonymous referees for their careful reading and helpful suggestions which led to an improvement of our original manuscript. Xiao-Qiang Zhao would like to thank the National Center for Theoretical Science, Tsing Hua University, Taiwan for its kind hospitality during his visit there.

REFERENCES

[1] D. G. Aronson and H. F. Weinberger, *Nonlinear diffusion in population genetics, combustion, and nerve pulse propagation*, in Partial Differential Equations and Related Topics, J. A. Goldstein, ed., Lecture Notes in Mathematics Ser. 446, Springer-Verlag, Berlin, 1975, pp. 5–49.

[2] D. G. Aronson and H. F. Weinberger, *Multidimensional nonlinear diffusion arising in population dynamics*, Adv. Math., 30 (1978), pp. 33–76.

[3] O. Diekmann, *Thresholds and traveling waves for the geographical spread of infection*, J. Math. Biol., 6 (1978), pp. 109–130.

[4] T. Faria, W. Huang, and J. Wu, *Traveling waves for delayed reaction-diffusion equations with global response*, Proc. Roy. Soc. Lond. Ser. A., 462 (2006), pp. 229–261.

[5] J. K. Hale and H. Kocak, *Dynamics and Bifurcations*, Springer-Verlag, New York, 1991.

[6] A. Hastings, K. Cuddington, K. F. Davies, C. J. Dugaw, S. Elmendorf, A. Freestone, S. Harrison, M. Holland, J. Lambrinos, U. Malvadkar, B. A. Melbourne, K. Moore, C. Taylor, and D. Thomson, *The spatial spread of invasions: New developments in theory and evidence*, Ecology Lett., 8 (2005), pp. 91–101.

[7] M. Kot, *Discrete-time traveling waves: Ecological examples*, J. Math. Biol., 30 (1992), pp. 413–436.

[8] J. M. Levine, E. Pachepsky, B. E. Kendall, S. G. Yelenik, and J. H. R. Lambers, *Plant-soil feedbacks and invasive spread*, Ecology Lett., 9 (2006), pp. 1005–1014.

[9] M. Kot, M. A. Lewis, and P. van den Driessche, *Dispersal data and the spread of invading organisms*, Ecology, 77 (1996), pp. 2027–2042.

[10] B. Li, H. F. Weinberger, and M. A. Lewis, *Spreading speeds as slowest wave speeds for cooperative systems*, Math. Biosci., 196 (2005), pp. 82–89.

[11] X. Liang and X.-Q. Zhao, *Asymptotic speeds of spread and traveling waves for monotone semiflows with applications*, Commun. Pure Appl. Math., 60 (2007), pp. 1–40. Erratum: 61 (2008), pp. 137–138.

[12] X. Liang, Y. Yi, and X.-Q. Zhao, *Spreading speeds and traveling waves for periodic evolution systems*, J. Differential Equations, 231 (2006), pp. 57–77.

[13] R. Lui, *Biological growth and spread modeled by systems of recursions,* I. *Mathematical theory*, Math. Biosci., 93 (1989), pp. 269–295.

[14] S. Ma, *Traveling waves for non-local delayed diffusion equations via auxiliary equations*, J. Differential Equations, 237 (2007), pp. 259–277.

[15] M. Neubert and H. Caswell, *Demography and dispersal: Calculation and sensitivity analysis of invasion speed for structured populations*, Ecology, 81 (2000), pp. 1613–1628.

[16] C. Ou and J. Wu, *Persistence of wavefronts in delayed nonlocal reaction-diffusion equations*, J. Differential Equations, 235 (2007), pp. 219–261.

[17] H. R. Thieme, *Density-dependent regulation of spatially distributed populations and their asymptotic speed of spread*, J. Math. Biol., 8 (1979), pp. 173–187.

[18] H. R. Thieme, *On a class of Hammerstein integral equations*, Manuscripta Math., 29 (1979), pp. 49–84.

[19] H. R. Thieme and X.-Q. Zhao, *Asymptotic speeds of spread and traveling waves for integral equations and delayed reaction-diffusion models*, J. Differential Equations, 195 (2003), pp. 430–470.

[20] H. F. Weinberger, *Long-time behavior of a class of biological models*, SIAM J. Math. Anal., 13 (1982), pp. 353–396.

[21] H. F. Weinberger, M. A. Lewis, and B. Li, *Analysis of linear determinacy for spread in cooperative models*, J. Math. Biol., 45 (2002), pp. 183–218.

[22] J. Wu and X. Zou, *Traveling wave fronts of reaction-diffusion systems with delay*, J. Dynam. Differential Equations, 13 (2001), pp. 651–687.

SIAM J. MATH. ANAL.
Vol. 40, No. 2, pp. 790–823

ESTIMATES ON THE HAUSDORFF DIMENSION OF THE RUPTURE SET OF A THIN FILM*

KAI-SENG CHOU† AND SHI-ZHONG DU†

Abstract. Upper bounds on the Hausdorff dimensions of the rupture set of a weak solution of the thin film equation in space-time and in space slices are derived. Finite time rupture is shown to occur for a class of thin films obeying the power law with power in $(0, 1/2)$ under periodic boundary conditions.

Key words. Hausdorff measure, Hausdorff dimension, finite time rupture, thin film equation, partial regularity of a suitable weak solution

AMS subject classifications. 35Q35, 76A20, 35B35, 93D20

DOI. 10.1137/070685348

1. Introduction.

In recent years the thin film type equations

$$h_t + (h^n h_{xxx})_x = 0, \quad n > 0, \quad h \geq 0, \tag{1.1}$$

and

$$h_t + (h^n (h_{xx} + f(h))_x)_x = 0, \quad n > 0, \ h \geq 0, \tag{1.2}$$

have attracted much attention. These equations, where h describes the height of an axisymmetric thin film in motion, arise from the lubrication approximation of the Navier–Stokes equations when n is equal to 3 under no-slip boundary conditions at the lower boundary. It also describes the motion of fluid in a Hele–Shaw cell when n is equal to 1. The f-term in the second equation describes the presence of other physical effects such as gravity, the van der Waals forces, and thermocapillary effect. It is called a power law if

$$f(z) = \begin{cases} B\dfrac{z^q}{q}, & q \neq 0, \\ B \log z, & q = 0, \end{cases} \tag{1.3}$$

for some positive B. One may consult the surveys Myers [M] and Oron, Davis, and Bankoff [ODB] for background on these equations and various models. Here we are concerned with the analytic aspects of these equations. To study them one needs to impose initial and boundary conditions so that they become well-posed at least for a short time. In the literature several boundary conditions such as Neumann-type conditions, pressure boundary conditions, and periodic boundary conditions have been used. Throughout this paper we shall employ the periodic boundary conditions. Thus given a nonnegative, periodic function h_0, we would like to consider the solution to (1.1) starting at h_0, which is of the same period as h_0. Observing that the equation is a parabolic equation of fourth order when h is positive, it follows from parabolic theory that, for any positive, sufficiently regular initial datum, the equation admits a classical solution for small time, and it continues to exist as long as the solution

*Received by the editors March 14, 2007; accepted for publication (in revised form) March 11, 2008; published electronically August 20, 2008. This research was supported by an Earmarked Grant for Research, Hong Kong.
http://www.siam.org/journals/sima/40-2/68534.html
†Department of Mathematics, The Chinese University of Hong Kong, Hong Kong (kschou@math.cuhk.edu.hk, szdu@math.cuhk.edu.hk).

is bounded away from zero. However, unlike second order parabolic equations where positivity is ensured by the maximum principle in an *a prior* way, it is not clear at all whether the film will remain positive for all time, or it will touch down at 0 at some finite time. The film is said to rupture when the latter occurs, and the equation loses its parabolicity. The mathematical foundation of (1.1) was laid down in Bernis-Friedman [BF] where they discovered the following remarkable fact: When $n \geq 4$ equation (1.1) preserves positivity. (Later the requirement $n \geq 4$ was relaxed to $n \geq 3.5$ in Bertozzi et al. [BBDK].) Using this fact, Bernis and Friedman was able to construct a nonnegative, global weak solution of (1.1) for some n less than 4 for nonnegative, H^1-data. After the works of Beretta, Bertsch, and Dal Passo [BBDP], Bertozzi-Pugh [BP1] and [BP2], nowadays it is known that nonnegative, global weak solutions exist for (1.1) and (1.2) for any positive n when f is in $C^1([0,\infty))$ and grows slower less than a cubic power in h at ∞.

A natural question is, For n in $(0, 3.5)$, will rupture occur in finite time for some positive initial data? Observing that, when n is equal to 0, finite time rupture occurs for easily constructed initial data, it would not be surprising that finite time rupture occurs for positive, small n. In fact, under the boundary conditions $h = h_{xx} = 1$ at endpoints, such property is established for (1.1) when n is less than $1/2$ in [BBDP]. However, for the periodic conditions this remains open. Numerical studies show that there is a critical number $n^* \in (1, 2)$ so that finite time rupture occurs for $n < n^*$, and positivity is preserved for $n > n^*$. Likewise, as for (1.2) under the power law $q < 3$ there should be a critical number $n_c(q)$, independent of B and the period, which separates finite time rupture and the preservation of positivity. Numerical results in Goldstein, Pesci, and Shelly [GPS] show that $1 \leq n_c(1) \leq 3$, and those in Laugesen-Pugh [LP2] show that $1.8 < n_c(0.5) < 1.85$ and $1.65 < n_c(2.5) < 1.6625$. Theoretical upper and lower bounds on $n_c(q)$ can be found in Chou-Kwong [CK] for negative q. In this paper we shall show that finite time rupture occurs for (1.2), where f is a power law (1.3) ($q \in (1, 3)$) under the periodic boundary conditions when n is less than $1/2$; see Corollary 3.1 for a precise statement. We shall not consider (1.2) when f grows faster than a cubic power. In addition to the possibility of rupture, it is believed that, in this case the solution may also blow up in finite time. We refer to [BP1] for a conjecture in this direction, and [BP3] and Slepčev-Pugh [SP] for more recent progress.

In this paper we approach the problem of finite time rupture via estimating the possible size of the rupture set. When rupture really occurs, surely our results yield information on the size of the rupture set. On the other hand, even if rupture does not occur, one may still build on these results further criteria for the absence of rupture. A similar situation can be found in the study of the regularity of the weak solution of the three-dimensional Navier–Stokes equations. One may consult Escauriaza, Seregin, and Šverák [ESS] for how the partial regularity estimates in Caffarelli, Kohn, and Nirenberg [CKN] are used in establishing regularity criteria.

Let us review how one excludes finite time rupture for thin films when $n \geq 3.5$. Recall that the weak solution constructed in [BF], [BBDP], [BP1], and [BP2] has many regularity properties including the following two: For any positive initial data, the weak solution of (1.1) or (1.2) ($1 < q < 3$) satisfies, for any $0 < t < T$,

$$\int h_x^2(t) + \int_0^T \int h^n \left[(h_{xx} + f(h))_x\right]^2 \leq C_1, \tag{1.4}$$

and

$$\int h(t)^{3/2-n} \leq C_2 \tag{1.5}$$

for some constants C_1 and C_2 depending on the initial data. It follows from the conservation of area

$$\int h(t) = \int h(0), \quad \text{all } t > 0,$$

and (1.4) that the solution h has a uniformly bounded $C^{1/2}-$norm. At a rupture point (x_0, t) the solution is dominated by a constant multiple of $|x - x_0|^{1/2}$. Putting this estimate into (1.5) yields a contradiction when $n \geq 3.5$, so no finite rupture can occur when $n \geq 3.5$. Even when n is less than 3.5, (1.5) tells us that the rupture set has null measure at every time as long as $n > 1.5$. Our first result gives a more precise estimate.

THEOREM 1.1. *Let h be a suitable weak solution of* (1.1) *or* (1.2)*, where $n \in (1.5, 3.5)$ and f is a power law with $q > 1 - n/2$. Then for any $t > 0$, the $(7 - 2n)/(2n + 1)$-Hausdorff measure of the set $\{x : h(x, t) = 0\}$ is finite, so its Hausdorff dimension cannot exceed $(7 - 2n)/(2n + 1)$.*

Since weak solutions to (1.1) or (1.2) may not be unique, we use the terminology "a suitable weak solution" to refer to a certain weak solution constructed by the methods in [BF], [BBDP], and [BP1] and [BP2]. We shall review its definition in the next section and sketch a construction of these solutions in Appendix A. The proof of our results including Theorem 1.1 and those stated below involves the use of (1.5) in an essential way, so it does not work when n goes below 1.5. We point out that our bound on the Hausdorff dimension $(7 - 2n)/(2n + 1)$ is always less than 1 and is sharp in the sense that it tends to 0 and 1 as n tends to 3.5 and 1.5, respectively.

The proof of Theorem 1.1 follows rather easily from a covering argument coupling with properties (1.4) and (1.5). However, for the proof of the following size estimate in space-time, our main result, we need to make use of the further regularity properties of the suitable weak solution.

THEOREM 1.2. *Let h be a suitable weak solution of* (1.1) *or* (1.2)*, where $n \in (1.5, 3.5)$ and f is a power law with $q > 1 - n/2$. The parabolic Hausdorff dimension of the rupture set $\{(x, t) : h(x, t) = 0\}$ cannot exceed $2n/(2n - 3)$ for $n \in [2, 3.5)$ and $6/n + 1$ for $n \in (1.5, 2)$.*

The parabolic Hausdorff measure, to be discussed in Appendix B, is defined through covering a set by cylinders with different weights in space and time. Here we take the parabolic fourth order weights, so the parabolic Hausdorff dimension is equal to 5 for every open set in space-time. Again we observe that our estimate on the dimension of the rupture set lies between 1.75 and 5 and tends to 1.75 and 5 as n tends to 3.5 and 1.5, respectively.

Finally, we give an estimate on the rupture times.

THEOREM 1.3. *Let h be a suitable weak solution of* (1.1) *or* (1.2)*, where $n \in (2, 3.5)$ and f is a power law with $q > 1 - n/2$. The Hausdorff dimension of the set $\{t > 0 : h(x, t) = 0$ for some $x\}$ cannot exceed $1 - 2(n - 2)^2/(8 - n)$.*

Again this estimate becomes sharp as n tends to 3.5.

Theorems 1.1–1.3 are contained in the results in sections 2 and 3, where we also consider more general f. In section 4 we establish the result on finite time rupture mentioned above.

The problem of rupture for thin films is closely related to the formation of singularities for solutions of (1.1) or (1.2). Indeed, let us assume that the solution remains to be C^2 at the rupture point (x_0, t). Then it is bounded by a constant multiple of $|x - x_0|^2$ near x_0. Putting this into (1.5) yields a contradiction unless $n \leq 2$. It demonstrates that the second derivatives of the solution blow up at any rupture point

when $n > 2$, so the rupture set coincides with the singular set of any weak solution of (1.1) or (1.2) in which (1.4) and (1.5) hold. Our proofs of Theorems 1.1, 1.2, and 1.3 make use of some ideas in known works in the partial regularity of solutions for evolution equations. There are quite a number of works on this topic in recent years. We wish to mention the works of Caffarelli–Kohn–Nirenberg [CKN] (later simplified substantially by Lin [L]) on the three-dimensional Navier–Stokes equations, Struwe [S], and Chen–Struwe [CS] on the harmonic heat flows, Lin–Liu [LL] on the motion of liquid crystals, and Chou–Du–Zheng [CDZ] on the semilinear heat equation with supercritical growth.

In Jiang-Lin [JL] estimates for the Hausdorff dimension of the rupture set of a steady state of a multidimensional thin film type equation are present.

Although we are mainly concerned with thin films whose initial data are not too degenerate in the sense that they have finite entropy (1.5), there are significant results for weak solutions starting from nonnegative initial data that could vanish in an open set. Our results do not apply to these solutions. By restricting to the literature on the one-dimensional case (axisymmetric films) only, we point out Bernis [B1], [B2], and Hulshof-Shishkov [HS], where results on the finite speed of the propagation of zero are established for (1.1). That such property continues to hold for (1.2) is explained in [BP2]. Moreover, a waiting time phenomenon which asserts that the support of a thin film does not increase in a small period of time is established for (1.1) ($n \in (0,3)$) in Dal Passo-Giacomelli-Grun [DPGG]. For $n > 4$, it is known in [BBDP] that the support remains the same all the time.

2. A size estimate on the rupture set in space-time. First we give the definition of a suitable weak solution of the thin film type equation under the periodic boundary conditions. Consider the problem

$$\begin{cases} h_t + \big[h^n\big(h_{xx} + f(h)\big)_x\big]_x = 0, \ (x,t) \in (-L/2, L/2) \times (0,\infty), \\ h \text{ is of period } L \text{ for each } t > 0, \end{cases} \tag{2.1}$$

where n is positive and f is a measurable function in $[0,\infty)$ which belongs to $C^{2,\alpha}\big(0,\infty\big)$, $\alpha \in (0,1)$. Let S_L denote the circle obtained by identifying the endpoints of $(-L/2, L/2)$. A nonnegative function h in

$$C_{x,t}^{\frac{1}{2},\frac{1}{8}}\big(S_L \times (0,\infty)\big) \cap L^\infty\big(0,\infty; H^1(S_L)\big) \cap L^2_{loc}\big(0,\infty; H^2(S_L)\big)$$

is called a suitable weak solution of (2.1) if it fulfills the following requirements. For every $T > 0$,

(H_1) $h^{\frac{n}{2}}\big(h_{xx} + f(h)\big)_x \in L^2(Q_T), Q_T = S_L \times (0,T)$, and

$$\iint_{Q_T} h\Phi_t = -\iint_{Q_T} h^n\big(h_{xx} + f(h)\big)_x \Phi_x$$

for all $\Phi \in H^1(Q_T)$ which are compactly supported in Q_T.

(H_2) The following quantities

$$K_1 = \sup_{[0,T]} \int h_x^2,$$

$$K_2 = \iint_{Q_T} \Big(h^n\big|(h_{xx} + f(h))_x\big|^2 + h^n h_{xxx}^2\Big),$$

$$K_3 = \sup_{[0,T]} \int h^{\frac{3}{2}-n},$$

and

$$K_4 = \iint_{Q_T} \left(h^{-\frac{1}{2}+\delta} h_{xx}^2 + h^{-\frac{5}{2}+\delta} h_x^4 \right) \quad \forall \delta \in \left(0, \min\{1.5, n-1.5\}\right)$$

are finite, where K_1–K_4 depend on T, and K_4 also depends on δ.

(H_3) There exists $\Gamma \subseteq [0,T]$ of full measure such that, for every $t_0 \in \Gamma$, the following local energy inequality and local entropy inequality hold: For $\forall t > t_0, \forall \phi \in C^2(S_L)$,

$$\begin{aligned} \frac{1}{2}\int h_x^2(t)\phi^4 - \int F(h(t))\phi^4 + \int_{t_0}^t \int h^n \left| \left(h_{xx} + f(h)\right)_x \right|^2 \phi^4 \\ \leq -\int_{t_0}^t \int h^n \left(h_{xx} + f(h)\right)_x \left(h_{xx} + f(h)\right)(\phi^4)_x \\ - \int_{t_0}^t \int h^n \left(h_{xx} + f(h)\right)_x \left[h_{xx}(\phi^4)_x + h_x(\phi^4)_{xx} \right] \\ + \frac{1}{2}\int h_x^2(t_0)\phi^4 - \int F(h(t_0))\phi^4, \end{aligned} \tag{2.2}$$

and

$$\begin{aligned} \int h^{\frac{3}{2}-n+\delta}(t)\phi^4 + \sigma \int_{t_0}^t \int \left(h^{-\frac{1}{2}+\delta} h_{xx}^2 \phi^4 + h^{-\frac{5}{2}+\delta} h_x^4 \phi^4 \right) \\ \leq C \int_{t_0}^t \int \left[h^{\frac{3}{2}+\delta}\left(\phi_x^4 + \phi^2 \phi_{xx}^2\right) + f^2(h) h^{-\frac{1}{2}+\delta}\phi^4 \right] \\ + C\int h(t_0)\phi^4 + \int h^{\frac{3}{2}-n+\delta}(t_0)\phi^4, \ \delta \in \left(0, \min\{1.5, n-1.5\}\right), \end{aligned} \tag{2.3}$$

where F is a primitive of f and σ and C are positive constants depending on δ.

From the expression for K_4 we see that it is implicitly assumed $n > 1.5$ in the definition of a suitable weak solution. On the other hand, when $n \geq 3.5$, K_1 and K_3 together imply that a suitable weak solution is always positive; hence by parabolic regularity, it is a classical solution. Consequently we will focus on the range $(1.5, 3.5)$. To state our existence result for these solutions we impose the following assumptions on f: For some constants $C_1 - C_3$,

$$|f(z)| \leq C_1(1 + z^p), \quad z \geq 1 \quad \text{for some } 0 \leq p < 3, \tag{2.4}$$

$$|f'(z)| \leq C_2, \quad z \in (0,1], \tag{2.5}$$

or,

$$|f(z)| + z|f'(z)| \leq C_3 z^q, \quad z \in (0,1] \ \text{ for some } q > 1 - \frac{n}{2}. \tag{2.6}$$

Equation (2.4) restricts the growth of f at infinity. When $f(z)$ grows faster than z^3, it is conjectured that some solutions of (2.1) blow up in finite time. So growth restriction like (2.4) is needed to ensure the existence of a global, weak solution. According to the behavior of f at 0 we impose (2.5) and (2.6), respectively. For power laws the former applies when $q \geq 1$, and the latter applies when $q \in (1 - n/2, 1)$.

THEOREM 2.1. *Consider problem* (2.1), *where* $n \in (1.5, 3.5)$, (2.4) *and* (2.5) *or* (2.6) *hold. Then for any nonnegative* $h_0 \in H^1(S_L)$ *with finite* $||h_0^{3/2-n}||_{L^1}$, (2.1) *admits a suitable weak solution* h *satisfying* $||h(t) - h_0||_{L^2}$ *tends to* 0 *as* $t \downarrow 0$.

An outline of the proof of this theorem can be found in Appendix A. We remark that this theorem is still valid for $p = 3$ in (2.4) provided that $||h_0||_{L^1}$ is sufficiently small.

Weak solutions for (1.1) have been constructed in [BF], [BBDP], and [BP1] and for (1.2) in [BP2]. In particular, in Theorem 3.3 of [BP2] the existence of a weak solution essentially satisfying (H_1) and (H_2) is proved. A slight difference is that our Theorem 2.1 applies to the condition $q > 1 - n/2$, which could be negative when $n > 2$, while q is considered to be positive in the previous works. In (H_3) we also need the local energy and entropy inequalities. The inequalities of this type were derived in [B1], [B2], and [BP2] to study the finite speed of the propagation of the weak solution. Unlike previous works where the initial time is always taken to be zero, here we need these inequalities to be valid starting at initial times which form a subset of the time axis of full measure.

We begin with an estimate on the size of the rupture set in a spatial slice.

THEOREM 2.2. *Let* h *be a nonnegative function in* $L^\infty(0, T; H^1(S_L))$ *satisfying, for some* $n \in (1.5, 3.5)$,

$$\sup_{0<t<T} \int h_x^2 < \infty,$$

and

$$\sup_{0<t<T} \int h^{3/2-n} < \infty.$$

Then for each $T > 0$, *the* $((7-2n)/(2n+1))-$*Hausdorff measure of the set* $\mathcal{R}_T = \{x : h(x, T) = 0\}$ *is finite.*

Proof. Let h be a function satisfying the conditions in the theorem. We claim that

$$\int_{B_R} h_x^2(T) \le R^{\gamma_0}, \quad \gamma_0 = \frac{7-2n}{2n+1}$$

and

$$\int_{B_R} h^{\frac{3}{2}-n}(T) \le \delta_0 R^{\gamma_0}, \quad \delta_0 = \frac{2}{(1+\gamma_0)(\frac{3}{2}-n)+2}$$

hold for all small B_R centered at x_0, then $h(x_0, T) > 0$. Assuming on the contrary that (x_0, T) is a rupture point, we would have

$$\begin{aligned} h(x, T) &= h(x, T) - h(x_0, T) \\ &\le \left| \int_{x_0}^x 1 \right|^{1/2} \left(\int_{x_0}^x h_x^2 \right)^{1/2} \\ &\le |x - x_0|^{\frac{1}{2}(1+\gamma_0)} \end{aligned}$$

for all x sufficiently close to x_0, but then

$$\begin{aligned} \delta_0 R^{\gamma_0} &\ge \int_{B_R} h^{\frac{3}{2}-n}(T) \\ &\ge \frac{2R^{\frac{1}{2}(1+\gamma_0)(\frac{3}{2}-n)+1}}{\frac{1}{2}(1+\gamma_0)(\frac{3}{2}-n)+1}. \end{aligned}$$

As $\gamma_0 = \frac{1}{2}(1+\gamma_0)(\frac{3}{2}-n)+1$, contradiction holds.

Henceforth, at a rupture point (x_0, T) there exists $\{R_j\}$ (depending on x_0) $\downarrow 0$ such that

$$\int_{B_{R_j}} \big(h_x^2 + h^{\frac{3}{2}-n}\big)(T) > \delta_0 R_j^{\gamma_0}.$$

For each $\delta > 0$, the collection $\mathcal{C}_\delta$ of all of these balls B_{R_j}, with radii less than δ forms a covering of the rupture set $\mathcal{R}_T$. By a version of the Vitali covering theorem (see Appendix B) we can select from $\mathcal{C}_\delta$ a mutually disjoint, countable subcollection $\mathcal{C}'_\delta$ such that

$$\mathcal{R}_T \subseteq \bigcup \big\{ B_{5R_j}(x_j) : B_{R_j}(x_j) \in \mathcal{C}'_\delta \big\}.$$

Therefore, according to the definition of the Hausdorff measure Federer [Fe]

$$\begin{aligned}
\mathcal{H}_\delta^{\gamma_0}(\mathcal{R}_T) &\le \alpha(\gamma_0) \sum (5R_j)^{\gamma_0} \\
&\le \frac{5^{\gamma_0}\alpha(\gamma_0)}{\delta_0} \sum \int_{B_{R_j}(x_j)} \big(h_x^2 + h^{\frac{3}{2}-n}\big)(T) \\
&\le \frac{5^{\gamma_0}\alpha(\gamma_0)}{\delta_0} \int \big(h_x^2 + h^{\frac{3}{2}-n}\big)(T), \quad \alpha(s) = \frac{\pi^s/2}{\Gamma(s/2+1)},
\end{aligned}$$

which leads to the desired conclusion after taking $\delta \to 0$. □

Clearly Theorem 1.1 is a special case of Theorem 2.2. Here we have used little beyond the finiteness of K_1 and K_3. Although the proof of Theorem 2.2 is quite simple, as one will see, the same idea works in the size estimate of the rupture set in space-time. In order to obtain such estimates, we need to make use of the finiteness of the integrals in space-time such as the quantities K_2 and K_4. We have the following.

THEOREM 2.3. *Consider any suitable weak solution of* (2.1)*, where* $n \in (2, 3.5)$ *and either* (2.5) *or* (2.6) *hold. The parabolic Hausdorff dimension of the rupture set of this solution cannot exceed* $2n/(2n-3)$.

The proof of this theorem depends on Lemmas 2.1–2.4 below. For the ease of presentation we shall prove the theorem first.

Proof. Let (x_0, T) be a rupture point of the solution h. We claim that, for every $\gamma_0 \in \big(\frac{7+2\delta}{4}, 4+2\delta\big)$, there exists a sequence $\{R_j\}$, $R_j \downarrow 0$, such that

$$\iint_{Q_{R_j}} h^n \Big| \big(h_{xx} + f(h)\big)_x \Big|^2 > R_j^{\gamma_0},$$

or

$$\iint_{Q_{R_j}} h^{-\frac{5}{2}+\delta} h_x^4 > R_j^{\gamma_0}$$

holds. For, were this not true, it means that (2.7) and (2.14) hold for all sufficiently small R. By Lemma 2.4, there exists some $t_1 \in \Gamma \cap [T - R^4, T]$ such that

$$\int_{B_{R/2}} h_x^2(t_1) \le C R^\gamma \quad \forall \gamma \in \Big(0, \frac{4\gamma_0 - 7 - 2\delta}{3+2\delta}\Big).$$

Then by Lemma 2.3,

$$\int_{B_{R/4}} h_x^2 \le C R^\gamma \quad \forall t \in [t_1, T].$$

Lemma 2.2 further gives

$$\sup_{B_{R/4}\times[t_1,T]} h \le CR^{\frac{1+\gamma}{2}},$$

which means that

$$h(x,T) \le C|x-x_0|^{\frac{1+\gamma}{2}}$$

for all x sufficiently close to x_0. However, on the other hand,

$$\begin{aligned}\infty &> \int h^{\frac{3}{2}-n}(T)\\ &\ge C\int |x-x_0|^{\frac{1+\gamma}{2}(\frac{3}{2}-n)} = \infty,\end{aligned}$$

if $(\frac{1+\gamma}{2})(\frac{3}{2}-n) \le -1$, i.e., $\gamma \ge (7-2n)/(2n-3)$. A contradiction will be drawn if we can verify that

$$\frac{7-2n}{2n-3} < \frac{4\gamma_0 - 7 - 2\delta}{3+2\delta}, \quad \text{i.e., } \gamma_0 > \frac{2n+2\delta}{2n-3}.$$

A direct computation shows that this is true if $n > 2$.

Thus, at a rupture point (x_0,T), there exists a sequence of cylinders $\{Q_{R_j}\}$, $R_j \downarrow 0$, satisfying

$$\iint_{Q_{R_j}} \left[h^n \left|\big(h_{xx}+f(h)\big)_x\right|^2 + h^{-\frac{5}{2}+\delta}h_x^4\right] > R_j^{\gamma_0}.$$

The collection of all of these cylinders forms a covering of the rupture set in space-time, and we can replace the arguments in the second half of the proof of Theorem 2.2 involving the Hausdorff measure by the parabolic Hausdorff measure (see Appendix B) to show that the γ_0-parabolic Hausdorff measure of the rupture set is finite for all $\gamma_0 > (2n+2\delta)/(2n-3)$. By letting $\delta \to 0$ we conclude that its parabolic Hausdorff dimension cannot exceed $2n/(2n-3)$. The proof of Theorem 2.3 is completed. □

LEMMA 2.1. *Let χ be a strictly increasing, continuous function in $[0,\infty)$, with $\chi(0) > 0$. Define a sequence $\{\alpha_n\}$ by $\alpha_{n+1} = \chi(\alpha_n)$, $\alpha_1 = 0$. Then $\{\alpha_n\}$ is increasing and*

(a) *it diverges to ∞ if $y = \chi(x)$ has no intersection with $y = x$, or*
(b) *it converges to some α^* which is the $x-$coordinate of the first intersection between $y = \chi(x)$ and $y = x$.*

Consequently, for any α (less than α^* if α^* exists), there exists some N such that $\alpha_N \ge \alpha$.

The proof of this lemma is elementary and is omitted.

LEMMA 2.2. *Let h be a suitable weak solution of (2.1), and (x_0,T) a rupture point of h. Suppose that*

$$\iint_{Q_R} h^n \left|(h_{xx}+f(h))_x\right|^2 \le R^{\gamma_0}, \quad Q_R = (x_0-R,x_0+R)\times(T-R^4,T), \tag{2.7}$$

and

$$\int_{B_R} h_x^2 \le C_4 R^{\gamma} \quad \forall t \in [t_1,T], \quad B_R = (x_0-R,x_0+R), \tag{2.8}$$

hold for some R, γ_0, γ, t_1, *and* C_4, *where* $t_1 \in [T - R^4, T]$, $R^4 < T$, *and* γ *satisfies*

$$\gamma < \frac{2\gamma_0 + n}{2 - n} \quad \text{for } n \in (1.5, 2), \text{ and } < \infty \quad \text{for } n \in [2, 3.5). \tag{2.9}$$

Then for every $\tau_0 \in (0,1)$, *there exists a constant* C *depending on* n, γ_0, γ, C_4, *and* τ_0 *such that*

$$\sup_{B_{\tau_0 R} \times [t_1, T]} h \le C R^{\frac{1+\gamma}{2}}.$$

Proof. By (H_1) in the definition of a suitable weak solution,

$$\iint h \Phi_t = -\iint h^n \big(h_{xx} + f(h)\big)_x \Phi_x \quad \forall \Phi \text{ test function.}$$

We choose $\Phi = \xi(x)\theta_\delta(t)$, where ξ is compactly supported in B_R, and $\theta_\delta(t) = \int_0^t \eta(t)dt$, with η given by

$$\eta = \begin{cases} -\dfrac{1}{2\delta}, & t \in [s - \delta, s + \delta], \\ \dfrac{1}{2\delta}, & t \in [T - \delta, T + \delta], \\ 0, & \text{otherwise}, \end{cases}$$

and s is a fixed number in $[t_1, T]$. Letting $\delta \downarrow 0$, as $\theta_\delta \le 1$, we obtain

$$\left| \int_{B_R} \big(h(x,T) - h(x,t)\big)\xi(x) \right| \le \int_t^T \int_{B_R} h^n \left| \big(h_{xx} + f(h)\big)_x \xi_x \right|, \tag{2.10}$$

where we have changed notation from s to t. Letting $\tau \in (\tau_0, 1)$ to be chosen later, we pick ξ so that $\xi = 1$ in $B_{\tau R}$ and $|\xi_x| \le 1/R(1-\tau)$ in B_R. Using (2.7),

$$\begin{aligned} \int_t^T \int_{B_R} h^n \left| \big(h_{xx} + f(h)\big)_x \xi_x \right| &\le \sqrt{\int_t^T \int_{B_R} h^n \big(h_{xx} + f(h)\big)_x^2} \ \sqrt{\int_t^T \int_{B_R} h^n \xi_x^2} \\ &\le \frac{1}{(1-\tau)} R^{\frac{\gamma_0 + 3}{2}} \sup_{Q_R} h^{\frac{n}{2}}. \end{aligned}$$

It follows from (2.10) that

$$\left| \fint_{B_R} \big(h(x,T) - h(x,t)\big)\xi(x) \right| \le \frac{1}{2\tau(1-\tau)} R^{\frac{\gamma_0 + 1}{2}} \sup_{Q_R} h^{\frac{n}{2}},$$

where we have set $\fint_{B_R} F \equiv \frac{\int_{B_R} F}{\int_{B_R} \xi}$. By (2.8) and (2.10),

$$\begin{aligned} h(x,t) &\le \left| \fint_{B_R} \big(h(x,t) - h(y,t)\big)\xi(y)dy \right| + \left| \fint_{B_R} \big(h(y,t) - h(y,T)\big)\xi(y)dy \right| \\ &\quad + \left| \fint_{B_R} \big(h(y,T) - h(x_0,T)\big)\xi(y)dy \right| \\ &\le 2\sqrt{C_4} R^{\frac{1+\gamma}{2}} + \frac{1}{2\tau(1-\tau)} R^{\frac{1+\gamma_0}{2}} \sup_{Q_R} h^{n/2} \quad \forall x \in B_{\tau R}. \end{aligned} \tag{2.11}$$

Hence in the case

$$\sup_{Q_R} h \le CR^{\alpha} \quad \text{for some } \alpha < \frac{1+\gamma}{2}, \tag{2.12}$$

(2.11) implies that

$$\begin{aligned} \sup_{Q'_R} h &\le C\Big(R^{\frac{1+\gamma}{2}} + R^{\frac{1+\gamma_0+n\alpha}{2}}\Big) \\ &\le CR^{\chi(\alpha)}, \quad Q'_R = B_{\tau R} \times [t_1, T], \end{aligned} \tag{2.13}$$

provided that $\chi(\alpha) \equiv (1+\gamma_0+n\alpha)/2 < (1+\gamma)/2$. Starting at $\alpha = 0$, we bootstrap between (2.12) and (2.13) to obtain

$$\sup_{B_{\tau^k R} \times [t_1, T]} h \le C_k R^{\chi(\alpha_k)}$$

after k times of iterations as long as $\chi(\alpha_k) < (1+\gamma)/2$. When $n \ge 2$, $y = \chi(x)$ does not intersect $y = x$, so there exists an N such that $\chi(\alpha_N) \ge (1+\gamma)/2$. Then the lemma follows by applying the Holder continuity of h in space. When $n \in (1.5, 2)$, the same conclusion holds if $(\gamma+1)/2 < \alpha_*$, the intersection of $y = \chi(x)$, and $y = x$. As α_* is given by $\frac{(1+\gamma_0)/2}{1-n/2} = \frac{1+\gamma_0}{2-n}$, $(\gamma+1)/2 < \alpha_*$ is equivalent to (2.9). ∎

LEMMA 2.3. *Let h be a suitable weak solution of* (2.1)*, where $n \in (1.5, 3.5)$ and* (2.5) *holds, and let (x_0, T) be a rupture point of h. Suppose that* (2.7)*,*

$$\iint_{Q_R} h^{-\frac{5}{2}+\delta} h_x^4 \le R^{\gamma_0} \tag{2.14}$$

and

$$\int_{B_R} h_x^2(t_1) \le C_5 R^{\gamma} \tag{2.15}$$

hold for some $R, \gamma_0, \gamma, t_1 \in \Gamma \cap [T - R^4, T]$, C_5, where γ and γ_0 also satisfy (2.9)*,*

$$\gamma_0 > \frac{7 - 2n + 2\delta}{4}, \quad \delta \in (0, n - 1.5), \tag{2.16}$$

and

$$0 \le \gamma < 3. \tag{2.17}$$

Then for every $\tau_0 \in (0,1)$, there exists a constant C depending on $n, \gamma_0, \gamma, \delta, C_5$, and τ_0 such that

$$\int_{B_{\tau_0 R}} h_x^2 \le CR^{\gamma} \quad \forall t \in [t_1, T].$$

When (2.6) *instead of* (2.5) *holds in the above assumptions, the same conclusion holds when* (2.17) *is replaced by*

$$\gamma < \frac{8-n}{n}. \tag{2.17}'$$

Proof. Let ϕ be a cut-off function supported in B_R, $\phi = 1$ in $B_{\tau R}$, where $\tau \in (\tau_0, 1)$ is to be chosen later. We plug ϕ into the local energy inequality (2.2) and use the Cauchy–Schwarz inequality to get

$$\begin{aligned}&\left|\iint h^n(h_{xx}+f(h))_x(h_{xx}+f(h))(4\phi^3\phi_x)\right|\\&\le \frac{1}{4}\iint h^n\left|(h_{xx}+f(h))_x\right|^2\phi^4 + C\iint h^n(h_{xx}+f(h))^2\phi^2\phi_x^2\end{aligned}$$

and

$$\begin{aligned}&\left|\iint h^n(h_{xx}+f(h))_x\left[4h_{xx}\phi^3\phi_x + h_x(4\phi^3\phi_{xx}+12\phi^2\phi_x^2)\right]\right|\\&\le \frac{1}{4}\iint h^n\left|(h_{xx}+f(h))_x\right|^2\phi^4 + C\iint\left[h^n(h_{xx}^2\phi^2\phi_x^2 + h_x^2(\phi_x^4+\phi^2\phi_{xx}^2)\right].\end{aligned}$$

It follows that

$$\begin{aligned}\frac{1}{2}\int_{B_R} h_x^2\phi^4 &\le \int_{B_R} F(h)\phi^4 + \frac{1}{2}\int_{B_R} h_x^2(t_1)\phi^4 - \int_{B_R} F(h(t_1))\phi^4\\&\quad - \frac{1}{2}\int_{t_1}^t\int_{B_R} h^n\left|\big(h_{xx}+f(h)\big)_x\right|^2\phi^4 + C\left[\int_{t_1}^t\int_{B_R} h^n\big(h_{xx}+f(h)\big)^2\phi^2\phi_x^2\right.\\&\quad\left. + \int_{t_1}^t\int_{B_R} h^n\big(h_{xx}^2\phi^2\phi_x^2 + h_x^2(\phi_x^4+\phi^2\phi_{xx}^2)\big)\right]\\&\le \int_{B_R} F(h)\phi^4 + \frac{1}{2}\int_{B_R} h_x^2(t_1)\phi^4\\&\quad - \int_{B_R} F(h(t_1))\phi^4 - \frac{1}{4}\int_{t_1}^t\int_{B_R} h^n\left|\big(h_{xx}+f(h)\big)_x\right|^2\phi^4\\&\quad + C\left[\int_{t_1}^t\int_{B_R} h^n f'^2(h)h_x^2\phi^4 + \int_{t_1}^t\int_{B_R} h^n f^2(h)\phi^2\phi_x^2\right.\\&\quad\left. + \int_{t_1}^t\int_{B_R} h^n h_x^2(\phi_x^4+\phi^2\phi_{xx}^2) + \int_{t_1}^t\int_{B_R} h^n h_{xx}^2\phi^2\phi_x^2\right]\\&\le \int_{B_R} F(h)\phi^4 + \frac{1}{2}\int_{B_R} h_x^2(t_1)\phi^4\\&\quad - \int_{B_R} F(h(t_1))\phi^4 - \frac{1}{4}\int_{t_1}^t\int_{B_R} h^n\left|\big(h_{xx}+f(h)\big)_x\right|^2\phi^4\\&\quad + C\left[\int_{t_1}^t\int_{B_R} h^n f'^2(h)h_x^2\phi^4 + \int_{t_1}^t\int_{B_R} h^n f^2(h)\phi^2\phi_x^2\right.\\&\quad\left. + \int_{t_1}^t\int_{B_R} h^n h_x^2(\phi_x^4+\phi^2\phi_{xx}^2) + \int_{t_1}^t\int_{B_R} h^{n-2}h_x^4\phi^2\phi_x^2\right].\end{aligned}\tag{2.18}$$

Notice that in the last inequality we have used

$$\begin{aligned}\int_{t_1}^{t}\int_{B_R} h^n h_{xx}^2\phi^2\phi_x^2 &= -\int_{t_1}^{t}\int_{B_R} h^n h_x h_{xxx}\phi^2\phi_x^2 - n\int_{t_1}^{t}\int_{B_R} h^{n-1}h_x^2 h_{xx}\phi^2\phi_x^2\\ &\quad -\int_{t_1}^{t}\int_{B_R} h^n h_x h_{xx}(2\phi\phi_x^3 + 2\phi^2\phi_x\phi_{xx})\\ &\le \varepsilon\int_{t_1}^{t}\int_{B_R} h^n h_{xxx}^2\phi^4 + \varepsilon\int_{t_1}^{t}\int_{B_R} h^n h_{xx}^2\phi^2\phi_x^2\\ &\quad + C_\varepsilon\left[\int_{t_1}^{t}\int_{B_R} h^n h_x^2(\phi_x^4 + \phi^2\phi_{xx}^2) + \int_{t_1}^{t}\int_{B_R} h^{n-2}h_x^4\phi^2\phi_x^2\right].\end{aligned}$$

Therefore, under (2.5), when $f \in C^1[0,\infty)$, we have

$$\begin{aligned}\frac{1}{2}\int_{B_R} h_x^2\phi^4 &\le \int_{B_R} F(h)\phi^4 + \frac{1}{2}\int_{B_R} h_x^2(t_1)\phi^4 - \int_{B_R} F(h(t_1))\phi^4\\ &\quad + C\left[\int_{t_1}^{t}\int_{B_R} h^n h_x^2(\phi^4 + \phi_x^4 + \phi^2\phi_{xx}^2) + \int_{t_1}^{t}\int_{B_R} h^{n-2}h_x^4\phi^2\phi_x^2\right.\\ &\quad \left. + \int_{t_1}^{t}\int_{B_R} h^n\phi^2\phi_x^2\right], \qquad (2.19)\end{aligned}$$

since a suitable weak solution is bounded. Here the constant C depends on f. Suppose that

$$\int_{B_R} h_x^2 \le CR^\alpha \quad \forall t\in[t_1,T] \quad \text{for some } \alpha<\gamma. \qquad (2.20)$$

We claim that

$$\int_{B_{\tau R}} h_x^2 \le C'R^{\chi(\alpha)} \quad \forall t\in[t_1,T], \qquad (2.21)$$

where

$$\chi(\alpha) = \min\left\{\frac{3}{2}+\frac{\alpha}{2}, \frac{n}{2}+\left(1+\frac{n}{2}\right)\alpha, \gamma_0 + \frac{n-\frac{7}{2}-\delta}{2} + \frac{n+\frac{1}{2}-\delta}{2}\alpha\right\}$$

as long as $\alpha<\gamma$. For, by Lemma 2.2 and (2.20), we have

$$\sup_{B_{\tau R}\times[t_1,T]} h \le CR^{\frac{1+\alpha}{2}}.$$

Using this estimate we control the integrals on the right of (2.19) as follows.

$$\begin{aligned}&\int_{B_R} F(h(t))\phi^4 \le CR^{\frac{3}{2}+\frac{\alpha}{2}},\\ &\int_{B_R} h_x^2(t_1)\phi^4 \le C_5R^\gamma, \qquad \text{(from (2.15))}\\ &\int_{t_1}^{t}\int_{B_R} h^n h_x^2\left(\phi^4+\phi_x^4+\phi^2\phi_{xx}^2\right) \le CR^{\alpha+\frac{n}{2}(1+\alpha)},\\ &\int_{t_1}^{t}\int_{B_R} h^{n-2}h_x^4\phi^2\phi_x^2 \le CR^{-2}\sup_{Q_R} h^{n+\frac{1}{2}-\delta}\int_{t_1}^{t}\int_{B_R} h^{-\frac{5}{2}+\delta}h_x^4\\ &\qquad\qquad \le CR^{\gamma_0+\frac{n-\frac{7}{2}-\delta}{2}+\frac{n+\frac{1}{2}-\delta}{2}\alpha},\\ &\int_{t_1}^{t}\int_{B_R} h^n\phi^2\phi_x^2 \le CR^{3+\frac{n}{2}(1+\alpha)} \le CR^{\alpha+\frac{n}{2}(1+\alpha)},\end{aligned}$$

so (2.21) follows after putting these estimates back to (2.19). As (2.20) holds for $\alpha = 0$, we can bootstrap between (2.20) and (2.21) to increases the power in R in (1.20). Observe that the function χ is piecewise linear with slopes equal to $1/2, 1+n/2$, or $(2n+1-2\delta)/4$. As $(2n+1-2\delta)/4 > 1$ for $\delta < n - 3/2$, χ can possibly intersect $y = x$ at $\alpha^* = 3$. The lemma now follows from Lemma 2.1 after taking $\tau^N = \tau_0$.

In case f satisfies (2.6), the local energy inequality reads as

$$\begin{aligned}\frac{1}{2}\int_{B_R} h_x^2\phi^4 \le{}& \frac{B}{q(q+1)}\int_{B_R} h^{q+1}\phi^4 + \frac{1}{2}\int_{B_R} h_x^2(t_1)\phi^4 - \frac{B}{q(q+1)}\int_{B_R} h^{q+1}(t_1)\phi^4\\ &+ C\bigg(\int_{t_1}^t\int_{B_R} h^n h_x^2\big(\phi_x^4 + \phi^2\phi_{xx}^2\big) + \int_{t_1}^t\int_{B_R} h^{n-2}h_x^4\phi^2\phi_x^2\\ &+ \int_{t_1}^t\int_{B_R} h^{n+2q}\phi^2\phi_x^2 + \int_{t_1}^t\int_{B_R} h^{n+2q-2}h_x^2\phi^4\bigg).\end{aligned}$$

Under (2.20), the integral terms involving q can be estimated as follows:

$$\begin{aligned}\int_{B_R} h^{q+1}\phi^4 &\le CR\sup_{B_R} h^{q+1}\\ &\le CR^{\frac{q+3}{2}+\frac{4-n}{4}\alpha}\\ &\le CR^{\frac{8-n}{4}+\frac{4-n}{4}\alpha}, \qquad (\because q > 1 - n/2)\\ \int_{t_1}^t\int_{B_R} h^{n+2q}\phi^2\phi_x^2 &\le CR^3\sup_{Q_R} h^{n+2q}\\ &\le CR^{4+\alpha}\end{aligned}$$

and

$$\begin{aligned}\int_{t_1}^t\int_{B_R} h^{n+2q-2}h_x^2\phi^4 &\le CR^4\bigg(\sup_{[t_1,t]}\int_{B_R} h_x^2\bigg)\sup_{Q_R} h^{n+2q-2}\\ &\le CR^{4+\alpha}, \qquad (\because q > 1 - n/2).\end{aligned}$$

So (2.21) holds after χ is replaced by the function χ':

$$\begin{aligned}\chi'(\alpha) = \min\bigg\{&\frac{8-n}{4} + \frac{4-n}{4}\alpha, 4+\alpha, \frac{n}{2} + \Big(1+\frac{n}{2}\Big)\alpha,\\ &\gamma_0 + \frac{n - \frac{7}{2} - \delta}{2} + \frac{n + \frac{1}{2} - \delta}{2}\alpha\bigg\}.\end{aligned}$$

Arguing as before, we can show that Lemma 2.3 holds when $q > 1 - n/2$ and (2.17) is replaced by (2.17)′. □

LEMMA 2.4. *Let h be a suitable weak solution of* (2.1)*, where $n \in (1.5, 3.5)$ and* (2.5) *holds, and let (x_0, T) be a rupture point of h. Suppose* (2.7)*,* (2.14)*,* (2.16)*, and* (2.17) *hold. Then there exists $t_1 \in \Gamma \cap [T - R^4, T]$ such that*

$$\int_{B_{R/2}} h_x^2(t_1) \le CR^\gamma \quad \forall\gamma \in \bigg(0, \frac{4\gamma_0 - 7 - 2\delta}{3 + 2\delta}\bigg),$$

if

$$\frac{7 + 2\delta}{4} < \gamma_0 \le 4 + 2\delta. \tag{2.22}$$

When (2.5) *is replaced by* (2.6) *and* (2.17) *by* (2.17)′*, respectively, the same conclusion holds when* (2.22) *is replaced by*

$$\frac{7+2\delta}{4} < \gamma_0 < 1 + \frac{6+4\delta}{n}. \tag{2.22}'$$

Proof. We only prove this lemma when f satisfies (2.5), for the other case can be proved in a similar way. From (2.14) there exists $t_1 \in \Gamma$ such that

$$\int_{B_R} h^{-\frac{5}{2}+\delta} h_x^4(t_1) \le R^{\gamma_0 - 4}.$$

Thus

$$\begin{aligned} \int_{B_R} h_x^2(t_1) &\le \sqrt{\int_{B_R} h^{-\frac{5}{2}+\delta}(t_1) h_x^4(t_1)} \ \sqrt{\int_{B_R} h^{\frac{5}{2}-\delta}(t_1)} \\ &\le R^{\frac{\gamma_0-3}{2}} \sup_{B_R} h^{\frac{5-2\delta}{4}}(t_1). \end{aligned} \tag{2.23}$$

Suppose that

$$\int_{B_R} h_x^2(t_1) \le C R^{\alpha} \tag{2.24}$$

for some α satisfying $\alpha < \gamma$. We claim that, for any fixed $\tau_0 \in (0,1)$,

$$\int_{B_{\tau_0 R}} h_x^2(t_1) \le C' R^{\chi(\alpha)}, \tag{2.25}$$

where $\chi(\alpha) = \frac{\gamma_0 - 3}{2} + \frac{5-2\delta}{8}(1+\alpha)$ as long as $\chi(\alpha) < \gamma$. For, using (2.24) and the fact that $\alpha < \gamma$, we can apply Lemma 2.3 to infer that

$$\int_{B_{\frac{1+\tau_0}{2} R}} h_x^2 \le C' R^{\alpha} \quad \forall t \in [t_1, T].$$

By Lemma 2.2,

$$\sup_{B_{\tau_0 R} \times [t_1, T]} h \le C R^{\frac{1+\alpha}{2}}.$$

Plugging this estimate into (2.23) yields (2.25). Now, using the fact that (2.24) holds for $\alpha = 0$, bootstrapping between (2.24) and (2.25) increases the power in R. Noting that the intersection of χ and $y = x$ is at $\alpha^* = (4\gamma_0 - 7 - 2\delta)/(3 + 2\delta)$, we conclude that the lemma holds if $\gamma < (4\gamma_0 - 7 - 2\delta)/(3 + 2\delta)$. Note that in (2.17) it is required that $\gamma < 3$, and this holds if $(4\gamma_0 - 7 - 2\delta)/(3 + 2\delta) < 3$, that is, $\gamma_0 \le 4 + 2\delta$. ∎

An estimate on the Hausdorff dimension of the rupture set is also available in the range $(1.5, 2]$. We need to work a bit harder for it though.

THEOREM 2.4. *Let h be a suitable weak solution of* (2.1)*, where $n \in (1.5, 2]$ and f satisfies either* (2.5) *or* (2.6)*. The Hausdorff dimension of the rupture set of h cannot exceed $1 + 6/n$.*

Proof. Let us consider the case where f satisfies (2.6) only. We claim that at a rupture point (x_0, T) there exists $\{R_j\} \downarrow 0$ such that

$$\iint_{Q_{R_j}} h^n \Big| \big(h_{xx} + f(h)\big)_x \Big|^2 > R_j^{\gamma_0},$$

$$\iint_{Q_{R_j}} h^{-\frac{5}{2}+\delta} h_x^4 > R_j^{\gamma_0},$$

or

$$\iint_{Q_{R_j}} h^{\frac{3}{2}-n+\delta} > R_j^{\gamma_0+\lambda}$$

hold when γ_0 and δ satisfy $(2.22)'$ and λ is a fixed positive number. For, were this not true, that means (2.7), (2.14), and

$$\iint_{Q_R} h^{\frac{3}{2}-n+\delta} \le R^{\gamma_0+\lambda} \tag{2.26}$$

hold for all sufficiently small R. By Lemma 2.4,

$$\int_{B_{R/2}} h_x^2(t_1) \le CR^{\gamma} \quad \text{for } \gamma < \min\left\{\frac{4\gamma_0 - 7 - 2\delta}{3+2\delta}, \frac{8-n}{n}\right\} = \frac{4\gamma_0 - 7 - 2\delta}{3+2\delta}.$$

By Lemma 2.3 we further infer that

$$\int_{B_{R/4}} h_x^2(t) \le C'R^{\gamma} \quad \forall t \in [t_1, T].$$

Consequently,

$$h(x,T) \le C|x - x_0|^{\frac{1+\gamma}{2}} \quad \forall x \text{ sufficiently close to } x_0.$$

On the other hand, (2.26) implies that there exists some $t_2 \in \Gamma$ such that

$$\int_{B_R} h^{\frac{3}{2}-n+\delta}(t_2) \le CR^{\gamma_0-4+\lambda}.$$

By Lemma 2.5,

$$\int_{B_{R/2}} h^{\frac{3}{2}-n+\delta}(t) \le C'R^{\gamma_0-4+\lambda} \quad \forall t \in [t_2, T].$$

So

$$\begin{aligned} 4R^{\gamma_0-4+\lambda} &\ge \int_{B_{R/2}} h^{\frac{3}{2}-n+\delta} \\ &\ge C \int_{B_{R/2}} |x - x_0|^{\frac{1+\gamma}{2}(\frac{3}{2}-n+\delta)} \\ &\ge CR^{\frac{1+\gamma}{2}(\frac{3}{2}-n+\delta)+1} \end{aligned}$$

if

$$\left(\frac{1+\gamma}{2}\right)\left(\frac{3}{2}-n+\delta\right)+1>0.$$

To derive a contradiction we choose

$$\gamma_0 = 1+\frac{6}{n}$$

and

$$\gamma = \frac{4\gamma_0-7}{3}-\sigma,$$

where σ is chosen to be small as δ is small and such that

$$\gamma < \frac{4\gamma_0-7-2\delta}{3+2\delta}$$

and

$$\frac{1+\gamma}{2}\left(\frac{3}{2}-n+\delta\right)+1>0.$$

As $\delta,\ \sigma\to 0$, and γ tends to $(4\gamma_0-7)/3$ and

$$\gamma_0-4+\lambda>\gamma_0-4=\frac{1+\gamma}{2}\left(\frac{3}{2}-n\right)+1,$$

the contradiction holds.

Now a covering argument as in the proof of Theorem 2.3 finishes the job. □

LEMMA 2.5. *Let h be a suitable weak solution* (2.1)*, where $n\in(1.5,2]$ and either* (2.5) *or* (2.6) *hold. There exists a small $R_0>0$ such that if*

$$\int_{B_R} h^{\frac{3}{2}-n+\delta}(t_1)\le C_6R^{\gamma},\quad 0<\delta<n-\frac{3}{2}\ \ \textit{small,}$$

for some $\gamma\in[0,1], R<R_0$ and $t_1\in\Gamma\cap[T-R^4,T]$, then for every $\tau_0\in(0,1)$,

$$\int_{B_{\tau_0 R}} h^{\frac{3}{2}-n+\delta}(t)\le CR^{\gamma}\quad \forall t\in[t_1,T],$$

where C depends on n,q,δ,C_6, and τ_0.

Proof. We focus on the case where (2.6) holds, since the other case can be handled similarly. By the local entropy inequality (2.3) $\forall t\ge t_1$,

$$\begin{aligned}\int_{B_R} h^{\frac{3}{2}-n+\delta}\phi^4 \le &-\sigma\left[\int_{t_1}^{t}\int_{B_R} h^{-\frac{1}{2}+\delta}h_{xx}^2\phi^4+\int_{t_1}^{t}\int_{B_R} h^{-\frac{5}{2}+\delta}h_x^4\phi^4\right]\\ &+C\left[\int_{t_1}^{t}\int_{B_R} h^{\frac{3}{2}+\delta}(\phi_x^4+\phi^2\phi_{xx}^2)+\int_{t_1}^{t}\int_{B_R} h^{2q-\frac{1}{2}+\delta}\phi^4\right]\\ &+\int_{B_R} h^{\frac{3}{2}-n+\delta}(t_1),\end{aligned}$$

where ϕ is supported in B_R. Using $q \geq 1 - n/2$, we have

$$\int_{t_1}^{t} \int_{B_R} h^{\frac{3}{2}+\delta} \left(\phi_x^4 + \phi^2 \phi_{xx}^2\right) \leq CR,$$
$$\int_{t_1}^{t} \int_{B_R} h^{2q-\frac{1}{2}+\delta} \phi^4 \leq C(t-t_1) \sup_{[t_1,t]} \int_{B_R} h^{\frac{3}{2}-n+\delta} \phi^4.$$

Therefore,

$$\sup_{[t_1,t]} \int_{B_R} h^{\frac{3}{2}-n+\delta} \phi^4 \leq C \left[(t-t_1) \sup_{[t_1,t]} \int_{B_R} h^{\frac{3}{2}-n+\delta} + R^{\gamma} \right].$$

By taking t_2 so that $C(t_2 - t_1) = 1/2$, and $\phi \equiv 1$ on $B_{\tau R}$, we get

$$\sup_{[t_1,t_2]} \int_{B_{\tau R}} h^{\frac{3}{2}-n+\delta} \leq CR^{\gamma}.$$

Now we may use t_2 as the initial time to extend the estimate beyond t_2. After finitely many times extensions the lemma follows. □

3. An estimate on rupture times. In this section we establish an estimate on the Hausdorff dimension of the rupture times. An instant t is a rupture time for a suitable weak solution u of (2.1) if $u(x,t) = 0$ for some x. As our derivation does not involve the use of the local energy or entropy inequalities, the condition (H_3) in the definition of a suitable weak solution is not needed. In other words, the main result, Theorem 3.1, holds for any weak solution of (2.1) satisfying (H_1) and (H_2) only.

We start with the following interpolating inequalities controlling $\sup u^{-1}$ in terms of integrals involving the derivatives of the function u and the integrals of its negative powers.

LEMMA 3.1. *Let u be a nonnegative H^1-function of period L. Then for any $p > 2$, there exists a constant C depending on p and L such that*

$$\sup u^{-1} \leq C \left[\left(\int u^{-p} \right)^{\frac{1}{p}} + \left(\int u^{-p} \right)^{\frac{1}{p-2}} \left(\int u_x^2 \right)^{\frac{1}{p-2}} \right].$$

Proof. We have, for $\alpha > 0$,

$$u^{-\alpha}(x) \leq u^{-\alpha}(y) + \alpha \left| \int u^{-\alpha-1} u_x \right|$$
$$\leq u^{-\alpha}(y) + \alpha \left(\int u^{-2\alpha-2} \right)^{\frac{1}{2}} \left(\int u_x^2 \right)^{\frac{1}{2}}.$$

Integrating in y yields

$$u^{-\alpha}(x) \leq \frac{1}{L} \int u^{-\alpha} + \alpha \left(\int u^{-2\alpha-2} \right)^{\frac{1}{2}} \left(\int u_x^2 \right)^{\frac{1}{2}}.$$

Letting $2\alpha = p - 2$, we get

$$u^{-1}(x) \leq C \left[\left(\int u^{\frac{-p+2}{2}} \right)^{\frac{2}{p-2}} + \left(\int u^{-p} \right)^{\frac{1}{p-2}} \left(\int u_x^2 \right)^{\frac{1}{p-2}} \right]$$
$$\leq C \left[\left(\int u^{-p} \right)^{\frac{1}{p}} + \left(\int u^{-p} \right)^{\frac{1}{p-2}} \left(\int u_x^2 \right)^{\frac{1}{p-2}} \right]. \qquad \square$$

LEMMA 3.2. *Let u be a nonnegative H^2-function of period L. Then for $p > 2/3$, there exists a constant C depending on p and L such that*

$$\sup u^{-1} \le C\left[\left(\int u^{-p}\right)^{\frac{3}{3p-2}}\left(\int u_{xx}^2+\beta\right)^{\frac{1}{3p-2}}+\left(\int u_{xx}^2+\beta\right)^{-\frac{1}{2}}\right] \quad \forall \beta \ge 0.$$

Proof. It suffices to establish this inequality by taking u to be positive and C^2. The general case follows from approximation. We assume $u(0) = \inf u$ and set

$$E = \int u^{-p} \text{ and } F = \int u_{xx}^2 + \beta.$$

Then by using Taylor's formula

$$u(x) = u(0) + u_x(0)x + \int_0^x u_{xx}(t)(x-t)dt,$$

we have, for all x,

$$u(x) \le u(0) + \frac{2}{3}F^{\frac{1}{2}}x^{\frac{3}{2}} \quad \forall x \in [0, L].$$

So

$$\int_0^L \frac{dx}{\left(u(0)+\frac{2}{3}F^{\frac{1}{2}}x^{\frac{3}{2}}\right)^p} \le \int u^{-p} \le E,$$

which implies that

$$\int_0^L \frac{dx}{\left(x+2u(0)^{\frac{2}{3}}F^{-\frac{1}{3}}\right)^{\frac{3p}{2}}} \le CEF^{\frac{p}{2}}.$$

After an integration over $(0, L/2)$, we have

$$\left(2u(0)^{\frac{2}{3}}F^{-\frac{1}{3}}\right)^{1-\frac{3p}{2}} \le C\left(EF^{\frac{p}{2}}+1\right),$$

i.e.,

$$\inf u^{-1} \le C\left(E^{\frac{3}{3p-2}}F^{\frac{1}{3p-2}}+F^{-\frac{1}{2}}\right). \qquad \square$$

To apply this lemma to our weak solution, we take advantage of the finite quantities K_4 and K_3. A direct calculation shows that $u = h^{\frac{3+2\delta}{4}}$ satisfies

$$\int u_{xx}^2 \le C_\delta\left(\int h^{-\frac{1}{2}+\delta}h_{xx}^2+\int h^{-\frac{5}{2}+\delta}h_x^4\right).$$

Now, using Lemma 3.2 ($\beta = 1$) we have, for any $p' > 2/3$,

(3.1)

$$\sup h^{-\frac{3+2\delta}{4}} \le C\left[\left(\int h^{-p'\frac{3+2\delta}{4}}\right)^{3/(3p'-2)}\left(\int h^{-\frac{1}{2}+\delta}h_{xx}^2+\int h^{-\frac{5}{2}+\delta}h_x^4\right)^{1/(3p'-2)}+1\right].$$

COROLLARY 3.1. *Let h be a suitable weak solution of* (2.1) *for $n \in (2, 3.5)$. Then*

$$K_5 \equiv \int_0^T \left(\int h^{-p} \right)^{\frac{3(n-2)-\delta}{p+\frac{3}{2}-n}} < \infty \quad \forall p > n - 3/2,\ \delta \in (0, 3(n-2)).$$

Proof. When $\delta \in (0, 3(n-2))$, we can take $p' = (4n-6)/(3+2\delta) > 2/3$ so that $p'(3+2\delta)/4 = n - 3/2$ in (3.1) to get, for $p > n - 3/2$,

$$\begin{aligned} \int h^{-p} &\le \sup(h^{-1})^{p+\frac{3}{2}-n} \int h^{\frac{3}{2}-n} \\ &\le CK_3 \left[K_3^{\frac{1}{n-2-\frac{\delta}{3}}} \left(\int h^{-\frac{1}{2}+\delta} h_{xx}^2 + \int h^{-\frac{5}{2}+\delta} h_x^4 + 1 \right)^{\frac{1}{3(n-2)-\delta}} + 1 \right]^{p+\frac{3}{2}-n}. \end{aligned}$$

Raise both sides to the power $[3(n-2)-\delta]/(p+\frac{3}{2}-n)$ and then integrate over time. The corollary follows from the boundedness of K_3 and K_4. □

Now, we have the following result which contains Theorem 1.3.

THEOREM 3.1. *Let h be a suitable weak solution of* (2.1)*, where $n \in (2, 3.5)$ and either* (2.5) *or* (2.6) *holds. The Hausdorff dimension of the set of rupture times cannot exceed $1 - 2(n-2)^2/(8-n)$.*

Once again we point out that the theorem holds for any weak solution of (2.1) satisfying (H_1) and (H_2) only.

Proof. Let n be given as above and $p,\ \delta$ satisfy

$$p \in (2, n], \quad p + \frac{3}{2} > n, \quad \delta \in (0, 3(n-2)).$$

We claim that there exists a small ε_0 depending on n, p, δ such that whenever

$$\int_{T-R}^T \left(\int h^{-p} \right)^{\beta} \le \varepsilon_0 R^{\gamma} \quad \forall \text{ small } R,$$

where

$$\beta = \frac{3(n-2)-\delta}{p+\frac{3}{2}-n}, \quad \gamma = 1 - \beta\Big(\frac{p-2}{8-n}\Big),$$

then $h(x_0, T) > 0$. To prove this claim we first observe that by the mean-value theorem there is some $t_1 \in [T-R, T]$ such that

$$\int h^{-p}(\cdot, t_1) \le 2\varepsilon_0^{\frac{1}{\beta}} R^{\frac{\gamma-1}{\beta}}.$$

By Lemma 3.1 and $K_1 < \infty$, h is positive at t_1. By continuity h is positive for all t close to t_1, in particular, it is a classical solution near t_1. Let

$$t^* = \sup\big\{ t_2 \in (t_1, T] :\ h \text{ is positive in } [t_1, t_2] \big\}.$$

We shall show that $t^* = T$. As for $t \in (t_1, t^*)$,

$$\begin{aligned} \frac{d}{dt} \int h^{-p} &= -p(p+1) \int h^{n-p-2} h_{xx}^2 - p(p+1)(n-p-2) \int h^{n-p-3} h_x^2 h_{xx} \\ &\quad + p(p+1) \int h^{n-p-2} f'(h) h_x^2. \end{aligned} \tag{3.2}$$

We estimate the integrals on the right as follows: For $p \in (2, n)$,

$$\begin{aligned}
\left| \int h^{n-p-3} h_x^2 h_{xx} \right| &= \left| \frac{n-p-3}{3} \int h^{n-p-4} h_x^4 \right| \\
&\le C \int h^{p-n} \left| \left(h^{\frac{n-p}{2}}\right)_x \right|^4 \\
&\le C \sup(h^{-1})^{n-p} \int \left| \left(h^{\frac{n-p}{2}}\right)_x \right|^4 \\
&\le C \sup(h^{-1})^{n-p} \left(\int \left(h^{\frac{n-p}{2}}\right)_x^2 \right)^{3/2} \left(\int \left(h^{\frac{n-p}{2}}\right)_{xx}^2 \right)^{1/2} \\
&\le \varepsilon \int \left(h^{\frac{n-p}{2}}\right)_{xx}^2 + C \sup(h^{-1})^{2(n-p)} \left(\int \left(h^{\frac{n-p}{2}}\right)_x^2 \right)^3 \\
&\le C\varepsilon \left(\int h^{n-p-2} h_{xx}^2 + \int h^{n-p-4} h_x^4 \right) + C \sup(h^{-1})^{6+p-n} \left(\int h_x^2 \right)^3,
\end{aligned}$$

which means, for small ε,

$$\int h^{n-p-4} h_x^4 \le C \left[\varepsilon \int h^{n-p-2} h_{xx}^2 + \sup(h^{-1})^{6+p-n} \right].$$

By Lemma 3.1 we further have

$$\int h^{n-p-4} h_x^4 \le C \left[\varepsilon \int h^{n-p-2} h_{xx}^2 + C \left(1 + \int h^{-p} \right)^{\frac{6+p-n}{p-2}} \right]. \tag{3.3}$$

By modifying the above arguments slightly we see that (3.3) also holds when $p = n$. For the third term on the right-hand side of (3.2) we have

$$\begin{aligned}
\int h^{n-p-2} f'(h) h_x^2 &\le C \left(\int h^{n-p-2} h_x^2 + \int h^{n-p+q-3} h_x^2 \right) \\
&\le C \left(\int h^{n-p-4} h_x^4 + \int h^{n-p} + \int h^{n-p+2q-2} \right) \\
&\le C \left[\int h^{n-p-4} h_x^4 + \left(\int h^{-p} \right)^{\frac{n-p+2q-2}{-p}} \right]. \qquad (3.4)
\end{aligned}$$

Putting (3.3) and (3.4) back to (3.2) yields, after taking ε small,

$$\frac{d}{dt} \int h^{-p} \le C \left(1 + \int h^{-p} \right)^{\mu},$$

where $\mu = \max\{\frac{6+p-n}{p-2}, \frac{p-n-2q+2}{p}\} = \frac{6+p-n}{p-2}$ (because $n \le 8$). By comparing $\int h^{-p}$ with the solution Y of

$$\frac{dY}{dt} = C(1+Y)^{\mu}, \quad Y(t_1) = 2\varepsilon_0^{\frac{1}{\beta}} R^{\frac{\gamma-1}{\beta}},$$

we see that Y exists in $[t_1, T]$, when $\varepsilon_0^{1/\beta}$ is sufficiently small and $\int h^{-p} \le Y(t)$ as long as both functions exist. So $\int h^{-p}$ is finite in $[t_1, T]$. It follows from Lemma 3.1 that h is positive up to T, that is, $t^* = T$.

We fix an ε_0 satisfying the property in the above claim. It follows that T is a rupture time if and only if there exists $R_j \downarrow 0$ such that

$$\int_{T-R_j}^{T} \left(\int h^{-p} \right)^{\beta} > \varepsilon_0 R_j^{\gamma}$$

for each p, n, δ in the range. Applying a covering argument to the time interval $(0, \infty)$ one can show that, using the finiteness of K_5, the $\gamma-$Hausdorff measure of the set of rupture times is finite for for all $\gamma = 1 - \beta(p-2)/(8-n)$. Letting $\delta \to 0$ and then taking $p = n$, we conclude the proof of this theorem. □

4. A result on finite time rupture. For any positive, smooth, periodic function h_0, it is known that

$$\begin{cases} h_t + \left[h^n \left(h_{xx} + f(h) \right)_x \right]_x = 0, \\ h(0) = h_0 \end{cases} \tag{4.1}$$

admits a unique, classical solution in $[0, \omega)$, which is of the same period in x and is maximal in the sense that whenever ω is finite,

$$\inf_{t \uparrow \omega} h(t) = 0 \quad \text{or} \quad \sup_{t \uparrow \omega} h(t) = \infty.$$

A proof of this fact, which can be found in [BF], is based on parabolic regularity theory (Eidelman [Ei] or Friedman [F1]) when there is no f-term in (4.1). Nevertheless, an examination of their arguments shows that it still works with the presence of the f-term as long as f is in $C^{2,\alpha}((0,\infty))$ for some $\alpha \in (0,1)$. In the following we denote by $X(L, c_0)$ the subset of $H^1(S_L)$ consisting of all of the positive functions with area c_0. As the area is conserved by the equation in (4.1), the classical solution of (4.1) forms a flow in $X(L, c_0)$. For our result the following condition on f will be imposed:

$$f(0) = 0, \quad |f'(z)| \le C_1 z^{q-1}, \quad z \in (0,1] \quad \text{for some } q \ge \frac{1-n}{2}. \tag{4.2}$$

THEOREM 4.1. *Let $n \in (0, 1/2)$ be in (4.1), and $f \in C([0,\infty)) \cap C^{2,\alpha}((0,\infty))$ satisfy (4.2). Suppose that there exists a bounded open set $\mathcal{U}$ in $X(L, c_0)$ satisfying the following properties:*

(P_1) The closure of $\mathcal{U}$ does not contain any positive steady state, fully supported droplet with zero contact angle, or an array of identical droplets with zero contact angle of (4.1); and

(P_2) Any maximal solution h starting at some positive smooth h_0 in $\mathcal{U}$ stays inside $\mathcal{U}$.

Then ω is finite, that is, h ruptures in finite time.

From this theorem we deduce the following result on finite time rupture.

COROLLARY 4.1. *Consider (4.1), where $n \in (0, 1/2)$ and f is a power law, with q satisfying $(1-n)/2 \le q < 3$. Then for any positive c_0, there exists some L_0 depending on q and c_0 such that, for any $L \ge L_0$, there are solutions of (4.1) in $X(L, c_0)$ which rupture in finite time.*

Steady states arise in the study of long-time behavior of the solutions of (4.1), and the definition of a steady state is motivated from the energy dissipation relation. Indeed, a nonnegative function h in the closure of $X(L, c_0)$ is called a steady state of (4.1) if, on each component of the positivity set of h, $h_{xx} + f(h) = c$ holds for some

constant c (which could be different on components). Besides the constant states c_0/L that always exist in any $X(L, c_0)$, there could be other steady states in the same $X(L, c_0)$. Among others, a droplet is a steady state that is positive in some $(-a, a), a \le L/2$ and equal to 0 in $[-L/2, L/2] \setminus (-a, a)$ after a suitable translation. It is of full support if a is equal to $L/2$. The droplet has zero contact angle if its derivatives vanish at $\pm a$. A configuration of droplets is a steady state, where the positivity set $\{x \in (-L/2, L/2) : h(x) > 0\}$ is a union of at least two disjoint, open subintervals of $(-L/2, L/2)$. An array of identical droplets is a configuration of droplets in which, after a suitable translation, there is a partition of $(-L/2, L/2)$ into subintervals with equal length so that the restrictions of the steady state to these subintervals are congruent droplets with full support.

Criteria for the existence and multiplicity of positive steady states and droplets in a given $X(L, c_0)$ for general f and for power laws can be found in [LP1], and the stability of positive steady states, especially those for power laws, is studied systematically in [LP2], [LP3], and [LP4].

Proof of Theorem 4.1. We let $h(t)$ be a solution of (4.1) starting inside $\mathcal{U}$. Under (P_2), h cannot blow up in any time, so either it ruptures in finite time or it is a positive global solution. Assuming that the latter holds we will derive a contradiction.

As h stays inside $\mathcal{U}$, it is uniformly bounded in H^1-norm. From the energy dissipation relation

$$\mathcal{E}(h) + \int_0^T \int h^n \Big| \Big(h_{xx} + f(h) \Big)_x \Big|^2 = \mathcal{E}(h_0),$$

where the energy is given by

$$\mathcal{E}(h) = \frac{1}{2} \int h_x^2 - \int F(h), \quad F'(z) = f(z), \quad F(0) = 0,$$

we know that

$$\int_0^\infty \int \Big| \Big(h_{xx} + f(h) \Big)_x \Big|^2 < \infty. \tag{4.3}$$

On the other hand, the entropy dissipation relation is given by

$$\frac{d}{dt} \int h^{2-n} = -(2-n)(1-n) \int \Big(h_{xx}^2 - f(h) h_{xx} \Big).$$

As h is uniformly bounded for all time and $n < 1$, we can find a constant C_2 so that

$$\int_T^{T+1} \int \Big(h_{xx}^2 - f(h) h_{xx} \Big) \le C_2 \tag{4.4}$$

holds for all $T > 0$. From (4.3) and (4.4) it is easy to find $\{t_j\} \to \infty$ and a constant C_3 such that, for $h_j = h(t_j)$,

$$\int h_j^n \Big(h_{jxx} + f(h_j) \Big)_x^2 \to 0 \tag{4.5}$$

and

$$\int h_{jxx}^2 \le C_3. \tag{4.6}$$

By passing to a subsequence if necessary, we may further assume that

$$\begin{aligned} &h_j \to \hat{h} \text{ in } C^{1,\frac{1}{2}}(S_L),\\ &h_j \rightharpoonup \hat{h} \text{ in } H^2(S_L),\\ &h_j \to \hat{h} \text{ in } C^{2,\frac{1}{2}}_{loc}\big(\{\hat{h}>0\}\big), \end{aligned}$$

where $\hat{h}$ is a steady state of (4.1). By (P_1) $\hat{h}$ cannot be positive, so after a suitable translation, we may assume that there is a component (a,b) of the positivity set of h in $(-L/2, L/2)$ on which

$$\hat{h}'' + f(\hat{h}) = c.$$

(4.6) implies that $\hat{h}'(a) = \hat{h}'(b) = 0$. By integrating this equation we obtain

$$\frac{\hat{h}_x^2}{2} + F(\hat{h}) = c\hat{h}.$$

As $f(0) = 0$, the constant c must be positive unless the derivatives of $\hat{h}$ vanish at the endpoints. But the derivative cannot be equal to 0, for the vanishing of both $\hat{h}$ and $\hat{h}_x$ at the endpoints would force $\hat{h}$ to vanish identically by the uniqueness of the solution to the ordinary differential equation satisfied by $\hat{h}$. Therefore c is positive. By (P_1) there are two possibilities: First, there exists some $\delta_0 > 0$ such that $\hat{h} = 0$ in $[a-\delta_0, a]$ or $[b, b+\delta_0]$. Second, there exist two subintervals (a,b) and (b,b_1) in $(-L/2, L/2)$ such that $\hat{h}_{xx} + f(\hat{h}) = c_1, x \in (a,b)$ and $\hat{h}_{xx} + f(\hat{h}) = c_2, x \in (b,b_1)$, where c_1 and c_2 are different positive constants. Moreover, $\hat{h}_x^-(b) = \hat{h}_x^+(b) = 0$ holds.

Let us treat the first case first. For simplicity we take b to be 0 and so $\hat{h} = 0$ in $[0, \delta_0]$. For each $0 < \delta \le \delta_0$, we rescale h_j and $\hat{h}$ by

$$h_j(x) = \delta^2 H_j\left(\frac{x}{\delta}\right), \quad \hat{h}(x) = \delta^2 \hat{H}\left(\frac{x}{\delta}\right).$$

Then, for each fixed δ,

$$\begin{aligned} H_j \to \hat{H} \qquad & \text{in } C^{1,\frac{1}{2}}([-1,1]) \cap C^{2,\frac{1}{2}}_{loc}((-1,1)),\\ \hat{H}(y),\ \hat{H}_y(y) \to 0 \qquad & \text{as } y \uparrow 0,\\ \hat{H}_{yy}(y) \to c \qquad & \text{as } y \uparrow 0. \end{aligned}$$

We claim that there exists a constant C_* independent of δ such that

$$\int_{-1}^{1} H_j^n H_{jyyy}^2 \ge C_* > 0 \quad \forall j \text{ large}. \tag{4.7}$$

To prove (4.7) we first fix $\sigma_0 \in (-1, 0)$ such that

$$0 > H_{jy}(\sigma_0) \ge -\frac{c}{32}$$

and

$$H_{jyy}(\sigma_0) \ge \frac{c}{2}$$

hold (σ_0 may depend on j). Next we claim that, for each large j, there exists some $\xi \in [\sigma_0, 1]$ such that

$$H_{jyy}(\xi) = \frac{c}{4}.$$

For, if not, then $H_{jyy} > c/4$ in $[\sigma_0, 1]$. By Taylor expansion at σ_0, there exists some $z \in (\sigma_0, 1)$ such that

$$\begin{aligned} H_j(1) &= H_j(\sigma_0) + H_{jy}(\sigma_0)(1-\sigma_0) + \frac{H_{jyy}(z)}{2}(1-\sigma_0)^2 \\ &\geq -\frac{c}{32}(1-\sigma_0) + \frac{c}{8}(1-\sigma_0)^2 \\ &\geq \frac{c}{16}, \end{aligned}$$

contradicting the fact that $H_j(1)$ tends to 0 as $j \to \infty$. Thus $H_{jyy} = c/4$ somewhere in $(\sigma_0, 1)$. Let

$$\sigma_1 = \inf\left\{\sigma > \sigma_0 : \quad H_{jyy}(\sigma) = c/4\right\} < 1.$$

We consider two subcases separately:

(a) H_{jy} does not change sign in (σ_0, σ_1), and

(b) $H_{jy} = 0$ somewhere in (σ_0, σ_1).

In subcase (a), $H_{jy} < 0$ in (σ_0, σ_1) because $H_{jy}(\sigma_0) < 0$. By Taylor expansion at σ_1, for $y \in (\sigma_0, \sigma_1)$,

$$\begin{aligned} H_j(y) &= H_j(\sigma_1) + H_{jy}(\sigma_1)(y-\sigma_1) + \frac{H_{jyy}(z)}{2}(y-\sigma_1)^2 \\ &\geq H_j(\sigma_1) + \frac{c}{8}(y-\sigma_1)^2 \\ &\geq \frac{c}{8}(y-\sigma_1)^2 \quad \forall j \text{ large.} \end{aligned} \tag{4.8}$$

On the other hand,

$$\begin{aligned} \int_{\sigma_0}^{\sigma_1} \left|H_{jyyy}\right| &\geq \left|H_{jyy}(\sigma_1) - H_{jyy}(\sigma_0)\right| \\ &\geq \frac{c}{4}. \end{aligned}$$

So,

$$\begin{aligned} \int_{\sigma_0}^{\sigma_1} H_j^n \left|H_{jyyy}\right|^2 \int_{\sigma_0}^{\sigma_1} H_j^{-n} &\geq \left(\int_{\sigma_0}^{\sigma_1} \left|H_{jyyy}\right|\right)^2 \\ &\geq \frac{c^2}{16}. \end{aligned}$$

By (4.8),

$$\int_{\sigma_0}^{\sigma_1} H_j^{-n} \leq \left(\frac{8}{c}\right)^n \int_{\sigma_0}^{\sigma_1} \frac{1}{(y-\sigma_1)^{2n}} = \left(\frac{8}{c}\right)^n \frac{2}{1-2n}.$$

It follows that

$$\int_{-1}^{1} H_j^n \left|H_{jyyy}\right|^2 \geq C_* \equiv \frac{c^{n+2}(1-2n)}{2^{3n+5}},$$

that is, (4.7) holds.

In subcase (b), let $\sigma_2 \in (\sigma_0, \sigma_1)$ be a point at which $H_{jy}(\sigma_2) = 0$. By Taylor expansion at σ_2, for $y \in (\sigma_0, \sigma_1)$,

$$\begin{aligned} H_j(y) &= H_j(\sigma_2) + \frac{H_{jyy}(z)}{2}(y-\sigma_2)^2 \\ &\geq \frac{c}{8}(y-\sigma_2)^2, \end{aligned}$$

and the above argument still works to show that (4.7) holds for the same C_*.

Now, we scale (4.7) back to get

$$\begin{aligned} \int_{-\delta}^{\delta} h_j^n h_{jxxx}^2 &= \delta^{2n-1} \int_{-1}^{1} H_j^n H_{jyyy}^2 \\ &\geq C_* \delta^{2n-1}. \end{aligned} \tag{4.9}$$

However, on the other hand, for small δ,

$$\begin{aligned} \int_{-\delta}^{\delta} h_j^n h_{jxxx}^2 &\leq \int_{-\delta}^{\delta} h_j^n \Big(h_{jxxx} + f'(h_j) h_{jx} \Big)^2 + 4 \int_{-\delta}^{\delta} f'(h_j)^2 h_{jx}^2 \\ &\leq \int_{-1}^{1} h_j^n \Big(h_{jxxx} + f'(h_j) h_{jx} \Big)^2 + C \int_{-\delta}^{\delta} \big(1 + h_j^{n+2q-2}\big) h_{jx}^2. \end{aligned}$$

As

$$\begin{aligned} \int h_j^{n+2q-2} h_{jx}^2 &= \frac{-1}{n+2q-1} \int h_j^{n+2q-1} h_{jxx} \\ &\leq C \left(\int h_j^{2(n+2q-1)} \right)^{1/2} \left(\int h_{jxx}^2 \right)^{1/2}, \end{aligned}$$

by (4.2), (4.5), and (4.6) we conclude that

$$\int_{-\delta}^{\delta} h_j^n h_{jxxx}^2$$

is bounded uniformly in j, but this is in conflict with (4.9) as $n < 1/2$ and δ could be arbitrarily small, so the classical solution cannot be a global one, contradiction holds.

The proof for the second case is similar. Without loss of generality let us assume that $c_1 > c_2$. Again we take $b = 0$ and fix some small δ_0 so that $\hat{h}$ is decreasing in $(-\delta_0, 0)$ and increasing in $(0, \delta_0)$. For each $\delta < \delta_0$ we use the same rescalings as before to define H_j and $\hat{H}$ and claim that (4.7) holds. For, we pick σ_0 in $(-1, 0)$ such that $H_{jyy}(\sigma_0) = c_1 - \varepsilon_0$, where ε_0 is a fixed number less than $(c_1 - c_2)/2$. As before one can show that there is some $\xi \in (\sigma_0, 1)$ such that $H_{jyy}(\xi) = (c_1 + c_2)/2$. So we can define

$$\sigma_1 = \inf\{\sigma > \sigma_0 : H_{jyy}(\sigma) = (c_1 + c_2)/2\} < 1.$$

We consider the two subcases as before. We have in subcase (a) $H_{jyy} \geq \frac{c_1+c_2}{4}(y-\sigma_1)^2$ and in subcase (b) $H_{jyy} \geq \frac{c_1+c_2}{4}(y-\sigma_2)^2$ for $y \in (\sigma_0, \sigma_1)$. In both cases

$$\int_{\sigma_0}^{\sigma_1} |H_{jyyy}| \geq \left| \frac{c_1 - c_2}{2} - \varepsilon_0 \right| > 0.$$

Now the same arguments as before would show that (4.7) holds, so the same contradiction can be drawn. □

Proof of Corollary 4.1. First of all, from the interpolation inequality [F2] for functions in $H^1(S_L)$:

$$\int u^4 \le C\left(\int |u|\right)^2 \left(\int u_x^2 + \int u^2\right),$$

it is easy to see that, for $q \in (-1, 3)$, the energy $\mathcal{E}(h)$ satisfies

$$\int h_x^2 \le C\mathcal{E}(h)$$

for some positive constant C. By a standard argument based on the direct method, the minimization problem

$$m = \inf\{\mathcal{E}(h) : \quad h \in X(L, c_0)\}$$

admits a minimizer in the closure of $X(L, c_0)$, which is a steady state of (4.1).

Let us examine the candidates for the minimizer. First, according to Theorem 10 in [LP3], the constant state c_0/L becomes energy unstable whenever L satisfies $L > L_1$, where L_1 satisfies $L_1^{3-q} c_0^{q-1} = 4\pi^2$, and hence it cannot be a minimizer. Next, the classification results in [LP1] shows that, for any pair (L, c_0), there are, up to translations, at most finitely many positive steady states with period $L/j, j \ge 2$ in $X(L, c_0)$, but they cannot be minimizers as they are also energy unstable due to Theorem 1 of [LP2]. Furthermore, the classification in section 5 of [LP1] asserts that positive steady states with minimal period and fully supported droplets with zero contact angle do not exist if $L > L_2$ for some L_2, depending on c_0 and q. From the same source we know that there are, up to translations, at most finitely many arrays of identical droplets with zero angle in $X(L, c_0)$. We claim that they too cannot be minimizers. For, consider an array of identical droplets h, which vanishes at $-L/2$ and $L/2$. Define h^* to be the function equal to h on $(-L/2, L/2)$ and zero outside this interval. Then the rearrangement of h^* gives a function which is positive inside and zero outside $(-L/2, L/2)$. Extending the restriction of h^* on $(-L/2, L/2)$ as an L-periodic function h^{**}. From the properties of rearrangement h^{**} belongs to the closure of $X(L, c_0)$, with energy strictly lower than that of h; see, for instance, [LiL]. Therefore h cannot be a minimizer. Since the energy is invariant under translation, we conclude that no array of identical droplets can be a minimizer of the energy.

As a result, for any L greater than L_1 and L_2, if we denote by μ the minimum of the energies of all of the possible positive steady states, fully supported droplets with zero contact angle and arrays of identical droplets with zero contact angle in $X(L, c_0)$, then $m < \mu$. Now the corollary follows from Theorem 4.1 by taking

$$\mathcal{U} = \{h \in X(L, c_0) : \mathcal{E}(h) < (m + \mu)/2\}. \qquad \square$$

Remark 4.1. A sharp version of the interpolation inequality used in the above proof is the Nagy's inequality [SP]:

$$\int u^4 \le \frac{9}{4\pi^2}\left(\int |u|\right)^2 \left(\int u_x^2\right),$$

which is valid for all $u \in H^1(\mathbf{R})$. By modifying this inequality one can show that the following inequality holds for all $u \in H^1(S_L)$:

$$\int u^4 \le C \left(\int |u| \right)^2 \left(\int u_x^2 + \int u^2 \right)$$

for any constant $C > 9/4\pi^2$. Using this result, one can show that Corollary 4.1 continues to hold when $q = 3$ provided c_0 satisfies

$$3Bc_0^2 < 8\pi^2.$$

Appendix A. Suitable weak solutions. Here we give a sketch of the proof of Theorem 2.1. We shall treat problem (2.1), where $n \in (1.5, 3.5)$ in (2.1) and f satisfies (2.4) and (2.6). The proof when f satisfies (2.5) instead of (2.6) can be handled in a similar and simpler way.

Step 1. (2.1) *admits a positive, classical solution for all time if* $n \ge 4$ *and* $f \in C^{2,\alpha}([0,\infty))$.

This is achieved by considering the approximation problem ($\varepsilon > 0$)

$$\begin{cases} h_t + \left[\left(|h|^n + \varepsilon \right) \left(h_{xx} + f(h) \right)_x \right]_x = 0, \\ h(0) = h_0 > 0 \text{ in } C^\infty(S_L). \end{cases}$$

One first shows that this problem admits a global classical solution h_ε. In fact, using the interpolation inequality on $H^1(S_L)$ [F2],

$$\int h^4 \le C \left(\int h \right)^2 \left(\int h_x^2 + \int h^2 \right)$$

for some positive constant C in the energy dissipation relation

$$\frac{1}{2} \int h_{\varepsilon x}^2 - \int F(h_\varepsilon) + \int_0^t \int (|h|^n + \varepsilon) \left| \left(h_{\varepsilon xx} + f(h_\varepsilon) \right)_x \right|^2 = \frac{1}{2} \int h_{0x}^2 - \int F(h_0),$$

where F is a primitive of f, with the help of the conservation of area

$$\int h = c_0 \equiv \int h_0,$$

and the growth condition $p < 3$ in (1.4) or $p = 3$ when c_0 is small one obtains a uniform H^1-bound on h_ε. Using the entropy inequality one further shows that, although h_ε may change sign, its limit of as $\varepsilon \downarrow 0$ through a subsequence is a positive function h which solves (2.1). We refer to [BF] and [BP2] for details.

Step 2. (2.1) *admits a positive, classical solution for all time if* $n \ge 4$ *and* $f \in C^{2,\alpha}((0,\infty))$ *satisfies* (2.4) *(*$p < 3$*) and* (2.6) *(*$q > 1 - n/2$*).*

Consider the approximation problem ($\delta > 0$)

$$\begin{cases} h_t + \left[h^n \left(h_{xx} + f_\delta(h) \right)_x \right]_x = 0, \quad f_\delta(z) = f(z + \delta), \\ h(0) = h_0 > 0 \text{ in } C^\infty(S_L). \end{cases}$$

By Step 1, this problem admits a positive, classical solution h_δ, which is uniformly bounded in H^1-norm again by (2.4) ($p < 3$). Further we compute the change rate of the entropy $\int h_\delta^{2-n}$ as follows:

$$\begin{aligned}\frac{d}{dt}\int h_\delta^{2-n} &= (2-n)(1-n)\int h_{\delta x}\big(h_{\delta xx}+f_\delta(h_\delta)\big)_x\\ &\le -\frac{1}{2}(2-n)(1-n)\int h_{\delta xx}^2 + C\int f_\delta^2(h_\delta)\\ &\le -\frac{1}{2}(2-n)(1-n)\int h_{\delta xx}^2 + C\left(\int h_\delta^{2q}+1\right)\\ &\le C\left(\int h_\delta^{2-n}+1\right)\end{aligned}$$

as $q \ge 1-n/2$. It follows that $||h_\delta^{2-n}||_{L^1}$ is uniformly bounded on every interval $[0,T]$. By passing to limit and using $n \ge 4$, h_δ subconverges to a positive, classical solution of (2.1), with initial datum h_0.

We remark that the assertions in Steps 1 and 2 hold when the coefficient h^n is replaced by some positive, smooth $a(h)$, which behaves like h^n near 0.

Step 3. Equation (2.1) *admits a suitable weak solution if* $n \in (1.5, 4)$, (2.4) *and* (2.6) *hold for* $0 \le p < 3$ *and* $q > 1 - n/2$.

Consider the approximation problem ($\varepsilon > 0$)

$$\begin{cases} h_t + \Big[a_\varepsilon(h)\big(h_{xx}+f(h)\big)_x\Big]_x = 0, \quad a_\varepsilon(z) = \dfrac{z^4}{z^{4-n}+\varepsilon},\\ h(0) = h_0 > 0 \text{ in } C^\infty(S_L).\end{cases}$$

As the coefficients behave like h^4 near 0, Step 2 guarantees that this problem admits a positive, classical solution h_ε for all $\varepsilon \le 1$. Moreover, these solutions satisfy

$$\sup_{[0,T]}\int h_{\varepsilon x}^2 \le C_1 \tag{A.1}$$

and

$$\iint_{Q_T} a_\varepsilon(h_\varepsilon)\Big|\big(h_{\varepsilon xx}+f(h_\varepsilon)\big)_x\Big|^2 \le C_2. \tag{A.2}$$

It follows that

$$||h_\varepsilon||_{C^{\frac{1}{2},\frac{1}{8}}_{x,t}(Q_T)} \le C_3 \tag{A.3}$$

(we refer to [BF] for the deduction of (A.3) from (A.1) and (A.2)), and, by (A.1), (A.2), and (2.6) ($q > 1 - n/2$),

$$\begin{aligned}\iint_{Q_T} a_\varepsilon(h_\varepsilon)h_{\varepsilon xxx}^2 &\le 2\iint_{Q_T} a_\varepsilon(h_\varepsilon)\left|\big(h_{\varepsilon xx}+f(h_\varepsilon)\big)_x\right|^2\\ &\quad + 2\iint_{Q_T} a_\varepsilon(h_\varepsilon)f'^2(h_\varepsilon)h_{\varepsilon x}^2\\ &\le C\left(1+\iint_{Q_T} h_\varepsilon^{n+2q-2}h_{\varepsilon x}^2\right)\\ &\le C\left(1+\iint_{Q_T} h_{\varepsilon x}^2\right) \le C_4.\end{aligned} \tag{A.4}$$

Define

$$G_{s,\varepsilon}(z)=\int_z^M\int_y^M\frac{\tau^s}{a_\varepsilon(\tau)}d\tau dy,$$

where M is a fixed number greater than $\sup_{[0,T]}h_\varepsilon$ for all $\varepsilon\le 1$. Then we have

$$\begin{aligned}\frac{d}{dt}\int G_{s,\varepsilon}(h_\varepsilon)=&-\int\left(h_\varepsilon^s h_{\varepsilon xx}^2+s\int h_\varepsilon^{s-2}h_{\varepsilon x}^2h_{\varepsilon xx}\right)\\&-\int f(h_\varepsilon)\left(h_\varepsilon^s h_{\varepsilon xx}+sh_\varepsilon^{s-1}h_{\varepsilon x}^2\right).\end{aligned}\tag{A.5}$$

We estimate the terms on the left of (A.5) as follows: For $s\in(-1/2,1)$, there exists a positive number σ depending on s such that

$$\int h_\varepsilon^s h_{\varepsilon xx}^2+s\int h_\varepsilon^{s-1}h_{\varepsilon x}^2h_{\varepsilon xx}\ge\sigma_s\left(\int h_\varepsilon^s h_{\varepsilon xx}^2+\int h_\varepsilon^{s-2}h_{\varepsilon x}^4\right),$$

and, for $s=-1/2$,

$$\int h_\varepsilon^s h_{\varepsilon xx}^2+s\int h_\varepsilon^{s-1}h_{\varepsilon x}^2h_{\varepsilon xx}\ge 0.$$

Therefore, as $q\ge 1-n/2$,

$$\begin{aligned}&\left|\int f(h_\varepsilon)\big(h_\varepsilon^s h_{\varepsilon xx}+sh_\varepsilon^{s-1}h_{\varepsilon x}^2\big)\right|\\&\qquad\le\delta\left(\int h_\varepsilon^s h_{\varepsilon xx}^2+\int h_\varepsilon^{s-2}h_{\varepsilon x}^4\right)+C_\delta\int h_\varepsilon^s f^2(h_\varepsilon)\\&\qquad\le\delta\left(\int h_\varepsilon^s h_{\varepsilon xx}^2+\int h_\varepsilon^{s-2}h_{\varepsilon x}^4\right)+C_\delta\Big(\int h_\varepsilon^{s-n+2}+1\Big)\\&\qquad\le\delta\left(\int h_\varepsilon^s h_{\varepsilon xx}^2+\int h_\varepsilon^{s-2}h_{\varepsilon x}^4\right)+C_\delta\left(1+\int h_\varepsilon+\int G_{s,\varepsilon}(h_\varepsilon)\right).\end{aligned}$$

In the last step we have used the relation

$$\begin{aligned}G_{s,\varepsilon}(z)&=\int_z^M\int_y^M\tau^{s-n}d\tau dy+\varepsilon\int_z^M\int_y^M\tau^{s-4}d\tau dy\\&=\frac{1}{(s-n+1)(s-n+2)}z^{s-n+2}-\frac{M^{s-n+1}}{s-n+1}z+\frac{M^{s-n+2}}{s-n+2}\\&\quad+\varepsilon\left[\frac{1}{(s-3)(s-2)}z^{s-2}-\frac{M^{s-3}}{s-3}z+\frac{M^{s-2}}{s-2}\right].\end{aligned}\tag{A.6}$$

Putting this identity back to (A.5) yields, for $s \in (-1/2, \min\{1, n-2\})$,

$$\sup_{[0,T]} \left[\int h_\varepsilon^{s-n+2} + \varepsilon \int h_\varepsilon^{s-2} \right] + \sigma \iint_{Q_T} \left(h_\varepsilon^s h_{\varepsilon xx}^2 + h_\varepsilon^{s-2} h_{\varepsilon x}^4 \right) \le C_5. \tag{A.7}$$

Next, we fix $s' > -1/2$ such that $-\frac{1}{2} - s' + 2q > -n + 2$ still holds. By (A.5) and (A.7),

$$\begin{aligned} \frac{d}{dt} \int G_{-\frac{1}{2},\varepsilon}(h_\varepsilon) &\le C \int \left(h_\varepsilon^q + 1\right) \left(h_\varepsilon^{-1/2} h_{\varepsilon xx} + s h_\varepsilon^{-3/2} h_{\varepsilon x}^2 \right) \\ &\le C \int \left(h_\varepsilon^{s'} h_{\varepsilon xx}^2 + h_\varepsilon^{s'-2} h_{\varepsilon x}^4 \right) + C \left(\int h_\varepsilon^{-1-s'+2q} + 1 \right) \\ &\le C \int \left(h_\varepsilon^{s'} h_{\varepsilon xx}^2 + h_\varepsilon^{s'-2} h_{\varepsilon x}^4 \right) + C \left(\int G_{-\frac{1}{2},\varepsilon}(h_\varepsilon) + \int h_\varepsilon + 1 \right), \end{aligned}$$

so

$$\sup_{[0,T]} \left[\int h_\varepsilon^{\frac{3}{2}-n} + \varepsilon \int h_\varepsilon^{-\frac{5}{2}} \right] \le C_6. \tag{A.8}$$

To let ε go to 0, first observe that, by (A.2) and (A.4),

$$\sqrt{a_\varepsilon(h_\varepsilon)} h_{\varepsilon xxx} \rightharpoonup h^{\frac{n}{2}} h_{xxx} \quad \text{in } L^2(Q_T), \tag{A.9}$$

and

$$\sqrt{a_\varepsilon(h_\varepsilon)} \left(h_{\varepsilon xx} + f(h_\varepsilon) \right)_x \rightharpoonup h^{\frac{n}{2}} \left(h_{xx} + f(h) \right)_x \quad \text{in } L^2(Q_T). \tag{A.10}$$

By parabolic regularity, for all $s > -1/2$ (since h_ε are uniformly bounded in ε, (A.7) holds for all $s > -1/2$),

$$\begin{aligned} &h_\varepsilon^{\frac{s}{2}} h_{\varepsilon xx} \to h^{\frac{s}{2}} h_{xx}, \\ &h_\varepsilon^{\frac{s-2}{4}} h_{\varepsilon x} \to h^{\frac{s-2}{4}} h_x, \text{ uniformly in any compact subset of } \{h > 0\}. \end{aligned}$$

Taking $s' \in (-1/2, s)$ and using (A.7), we have

$$\begin{aligned} \iint_{Q_T} \left| h_\varepsilon^{\frac{s}{2}} h_{\varepsilon xx} - h^{\frac{s}{2}} h_{xx} \right|^2 &= \left(\iint_{\{h \ge \delta\}} + \iint_{\{h < \delta\}} \right) \left| h_\varepsilon^{\frac{s}{2}} h_{\varepsilon xx} - h^{\frac{s}{2}} h_{xx} \right|^2 \\ &\le o(\varepsilon) + C\delta^{s-s'} \to 0, \text{ as } \varepsilon \to 0 \text{ and } \delta \to 0, \end{aligned}$$

so

$$h_\varepsilon^{\frac{s}{2}} h_{\varepsilon xx} \to h^{\frac{s}{2}} h_{xx} \quad \text{in } L^2(Q_T). \tag{A.11}$$

Similarly, we have

$$h_\varepsilon^{\frac{s-2}{4}} h_{\varepsilon x} \to h^{\frac{s-2}{4}} h_x \quad \text{in } L^4(Q_T). \tag{A.12}$$

Now, using (A.9)–(A.12) to pass to limit we obtain a subsequence of h_ε converging uniformly to some h in Q_T. From (A.1)–(A.4), (A.7), and (A.8) we see that (H_1) and (H_2) hold. It remains to establish (H_3).

LEMMA A.1. *There exists a set* Γ *of full measure in* $[0,T]$ *such that, for a sequence* $\varepsilon_j \to 0$,

$$h_{\varepsilon_j}(x,t) \to h(x,t) \quad \text{in } H^1(S_L)\ \forall t \in \Gamma,$$

and

$$\int G_{s,\varepsilon_j}(h_{\varepsilon_j})(t)\phi^4 \to \int G_s(h)(t)\phi^4 \quad \forall t \in [0,T]\ \forall \phi \in C^\infty(S_L),\ s \in (-1/2, n-2).$$

Proof. It follows from (A.12) (taking $s=2$) that

$$\iint_{Q_T} |h_{\varepsilon x} - h_x|^2 \to 0 \quad \text{as } \varepsilon \to 0.$$

Therefore, there exists a subset $\Gamma \subset [0,T]$ with full measure such that, for a subsequence $\varepsilon = \varepsilon_j$,

$$\int |h_{\varepsilon x} - h_x|^2(t) \to 0 \quad \forall t \in \Gamma.$$

To prove the second statement, observe that each $t \in [0,T]$, K_3 implies that $h(x,t) > 0$ a.e. in S_L. So it follows (A.6), (A.7), the pointwise convergence

$$G_{s,\varepsilon}(h_\varepsilon)(t) \to G_s(h)(t) \quad \text{a.e. in } S_L,$$

and the dominated convergence theorem. ∎

Now we can prove (H_3). First, by a direct computation we have

$$\begin{aligned}
&\frac{1}{2}\int h_{\varepsilon x}^2\phi^4 - \int F(h_\varepsilon)\phi^4 + \int_{t_0}^t \int a_\varepsilon(h_\varepsilon)\Big|\big(h_{\varepsilon xx} + f(h_\varepsilon)\big)_x\Big|^2\phi^4 \\
&\qquad = -\int_{t_0}^t \int a_\varepsilon(h_\varepsilon)\big(h_{\varepsilon xx} + f(h_\varepsilon)\big)_x\big(h_{\varepsilon xx} + f(h_\varepsilon)\big)(\phi^4)_x \\
&\qquad\quad -\int_{t_0}^t \int a_\varepsilon(h_\varepsilon)\big(h_{\varepsilon xx} + f(h_\varepsilon)\big)_x\Big[h_{\varepsilon xx}(\phi^4)_x + h_{\varepsilon x}(\phi^4)_{xx}\Big] \\
&\qquad\quad +\frac{1}{2}\int h_{\varepsilon x}^2(t_0)\phi^4 - \int F(h_\varepsilon(t_0))\phi^4, \qquad \text{(A.13)}
\end{aligned}$$

and, for $s = -\frac{1}{2} + \delta \in (-\frac{1}{2}, \min\{1, n-2\})$,

$$\begin{aligned}
\frac{d}{dt}\int G_{s,\varepsilon}(h_\varepsilon)\phi^4 &= \int G'_{s,\varepsilon}(h_\varepsilon)h_{\varepsilon t}\phi^4 \\
&= -\Big[\int h_\varepsilon^s h_{\varepsilon xx}^2\phi^4 + s\int h_\varepsilon^{s-1}h_{\varepsilon x}^2 h_{\varepsilon xx}\phi^4\Big] - \int h_\varepsilon^s h_{\varepsilon x}h_{\varepsilon xx}(\phi^4)_x \\
&\quad -\int f(h_\varepsilon)\big[h_\varepsilon^s h_{\varepsilon xx}\phi^4 + sh_\varepsilon^{s-1}h_{\varepsilon x}^2\phi^4 + h_\varepsilon^s h_{\varepsilon x}(\phi^4)_x\big] \\
&\quad -\int \big[a'_\varepsilon(h_\varepsilon)G'_{s,\varepsilon}(h_\varepsilon) + a_\varepsilon(h_\varepsilon)G''_{s,\varepsilon}(h_\varepsilon)\big]h_{\varepsilon x}\big[h_{\varepsilon xx} + f(h_\varepsilon)\big](\phi^4)_x \\
&\quad -\int a_\varepsilon(h_\varepsilon)G'_{s,\varepsilon}(h_\varepsilon)\big[h_{\varepsilon xx} + f(h_\varepsilon)\big](\phi^4)_{xx} \\
&\le -\sigma\Big[\int h_\varepsilon^s h_{\varepsilon xx}^2\phi^4 + \int h_\varepsilon^{s-2}h_{\varepsilon x}^4\phi^4\Big] \\
&\quad + C\Big[\int h_\varepsilon^{s+2}\big(\phi_x^4 + \phi^2\phi_{xx}^2\big) + \int h_\varepsilon^s f^2(h_\varepsilon)\phi^4\Big], \qquad \text{(A.14)}
\end{aligned}$$

where the constants σ and C are positive constants. Note that, in the last inequality of (A.14), we have used

$$\begin{aligned} a_\varepsilon(z)G_{s,\varepsilon}(z) &= z^s, \\ |a'_\varepsilon(z)G'_{s,\varepsilon}(z)| &\le C_7 z^s, \\ |a_\varepsilon(z)G'_{s,\varepsilon}(z)| &\le C_8 z^{s+1}, \end{aligned}$$

and, for $-1/2 < s < 1$,

$$\begin{aligned} &\int h_\varepsilon^s h_{\varepsilon xx}^2 \phi^4 + s\int h_\varepsilon^{s-1} h_{\varepsilon x}^2 h_{\varepsilon xx}\phi^4 \\ &\qquad \ge \sigma_1 \left[\int h_\varepsilon^s h_{\varepsilon xx}^2\phi^4 + \int h_\varepsilon^{s-2} h_{\varepsilon x}^4 \phi^4\right] - C_1 \left|\int h_\varepsilon^{s-1} h_{\varepsilon x}^3 (\phi^4)_x\right| \\ &\qquad \ge \sigma_2 \left[\int h_\varepsilon^s h_{\varepsilon xx}^2\phi^4 + \int h_\varepsilon^{s-2} h_{\varepsilon x}^4 \phi^4\right] - C_2 \int h_\varepsilon^{s+2}\phi_x^4. \end{aligned}$$

Using (A.9)–(A.12) we can pass limit in (A.13) to obtain the local energy inequality. Similarly, by passing limit in the integral form of (A.14) using also Lemma A.1, we see that the local entropy inequality also holds.

Finally, by an approximation argument the initial data can be taken to be nonnegative H^1-functions, with finite $||h_0^{3/2-n}||_{L^1}$. One may consult [BF] for the proof of $\lim_{t\to 0} h(t) = h_0$ in H^1. We have finished our outline of the proof of Theorem 2.1.

Appendix B. The parabolic Hausdorff dimension. Same as the Hausdorff measures, fourth order parabolic Hausdorff measures can be defined by a general construction due to Carathéordory; see [Fe]. Here we state their definitions and list some of their basic properties.

Let s be nonnegative, and $\delta > 0$. For any set X in $\mathbf{R}^2$, we let $\mathcal{C}$ be a collection of cylinders Q_α of the form $\{(x,t) : |x - x_0| < \rho_\alpha, \text{ and } |t - t_0| < \rho_\alpha^4\}$ $(\rho_\alpha > 0)$ whose union contains X. Then let

$$\mathcal{P}_\delta^s(X) \equiv \inf\left\{\sum \rho_\alpha^s : \text{ All collections } \mathcal{C} \text{ that cover } X \text{ with } \rho_\alpha < \delta\right\},$$

and define the s-parabolic Hausdorff measure to be

$$\mathcal{P}^s(X) = \sup_{\delta\to 0} \mathcal{P}_\delta^s(X).$$

Then $\mathcal{P}_\delta^s$ and $\mathcal{P}^s$ are outer measures for which all Borel subsets are measurable. We define the s-parabolic Hausdorff dimension to be the infimum of $\{s : \mathcal{P}^s(X) = 0.\}$. It is clear that $\mathcal{P}^t(X) = 0$ if $\mathcal{P}^s(X) < \infty$ whenever $t > s$, and $\mathcal{P}^5(X)$ is positive for any bounded, open X.

We state the following version of the Vitali covering theorem; see section 6 in [CKN] for a proof.

THEOREM B.1. *Let $\mathcal{C}$ be a collection of parabolic cylinders described as above. Assuming that there is a uniform bound on the diameters of these cylinders, we can find a countable subcollection consisting of mutually disjoint cylinders $Q_j = \{(x,t) : |x - x_0| < \rho_j, |t - t_0| < \rho_j^4\}$ so that the union of the enlarged cylinders $Q'_j = \{(x,t) : |x - x_0| < 5\rho_j, |t - t_0| < 625\rho_j^4\}$ covers $\mathcal{C}$.*

Acknowledgments. We are grateful to the referees whose suggestions improve the presentation of this work.

REFERENCES

[BBDP] E. Beretta, M. Bertsch, and R. Dal Passo, *Nonnegative solutions of a fourth order nonlinear degenerate parabolic equation*, Arch. Rat. Mech. Anal., 129 (1995), pp. 175–200.

[B1] F. Bernis, *Finite speed of propagation and continuity of the interface for thin viscous flows*, Adv. Differential Equations, 1 (1996), pp. 337–368.

[B2] F. Bernis, *Finite speed of propagation for thin viscous flows when* $2 \leq n < 3$, C. R. Acad. Sci. Paris Sér. I Math., 322 (1996), pp. 1169–1174.

[BF] F. Bernis and A. Friedman, *Higher order nonlinear degenerate parabolic equations*, J. Diff. Eq., 83 (1990), pp. 179–206.

[BBDK] A.L. Bertozzi, M.P. Brenner, T.F. Dupont, and L.P. Kadanoff, *Singularities and similarities in interface flow*, in Trends and Perspectives in Applied Mathematics, v. 100, L. Sirovich, ed., Springer-Verlag, New York, 1994, pp. 155–208.

[BP1] A.L. Bertozzi and M.C. Pugh, *The lubrication approximation for their viscous films: Regularity and long time behavior of weak solutions*, Comm. Pure Appl. Math., 49 (1996), pp. 85–123.

[BP2] A.L. Bertozzi and M.C. Pugh, *Long-wave instabilities and saturation in thin film equations*, Comm. Pure Appl. Math., 51 (1998), pp. 625–661.

[BP3] A.L. Bertozzi and M.C. Pugh, *Finite-time blow-up of solutions of some long-wave unstable thin film equations*, Indiana Univ. Math. J., 49 (2000), pp. 1323–1366.

[CKN] L. Caffarelli, R. Kohn, and L. Nirenberg, *Partial regularity of suitable weak solutions of the Navier-Stokes equations*, Comm. Pure Appl. Math., 35 (1982), pp. 771–831.

[CS] Y.M. Chen and M. Struwe, *Existence and partial regularity results for the heat flow for harmonic maps*, Math. Z., 201 (1989), pp. 83–103.

[CDZ] K.S. Chou, S.Z. Du, and G.F. Zheng, *On partial regularity of the borderline solution of semilinear parabolic problems*, Calc. Var. Partial Differential Equations, 30 (2007), pp. 251–275.

[CK] K.S. Chou and Y.C. Kwong, *Finite time rupture for thin films under Van Der Waals forces*, Nonlinearity, 20 (2007), pp. 299–317.

[DPGG] R. Dal Passo, L. Giacomelli, and G. Grün, *A waiting time phenomenon for thin film equations*, Ann. Sc. Norm. Super. Pisa Cl. Sci., 30 (2001), pp. 437–463.

[E] J. Eggers, *Nonlinear dynamics and breakup of free-surface flows*, Rev. Modern Phys., 69 (1997), pp. 865–929.

[Ei] S.D. Eidelman, *Parabolic Systems*, North-Holland, Amsterdam, 1969.

[ESS] L. Escauriaza, G.A. Seregin, and V. Šverák, *$L_{3,\infty}$-solutions of Navier-Stokes equations and backward uniqueness*, Uspekhi Mat. Nauk, 58 (2003), pp. 3–44 (in Russian); Russian Math. Surveys, 58 (2003), pp. 211–250 (in English).

[Fe] H. Federer, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.

[F1] A. Friedman, *Interior estimates for parabolic systems of partial differential equations*, J. Math. Mech., 7 (1958), pp. 393–418.

[F2] A. Friedman, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.

[GPS] R.E. Goldstein, A.I. Pesci, and M.J. Shelly, *Instabilities and singularities in Hele-Shaw flow*, Phy. Fluids, 10 (1998), pp. 2701–2723.

[HS] J. Hulshof and A. Shishkov, *The thin film equation with* $2 \leq n < 3$*: Finite speed of propagation in terms of the* L^1*-norm*, Adv. Differential Equations, 3 (1998), pp. 625–642.

[JL] H.Q. Jiang and F.H. Lin, *Zero set of Sobolev functions with negative power of integrability*, Chinese Ann. Math. Ser. B, 25 (2004), pp. 65–72.

[LP1] R.S. Laugesen and M.C. Pugh, *Properties of steady states for thin film equations*, European J. Appl. Math., 11 (2000), pp. 293–351.

[LP2] R.S. Laugesen and M.C. Pugh, *Linear stability of steady states for thin film and Cahn-Hilliard type equations*, Arch. Ration. Mech. Anal., 154 (2000), pp. 3–51.

[LP3] R.S. Laugesen and M.C. Pugh, *Energy levels of steady states for thin film type equations*, J. Diff. Eq., 182 (2002), pp. 377–415.

[LP4] R.S. Laugesen and M.C. Pugh, *Hetroclinic orbits, mobility parameters and stability for thin film type equations*, Electron. J. Diff. Eq., 95 (2002), p. 29.

[LiL] E.H. Lieb and M. Loss, *Analysis*, 2nd ed., American Mathematical Society, Providence, RI, 2001.

[L] F.H. Lin, *A new proof of the Caffarelli-Kohn-Nirenberg theorem*, Comm. Pure Appl. Math., 51 (1998), pp. 241–257.

[LL] F.H. Lin and C. Liu, *Partial regularity of the dynamic system modeling the flow of liquid crystals*, Discrete Contin. Dyn. Syst., 2 (1996), pp. 1–22.

[M] T.G. Myers, *Thin films with high surface tension*, SIAM Rev., 40 (1998), pp. 441–462.

[ODB] A. Oron, S.H. Davis, and S.G. Bankoff, *Long-scale evolution of thin liquid films*, Rev. Modern Phys., 69 (1997), pp. 931–980.

[SP] D. Slepčev and M.C. Pugh, *Self-similar blow-up of unstable thin film equations*, Indiana Univ. Math. J., 54 (2005), pp. 1697–1738.

[S] M. Struwe, *On the evolution of harmonic maps in higher dimensions*, J. Differential Geom., 28 (1988), pp. 485–502.

SIAM J. MATH. ANAL.
Vol. 40, No. 2, pp. 824–850

THE STABILITY OF THE NORMAL STATE OF SUPERCONDUCTORS IN THE PRESENCE OF ELECTRIC CURRENTS*

Y. ALMOG†

Abstract. The stability of the normal state of superconductors in the presence of electric currents is studied in the large domain limit. The model being used is the time-dependent Ginzburg–Landau model, in the absence of an applied magnetic field, and with the effect of the induced magnetic field being neglected. We find that if the current is nowhere perpendicular to the boundary, or if the minimal current on the boundary, at points where it is perpendicular to it, is greater than the critical current in the one-dimensional case, then the normal state is stable. We also prove some short-time instability when the current is both perpendicular to the boundary and smaller than the one-dimensional critical current.

Key words. superconductivity, Ginzburg–Landau, electric current, non–self-adjoint

AMS subject classifications. 82D55, 35P10, 35P15

DOI. 10.1137/070699755

1. Introduction.

It is well known that when a superconductor is placed at a temperature lower than the critical one, it loses its electrical resistivity. This means that current can flow through a superconducting sample with a vanishingly small voltage drop. If one raises the current above a certain critical level, superconductivity will be destroyed and the material would revert to the normal state, even if the temperature is kept fixed below the critical one.

The reverse experiment can also be considered. One can flow a strong current through the sample which would set it in the normal state. Then, if we lower the current, there is a critical current where the sample would abruptly become purely superconducting. Though the two experiments substantially differ from each other from a theoretical point of view, hysteresis was not experimentally observed in the current-voltage characteristics of the sample [8, 14, 15].

We consider here the second experiment. To this end we must analyze the stability of the normal state. The model we use in this work is the time-dependent Ginzburg–Landau model [7, 4], presented here in a dimensionless form as follows:

$$\frac{\partial \psi}{\partial t} + i\phi\psi = (\nabla - iA)^2 \psi + \psi\left(1 - |\psi|^2\right) \quad \text{in } \Omega, \tag{1.1a}$$

$$-\kappa^2 \nabla \times (\nabla \times A) - \sigma\left(\frac{\partial A}{\partial t} + \nabla\phi\right) = \frac{i}{2}(\bar{\psi}\nabla\psi - \psi\nabla\bar{\psi}) + |\psi|^2 A \quad \text{in } \Omega, \tag{1.1b}$$

$$\psi = 0 \quad \text{on } \partial\Omega_c, \tag{1.1c}$$

$$(i\nabla + A)\psi \cdot \nu = 0 \quad \text{on } \partial\Omega_i\,. \tag{1.1d}$$

*Received by the editors August 9, 2007; accepted for publication (in revised form) April 8, 2008; published electronically August 20, 2008. The author was supported by NSF grant DMS 0604467 and by the Summer Stipend Program at LSU.
http://www.siam.org/journals/sima/40-2/69975.html
†Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803 (almog@math.lsu.edu).

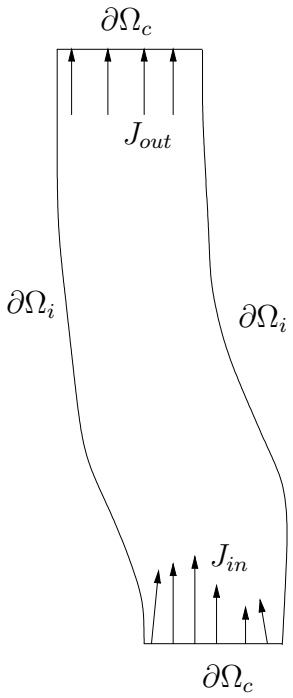


FIG. 1. *Typical superconducting sample. The arrows denote the direction of the current flow (J_{in} for the inlet and J_{out} for the outlet).*

In (1.1) ψ is the superconducting order parameter, so that $|\psi|$ represents the number density of superconducting electrons. Superconductors for which $|\psi| = 1$ are said to be wholly superconducting, and those for which $\psi = 0$ are said to be at the normal state. A is the magnetic vector potential and ϕ is the electric scalar potential. The constant σ is a measure of the normal conductivity of the superconducting material so that $-\sigma\nabla\phi$ is the normal current, and κ is the Ginzburg–Landau parameter. Length has been scaled with the coherence length ξ, which is the natural length-scale for variations in ψ. The domain $\Omega \subset\subset \mathbb{R}^n$ $(n = 1, 2, 3)$, where the superconducting sample resides, is smooth and has an interface, denoted by $\partial\Omega_c$, with a conducting metal which is at the normal state. The rest of the boundary, denoted by $\partial\Omega_i$, is adjacent to an insulator. We allow nonsmoothness of $\partial\Omega$ in the sense that $\partial\Omega_c$ and $\partial\Omega_i$ are required to be perpendicular to each other in order to include cylindrical-like domains. Figure 1 presents a typical two-dimensional sample, where the current flows into the sample from one part of $\partial\Omega_c$ and exits from another part, disconnected from the first one. Most wires would fall into the above class of domains.

Equations (1.1) are gauge invariant in the sense that they are invariant under transformations of the form

$$A \to A + \nabla\omega\,, \quad \phi \to \phi - \frac{\partial\omega}{\partial t}\,, \quad \psi \to \psi e^{i\kappa\omega}\,.$$

Note that none of the important physical properties, i.e., $|\psi|$, the magnetic field $H = \nabla \times A$, and the electric field $E = -\partial A/\partial t - \nabla\phi$, are altered by the above transformation.

To obtain a well-posed problem one must add to (1.1) initial conditions, and the equations satisfied by A and ϕ outside Ω, that is, the Maxwell equations. Continuity

of the tangential components of A and $\nabla \times A$ through $\partial\Omega$ and some conditions on $\nabla \times A$ at infinity should be required as well. Since these details are irrelevant in the context of the present contribution, we omit them. Interested readers may be able to find them in [4, 6].

We consider here the stability of the normal state. If $\psi \equiv 0$, we obtain that the steady state solution must satisfy

$$\nabla \times H = -\frac{\sigma}{\kappa^2}\nabla\phi\,,$$

and hence ϕ is harmonic in Ω. To obtain ϕ we thus need to solve the following problem:

$$\begin{cases} \Delta\phi = 0 & \text{in } \Omega, \\ \phi = \phi_0(x) & \text{on } \partial\Omega_c, \\ \frac{\partial\phi}{\partial\nu} = 0 & \text{on } \partial\Omega_i. \end{cases} \tag{1.2}$$

We note that instead of prescribing the potential on $\partial\Omega_c$ we can prescribe the current in the normal direction to the boundary $J_n(x) = -\sigma\partial\phi/\partial\nu$. For simplicity we assume that $\nabla\phi \neq 0$ everywhere in Ω. This can easily be achieved; for instance, for the samples described in Figure 1, if $\inf\phi$ on one connected component of $\partial\Omega_c$ is greater than the $\sup\phi$ on the other connected component, then $\nabla\phi$ never vanishes.

Once (1.2) is solved, one can solve for the magnetic field. Here we need to solve a problem in $\mathbb{R}^n$,

$$\begin{cases} \nabla \times H = -\frac{\sigma}{\kappa^2}\nabla\phi = \frac{1}{\kappa^2}J & \text{in } \Omega, \\ \nabla \times H = 0 & \text{in } \mathbb{R}^n \setminus \Omega, \\ H \to h_{ex} & \text{as } |x| \to \infty, \end{cases} \tag{1.3}$$

with H continuous across $\partial\Omega$. We assume zero applied magnetic field ($h_{ex} = 0$). To simplify our problem further we shall assume $A = H = 0$. This assumption can be justified in the case where $\kappa \gg 1$ in view of (1.3). However, since we intend to consider large domains, one must assume that $\kappa \gg \operatorname{diam}\Omega$. In real-world coordinates this means that our domain size must be much larger than the coherence length ξ but also much smaller than the penetration depth λ, which is the length-scale characterizing variations in H ($\kappa = \lambda/\xi$). While this assumption significantly limits the validity of our results, it has been made very often by physicists [15, 8] and is reasonable to adopt as a starting point.

Once the above assumption is adopted one obtains

$$\begin{cases} \frac{\partial\psi}{\partial t} - \Delta\psi + i\phi\psi - \psi(1 - |\psi|^2) = 0 & \text{in } \Omega, \\ \psi = 0 & \text{on } \partial\Omega_c, \\ \frac{\partial\psi}{\partial\nu} = 0 & \text{on } \partial\Omega_i, \\ \psi(x, 0) = \psi_0(x) & \text{in } \Omega. \end{cases}$$

It should be noted that in [6] Du and Gray prove, within the framework of a more general case, convergence in the limit $\kappa \to \infty$ of (1.1) to a different limit problem where the magnetic field is not negligible. The domain size considered there is, however, much larger than in our case, as it is comparable with the penetration depth.

Linearizing the above near the normal state, we obtain

$$\begin{cases} \frac{\partial \psi}{\partial t} - \Delta \psi + i\phi\psi - \psi = 0 & \text{in } \Omega, \\ \psi = 0 & \text{on } \partial\Omega_c, \\ \frac{\partial \psi}{\partial \nu} = 0 & \text{on } \partial\Omega_i, \\ \psi(x,0) = \psi_0(x) & \text{in } \Omega. \end{cases} \tag{1.4}$$

The above problem has been analyzed by physicists in one-dimensional settings. Ivlev and Kopnin review these results in [8]. In these settings we have $\phi = Jx + \mu$, where J (denoting the current) and μ are constants. Previous results include the closed form solution of (1.4) in $\mathbb{R}$ for any initial condition. In $\mathbb{R}_+$, the first critical current J_c for which a steady state solution ($\partial|\psi|/\partial t = 0$) exists is found. Then weakly nonlinear analysis is performed, where it is shown that [9] the bifurcation taking place near $J = J_c$ is subcritical (i.e., unstable).

We consider (1.4) in three-dimensional settings. While all the results are stated for three-dimensional objects, they are equally valid for two-dimensional objects as well. We deal with (1.4) in the large domain limit; i.e., we consider a domain Ω_R which is obtained from Ω via the transformation

$$x \to Rx\,. \tag{1.5}$$

The portions of the boundaries $\partial\Omega_R^c$ and $\partial\Omega_R^i$ are similarly obtained from $\partial\Omega_c$ and $\partial\Omega_i$, respectively. To keep $\nabla\phi$ unaltered we consider also potentials of the form

$$\phi_R = R\phi(x/R)\,.$$

Thus, we consider the following problem:

$$\begin{cases} \frac{\partial \psi_R}{\partial t} - \Delta \psi_R + i\phi_R\psi_R - \psi_R = 0 & \text{in } \Omega_R, \\ \psi_R = 0 & \text{on } \partial\Omega_R^c, \\ \frac{\partial \psi_R}{\partial \nu} = 0 & \text{on } \partial\Omega_R^i, \\ \psi_R(x,0) = \psi_0(x) & \text{in } \Omega_R. \end{cases} \tag{1.6}$$

Our main result is the following.

THEOREM 1.1. *Let ϕ satisfy* (1.2). *Let $\partial\Omega_c^n$ denote the portion of $\partial\Omega_c$ on which $\nabla\phi$ is perpendicular to the boundary. Let further*

$$J_m = \min_{x\in\partial\Omega_c^n} \left|\frac{\partial\phi}{\partial\nu}\right|,$$

and let J_c denote the critical current for the problem in $\mathbb{R}_+$ (which is precisely defined in (2.19)*). Suppose further that $|J| > 0$ everywhere on $\partial\Omega_c$. Then, if $J_m > J_c$, or if $\partial\Omega_c^n$ is empty, there exists $R_0 > 0$ such that $\psi_R \equiv 0$ is a stable solution of* (1.6) *in the sense of $L^2(\Omega_R)$ for all $R > R_0$.*

Furthermore, if $J_m < J_c$, there exists $\psi_0 \in L^2(\Omega_R)$ and $T_R > 0$ such that

$$\liminf_{R\to\infty} \frac{T_R}{\ln R} > 0\,,$$

and such that the solution ψ_R of (1.6) *has the following property:*

$$t < T_R \Rightarrow \|\psi_R\|_{L^2(\Omega_R)} > \frac{1}{2}\|\psi_0\|_{L^2(\Omega_R)} e^{\beta t}\,, \tag{1.7}$$

where

$$\beta = 1 - (J/J_c)^{2/3}\,.$$

In the next section we review and enhance the results in [8] in the one-dimensional case. In section 3 we extend some of the results in section 2 to unbounded three-dimensional domains. In section 4 we provide the proof of the theorem. Finally, in the last section we highlight some possible directions for future research. The appendix provides a technical result needed in section 2.

2. One-dimensional problems. In this section we consider (1.4) in two different one-dimensional settings: on $\mathbb{R}$ and on $\mathbb{R}_+$. The solution of these simple problems would provide us with some important intuition on the solution of (1.6) in three-dimensional bounded domains in the large domain limit. In both cases we shall assume $\phi = Jx$, i.e., that the current is uniform and equals J throughout the sample.

2.1. Infinite one-dimensional domain. Here we consider the problem

$$\frac{\partial \psi}{\partial t} - \psi'' - \psi + iJx\psi = 0\,. \tag{2.1}$$

We consider here only the case $J > 0$. Otherwise, if $J < 0$, we can consider the complex conjugate of (2.1). Applying the coordinate transformation

$$x \to J^{1/3}x; \qquad t \to J^{2/3}t, \tag{2.2}$$

we obtain the problem

$$\frac{\partial \psi}{\partial t} + \mathcal{L}\psi = \lambda_J \psi\,,$$

where

$$\mathcal{L}\psi = -\psi'' + ix\psi\,, \tag{2.3}$$

and $\lambda_J = J^{-2/3}$.

We first focus our interest on the spectrum of the operator $\mathcal{L} : \mathcal{D}_{\mathbb{R}}(\mathcal{L}) \to L^2(\mathbb{R}, \mathbb{C})$, where $\mathcal{D}_{\mathbb{R}}(\mathcal{L})$ is the dense subset of $L^2(\mathbb{R}, \mathbb{C})$ defined as

$$\mathcal{D}_{\mathbb{R}}(\mathcal{L}) = \{u \in L^2(\mathbb{R}, \mathbb{C}) \,|\, -u'' + ixu \in L^2(\mathbb{R}, \mathbb{C})\}\,.$$

LEMMA 2.1. *The operator $\mathcal{L} - \lambda I$ is invertible for all $\lambda \in \mathbb{C}$. (Thus $\sigma(\mathcal{L}) = \phi$.)*

Proof. It is sufficient to consider here $\lambda \in \mathbb{R}$. Otherwise, if $\lambda = \lambda_r + i\lambda_i$, we apply the transformation $x \to x - \lambda_i$.

Though it is not necessary, we first prove injectivity of $\mathcal{L} - \lambda I$. We shall later make use of a similar argument in three dimensions. To the problem

$$\mathcal{L}u = \lambda u$$

we apply the Fourier transform

$$\hat{u}(\omega) = \mathcal{F}(u) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-i\omega x} u(x) dx \tag{2.4}$$

to obtain

$$\omega^2 \hat{u} - \frac{\partial \hat{u}}{\partial \omega} = \lambda \hat{u}\,.$$

Since the above equation doesn't possess any nontrivial solutions with bounded $L^2(\mathbb{R})$ norm, $\mathcal{L} - \lambda I$ must be injective.

To find the spectrum of $\mathcal{L}$ we construct the Green's function of $\mathcal{L} - \lambda I$. Let

$$\begin{cases} w_1(x,\lambda) = A_i\big(e^{i\pi/6}x + \lambda e^{i2\pi/3}\big), \\ w_2(x,\lambda) = A_i\big(-e^{-i\pi/6}x + \lambda e^{-i2\pi/3}\big), \end{cases} \tag{2.5}$$

where A_i denotes Airy's function [1] (cf. (A.2) in Appendix A for the asymptotic behavior of A_i). It is easy to show that u_1, u_2 constitute a fundamental set of solutions to $(\mathcal{L} - \lambda I)u = 0$ [13]. We can now write the Green's function in the form

$$G(x,\xi,\lambda) = \begin{cases} \frac{w_2(\xi,\lambda)}{W(w_1,w_2)} w_1(x,\lambda), & x > \xi, \\ \frac{w_1(\xi,\lambda)}{W(w_1,w_2)} w_2(x,\lambda), & x < \xi, \end{cases}$$

where $W(w_1, w_2)$ denote the Wronskian (which is a constant that clearly doesn't vanish since w_1 and w_2 are linearly independent).

Using the asymptotic behavior of Airy's functions [1], we can show, using the same procedure applied in Appendix A to the semi-infinite case, that $G(\cdot,\lambda) \in L^2(\mathbb{R}^2,\mathbb{C})$ for all $\lambda \in \mathbb{R}$. The lemma is proved. □

The spectrum of the operator $\mathcal{L} : \mathcal{D}_{\mathbb{R}}(\mathcal{L}) \to L^2(\mathbb{R},\mathbb{C})$ can therefore teach us very little on the stability of the normal state in this case. The following lemma proves its stability in a direct manner.

LEMMA 2.2. *The trivial solution of* (2.1), $\psi \equiv 0$, *is globally stable in* L^2*; i.e., if* $f(x) \in L^2(\mathbb{R},\mathbb{C})$, *then the solution of* (2.1), $\psi(x,t)$, *satisfying the initial condition* $\psi(x,0) = f(x)$ *satisfies* $\psi(x,t) \xrightarrow[t\to\infty]{} 0$ *in* $L^2(\mathbb{R},\mathbb{C})$.

Proof. As long as $\psi(x,t) \in L^2(\mathbb{R},\mathbb{C})$, we can apply to (2.1) the Fourier transform (2.4). We obtain

$$\frac{\partial \hat{\psi}}{\partial t} + \omega^2 \hat{\psi} - \frac{\partial \hat{\psi}}{\partial \omega} - \lambda_J \hat{\psi} = 0\,.$$

The unique solution to the above problem is given by

$$\hat{\psi}(\omega,t) = \hat{f}(\omega) \exp\left\{ -\omega^2 t - \omega t^2 - \frac{1}{3}t^3 + \lambda_J t \right\}, \tag{2.6}$$

in which $\hat{f}(\omega)$ denotes the Fourier transform of f. Integrating the modulus square of the above over $\mathbb{R}$ with respect to ω, we obtain

$$\|\psi(.,t)\|_2 \le C\|f\|_2 \exp\left\{ -\frac{1}{12}t^3 + \lambda_J t \right\},$$

where $\|\cdot\|_2$ denotes the $L^2(\mathbb{R},\mathbb{C})$ norm. □

The above superlinear convergence is in accordance with the result proved in the previous lemma by which the spectrum of $\mathcal{L}$ is empty. We note that in [8] the inverse transform of (2.6) is obtained, but no decay proof is given.

2.2. Semi-infinite one-dimensional domain. We now consider (2.1) on $\mathbb{R}_+$. We concentrate here on the Dirichlet boundary condition $\psi(0)=0$; however, the same analysis applies to Neumann and mixed boundary conditions as well.

We start by proving the following result on the spectrum of the operator $\mathcal{L}:\mathcal{D}_{\mathbb{R}_+}(\mathcal{L})\to L^2(\mathbb{R},\mathbb{C})$, where

$$\mathcal{D}_{\mathbb{R}_+}(\mathcal{L})=\{u\in L^2(\mathbb{R}_+,\mathbb{C})\,|\,-u''+ixu\in L^2(\mathbb{R}_+,\mathbb{C}),\ u\in H_0^1(\mathbb{R}_+,\mathbb{C})\}\,.$$

LEMMA 2.3.

1. *There exists a sequence of eigenvalues* $\{\lambda_n\}_{n=1}^{\infty}$ *and eigenfunctions of* $\mathcal{L}$*, with unity norm,* $\{u_n\}_{n=1}^{\infty}\subset\mathcal{D}_{\mathbb{R}_+}(\mathcal{L})$*, i.e.,*
$$\mathcal{L}u_n=\lambda_n u_n\,.$$
2. *We have*
$$m=\max_{n\in\mathbb{N}}\Re\lambda_n>0. \tag{2.7}$$
3. $\operatorname{span}\{u_n\}_{n=1}^{\infty}=L^2(\mathbb{R}_+,\mathbb{C})$.
4. *Suppose that* $u,v\in L^2(\mathbb{R}_+,\mathbb{C})$ *can be represented in the form*
$$v=\sum_{n=1}^{\infty}\alpha_n u_n;\qquad w=\sum_{n=1}^{\infty}\beta_n u_n,$$
where the convergence is in the L^2 *sense. Let then*
$$\langle v,w\rangle_U=\sum_{n=1}^{\infty}\alpha_n\bar{\beta}_n\,. \tag{2.8}$$
Then,
$$\inf_{v\in\mathcal{D}_{\mathbb{R}_+}(\mathcal{L})}\Re\langle v,\mathcal{L}v\rangle_U\geq m\|v\|_U^2\,, \tag{2.9}$$
where $\|\cdot\|_U$ *denotes the norm induced by* (2.8).

Proof. Let
$$z=-ix+\lambda.$$
Let $u(x,\lambda)\in\mathcal{D}_{\mathbb{R}_+}(\mathcal{L})$ denote an eigenfunction of $\mathcal{L}$, i.e., $\mathcal{L}u=\lambda u$. Let $v(z,\lambda)=u(x,\lambda)$. We have
$$\begin{cases}\frac{\partial^2 v}{\partial z^2}-zv=0, & z\in\mathbb{C},\\ v(\lambda,\lambda)=0.\end{cases} \tag{2.10}$$
Since $u\in L^2(\mathbb{R}_+,\mathbb{C})$, v must be subdominant (i.e., it decays exponentially fast [13]) in the sector
$$S_1:\ -\pi<\arg z<-\frac{\pi}{3}\,.$$
The decaying solution of (2.10) in S_1 is given by (cf. [13])
$$v=A_i\big(e^{2\pi i/3}z\big)\,.$$

Since the zeros of Airy's functions are eigenvalues of the self-adjoint operator $d^2/dx^2 - x$ in $\mathcal{D}_{\mathbb{R}_+}(\mathcal{L})$, they must all be real. Let $\{\mu_n\}_{n=1}^{\infty} \subset \mathbb{R}$ denote the zeroes of Airy's function on the real axis. By the maximum principle they must all be strictly negative. We arrange them so that $\mu_n \downarrow -\infty$. As every eigenvalue of $\mathcal{L}$ must satisfy $v(\lambda, \lambda) = 0$, the set $\{\lambda_n\}_{n=1}^{\infty}$, where

$$\lambda_n = e^{-i2\pi/3}\mu_n, \tag{2.11}$$

contains all the eigenvalues of $\mathcal{L}$. Since $\mu_1 < 0$ we have that

$$m = \Re\lambda_1 = -\frac{\mu_1}{2} > 0\,.$$

The set $\{\tilde{u}_n\}_{n=1}^{\infty}$ of eigenfunctions of $\mathcal{L}$ in $\mathcal{D}_{\mathbb{R}_+}(\mathcal{L})$ is given by

$$\tilde{u}_n = A_i\big(e^{i2\pi/3}(-ix + \lambda_n)\big) = A_i\big(e^{i\pi/6}x + \mu_n\big) \quad \forall n \in \mathbb{N}\,. \tag{2.12}$$

We then set

$$u_n = \frac{\tilde{u}_n}{\|\tilde{u}_n\|_{L^2(\mathbb{R}_+)}}\,. \tag{2.13}$$

To prove that $\{u_n\}_{n=1}^{\infty}$ is complete in $L^2(\mathbb{R}, \mathbb{C})$ we consider the resolvent $\mathcal{L}_\lambda^{-1} = (\mathcal{L} - \lambda I)^{-1}$ (which is also the modified resolvent of $\mathcal{L}^{-1}$). We have

$$\mathcal{L}_\lambda^{-1} f = \int_0^\infty \tilde{G}(x, \xi, \lambda) f(\xi) d\xi\,, \tag{2.14a}$$

in which

$$\tilde{G}(x, \xi, \lambda) = \begin{cases} \frac{\tilde{w}_2(\xi,\lambda)}{W(w_1, \tilde{w}_2)} w_1(x, \lambda), & x > \xi, \\ \frac{w_1(\xi,\lambda)}{W(w_1, \tilde{w}_2)} \tilde{w}_2(x, \lambda), & x < \xi, \end{cases} \tag{2.14b}$$

where

$$\tilde{w}_2(x, \lambda) = \frac{w_1(0, \lambda)}{w_2(0, \lambda)} w_1(x, \lambda) - w_2(x, \lambda)\,, \tag{2.15}$$

and w_1 and w_2 are given in (2.5). In Appendix A, we prove that $\tilde{G} \in L^2(\mathbb{R}_+ \times \mathbb{R}_+)$, and that

$$\|\tilde{G}(., ., \lambda)\|_{L^2(\mathbb{R}_+ \times \mathbb{R}_+)} \leq e^{M|\lambda|^{3/2}}\,, \tag{2.16}$$

as long as $\lambda \notin \{\lambda_n\}_{n=1}^{\infty}$.

Let $u \in \mathcal{D}_{\mathbb{R}_+}(\mathcal{L})$. We now multiply $\mathcal{L}u$ by $e^{i\theta}\bar{u}$ and integrate over $\mathbb{R}_+$ to obtain

$$\Re < e^{i\theta}\mathcal{L}u, u >= \int_0^\infty (\cos\theta |u'|^2 - \sin\theta x |u|^2) dx\,.$$

By Theorem 12.8 in [2] we have that every direction $e^{i \arg\lambda}$ with

$$\pi/2 < \arg\lambda < 3\pi/2$$

is a direction of minimal growth of the resolvent of $e^{i\theta}\mathcal{L}$ for every $-\pi/2 < \theta < 0$. Consequently, every direction $e^{i \arg \lambda}$ with

$$\pi/2 < \arg \lambda < 2\pi$$

is a direction of minimal growth of $\mathcal{L}_\lambda^{-1}$, i.e. (cf. [2]),

$$\|\mathcal{L}_\lambda^{-1}\| \sim \mathcal{O}(|\lambda|^{-1}) \quad \pi/2 < \arg \lambda < 2\pi\,. \tag{2.17}$$

We now apply the same argument used in the proof of Theorem 16.4 in [2]. Let $f \in L^2(\mathbb{R}_+)$, and let $g \in V^\perp$, where $V = \operatorname{span}\{A_i(e^{i\pi/6}x + \mu_n)\}_{n=1}^\infty = \operatorname{sp}'(\mathcal{L}^{-1})$. Then,

$$F(\lambda) = \langle \mathcal{L}_\lambda^{-1} f, g\rangle$$

is an entire function (cf. [2]) of λ satisfying, by (2.16) and (2.17),

$$\begin{cases} |F(\lambda)| \le \frac{C}{|\lambda|}, & \pi/2 < \arg \lambda < 2\pi, \\ |F(\lambda)| \le Ce^{M|\lambda|^{3/2}}, & \lambda \in \mathbb{C}. \end{cases}$$

By Theorem 16.1 in [2] (or the Phragmen–Leindelöf theorem) and Liouville's theorem, we must have $F(\lambda) \equiv 0$. Hence, $\overline{V} = \overline{\operatorname{range}(\mathcal{L}^{-1})} = \overline{\mathcal{D}_{\mathbb{R}_+}(\mathcal{L})}$.

It remains still necessary to prove (2.9). Let $w \in \mathcal{D}_{\mathbb{R}_+}(\mathcal{L})$. Then,

$$w = \sum_{n=1}^\infty \alpha_n u_n$$

and

$$\mathcal{L}w = \sum_{n=1}^\infty \alpha_n \lambda_n u_n\,.$$

Hence

$$\Re \langle w, \mathcal{L}w\rangle_U = \sum_{n=1}^\infty \Re\lambda_n |\alpha_n|^2 \ge m\|w\|_U^2\,. \qquad \square$$

Similar techniques were used in [12, 10] to prove completeness of the system of eigenfunctions of some nonlinear eigenvalue problems in $\mathbb{R}$. We note that the set $\{u_n\}_{n=1}^\infty$ is not a basis in the usual sense in Banach spaces. In fact, it has been demonstrated in [5] that the system $\{\tilde{u}_n, \overline{\tilde{u}_n}\}_{n=1}^\infty$, which is a biorthogonal system after we appropriately normalize it, is wild. This means that $\|\tilde{u}_n\|_2$ grows faster than any algebraic rate as $n \to \infty$.

We now prove the existence of a critical current J_c obtained in [8, 9].

LEMMA 2.4. *Let $\psi(x,t) \in H_0^2(\mathbb{R}_+ \times \mathbb{R}_+, \mathbb{C})$ denote a solution of the equation*

$$\frac{\partial \psi}{\partial t} - \frac{\partial^2 \psi}{\partial x^2} - \psi + iJx\psi = 0 \quad \text{in } \mathbb{R}_+ \times \mathbb{R}_+\,. \tag{2.18}$$

If

$$J > J_c = \left(-\frac{\mu_1}{2}\right)^{-3/2}, \tag{2.19}$$

in which μ_1 is the rightmost zero of Airy's function, then $\|\psi(\cdot,t)\|_U \xrightarrow[t\to\infty]{} 0$. Otherwise, if $J < J_c$, then $\psi \equiv 0$ is an unstable solution of (2.18).

Proof. We first apply (2.2) to obtain

$$\begin{cases} \frac{\partial \psi}{\partial t} + \mathcal{L}\psi - \lambda_J \psi = 0, \\ \psi(0,t) = 0, \\ \psi(x,0) = \psi_0(x), \end{cases}$$

where $\|\psi_0\|_U < \infty$. Taking the inner product (2.8) of the above equation with ψ, we obtain

$$\frac{1}{2}\frac{d}{dt}\|\psi\|_U^2 \le \left(\lambda_J + \frac{\mu_1}{2}\right)\|u\|_U^2 .$$

Hence, if (2.19) is satisfied, we have $\lambda_J + \mu_1/2 < 0$, and hence ψ tends to 0 exponentially fast as $t \to \infty$, in the $\|\cdot\|_U$ sense.

If $J < J_c$, let $\psi(x,0) = u_1(x)$, where u_1 is given in (2.13). Then

$$\psi(x,t) = u_1(x) e^{(\lambda_J + \mu_1/2)t} .$$

Returning to the original variables by applying the inverse of (2.2), we obtain

$$\psi(x,t) = u_1(J^{1/3}x) e^{[1-(J/J_c)^{2/3}]t} ,$$

and hence $\psi \equiv 0$ is unstable. ∎

3. Unbounded domains in $\mathbb{R}^3$. In this section we consider several different problems: in $\mathbb{R}^3$, in $\mathbb{R}^3_+$, and in a quarter-space. In contrast with the previous section, we consider these problems only to the extent needed in the next section, that is, we analyze the existence of eigenvalues with nonpositive real part of the elliptic operator in the right-hand side of (1.4).

3.1. Eigenfunctions in $\mathbb{R}^3$. We consider here (1.4) with $\phi = Jx_1$. This choice summarizes all possible electric potentials with constant gradients, as the problem is invariant to translations and rotations. Thus, we have for every eigenfunction u,

$$-\Delta u - u + iJx_1 u = -\lambda u . \tag{3.1}$$

We shall assume here that $\lambda \in \mathbb{R}$; otherwise we can apply the transformation $x_1 \to x_1 - \Im\lambda/J$.

It is easy to find all L^2 solutions of (3.1) in $\mathbb{R}^3$ as follows: apply the Fourier transform (2.4) in the x_2 and x_3 directions (using the respective Fourier coordinates ω_2 and ω_3) to obtain

$$\mathcal{L}\hat{u} = \big((1-\lambda)J^{-2/3} - \omega_2^2 - \omega_3^2\big)\hat{u}, \tag{3.2}$$

where assuming $J > 0$ we have applied the transformation

$$x \to J^{1/3}x .$$

(We confine the discussion in what follows to the case $J > 0$. If $J < 0$, we can consider the complex conjugate of (3.2) to obtain a new problem with $J > 0$.)

By Lemma 2.1 we have that $\hat{u} \equiv 0$ is the unique L^2 solution of (3.2). However, for the blow-up arguments employed in the next section we need to obtain the above result for any uniformly bounded solution of (3.1) in $\mathbb{R}^3$. This is exactly what the next lemma states.

LEMMA 3.1. *Let u denote a uniformly bounded solution of* (3.1) *in* $\mathbb{R}^3$. *Then,* $u \equiv 0$.

Proof. We first show that $u(., x_2, x_3) \in L^2(\mathbb{R}, \mathbb{C})$. Let $\chi_r \in C^\infty(\mathbb{R}_+, [0,1])$ satisfy

$$\chi_r(x) = \begin{cases} 1 & x < r/2, \\ 0 & x > r, \end{cases} \qquad |\chi'| \leq \frac{C}{r}\,. \tag{3.3}$$

Multiplying (3.1) by $\chi_r^2(|x - x_0|)\bar{u}$ we obtain, taking the real part of identity, that

$$\int_{B(x_0,r)} |\nabla(\chi_r u)|^2 \leq \int_{B(x_0,r)} \left[\chi^2 + |\nabla\chi|^2\right]|u|^2\,. \tag{3.4}$$

Consequently, since u is bounded in $L^\infty(\mathbb{R}^3)$, we have

$$\int_{B(x_0,r/2)} |\nabla u|^2 \leq C \quad \forall x_0 \in \mathbb{R}^3\,. \tag{3.5}$$

From the imaginary part of the identity, we obtain that

$$\int_{B(x_0,r)} \left(\nabla(\chi_r^2) \cdot \Im\left(\bar{u}\nabla u\right) + J x_1 \chi_r |u|^2\right) = 0\,.$$

Let $x_0 = (x_0^1, x_0^2, x_0^3)$. Then,

$$\int_{B(x_0,r/2)} |x_1||u|^2 dx_1 \leq C \int_{B(x_0,r)} \left[|u|^2 + |\nabla u|^2\right]\,.$$

Consequently, for $|x_0^1| > r$, since u is bounded and in view of (3.5), we have

$$\int_{B(x_0,r/2)} |u|^2 \leq \frac{C}{|x_0^1| - \frac{r}{2}} \leq \frac{C}{|x_0^1|}\,.$$

Repeating the above steps (from (3.4) to the above inequality) k times we obtain that

$$\int_{B(x_0,r/2^k)} |u|^2 \leq \frac{C_k}{|x_0^1|^k}.$$

Hence,

$$\int_{B(x_0,r/2^k)} |x_1|^2|u|^2 \leq \frac{C_k}{|x_0^1|^{k-2}}\,,$$

which allows us to apply standard elliptic estimates [3] to obtain that

$$|u| \leq \frac{C_k}{(|x_1| + 1)^k} \quad \forall k \in \mathbb{N}\,.$$

In view of the above we have that $x_1^k u(x_1, x') \in L^2(\mathbb{R}, \mathbb{C})$ for all fixed $x' \in \mathbb{R}^2$. Thus, one can apply to (3.1) the Fourier transform (2.4) to obtain

$$-\Delta_\perp \hat{u} + (\omega^2 - 1 + \lambda)\hat{u} - J\frac{\partial \hat{u}}{\partial \omega} = 0\,, \tag{3.6}$$

where

$$-\Delta_\perp = \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2}\,.$$

Let

$$\tilde{x}_r(x) = \begin{cases} 1, & |x| < r, \\ e^{-\frac{1}{r}(|x|-r)}, & |x| > r. \end{cases} \tag{3.7}$$

Multiplying (3.6) by $\tilde{\chi}_r^2(|x' - x^0|)\overline{\hat{u}}$ and integrating over $\mathbb{R}^2$, we obtain

$$-J\frac{dU_r}{d\omega} + (\omega^2 - 1 + \lambda)U_r = \int_{\mathbb{R}^2} \left(-|\nabla(\tilde{\chi}_r\hat{u})|^2 + |\nabla\tilde{\chi}_r|^2|\hat{u}|^2\right) dx',$$

where

$$U_r(\omega) = \int_{\mathbb{R}^2} \tilde{\chi}_r^2|\hat{u}|^2 dx'\,.$$

For the last term, we have

$$\int_{\mathbb{R}^2} |\nabla\tilde{\chi}_r|^2|\hat{u}|^2 \le \frac{1}{r^2}U_r\,.$$

Consequently,

$$-J\frac{dU_r}{d\omega} + (\omega^2 - 1 + \lambda - r^{-2})U_r \le 0\,,$$

and therefore, for every $\omega_0 \in \mathbb{R}$ and $\omega > \omega_0$, we have

$$U_r(\omega) \ge U_r(\omega_0)\exp\left\{\frac{1}{3}(\omega^3 - \omega_0^3) - (1 + \lambda + r^{-2})(\omega - \omega_0)\right\}.$$

Thus, since U_r is positive, it must diverge exponentially fast, unless

$$U_r \equiv 0,$$

from which the lemma easily follows. □

3.1.1. Eigenfunctions in $\mathbb{R}^3_+$: Perpendicular current. Let

$$\mathbb{R}^3_+ = \{(x_1, x_2, x_3)\,|\,x_1 > 0\}\,.$$

We consider here solutions of (3.1) in $\mathbb{R}^3_+$ satisfying a Dirichlet boundary condition on $\partial\mathbb{R}^3_+$. Instead of considering complex eigenvalues, we consider only real ones and treat their imaginary part as part of the electric potential.

LEMMA 3.2. *Let $u \in H^2(\mathbb{R}^3_+)$ denote a uniformly bounded solution of*

$$\begin{cases} -\Delta u - u + iJ(x_1 - \mu)u = -\lambda u & \text{in } \mathbb{R}^3_+, \\ u = 0 & \text{on } \partial\mathbb{R}^3_+\,, \end{cases} \tag{3.8}$$

with $\lambda \in \mathbb{R}_+$. Then, if $J > J_c$, where J_c is defined as in (2.19)*, u must vanish identically.*

Proof. Let $u_n(x_1)$ be defined as in (2.12). Let further

$$a_n(x_2,x_3)=\int_0^\infty u_n(x_1)u(x_1,x_2,x_3)dx_1=\langle u_n,\bar{u}\rangle\,,$$

where the inner product is in the regular L^2 sense. Clearly, a_n is uniformly bounded in $\mathbb{R}^2$ as $u\in L^\infty(\mathbb{R}^3,\mathbb{C})$ and $u_n\in L^1(\mathbb{R},\mathbb{C})$.

Applying the transformation

$$x\to J^{1/3}x\,, \tag{3.9}$$

multiplying (3.8) by u_n, and integrating over $\mathbb{R}_+$ with respect to x_1 we obtain, in view of the boundedness of u and the exponential rate of decay of u_n as $x_1\to\infty$, that

$$-\Delta_\perp a_n+(\lambda_n-\tilde{\lambda}_J-i\mu)a_n=0\,, \tag{3.10}$$

where $\tilde{\lambda}_J=(1-\lambda)J^{-2/3}$ and the definition of λ_n is given as in Lemma 2.3. Multiplying (3.10) by $\tilde{\chi}_r(|x'|)$, we obtain for the real part

$$\left(\frac{|\mu_n|}{2}-\tilde{\lambda}_J\right)\int_{\mathbb{R}^2}|\tilde{\chi}_r|^2|a_n|^2=-\int_{\mathbb{R}^2}|\nabla(\tilde{\chi}_r a_n)|^2+\int_{\mathbb{R}^2}|\nabla\tilde{\chi}_r|^2|a_n|^2\le\frac{C}{r^2}\int_{\mathbb{R}^2}|\tilde{\chi}_r|^2|a_n|^2\,.$$

Since $\tilde{\lambda}_J<|\mu_1|/2\le|\mu_n|/2$ by our assumption we obtain that for sufficiently large r we must have

$$\int_{\mathbb{R}^2}|\tilde{\chi}_r|^2|a_n|^2=0\,.$$

Hence, $a_n\equiv 0$ in $\mathbb{R}^2$. Since by Lemma 2.3 $\{u_n\}_{n=1}^\infty$ is a basis for $L^2(\mathbb{R}_+,\mathbb{C})$, we must have $u\equiv 0$. ∎

3.1.2. Steady solutions in $\mathbb{R}^3_+$: Nonperpendicular current. This problem is very similar to the problem in $\mathbb{R}^3$. Consider the equation

$$\begin{cases}-\Delta u-u+i(J_1x_1+J_2x_2-\mu)u=-\lambda u & \text{in } \mathbb{R}^3_+,\\ u=0 & \text{on } \partial\mathbb{R}^3_+,\end{cases} \tag{3.11}$$

with $J_2\neq 0$ and $\lambda\in\mathbb{R}_+$. Like the problem in $\mathbb{R}^3$, there is no need to consider $\mu\neq 0$ here since the transformation

$$x_2\to x_2+\frac{\mu}{J_2}$$

sets $\mu=0$ in the transformed problem. Furthermore, we also obtain the following result, which is exactly the same as the result obtained in $\mathbb{R}^3$.

LEMMA 3.3. *Let u denote a bounded solution of* (3.11) *with $J_2\neq 0$. Then $u\equiv 0$.*

Proof. Consider first the case where $J_1\neq 0$. We first apply the transformation (3.9) with $J=J_1$ to obtain

$$-\Delta_\perp u+\mathcal{L}u-\tilde{\lambda}_{J_1}u+i\gamma x_2u=0\,, \tag{3.12}$$

where $\tilde{\lambda}_{J_1}=(1-\lambda)J_1^{-2/3}$ and $\gamma=J_2/J_1$. Multiplying (3.12) by u_n and integrating over $\mathbb{R}_+$ with respect to x_1, we obtain

$$-\Delta_\perp a_n+(\lambda_n-\tilde{\lambda}_{J_1})a_n+i\gamma x_2a_n=0\quad\text{in } \mathbb{R}^2\,,$$

where $a_n = \langle u_n, \bar{u} \rangle$. The above equation cannot have any nontrivial bounded solution in $\mathbb{R}^2$; otherwise it would also be a bounded solution in $\mathbb{R}^3$, which by Lemma 3.1 must identically vanish. Consequently, since a_n must be bounded, we must have $a_n \equiv 0$ for all n, from which the lemma easily follows in this case.

Consider now the case where $J_1 = 0$. Here we define

$$\tilde{u}(x_1, x_2, x_3) = \begin{cases} u(x_1, x_2, x_3), & x_1 > 0, \\ -u(-x_1, x_2, , x_3), & x_1 < 0. \end{cases}$$

Clearly, $\tilde{u}$ is a bounded weak solution of (3.1) in $\mathbb{R}^3$ and hence, by Lemma 3.1, $\tilde{u} \equiv 0$. □

For later reference we shall also need the following lemma.

LEMMA 3.4. *Let u denote a bounded solution of*

$$\begin{cases} -\Delta u - u + iJ_2 x_2 u = -\lambda u & \text{in } \mathbb{R}^3_+, \\ \frac{\partial u}{\partial x_1} = 0 & \text{on } \partial\mathbb{R}^3_+, \end{cases}$$

with $\lambda \in \mathbb{R}_+$. Then $u \equiv 0$.

Proof. Once again we define, this time an even function,

$$\tilde{u}(x_1, x_2, x_3) = \begin{cases} u(x_1, x_2, x_3), & x_1 > 0, \\ u(-x_1, x_2, , x_3), & x_1 < 0. \end{cases}$$

Clearly, $\tilde{u}$ is a bounded weak solution of (3.1) in $\mathbb{R}^3$ and hence, by Lemma 3.1, $\tilde{u} \equiv 0$. □

The last result in this section is needed in the next section in order to deal with the interface between $\partial\Omega^i$ and $\partial\Omega_n$ (that are perpendicular by assumption).

LEMMA 3.5. *Let*

$$Q = \{(x_1, x_2, x_3) \in \mathbb{R}^3_+ \,|\, x_2 > 0\}.$$

Let u denote a bounded solution of

$$\begin{cases} -\Delta u - u + i(J_2 x_2 + J_3 x_3)u = -\lambda u & \text{in } Q, \\ \frac{\partial u}{\partial x_1}(0, x_2, x_3) = 0, & x_2 > 0, \quad x_3 \in \mathbb{R}, \\ u(x_1, 0, x_3) = 0, & x_1 > 0, \quad x_3 \in \mathbb{R}, \end{cases}$$

with $\lambda \in \mathbb{R}_+$. Then $u \equiv 0$.

Proof. Once again we define an even extension of u,

$$\tilde{u}(x_1, x_2, x_3) = \begin{cases} u(x_1, x_2, x_3), & x_1 > 0, \\ u(-x_1, x_2, , x_3), & x_1 < 0. \end{cases}$$

By Lemma 3.3, $\tilde{u} = 0$. □

4. Large bounded domains in $\mathbb{R}^3$. We consider here (1.6) in the limit $R \to \infty$ which is the large domain limit. We first show that any eigenfunctions of the elliptic operator in (1.6) must decay exponentially fast, as $R \to \infty$ away from the boundary. As in the previous section we insert the imaginary part of λ into the electric potential

(and consequently consider a family of potentials) and then confine the discussion to real values of λ.

Before getting into the main discussion we repeat here the definition of ϕ_R from section 1 and list some of its properties. Recall that ϕ is the unique solution of (1.2), and that

$$\phi_R(x) = R\phi(x/R)\,.$$

The following proposition lists some of the properties of ϕ and ϕ_R.

PROPOSITION 4.1. *Let ϕ denote a solution of* (1.2) *and $\phi_R(x) = R\phi(x/R)$. Let $\mu \in \mathbb{R}$ and $\mu_R = R\mu$. Let further $x_j \in \Omega_{R_j}$ where $R_j \uparrow \infty$, and $\{\mu_j\}_{j=1}^{\infty} \subset \mathbb{R}$. Then*

(i) *we have either $|\phi_{R_j}(x_j) - \mu_j|$ is unbounded, or else we must have, up to a subsequence,*

$$\exists (b, J) \in \mathbb{R} \times \mathbb{R}^3 : \ \|\phi_{R_j} - \mu_j - J \cdot x - b\|_{L^\infty(D_r(x_j))} \to 0 \quad \forall r > 0\,,$$

where $D_r(x_j) = \Omega_R \cap B(x_j, r)$.

(ii) *we let Γ_μ denote the level set $\phi = \mu$. Let $M = \max_{x \in \partial\Omega_c} \phi(x)$ and $m = \min_{x \in \partial\Omega_c} \phi(x)$. If $\mu \notin [m, M]$, then Γ_μ is empty.*

(iii) *assume that $\partial\Omega_c$ is composed of exactly two connected sets as in Figure* 1*, and that Ω is diffeomorphic to a cylinder. If $\mu \in [m, M]$, but $\mu \notin \phi(\partial\Omega_c)$, then $\Gamma_\mu \cap \partial\Omega_i$ is a simple closed contour, separating $\partial\Omega$ into two subsets, such that none of them is a subset of $\partial\Omega_i$. Furthermore, $\nabla\phi \neq 0$ on Γ_μ.*

Proof. Let $b_j = \phi_{R_j}(x_j) - \mu_j$ and $J_j = \nabla\phi_{R_j}(x_j)$. To prove (i) we first choose a subsequence such that $(b_j, J_j) \to (b, J)$. The claim then follows from the Taylor expansion of ϕ_{R_j} near x_j.

The proof of (ii) follows immediately from the maximum principle.

To prove (iii) we notice first that since ϕ is real analytic, Γ_μ must either be closed or intersect $\partial\Omega$. If it is closed, then $\phi \equiv \mu$ inside Γ_μ, and hence also outside Γ_μ (in view of its analyticity), which is clearly a contradiction. Thus Γ_μ must intersect the boundary on $\partial\Omega_i$.

Since ϕ is continuous and since on one of the connected sets we must have $\phi < \mu$ and on the other one $\phi > \mu$, the intersection of $\partial\Omega_i$ with Γ_μ must contain at least one closed contour. This contour separates $\partial\Omega$ into two disjoints subsets $\partial\Omega_+$ and $\partial\Omega_-$, the first of them contains the connected subset of $\partial\Omega_c$; over which $\phi > \mu$, and contains a portion of $\partial\Omega^i$. Moreover, this contour is the boundary of a continuous subset of Γ_μ which we denote by A.

To see this, define cylindrical coordinates (r, θ, z) in Ω, where $\theta \in [-\pi, \pi)$ and $0 \leq r < R(\theta, z)$. Then, for each (r, θ) there exists a finite (as ϕ is real analytic) set $\{z_j\}$ such that $(r, \theta, z_j) \in \Gamma_\mu$. Denote the minimum in this set by z_m. Clearly, $z = z_m(r, \theta)$ is continuous, and thus we can define

$$A = \{(r, \theta, z) \,|\, z = z_m(r, \theta)\}\,.$$

We now consider the problem for ϕ in a subdomain of Ω whose boundary consists of A and $\partial\Omega_+$. By the maximum principle and Hopf's lemma we have $\phi > \mu$ on $\partial\Omega_+$ and in the interior. In a similar manner we show that $\phi < \mu$ both on $\partial\Omega_-$ and in the interior of the subdomain surrounded by A and $\partial\Omega_-$. This shows that $A = \Gamma_\mu$. From Hopf's lemma it follows that $\nabla\phi \neq 0$ on Γ_μ. ∎

Remark 4.1. While property (i) will be used extensively throughout this section, properties (ii) and (iii) are brought here to provide the reader with some intuition of the behavior of ϕ in the "wire-like" domain presented in Figure 1.

Since (1.6) contains the term $i(\phi_R-\mu_R)\psi_R$ which might be unbounded as $R\to\infty$, we must provide here the following elliptic estimate.

LEMMA 4.1. *Let u_R denote a solution of*

$$\begin{cases} -\Delta u_R - u_R + i(\phi_R-\mu_R)u_R = -\lambda u_R & \text{in } \Omega_R, \\ \frac{\partial \psi_R}{\partial \nu} = 0, & \text{on } \partial\Omega_R^i, \\ \psi_R = 0, & \text{on } \partial\Omega_R^c, \end{cases} \tag{4.1}$$

in which $\lambda\in\mathbb{R}_+$. Let $D_r(x_0)=\Omega_R\cap B_r(x_0)$, where $x_0\in\Omega_R$ is chosen such that either

(i) *$D_r(x_0)=B_r(x_0)$, or*

(ii) *$x_0\in\partial\Omega_R$ and either $(\partial D_r(x_0)\cap\partial\Omega_R)\subset\partial\Omega_R^c$ or $(\partial D_r(x_0)\cap\partial\Omega_R)\subset\partial\Omega_R^i$.*

Then,

$$\exists \tilde r>0:\ \|u_R\|_{L^\infty(D_r(x_0))}\le C_r\|u_R\|_{L^2(D_{2r}(x_0))} \quad \forall x_0\in\Omega_R,\ r<\tilde r\,, \tag{4.2}$$

where C_r is independent of R and x_0.

Proof. Let $\rho_R=|u_R|$. By (4.1) we have that

$$-\Delta\rho_R-\rho_R\le 0$$

in Ω_R. Let $x_0\in\Omega_R$. We set U_r to be the solution (we discuss its existence below) of

$$\begin{cases} -\Delta U_r - U_r = 0 & \text{in } D_r(x_0), \\ U_r = 0 & \text{on } \partial D_r^c(x_0), \\ \frac{\partial U_r}{\partial \nu} = 0 & \text{on } \partial D_r^i(x_0), \\ U_r = \rho_R & \text{on } \partial D_r^s(x_0), \end{cases} \tag{4.3}$$

where $\partial D_r^c(x_0)=\partial D_r(x_0)\cap\partial\Omega_R^c$, $\partial D_r^i(x_0)=\partial D_r(x_0)\cap\partial\Omega_R^i$, and $\partial D_r^s(x_0)=\partial D_r(x_0)\setminus(\partial D_r(x_0)\cap\partial\Omega_R)$. Note that either $\partial D_r^c(x_0)=\phi$ or $\partial D_r^i(x_0)=\phi$. Clearly, there exists $\tilde r$, independent of x_0 and R, such that for every $r<\tilde r$, we have

$$\inf_{u\in\mathcal{D}}\frac{\text{int}_{D_r(x_0)}|\nabla u|^2}{\text{int}_{D_r(x_0)}|u|^2}>1\,, \tag{4.4}$$

where

$$\mathcal{D}=\{u\in H^1(D_r)\,|\,u=0 \text{ on } \partial D_r(x_0)\setminus\partial D_r^i(x_0)\}\,.$$

For $r<\tilde r$ the elliptic operator in (4.3) is invertible, and hence a unique U_r exists.

Let then $V=\rho_R-U_r$. Clearly,

$$\begin{cases} -\Delta V - V\le 0 & \text{in } D_r(x_0), \\ \frac{\partial V}{\partial \nu} = 0 & \text{on } \partial D_r^i(x_0), \\ V = 0 & \text{on } \partial D_r(x_0)\setminus\partial D_r^i(x_0). \end{cases} \tag{4.5}$$

Let further

$$V_+=\begin{cases} V, & V\ge 0, \\ 0, & V<0. \end{cases}$$

Multiplying (4.5) by V_+ and integrating over $D_r(x_0)$, we obtain that

$$\int_{D_r} |\nabla V_+|^2 - |V_+|^2 \leq 0\,,$$

and since $V \in \mathcal{D}$, we obtain by (4.4) that $V_+ = 0$. Consequently we have

$$\rho_R \leq U_r \quad \text{in } D_r(x_0). \tag{4.6}$$

To complete the proof of (4.2) it is thus necessary to obtain an estimate of $\|U_r\|_{L^\infty(D_r)}$ in terms of $\|\rho_R\|_{L^2(D_r)}$. In case (i) above, since U_r is unique, we can use Theorem 10.5 in [3], together with Sobolev embeddings, to obtain

$$\|U_r\|_{L^\infty(D_r(x_0))} \leq C_r \|\rho_R\|_{H^{1/2}(\partial D_r^s)}\,, \tag{4.7}$$

where C_r is independent of x_0 and R.

In case (ii) $D_r(x_0)$ is diffeomorphic to a hemisphere with radius r. Denote the diffeomorphism by T_{R,x_0}. Note that $T_{R,x_0} \to I$ as $R \to \infty$ uniformly in x_0 (as long as the assumption in (ii) holds) in view of the transformation (1.5). Let $B_+ = T_{R,x_0}(D_r)$. On the flat surface of B_+ we have either $\partial U_r/\partial \nu = 0$ or $U_r = 0$. In the first case one can extend U_r evenly to a sphere B, whereas in the second case we use an odd extension for that matter. In both cases U_r satisfies

$$A : \nabla(A^t \nabla U_r) + U_r = 0 \quad \text{in } B\,,$$

where $A = DT_{R,x_0}$ is smooth in B and satisfies $A \to I$ uniformly in B as $R \to \infty$. On ∂B, U_r is equal to either the even extension or the odd extension of ρ_R. Hence, for sufficiently large R, U_r must satisfy (4.7) in case (ii) as well.

Clearly,

$$\|\rho_R\|_{H^{1/2}(\partial D_r^s)} \leq \|\rho_R\|_{H^1(D_r(x_0))}\,. \tag{4.8}$$

Multiplying (4.1) by $\chi_{2r}^2(|x-x_0|)\bar{u}_R$ and integrating over Ω_R, we obtain from the real part of the identity

$$\int_{D_{2r}(x_0)} |\nabla(\chi_{2r} u_R)|^2 - \big(|\nabla \chi_{2r}|^2 + (1-\lambda)\chi_{2r}^2\big)|u_R|^2 = 0\,.$$

Hence,

$$\int_{D_r(x_0)} |\nabla \rho_R|^2 \leq C_r \int_{D_{2r}(x_0)} |\rho_R|^2\,. \tag{4.9}$$

Combining the above with (4.6)–(4.8) yields (4.2). □

As an immediate conclusion of Lemma 4.1 we prove the following lemma

LEMMA 4.2. *Let u_R denote a solution of (4.1) with $\lambda \in \mathbb{R}_+$. Let X_R denote a maximum point of $|u_R|$ in Ω_R. Then, $|\phi_R(x_R) - \mu_R|$ is bounded as $R \to \infty$.*

Proof. We first normalize u_R by $|u_R(x_R)|$ so that $\|u_R\|_\infty = 1$. Let $D_r(x_R) = B(x_R, r) \cap \Omega_R$. Multiplying (4.1) by $\chi_r^2(|x - x_R|)\bar{u}_R$ and integrating over Ω_R, we obtain from the imaginary part of the identity

$$\int_{D_r} \nabla(\chi_r^2) \cdot \frac{1}{2i}(\bar{u_R}_j \nabla u_R - \nabla \bar{u}_R) + \int_{D_r} \chi_r^2(\phi_R - \mu_R)|u_R|^2 = 0\,.$$

Let $b_R = \phi_R(x_R) - \mu_R$. Since $\nabla\phi_R$ is bounded in Ω_R, and since r is fixed, we have

$$\inf_{x\in D_r} |\phi_R - \mu_R| \geq \frac{1}{2} b_R \,.$$

Thus,

$$b_R \int_{D_{r/2}} |u_R|^2 \leq C_r \int_{D_r} |u_R|^2 + |\nabla u_R|^2 \,.$$

Using (4.9) and the fact that $|u_R| \leq 1$, we obtain that

$$\int_{D_{r/2}} |u_R|^2 \leq \frac{C}{b_R} \,.$$

By Lemma 4.1 we then have that

$$1 = |u_R(x_R)| \leq \frac{C_r}{b_R},$$

from which the lemma immediately follows. □

Remark 4.2. Note that if $\phi_R \neq \mu_R$ for all $x \in \Omega_R$, then $u_R \equiv 0$ must be the unique solution of (4.1). To see this multiply (4.1) by $\bar{u}_R$ and integrate over Ω_R to obtain from the imaginary part,

$$\int_{\Omega_R} (\phi_R - \mu_R)|u_R|^2 = 0 \,.$$

Since $\phi_R - \mu_R$ is either positive or negative throughout Ω_R, u_R must vanish everywhere.

Denote the curve in Ω along which we have $\phi = \mu$ by Γ_μ. Denote its image under the mapping (1.5) by Γ_R. By the previous lemma we have that $d(x_R, \Gamma_R)$ is bounded as $R \to \infty$. We now prove that u_R must decay exponentially fast away from Γ_R.

LEMMA 4.3. *Let u_R denote a solution of* (4.1) *with $\lambda \in \mathbb{R}_+$. Then, there exists $\alpha > 0$ such that*

$$\int_{\Omega_R} |u_R|^2 e^{2\alpha s} \leq C \,, \tag{4.10}$$

where $s = d(x, \Gamma_R)$ and C is independent of R.

Recall that by Proposition 4.1, for domains that are diffeomorphic to a cylinder, when $\phi \neq \mu$ for every $x \in \partial\Omega^c$, Γ_μ must be a surface whose boundary is a closed simple contour on $\partial\Omega^i$. We also have $\nabla\phi \neq 0$ on Γ_μ. Note also that by the above remark, if Γ_R is empty, then $u_R \equiv 0$.

Proof. It is convenient to consider here u_R for which $\|u_R\|_{L^2(\Omega_R)} = 1$. Let Ω_β^+ and Ω_β^- be, respectively, defined by

$$\Omega_\beta^+ = \{x \in \Omega_R \,|\, \phi_R - \mu_R > \beta\},$$

$$\Omega_\beta^- = \{x \in \Omega_R \,|\, \phi_R - \mu_R < -\beta\}$$

for some $\beta > 0$ which is independent of R.

Let $\eta_\beta^+ \in C^\infty(\Omega_R,[0,1])$ and η_β^-, respectively, be defined by

$$\eta_\beta^+ = \begin{cases} 1, & x \in \Omega_\beta^+, \\ 0, & x \in \Omega_0^-, \end{cases}$$

and

$$\eta_\beta^- = \begin{cases} 1, & x \in \Omega_\beta^-, \\ 0, & x \in \Omega_0^+. \end{cases}$$

Let $\mathcal{C}_\beta^\pm$ denote the portion of $\partial\Omega_\beta^\pm$ which is not on $\partial\Omega_R$. As $\nabla\phi_R(x) = \nabla\phi(x/R)$ and since $\nabla\phi$ is bounded in Ω, we have that

$$d(\mathcal{C}_\beta^\pm, \Gamma_\mu) \geq \frac{\beta}{\|\nabla\phi\|_{L^\infty(\Omega)}}\,.$$

Hence, we can choose $\eta_\beta^\pm$ such that $|\nabla\eta_\beta^\pm| < C$. Let

$$D_\beta^+ = \{x \in \Omega_R \,|\, 0 < \eta_\beta^+ < 1\}\,.$$

We choose $\eta_\beta^\pm$ such that

$$\sup_{x\in D_\beta^+} s \leq 1\,.$$

Multiplying (4.1) by $(\eta_\beta^+)^2 e^{2\alpha s}\bar{u}_R$ and integrating over Ω_R we obtain, for the imaginary part and the real part, respectively,

(4.11a)
$$\int_{\Omega_R} \nabla((\eta_\beta^+)^2 e^{2\alpha s})\cdot\frac{1}{2i}(\bar{u}_R\nabla u_R - u_R\nabla\bar{u}_R) + \int_{\Omega_R}(\phi_R-\mu_R)(\eta_\beta^+)^2|u_R|^2 e^{2\alpha s} = 0\,,$$

(4.11b)
$$\int_{\Omega_R}(\eta_\beta^+)^2|\nabla u_R|^2 e^{2\alpha s} = (1-\lambda)\int_{\Omega_R}(\eta_\beta^+)^2|u_R|^2 e^{2\alpha s} - \frac{1}{2}\int_{\Omega_R}\nabla((\eta_\beta^+)^2 e^{2\alpha s})\cdot\nabla|u_R|^2\,.$$

From the real part, (4.11b), we obtain that for every $\epsilon > 0$ we have

(4.12)
$$(1-2\alpha\epsilon)\int_{\Omega_R}(\eta_\beta^+)^2 e^{2\alpha s}|\nabla u_R|^2 \leq \Big(1+\frac{\alpha}{2\epsilon}\Big)\int_{\Omega_R}(\eta_\beta^+)^2 e^{2\alpha s}|u_R|^2 + C\int_{D_\beta^+} e^{2\alpha s}|\nabla u_R|^2\,.$$

From (4.11a), or the imaginary part, we obtain

$$\beta\int_{\Omega_\beta^+}|u_R|^2 e^{2\alpha s} \leq \alpha\int_{\Omega_R}(\eta_\beta^+)^2 e^{2\alpha s}[|u_R|^2+|\nabla u_R|^2] + C\int_{D_\beta^+} e^{2\alpha s}[|u_R|^2+|\nabla u_R|^2].$$

Combining the above with (4.12) for $\epsilon = 4\alpha^{-1}$, we obtain

$$(\beta - 2\alpha - 4\alpha^3)\int_{\Omega_\beta^+}|u_R|^2 e^{2\alpha s} \leq C\int_{D_\beta^+} e^{2\alpha s}[|u_R|^2+|\nabla u_R|^2]\,. \tag{4.13}$$

Multiplying (4.1) by $\bar{u}_R$ and integrating over Ω_R, we obtain (for the real part)

$$\int_{\Omega_R} |\nabla u_R|^2 = (1-\lambda) \int_{\Omega_R} |u_R|^2 = 1-\lambda\,.$$

Consequently, we obtain from (4.13) that for any given α we may choose β to be sufficiently large (but still independent of R) so that $\beta > 2\alpha + 4\alpha^3$, and hence

$$\int_{\Omega_\beta^+} |u_R|^2 e^{2\alpha s} \le Ce^{2\alpha} \int_{D_\beta^+} [|u_R|^2 + |\nabla u_R|^2] \le 2Ce^{2\alpha}\,. \qquad \square$$

Remark 4.3. By Lemma 4.1 it follows that $|u_R|$ decays exponentially fast away from Γ_R also in a pointwise sense.

We now prove that any eigenfunction corresponding to a nonpositive eigenvalue must decay exponentially fast away from the boundary.

LEMMA 4.4. *Let u_R denote a solution of* (4.1). *Then*

$$\exists \alpha > 0 : \ |u_R| \le Ce^{-\alpha d(x,\partial\Omega)} \quad \forall \mu_R \in \mathbb{R}\,, \tag{4.14}$$

where α is independent of R, μ_R, and λ. Furthermore, denote by x_R the maximum point of $|u_R|$. Then $d(x_R, \partial\Omega_R)$ is bounded as $R \to \infty$.

Proof. We apply standard blow-up arguments to prove the lemma. Let

$$\Omega(R,k,s) = \{x \in \Omega_R |\ d(x,\partial\Omega_R) \ge ks\}\,.$$

We prove the exponential rate of decay by showing first that

$$\exists R_0, s_0 : \ \|u_R\|_{L^\infty\left(\Omega(R,k+1,s)\right)} \le \frac{1}{2} \|u_R\|_{L^\infty\left(\Omega(R,k,s)\right)} \quad \forall s > s_0,\ R > R_0,\ k \in \mathbb{N}. \tag{4.15}$$

Suppose, for contradiction, that (4.15) does not hold. Then, there exist sequences $\{R_j\}_{j=1}^\infty$, $\{s_j\}_{j=1}^\infty$, and $\{k_j\}_{j=1}^\infty$ satisfying $R_j \uparrow \infty$, $s_j \uparrow \infty$, $k_j \in \mathbb{N}$, and

$$\left\|u_{R_j}\right\|_{L^\infty\left(\Omega(R_j,k_j+1,s_j)\right)} \ge \frac{1}{2} \left\|u_{R_j}\right\|_{L^\infty\left(\Omega(R_j,k_j,s_j)\right)} \stackrel{def}{=} \frac{1}{2} m_j. \tag{4.16}$$

Let

$$\tilde{u}_{R_j} = \frac{u_{R_j}}{m_j}.$$

By (4.15) there exists $x_j \in \Omega(R_j, k_j+1, s_j)$ such that

$$|\tilde{u}_{R_j}(x_j)| \ge \frac{1}{2}\,. \tag{4.17}$$

For notational convenience we also let $f_j(x) = \tilde{u}_{R_j}(x_j + x)$.

We now distinguish between two different cases.

Case 1.

$$b_j = \inf_{x \in B(x_j, s_j)} |\phi_{R_j} - \mu_{R_j}| \to \infty$$

up to a subsequence.

Let $\chi_r \in C^\infty(\mathbb{R}_+, [0,1])$ be defined by (3.3). Since f_j satisfies (4.1) we multiply it by $\chi_r^2(0)\overline{f}_j$ and integrate over $B(0,r)$ to obtain, for the imaginary part,

$$\int_{B(0,r)} \nabla(\chi_r^2) \cdot \frac{1}{2i}(\bar{f}_j \nabla f_j - f_j \nabla \bar{f}_j) + \int_{B(0,r)} \chi_r^2(\phi_{R_j} - \mu_{R_j})|f_j|^2 = 0,$$

yielding

$$b_j \int_{B(0,r/2)} |f_j|^2 \leq C_r \int_{B(0,r)} |f_j|^2 + |\nabla f_j|^2 \tag{4.18}$$

for all $r < s_j$. For the real part we obtain that

$$\int_{B(0,r)} |\nabla(\chi_r f_j)|^2 - \left(|\nabla \chi_r|^2 + (1-\lambda)\chi_r^2\right)|f_j|^2 = 0\,,$$

and hence

$$\int_{B(0,r/2)} |\nabla f_j|^2 \leq C \int_{B(0,r)} |f_j|^2\,. \tag{4.19}$$

Combining (4.18) with (4.19), we obtain

$$\int_{B(0,r/2)} |f_j|^2 \leq \frac{C_r}{b_j} \int_{B(0,2r)} |f_j|^2\,.$$

As $|f_j| \leq 1$ we obtain that

$$\int_{B(0,r/2)} |f_j|^2 \to 0$$

as $j \to \infty$, and by Lemma 4.1 also that $f_j(0) \to 0$, a contradiction.

Case 2. $\limsup_{j\to\infty} b_j < \infty$.

Let $J = |\nabla\phi|$. We choose a coordinate system, where $\nabla\phi(x_j)$ is parallel to the x_1 axis. Then, by Lemma 4.2 we have a subsequence for which

$$\phi_{R_j} - \mu_{R_j} \to Jx_1 + b$$

uniformly in $B(0,r)$ for all $r > 0$, where b is a constant. Thus, by standard elliptic estimates and Sobolev embeddings, there exists a subsequence $\{f_{j_k}\}_{k=1}^\infty$ such that $f_{j_k} \to f_\infty$ uniformly on every compact set in $\mathbb{R}^3$ and such that f_∞ is a bounded solution of

$$-\Delta f_\infty - f_\infty + i(Jx_1 + b)f_\infty = -\lambda f_\infty \quad \text{in } \mathbb{R}^3\,.$$

By Lemma 3.1 we must have $f_\infty \equiv 0$, a contradiction.

Thus, we have proved (4.15), and hence also (4.14). The boundedness of $d(x_R, \partial\Omega_R)$ follows from (4.14) as well. □

The following lemma provides the basis for our main stability result.

LEMMA 4.5. *Let $\partial\Omega_n$ denote the subset of $\partial\Omega_c$, where $\nabla\phi$ is perpendicular to $\partial\Omega$. Suppose that either $|\nabla\phi| > J_c$ for all $x \in \partial\Omega_n$ or that $\partial\Omega_n$ is empty. Then, for sufficiently large R, $u_R \equiv 0$ is the unique solution of* (4.1) *for all $\mu_R \in \mathbb{R}$.*

Proof. Let x_R be the point where u_R obtains its maximum in Ω_R. Let x_0^R denote its projection on $\partial\Omega_R$. Recall that by Lemma 4.4 $|x_R - x_0^R|$ is bounded as $R \to \infty$. Note that by Lemma 4.2 ϕ_R converges uniformly to a linear function in $D_r(x_0^R)$ for all fixed $r > 0$ as $R \to \infty$. Suppose first that $x_0^R \in \partial\Omega_i^R$. Following [11], let (t_1, t_2, t_3) denote a local curvilinear coordinate system, whose origin lies at x_0^R, such that $t_3 = d(x, \partial\Omega_R)$ when $x \in \Omega_R$ and such that the t_1 and t_2 curves on $\partial\Omega$ are the lines of curvature. Let further κ_1^R and κ_2^R denote the respective principal curvatures on $\partial\Omega_R$. Clearly, $\kappa_i^R = \kappa_i/R$ $(i = 1, 2)$, where κ_i is the corresponding principal curvature on $\partial\Omega$, at $x_0 = x_0^R/R$.

Since $\partial\Omega$ is smooth near x_0^R, this curvilinear coordinate system is properly defined in some neighborhood of x_0^R. Let

$$B^+(0, r) = \{(t_1, t_2, t_3) \in B(0, r) \,|\, t_3 > 0\}\,.$$

Then, the above coordinate system is well defined in $B^+(0, \delta R)$ for some $\delta > 0$. We can now present any x in this neighborhood by

$$x = r(t_1, t_2) - t_3\nu\,,$$

where ν is the outward normal at $(t_1, t_2, 0)$. Let

$$g_{ij}(t_1, t_2) = \frac{\partial r}{\partial t_i} \cdot \frac{\partial r}{\partial t_j}, \quad i, j = 1, 2,$$

and

$$G_{ij} = [1 - \kappa_i t_3/R] g_{ij}, \quad i, j = 1, 2\,.$$

Since our coordinate system is orthogonal, we have

$$g_{12} = 0\,.$$

Furthermore, we can scale t_1 and t_2 so that

$$g_{11} = g_{22} = 1 + O(1/R) \quad \text{as } R \to \infty$$

uniformly in $B^+(0, r)$ for every fixed $r > 0$. Finally, we define

$$\begin{cases} G = \sqrt{G_{11} G_{22}}, \\ \alpha_j = \frac{G}{G_{jj}}, \qquad j = 1, 2. \\ \alpha_3 = G, \end{cases}$$

Let $w_R(x) = u_R(x)/|u_R(x_R)|$. In the new coordinates, (4.1) takes the form

$$-\sum_{j=1}^{3} \frac{1}{G} \frac{\partial}{\partial t_j} \left(\alpha_j \frac{\partial w_R}{\partial t_j} \right) - w_R + i(\phi_R - \mu_R) w_R = -\lambda w_R$$

in $B^+(0, \delta R)$. Standard elliptic estimates then prove the existence of a sequence $\{w_{R_j}\}_{j=1}^{\infty}$ such that $w_{R_j} \to w_\infty$ uniformly on every compact set in $\mathbb{R}^3_+$, where here

$$\mathbb{R}^3_+ = \{(t_1, t_2, t_3) \,|\, t_3 > 0\}\,.$$

By standard elliptic estimates again we have that w_∞ satisfies the following problem:

$$\begin{cases} -\Delta w_\infty - w_\infty + i(J_1 t_1 + J_2 t_2 + b) w_\infty = -\lambda w_\infty & \text{in } \mathbb{R}^3_+, \\ \frac{\partial w_\infty}{\partial \nu} = 0 & \text{on } \partial\mathbb{R}^3_+. \end{cases}$$

By Lemma 3.4 we have $w_\infty \equiv 0$ in $\mathbb{R}^3_+$, a contradiction since $|w_R(x_R)| = 1$ and $|x_R - x_0^R|$ is bounded.

Consider now the case where $x_0^R \in \partial\Omega_c^R$. Following the same procedure as before we obtain that $w_R \to w_\infty$ uniformly on every compact set in $\mathbb{R}^3_+$, where w_∞ must satisfy

$$\begin{cases} -\Delta w_\infty - w_\infty + i(J_1 t_1 + J_2 t_2 + J_3 t_3 + b) w_\infty = -\lambda w_\infty & \text{in } \mathbb{R}^3_+, \\ w_\infty = 0 & \text{on } \partial\mathbb{R}^3_+. \end{cases}$$

If $\partial\Omega_n$ is empty, we have $J_1^2 + J_2^2 > 0$. By Lemma 3.3 we then have $w_\infty \equiv 0$. Otherwise, if $x_0 \in \partial\Omega_n$, we must have $w_\infty \equiv 0$ by Lemma 3.2 since $J_3 > J_c$.

Finally, if x_0 lies on the interface between $\partial\Omega_i$ and $\partial\Omega_c$, we obtain that w_∞ must satisfy, since the two surfaces are perpendicular to each other at the interface, a problem in Q, where

$$Q = \{(t_1, t_2, t_3) \,|\, t_3 > 0,\ t_1 > 0\}\,.$$

We have

$$\begin{cases} -\Delta w_\infty - w_\infty + i(J_1 t_1 + J_2 t_2 + b) w_\infty = 0 & \text{in } Q, \\ \frac{\partial w_\infty}{\partial \nu} = 0, & x \in \partial Q\ t_3 = 0, \\ w_\infty = 0, & x \in \partial Q :\ t_1 = 0. \end{cases}$$

By Lemma 3.5 we have $w_\infty \equiv 0$. (Note that some modification of the local coordinate system is necessary in this case.) □

Proof of Theorem 1.1. Since the principal part of the differential operator on the left-hand side of (4.1) is the Laplacian, it can be regarded as a perturbation of a self-adjoint operator. Thus, it follows from the discussion below Theorem 15.2 in [2] regarding such perturbations that it has exactly one direction which is not a direction of minimal growth, that is, $\arg\lambda = 0$. Hence, it follows from Theorem 15.1 in [2] that the spectrum of this differential operator must be discrete and that all its eigenvalues must have finite multiplicity. Furthermore, by Theorem 16.5 in [2], the eigenfunctions span $L^2(\Omega_R, \mathbb{C})$. Hence, for $J > J_c$, since all the eigenvalues of the above operator must have positive real part, the normal state must be stable.

Consider now the case when a point $x \in \partial\Omega_c$ exists, where

$$\left|\frac{\partial\phi}{\partial\nu}\right| = |\nabla\phi| = J < J_c\,.$$

To prove the short-time instability we look at the solution of (1.6), after applying to it the transformation (2.2), with the initial condition

$$\psi(x, 0) = u_1(t_3)\chi_{R^{1/2}}(t_1, t_2)\eta_R(t_3)\,,$$

where (t_1, t_2, t_3) are the above-defined system of local curvilinear coordinates, χ_r is defined in (3.3), and

$$\eta_R(x) = \begin{cases} 1, & x < \frac{1}{2}\delta R, \\ 0, & x > \delta R, \end{cases}$$

is a smooth cutoff function.

Let $\tilde{\beta} = \lambda_J - \lambda_{J_c}$. We write

$$\psi_R = v + \psi_0(x) e^{\tilde{\beta} t} \tag{4.20}$$

to obtain

$$\frac{\partial v}{\partial t} - \Delta v - \lambda_J v = f\,. \tag{4.21}$$

The precise form of f need not concern us except for the fact that

$$\|f\|_2 \le \frac{C_\alpha}{R^\alpha} \|\psi_0\|_2 e^{\tilde{\beta} t} \quad \forall \alpha < 1\,. \tag{4.22}$$

Multiplying (4.21) by $\bar{v}$ and integrating by parts, we obtain for the real part

$$\begin{cases} \frac{\partial \|v\|_2}{\partial t} - \lambda_J \|v\|_2 = \|f\|_2, \\ \|v\|_2(0) = 0. \end{cases}$$

Consequently,

$$\|v\|_2(t) \le \int_0^t e^{\lambda(t-\tau)} \|f\|_2(\tau) d\tau\,,$$

and hence,

$$\|v\|_2 \le \frac{C_\alpha}{R^\alpha} \|\psi_0\|_2 e^{(\lambda_J + \tilde{\beta}) t}\,.$$

Clearly, there exist $T_R \sim \mathcal{O}(\ln R)$, as $R \to \infty$, such that

$$t < T_R \Rightarrow \frac{C_\alpha}{R^\alpha} e^{\lambda_J t} < \frac{1}{2}\,,$$

and thus

$$t < T_R \Rightarrow \|\psi\|_2 \ge \frac{1}{2} \|\psi_0\|_2 e^{\tilde{\beta} t}\,.$$

Applying the inverse of (2.2), we obtain (1.7). □

The above instability result is valid, of course, only for $T < T_R$. Proving long-time instability appears to be a much more difficult problem. The stability proof presented above relies on the convergence of any solution of (4.1) to a solution of (3.8) uniformly on every compact set near the point on the boundary where J is perpendicular to it. This, however, does not prove convergence of the spectrum or even of its bottom. What has been demonstrated is only the upper semicontinuity of the spectrum of the differential operator on the left-hand side of (4.1). Lower semicontinuity of the spectrum appears to be much harder to prove, especially since the operator is not self-adjoint. Nevertheless, it does seem reasonable to conjecture that the solution would continue to grow exponentially fast as $t \to \infty$, in view of the above short-time instability result. Further research is necessary in order to establish that result.

5. Concluding remarks. In the previous section we proved that the normal state remains stable in the large domain limit, as long as the current on the boundary, at points where it is perpendicular to it, is greater than J_c. If the current is nowhere perpendicular to the boundary, then as long as it doesn't vanish there, the normal state must be stable. We also demonstrate short-time instability when $J_m < J_c$.

In the following we provide a short list of interesting problems that are waiting to be resolved:

1. Proving long-time instability when $J > J_c$. We have elaborated on this matter at the end of the preceding section.
2. Adding the effect of magnetic fields. This magnetic field can be either induced by the electric current (via (1.3)) or else be applied externally (or both). It has been verified experimentally that an induced magnetic field can generate vortices [14] if the current is sufficiently large and the material is close to the wholly superconducting state. However, its effect on the critical current J_c has not been investigated. It is reasonable to believe that J_c would become smaller if we add the effect of the magnetic field.
3. Adding the effect of temperature, since electric currents have the tendency to heat the sample, thereby creating vortices [14]. Incorporating this effect requires modification of (1.1), and the use of different nondimensionalization; otherwise the domain would become temperature-dependent.
4. Proving that the bifurcating branch (at $J = J_c$) is unstable.

Appendix A. The Hilbert–Schmidt norm of $\mathcal{L}_\lambda^{-1}$.

LEMMA A.1. *Let $\tilde{G}$ be as given by* (2.14b). *Then, for any $\lambda \in \mathbb{C} \setminus \{\lambda_n\}_{n=1}^{\infty}$, where $\{\lambda_n\}_{n=1}^{\infty}$ are given as in* (2.11),

$$\exists C, M > 0 : \ \|\tilde{G}\| \leq C e^{M\lambda^{3/2}} . \tag{A.1}$$

Proof. We first note that

$$W(w_1, \tilde{w}_2)(x,\lambda) = w_2'(x,\lambda) w_1(x,\lambda) - w_1'(x,\lambda) w_2(x,\lambda)$$

is independent of x by Abel's formula. Furthermore, for $y = x + i\lambda$ we have

$$W(w_1, \tilde{w}_2)(x,\lambda) = W(w_1, \tilde{w}_2)(y,0) = W(w_1, \tilde{w}_2)(0,0) .$$

Therefore, W is independent of both x and λ.

Since W is constant, it follows that $\tilde{G}$ is symmetric, i.e.,

$$\tilde{G}(x,\xi,\lambda) = \tilde{G}(\xi,x,\lambda) .$$

Consequently, it suffices to prove that

$$\int_0^\infty d\xi \int_\xi^\infty dx |\tilde{G}(x,\xi)|^2 = C_1 \int_0^\infty d\xi |\tilde{w}_2(\xi)|^2 \int_\xi^\infty dx |\tilde{w}_1(x)|^2 \leq C e^{M\lambda^{3/2}} ,$$

where C, C_1, and M are all independent of λ.

To obtain the above estimate we use asymptotic properties of Airy's functions [1, 13], from which it follows that

$$|A_i(z)| \leq \frac{C}{|z|^{1/4}} \left| e^{-\frac{2}{3} z^{3/2}} \right| . \tag{A.2}$$

Consider then first the domain $\xi > M_0\lambda$ for sufficiently large $M_0 > 0$. We have

$$\int_\xi^\infty |w_1(x)|^2 dx \le \int_\xi^\infty \frac{dx}{|x+i\lambda|^{1/2}} e^{-\beta(x)|x+i\lambda|^{3/2}},$$

where

$$\beta(x) = \frac{4}{3}\cos\left(\frac{3}{2}\arg(x+i\lambda) + \frac{\pi}{4}\right).$$

It is easy to show that

$$\begin{cases} |\beta'(x)| \le C\frac{|\lambda|}{|x+i\lambda|^2}, \\ |\beta'(x)| \le C\frac{|\lambda|}{|x+i\lambda|^3}. \end{cases} \tag{A.3}$$

Let then

$$\gamma(x) = \beta(x)|x+i\lambda|^{3/2}.$$

By (A.3) we have

$$\begin{cases} |\gamma'(x)| \ge \frac{3}{2}\beta(x)|x+i\lambda|^{1/2}\left[1 - C\frac{|\lambda|}{|x+i\lambda|}\right], \\ |\gamma''(x)| \le \frac{3}{4}\beta(x)|x+i\lambda|^{-1/2}\left[1 + C\frac{|\lambda|}{|x+i\lambda|}\right]. \end{cases} \tag{A.4}$$

Using (A.4) and integration by parts, we obtain

$$\begin{aligned} \int_\xi^\infty e^{-\gamma(x)}dx &\le \frac{1}{|\gamma'(\xi)|}e^{-\gamma(\xi)} + \int_\xi^\infty \frac{|\gamma''(x)|}{|\gamma'(x)|}e^{-\gamma(x)}dx \\ &\le \frac{C}{|\xi+i\lambda|^{1/2}}e^{-\gamma(\xi)} + \frac{C}{|\xi+i\lambda|^{3/2}}\int_\xi^\infty e^{-\gamma(x)}dx. \end{aligned}$$

Hence,

$$\int_\xi^\infty e^{-\gamma(x)}dx \le \frac{C}{|\xi+i\lambda|^{1/2}}e^{-\gamma(\xi)},$$

from which we easily obtain that

$$\int_\xi^\infty |w_1(x)|^2 dx \le \frac{C}{|\xi+i\lambda|}e^{-\gamma(\xi)}. \tag{A.5}$$

From the asymptotic behavior of Airy's function (A.2), we obtain again

$$|\tilde{w}_2(\xi)|^2 \le \frac{C}{|\xi+i\lambda|^{1/2}}e^{\gamma(\xi)}.$$

Thus,

$$\int_{M_0\lambda}^\infty d\xi|\tilde{w}_2(\xi)|^2 \int_\xi^\infty dx|\tilde{w}_1(x)|^2 \le \frac{C}{\lambda^{1/2}}.$$

To complete the proof we need to bound the norm for $0 < \xi < M_0\lambda$. By (A.2) and (2.14b) we have

$$|\tilde{G}(x,\xi,\lambda)| \leq C \exp\left\{\frac{2}{3}(M_0+1)^{3/2}|\lambda|^{3/2}\right\}.$$

Consequently, from the above and (A.5) we obtain

$$\int_0^{M_0\lambda} d\xi |\tilde{w}_2(\xi)|^2 \int_\xi^\infty dx |\tilde{w}_1(x)|^2 = \int_0^{M_0\lambda} d\xi |\tilde{w}_2(\xi)|^2 \int_\xi^{M_0\lambda} dx |\tilde{w}_1(x)|^2$$
$$+ \int_0^{M_0\lambda} d\xi |\tilde{w}_2(\xi)|^2 \int_{M_0\lambda}^\infty dx |\tilde{w}_1(x)|^2 \leq C M_0^2 \lambda^2 \exp\left\{\frac{2}{3}(M_0+1)^{3/2}|\lambda|^{3/2}\right\},$$

from which (A.1) easily follows. □

Acknowledgments. The author wishes to thank Professor Bernard Helffer and Professor Yehuda Pinchover for their comments.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, Mineola, NY, 1972.
[2] S. Agmon, *Lectures on Elliptic Boundary Value Problems*, prepared for publication by B. Frank Jones, Jr., with the assistance of George W. Batten, Jr., Van Nostrand Math. Stud. 2, D. Van Nostrand Co., Inc., Princeton, NJ, 1965.
[3] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions*. II, Comm. Pure Appl. Math., 17 (1964), pp. 35–92.
[4] S. J. Chapman and D. R. Heron, *A hierarchy of models for superconducting thin films*, SIAM J. Appl. Math., 63 (2003), pp. 2087–2127.
[5] E. B. Davies, *Wild spectral behaviour of anharmonic oscillators*, Bull. London Math. Soc., 32 (2000), pp. 432–438.
[6] Q. Du and P. Gray, *High-kappa limits of the time-dependent Ginzburg–Landau model*, SIAM J. Appl. Math., 56 (1996), pp. 1060–1093.
[7] L. P. Gor'kov and G. M. Éliashberg, *Generalisation of the Ginzburg-Landau equations for non-stationary problems in the case of alloys with paramagnetic impurities*, Soviet Phys. JETP, 27 (1968), p. 328.
[8] B. I. Ivlev and N. B. Kopnin, *Electric currents and resistive states in thin superconductors*, Adv. in Phys., 33 (1984), pp. 47–114.
[9] B. I. Ivlev, N. B. Kopnin, and L. A. Maslova, *Stability of current-carrying states in narrow finite-length superconducting channels*, Soviet Phys. JETP, 56 (1982), pp. 884–890.
[10] P. T. Lại and D. Robert, *Sur un problème aux valeurs propres non linéaire*, Israel J. Math., 36 (1980), pp. 169–186.
[11] X.-B. Pan, *Surface superconductivity in* 3 *dimensions*, Trans. Amer. Math. Soc., 356 (2004), pp. 3899–3937.
[12] D. Robert, *Non-linear eigenvalue problems*, Mat. Contemp., 26 (2004), pp. 109–127.
[13] Y. Sibuya, *Global Theory of a Second Order Linear Ordinary Differential Equation with a Polynomial Coefficient*, North–Holand, Amsterdam, 1975.
[14] A. G. Sivakov, A. M. Glukhov, A. N. Omelyanchouk, Y. Koval, P. Müller, and A. V. Ustinov, *Josephson behavior of phase-slip lines in wide superconducting strips*, Phys. Rev. Lett., 91 (2003), 267001.
[15] D. Y. Vodolazov, F. M. Peeters, L. Piraux, S. Matefi-Tempfli, and S. Michotte, *Current-voltage characteristics of quasi-one-dimensional superconductors: An s-shaped curve in the constant voltage regime*, Phys. Rev. Lett., 91 (2003), 157001.

SIAM J. MATH. ANAL.
Vol. 40, No. 2, pp. 851–868

CARLEMAN AND OBSERVABILITY ESTIMATES FOR STOCHASTIC WAVE EQUATIONS*

XU ZHANG†

Abstract. Based on a fundamental identity for stochastic hyperbolic-like operators, we derive in this paper a global Carleman estimate (with singular weight function) for stochastic wave equations. This leads to an observability estimate for stochastic wave equations with nonsmooth lower order terms. Moreover, the observability constant is estimated by means of an explicit function of the norm of the coefficients involved in the equation. An application to the state observation problem for semilinear stochastic wave equations is also given.

Key words. Carleman estimate, singular weight function, observability estimate, stochastic wave equation, state observation problem

AMS subject classifications. Primary, 60H15; Secondary, 93B07, 35B45

DOI. 10.1137/070685786

1. Introduction.

Let $T > 0$, $G \subset \mathbb{R}^n$ ($n \in \mathbb{N}$) be a given bounded domain with a C^2 boundary Γ, with Γ_0 a given nonempty open subset of Γ. Put $Q \triangleq (0,T) \times G$, $\Sigma \triangleq (0,T) \times \Gamma$, and $\Sigma_0 \triangleq (0,T) \times \Gamma_0$. Throughout this paper, we will use C to denote a generic positive constant depending only on T, G, and Γ_0, which may change from line to line.

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, P)$ be a complete filtered probability space on which a one dimensional standard Brownian motion $\{B(t)\}_{t\geq 0}$ is defined. Let H be a Banach space. We denote by $L^2_{\mathcal{F}}(0,T;H)$ the Banach space consisting of all H-valued $\{\mathcal{F}_t\}_{t\geq 0}$-adapted processes $X(\cdot)$ such that $\mathbb{E}(|X(\cdot)|^2_{L^2(0,T;H)}) < \infty$, with the canonical norm; by $L^\infty_{\mathcal{F}}(0,T;H)$ the Banach space consisting of all H-valued $\{\mathcal{F}_t\}_{t\geq 0}$-adapted bounded processes; and by $L^2_{\mathcal{F}}(\Omega; C([0,T];H))$ the Banach space consisting of all H-valued $\{\mathcal{F}_t\}_{t\geq 0}$-adapted continuous processes $X(\cdot)$ such that $\mathbb{E}(|X(\cdot)|^2_{C([0,T];H)}) < \infty$, with the canonical norm (similarly, one can define $L^2_{\mathcal{F}}(\Omega; C^k([0,T];H))$ for any positive integer k).

Let us consider the following stochastic wave equation:

$$\begin{cases} dy_t - \Delta y dt = (a_1 y_t + a_2 \cdot \nabla y + a_3 y + f)dt + (a_4 y + g)dB(t) & \text{in } Q, \\ y = 0 & \text{on } \Sigma, \\ y(0) = y_0, \quad y_t(0) = y_1 & \text{in } G, \end{cases} \tag{1.1}$$

with initial data $(y_0, y_1) \in L^2(\Omega, \mathcal{F}_0, P; H^1_0(G) \times L^2(G))$, suitable coefficients a_i ($i = 1,2,3,4$), and source terms f and g. Here, $y_t = \frac{dy}{dt}$.

*Received by the editors March 20, 2007; accepted for publication (in revised form) April 3, 2008; published electronically August 27, 2008. This work was supported by the NSF of China under grant 10525105, grant MTM2005-00714 of the Spanish MEC, and the SIMUMAT projet of the CAM (Spain). Part of this work was done when the author visited the Departamento de Matemáticas at Universidad Autónoma de Madrid and the Shanghai Key Laboratory for Contemporary Applied Mathematics at Fudan University.

http://www.siam.org/journals/sima/40-2/68578.html

†Key Laboratory of Systems and Control, Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences, Beijing 100080, China, and Yangtze Center of Mathematics, Sichuan University, Chengdu 610064, China (xuzhang@amss.ac.cn).

Put

$$\mathcal{H}_T \stackrel{\triangle}{=} L^2_{\mathcal{F}}(\Omega; C([0,T]; H_0^1(G))) \bigcap L^2_{\mathcal{F}}(\Omega; C^1([0,T]; L^2(G))). \tag{1.2}$$

Clearly, $\mathcal{H}_T$ is a Banach space with the canonical norm. We begin with the following notion.

DEFINITION 1.1. *We call $y \in \mathcal{H}_T$ a solution of* (1.1) *if the following hold:*
(i) $y(0) = y_0$ *in* G, P*-a.s.*
(ii) *For any $t \in [0,T]$ and any $\eta \in H_0^1(G)$, it holds that*

$$\begin{aligned}
&\int_G y_t(t,x)\eta(x)dx - \int_G y_t(0,x)\eta(x)dx \\
&= \int_0^t \int_G \Big\{ -\nabla y(s,x)\cdot\nabla\eta(x) + \Big[a_1(s,x)y_t(s,x) + a_2(s,x)\cdot\nabla y(s,x) \\
&\qquad + a_3(s,x)y(s,x) + f(s,x)\Big]\eta(x)\Big\}dxds \\
&\quad + \int_0^t \int_G \Big[a_4(s,x)y(s,x) + g(s,x)\Big]\eta(x)dxdB(s), \qquad P\text{-a.s.}
\end{aligned} \tag{1.3}$$

Under some assumptions, for any initial data $(y_0, y_1) \in L^2(\Omega, \mathcal{F}_0, P;\ H_0^1(G) \times L^2(G))$, one can show that system (1.1) admits one and only one solution $y \in \mathcal{H}_T$ (see Proposition 3.1 in section 3).

The main purpose of this paper is to derive a (partial) boundary observability estimate for system (1.1), i.e., find (if possible) a constant $\mathcal{C}(a_1, a_2, a_3, a_4) > 0$ such that solutions of system (1.1) satisfy

$$\begin{aligned}
&|(y(T), y_t(T))|_{L^2(\Omega,\mathcal{F}_T,P;H_0^1(G)\times L^2(G))} \\
&\le \mathcal{C}(a_1,a_2,a_3,a_4)\left[\left|\frac{\partial y}{\partial \nu}\right|_{L^2_{\mathcal{F}}(0,T;L^2(\Gamma_0))} + |f|_{L^2_{\mathcal{F}}(0,T;L^2(G))} + |g|_{L^2_{\mathcal{F}}(0,T;L^2(G))}\right] \\
&\qquad \forall\ (y_0,y_1) \in L^2(\Omega,\mathcal{F}_0,P;H_0^1(G)\times L^2(G)).
\end{aligned} \tag{1.4}$$

As we shall see in the last section of this paper, inequality (1.4) is strongly related to the state observation problem of semilinear stochastic wave equations.

It is well known that the observability estimate is an important tool for the study of stabilization and controllability problems for deterministic PDEs. We refer the reader to [16] for a recent survey in this respect. Although there are numerous references addressing the observability problems for deterministic PDEs, very little is known for the stochastic counterpart, and it remains to be further understood. Indeed, to the best of our knowledge, [1] is the only publication in this direction, which is devoted to the controllability/observability for the stochastic heat equation. As far as we know, nothing is known for the observability estimate on the stochastic wave equation.

Similar to the deterministic setting, we shall use a stochastic version of the global Carleman estimate to establish inequality (1.4). The difficulty in doing this is the very fact that, unlike the deterministic situation, system (1.1), a stochastic wave equation, is *time-irreversible*. Therefore, one cannot simply mimic the usual Carleman inequality for the deterministic wave equations (see [3, 4, 5, 8, 9, 13] and the references cited therein). Rather, in order to overcome this difficulty, instead of the usual smooth

weight function, we need to introduce another singular weight function in this paper to derive the desired Carleman estimate for system (1.1) (see (2.13) in the next section).

On the other hand, the Carleman estimate is itself a fundamental tool for the study of control and inverse problems for deterministic PDEs (see [7, 16] and the references cited therein). Similar to the situation for the observability estimate, although there are numerous references addressing the Carleman estimate for deterministic PDEs, to the best of our knowledge, [1, 12] are the only references for the stochastic counterpart, which are devoted to the stochastic heat equation. It would be quite interesting to extend the deterministic Carleman estimate for other PDEs to the stochastic ones, but there are many things to be done, and some of them seem to be challenging. In this paper, in order to present the key idea in the simplest way, we do not pursue the full technical generality.

The rest of this paper is organized as follows. The main results of this paper are stated in section 2. In section 3, we show some preliminary results. In section 4, we present a crucial identity for a stochastic hyperbolic-like operator. Then, in section 5, we derive pointwise Carleman-type estimates for the stochastic wave operator. Section 6 is devoted to the proof of Theorems 2.1–2.2. Finally, section 7 addresses giving an application of our observability result.

2. Statement of the main results. Fix any $x_0 \in \mathbb{R}^d \setminus \overline{G}$. It is clear that

$$0 < R_0 \triangleq \min_{x\in G} |x - x_0| < R_1 \triangleq \max_{x\in G} |x - x_0|. \tag{2.1}$$

Put

$$\Gamma_0 \triangleq \big\{x \in \Gamma \,\big|\, (x - x_0) \cdot \nu(x) > 0\big\}, \tag{2.2}$$

where $\nu(x)$ is the unit outward normal vector of G at $x \in \Gamma$.

Assume that

$$\begin{aligned} &a_1 \in L^\infty_{\mathcal{F}}(0,T;L^\infty(G)), \qquad && a_2 \in L^\infty_{\mathcal{F}}(0,T;L^\infty(G;\mathbb{R}^n)),\\ &a_3 \in L^\infty_{\mathcal{F}}(0,T;L^n(G)), && a_4 \in L^\infty_{\mathcal{F}}(0,T;L^\infty(G)) \end{aligned} \tag{2.3}$$

and that

$$f \in L^2_{\mathcal{F}}(0,T;L^2(G)), \qquad g \in L^2_{\mathcal{F}}(0,T;L^2(G)). \tag{2.4}$$

In what follows, we use the following notation:

$$\begin{aligned} \mathcal{A}(a_1,a_2,a_3,a_4) \triangleq\ & |(a_1,a_4)|^2_{L^\infty_{\mathcal{F}}(0,T;(L^\infty(G))^2)} + |a_2|^2_{L^\infty_{\mathcal{F}}(0,T;L^\infty(G;\mathbb{R}^n))}\\ &+ |a_3|^2_{L^\infty_{\mathcal{F}}(0,T;L^n(G))}. \end{aligned} \tag{2.5}$$

We choose a sufficiently small constant $c \in (0,1)$ so that (recall (2.1) for R_0 and R_1)

$$\frac{(4+5c)R_0^2}{9c} > R_1^2. \tag{2.6}$$

In what follows, we take $T(> 2R_1)$ sufficiently large such that

$$\frac{4(4+5c)R_0^2}{9c} > c^2T^2 > 4R_1^2. \tag{2.7}$$

Our observability estimate for system (1.1) is stated as follows.

THEOREM 2.1. *Let* (2.3)–(2.4) *hold,* Γ_0 *be given by* (2.2), *and* T *satisfy* (2.7). *Then solutions of system* (1.1) *satisfy* (1.4) *with*

$$\mathcal{C}(a_1,a_2,a_3,a_4)=Ce^{C\mathcal{A}(a_1,a_2,a_3,a_4)}. \tag{2.8}$$

Remark 2.1. From the proof of Theorem 2.1, it is easy to see that the conclusion can be slightly strengthened as follows: for any $t\in(0,T]$, solutions of system (1.1) satisfy

(2.9)

$$\begin{aligned}&|(y(t),y_t(t))|_{L^2(\Omega,\mathcal{F}_t,P;H_0^1(G)\times L^2(G))}\\&\le e^{Ct^{-1}}\mathcal{C}(a_1,a_2,a_3,a_4)\left[\left|\frac{\partial y}{\partial\nu}\right|_{L^2_{\mathcal{F}}(0,T;L^2(\Gamma_0))}+|f|_{L^2_{\mathcal{F}}(0,T;L^2(G))}+|g|_{L^2_{\mathcal{F}}(0,T;L^2(G))}\right]\\&\qquad\forall\ (y_0,y_1)\in L^2(\Omega,\mathcal{F}_0,P;H_0^1(G)\times L^2(G)),\end{aligned}$$

where $\mathcal{C}(a_1,a_2,a_3,a_4)$ is given by (2.8).

Remark 2.2. A deterministic version of (1.1) reads

$$\begin{cases} w_{tt}-\Delta w=b_1w_t+b_2\cdot\nabla w+b_3w+h & \text{in } Q,\\ w=0 & \text{on } \Sigma,\\ w(0)=w_0,\quad w_t(0)=w_1 & \text{in } G,\end{cases} \tag{2.10}$$

where $b_1\in L^\infty(Q)$, $b_2\in L^\infty(Q;\mathbb{R}^n)$, $b_3\in L^\infty(0,T;L^p(G))$ with $p\in[n,\infty]$, and $h\in L^2(Q)$. As a special case of [3, Theorems 2.2 and 2.3] and noting the time reversibility of system (2.10), the following counterpart of Theorem 2.1 holds: If $T>2R_1$, then solutions of (2.10) satisfy

$$\begin{aligned}&|(w(T),w_t(T))|_{H_0^1(G)\times L^2(G)}\\&\le Ce^{C\left[|b_1|^2_{L^\infty(Q)}+|b_2|^2_{L^\infty(Q;\mathbb{R}^n)}+|b_3|^{\frac{1}{3/2-n/p}}_{L^\infty(0,T;L^p(G))}\right]}\left[\left|\frac{\partial w}{\partial\nu}\right|_{L^2(\Sigma_0)}+|h|_{L^2(Q)}\right]\\&\qquad\forall\ (w_0,w_1)\in H_0^1(G)\times L^2(G).\end{aligned} \tag{2.11}$$

There are three main differences between Theorem 2.1 and this result. The first is that one can replace the left-hand side of (2.11) by $|(w_0,w_1)|_{H_0^1(G)\times L^2(G)}$. However, due to the time irreversibility of system (1.1), one cannot do the same in the stochastic setting, i.e., replacing the left-hand side of (1.4) by $|(y_0,y_1)|_{L^2(\Omega,\mathcal{F}_0,P;H_0^1(G)\times L^2(G))}$. The second is that we assume that T in Theorem 2.1 satisfies (2.7), which is usually much more restrictive than that $T>2R_1$ for the deterministic setting (see Remark 2.4 below for more explanation about this). The third is that the observability constant $\mathcal{C}(a_1,a_2,a_3,a_4)$ in Theorem 2.1 is not as sharp as that in (2.11). Indeed, it is clear that (recall (2.5) for $\mathcal{A}(\cdot,\cdot,\cdot,\cdot)$)

$$|b_1|^2_{L^\infty(Q)}+|b_2|^2_{L^\infty(Q;\mathbb{R}^n)}+|b_3|^{\frac{1}{3/2-n/p}}_{L^\infty(0,T;L^p(G))}\le\mathcal{A}(b_1,b_2,b_3,0)+C\qquad\forall\ p\in[n,\infty].$$

Remark 2.3. It is well known that a sharp condition guaranteeing observability inequality (2.11) (at least when b_1, b_2, and b_3 are time-invariant) is that the triple

(G, Γ_0, T) satisfies the geometric optic condition introduced in [2]. It would be quite interesting to extend this result to the stochastic setting, but this is an open problem.

As mentioned in section 1, in order to prove Theorem 2.1, we need to derive a Carleman estimate (with singular weight function) for system (1.1). For this purpose, for any (large) $\lambda > 0$ and $c \in (0,1)$ given in (2.6), set

$$\ell = \ell(t,x) \triangleq \lambda \left[|x - x_0|^2 - c\left(t - \frac{T}{2}\right)^2 \right], \qquad \theta \triangleq e^{\ell}. \tag{2.12}$$

Also, for any $\beta > 0$, set

$$\Theta = \Theta(t) \triangleq \exp\left\{ -\frac{\beta}{t(T-t)} \right\}, \qquad 0 < t < T. \tag{2.13}$$

It is easy to see that $\Theta(t)$ decays rapidly to 0 as $t \to 0$ or $t \to T$. Our Carleman estimate for system (1.1) is stated as follows.

THEOREM 2.2. *Let* (2.3)–(2.4) *hold,* Γ_0 *be given by* (2.2), *and* c *and* T *satisfy, respectively,* (2.6) *and* (2.7). *Then there exist a constant* $\beta > 0$ *(which is very small) and a constant*

$$\lambda^* = C\Big[1 + \mathcal{A}(a_1, a_2, a_3, a_4)\Big]$$

such that solutions of system (1.1) *satisfy*

$$\begin{aligned} &\lambda \mathbb{E} \int_Q \Theta\theta^2 (y_t^2 + |\nabla y|^2 + \lambda^2 y^2) dxdt \\ &\le C\mathbb{E}\left\{ \lambda \int_{\Sigma_0} \Theta\theta^2 \left| \frac{\partial y}{\partial \nu} \right|^2 d\Sigma_0 + \int_Q \Theta\theta^2 (f^2 + \lambda g^2) dxdt \right\} \\ &\qquad \forall\, (y_0, y_1) \in L^2(\Omega, \mathcal{F}_0, P; H_0^1(G) \times L^2(G)), \quad \forall\, \lambda \ge \lambda^*. \end{aligned} \tag{2.14}$$

Remark 2.4. The restriction on T in (2.7) is a technical condition, which does not seem to be natural. However, this condition plays a key role in Step 3 in the proof of Theorem 5.1, a crucial preliminary for the proof of Theorem 2.2. As in the deterministic setting (recall Remark 2.2), it is reasonable to expect that it should be improved to be $T > 2R_1$. But this is an open problem.

Remark 2.5. We now recall the Carleman estimate for the deterministic wave equation, i.e., system (2.10). Fix a constant $c \in (0,1)$ so that $\left(\frac{2R_1}{T}\right)^2 < c < \frac{2R_1}{T}$ (which is possible since $T > 2R_1$ for the deterministic situation). By [5, equation (11.12) in the proof of Theorem 5.1] and similar to [3, Theorem 2.3], we conclude that there exists a constant

$$\lambda_* = C\Big[1 + |b_1|^2_{L^\infty(Q)} + |b_2|^2_{L^\infty(Q;\mathbb{R}^n)} + |b_3|^{\frac{1}{3/2 - n/p}}_{L^\infty(0,T;L^p(G))}\Big]$$

such that solutions of system (2.10) satisfy

$$\begin{aligned} &\lambda \int_Q \theta^2 (\lambda^2 w^2 + w_t^2 + |\nabla w|^2) dxdt \le C\left[\lambda \int_{\Sigma_0} \theta^2 \Big| \frac{\partial w}{\partial \nu} \Big|^2 d\Sigma_0 + \int_Q \theta^2 |h|^2 dxdt \right] \\ &\qquad \forall\, (w_0, w_1) \in H_0^1(G) \times L^2(G), \quad \forall\, \lambda \ge \lambda_*. \end{aligned} \tag{2.15}$$

The main difference between Theorem 2.2 and this result is, as mentioned before, one has to introduce a singular weight Θ in (2.14).

Remark 2.6. We consider here the simplest case of one dimensional standard Brownian motion. It would be interesting to extend the results in this paper to the case of colored (infinite dimensional) noise, or even with both state- and control-dependent noise. But these remain to be done.

Remark 2.7. It would be quite interesting to study Carleman and observability estimates for backward stochastic wave equations. To the best of our knowledge, this is a challenging problem, and nothing is known in this respect.

3. Preliminaries. In this section, we show some preliminary results that will subsequently be used.

In what follows, for simplicity, we denote $\sum_{i,j=1}^n$ and $\sum_{i=1}^n$ simply by $\sum_{i,j}$ and $\sum_i$, respectively. Also, we will use the notation $u_i = u_{x_i}$, where x_i is the ith coordinate of a generic point $x = (x_1, \dots, x_n)$ in $\mathbb{R}^n$. In a similar manner, we use the notation ℓ_i, v_i, etc., for the partial derivatives of ℓ and v with respect to x_i.

First of all, we have the following well-posedness result for system (1.1).

PROPOSITION 3.1. *Under assumptions* (2.3)–(2.4), *for any* $(y_0, y_1) \in L^2(\Omega, \mathcal{F}_0, P; H_0^1(G) \times L^2(G))$, *system* (1.1) *admits one and only one solution* $y \in \mathcal{H}_T$. *Moreover (recall* (2.8) *for* $\mathcal{C}(a_1, a_2, a_3, a_4)$*),*

$$
\begin{aligned}
|y|_{\mathcal{H}_T} \le \mathcal{C}(a_1, a_2, a_3, a_4) \Big[& |(y_0, y_1)|_{L^2(\Omega, \mathcal{F}_0, P; H_0^1(G) \times L^2(G))} + |f|_{L^2_{\mathcal{F}}(0,T; L^2(G))} \\
& + |g|_{L^2_{\mathcal{F}}(0,T; L^2(G))} \Big].
\end{aligned} \tag{3.1}
$$

Proof. The detailed proof is lengthy but almost standard. Therefore, we give below only a sketch.

Step 1. First of all, following the *parabolic regularization approach* in the proof of [6, Theorem 4.1, p. 159] (for this, one needs to make numerous but small changes), one can show that, for any $(w_0, w_1) \in L^2(\Omega, \mathcal{F}_0, P; H_0^1(G) \times L^2(G))$, $\hat{f} \in L^2_{\mathcal{F}}(0, T; L^2(G))$, and $\hat{g} \in L^2_{\mathcal{F}}(0, T; L^2(G))$, the system

$$
\begin{cases}
dw_t - \Delta w dt = \hat{f} dt + \hat{g} dB(t) & \text{in } Q, \\
w = 0 & \text{on } \Sigma, \\
w(0) = w_0, \quad w_t(0) = w_1 & \text{in } G
\end{cases} \tag{3.2}
$$

admits a unique solution $w \in \mathcal{H}_T$. Furthermore, there is a constant $C = C(T) > 0$ such that, for any $\tau \in [0, T]$, it holds that

$$
|w|_{\mathcal{H}_\tau} \le C \Big[|(w_0, w_1)|_{L^2(\Omega, \mathcal{F}_0, P; H_0^1(G) \times L^2(G))} + |\hat{f}|_{L^2_{\mathcal{F}}(0,\tau; L^2(G))} + |\hat{g}|_{L^2_{\mathcal{F}}(0,\tau; L^2(G))} \Big], \tag{3.3}
$$

where $\mathcal{H}_\tau$ is defined similarly to (1.2).

Step 2. Next, we show that system (1.1) admits a local (in time) solution $y \in \mathcal{H}_\tau$ for small $\tau \in (0, T]$. For this purpose, fixing any $y \in \mathcal{H}_\tau$, we solve the following system:

$$
\begin{cases}
dz_t - \Delta z dt = (a_1 y_t + a_2 \cdot \nabla y + a_3 y + f) dt + (a_4 y + g) dB(t) & \text{in } Q, \\
z = 0 & \text{on } \Sigma, \\
z(0) = y_0, \quad z_t(0) = y_1 & \text{in } G.
\end{cases} \tag{3.4}
$$

By Step 1, system (3.4) (by viewing, respectively, $a_1y_t+a_2\cdot\nabla y+a_3y+f$ and a_4y+g as nonhomogeneous terms $\hat{f}$ and $\hat{g}$) admits a unique solution $z\in\mathcal{H}_\tau$. We define a map $\mathcal{F}$ from $\mathcal{H}_\tau$ into itself by

$$\mathcal{F}(y)=z.$$

For any $y^1,y^2\in\mathcal{H}_\tau$, by means of (3.3), it is easy to verify that

$$|\mathcal{F}(y^1)-\mathcal{F}(y^2)|_{\mathcal{H}_\tau}\le C\tau|y^1-y^2|_{\mathcal{H}_\tau}.$$

This shows that the map $\mathcal{F}$ is contractive when τ is small enough. Therefore, $\mathcal{F}$ admits a fixed point $y\in\mathcal{H}_\tau$. Hence, system (1.1) admits a local (in time) solution $y\in\mathcal{H}_\tau$.

Step 3. To conclude the proof of Proposition 3.1, it suffices to show that estimate (3.1) holds.

For any $\tau\in[0,T]$, applying estimate (3.3) to system (1.1) (by viewing, respectively, $a_1y_t+a_2\cdot\nabla y+a_3y+f$ and a_4y+g as nonhomogeneous terms $\hat{f}$ and $\hat{g}$), we obtain (recall (2.5) for $\mathcal{A}(a_1,a_2,a_3,a_4)$)

(3.5)

$$\begin{aligned}|y|^2_{\mathcal{H}_\tau}&\le C\Big[|(y_0,y_1)|^2_{L^2(\Omega,\mathcal{F}_0,P;H_0^1(G)\times L^2(G))}\\&\qquad+|a_1y_t+a_2\cdot\nabla y+a_3y+f|^2_{L^2_{\mathcal{F}}(0,\tau;L^2(G))}+|a_4y+g|^2_{L^2_{\mathcal{F}}(0,\tau;L^2(G))}\Big]\\&\le C\bigg[|(y_0,y_1)|^2_{L^2(\Omega,\mathcal{F}_0,P;H_0^1(G)\times L^2(G))}+|f|^2_{L^2_{\mathcal{F}}(0,\tau;L^2(G))}+|g|^2_{L^2_{\mathcal{F}}(0,\tau;L^2(G))}\\&\qquad+\mathcal{A}(a_1,a_2,a_3,a_4)\int_0^\tau|y|^2_{\mathcal{H}_s}ds\bigg].\end{aligned}$$

Now the desired result (3.1) follows from (3.5) and Gronwall's inequality. □

Next, we show the following identity.

PROPOSITION 3.2. *Let $\mu=\mu(x)\overset{\triangle}{=}(\mu^1,\dots,\mu^n):\mathbb{R}^n\to\mathbb{R}^n$ be a vector field of class C^1 and w an $H^2_{loc}(\mathbb{R}^n)$-valued $\{\mathcal{F}_t\}_{t\ge0}$-adapted process such that w_t is an $L^2_{loc}(\mathbb{R}^n)$-valued semimartingale. Then, for a.e. $x\in\mathbb{R}^n$ and P-a.s. $\omega\in\Omega$, it holds that*

$$\begin{aligned}&-\nabla\cdot\left[2(\mu\cdot\nabla w)\nabla w+\mu\left(w_t^2-|\nabla w|^2\right)\right]dt\\&=2\left[\mu\cdot\nabla w(dw_t-\Delta w dt)-d(w_t\mu\cdot\nabla w)-\sum_{i,j}w_iw_j\frac{\partial\mu^j}{\partial x_i}dt\right]\\&\quad+(|\nabla w|^2-w_t^2)\nabla\cdot\mu dt.\end{aligned}\tag{3.6}$$

Proof. Noting that μ is time-independent, it follows that

$$2\left[\mu\cdot\nabla w\Delta w dt+d(w_t\mu\cdot\nabla w)\right]=2\left[\sum_{i,j}\mu^j w_j w_{ii}dt+\mu\cdot\nabla w dw_t+w_t\mu\cdot\nabla w_t dt\right]$$

$$=2\sum_{i,j}\left[\left(w_i\mu^j w_j\right)_i-w_i w_j\frac{\partial\mu^j}{\partial x_i}\right]dt-\mu\cdot\nabla|\nabla w|^2dt+2\mu\cdot\nabla w dw_t+\mu\cdot\nabla w_t^2dt$$

$$=\nabla\cdot\left[2(\mu\cdot\nabla w)\nabla w+\mu\left(w_t^2-|\nabla w|^2\right)\right]dt+2\left[\mu\cdot\nabla w dw_t-\sum_{i,j}w_i w_j\frac{\partial\mu^j}{\partial x_i}dt\right]$$

$$+(|\nabla w|^2-w_t^2)\nabla\cdot\mu dt.$$

This gives (3.6). □

Finally, we show the following hidden regularity for the solution of system (1.1) (here, by hidden regularity we mean that it follows from the equation rather than from the usual trace theorem in the theory of Sobolev spaces).

PROPOSITION 3.3. *Under assumptions* (2.3)–(2.4), *for any* $(y_0,y_1)\in L^2(\Omega,\mathcal{F}_0,P;H_0^1(G)\times L^2(G))$, *the solution of system* (1.1) *satisfies* $\frac{\partial y}{\partial\nu}\in L^2_{\mathcal{F}}(0,T;L^2(\Gamma))$. *Moreover,*

(3.7)

$$\left|\frac{\partial y}{\partial\nu}\right|_{L^2_{\mathcal{F}}(0,T;L^2(\Gamma))}$$

$$\le C\Big[|(y_0,y_1)|_{L^2(\Omega,\mathcal{F}_0,P;H_0^1(G)\times L^2(G))}+|f|_{L^2_{\mathcal{F}}(0,T;L^2(G))}+|g|_{L^2_{\mathcal{F}}(0,T;L^2(G))}\Big]$$

$$\times\exp\Big\{C\Big[|(a_1,a_4)|^2_{L^\infty_{\mathcal{F}}(0,T;(L^\infty(G))^2)}+|a_2|^2_{L^\infty_{\mathcal{F}}(0,T;L^\infty(G;\mathbb{R}^n))}+|a_3|^2_{L^\infty_{\mathcal{F}}(0,T;L^n(G))}\Big]\Big\}.$$

Proof. Since $\Gamma\in C^2$, one can find a vector field $\mu_0=(\mu_0^1,\dots,\mu_0^n)\in C^1(\overline{\Omega};\mathbb{R}^n)$ such that $\mu_0=\nu$ on Γ (see [10]). Applying Proposition 3.2 with $\mu=\mu_0$ and $w=y$, by means of Proposition 3.1, following [10, 13], it is not difficult to show $\frac{\partial y}{\partial\nu}\in L^2_{\mathcal{F}}(0,T;L^2(\Gamma))$ and the desired estimate (3.7) (hence we omit the details). □

4. Identity for a stochastic hyperbolic-like operator. In this section, we show the following fundamental identity for a stochastic hyperbolic-like operator.

THEOREM 4.1. *Let* $b^{ij}\in C^1((0,T)\times\mathbb{R}^n)$ *satisfy*

$$b^{ij}=b^{ji},\qquad i,j=1,2,\dots,n, \tag{4.1}$$

$\ell,\ \Psi\in C^2((0,T)\times\mathbb{R}^n)$. *Assume* u *is an* $H^2_{loc}(\mathbb{R}^n)$*-valued* $\{\mathcal{F}_t\}_{t\ge0}$*-adapted process such that* u_t *is an* $L^2_{loc}(\mathbb{R}^n)$*-valued semimartingale. Set* $\theta=e^\ell$ *and* $v=\theta u$. *Then, for*

a.e. $x \in \mathbb{R}^n$ *and* P*-a.s.* $\omega \in \Omega$,

$$
\begin{aligned}
&\theta\left(-2\ell_t v_t + 2\sum_{i,j} b^{ij}\ell_i v_j + \Psi v\right)\left[du_t - \sum_{i,j}(b^{ij}u_i)_j dt\right] \\
&\quad + \sum_{i,j}\Bigg[\sum_{i',j'}\left(2b^{ij}b^{i'j'}\ell_{i'}v_i v_{j'} - b^{ij}b^{i'j'}\ell_i v_{i'}v_{j'}\right) - 2b^{ij}\ell_t v_i v_t + b^{ij}\ell_i v_t^2 \\
&\qquad\qquad + \Psi b^{ij}v_i v - \left(A\ell_i + \frac{\Psi_i}{2}\right)b^{ij}v^2\Bigg]_j dt \\
&\quad + d\left[\sum_{i,j} b^{ij}\ell_t v_i v_j - 2\sum_{i,j} b^{ij}\ell_i v_j v_t + \ell_t v_t^2 - \Psi v_t v + \left(A\ell_t + \frac{\Psi_t}{2}\right)v^2\right] \\
&= \Bigg\{\left[\ell_{tt} + \sum_{i,j}(b^{ij}\ell_i)_j - \Psi\right]v_t^2 - 2\sum_{i,j}\left[(b^{ij}\ell_j)_t + b^{ij}\ell_{tj}\right]v_i v_t \\
&\quad + \sum_{i,j}\left[(b^{ij}\ell_t)_t + \sum_{i',j'}\left(2b^{ij'}(b^{i'j}\ell_{i'})_{j'} - (b^{ij}b^{i'j'}\ell_{i'})_{j'}\right) + \Psi b^{ij}\right]v_i v_j \\
&\quad + Bv^2 + \left(-2\ell_t v_t + 2\sum_{i,j} b^{ij}\ell_i v_j + \Psi v\right)^2\Bigg\}dt + \theta^2\ell_t(du_t)^2,
\end{aligned}
\tag{4.2}
$$

where $(du_t)^2$ *denotes the quadratic variation process of* u_t,

$$
\begin{cases}
A \triangleq (\ell_t^2 - \ell_{tt}) - \sum_{i,j}(b^{ij}\ell_i\ell_j - b^{ij}_j\ell_i - b^{ij}\ell_{ij}) - \Psi, \\
B \triangleq A\Psi + (A\ell_t)_t - \sum_{i,j}(Ab^{ij}\ell_i)_j + \frac{1}{2}\left[\Psi_{tt} - \sum_{i,j}(b^{ij}\Psi_i)_j\right].
\end{cases}
\tag{4.3}
$$

Proof. By $v(t,x) = \theta(t,x)u(t,x)$, we have $u_t = \theta^{-1}(v_t - \ell_t v)$ and $u_j = \theta^{-1}(v_j - \ell_j v)$ for $j = 1, 2, \ldots, n$. Hence,

$$
du_t = \theta^{-1}[dv_t - 2\ell_t v_t dt + (\ell_t^2 - \ell_{tt})v dt]. \tag{4.4}
$$

Similarly, by symmetry condition (4.1), one may check that

$$
\sum_{i,j}(b^{ij}u_i)_j = \theta^{-1}\sum_{i,j}\left[(b^{ij}v_i)_j - 2b^{ij}\ell_i v_j + (b^{ij}\ell_i\ell_j - b^{ij}_j\ell_i - b^{ij}\ell_{ij})v\right]. \tag{4.5}
$$

Therefore, by (4.4)–(4.5) and recalling the definition of A in (4.3), we get

$$
\begin{aligned}
&\theta\left(-2\ell_t v_t+2\sum_{i,j}b^{ij}\ell_i v_j+\Psi v\right)\left[du_t-\sum_{i,j}(b^{ij}u_i)_j dt\right]\\
&=\left(-2\ell_t v_t+2\sum_{i,j}b^{ij}\ell_i v_j+\Psi v\right)\left\{dv_t-\left[\sum_{i,j}(b^{ij}v_i)_j-Av\right.\right.\\
&\qquad\left.\left.+2\ell_t v_t-2\sum_{i,j}b^{ij}\ell_i v_j-\Psi v\right]dt\right\}\\
&=\left(-2\ell_t v_t+2\sum_{i,j}b^{ij}\ell_i v_j+\Psi v\right)dv_t\\
&\quad+\left(-2\ell_t v_t+2\sum_{i,j}b^{ij}\ell_i v_j+\Psi v\right)\left[-\sum_{i,j}(b^{ij}v_i)_j+Av\right]dt\\
&\quad+\left(-2\ell_t v_t+2\sum_{i,j}b^{ij}\ell_i v_j+\Psi v\right)^2 dt.
\end{aligned}
\tag{4.6}
$$

We now analyze the first two terms in the right-hand side of (4.6). First, using Itô's formula, we have

$$
\begin{aligned}
&\left(-2\ell_t v_t+2\sum_{i,j}b^{ij}\ell_i v_j+\Psi v\right)dv_t\\
&=d\left[\left(-\ell_t v_t+2\sum_{i,j}b^{ij}\ell_i v_j+\Psi v\right)v_t\right]\\
&\quad-\left[-\ell_{tt}v_t^2+2\sum_{i,j}(b^{ij}\ell_i)_t v_j v_t+2\sum_{i,j}b^{ij}\ell_i v_{tj}v_t+\Psi v_t^2+\Psi_t v v_t\right]dt+\ell_t(dv_t)^2\\
&=d\left(-\ell_t v_t^2+2\sum_{i,j}b^{ij}\ell_i v_j v_t+\Psi v v_t-\frac{\Psi_t}{2}v^2\right)\\
&\quad+\left\{-\sum_{i,j}(b^{ij}\ell_i v_t^2)_j+\left[\ell_{tt}+\sum_{i,j}(b^{ij}\ell_i)_j-\Psi\right]v_t^2-2\sum_{i,j}(b^{ij}\ell_j)_t v_i v_t+\frac{\Psi_{tt}}{2}v^2\right\}dt\\
&\quad+\theta^2\ell_t(du_t)^2.
\end{aligned}
\tag{4.7}
$$

Next,

$$
\begin{aligned}
&-2\ell_t v_t\left[-\sum_{i,j}(b^{ij}v_i)_j+Av\right]\\
&=2\left[\sum_{i,j}(b^{ij}\ell_t v_i v_t)_j-\sum_{i,j}b^{ij}\ell_{tj}v_i v_t\right]-\sum_{i,j}b^{ij}\ell_t(v_i v_j)_t-A\ell_t(v^2)_t\\
&=2\left[\sum_{i,j}(b^{ij}\ell_t v_i v_t)_j-\sum_{i,j}b^{ij}\ell_{tj}v_i v_t\right]+\sum_{i,j}(b^{ij}\ell_t)_t v_i v_j\\
&\quad-\left(\sum_{i,j}b^{ij}\ell_t v_i v_j+A\ell_t v^2\right)_t+(A\ell_t)_t v^2.
\end{aligned}
\tag{4.8}
$$

Further, by means of a direct computation, one may check that

$$
\begin{aligned}
&2\sum_{i,j}b^{ij}\ell_i v_j\left[-\sum_{i,j}(b^{ij}v_i)_j+Av\right]\\
&=-\sum_{i,j}\left[\sum_{i',j'}\left(2b^{ij}b^{i'j'}\ell_{i'}v_i v_{j'}-b^{ij}b^{i'j'}\ell_i v_{i'}v_{j'}\right)-Ab^{ij}\ell_i v^2\right]_j\\
&\quad+\sum_{i,j,i',j'}\left[2b^{ij'}(b^{i'j}\ell_{i'})_{j'}-(b^{ij}b^{i'j'}\ell_{i'})_{j'}\right]v_i v_j-\sum_{i,j}(Ab^{ij}\ell_i)_j v^2
\end{aligned}
\tag{4.9}
$$

and

$$
\begin{aligned}
\Psi v\left[-\sum_{i,j}(b^{ij}v_i)_j+Av\right]&=-\sum_{i,j}\left(\Psi b^{ij}v_i v-\frac{\Psi_i}{2}b^{ij}v^2\right)_j+\Psi\sum_{i,j}b^{ij}v_i v_j\\
&\quad+\left[-\frac{1}{2}\sum_{i,j}(b^{ij}\Psi_i)_j+A\Psi\right]v^2.
\end{aligned}
\tag{4.10}
$$

Finally, combining (4.6)–(4.10), we arrive at the desired equality (4.2). ∎

5. Pointwise Carleman-type estimates for the stochastic wave operator. In this section, we show a pointwise Carleman-type estimate (with singular weight) for the stochastic wave operator "$du_t-\Delta u dt$."

To begin with, by taking $(b^{ij})_{n\times n}=I$, the identity matrix, and $\theta=e^{\ell}$ (with ℓ given in (2.12)) in Theorem 4.1, one obtains the following pointwise Carleman-type estimate for the stochastic wave operator.

LEMMA 5.1. *Let $\ell,\ \Psi\in C^2((0,T)\times\mathbb{R}^n)$ and $k\in\mathbb{R}$. Assume u is an $H^2_{loc}(\mathbb{R}^n)$-valued $\{\mathcal{F}_t\}_{t\geq 0}$-adapted process such that u_t is an $L^2_{loc}(\mathbb{R}^n)$-valued semimartingale.*

Set $v=\theta u$. Then, for a.e. $x\in\mathbb{R}^n$ and P-a.s. $\omega\in\Omega$, it holds that

$$
\begin{aligned}
&\theta(-2\ell_t v_t+2\nabla\ell\cdot\nabla v+\psi v)(du_t-\Delta u dt)\\
&\quad+d\Big[\ell_t(v_t^2+|\nabla v|^2)-2(\nabla\ell)\cdot(\nabla v)v_t-\Psi v v_t+A\ell_t v^2\Big]\\
&\quad+\sum_i\Big\{2v_i(\nabla\ell)\cdot(\nabla v)-\ell_i|\nabla v|^2-2\ell_t v_t v_i+\ell_i v_t^2+\Psi v v_i-A\ell_i v^2\Big\}_i dt\\
&\geq\Big[(1-k)\lambda v_t^2+(k+3-4c)\lambda|\nabla v|^2+Bv^2\\
&\quad+\Big(-2\ell_t v_t+2\nabla\ell\cdot\nabla v+\psi v\Big)^2\Big]dt+\theta^2\ell_t(du_t)^2,
\end{aligned}
\tag{5.1}
$$

where

$$
\begin{cases}
\Psi\triangleq(2n-2c-1+k)\lambda,\\
A=4\left[c^2\left(t-\dfrac{T}{2}\right)^2-|x-x_0|^2\right]\lambda^2+\lambda(4c+1-k),\\
B=4\left[(4c+5-k)|x-x_0|^2-(8c+1-k)c^2\left(t-\dfrac{T}{2}\right)^2\right]\lambda^3+O(\lambda^2).
\end{cases}
\tag{5.2}
$$

The desired pointwise Carleman-type estimate (with singular weight function Θ) for the stochastic wave operator reads as follows.

THEOREM 5.1. *Assume u is an $H^2(G)$-valued $\{\mathcal{F}_t\}_{t\geq0}$-adapted process such that u_t is an $L^2(G)$-valued semimartingale. Let $v=\theta u$ and T satisfy* (2.7)*. Then there exist three constants $\lambda_0>0$, $\beta_0>0$, and $c_0>0$, independent of u, such that, for all $\beta\in(0,\beta_0)$ and $\lambda\geq\lambda_0$ and a.e. $x\in G$ and P-a.s. $\omega\in\Omega$, it holds that*

(5.3)

$$
\begin{aligned}
&\Theta\theta(-2\ell_t v_t+2\nabla\ell\cdot\nabla v+\psi v)(du_t-\Delta u dt)\\
&\quad+d\Big\{\Theta\Big[\ell_t(v_t^2+|\nabla v|^2)-2(\nabla\ell)\cdot(\nabla v)v_t-\Psi v v_t+A\ell_t v^2\Big]\Big\}\\
&\quad+\sum_i\Big\{\Theta\Big[2v_i(\nabla\ell)\cdot(\nabla v)-\ell_i|\nabla v|^2-2\ell_t v_t v_i+\ell_i v_t^2+\Psi v v_i-A\ell_i v^2\Big]\Big\}_i dt\\
&\geq\Big[c_0\lambda\Theta\theta^2(u_t^2+|\nabla u|^2+\lambda^2u^2)+\Theta\Big(-2\ell_t v_t+2\nabla\ell\cdot\nabla v+\psi v\Big)^2\Big]dt+\Theta\theta^2\ell_t(du_t)^2,
\end{aligned}
$$

with A and Ψ given by (5.2)*.*

Remark 5.1. The main difference between the pointwise estimates (5.1) and (5.3) is that we introduce a singular "pointwise" weight in (5.3). Another difference between (5.1) and (5.3) is that T is arbitrary in the former estimate, while for the latter one needs to take T to be large enough.

Proof of Theorem 5.1. We borrow some idea from the proof of [15, Theorem 1]. The proof is divided into several steps.

Step 1. We multiply both sides of inequality (5.1) by Θ. Obviously, we have (recall (5.2) for A and Ψ)

$$\begin{aligned}
&\Theta d\Big[\ell_t(v_t^2+|\nabla v|^2)-2(\nabla\ell)\cdot(\nabla v)v_t-\Psi vv_t+A\ell_t v^2\Big]\\
&=d\Big\{\Theta\Big[\ell_t(v_t^2+|\nabla v|^2)-2(\nabla\ell)\cdot(\nabla v)v_t-\Psi vv_t+A\ell_t v^2\Big]\Big\}\\
&\quad-\frac{\beta(T-2t)}{t^2(T-t)^2}\Theta\Big[\ell_t(v_t^2+|\nabla v|^2)-2(\nabla\ell)\cdot(\nabla v)v_t-\Psi vv_t+A\ell_t v^2\Big]dt.
\end{aligned}\tag{5.4}$$

Note that

$$\begin{aligned}
&\left|-\frac{\beta(T-2t)}{t^2(T-t)^2}\Theta\left[-2(\nabla\ell)\cdot(\nabla v)v_t-\Psi vv_t\right]\right|\\
&\le\frac{\beta|T-2t|}{t^2(T-t)^2}\Theta\left[2|(\nabla\ell)\cdot(\nabla v)v_t|+|\Psi v_t v|\right]\\
&\le\frac{\beta|T-2t|}{t^2(T-t)^2}\Theta\left[(|\nabla\ell|+1)v_t^2+|\nabla\ell||\nabla v|^2+\frac{1}{4}\Psi^2v^2\right].
\end{aligned}\tag{5.5}$$

Thus, by (5.1) and using (5.4)–(5.5), we get

$$\begin{aligned}
&\Theta\theta(-2\ell_t v_t+2\nabla\ell\cdot\nabla v+\psi v)(du_t-\Delta u dt)\\
&\quad+d\Big\{\Theta\Big[\ell_t(v_t^2+|\nabla v|^2)-2(\nabla\ell)\cdot(\nabla v)v_t-\Psi vv_t+A\ell_t v^2\Big]\Big\}\\
&\quad+\sum_i\Big\{\Theta\Big[2v_i(\nabla\ell)\cdot(\nabla v)-\ell_i|\nabla v|^2-2\ell_t v_t v_i+\ell_i v_t^2+\Psi vv_i-A\ell_i v^2\Big]\Big\}_i dt\\
&\ge\Big\{\Theta(1-k)\lambda v_t^2+\Theta(k+3-4c)\lambda|\nabla v|^2+\frac{\beta(T-2t)}{t^2(T-t)^2}\ell_t\Theta(v_t^2+|\nabla v|^2)\\
&\quad-\frac{\beta|T-2t|}{t^2(T-t)^2}\Theta\left[(|\nabla\ell|+1)v_t^2+|\nabla\ell||\nabla v|^2\right]\\
&\quad+\left[B+\frac{\beta(T-2t)}{t^2(T-t)^2}\ell_t A-\frac{\beta|T-2t|}{4t^2(T-t)^2}\Psi^2\right]\Theta v^2\\
&\quad+\Theta\Big(-2\ell_t v_t+2\nabla\ell\cdot\nabla v+\psi v\Big)^2\Big\}dt+\Theta\theta^2\ell_t(du_t)^2,
\end{aligned}\tag{5.6}$$

where B is given by (5.2).

Step 2. Recalling that ℓ and Ψ are given, respectively, by (2.12) and (5.2), we get

$$\begin{aligned}
\text{right-hand side of (5.6)}=&\Big[\lambda\Theta(F_1v_t^2+F_2|\nabla v|^2)+\lambda^3\Theta Gv^2\\
&+\Theta\Big(-2\ell_t v_t+2\nabla\ell\cdot\nabla v+\psi v\Big)^2\Big]dt+\Theta\theta^2\ell_t(du_t)^2,
\end{aligned}\tag{5.7}$$

where

$$F_1\triangleq 1-k+\frac{c\beta(T-2t)^2}{t^2(T-t)^2}-\frac{\beta|T-2t|}{t^2(T-t)^2}(2|x-x_0|+\lambda^{-1}),\tag{5.8}$$

$$F_2\triangleq k+3-4c+\frac{c\beta(T-2t)^2}{t^2(T-t)^2}-\frac{2\beta|T-2t||x-x_0|}{t^2(T-t)^2},\tag{5.9}$$

and

$$
\begin{aligned}
G \triangleq\ & 4\left[(4c+5-k)|x-x_0|^2-(8c+1-k)c^2\left(t-\frac{T}{2}\right)^2\right]+O(\lambda^{-1})\\
&+\frac{\beta|T-2t|}{t^2(T-t)^2}\Big\{4c|T-2t|\Big[c^2(t-T/2)^2-|x-x_0|^2\Big]+O(\lambda^{-1})\Big\}.
\end{aligned}
\tag{5.10}
$$

Step 3. Let us show that F_1, F_2, and G are positive when λ is large enough and β is sufficiently small. For this, put

$$
\begin{aligned}
&F_1^0 \triangleq 1-k, \qquad F_2^0 \triangleq k+3-4c,\\
&G^0 \triangleq 4\left[(4c+5-k)|x-x_0|^2-(8c+1-k)c^2\left(t-\frac{T}{2}\right)^2\right]+O(\lambda^{-1}),
\end{aligned}
$$

which are, respectively, the nonsingular part of F_1, F_2, and G. Similarly, put

$$
\begin{aligned}
&F_1^1 \triangleq \frac{c\beta(T-2t)^2}{t^2(T-t)^2}-\frac{\beta|T-2t|}{t^2(T-t)^2}(2|x-x_0|+\lambda^{-1}),\\
&F_2^1 \triangleq \frac{c\beta(T-2t)^2}{t^2(T-t)^2}-\frac{2\beta|T-2t||x-x_0|}{t^2(T-t)^2},\\
&G^1 \triangleq \frac{\beta|T-2t|}{t^2(T-t)^2}\Big\{4c|T-2t|\Big[c^2(t-T/2)^2-|x-x_0|^2\Big]+O(\lambda^{-1})\Big\},
\end{aligned}
$$

which are, respectively, the singular part of F_1, F_2, and G.

Further, we choose $k=1-c$. It is easy to see that both F_1^0 and F_1^0 are positive, and

$$
G^0 \geq 4(4+5c)R_0^2-9c^3T^2+O(\lambda^{-1}),
$$

which, via the first inequality in (2.7), is positive, provided that λ is sufficiently large.

When t is close to 0 or T, i.e., $t\in I_0 \triangleq (0,\delta_0)\cup(T-\delta_0,T)$ for some sufficiently small $\delta_0\in(0,T/2)$, the dominant terms in F_i $(i=1,2)$ and G are the singular ones. For $t\in I_0$,

$$
F_1^1 \geq \frac{\beta|T-2t|}{t^2(T-t)^2}[c(T-2\delta_0)-2R_1-\lambda^{-1})] = \frac{\beta|T-2t|}{t^2(T-t)^2}(cT-2R_1-2c\delta_0-\lambda^{-1}),
$$

which, via the second inequality in (2.7), is positive, provided that both δ_0 and λ^{-1} are sufficiently small. Similarly, for $t\in I_0$, F_2^0 is positive, provided that δ_0 is sufficiently small. Further, for $t\in I_0$,

$$
\begin{aligned}
G^1 &\geq \frac{\beta|T-2t|}{t^2(T-t)^2}\Big\{4c|T-2\delta_0|\Big[c^2(\delta_0-T/2)^2-R_1^2\Big]+O(\lambda^{-1})\Big\}\\
&\geq \frac{\beta|T-2t|}{t^2(T-t)^2}\Big\{4c|T-2\delta_0|\Big[c^2T^2/4-R_1^2+c^2\delta_0(\delta_0-T)\Big]+O(\lambda^{-1})\Big\},
\end{aligned}
$$

which, via the second inequality in (2.7), is positive, provided that both δ_0 and λ^{-1} are sufficiently small.

By (5.8)–(5.10), we see that $F_1 = F_1^0 + F_1^1$, $F_2 = F_2^0 + F_2^1$, and $G = G^0 + G^1$. Noting the positivity of F_1^0, F_2^0, and G^0, by the above argument, we see that F_1, F_2, and G are positive for $t \in I_0$. For $t \in (0,T) \setminus I_0$, noting again the positivity of F_1^0, F_2^0, and G^0, one can choose $\beta > 0$ small enough such that F_1^1, F_2^1, and G^1 are very small so that F_1, F_2, and G are positive. Hence (5.6)–(5.7) yield the desired (5.3). This completes the proof of Theorem 5.1. ∎

6. Proof of Theorems 2.1–2.2. We are now in a position to prove Theorems 2.1–2.2.

Proof of Theorem 2.2. The key idea is to apply Theorem 5.1. Integrating both sides of (5.3) (with u replaced by y, and $v = \theta y$), using integration by parts, and recalling that $\Theta(t)$ decays exponentially to 0 as $t \to 0$ or $t \to T$, noting that $v|_\Sigma = 0$ (and hence $\nabla v = \frac{\partial v}{\partial \nu}\nu$ on Σ), we arrive at

$$\begin{aligned}
&\mathbb{E}\int_Q \left[c_0\lambda\Theta\theta^2(y_t^2 + |\nabla y|^2 + \lambda^2 y^2) + \Theta\Big(-2\ell_t v_t + 2\nabla\ell\cdot\nabla v + \psi v\Big)^2\right] dxdt \\
&\le \mathbb{E}\int_Q \Theta\theta\left[(-2\ell_t v_t + 2\nabla\ell\cdot\nabla v + \psi v)(dy_t - \Delta y dt) - \theta\ell_t(dy_t)^2\right] dx \\
&\quad + \mathbb{E}\int_\Sigma \Theta\frac{\partial\ell}{\partial\nu}\left|\frac{\partial v}{\partial\nu}\right|^2 d\Gamma dt.
\end{aligned} \tag{6.1}$$

By the first equation of system (1.1), we get

(6.2)

$$\begin{aligned}
&\mathbb{E}\int_Q \Theta\theta\left[(-2\ell_t v_t + 2\nabla\ell\cdot\nabla v + \psi v)(dy_t - \Delta y dt) - \theta\ell_t(dy_t)^2\right] dx \\
&= \mathbb{E}\int_Q \Theta\theta\Big[(-2\ell_t v_t + 2\nabla\ell\cdot\nabla v + \psi v)(a_1 y_t + a_2\cdot\nabla y + a_3 y + f) \\
&\qquad\qquad - \theta\ell_t(a_4 y + g)^2\Big] dxdt \\
&\le \mathbb{E}\int_Q \Theta\Big(-2\ell_t v_t + 2\nabla\ell\cdot\nabla v + \psi v\Big)^2 dxdt \\
&\quad + C\left\{\mathbb{E}\int_Q \Theta\theta^2\Big(a_1 y_t + a_2\cdot\nabla y + a_3 y + f\Big)^2 dxdt + \lambda\mathbb{E}\int_Q \Theta\theta^2(a_4 y + g)^2 dxdt\right\} \\
&\le \mathbb{E}\int_Q \Theta\Big(-2\ell_t v_t + 2\nabla\ell\cdot\nabla v + \psi v\Big)^2 dxdt \\
&\quad + C\left\{\mathbb{E}\int_Q \Theta\theta^2(f^2 + \lambda g^2) dxdt + |a_1|^2_{L^\infty_{\mathcal{F}}(0,T;(L^\infty(G))}\mathbb{E}\int_Q \Theta\theta^2 y_t^2 dxdt\right. \\
&\quad + \lambda\Big[\lambda|a_3|^2_{L^\infty_{\mathcal{F}}(0,T;L^n(G))} + |a_4|^2_{L^\infty_{\mathcal{F}}(0,T;(L^\infty(G))}\Big]\mathbb{E}\int_Q \Theta\theta^2 y^2 dxdt \\
&\quad \left. + \Big[|a_2|^2_{L^\infty_{\mathcal{F}}(0,T;L^\infty(G;\mathbb{R}^n))} + |a_3|^2_{L^\infty_{\mathcal{F}}(0,T;L^n(G))}\Big]\mathbb{E}\int_Q \Theta\theta^2|\nabla y|^2 dxdt\right\}.
\end{aligned}$$

On the other hand, recalling (2.2), we have

$$\begin{aligned}(6.3)\qquad &\mathbb{E}\int_\Sigma \Theta\frac{\partial \ell}{\partial \nu}\left|\frac{\partial v}{\partial \nu}\right|^2 d\Gamma dt = 2\lambda\mathbb{E}\int_\Sigma \Theta\theta^2(x-x_0)\cdot\nu(x)\left|\frac{\partial y}{\partial \nu}\right|^2 d\Gamma dt\\ &\le 2\lambda\mathbb{E}\int_{\Sigma_0}\Theta\theta^2(x-x_0)\cdot\nu(x)\left|\frac{\partial y}{\partial \nu}\right|^2 d\Gamma_0 dt \le C\lambda\mathbb{E}\int_{\Sigma_0}\Theta\theta^2\left|\frac{\partial y}{\partial \nu}\right|^2 d\Gamma_0 dt.\end{aligned}$$

Finally, combining (6.1), (6.2), and (6.3), we arrive at the desired estimate (2.14). This completes the proof of Theorem 2.2. □

Proof of Theorem 2.1. The proof follows easily from Theorem 2.2 and the energy estimate in Proposition 3.1. We omit the details. □

7. Application to state observation problem of semilinear stochastic wave equations. This section is devoted to giving an application of the observability inequality in section 2. Fixing two known nonlinear functions $F(\eta, v, \zeta) : \mathbb{R}^1 \times \mathbb{R}^1 \times \mathbb{R}^n \to \mathbb{R}^1$ and $G(\eta) : \mathbb{R}^1 \to \mathbb{R}^1$, and unknown initial data $(w_0, w_1) \in L^2(\Omega, \mathcal{F}_0, P; H_0^1(G) \times L^2(G))$, for the semilinear stochastic wave equation

$$(7.1)\qquad \begin{cases} dw_t - \Delta w dt = F(w, w_t, \nabla w)dt + G(w)dB(t) & \text{in } Q,\\ w = 0 & \text{on } \Sigma,\\ w(0) = w_0, \quad w_t(0) = w_1 & \text{in } G,\end{cases}$$

we consider the following.

State observation problem: Let $T > 0$ be given. Determine the state $(w(t), w_t(t)) \in L^2(\Omega, \mathcal{F}_t, P; H_0^1(G) \times L^2(G))$ of (7.1) (for $t \in (0, T]$) from the observed boundary data

$$\frac{\partial w}{\partial \nu}(t, x), \qquad (t, x) \in \Sigma_0, \quad P\text{-a.s.}$$

We refer the reader to [11, 14] for some studies on the state observation problems for (deterministic) semilinear wave equations. To the best of our knowledge, nothing is published for the stochastic setting. We need the following assumption:

(H) The nonlinear functions $F(\cdot)$ and $G(\cdot)$ satisfy the following:

(1)

$$\begin{aligned}|F(\eta_1, v, \zeta) - F(\eta_2, v, \zeta)| + |G(\eta_1) - G(\eta_2)| &\le L(1 + |\eta_1|^{p-1} + |\eta_2|^{p-1})|\eta_1 - \eta_2|\\ &\forall\ \eta_1, \eta_2, v \in \mathbb{R}^1,\ \zeta \in \mathbb{R}^n\end{aligned}$$

with $1 \le p \le \frac{n}{n-2}$ if $n \ge 3$; $1 \le p < \infty$ if $n = 1, 2$;

(2)

$$\begin{aligned}&|F(\eta, v_1, \zeta_1) - F(\eta, v_2, \zeta_2)| \le L(|v_1 - v_2| + |\zeta_1 - \zeta_2|)\\ &\qquad\forall\ (\eta, v_i, \zeta_i) \in \mathbb{R}^1 \times \mathbb{R}^1 \times \mathbb{R}^n,\ i = 1, 2,\\ &|F(0, v, \zeta)| \le L(|v| + |\zeta|)\ \ \forall\ (v, \zeta) \in \mathbb{R}^1 \times \mathbb{R}^n\end{aligned}$$

for some constant $L > 0$;

(3) for any given initial data $(w_0, w_1) \in L^2(\Omega, \mathcal{F}_0, P; H_0^1(G) \times L^2(G))$, (7.1) admits a unique solution $w = w(\cdot\,; w_0, w_1) \in \mathcal{H}_T$ (the solution of (7.1) is defined similarly to Definition 1.1).

Since we do not introduce any sign condition on the nonlinearities $F(\cdot)$ and $G(\cdot)$, the global existence of a solution of (7.1) is not guaranteed. This is why we need to impose the third assumption in (H). The (global) well-posedness of (7.1) is an interesting but difficult problem, which is beyond the scope of this paper.

It is easy to see that, by the Sobolev embedding theorem, assumption (H) implies that

$$F(z, z_t, \nabla z) \in L^2_{\mathcal{F}}(0,T;L^2(G)), \quad G(z) \in L^2_{\mathcal{F}}(0,T;L^2(G))$$

for any $z \in \mathcal{H}_T$. Thus, thanks to Proposition 3.3, we can define a nonlinear map $\gamma:\ L^2(\Omega,\mathcal{F}_0,P;H_0^1(G)\times L^2(G)) \to L^2_{\mathcal{F}}(0,T;L^2(\Gamma_0))$ by

$$\gamma(w_0,w_1) = \left.\frac{\partial w}{\partial \nu}\right|_{\Sigma_0\times\Omega},$$

where w is the solution of (7.1). Now we have the following result.

THEOREM 7.1. *Let* (H) *hold,* Γ_0 *be given by* (2.2), *and* T *satisfy* (2.7). *Then, for any* $(w_0,w_1),(\hat{w}_0,\hat{w}_1) \in L^2(\Omega,\mathcal{F}_0,P;H_0^1(G)\times L^2(G))$ *and any* $t\in(0,T]$, *there exists a constant* $\widetilde{C} = \widetilde{C}(F,G,w_0,\hat{w}_0,w_1,\hat{w}_1) > 0$ *such that*

$$\begin{aligned}&|(w(t)-\hat{w}(t), w_t(t)-\hat{w}_t(t))|_{L^2(\Omega,\mathcal{F}_t,P;H_0^1(G)\times L^2(G))}\\ &\le e^{Ct^{-1}}\widetilde{C}|\gamma(w_0,w_1)-\gamma(\hat{w}_0,\hat{w}_1)|_{L^2_{\mathcal{F}}(0,T;L^2(\Gamma_0))},\end{aligned}$$

where $\hat{w} = \hat{w}(\cdot\,;\hat{w}_0,\hat{w}_1) \in \mathcal{H}_T$ *is the solution of* (7.1) *with* (w_0,w_1) *replaced by* $(\hat{w}_0,\hat{w}_1)$.

Remark 7.1. Theorem 7.1 indicates that the state $(w(t),w_t(t)) \in L^2(\Omega,\mathcal{F}_t,P;H_0^1(G)\times L^2(G))$ of (7.1) (for $t\in(0,T]$) can be uniquely determined from the observed boundary data $\frac{\partial w}{\partial\nu}(t,x)\big|_{\Sigma_0}$, P-a.s., and continuously depend on it.

Proof of Theorem 7.1. Set

$$y = \hat{w} - w.$$

It is easy to see that y is a solution of (1.1) with

$$\begin{cases} a_1 = \displaystyle\int_0^1 \partial_\eta F(w+s(\hat{w}-w),w_t,\nabla w)ds, \quad a_2 = \int_0^1 \partial_v(\hat{w},w_t+s(\hat{w}_t-w_t),\nabla w)ds,\\ a_3 = \displaystyle\int_0^1 \partial_\zeta F(\hat{w},\hat{w}_t,\nabla w+s(\nabla\hat{w}-\nabla w))ds, \quad a_4 = \int_0^1 \partial_\eta G(w+s(\hat{w}-w))ds.\end{cases}$$

The desired result follows immediately from Remark 2.1. ∎

Acknowledgments. The author acknowledges Professor Shanjian Tang for stimulating discussion and the anonymous referees for helpful comments.

REFERENCES

[1] V. BARBU, A. RĂSCANU, AND G. TESSITORE, *Carleman estimate and controllability of linear stochastic heat equations*, Appl. Math. Optim., 47 (2003), pp. 97–120.

[2] C. BARDOS, G. LEBEAU, AND J. RAUCH, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[3] T. Duyckaerts, X. Zhang, and E. Zuazua, *On the optimality of the observability inequalities for parabolic and hyperbolic systems with potentials*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 25 (2008), pp. 1–41.
[4] X. Fu, *A weighted identity for partial differential operators of second order and its applications*, C. R. Math. Acad. Sci. Paris, 342 (2006), pp. 579–584.
[5] X. Fu, J. Yong, and X. Zhang, *Exact controllability for the multidimensional semilinear hyperbolic equations*, SIAM J. Control Optim., 46 (2007), pp. 1578–1614.
[6] W. Grecksch and C. Tudor, *Stochastic Evolution Equations: A Hilbert Space Approach*, Math. Res. 85, Akademie-Verlag, Berlin, 1995.
[7] V. Isakov, *Carleman estimates and applications to inverse problems*, Milan J. Math., 72 (2004), pp. 249–271.
[8] M. Kazemi and M. V. Klibanov, *Stability estimates for ill-posed Cauchy problem involving hyperbolic equations and inequalities*, Appl. Anal., 50 (1993), pp. 93–102.
[9] M. M. Lavrent'ev, V. G. Romanov, and S. P. Shishat·skiĭ, *Ill-Posed Problems of Mathematical Physics and Analysis*, Transl. Math. Monogr. 64, AMS, Providence, RI, 1986.
[10] J.-L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués, Tome* 1*, Contrôlabilité exacte*, Rech. Math. Appl. 8, Masson, Paris, 1988.
[11] L. Pan, K. L. Teo, and X. Zhang, *State-observation problem for a class of semi-linear hyperbolic systems*, Chinese Ann. Math. Ser. A, 25 (2004), pp. 189–198 (in Chinese); Chinese J. Contemp. Math., 25 (2004), pp. 163–172 (in English).
[12] S. Tang and X. Zhang, *Carleman inequality for backward stochastic parabolic equations with general coefficients*, C. R. Math. Acad. Sci. Paris, 339 (2004), pp. 775–780.
[13] X. Zhang, *Explicit observability inequalities for the wave equation with lower order terms by means of Carleman inequalities*, SIAM J. Control Optim., 39 (2000), pp. 812–834.
[14] X. Zhang, *Exact Controllability of Semi-linear Distributed Parameter Systems*, Gaodeng Jiaoyu Chubanshe (Higher Education Press), Beijing, China, 2004 (in Chinese).
[15] X. Zhang and E. Zuazua, *A sharp observability inequality for Kirchoff plate systems with potentials*, Comput. Appl. Math., 25 (2006), pp. 353–373.
[16] E. Zuazua, *Controllability and observability of partial differential equations: Some results and open problems*, in Handbook of Differential Equations: Evolutionary Equations, Vol. 3, Elsevier Science, Amsterdam, 2006, pp. 527–621.

SIAM J. MATH. ANAL.
Vol. 40, No. 3, pp. 869–904

GLOBAL BEHAVIOR OF SPHERICALLY SYMMETRIC NAVIER–STOKES EQUATIONS WITH DEGENERATE VISCOSITY COEFFICIENTS*

MINGJUN WEI†, TING ZHANG†, AND DAOYUAN FANG†

Abstract. In this paper, we study a free boundary problem for compressible spherically symmetric Navier–Stokes equations with a gravitational force and degenerate viscosity coefficients. Under certain assumptions that are imposed on the initial data, we obtain the global existence and uniqueness of the weak solution and give some uniform bounds (with respect to time) of the solution. Moreover, we obtain some stabilization rate estimates in L^∞-norm and weighted H^1-norm of the solution. The results show that such a system is stable under small perturbations and could be applied to the astrophysics.

Key words. compressible Navier–Stokes equations, density-dependent viscosity, free boundary, asymptotic behavior

AMS subject classifications. 35Q35, 35D05, 76N10

DOI. 10.1137/070681703

1. Introduction.

We consider the compressible Navier–Stokes equations with density-dependent viscosity in $\mathbb{R}^n (n \geq 2)$, which can be written in Eulerian coordinates as

$$\begin{cases} \partial_\tau \rho + \nabla \cdot (\rho \vec{u}) = 0, \\ \partial_\tau (\rho \vec{u}) + \nabla \cdot (\rho \vec{u} \otimes \vec{u}) + \nabla P = \mathrm{div}(\mu(\nabla \vec{u} + \nabla \vec{u}^\top)) + \nabla(\lambda \mathrm{div} \vec{u}) - \rho \vec{f}, \end{cases} \tag{1.1}$$

in a domain $\{(\vec{\xi}, \tau) \big| \vec{\xi} \in \Omega_\tau \subset \mathbb{R}^n, \tau > 0\}$, the initial conditions and boundary conditions are

$$(\rho, \vec{u})(\vec{\xi}, 0) = (\rho_0, u_0)(\vec{\xi}), \ \vec{\xi} \in \Omega_0 = \{\vec{\xi} \in \mathbb{R}^n \big| a < |\vec{\xi}| < b\}, \tag{1.2}$$

$$\vec{u}|_{|\vec{\xi}|=a} = 0, \ \rho|_{\vec{\xi} \in \partial\Omega_\tau \setminus \{|\vec{\xi}|=a\}} = 0, \tag{1.3}$$

where $0 < a < b < \infty$, $\Omega_\tau = \psi(\Omega_0, \tau)$ and ψ is the flow of $\vec{u}$:

$$\begin{cases} \partial_\tau \psi(\vec{\xi}, \tau) = \vec{u}(\psi(\vec{\xi}, \tau), \tau), \quad \vec{\xi} \in \mathbb{R}^n, \\ \psi(\vec{\xi}, 0) = \vec{\xi}. \end{cases} \tag{1.4}$$

Here ρ, P, $\vec{u} = (\mathrm{u}_1, \ldots, \mathrm{u}_n)$, and $\vec{f}$ are the density, pressure, velocity, and external force, respectively; $\lambda = \lambda(\rho)$ and $\mu = \mu(\rho)$ are the viscosity coefficients.

*Received by the editors February 3, 2007; accepted for publication (in revised form) May 21, 2008; published electronically September 8, 2008. This work is supported by NSFC 10571158, Zhejiang Provincial NSF of China (Y605076), and China Postdoctoral Science Foundation 20060400335.
http://www.siam.org/journals/sima/40-3/68170.html
†Department of Mathematics, Zhejiang University, Hangzhou 310027, People's Republic of China (m.j.wei@126.com, zhangting79@hotmail.com (corresponding author), dyf@zju.edu.cn).

For the initial-boundary value problem (1.1)–(1.3) with the spherically symmetric initial data and external force

$$(\rho,\vec{u})(\vec{\xi},0)=\left(\rho_0(r),u_0(r)\frac{\vec{\xi}}{r}\right),\ \vec{\xi}\in\Omega_0,$$
$$\vec{f}=f(m,r,\tau)\frac{\vec{\xi}}{r},\ m(\rho,r)=\int_a^r\rho(s,\tau)s^{n-1}ds,\ \vec{\xi}\in\Omega_\tau,$$

where $r=|\vec{\xi}|=\sqrt{\xi_1^2+\cdots+\xi_n^2}$, we are looking for spherically symmetric solutions $(\rho,\vec{u})$:

$$\rho(\vec{\xi},\tau)=\rho(r,\tau),\quad \vec{u}(\vec{\xi},\tau)=u(r,\tau)\frac{\vec{\xi}}{r},\ \vec{\xi}\in\Omega_\tau,$$

where $\Omega_\tau=\{\vec{\xi}\in\mathbb{R}^n\ \big|\ a<|\vec{\xi}|<b(\tau),b(0)=b,b'(\tau)=u(b(\tau),\tau)\}$. Then $(\rho,u)(r,\tau)$ is determined by

$$\text{(1.5)}\quad\begin{cases}\partial_t\rho+\partial_r(\rho u)+\frac{n-1}{r}\rho u=0,\\ \rho(\partial_t u+u\partial_r u)+\partial_r P=\partial_r[(\lambda+2\mu)(\partial_r u+\frac{n-1}{r}u)]-2(n-1)\frac{u}{r}\partial_r\mu-\rho f,\end{cases}$$

where $(r,\tau)\in(a,b(\tau))\times(0,\infty)$, with the initial data

$$(\rho,u)|_{\tau=0}=(\rho_0,u_0)(r),\ a\le r\le b,\tag{1.6}$$

the boundary conditions

$$u|_{r=a}=0,\ \rho|_{r=b(\tau)}=0,\tag{1.7}$$

where $b(0)=b$, $b'(\tau)=u(b(\tau),\tau)$, $\tau>0$.

To simplify the presentation, we consider only the famous polytropic model, i.e., $P(\rho)=A\rho^\gamma$, with $\gamma>1$ and $A>0$ being constants. And we assume that the viscosity coefficients μ and λ are proportional to ρ^θ, i.e., $\mu(\rho)=c_1\rho^\theta$ and $\lambda(\rho)=c_2\rho^\theta$, where c_1,c_2, and θ are three constants.

Additionally, we assume that the external force $f(m,r,\tau)$ satisfies

$$f(m,r,\tau)=f_\infty(m,r)+\Delta f(m,r,\tau),\tag{1.8}$$

for all $m\ge0$, $r\ge a$, and $\tau\ge0$, with

$$f_\infty(m,r)=G\frac{M_0+m}{r^{n-1}},\ \Delta f(m,r,\tau)\in C^1(\mathbb{R}_+\times\mathbb{R}_+\times\mathbb{R}_+)\tag{1.9}$$

$$\|\Delta f(\cdot,\cdot,\tau)\|_{L^\infty(\mathbb{R}_+\times\mathbb{R}_+)}\le f_1(\tau),\ \|(\partial_r\Delta f,\partial_\tau\Delta f)(\cdot,\cdot,\tau)\|_{L^\infty(\mathbb{R}_+\times\mathbb{R}_+)}\le f_2(\tau),\tag{1.10}$$

$$f_1\in L^\infty\cap L^2(\mathbb{R}_+),\quad f_2\in L^2(\mathbb{R}_+),\tag{1.11}$$

where $\mathbb{R}_+=[0,\infty)$, $G>0$ is a constant, $M_0\ge0$ is the total mass of the solid core surrounded by the gas, and the perturbation Δf tends to 0 as $\tau\to\infty$ in some weak sense. If $M_0=0$, we ignore the gravitational effect of the solid core. Δf expresses the influence of the outside gravitational force, f_∞ is the precise expression for its own gravitational force and the gravitational force of the solid core, in the astrophysical case (with spherical symmetry).

It is very interesting to study the stabilization problem for the model of the atmosphere of a planet, whose evolution is influenced by the gravitational force. In [22], we considered a simple model that is one-dimensional Navier–Stokes equations. In this paper, we study the stabilization problem of the spherically symmetric model (1.5). We should consider the essential multidimensional model in the future.

Now, we consider the stationary problem, namely

$$(P(\rho_\infty))_r = -\rho_\infty f_\infty(m(\rho_\infty, r), r) \tag{1.12}$$

in an interval $r \in (a, l_\infty)$, with the end l_∞ satisfying

$$\rho_\infty(l_\infty) = 0, \ \int_a^{l_\infty} \rho_\infty r^{n-1} dr = M := \int_a^b \rho_0 r^{n-1} dr. \tag{1.13}$$

The unknown quantities are the stationary density $\rho_\infty \geq 0$ and free boundary $l_\infty > 0$. If $\gamma > \frac{2n-2}{n}$, from Proposition 2.5, we know that there exists a unique solution (ρ_∞, l_∞) to the stationary system (1.12)–(1.13), satisfying $\rho_\infty(r) \sim (l_\infty^n - r^n)^{\frac{1}{\gamma-1}}$, $(\rho_\infty)_r(r) < 0$, $a < r < l_\infty$, with $l_\infty < +\infty$.

To handle the free boundary problem (1.5)–(1.7), it is convenient to reduce the problem in Eulerian coordinates (r, τ) to the problem in Lagrangian coordinates (x, t) moving with the fluids, via the transformation:

$$x = \int_a^r y^{n-1} \rho(y, \tau) dy, \ t = \tau. \tag{1.14}$$

Then the fixed boundary $r = a$ and the free boundary $r = b(\tau)$ become

$$x = 0 \ \text{ and } x = \int_a^{b(\tau)} y^{n-1} \rho(y, \tau) dy = \int_a^b y^{n-1} \rho_0(y) dy = M,$$

where M is the total mass initially so that the region $\{(r, \tau) | a \leq r \leq b(\tau), \tau \geq 0\}$ under consideration is transformed into the region $\{(x, t) | 0 \leq x \leq M, t \geq 0\}$, and the function $m(\rho, r)$ becomes x. Under the coordinate transformation (1.14), the equations (1.5)–(1.7) are transformed into

$$\begin{cases} \partial_t \rho(x,t) = -\rho^2 \partial_x (r^{n-1} u), \\ \partial_t u(x,t) = r^{n-1} \left\{ \partial_x [\rho(\lambda + 2\mu) \partial_x (r^{n-1} u) - P] - 2(n-1) \frac{u}{r} \partial_x \mu \right\} - f(x, r, t), \\ r^n(x,t) = a^n + n \int_0^x \rho^{-1}(y, t) dy, \end{cases} \tag{1.15}$$

where $(x, t) \in (0, M) \times (0, \infty)$, with the initial data

$$(\rho, u)|_{t=0} = (\rho_0, u_0)(x), r|_{t=0} = r_0(x) = \left(a^n + n \int_0^x \rho_0^{-1}(y) dy \right)^{\frac{1}{n}}, \tag{1.16}$$

and the boundary conditions

$$u|_{x=0} = 0, \ \rho|_{x=M} = 0, \ t > 0. \tag{1.17}$$

It is standard that if we can solve the problem (1.15)–(1.17), then the free boundary problem (1.1)–(1.3) has a solution.

From (1.12)–(1.13), it is easy to see that $\rho_\infty(x)$ is the solution to the stationary system

$$Ar_\infty^{n-1}(\rho_\infty^\gamma)_x = -f_\infty(x, r_\infty),\ r_\infty^n(x) = a^n + n\int_0^x \rho_\infty^{-1}(y)dy,\ x \in (0, M), \tag{1.18}$$

$$\rho_\infty(M) = 0. \tag{1.19}$$

The results in [8, 20] show that the compressible Navier–Stokes system with the constant viscosity coefficient has the singularity at the vacuum. Considering the modified Navier–Stokes system in which the viscosity coefficient depends on the density, Liu, Xin, and Yang [11] proved that such system is local well posed. It is motivated by the physical consideration that, in the derivation of the Navier–Stokes equations from the Boltzmann equation through the Chapman–Enskog expansion to the second order, cf. [6], viscosity is a function of temperature. If we consider the case of isentropic fluids, this dependence is reduced to the dependence on the density function.

Since $n \geq 2$ and the viscosity coefficient μ depends on ρ, the nonlinear term $2(n-1)\frac{1}{r}u\partial_x\mu$ in $(1.15)_2$ makes the analysis significantly different from the one-dimensional case [11, 15, 19, 21, 22]. When $\mu \geq \underline{\mu} > 0$ and $\rho_0 \geq \underline{\rho} > 0$, authors in [4, 24] obtained the existence, uniqueness, and global behavior of the solution for compressible spherically symmetric Navier–Stokes equations with an external pressure and without the nonlinear term $2(n-1)\frac{1}{r}u\partial_x\mu$. Following the ideas in [24], we can obtain the existence and uniqueness results for the stationary problem in section 2. In this paper, since viscosity coefficients and density will degenerate at the free boundary, we need to use the weighted function $(M-x)$ to control the lower bound of the density in section 3.

Considering the system (1.15)–(1.17) with a general external force, Chen-Zhang obtained the local existence and uniqueness of the solution in [3]. In this paper, when the initial data (ρ_0, u_0, r_0) are close to the stationary state $(\rho_\infty, 0, r_\infty)$, we will obtain some appropriate a priori estimates and prove that the maximum existence time $T^* = \infty$. The difficulty of this problem is to obtain the lower bound of the density ρ. The key ideas are using the classical continuation method and the result of Claim 1. In Claim 1, we want to prove that there is a small positive constant ϵ_1, such that, for any $T > 0$, if

$$I(t) = \|g(\cdot, t) - g_\infty\|_{L^\infty} \leq 2\epsilon_1\ \forall\ t \in [0, T], \tag{1.20}$$

where $g(x,t) = (M-x)^{-\frac{1}{\gamma}}\rho(x,t)$ and $g_\infty(x) = (M-x)^{-\frac{1}{\gamma}}\rho_\infty(x)$, then

$$I(t) \leq \epsilon_1\ \forall\ t \in [0, T]. \tag{1.21}$$

Using the energy method and induction method, we can estimate the weighted L^2- norm of $g - g_\infty$ in Lemma 3.7. In such a process (see Lemmas 3.5–3.6), we use the weight function $(1+t)^\alpha$ (with $\alpha = -\frac{5}{8}$) to remedy the disadvantage of the nonlinear term $2(n-1)\frac{1}{r}u\partial_x\mu$, and we use the induction method to increase α to $-\epsilon_2$. Then, by the reduction to absurdity, we can finish the proof of Claim 1 in Lemma 3.8.

Our results show that such a system is stable under small perturbations, it does not develop vacuum states or concentration states for all time, and the interface $\partial\Omega_\tau$ propagates with finite speed.

The assumptions on c_1, c_2, θ, γ, and initial data can be stated as follows:

(A1) $\gamma > \frac{2n-2}{n}$, $\theta \in (0, \gamma-1) \cap (0, \frac{\gamma}{2}]$, c_1 and c_2 satisfy that $c_1 > 0$ and $2c_1 + nc_2 > 0$;

(A2) $N_1(M-x)^{1/\gamma} \le \rho_0 \le N_2(M-x)^{1/\gamma}$, with some positive constants $0 < N_1 < N_2$, and $(M-x)^{1-\frac{\theta}{\gamma}}(\rho_0^\theta)_x^2 \in L^1([0,M])$, $(\rho_0^\gamma)_x \in L^2([0,M])$;

(A3) $u_0 \in L^2([0,M])$, $\rho_0^{\frac{\theta+1}{2}}(u_0)_x \in L^2([0,M])$, $u_0(0) = 0$,

$$\big((2c_1+c_2)\rho_0^{\theta+1}(r_0^{n-1}u_0)_x\big)_x - 2c_1(n-1)\frac{u_0}{r_0}(\rho_0^\theta)_x \in L^2([0,M]). \tag{1.22}$$

Under the above assumptions (A1)–(A3), we will prove the existence of the global weak solution to the initial-boundary value problem (1.15)–(1.17) in the sense of the following definition.

DEFINITION 1.1. *A pair of functions $(\rho,u,r)(x,t)$ is called a global weak solution to the initial-boundary value problem* (1.15)–(1.17) *if, for any $T > 0$,*

$$\rho, u \in L^\infty([0,M]\times[0,T]) \cap C^1([0,T]; L^2([0,M])),$$

$$r \in C^1([0,T]; L^\infty([0,M])),$$

$$\rho^{-1},\ (r^{n-2}u)_x,\ (r^{n-1})_x \in L^\infty([0,T]; L^1([0,M])),$$

and

$$\rho^{1+\theta}(r^{n-1}u)_x \in L^\infty([0,M]\times[0,T]) \cap C^{\frac{1}{2}}([0,T]; L^2([0,M])).$$

Furthermore, the following equations hold:

$$\rho_t + \rho^2(r^{n-1}u)_x = 0,\ \rho(x,0) = \rho_0(x),\ a.e.$$

$$r_t = u,\ r(x,0) = r_0(x),\ r^n(x,t) = a^n + n\int_0^x \rho^{-1}(y,t)dy,\ a.e.$$

$$\int_0^\infty\int_0^M [u\psi_t + (P - \rho(\lambda+2\mu)(r^{n-1}u)_x)(r^{n-1}\psi)_x$$
$$+ 2(n-1)\mu(r^{n-2}u\psi)_x - f(x,r,t)\psi]dxdt + \int_0^M u_0(x)\psi(x,0)dx = 0$$

for any test function $\psi(x,t) \in C_0^\infty(\Omega)$, with $\Omega = \{(x,t) \mid 0 < x \le M,\ t \ge 0\}$ and

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_0^\epsilon u dx = \lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_{1-\epsilon}^1 \rho dx = 0.$$

In what follows, we always use $C(C_i)$ to denote a generic positive constant depending only on γ, θ, f_1, f_2, and the initial data, independent of the given time T.

We now state the main theorems in this paper.

THEOREM 1.1. *Under the conditions* (1.8)–(1.11) *and* (A1)–(A3)*, there exists a constant $\epsilon_0 > 0$ such that if*

$$\|f_1\|_{L^\infty}^2 + \int_0^\infty (1+t)f_1^2(t)dt \le \epsilon_0^2 \tag{1.23}$$

and

$$\|u_0\|_{L^2} + \|(M-x)^{-\frac{1}{\gamma}}(\rho_0 - \rho_\infty)\|_{L^\infty} \le \epsilon_0, \tag{1.24}$$

then the system (1.15)–(1.17) *has a unique global weak solution* (ρ, u, r) *satisfying*

$$C^{-1}(M-x)^{\frac{1}{\gamma}} \le \rho(x,t) \le C(M-x)^{\frac{1}{\gamma}}, \tag{1.25}$$

$$r(x,t) \in [a, C], \tag{1.26}$$

$$\int_0^M (M-x)^{1-\frac{\theta}{\gamma}}(\rho^\theta - \rho_\infty^\theta)_x^2 dx \le C \tag{1.27}$$

and

$$\|u(\cdot,t)\|_{L^\infty} + \|[\rho(r^{n-1}u)_x](\cdot,t)\|_{L^\infty} \le C \tag{1.28}$$

for all $t \ge 0$ *and* $x \in [0, M]$. *Furthermore, if* $(1+t)^{\frac{2(\gamma+\theta)}{\gamma+\theta+1}} f_1^2(t) \in L^1(\mathbb{R}_+)$, *for any* $\eta > 0$, *we have*

$$\int_0^M \left\{u^2 + (M-x)^{\frac{\gamma-1}{\gamma}}(g-g_\infty)^2 + (r-r_\infty)^2\right\} dx \le C_\eta (1+t)^{-\frac{2(\gamma+\theta)}{\gamma+\theta+1}+\eta}, \tag{1.29}$$

$$\int_0^M \left(\rho^{\theta-1}u^2 + \rho^{\theta+1}u_x^2\right) dx \le C_\eta (1+t)^{-\frac{2(\gamma+\theta)}{\gamma+\theta+1}+\eta}, \tag{1.30}$$

$$\int_0^M (M-x)^{2-\frac{2\theta}{\gamma}}(\rho^\theta - \rho_\infty^\theta)_x^2 dx \le C_\eta (1+t)^{-\frac{\gamma+\theta-1}{\gamma+\theta+1}+\eta}, \tag{1.31}$$

$$\|\rho^\gamma(\cdot,t) - \rho_\infty^\gamma(\cdot)\|_{L^\infty} \le C_\eta (1+t)^{-\frac{3\gamma+3\theta-1}{4(\gamma+\theta+1)}+\frac{\eta}{2}}, \tag{1.32}$$

and

$$\|\rho^{\frac{\gamma+\theta}{2}}(\cdot,t) - \rho_\infty^{\frac{\gamma+\theta}{2}}(\cdot)\|_{L^\infty}^2 + \|u(\cdot,t)\|_{L^\infty} \le C_\eta (1+t)^{-\frac{\gamma+\theta}{\gamma+\theta+1}+\frac{\eta}{2}}, \tag{1.33}$$

for all $t \ge 0$, *where* C_η *is a positive constant depending on* η.

Remark 1.1. The constant $\epsilon_0 = \epsilon_0(G, A, M_0, M, a, n, \gamma, \theta, c_1, c_2)$ is defined in (3.55). There are no smallness assumptions on $\|(M-x)^{1-\frac{\theta}{\gamma}}(\rho_0^\theta)_x^2\|_{L^1}$ and $\|\rho_0^{\frac{1+\theta}{2}}(u_0)_x\|_{L^2}$.

Remark 1.2. The uniqueness of the solution in Theorems 1.1 or 3.1 means that if (ρ_1, u_1, r_1) and (ρ_2, u_2, r_2) are two solutions to the system (1.15)–(1.17) with the same initial data (ρ_0, u_0, r_0) and satisfy regularity conditions in the theorem, then we have that $(\rho_1, u_1, r_1) = (\rho_2, u_2, r_2)$.

Remark 1.3. In particular, the viscosity of the gas is proportional to the square root of the temperature for the hard sphere model (as pointed out in [15, 21]), and the relation between θ and γ is

$$\theta = \frac{\gamma - 1}{2}.$$

Our condition (A1) covers it. Since the Navier–Stokes system with constant viscosity coefficients has the singularity at the vacuum [8, 20], we assume that $\theta > 0$ in (A1).

Remark 1.4. Considering the no-vacuum system in an exterior domain in $\mathbb{R}^3$, Kobayashi and Shibata [9] obtained $\|(\rho-\rho_\infty, u)(\cdot,t)\|_{L^2} \lesssim (1+t)^{-\frac{3}{4}}$ and $\|(\rho-\rho_\infty, u)(\cdot,t)\|_{L^\infty} \lesssim (1+t)^{-\frac{3}{2}}$ when ρ_∞ is a positive constant. Considering the no-vacuum system in $\mathbb{R}^n$ $(n \ge 3)$, Ukai, Yang, and Zhao [18] obtained $\|(\rho-\rho_\infty, u)(\cdot,t)\|_{L^2\cap L^\infty} \le C(1+t)^{-\frac{n}{4}+\epsilon}$ when ρ_∞ is close to a positive constant. Considering the one-dimensional system with a degenerate viscosity coefficient, we [22] obtained $\|(\rho^\gamma - \rho_\infty^\gamma, u)(\cdot,t)\|_{L^\infty} \le$

$C(1+t)^{-\frac{1}{2}}$ when the external force f_∞ is close to a positive constant N_0 and the stationary density ρ_∞ is close to $(\frac{N_0(M-x)}{A})^{\frac{1}{\gamma}}$. Since $\frac{\gamma+\theta}{\gamma+\theta+1} > \frac{1}{2}$ and $\frac{3\gamma+3\theta-1}{4(\gamma+\theta+1)} > \frac{1}{2}$ (if $\gamma+\theta > 3$), it is easy to see that our results in Theorem 1.1 are better than that in [22]. Using similar arguments as that in Theorem 1.1, we can also obtain similar results in the one-dimensional case, which are better than that in [22]. For example, the stabilization rate estimate $\|(\rho^\gamma - \rho_\infty^\gamma, u)(\cdot,t)\|_{L^\infty} \le C(1+t)^{-\frac{1}{2}}$ in [22] can be replaced by $\|(\rho^\gamma - \rho_\infty^\gamma, u)(\cdot,t)\|_{L^\infty} \le C_\eta(1+t)^{-\frac{\gamma+\theta}{\gamma+\theta+1}+\frac{\eta}{2}}$, $\|\rho^{\frac{\gamma+\theta}{2}}(\cdot,t) - \rho_\infty^{\frac{\gamma+\theta}{2}}(\cdot)\|_{L^\infty} \le C_\eta(1+t)^{-\frac{\gamma+\theta}{2(\gamma+\theta+1)}+\frac{\eta}{4}}$, $t \ge 0$ for any $\eta > 0$.

Remark 1.5. Since the information of the dimension n mainly appear in the index of the radii r, and we consider only the system with a solid core $r \ge a > 0$ in this paper; our results cannot show the effect of the dimension. In [23], we studied the global behavior of the solution to a similar problem with a positive external pressure and without a solid core, and we obtained the stabilization rate estimates for the solution of exponential type. The admissible range of the parameter $\frac{\lambda}{\mu}$ depends on the dimension n in [23]. We will study the system without a solid core $r \ge 0$ and with degenerate viscosity coefficients in the future, and we will guess that stabilization rate estimates of the solution cannot be better than that in Theorem 1.1.

THEOREM 1.2 (continuous dependence). *For each $i = 1, 2$, let (ρ_i, u_i, r_i) be the solution to the system* (1.15)–(1.17) *with the initial data $(\rho_{0i}, u_{0i}, r_{0i})$, which satisfies regularity conditions in Theorem* 1.1. *Then, we have*

$$\int_0^M \left((u_1-u_2)^2 + \rho_1^{1-\theta}\rho_2^{2\theta-4}(\rho_1-\rho_2)^2 + \rho_1^\theta\rho_2^{-1}(r_1-r_2)^2\right)dx$$
$$\le Ce^{Ct}\int_0^M \left((u_{01}-u_{02})^2 + \rho_{01}^{1-\theta}\rho_{02}^{2\theta-4}(\rho_{01}-\rho_{02})^2 + \rho_{01}^\theta\rho_{02}^{-1}(r_{01}-r_{02})^2\right)dx$$

for all $t \ge 0$.

Remark 1.6. Using similar arguments as that in [3], we can easily obtain such continuous dependence of the solution on the initial data and omit the details.

Remark 1.7. If we ignore the influence of self-gravitation, i.e., $f_\infty(m,r) = G\frac{M_0}{r^{n-1}}$, with $M_0 > 0$, then we can also obtain the same results in Theorems 1.1–1.2.

We now briefly review the previous works in this direction. For the related free boundary problem of one-dimensional isentropic fluids with density-dependent viscosity (like $\mu(\rho) = c\rho^\theta$), see [11, 15, 19, 21, 22] and the references therein. For the related stabilization rate estimates of the one-dimensional free boundary problem, see [5, 12, 16, 22], etc. For the spherically symmetric solutions of the Navier–Stokes equations with a free boundary, see [2, 4, 13, 14, 23, 24], etc. Also see Bresch and Desjardins [1], Lions [10], and Vaigant and Kazhikhov [17] for multidimensional isentropic fluids.

The rest of this paper is organized as follows. First, we obtain the existence and the uniqueness of the solution to the stationary problem in section 2. In section 3, we will prove some a priori estimates and extend the local solution in [3] to the global solution in time. In section 4, we obtain the stabilization rate estimates of the solution.

2. The stationary problem. Zlotnik and Ducomet [24] obtained the existence of the positive solution to the stationary problem with a positive external pressure. Using similar arguments as that in [24], we can obtain the following results for the stationary problem without an external pressure. We start with a proof of the existence of a nonnegative solution to the Lagrangian stationary problem.

PROPOSITION 2.1. *If*

$$\gamma > \frac{2n-2}{n} \tag{2.1}$$

or

$$\gamma = \frac{2n-2}{n} \quad \textit{and} \quad \left(\frac{n\gamma}{(\gamma-1)}M^{\frac{\gamma-1}{\gamma}}\right)^{\frac{2n-2}{n}} < \frac{G}{A}\left(\frac{M}{2}+M_0\right), \tag{2.2}$$

or

$$n>2,\ 1<\gamma<\frac{2n-2}{n} \quad \textit{and} \quad \delta_6^{\gamma}\left(a^n+\frac{n\gamma}{\delta_6(\gamma-1)}M^{\frac{\gamma-1}{\gamma}}\right)^{\frac{2n-2}{n}} \le \frac{G}{A}\left(\frac{M}{2}+M_0\right), \tag{2.3}$$

where $\delta_6 = a^{-n}(1-\frac{\gamma n}{2n-2})\frac{2n-2}{\gamma-1}M^{\frac{\gamma-1}{\gamma}}$, *then the Lagrangian stationary problem* (1.18)–(1.19) *has a nonnegative solution* $\rho_\infty \in W^{1,\beta}([0,M])$ *satisfying* $C^{-1}(M-x)^{\frac{1}{\gamma}} \le \rho_\infty(x) \le C(M-x)^{\frac{1}{\gamma}}$, *where* $\beta \in [1, \frac{\gamma}{\gamma-1})$ *is a constant.*

Proof. We introduce the nonlinear operator

$$I : K \to W^{1,\beta}([0,M]),$$

where $K = \{f \in C([0,M]) | f \ge 0, \|\frac{(M-x)^{\frac{1}{\gamma}}}{f(x)}\|_{L^\infty} < \infty, \|\frac{f(x)}{(M-x)^{\frac{1}{\gamma}}}\|_{L^\infty} < \infty,\}$, by setting

$$I(f)(x) = \left(\frac{\int_x^M G\frac{M_0+y}{r_f^{2n-2}(y)}dy}{A}\right)^{\frac{1}{\gamma}}, \text{with } r_f^n(x) = a^n + n\int_0^x f^{-1}(y)dy,\ x\in[0,M].$$

We can restate the problem (1.18)–(1.19) as the fixed-point problem

$$\rho_\infty = I(\rho_\infty). \tag{2.4}$$

For all $f \in K_{\delta_1,\delta_2} = \{f \in K | \delta_1(M-x)^{\frac{1}{\gamma}} \le f(x) \le \delta_2(M-x)^{\frac{1}{\gamma}}\}$, with $0 < \delta_1 \le \delta_2 < \infty$, we have

$$a^n \le r_f^n(x) \le a^n + \frac{n\gamma}{\delta_1(\gamma-1)}M^{\frac{\gamma-1}{\gamma}} := B^n$$

and

$$\left(\frac{G(\frac{M}{2}+M_0)}{AB^{2n-2}}\right)^{\frac{1}{\gamma}}(M-x)^{\frac{1}{\gamma}} \le I(f)(x) \le \left(\frac{G(M+M_0)}{Aa^{2n-2}}\right)^{\frac{1}{\gamma}}(M-x)^{\frac{1}{\gamma}}, x\in[0,M].$$

If $\gamma > \frac{2n-2}{n}$, then $I(K_{\delta_3,\delta_4}) \subset K_{\delta_3,\delta_4}$, where $\delta_4 = (\frac{G(M+M_0)}{Aa^{2n-2}})^{\frac{1}{\gamma}}$ and δ_3 is a positive constant satisfying $\delta_3^{\gamma}(a^n + \frac{n\gamma}{\delta_3(\gamma-1)}M^{\frac{\gamma-1}{\gamma}})^{\frac{2n-2}{n}} \le \frac{G}{A}(\frac{M}{2}+M_0)$. And one can immediately verify that I is a compact operator on K_{δ_3,δ_4}. Since K_{δ_3,δ_4} is a convex closed bounded nonempty subset of $C([0,M])$, the problem (2.4) has a solution $\rho \in K_{\delta_3,\delta_4}$ by Schauder's fixed-point theorem.

Similarly, if $\gamma = \frac{2n-2}{n}$ and $(\frac{n\gamma}{(\gamma-1)}M^{\frac{\gamma-1}{\gamma}})^{\frac{2n-2}{n}} < \frac{G}{A}(\frac{M}{2}+M_0)$, then $I(K_{\delta_5,\delta_4}) \subset K_{\delta_5,\delta_4}$, where $\delta_5 = a^{-n}[(\frac{G}{A}(\frac{M}{2}+M_0))^{\frac{n}{2n-2}} - \frac{n\gamma}{(\gamma-1)}M^{\frac{\gamma-1}{\gamma}}]$, and problem (2.4) has a solution $\rho \in K_{\delta_5,\delta_4}$.

Similarly, if $n>2$, $1<\gamma<\frac{2n-2}{n}$ and $\delta_6^{\gamma}(a^n + \frac{n\gamma}{\delta_6(\gamma-1)}M^{\frac{\gamma-1}{\gamma}})^{\frac{2n-2}{n}} \le \frac{G}{A}(\frac{M}{2}+M_0)$, then $I(K_{\delta_6,\delta_4}) \subset K_{\delta_6,\delta_4}$, and problem (2.4) has a solution $\rho \in K_{\delta_6,\delta_4}$. □

Similar to [24], we say a stationary solution $(\rho_\infty, r_\infty^n)$ is *statically stable* if

$$
\begin{aligned}
J[W] &:= \int_0^M \left(\gamma A\rho_\infty^{1+\gamma}W_x^2 - (2n-2)G(M_0+x)r_\infty^{2-3n}W^2\right)dx \\
&\ge \delta_7\int_0^M \left((M-x)^{\frac{1+\gamma}{\gamma}}W_x^2 + W^2\right)dx
\end{aligned} \tag{2.5}
$$

for some $\delta_7>0$ and all $W\in K_1$,

$$
K_1 = \left\{ f\in C([0,M]) \,\middle|\, f(0)=0, \|(M-x)^{\frac{1}{\gamma}}f'(x)\|_{L^\infty} < \infty, \left\|\frac{1}{(M-x)^{\frac{1}{\gamma}}f'(x)}\right\|_{L^\infty} < \infty \right\}.
$$

Now, the static potential energy takes the following form:

$$
S[V] = \int_0^M \left(\frac{A}{\gamma-1}(V_x)^{1-\gamma} + \int_{\frac{a^n}{n}}^V G(M_0+x)(nh)^{\frac{2-2n}{n}}dh\right)dx. \tag{2.6}
$$

We call

$$
\begin{aligned}
V\in K_2 = \Bigg\{ f\in C([0,M]) \,\Big|\, & f(0) = \frac{a^n}{n}, \|(M-x)^{\frac{1}{\gamma}}f'(x)\|_{L^\infty} < \infty, \\
& \left\|\frac{1}{(M-x)^{\frac{1}{\gamma}}f'(x)}\right\|_{L^\infty} < \infty \Bigg\}
\end{aligned}
$$

is a point of *local quadratic minimum of* S if

$$
S[V+W] - S[V] \ge \delta_8\int_0^M \left((M-x)^{\frac{1+\gamma}{\gamma}}W_x^2 + W^2\right)dx, \tag{2.7}
$$

for all $W\in K_1$ and $\|(M-x)^{\frac{1}{\gamma}}W_x\|_{L^\infty([0,M])} + \|W\|_{L^\infty} \le \delta_9$, for some $\delta_8>0$ and $\delta_9>0$.

PROPOSITION 2.2. *If* $\gamma > \frac{2n-2}{n}$ *and* ρ_∞ *is a solution of the problem* (1.18)–(1.19) *satisfying* $\rho_\infty \in W^{1,\beta}([0,M])$ *and* $C^{-1}(M-x)^{\frac{1}{\gamma}} \le \rho_\infty(x) \le C(M-x)^{\frac{1}{\gamma}}$, *then we have that* (2.5) *and* (2.7) *hold, with* $V = V_\infty = \frac{r_\infty^n}{n}$.

Proof. From $r_\infty \ge a$, $(A\rho_\infty^\gamma)_x = -G\frac{M_0+x}{r_\infty^{2n-2}}$ and $(r_\infty^n)_x = n\rho_\infty^{-1}$, using integration by parts, we have

$$
\begin{aligned}
J[W] &= \int_0^M \left(\gamma A\rho_\infty^{1+\gamma}W_x^2 + (2n-2)A(\rho_\infty^\gamma)_x r_\infty^{-n}W^2\right)dx \\
&= \int_0^M \Big(\gamma A\rho_\infty^{1+\gamma}W_x^2 - 2(2n-2)A\rho_\infty^\gamma r_\infty^{-n}WW_x \\
&\quad + n(2n-2)A\rho_\infty^{\gamma-1}r_\infty^{-2n}W^2\Big)dx \text{ for all } W\in K_1.
\end{aligned}
$$

If $\gamma > \frac{2n-2}{n}$, we have

$$J[W] \geq C^{-1} \int_0^M \left((M-x)^{\frac{1+\gamma}{\gamma}} W_x^2 + (M-x)^{\frac{\gamma-1}{\gamma}} W^2 \right) dx.$$

From $r_\infty \geq a$ and $(A\rho_\infty^\gamma)_x = -G\frac{M_0+x}{r_\infty^{2n-2}}$, using integrating by parts and the Cauchy–Schwarz inequality, we have

$$\begin{aligned} \int_0^M W^2 dx &\leq C \int_0^{\frac{M}{2}} (M-x)^{1-\frac{1}{\gamma}} W^2 dx + C \int_0^M G \frac{M_0+x}{r_\infty^{2n-2}} W^2 dx \\ &= C \int_0^{\frac{M}{2}} (M-x)^{1-\frac{1}{\gamma}} W^2 dx - C \int_0^M A(\rho_\infty^\gamma)_x W^2 dx \\ &\leq C \int_0^M \left((M-x)^{\frac{1+\gamma}{\gamma}} W_x^2 + (M-x)^{\frac{\gamma-1}{\gamma}} W^2 \right) dx, \end{aligned} \tag{2.8}$$

then we can immediately obtain (2.5), with some $\delta_7 = \delta_7(G, A, M_0, M, a, n, \gamma)$.

Similarly, we obtain

$$\begin{aligned} & S[V_\infty + W] - S[V_\infty] \\ &= \frac{1}{2} \int_0^M \Big\{ A[\gamma + O(|(M-x)^{\frac{1}{\gamma}} W_x|)] \rho_\infty^{1+\gamma} W_x^2 \\ &\quad - [2n-2+O(|W|)] G(M_0+x) r_\infty^{2-3n} W^2 \Big\} dx \\ &\geq \frac{\delta_7}{2} \int_0^M \left((M-x)^{\frac{1+\gamma}{\gamma}} W_x^2 + W^2 \right) dx \\ &\quad + \int_0^M \Big\{ O(|(M-x)^{\frac{1}{\gamma}} W_x|) \rho_\infty^{1+\gamma} W_x^2 + O(|W|) G(M_0+x) r_\infty^{2-3n} W^2 \Big\} dx \end{aligned}$$

for all $W \in K_1$. Here, $O(d)$ means $O(d) \to 0$ as $d \to 0$. If $\gamma > \frac{2n-2}{n}$, choosing δ_9 small enough, we can immediately obtain (2.7), with some $(\delta_8, \delta_9)(G, A, M_0, M, a, n, \gamma)$. ∎

Using a similar argument as that in Proposition 2.2, we could obtain the following uniqueness result.

PROPOSITION 2.3. *Let ρ_∞ be a solution obtained in Proposition* 2.1, *and ρ_2 be another solution of the problem* (1.18)–(1.19) *satisfying $\rho_2 \in W^{1,\beta}([0,M])$ and $C^{-1}(M-x)^{\frac{1}{\gamma}} \leq \rho_2(x) \leq C(M-x)^{\frac{1}{\gamma}}$. If $\gamma > \frac{2n-2}{n}$ and $\|(M-x)^{-\frac{1}{\gamma}}(\rho_\infty - \rho_2)(x)\|_{L^\infty} \leq \delta_{10}$ with a small enough positive constant δ_{10}, then we have that $\rho_\infty(x) = \rho_2(x)$, a.e. $x \in [0,M]$.*

Proof. From (1.18)–(1.19), we have

$$\begin{aligned} A\rho_\infty^\gamma(x) &= \int_x^M G \frac{M_0+y}{r_\infty^{2n-2}} dy, \ r_\infty^n(x) = a^n + n \int_0^x \rho_\infty^{-1}(y) dy, \\ A\rho_2^\gamma(x) &= \int_x^M G \frac{M_0+y}{r_2^{2n-2}} dy, \ r_2^n(x) = a^n + n \int_0^x \rho_2^{-1}(y) dy, \end{aligned}$$

and

$$A(\rho_\infty^\gamma - \rho_2^\gamma) = \int_x^M G(M_0+y) \left(r_\infty^{2-2n} - r_2^{2-2n} \right) dy.$$

Multiplying the above equality by $(\rho_\infty^{-1}-\rho_2^{-1})$, integrating over $[0,M]$, and using the fact that $\int_0^M n(\rho_\infty^{-1}-\rho_2^{-1})(x)\int_x^M g(y)dydx = \int_0^M g(r_\infty^n - r_2^n)dx$, we obtain

$$\begin{aligned}
0 &= \int_0^M \left\{ A(\rho_\infty^\gamma - \rho_2^\gamma)(\rho_\infty^{-1}-\rho_2^{-1}) - G\frac{(M_0+x)}{n}(r_\infty^{2-2n}-r_2^{2-2n})(r_\infty^n - r_2^n)\right\} dx \\
&= \int_0^M \Big\{ - A[\gamma + O(|(M-x)^{-\frac{1}{\gamma}}(\rho_\infty-\rho_2)|)]\rho_\infty^{1+\gamma}(\rho_\infty^{-1}-\rho_2^{-1})^2 \\
&\quad - \frac{A}{n^2}[2n-2+O(|r_\infty^n - r_2^n|)](\rho_\infty^\gamma)_x r_\infty^{-n}(r_\infty^n - r_2^n)^2\Big\} dx \\
&\le -C^{-1}\int_0^M \left((M-x)^{\frac{1+\gamma}{\gamma}}(\rho_\infty^{-1}-\rho_2^{-1})^2 + (r_\infty^n - r_2^n)^2\right) dx,
\end{aligned}$$

when $\gamma > \frac{2n-2}{n}$ and δ_{10} is small enough. Thus, we can immediately obtain that $\rho_\infty = \rho_2$. □

Now, we shall use the shooting method to prove the uniqueness of the solution $\rho_\infty \in K$.

PROPOSITION 2.4. *Under the assumption* (2.1)*, the Lagrangian stationary problem* (1.18)–(1.19) *has a unique solution* $\rho_\infty \in K$.

Proof. We consider the Cauchy problem

$$(A\rho^\gamma)_x = -G(M_0+x)(nV)^{\frac{2-2n}{n}},\ (V)_x = \rho^{-1},\ x\in(0,M), \tag{2.9}$$

$$\rho\big|_{x=0} = \sigma,\ V\big|_{x=0} = \frac{a^n}{n} \tag{2.10}$$

for the unknown functions $\rho(\sigma,x)$ and $V(\sigma,x)$, where $\sigma>0$ is the shooting parameter. Thus, for each $\sigma>0$, using the classical ODE theory, there exists a unique solution to the problem (2.9)–(2.10) satisfying $\rho(\sigma,x)>0$ for $x\in[0,M_\sigma)$, where either $\rho\big|_{x=M_\sigma} = 0$ and $M_\sigma\in(0,M)$ or $M_\sigma = M$.

Clearly, if $\rho_\infty\in K$ is a solution to the problem (1.18)–(1.19), then ρ_∞ satisfies (2.9)–(2.10) for some $\sigma_0>0$, and $M_{\sigma_0} = M$. We will show that it is possible only for one value of σ. Using similar arguments as that in [7, section V.3], we obtain that $(\partial_\sigma\rho^\gamma, \partial_\sigma V)$ is well defined and satisfies the linear Cauchy problem

$$A(\partial_\sigma\rho^\gamma)_x = (2n-2)G(M_0+x)(nV)^{\frac{2-3n}{n}}\partial_\sigma V,\ (\partial_\sigma V)_x = -\frac{1}{\gamma}\rho^{-\gamma-1}\partial_\sigma\rho^\gamma,\ x\in[0,M_\sigma), \tag{2.11}$$

$$\partial_\sigma\rho^\gamma\big|_{x=0} = 1,\ \partial_\sigma V\big|_{x=0} = 0. \tag{2.12}$$

It is easy to see that

$$\partial_\sigma\rho^\gamma > 0,\ (\partial_\sigma V)_x < 0,\ \text{and } \partial_\sigma V < 0$$

hold on $(0,M_4)$, where either $\partial_\sigma\rho^\gamma\big|_{x=M_4} = 0$ and $M_4\in(0,M_\sigma)$ or $M_4 = M_\sigma$. We claim that only $M_4 = M_\sigma$ can occur.

Assume that $M_4\in(0,M_\sigma)$. Letting $\phi = A\rho^\gamma(\partial_\sigma V)_x + \frac{n}{2n-2}A\partial_\sigma\rho^\gamma(V)_x$, from (2.9) and (2.11), we have

$$\int_0^{M_4}\phi dx = \left\{A\rho^\gamma\partial_\sigma V + \frac{n}{2n-2}A\partial_\sigma\rho^\gamma V\right\}\Big|_0^{M_4}.$$

By the estimates $\rho(\sigma, M_4) > 0$, $\partial_\sigma \rho^\gamma\big|_{x=M_4} = 0$, $\partial_\sigma V\big|_{x=M_4} < 0$ and the initial conditions (2.10) and (2.12), we get

$$\int_0^{M_4} \phi dx < 0.$$

On the other hand, from (2.9) and (2.11), we have

$$\phi = A\rho^{-1}\partial_\sigma \rho^\gamma \left(\frac{n}{2n-2} - \frac{1}{\gamma}\right) > 0,\ x \in (0, M_4).$$

It is a contradiction.

Thus, we obtain

$$\rho(\sigma, x) > 0,\ \partial_\sigma \rho(\sigma, x) > 0,\ x \in (0, M_\sigma),$$

and M_σ is nondecreasing on $\sigma \in (0, \infty)$. Therefore, for each fixed point $x \in [0, M_b)$, the function $\rho(\sigma, x)$ is strictly increasing on $\sigma \geq b$.

If there exists $\sigma_1 \neq \sigma_0$ such that $M_{\sigma_1} = M_{\sigma_0} = M$ and $\rho(\sigma_1, x) \in K$, then there exists $\min\{\sigma_0, \sigma_1\} < \sigma_2 < \max\{\sigma_0, \sigma_1\}$ such that $0 < \|(M-x)^{-\frac{1}{\gamma}}(\rho(\sigma_2, x) - \rho(\sigma_0, x))\|_{L^\infty} \leq \delta_{10}$. From Proposition 2.3, we have that $\rho(\sigma_2, x) = \rho(\sigma_0, x) = \rho_\infty(x)$, which is a contradiction. Thus, we finish the proof of Proposition 2.4. □

Using the properties of the transformation (1.14) and Propositions 2.1–2.4, we can immediately obtain the following proposition.

PROPOSITION 2.5. *Under the assumption* (2.1)*, the Eulerian stationary problem* (1.12)–(1.13) *has a unique solution* (ρ_∞, l_∞) *satisfying* $\rho_\infty(r) \sim (l_\infty^n - r^n)^{\frac{1}{\gamma-1}}$, $(\rho_\infty)_r < 0$, $a < r < l_\infty$, *with* $l_\infty < +\infty$.

Remark 2.1. The uniqueness of the solution in Proposition 2.5 means that if $(\rho_{\infty 1}, l_{\infty 1})$ and $(\rho_{\infty 2}, l_{\infty 2})$ are two solutions to the Eulerian stationary problem (1.12)–(1.13) with the same total mass M, and satisfy $\rho_{\infty i}(r) \sim (l_{\infty i}^n - r^n)^{\frac{1}{\gamma-1}}$, $i = 1, 2$, then we have that $(\rho_{\infty 1}, l_{\infty 1}) = (\rho_{\infty 2}, l_{\infty 2})$.

3. Global existence. Using similar arguments as that in [3], we obtain the following local existence and uniqueness result and omit the proof.

THEOREM 3.1 (local result). *Under the assumptions in Theorem* 1.1*, there is a positive constant* $T_1 > 0$ *such that the free boundary problem* (1.15)–(1.17) *admits a unique weak solution* $(\rho, u, r)(x, t)$ *on* $[0, M] \times [0, T_1]$ *in the sense that*

$$\rho(x,t), u(x,t), r(x,t) \in L^\infty([0,M] \times [0,T_1]) \cap C^1([0,T_1]; L^2([0,M])),$$

$$\rho^{\theta+1}\partial_x(r^{n-1}u) \in L^\infty([0,M] \times [0,T_1]) \cap C^{\frac{1}{2}}([0,T_1]; L^2([0,M])),$$

$$\partial_x r^{n-1}, \partial_x(r^{n-2}u) \in L^\infty([0,T_1], L^1([0,M])),$$

and the following equations hold:

$$\partial_t \rho = -\rho^2 \partial_x(r^{n-1}u), \qquad \rho(x,0) = \rho_0,$$

$$\partial_t r(x,t) = u(x,t),\ r^n(x,t) = a^n + n\int_0^x \rho^{-1}(y,t)dy, \tag{3.1}$$

$$(r^\beta(\rho^\theta)_x)_t = -\frac{\theta r^{1+\beta-n}}{2c_1+c_2}u_t - \frac{\theta}{2c_1+c_2}\left(Ar^\beta(\rho^\gamma)_x + r^{1+\beta-n}f\right), \tag{3.2}$$

$$
\begin{aligned}
&(2c_1+c_2)\rho^{1+\theta}(r^{n-1}u)_x\\
&= A\rho^\gamma + 2c_1(n-1)\rho^\theta\frac{u}{r} + \int_x^M \left\{-\frac{u_t}{r^{n-1}} + 2c_1(n-1)\rho^\theta\left(\frac{u}{r}\right)_x - \frac{f}{r^{n-1}}\right\}dy,
\end{aligned}
\tag{3.3}
$$

for almost all $x\in[0,M]$, *any* $t\in[0,T_1]$, *where* $\beta=\frac{2(n-1)c_1\theta}{2c_1+c_2}$,

$$
\begin{aligned}
&\int_0^\infty\int_0^M [u\psi_t + (P-\rho(\lambda+2\mu)(r^{n-1}u)_x)(r^{n-1}\psi)_x\\
&+2(n-1)\mu(r^{n-2}u\psi)_x - f(x,r,t)\psi]dxdt + \int_0^M u_0(x)\psi(x,0)dx = 0
\end{aligned}
\tag{3.4}
$$

for any test function $\psi(x,t)\in C_0^\infty((0,M]\times[0,T_1))$. *Furthermore, we have*

$$
\frac{N_1}{3}(1-x)^{\frac{1}{\gamma}} \le \rho(x,t) \le 3N_2(1-x)^{\frac{1}{\gamma}},\ (x,t)\in[0,M]\times[0,T_1],
\tag{3.5}
$$

$$
\begin{aligned}
&(M-x)^{-\frac{1}{\gamma}}\rho(x,t)\in C([0,T_1];L^\infty([0,M])),\\
&(M-x)^{\frac{\gamma-\theta}{2\gamma}}(\rho^\theta)_x,\ \rho_t,\ u_t\in L^\infty([0,T_1];L^2([0,M])),
\end{aligned}
\tag{3.6}
$$

and

$$
\rho^{\frac{\theta+1}{2}}u_{xt}\in L^2([0,M]\times[0,T_1]),\ \rho\partial_x u\in L^\infty([0,M]\times[0,T_1]).
$$

Remark 3.1. From $(1.15)_1$, (3.5), and $\rho\partial_x u\in L^\infty_{t,x}$, we have that $(M-x)^{-\frac{1}{\gamma}}\partial_t\rho\in L^\infty_{tx}$. Thus, (3.6) holds.

Assume the maximum existence time of the weak solution in Theorem 3.1 is T^*. In this section, under the small assumptions on the initial data, we will obtain the following a priori estimates and prove that $T^*=\infty$. In the following, we may assume that $(\rho,u,r)(x,t)$ is suitably smooth since the following estimates are valid for the solutions with the regularities indicated in Theorem 3.1 by using the Friedrichs mollifier.

From (1.9), (1.18), and Proposition 2.1, we could easily obtain the following lemma.

LEMMA 3.1. *Under the assumptions of Theorem* 1.1, *we have*

$$
A\rho_\infty^\gamma(x) = \int_x^M G\frac{M_0+y}{r_\infty^{2n-2}}dy,
\tag{3.7}
$$

$$
C^{-1}(M-x)^{\frac{1}{\gamma}} \le \rho_\infty \le C(M-x)^{\frac{1}{\gamma}},\ r_\infty(x)\in[a,C],
\tag{3.8}
$$

and

$$
\frac{d}{dx}\left(A\rho_\infty^\gamma(x)\right) = -G\frac{M_0+x}{r_\infty^{2n-2}},\ C^{-1}\le(M-x)^{1-\frac{1}{\gamma}}\frac{d}{dx}\rho_\infty(x)\le C
\tag{3.9}
$$

for all $x\in[0,M]$.

LEMMA 3.2. *Under the assumptions of Theorem* 1.1, *we have*

$$
\begin{aligned}
&\frac{d}{dt}\int_0^M\left(\frac{1}{2}u^2 + \frac{A\rho^{\gamma-1}}{\gamma-1} + \int_a^r G\frac{M_0+x}{s^{n-1}}ds\right)dx\\
&+\int_0^M\left\{\left(\frac{2}{n}c_1+c_2\right)\rho^{1+\theta}[\partial_x(r^{n-1}u)]^2 + \frac{2(n-1)}{n}c_1\rho^{1+\theta}\left(r^{n-1}u_x - \frac{u}{r\rho}\right)^2\right\}dx\\
&= -\int_0^M \Delta f u dx,\ t\in[0,T^*).
\end{aligned}
\tag{3.10}
$$

Proof. Multiplying $(1.15)_2$ by u, integrating the resulting equation over $[0, M]$, and using integration by parts and the boundary conditions (1.17), we obtain

(3.11)
$$\frac{d}{dt}\int_0^M \frac{1}{2}u^2 dx + \int_0^M \left\{(2c_1+c_2)\rho^{1+\theta}[\partial_x(r^{n-1}u)]^2 - 2c_1(n-1)\rho^\theta \partial_x(r^{n-2}u^2)\right\} dx$$
$$= \int_0^M A\rho^\gamma \partial_x(r^{n-1}u)dx - \int_0^M f u dx.$$

From (1.15), we have

$$\int_0^M A\rho^\gamma \partial_x(r^{n-1}u)dx = -\frac{d}{dt}\int_0^M \frac{A}{\gamma-1}\rho^{\gamma-1}dx, \tag{3.12}$$

$$-\int_0^M f u dx = -\frac{d}{dt}\int_0^M \int_a^r G\frac{M_0+x}{s^{n-1}}dsdx - \int_0^M \Delta f u dx, \tag{3.13}$$

and

$$\begin{aligned}&(2c_1+c_2)\rho^{1+\theta}[\partial_x(r^{n-1}u)]^2 - 2c_1(n-1)\rho^\theta\partial_x(r^{n-2}u^2)\\ &= \left(\frac{2}{n}c_1+c_2\right)\rho^{1+\theta}[\partial_x(r^{n-1}u)]^2 + \frac{2(n-1)}{n}c_1\rho^{1+\theta}\left(r^{n-1}u_x - \frac{u}{r\rho}\right)^2.\end{aligned} \tag{3.14}$$

From (3.11)–(3.14), we can immediately obtain (3.10). ∎

Now, using the classical continuation method, we will obtain the estimate of $\|(M-x)^{-\frac{1}{\gamma}}(\rho-\rho_\infty)\|_{L^\infty}$.

Claim 1. *Under the assumptions of Theorem* 1.1*, for any* $T \in (0, T^*)$*, if*

$$I(t) = \|g(\cdot,t) - g_\infty\|_{L^\infty} \le 2\epsilon_1 \ \forall\ t \in [0,T], \tag{3.15}$$

where $g(x,t) = (M-x)^{-\frac{1}{\gamma}}\rho(x,t)$ *and* $g_\infty(x) = (M-x)^{-\frac{1}{\gamma}}\rho_\infty(x)$*, then*

$$I(t) \le \epsilon_1 \ \forall\ t \in [0,T], \tag{3.16}$$

where

$$\epsilon_1 = \epsilon_0 + C_{10}\epsilon_0^{\frac{\theta}{\gamma}2^{-N_4-1}},$$

and positive constants N_4 *and* C_{10} *are defined in Lemmas* 3.6 *and* 3.8*, respectively.*

Using the results in Lemmas 3.3–3.8, we can finish the proof of Claim 1.

Lemma 3.3. *Under the assumptions of Theorem* 1.1 *and* (3.15)*, if* ϵ_1 *is small enough, then there exists a positive constant* $C_1 = C_1(G, A, M_0, M, a, n, \gamma)$ *such that*

$$C_1^{-1}(M-x)^{\frac{1}{\gamma}} \le \rho(x,t) \le C_1(M-x)^{\frac{1}{\gamma}}, \tag{3.17}$$

$$r(x,t) \in [a, C_1] \tag{3.18}$$

for all $t \in [0,T]$ *and* $x \in [0,M]$.

Proof. From $(1.15)_3$, (3.15), and Lemma 3.1, we can easily obtain the estimate (3.17) and (3.18), when $4\epsilon_1 < \min_{x\in[0,M]} g_\infty$. ∎

LEMMA 3.4. *Under the assumptions of Lemma* 3.3*, if* ϵ_1 *is small enough, then there exists a positive constant* $C_2 = C_2(G, A, M_0, M, a, n, \gamma, \theta, c_1, c_2)$ *such that*

$$\int_0^M \left\{u^2 + (M-x)^{\frac{\gamma-1}{\gamma}}(g-g_\infty)^2 + (r-r_\infty)^2\right\} dx \le C_2\epsilon_0^2, \tag{3.19}$$

and

$$\int_0^t \|u(\cdot,s)\|_{L^\infty}^2 ds + \int_0^t \int_0^M \left(\rho^{\theta+1}u_x^2 + \rho^{\theta-1}u^2\right)(x,s)dxds \le C_2\epsilon_0^2 \tag{3.20}$$

for all $t \in [0,T]$.

Proof. From (2.6), (3.7), and (3.10), we have

$$\begin{aligned}
&\frac{d}{dt}\left(\int_0^M \frac{1}{2}u^2dx + S[V] - S[V_\infty]\right) \\
&+ \int_0^M \left\{\left(\frac{2}{n}c_1 + c_2\right)\rho^{1+\theta}[\partial_x(r^{n-1}u)]^2 + \frac{2(n-1)}{n}c_1\rho^{1+\theta}\left(r^{n-1}u_x - \frac{u}{r\rho}\right)^2\right\} dx \\
&= -\int_0^M \Delta f u dx,
\end{aligned} \tag{3.21}$$

where $V_\infty = \frac{r_\infty^n}{n}$ and $V = \frac{r^n}{n}$. From (2.7), (3.17)–(3.18), and Proposition 2.2, we have

$$\begin{aligned}
&C^{-1}\int_0^M (M-x)^{\frac{\gamma-1}{\gamma}}(g-g_\infty)^2 + (r-r_\infty)^2 dx \\
&\le S[V] - S[V_\infty] \le C\int_0^M (M-x)^{\frac{\gamma-1}{\gamma}}(g-g_\infty)^2 + (r-r_\infty)^2 dx,
\end{aligned} \tag{3.22}$$

when $\|(M-x)^{\frac{1}{\gamma}}(\rho^{-1} - \rho_\infty^{-1})\|_{L_x^\infty} + \|\frac{1}{n}(r^n - r_\infty^n)\|_{L_x^\infty} \le C_3\epsilon_1 \le \delta_9$. From (1.24), (3.17)–(3.18), and (3.21)–(3.22), we obtain

$$\begin{aligned}
&\int_0^M \left\{u^2 + (M-x)^{\frac{\gamma-1}{\gamma}}(g-g_\infty)^2 + (r-r_\infty)^2\right\} dx \\
&+ \int_0^t \int_0^M \left\{\rho^{\theta+1}u_x^2 + \rho^{\theta-1}u^2\right\} dxds \\
&\le C\epsilon_0^2 + C\int_0^t f_1(s)\|u(\cdot,s)\|_{L^\infty} ds.
\end{aligned} \tag{3.23}$$

Since $\theta \in (0, \gamma-1)$, we obtain

$$\begin{aligned}
|u(x,t)| &= \left|\int_0^x u_x dy\right| \le C\left(\int_0^x \rho^{\theta+1}u_x^2 dy\right)^{\frac{1}{2}}\left(\int_0^x \rho^{-\theta-1}dy\right)^{\frac{1}{2}} \\
&\le C\left(\int_0^x \rho^{\theta+1}u_x^2 dy\right)^{\frac{1}{2}}\left(\int_0^x (M-y)^{-\frac{\theta+1}{\gamma}}dy\right)^{\frac{1}{2}} \le C\left(\int_0^x \rho^{\theta+1}u_x^2 dy\right)^{\frac{1}{2}}
\end{aligned} \tag{3.24}$$

and

$$C\int_0^t f_1(s)\|u(\cdot,s)\|_{L^\infty} ds \le \frac{1}{2}\int_0^t \int_0^M \rho^{\theta+1}u_x^2 dyds + C^2\int_0^t f_1^2 dt. \tag{3.25}$$

From (1.23) and (3.23)–(3.25), we can immediately obtain (3.19)–(3.20). ∎

LEMMA 3.5. *Under the assumptions of Lemma* 3.3*, if* ϵ_1 *is small enough, then there exists a positive constant* $C_4 = C_4(G, A, M_0, M, a, n, \gamma, \theta, c_1, c_2)$ *such that*

$$\begin{aligned}&(1+t)^\alpha \int_0^M \rho_\infty^{\theta-1}(g-g_\infty)^2 dx\\ &+ \int_0^t \int_0^M (1+s)^\alpha \left[\rho_\infty^{\gamma-1}(g-g_\infty)^2 + (r-r_\infty)^2\right] dxds \le C_4\epsilon_0\end{aligned} \tag{3.26}$$

for all $t \in [0,T]$*, where* $\alpha = -\frac{5}{8}$.

Proof. Multiplying $(1.15)_2$ by $(1+t)^\alpha r^{1-n}(\frac{r^n}{n} - \frac{r_\infty^n}{n})$, and integrating over $[0,M]$, using integration by parts and the boundary conditions (1.17), we obtain

$$\begin{aligned}&(1+t)^\alpha \int_0^M \left[A(\rho_\infty^\gamma - \rho^\gamma)(\rho^{-1} - \rho_\infty^{-1}) + G(M_0+x)(r^{2-2n} - r_\infty^{2-2n})\left(\frac{r^n}{n} - \frac{r_\infty^n}{n}\right)\right] dx\\ &= -(1+t)^\alpha \int_0^M \frac{u_t}{r^{n-1}}\left(\frac{r^n}{n} - \frac{r_\infty^n}{n}\right) dx - (1+t)^\alpha \int_0^M \Delta f r^{1-n}\left(\frac{r^n}{n} - \frac{r_\infty^n}{n}\right) dx\\ &+ (1+t)^\alpha \int_0^M (2c_1+c_2)\rho^{1+\theta}\partial_x(r^{n-1}u)(\rho_\infty^{-1} - \rho^{-1}) dx\\ &+ (1+t)^\alpha \int_0^M 2c_1(n-1)\rho^\theta \left(\frac{u}{r}\left(\frac{r^n}{n} - \frac{r_\infty^n}{n}\right)\right)_x dx := \sum_{i=1}^4 B_i.\end{aligned} \tag{3.27}$$

We can rewrite the left-hand side (L.H.S.) of (3.27) as follows:

$$\begin{aligned}\text{L.H.S. of (3.27)} = (1+t)^\alpha \int_0^M \Bigg[&A(\gamma + O(\epsilon_1))\rho_\infty^{\gamma+1}(\rho^{-1} - \rho_\infty^{-1})^2\\ &-(2n-2+O(\epsilon_1))G(M_0+x)r_\infty^{2-3n}\left(\frac{r^n}{n} - \frac{r_\infty^n}{n}\right)^2\Bigg] dx.\end{aligned}$$

Similar to (2.5), we have

$$\text{L.H.S. of (3.27)} \ge C_5(1+t)^\alpha \int_0^M \left[\rho_\infty^{\gamma-1}(g-g_\infty)^2 + (r-r_\infty)^2\right] dx, \tag{3.28}$$

when $\epsilon_1 \le \delta_{10}$ is small enough.

Using (3.17)–(3.19), integration by parts, and Hölder's inequality, we can estimate B_i as follows:

$$\begin{aligned}B_1 &= -\frac{d}{dt}\int_0^M (1+t)^\alpha \frac{u}{r^{n-1}}\left(\frac{r^n}{n} - \frac{r_\infty^n}{n}\right) dx + \alpha(1+t)^{\alpha-1}\int_0^M \frac{u}{r^{n-1}}\left(\frac{r^n}{n} - \frac{r_\infty^n}{n}\right) dx\\ &\quad + (1+t)^\alpha \int_0^M u^2\left(\frac{1}{n} + \frac{(n-1)r_\infty^n}{nr^n}\right) dx\\ &\le -\frac{d}{dt}\int_0^M (1+t)^\alpha \frac{u}{r^{n-1}}\left(\frac{r^n}{n} - \frac{r_\infty^n}{n}\right) dx + C\int_0^M u^2 dx + C\epsilon_0^2(1+t)^{\alpha-1},\end{aligned} \tag{3.29}$$

$$B_2 \le C\epsilon_0(1+t)^{\alpha} f_1, \tag{3.30}$$

$$\begin{aligned} B_3 &= -\frac{2c_1+c_2}{\theta}\int_0^M (\rho^{\theta})_t (1+t)^{\alpha}\left(\frac{1}{\rho_\infty}-\frac{1}{\rho}\right)dy \\ &= -\frac{2c_1+c_2}{\theta}\int_0^M \partial_t h(\rho,\rho_\infty)(1+t)^{\alpha}dx \\ &= -\frac{2c_1+c_2}{\theta}\frac{d}{dt}\int_0^M h(\rho,\rho_\infty)(1+t)^{\alpha}dx + \frac{\alpha(2c_1+c_2)}{\theta}\int_0^M h(\rho,\rho_\infty)(1+t)^{\alpha-1}dx, \end{aligned} \tag{3.31}$$

where $h(\rho,\rho_\infty)=\int_{\rho_\infty}^{\rho}\theta s^{\theta-1}(\frac{1}{\rho_\infty}-\frac{1}{s})ds \sim \rho_\infty^{\theta-1}(g-g_\infty)^2$ and

$$\begin{aligned} B_4 &\le C(1+t)^{\alpha}\int_0^M \left[\rho^{\theta}|u_x| + \rho^{\theta-1}|u|\right]dx \\ &\le C(1+t)^{\alpha}\left[\int_0^M \left(\rho^{\theta+1}u_x^2+\rho^{\theta-1}u^2\right)dx\right]^{\frac{1}{2}}\left[\int_0^M (M-x)^{\frac{\theta-1}{\gamma}}\right]^{\frac{1}{2}} \\ &\le C(1+t)^{\alpha}\left[\int_0^M \left(\rho^{\theta+1}u_x^2+\rho^{\theta-1}u^2\right)dx\right]^{\frac{1}{2}}, \end{aligned} \tag{3.32}$$

since $\gamma+\theta-1>0$. From (3.27)–(3.32), we get

$$\begin{aligned} &\frac{d}{dt}\int_0^M (1+t)^{\alpha}\left\{\frac{u}{r^{n-1}}\left(\frac{r^n}{n}-\frac{r_\infty^n}{n}\right)+\frac{2c_1+c_2}{\theta}h(\rho,\rho_\infty)\right\}dx \\ &+C^{-1}\int_0^M (1+t)^{\alpha-1}\rho_\infty^{\theta-1}(g-g_\infty)^2 + (1+t)^{\alpha}\left\{\rho_\infty^{\gamma-1}\left[(g-g_\infty)^2+(r-r_\infty)^2\right]\right\}dx \\ &\le C(1+t)^{\alpha}\left[\int_0^M \rho_\infty^{\theta+1}u_x^2dx + \|u(\cdot,t)\|_{L^\infty}^2\right]^{\frac{1}{2}} + C\epsilon_0(1+t)^{\alpha}f_1 \\ &+C\int_0^M u^2dx + C\epsilon_0^2(1+t)^{\alpha-1}. \end{aligned}$$

And using (1.23), (3.17)–(3.20), and Hölder's inequality, we can immediately obtain (3.26). □

Let $\epsilon_2 \in (0,\min\{\frac{1}{4},\ \frac{\gamma-\theta-1}{\gamma-\theta},\frac{\gamma-1}{2(3\gamma-1)}\})$ be a constant. Define $\{\beta_j\}$ and $\{\alpha_j\}$ by $\beta_{j+1}=\frac{\beta_j}{2}+\frac{1}{2}-\frac{\epsilon_2}{4}$, $\alpha_j=\frac{\beta_j}{2}-\frac{1}{2}-\frac{\epsilon_2}{4}$ and $\alpha_0=\alpha=-\frac{5}{8}$, $j=0,1,\dots$. Let N_4 be an integer satisfying $\beta_{N_4}\in[1-\epsilon_2,1-\frac{3\epsilon_2}{4})$ and $\alpha_{N_4}\in(-\epsilon_2,-\frac{\epsilon_2}{4})$. It is easy to see that $\beta_0=-\frac{1}{4}+\frac{\epsilon_2}{2}<0$, $\alpha_j\in[-\frac{5}{8},-\frac{\epsilon_2}{4})$ and $\beta_j\in(-\frac{1}{4},1-\frac{3\epsilon_2}{4})$, $j=0,1,\dots,N_4$. Then, the following lemma can be proved by induction.

LEMMA 3.6. *Under the assumptions of Lemma* 3.3, *if* ϵ_1 *is small enough, then there exists a positive constant* $C_7=C_7(G,A,M_0,M,a,n,\gamma,\theta,c_1,c_2)$ *such that*

$$\int_0^M \left\{u^2+(M-x)^{\frac{\gamma-1}{\gamma}}(g-g_\infty)^2+(r-r_\infty)^2\right\}dx \le C_7\epsilon_0^{2^{1-N_4}}(1+t)^{\epsilon_2-1}, \tag{3.33}$$

$$\int_0^t (1+s)^{1-\epsilon_2}\|u(\cdot,s)\|_{L^\infty}^2 ds + \int_0^t\int_0^M (1+s)^{1-\epsilon_2}\left(\rho^{\theta+1}u_x^2+\rho^{\theta-1}u^2\right)dxds \le C_7\epsilon_0^{2^{1-N_4}}, \tag{3.34}$$

and

(3.35)
$$\int_0^M \frac{(M-x)^{\frac{\theta-1}{\gamma}}}{(1+t)^{\epsilon_2}}(g-g_\infty)^2dx+\int_0^t\int_0^M \frac{\rho_\infty^{\gamma-1}(g-g_\infty)^2+(r-r_\infty)^2}{(1+s)^{\epsilon_2}}dxds\le C_7\epsilon_0^{2^{-N_4}}$$

for all $t\in[0,T]$.

Proof. The following estimates can be proved by induction:

$$\int_0^M \left\{u^2+(M-x)^{\frac{\gamma-1}{\gamma}}(g-g_\infty)^2+(r-r_\infty)^2\right\}dx\le C_{j,\epsilon_2}\epsilon_0^{2^{1-j}}(1+t)^{-\beta_j}, \tag{3.36}$$

(3.37)
$$\int_0^t(1+s)^{\beta_j}\|u(\cdot,s)\|_{L^\infty}^2ds+\int_0^t\int_0^M(1+s)^{\beta_j}\left(\rho^{\theta+1}u_x^2+\rho^{\theta-1}u^2\right)(x,s)dxds\le C_{j,\epsilon_2}\epsilon_0^{2^{1-j}},$$

and

$$\begin{aligned}&(1+t)^{\alpha_j}\int_0^M\rho_\infty^{\theta-1}(g-g_\infty)^2dx\\&+\int_0^t\int_0^M(1+s)^{\alpha_j}\left[\rho_\infty^{\gamma-1}(g-g_\infty)^2+(r-r_\infty)^2\right]dxds\le C_{j,\epsilon_2}\epsilon_0^{2^{-j}}\end{aligned} \tag{3.38}$$

for all $t\ge0$, where C_{j,ϵ_2} is a constant depending on j and ϵ_2, $j=0,1,\dots,N_4$.

From (3.19)–(3.20) and (3.26), we obtain that (3.36)–(3.38) hold with $j=0$. Now, suppose that (3.36)–(3.38) hold with $j=k\ge0$. To show (3.36)–(3.37) hold with $j=k+1$, from (3.21), we have

$$\begin{aligned}&\frac{d}{dt}\left\{(1+t)^{\beta_{k+1}}\left(\int_0^M\frac{1}{2}u^2dx+S[V]-S[V_\infty]\right)\right\}\\&+(1+t)^{\beta_{k+1}}\int_0^M\left(\frac{2}{n}c_1+c_2\right)\rho^{1+\theta}[\partial_x(r^{n-1}u)]^2+\frac{2(n-1)}{n}c_1\rho^{1+\theta}\left(r^{n-1}u_x-\frac{u}{r\rho}\right)^2dx\\&=\beta_{k+1}(1+t)^{\alpha_k}\left(\int_0^M\frac{1}{2}u^2(x,t)dx+S[V]-S[V_\infty]\right)-(1+t)^{\beta_{k+1}}\int_0^M\Delta fudx.\end{aligned}$$

Integrating the above equality in $[0,t]$, using (1.24), (3.17)–(3.20), (3.22), and (3.38) with $j=k$ and the fact that $\alpha_k<0$, we obtain

$$\begin{aligned}&(1+t)^{\beta_{k+1}}\int_0^M\left\{u^2(x,t)+(M-x)^{\frac{\gamma-1}{\gamma}}(g-g_\infty)^2+(r-r_\infty)^2\right\}dx\\&+\int_0^t\int_0^M(1+s)^{\beta_{k+1}}\left\{\rho^{\theta+1}u_x^2+\rho^{\theta-1}u^2\right\}dxds\\&\le C\epsilon_0^{2^{-k}}+C\int_0^t(1+s)^{\beta_{k+1}}f_1(s)\|u(\cdot,s)\|_{L^\infty}ds.\end{aligned} \tag{3.39}$$

From (1.23) and (3.25), we can immediately obtain (3.36)–(3.37) with $j=k+1$.

To show (3.38) with $j = k+1$, from (3.27)–(3.28), we have

$$
\begin{aligned}
&(1+t)^{\alpha_{k+1}} \int_0^M \Big[A(\rho_\infty^\gamma - \rho^\gamma)(\rho^{-1} - \rho_\infty^{-1}) \\
&+ G(M_0+x)(r^{2-2n} - r_\infty^{2-2n}) \left(\frac{r^n}{n} - \frac{r_\infty^n}{n} \right) \Big] dx \\
&= -(1+t)^{\alpha_{k+1}} \int_0^M \frac{u_t}{r^{n-1}} \left(\frac{r^n}{n} - \frac{r_\infty^n}{n} \right) dx - (1+t)^{\alpha_{k+1}} \int_0^M \Delta f r^{1-n} \left(\frac{r^n}{n} - \frac{r_\infty^n}{n} \right) dx \\
&+ (1+t)^{\alpha_{k+1}} \int_0^M (2c_1 + c_2) \rho^{1+\theta} \partial_x (r^{n-1} u)(\rho_\infty^{-1} - \rho^{-1}) dx \\
&+ (1+t)^{\alpha_{k+1}} \int_0^M 2c_1 (n-1) \rho^\theta \left(\frac{u}{r} \left(\frac{r^n}{n} - \frac{r_\infty^n}{n} \right) \right)_x dx := \sum_{i=1}^4 E_i, \qquad (3.40)
\end{aligned}
$$

and

$$
\text{L.H.S. of (3.40)} \geq C_8 (1+t)^{\alpha_{k+1}} \int_0^M \left[\rho_\infty^{\gamma-1} (g - g_\infty)^2 + (r - r_\infty)^2 \right] dx. \qquad (3.41)
$$

Similar to (3.29)–(3.32), applying the estimates (3.17)–(3.19), integration by parts, Hölder's inequality, and the fact that $\alpha_{k+1} < 0$, we can estimate E_i as follows:

$$
E_1 \leq -\frac{d}{dt} \int_0^M (1+t)^{\alpha_{k+1}} \frac{u}{r^{n-1}} \left(\frac{r^n}{n} - \frac{r_\infty^n}{n} \right) dx + C \|u\|_{L_x^\infty}^2 + C \epsilon_0^2 (1+t)^{\alpha_{k+1}-1}, \qquad (3.42)
$$

$$
E_2 \leq C \epsilon_0 f_1 (1+t)^{\alpha_{k+1}}, \qquad (3.43)
$$

$$
\begin{aligned}
E_3 = &-\frac{2c_1 + c_2}{\theta} \frac{d}{dt} \int_0^M h(\rho, \rho_\infty)(1+t)^{\alpha_{k+1}} dx \\
&+ \frac{\alpha_{k+1}(2c_1 + c_2)}{\theta} \int_0^M h(\rho, \rho_\infty)(1+t)^{\alpha_{k+1}-1} dx, \qquad (3.44)
\end{aligned}
$$

and

$$
E_4 \leq C (1+t)^{-\frac{1}{2} - \frac{\epsilon_2}{4}} \left[(1+t)^{\beta_{k+1}} \int_0^M \rho^{\theta+1} u_x^2 dx + \|u(\cdot,t)\|_{L^\infty}^2 \right]^{\frac{1}{2}}. \qquad (3.45)
$$

Using (1.23), (3.40)–(3.45), (3.36)–(3.37) with $j = k+1$ and Hölder's inequality, we get

$$
\begin{aligned}
&\int_0^M (1+t)^{\alpha_{k+1}} \rho_\infty^{\theta-1} (g - g_\infty)^2 dx + \int_0^t \int_0^M (1+s)^{\alpha_{k+1}-1} \rho_\infty^{\theta-1} (g - g_\infty)^2 dx ds \\
&+ \int_0^t \int_0^M (1+s)^{\alpha_{k+1}} \left\{ \rho_\infty^{\gamma-1} \left[(g - g_\infty)^2 + (r - r_\infty)^2 \right] \right\} dx ds \\
&\leq C (1+t)^{\alpha_{k+1}} \int_0^M |u||r - r_\infty| dx + C \int_0^M |u_0||r_0 - r_\infty| dx \\
&+ C \int_0^t \left[f_1 (1+s)^{\alpha_{k+1}} + \epsilon_0^2 (1+s)^{\alpha_{k+1}-1} + \|u\|_{L_x^\infty}^2 \right] ds \\
&+ C \int_0^t (1+s)^{-\frac{1}{2} - \frac{\epsilon_2}{4}} \left[(1+t)^{\beta_{k+1}} \left(\int_0^M \rho^{\theta+1} u_x^2 dx + \|u(\cdot,t)\|_{L^\infty}^2 \right) \right]^{\frac{1}{2}} ds \\
&\leq C \epsilon_0^{2^{-(k+1)}} \qquad (3.46)
\end{aligned}
$$

and finish the proof of (3.38) with $j = k+1$. Thus, we show that (3.36)–(3.38) hold for $j = 0, 1, \ldots, N_4$ and obtain (3.33)–(3.35). ∎

From Lemma 3.6, we can obtain the following estimate of the weighted L^2-norm of $g - g_\infty$.

LEMMA 3.7. *Under the assumptions of Lemma* 3.3*, then there exists a positive constant* $C_9 = C_9(G, A, M_0, M, a, n, \gamma, \theta, c_1, c_2)$ *such that*

$$\int_0^M (M-x)^{\frac{\theta-1+(\gamma-\theta)\epsilon_2}{\gamma}}(g-g_\infty)^2 dx \le C_9\epsilon_0^{2^{-N_4}},\ t\in[0,T]. \tag{3.47}$$

Proof. Using (3.33), (3.35), and Hölder's inequality, we have

$$\begin{aligned}
&\int_0^M (M-x)^{\frac{\theta-1+(\gamma-\theta)\epsilon_2}{\gamma}}(g-g_\infty)^2 dx\\
&\le C\left[\int_0^M (1+t)^{1-\epsilon_2}(M-x)^{\frac{\gamma-1}{\gamma}}(g-g_\infty)^2 dx\right]^{\epsilon_2}\left[\int_0^M \frac{(M-x)^{\frac{\theta-1}{\gamma}}}{(1+t)^{\epsilon_2}}(g-g_\infty)^2 dx\right]^{1-\epsilon_2}\\
&\le C\epsilon_0^{2^{-N_4}}. \qquad ∎
\end{aligned}$$

Then, using a similar argument as that in [22], we can finish the proof of Claim 1 in the following lemma.

LEMMA 3.8. *Under the assumptions of Lemma* 3.3*, if* ϵ_0 *is small enough, then there exists a positive constant* $C_{10} = C_{10}(G, A, M_0, M, a, n, \gamma, \theta, c_1, c_2)$ *such that*

$$|g(x,t) - g_\infty(x)| \le C_{10}\epsilon_0^{\frac{\theta}{\gamma}2^{-N_4-1}} \quad \text{for all } (x,t)\in[0,M]\times[0,T]. \tag{3.48}$$

Proof. From (3.2), for any fixed $x \in [0, M]$, we have

$$\begin{aligned}
&I_1(x,t) + \frac{\theta}{2c_1+c_2}\int_0^t Ar^\beta(x,\tau)(\rho^\gamma(x,\tau) - \rho_\infty^\gamma(x))d\tau\\
&= r_0^\beta(x)\rho_0^\theta(x) + I_2(x,t),\ x\in[0,M],\ t\in[0,T],
\end{aligned} \tag{3.49}$$

where

$$\begin{aligned}
I_1(x,t) = {}& r_\infty^\beta(x)\rho^\theta(x,t) - (r_\infty^\beta(x) - r^\beta(x,t))\rho^\theta(x,t)\\
&+ \int_x^M \beta[(r^{\beta-n}\rho^{\theta-1})(y,t) - (r_0^{\beta-n}\rho_0^{\theta-1})(y)]dy\\
&- \frac{\theta}{2c_1+c_2}\int_x^M [(r^{\beta-n+1}u)(y,t) - (r_0^{\beta-n+1}u_0)(y)]dy\\
&+ \frac{\theta(\beta-n+1)}{2c_1+c_2}\int_0^t\int_x^M r^{\beta-n}u^2 dy d\tau
\end{aligned}$$

and

$$\begin{aligned}
&I_2(x,t)\\
&= -\frac{\theta A\beta}{2c_1+c_2}\int_0^t\int_x^M r^{\beta-n}\frac{\rho^\gamma - \rho_\infty^\gamma}{\rho}dy d\tau\\
&\quad + \frac{\theta}{2c_1+c_2}\int_0^t\int_x^M \left\{r^\beta G(M_0+y)(r^{2-2n} - r_\infty^{2-2n}) + r^{\beta-n+1}\Delta f\right\}dy d\tau.
\end{aligned}$$

Using (3.17)–(3.18), (3.47), Hölder's inequality, and the condition $\epsilon_2 < \frac{\gamma-\theta-1}{\gamma-\theta}$, i.e., $\frac{\theta+1+(\gamma-\theta)\epsilon_2}{\gamma} < 1$, we have

$$\begin{aligned}
&|(r-r_\infty)(x)| \le C|r^n - r_\infty^n| \\
&\le C\int_0^x |\rho^{-1}-\rho_\infty^{-1}|dy \le C\int_0^x (M-y)^{-\frac{1}{\gamma}}|g-g_\infty|dy \\
&\le C\left(\int_0^x (M-y)^{\frac{\theta-1+(\gamma-\theta)\epsilon_2}{\gamma}}(g-g_\infty)^2 dy\right)^{\frac{1}{2}}\left(\int_0^x (M-y)^{-\frac{\theta+1+(\gamma-\theta)\epsilon_2}{\gamma}}dy\right)^{\frac{1}{2}} \\
&\le C\epsilon_0^{2^{-N_4-1}} \qquad (3.50)
\end{aligned}$$

and

$$\begin{aligned}
&\int_x^M |\rho^{\theta-1}-\rho_\infty^{\theta-1}|dy \le C\int_x^M (M-y)^{\frac{\theta-1}{\gamma}}|g-g_\infty|dy \\
&\le C\left(\int_x^M (M-y)^{\frac{\theta-1+(\gamma-\theta)\epsilon_2}{\gamma}}(g-g_\infty)^2 dy\right)^{\frac{1}{2}}\left(\int_x^M (M-y)^{\frac{2\theta}{\gamma}-\frac{\theta+1+(\gamma-\theta)\epsilon_2}{\gamma}}dy\right)^{\frac{1}{2}} \\
&\le C\epsilon_0^{2^{-N_4-1}}(M-x)^{\frac{\theta}{\gamma}}. \qquad (3.51)
\end{aligned}$$

From the fact $\theta \in (0, \frac{\gamma}{2}]$ and the estimate (3.18)–(3.19), we have

$$\begin{aligned}
&\left|-\frac{\theta}{2c_1+c_2}\int_x^M [(r^{\beta-n+1}u)(y,t) - (r_0^{\beta-n+1}u_0)(y)]dy\right| \\
&\le C(M-x)^{\frac{1}{2}}(\|u\|_{L_x^2} + \|u_0\|_{L_x^2}) \\
&\le C\epsilon_0^{2^{-N_4-1}}(M-x)^{\frac{\theta}{\gamma}},\ x\in[0,M]. \qquad (3.52)
\end{aligned}$$

Thus, from (1.23), (3.17)–(3.20), and (3.50)–(3.52), we obtain

$$|I_1(x,t) - r_\infty^\beta \rho^\theta| \le C_{1,1}(M-x)^{\frac{\theta}{\gamma}}\epsilon_0^{2^{-N_4-1}}, \qquad (3.53)$$

and

$$|I_2(x,t_1) - I_2(x,t_2)| \le C_{1,2}(M-x)\epsilon_0^{2^{-N_4-1}}|t_2-t_1|,\ x\in[0,M]. \qquad (3.54)$$

Claim 2. *For any fixed $x \in [0,M]$, we have*

$$\begin{aligned}
I_1(x,t) &\ge \min\left\{I_1(x,0), r_\infty^\beta\left(\rho_\infty^\gamma - \frac{C_{1,2}}{C_{1,3}}\epsilon_0^{2^{-N_4-1}}(M-x)\right)^{\frac{\theta}{\gamma}} - C_{1,1}\epsilon_0^{2^{-N_4-1}}(M-x)^{\frac{\theta}{\gamma}}\right\} \\
&:= M_{1,1},
\end{aligned}$$

where $C_{1,3} := \frac{A\theta a^\beta}{2c_1+c_2} \le \frac{A\theta r^\beta}{2c_1+c_2}$.

Proof of Claim 2. If not, there exists $t_{1,1} > 0$ such that $I_1(x,t_{1,1}) < M_{1,1}$, then we can find $t_{1,2} \in (0,t_{1,1})$ such that $I_1(x,t_{1,2}) = M_{1,1}$ and $I_1(x,t) < M_{1,1}$ for all $t \in (t_{1,2}, t_{1,1})$. From (3.54), we have

$$I_1(x,t_{1,1}) - I_1(x,t_{1,2}) + \frac{A\theta}{2c_1+c_2}\int_{t_{1,2}}^{t_{1,1}} r^\beta(\rho^\gamma - \rho_\infty^\gamma) \ge -C_{1,2}\epsilon_0^{2^{-N_4-1}}(M-x)(t_{1,1}-t_{1,2}).$$

From (3.53), we have

$$\rho^\theta(x,t) = r_\infty^{-\beta}(I_1(x,t) - (I_1(x,t) - r_\infty^\beta \rho^\theta))$$
$$\le r_\infty^{-\beta}(M_{1,1} + C_{1,1}\epsilon_0^{2^{-N_4-1}}(M-x)^{\frac{\theta}{\gamma}}) \le \left(\rho_\infty^\gamma - \frac{C_{1,2}}{C_{1,3}}\epsilon_0^{2^{-N_4-1}}(M-x)\right)^{\frac{\theta}{\gamma}}$$

and

$$\rho^\gamma \le \rho_\infty^\gamma - \frac{C_{1,2}}{C_{1,3}}\epsilon_0^{2^{-N_4-1}}(M-x),$$

then $I_1(x,t_{1,1}) \ge I_1(x,t_{1,2})$. It is a contradiction. Thus, Claim 2 holds.

Similarly, we can obtain the following claim.

Claim 3. *For any fixed $x \in [0,M]$, we have*

$$I_1(x,t) \le \max\left\{I_1(x,0), r_\infty^\beta\left(\rho_\infty^\gamma + \frac{C_{1,2}}{C_{1,4}}\epsilon_0^{2^{-N_4-1}}(M-x)\right)^{\frac{\theta}{\gamma}} + C_{1,1}\epsilon_0^{2^{-N_4-1}}(M-x)^{\frac{\theta}{\gamma}}\right\}$$
$$:= M_{1,2},$$

where $C_{1,4}$ is a positive constant satisfying $C_{1,4} \ge \frac{A\theta r^\beta}{2c_1+c_2}$.

From Claims 2 and 3, we have

$$|g(x,t) - g_\infty(x)| \le C_{1,5}\epsilon_0^{\frac{\theta}{\gamma}2^{-N_4-1}}, \ x \in [0,M], \ t \in [0,T],$$

when $\epsilon_0 \le \delta_{11}$, with a small positive constant $\delta_{11}(G, A, M_0, M, a, n, \gamma, \theta, c_1, c_2)$. □

Using the results in Lemmas 3.3–3.8, we have that there exists ϵ_0 satisfying

$$4\epsilon_1 < \min_{x\in[0,M]} g_\infty, \ C_3\epsilon_1 \le \delta_9, \ \epsilon_1 \le \delta_{10} \ \text{ and } \ \epsilon_0 \le \delta_{11} \tag{3.55}$$

such that Claim 1 holds with

$$\epsilon_1 = \epsilon_0 + C_{10}\epsilon_0^{\frac{\theta}{\gamma}2^{-N_4-1}}.$$

From (3.6) and Claim 1, using the classical continuation method, we can easily obtain the following lemma.

Lemma 3.9. *Under the assumptions of Theorem* 1.1*, we obtain that* (3.17)–(3.18)*,* (3.33)–(3.35)*,* (3.48)*, and*

$$|r(x,t) - r_\infty(x)| \le C_{11}\epsilon_0^{\frac{\theta}{\gamma}2^{-N_4-1}} \tag{3.56}$$

hold for all $x \in [0,M]$ and $t \in [0,T^*)$.

Proof. Let $\mathcal{A} = \{T \in [0,T^*) \mid I(t) \le \epsilon_1 \ \text{ for all } \ t \in [0,T]\}$. Since $I(0) \le \epsilon_0 < \epsilon_1$ and $I(t) \in C([0,T^*))$, then there exists a constant $T_0 > 0$ such that $I(t) \le \epsilon_1$ for all $t \in [0,T_0]$. Thus, $\mathcal{A}$ is not empty and relatively closed in $[0,T^*)$. To show that $\mathcal{A}$ is also relatively open in $[0,T^*)$, and hence the entire interval, it therefore suffices to show that the weaker bound

$$I(t) \le 2\epsilon_1, \ \text{ for all } \ t \in [0,T'] \subset [0,T^*),$$

implies $I(t) \le \epsilon_1$ for all $t \in [0,T']$. From Claim 1, we have that $\mathcal{A} = [0,T^*)$.

Then, from Lemmas 3.3–3.8, we obtain that (3.17)–(3.18), (3.33)–(3.35), (3.48), and (3.56) hold for all $x \in [0,M]$ and $t \in [0,T^*)$. □

We will prove an estimate in weighted $L^2([0,M]\times[0,T^*))$-norm of the function $g - g_\infty$.

LEMMA 3.10. *Under the assumptions of Theorem* 1.1*, we obtain*

$$\int_0^t \int_0^M (1+s)^{-\epsilon_2}(g-g_\infty)^2 dxds \le C, \tag{3.57}$$

where $t \in [0, T^*)$.

Proof. From (1.15), we have

$$\begin{aligned} A(\rho^\gamma - \rho_\infty^\gamma) = &\int_x^M \left(\frac{u_t}{r^{n-1}} + \frac{\Delta f}{r^{n-1}}\right) dy + \int_x^M G(M_0+y)(r^{2-2n} - r_\infty^{2-2n}) dy \\ &+ (2c_1+c_2)\rho^{\theta+1}(r^{n-1}u)_x - 2c_1(n-1)\rho^\theta \frac{u}{r} - 2c_1(n-1)\int_x^M \rho^\theta \left(\frac{u}{r}\right)_x dy. \end{aligned}$$

Multiplying the above equality by $(1+t)^{-\epsilon_2}(M-x)^{-2}(\rho^\gamma - \rho_\infty^\gamma)$, and integrating the resulting equation over $[0,M]\times[0,t]$, we obtain

$$\begin{aligned} &\int_0^t \int_0^M A(1+s)^{-\epsilon_2}(M-x)^{-2}(\rho^\gamma - \rho_\infty^\gamma)^2 dxds \\ &= \int_0^t \int_0^M (1+s)^{-\epsilon_2}(M-x)^{-2}(\rho^\gamma - \rho_\infty^\gamma) \int_x^M \frac{u_t}{r^{n-1}} dydxds \\ &\quad + \int_0^t \int_0^M (1+s)^{-\epsilon_2}(M-x)^{-2}(\rho^\gamma - \rho_\infty^\gamma) \int_x^M \frac{\Delta f}{r^{n-1}} dydxds \\ &\quad + \int_0^t \int_0^M (1+s)^{-\epsilon_2}(M-x)^{-2}(\rho^\gamma - \rho_\infty^\gamma) \int_x^M G(M_0+y)(r^{2-2n} - r_\infty^{2-2n}) dydxds \\ &\quad + \int_0^t \int_0^M (2c_1+c_2)(1+s)^{-\epsilon_2}(M-x)^{-2}(\rho^\gamma - \rho_\infty^\gamma)\rho^{\theta+1}(r^{n-1}u)_x dxds \\ &\quad - \int_0^t \int_0^M 2c_1(n-1)(1+s)^{-\epsilon_2}(M-x)^{-2}(\rho^\gamma - \rho_\infty^\gamma)\rho^\theta \frac{u}{r} dxds \\ &\quad - \int_0^t \int_0^M 2c_1(n-1)(1+s)^{-\epsilon_2}(M-x)^{-2}(\rho^\gamma - \rho_\infty^\gamma) \int_x^M \rho^\theta \left(\frac{u}{r}\right)_x dydxds \\ &:= \sum_{i=1}^6 F_i. \end{aligned} \tag{3.58}$$

Using (1.15), (1.23), Lemma 3.9, integration by parts, and the Cauchy–Schwarz inequality, we can estimate F_i as follows.

$$\begin{aligned} F_1 = &\left\{ \int_0^M (1+s)^{-\epsilon_2}(M-x)^{-2}(\rho^\gamma - \rho_\infty^\gamma) \int_x^M \frac{u}{r^{n-1}} dydx \right\}\Bigg|_0^t \\ &+ \int_0^t \int_0^M \epsilon_2(1+s)^{-\epsilon_2-1}(M-x)^{-2}(\rho^\gamma - \rho_\infty^\gamma) \int_x^M \frac{u}{r^{n-1}} dydxds \\ &+ \int_0^t \int_0^M \gamma(1+s)^{-\epsilon_2}(M-x)^{-2}\rho^{\gamma+1}(r^{n-1}u)_x \int_x^M \frac{u}{r^{n-1}} dydxds \\ &+ \int_0^t \int_0^M (n-1)(1+s)^{-\epsilon_2}(M-x)^{-2}(\rho^\gamma - \rho_\infty^\gamma) \int_x^M \frac{u^2}{r^n} dydxds \end{aligned} \tag{3.59}$$

$$
\begin{aligned}
&\le C\|g-g_\infty\|_{L^\infty_{xt}}\left(\int_0^M u^2dx\right)^{\frac12}\left(\int_0^M (M-x)^{-\frac12}dx\right)^{\frac12}+C\\
&\quad+C\left(\int_0^t\int_x^M \frac{\rho_\infty^{\gamma-1}}{(1+s)^{\epsilon_2}}(g-g_\infty)^2dxds\right)^{\frac12}\\
&\qquad\left(\int_0^t \|u(\cdot,s)\|^2_{L^\infty}ds\int_0^M (M-x)^{-\frac{\gamma-1}{\gamma}}dx\right)^{\frac12}\\
&\quad+C\left(\int_0^t\int_x^M \rho^{\theta+1}(r^{n-1}u)_x^2dxds\right)^{\frac12}\left(\int_0^t \|u(\cdot,s)\|^2_{L^\infty}ds\int_0^M (M-x)^{-\frac{\theta-1}{\gamma}}dx\right)^{\frac12}\\
&\quad+C\|g-g_\infty\|_{L^\infty_{xt}}\int_0^t\|u(\cdot,s)\|^2_{L^\infty}ds\\
&\le C,
\end{aligned}
$$

$$
F_2\le C\left(\int_0^t(1+t)^{-1-\epsilon_2}\right)^{\frac12}\left(\int_0^t(1+t)f_1^2dt\right)^{\frac12}\le C, \tag{3.60}
$$

$$
\begin{aligned}
&|r(x,t)-r_\infty(x)|\\
&\le C\int_0^x\rho_\infty^{-1}|g-g_\infty|dy\le C\left(\int_0^x\rho_\infty^{\gamma-1}(g-g_\infty)^2dy\right)^{\frac12}\left(\int_0^x\rho_\infty^{-\gamma-1}dy\right)^{\frac12}\\
&\le C(M-x)^{-\frac{1}{2\gamma}}\left(\int_0^M\rho_\infty^{\gamma-1}(g-g_\infty)^2dx\right)^{\frac12}, \qquad (3.61)
\end{aligned}
$$

$$
\begin{aligned}
F_3&\le C\int_0^t\int_0^M(1+s)^{-\epsilon_2}(M-x)^{-2}|\rho^\gamma-\rho_\infty^\gamma|\\
&\qquad\times\left(\int_0^M\rho_\infty^{\gamma-1}(g-g_\infty)^2dz\right)^{\frac12}\int_x^M(M-y)^{-\frac{1}{2\gamma}}dydx\\
&\le\frac{A}{4}\int_0^t\int_0^M(1+s)^{-\epsilon_2}(M-x)^{-2}(\rho^\gamma-\rho_\infty^\gamma)^2dxds\\
&\quad+C\int_0^t\int_0^M(1+s)^{-\epsilon_2}\rho_\infty^{\gamma-1}(g-g_\infty)^2dxds\int_0^M(M-z)^{-\frac1\gamma}dz\\
&\le\frac{A}{4}\int_0^t\int_0^M(1+s)^{-\epsilon_2}(M-x)^{-2}(\rho^\gamma-\rho_\infty^\gamma)^2dxds+C, \qquad (3.62)
\end{aligned}
$$

$$
\begin{aligned}
F_4&=-\frac{2c_1+c_2}{\theta}\int_0^t\int_0^M(1+s)^{-\epsilon_2}(M-x)^{-2}(\rho^\gamma-\rho_\infty^\gamma)(\rho^\theta)_tdxds\\
&=-\frac{2c_1+c_2}{\theta}\left\{(1+s)^{-\epsilon_2}\int_0^M(M-x)^{-2}\left(\frac{\theta}{\gamma+\theta}\rho^{\gamma+\theta}-\rho_\infty^\gamma\rho^\theta\right)dx\right\}\Big|_0^t\\
&\quad-\frac{\epsilon_2(2c_1+c_2)}{\theta}\int_0^t\int_0^M(1+s)^{-\epsilon_2-1}(M-x)^{-2}\left(\frac{\theta}{\gamma+\theta}\rho^{\gamma+\theta}-\rho_\infty^\gamma\rho^\theta\right)dx\\
&\le C\|g-g_\infty\|_{L^\infty_{t,x}}\int_0^M(M-x)^{\frac{\theta}{\gamma}-1}dx\left(1+\int_0^t(1+s)^{-1-\epsilon_2}ds\right)+C\\
&\le C, \qquad (3.63)
\end{aligned}
$$

$$
\begin{aligned}
F_5 &\le C\int_0^t\int_0^M (1+s)^{-\epsilon_2}\|u(\cdot,t)\|_{L^\infty}(M-x)^{\frac{\theta}{\gamma}-1}dxds\\
&\le C\left\{\int_0^t (1+s)^{-1-\epsilon_2}ds\right\}^{\frac12}\left\{\int_0^t (1+s)^{1-\epsilon_2}\|u(\cdot,t)\|_{L^\infty}^2 ds\right\}^{\frac12}\le C
\end{aligned}\tag{3.64}
$$

and

$$
\begin{aligned}
F_6 &\le C\int_0^t\int_0^M (1+s)^{-\epsilon_2}(M-x)^{-1}\int_x^M \left(|\rho^\theta u_x|+|\rho^{\theta-1}u|\right)dydxds\\
&\le C\int_0^t\int_0^M (1+s)^{-\epsilon_2}(M-x)^{-1}\left[\int_x^M(\rho^{\theta+1}u_x^2+\rho^{\theta-1}u^2)dy\right]^{\frac12}\left[\int_x^M\rho^{\theta-1}dy\right]^{\frac12}dxds\\
&\le C\int_0^t\int_0^M (1+s)^{-\epsilon_2}(M-x)^{\frac{\theta-1}{2\gamma}-\frac12}\left[\int_x^M\left(\rho^{\theta+1}u_x^2+\rho^{\theta-1}u^2\right)dy\right]^{\frac12}dxds\\
&\le C\left[\int_0^t\int_0^M (1+s)^{1-\epsilon_2}\left(\rho^{\theta+1}u_x^2+\rho^{\theta-1}u^2\right)dxds\right]^{\frac12}\left[\int_0^t(1+s)^{-1-\epsilon_2}dt\right]^{\frac12}\\
&\le C.
\end{aligned}\tag{3.65}
$$

From (3.58)–(3.65), we can immediately get (3.57). □

LEMMA 3.11. *Under the assumptions of Theorem* 1.1*, we obtain*

$$
\begin{aligned}
&(1+t)^{-\epsilon_2}\int_0^M (M-x)^{1-\frac{\theta}{\gamma}}(\rho^\theta-\rho_\infty^\theta)_x^2(x,t)dx\\
&+\int_0^t\int_0^M (1+s)^{-\epsilon_2}(M-x)^{2-\frac{2\theta}{\gamma}}(\rho^\theta-\rho_\infty^\theta)_x^2(x,s)dxds\le C
\end{aligned}\tag{3.66}
$$

for all $t\in[0,T^*)$.

Proof. From (3.2), we have

$$
\begin{aligned}
&\partial_t\left[\frac{\theta}{2c_1+c_2}r^{1+\beta-n}u+r^\beta(\rho^\theta)_x-r_\infty^\beta(\rho_\infty^\theta)_x\right]\\
&+\frac{A\gamma\rho^{\gamma-\theta}}{2c_1+c_2}\left[\frac{\theta}{2c_1+c_2}r^{1+\beta-n}u+r^\beta(\rho^\theta)_x-r_\infty^\beta(\rho_\infty^\theta)_x\right]\\
&=\frac{A\gamma\theta}{(2c_1+c_2)^2}\rho^{\gamma-\theta}r^{1+\beta-n}u+\frac{\theta(1+\beta-n)}{2c_1+c_2}r^{\beta-n}u^2-\frac{\theta}{2c_1+c_2}r^{1+\beta-n}\Delta f\\
&-\frac{\theta}{2c_1+c_2}\left(r^\beta\frac{G(M_0+x)}{r^{2n-2}}+r_\infty^\beta\frac{\rho^{\gamma-\theta}}{\rho_\infty^{\gamma-\theta}}(A\rho_\infty^\gamma)_x\right).
\end{aligned}\tag{3.67}
$$

Let $H=\frac{\theta}{2c_1+c_2}r^{1+\beta-n}u+r^\beta(\rho^\theta)_x-r_\infty^\beta(\rho_\infty^\theta)_x$. Multiplying (3.67) by $(1+t)^{-\epsilon_2}(M-x)^{1-\frac{\theta}{\gamma}}H$, integrating the resulting equation over $[0,M]$, and using the Cauchy–Schwarz

inequality, we obtain

$$\begin{aligned}
&\frac{d}{dt}\int_0^M (1+t)^{-\epsilon_2}(M-x)^{1-\frac{\theta}{\gamma}}H^2dx + C_{13}\int_0^M (1+t)^{-\epsilon_2}(M-x)^{2-\frac{2\theta}{\gamma}}H^2dx\\
&\le C\int_0^M (1+t)^{-\epsilon_2}(M-x)^{1-\frac{\theta}{\gamma}}\left((M-x)^{1-\frac{\theta}{\gamma}}|Hu| + |u^2H| + |\Delta fH|\right)dx\\
&\quad + \int_0^M (1+t)^{-\epsilon_2}(M-x)^{1-\frac{\theta}{\gamma}}\left|r^\beta G\frac{M_0+x}{r^{2n-2}} + \frac{\rho^{\gamma-\theta}r_\infty^\beta}{\rho_\infty^{\gamma-\theta}}(A\rho_\infty^\gamma)_x\right||H|dx\\
&\le \frac{C_{13}}{4}\int_0^M (1+t)^{-\epsilon_2}(M-x)^{2-\frac{2\theta}{\gamma}}H^2dx + C\|u(\cdot,t)\|_{L^\infty}^2 + C\|u(\cdot,t)\|_{L^\infty}^2\|u(\cdot,t)\|_{L^2}^2\\
&\quad + Cf_1^2 + C\int_0^M (1+t)^{-\epsilon_2}\left|r^\beta G\frac{M_0+x}{r^{2n-2}} + \frac{\rho^{\gamma-\theta}r_\infty^\beta}{\rho_\infty^{\gamma-\theta}}(A\rho_\infty^\gamma)_x\right|^2dx. \qquad (3.68)
\end{aligned}$$

Here, we use the estimates (3.17)–(3.18) and the condition $\theta \in (0, \gamma-1)$. From (3.9) and (3.17)–(3.18), we have

$$\begin{aligned}
&\int_0^M (1+t)^{-\epsilon_2}\left|r^\beta G\frac{M_0+x}{r^{2n-2}} + \frac{\rho^{\gamma-\theta}r_\infty^\beta}{\rho_\infty^{\gamma-\theta}}(A\rho_\infty^\gamma)_x\right|^2dx\\
&= \int_0^M (1+t)^{-\epsilon_2}\left|r^\beta G\frac{M_0+x}{r^{2n-2}} - G\frac{(\rho^{\gamma-\theta}r_\infty^\beta)(M_0+x)}{(\rho_\infty^{\gamma-\theta})(r_\infty^{2n-2})}\right|^2dx\\
&\le C\int_0^M (1+t)^{-\epsilon_2}\left[(r-r_\infty)^2 + (g-g_\infty)^2\right]dx. \qquad (3.69)
\end{aligned}$$

From (3.33) and (3.68)–(3.69), we obtain

$$\begin{aligned}
&\frac{d}{dt}\int_0^M (1+t)^{-\epsilon_2}(M-x)^{1-\frac{\theta}{\gamma}}H^2dx + \frac{C_{13}}{2}\int_0^M (1+t)^{-\epsilon_2}(M-x)^{2-\frac{2\theta}{\gamma}}H^2dx\\
&\le C\int_0^M (1+t)^{-\epsilon_2}\left((r-r_\infty)^2 + (g-g_\infty)^2\right)dx + C\|u(\cdot,t)\|_{L^\infty}^2 + Cf_1^2. \qquad (3.70)
\end{aligned}$$

From (A2), (1.23), (3.33)–(3.35), (3.56)–(3.57), and (3.70), we can immediately obtain (3.66). □

LEMMA 3.12. *Under the assumptions of Theorem* 1.1*, we obtain*

$$(1+t)^{1-\epsilon_2}\int_0^M \left(\rho^{\theta-1}u^2 + \rho^{\theta+1}u_x^2\right)(x,t)dx + \int_0^t\int_0^M (1+s)^{1-\epsilon_2}u_t^2(x,s)dxds \le C, \qquad (3.71)$$

$$\|u(\cdot,t)\|_{L^\infty} \le C(1+t)^{-\frac{1}{2}+\frac{\epsilon_2}{2}} \qquad (3.72)$$

for all $t \in [0, T^*)$.

Proof. Multiplying $(1.15)_2$ by $(1+t)^{1-\epsilon_2}u_t$, integrating the resulting equation over $[0,M]\times[0,t]$, and using integration by parts and the boundary conditions (1.17), we obtain

(3.73)
$$\begin{aligned}
&\int_0^t\int_0^M (1+s)^{1-\epsilon_2}u_t^2(x,s)dxds\\
&= \int_0^t\int_0^M A(1+s)^{1-\epsilon_2}\rho^\gamma\partial_x(r^{n-1}u_t)dxds\\
&\quad - \int_0^t\int_0^M (2c_1+c_2)(1+s)^{1-\epsilon_2}\rho^{1+\theta}\partial_x(r^{n-1}u)\partial_x(r^{n-1}u_t)dxds\\
&\quad + \int_0^t\int_0^M 2c_1(n-1)(1+s)^{1-\epsilon_2}\rho^\theta\partial_x(r^{n-2}uu_t)dxds - \int_0^t\int_0^M (1+s)^{1-\epsilon_2}fu_tdxds\\
&:= \sum_{i=1}^4 H_i.
\end{aligned}$$

Using (A3), (1.23), (3.17)–(3.18), (3.33)–(3.34), and the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}
&H_2+H_3\\
&= \left\{(1+s)^{1-\epsilon_2}\int_0^M\left[-\frac{2c_1+c_2}{2}\rho^{1+\theta}[\partial_x(r^{n-1}u)]^2\right.\right.\\
&\qquad\qquad \left.\left. + \ c_1(n-1)\rho^\theta\partial_x(r^{n-2}u^2)\right]dx\right\}\Bigg|_0^t\\
&\quad + \int_0^t\int_0^M \frac{(2c_1+c_2)(1-\epsilon_2)}{2}(1+s)^{-\epsilon_2}\rho^{1+\theta}(r^{n-1}u)_x^2dxds\\
&\quad - \int_0^t\int_0^M c_1(n-1)(1-\epsilon_2)(1+s)^{-\epsilon_2}\rho^\theta\partial_x(r^{n-2}u^2)dxds\\
&\quad + \int_0^t\int_0^M (1+s)^{1-\epsilon_2}\left\{(2c_1+c_2)(n-1)\rho^{1+\theta}\partial_x(r^{n-1}u)\partial_x(r^{n-2}u^2)\right.\\
&\quad - \frac{(2c_1+c_2)}{2}(1+\theta)\rho^{2+\theta}[\partial_x(r^{n-1}u)]^3 + 2\theta c_1(n-1)\rho^{\theta+1}\frac{u}{r}[\partial_x(r^{n-1}u)]^2\\
&\quad - \theta c_1n(n-1)\rho^\theta\frac{u^2}{r^2}\partial_x(r^{n-1}u) + 2nc_1(n-1)(n-2)\rho^{\theta-1}\frac{u^3}{r^3}\\
&\quad \left. - 3c_1(n-1)(n-2)\rho^\theta\ \frac{u^2}{r^2}\partial_x(r^{n-1}u)\right\}dxds\\
&\le C + C\int_0^t\left(\|u\|_{L_x^\infty} + \|\rho(r^{n-1}u)_x\|_{L_x^\infty}\right)(1+s)^{1-\epsilon_2}\\
&\qquad\qquad \int_0^M\left[\rho^{1+\theta}(r^{n-1}u)_x^2 + \rho^{\theta-1}u^2\right]dxds\\
&\quad - C_{14}(1+t)^{1-\epsilon_2}\int_0^M[\rho^{1+\theta}(r^{n-1}u)_x^2 + \rho^{\theta-1}u^2]dx, \qquad (3.74)
\end{aligned}$$

$$
\begin{aligned}
H_1 = & \left\{(1+s)^{1-\epsilon_2}\int_0^M A\rho^\gamma \partial_x(r^{n-1}u)dx\right\}\Big|_0^t \\
& + \int_0^t\int_0^M A\gamma(1+s)^{1-\epsilon_2}\rho^{\gamma+1}[\partial_x(r^{n-1}u)]^2 dxds \\
& - \int_0^t\int_0^M 2A(n-1)(1+s)^{1-\epsilon_2}\rho^\gamma \frac{u}{r}\partial_x(r^{n-1}u)dxds \\
& + \int_0^t\int_0^M An(n-1)(1+s)^{1-\epsilon_2}\rho^{\gamma-1}\frac{u^2}{r^2}dxds \\
& - \int_0^t\int_0^M A(1-\epsilon_2)(1+s)^{-\epsilon_2}\rho^\gamma \partial_x(r^{n-1}u)dxds \\
\le & (1+s)^{1-\epsilon_2}\int_0^M A\rho^\gamma\partial_x(r^{n-1}u)dx + C, \qquad (3.75)
\end{aligned}
$$

and

$$
\begin{aligned}
H_4 = & -\left\{(1+s)^{1-\epsilon_2}\int_0^M G\frac{u(M_0+x)}{r^{n-1}}dx\right\}\Big|_0^t - \int_0^t\int_0^M (1+s)^{1-\epsilon_2}\Delta f u_t dxds \\
& + (1-\epsilon_2)\int_0^t\int_0^M (1+s)^{-\epsilon_2}\frac{Gu(M_0+x)}{r^{n-1}}dxds \\
& + \int_0^t\int_0^M (1-n)(1+s)^{1-\epsilon_2}G(M_0+x)r^{-n}u^2 dxds \\
\le & -(1+s)^{1-\epsilon_2}\int_0^M G\frac{u(M_0+x)}{r^{n-1}}dx + C + \frac{1}{2}\int_0^t\int_0^M (1+s)^{1-\epsilon_2}u_t^2 dxds. \qquad (3.76)
\end{aligned}
$$

Using (3.17)–(3.18), (3.24), (3.33), integration by parts, and the Cauchy–Schwarz inequality, we obtain

(3.77)

$$
\begin{aligned}
& (1+s)^{1-\epsilon_2}\int_0^M A\rho^\gamma\partial_x(r^{n-1}u)dx - (1+s)^{1-\epsilon_2}\int_0^M G\frac{u(M_0+x)}{r^{n-1}}dx \\
= & (1+s)^{1-\epsilon_2}\int_0^M \left(A(\rho^\gamma-\rho_\infty^\gamma)\partial_x(r^{n-1}u) - Gr^{n-1}u(M_0+x)(r^{2-2n}-r_\infty^{2-2n})\right)dx \\
\le & \frac{C_{14}}{4}(1+t)^{1-\epsilon_2}\int_0^M \rho^{1+\theta}(r^{n-1}u)_x^2 dx + C.
\end{aligned}
$$

From (3.73)–(3.77), we can obtain

$$
\begin{aligned}
& (1+t)^{1-\epsilon_2}\int_0^M \left[\rho^{1+\theta}(r^{n-1}u)_x^2 + \rho^{\theta-1}u^2\right]dx + \int_0^t\int_0^M (1+s)^{1-\epsilon_2}u_t^2 dxds \\
\le & C + C\int_0^t \left(\|u\|_{L_x^\infty} + \|\rho(r^{n-1}u)_x\|_{L_x^\infty}\right)(1+s)^{1-\epsilon_2} \\
& \times \int_0^M \left[\rho^{1+\theta}(r^{n-1}u)_x^2 + \rho^{\theta-1}u^2\right]dxds. \qquad (3.78)
\end{aligned}
$$

From (3.3), we have

$$\rho(r^{n-1}u)_x = \frac{1}{(2c_1+c_2)\rho^\theta}\left\{A\rho^\gamma + 2c_1(n-1)\rho^\theta\frac{u}{r} + \int_x^M\left[-\frac{u_t}{r^{n-1}} + 2c_1(n-1)\rho^\theta\left(\frac{u}{r}\right)_x - \frac{f}{r^{n-1}}\right]dy\right\}.$$

Using conditions $\theta \in (0, \frac{\gamma}{2}] \cap (0, \gamma-1)$, (1.23), estimates (3.17)–(3.20), and Hölder's inequality, we conclude that

$$\|\rho\partial_x(r^{n-1}u)\|_{L_x^\infty} \le C + C\left(\|u(\cdot,t)\|_{L^\infty}^2 + \int_0^M\left[\rho^{1+\theta}(r^{n-1}u)_x^2 + u_t^2\right]dx\right)^{\frac{1}{2}}. \tag{3.79}$$

Using (3.24), (3.34), (3.78)–(3.79), and the Cauchy–Schwarz inequality, we can obtain

$$\begin{aligned}&(1+t)^{1-\epsilon_2}\int_0^M\left[\rho^{1+\theta}(r^{n-1}u)_x^2 + \rho^{\theta-1}u^2\right]dx + \int_0^t\int_0^M(1+s)^{1-\epsilon_2}u_t^2dxds\\ &\le C + C\int_0^t\left(\int_0^M(1+s)^{1-\epsilon_2}\left[\rho^{1+\theta}(r^{n-1}u)_x^2 + \rho^{\theta-1}u^2\right]dx\right)^2 ds.\end{aligned}$$

Using Gronwall's inequality and the estimate (3.34), we can immediately get (3.71).

From (3.17), (3.71), and the fact $\theta \in (0, \gamma-1)$, we can obtain

$$\begin{aligned}|u(x,t)| &\le \left|\int_0^x u_x dy\right| \le C\left(\int_0^x \rho^{\theta+1}u_x^2 dy\right)^{\frac{1}{2}}\left(\int_0^x (M-y)^{-\frac{\theta+1}{\gamma}}dy\right)^{\frac{1}{2}}\\ &\le C(1+t)^{-\frac{1}{2}+\frac{\epsilon_2}{2}},\ (x,t)\in[0,M]\times[0,T^*).\end{aligned}$$ □

LEMMA 3.13. *Under the assumptions of Theorem* 1.1*, we obtain*

$$\int_0^M u_t^2(x,t)dx + \int_0^t\int_0^M\left[\rho^{\theta+1}u_{xt}^2 + \rho^{\theta-1}u_t^2\right]dxds \le C_{11}, \tag{3.80}$$

$$\|\rho(r^{n-1}u)_x(\cdot,t)\|_{L^\infty} \le C_{11} \tag{3.81}$$

for all $t \in [0, T^*)$.

Proof. We differentiate the equation $(1.15)_2$ with respect to t, multiply it by u_t, and integrate it over $[0,M]\times[0,t]$ using the boundary conditions (1.17), then derive

(3.82)

$$\begin{aligned}&\int_0^M\frac{1}{2}u_t^2dx\\ &= \int_0^M\frac{1}{2}u_t^2(x,0)dx - \int_0^t\int_0^M\left[(2c_1+c_2)\rho^{1+\theta}\partial_x(r^{n-1}u) - A\rho^\gamma - 2c_1(n-1)\rho^\theta\frac{u}{r}\right]\\ &\times\partial_x((n-1)r^{n-2}uu_t)dxds - \int_0^t\int_0^M\partial_t\left[(2c_1+c_2)\rho^{1+\theta}\partial_x(r^{n-1}u) - A\rho^\gamma\right.\\ &\left.- 2c_1(n-1)\rho^\theta\frac{u}{r}\right]\partial_x(r^{n-1}u_t)dxds + \int_0^t\int_0^M 2c_1(n-1)\partial_t\left(r^{n-1}\rho^\theta\partial_x\left(\frac{u}{r}\right)\right)u_t dxds\end{aligned}$$

$$
\begin{aligned}
&- \int_0^t \int_0^M f_t u_t dxds \\
&:= \sum_{i=1}^{5} J_i.
\end{aligned}
$$

From (A2)–(A3), we have

$$
\begin{aligned}
J_1 &\le C \left(\left\| \left((2c_1 + c_2)\rho_0^{\theta+1} (r_0^{n-1} u_0)_x \right)_x - 2c_1(n-1) \frac{u_0}{r_0} (\rho_0^\theta)_x \right\|_{L^2} \right. \\
&\quad \left. + \|(\rho_0^\gamma)_x\|_{L^2} + \|f(x, r_0, 0)\|_{L^2} \right)^2 \\
&\le C. \qquad (3.83)
\end{aligned}
$$

From (3.17)–(3.20) and the Cauchy–Schwarz inequality, we get

$$
\begin{aligned}
&J_3 + J_4 \\
&= - \int_0^t \int_0^M \left[(2c_1 + c_2)\rho^{1+\theta} (r^{n-1} u_t)_x^2 - 2c_1(n-1)\rho^\theta (r^{n-2} u_t^2)_x \right] dxds \\
&\quad + \int_0^t \int_0^M \left\{ (2c_1 + c_2)(1+\theta)\rho^{\theta+2} [\partial_x (r^{n-1} u)]^2 - (n-1)(2c_1 + c_2)\rho^{1+\theta} \partial_x (r^{n-2} u^2) \right. \\
&\quad \left. - \gamma \rho^{\gamma+1} \partial_x (r^{n-1} u) - 2c_1(n-1)\theta \rho^{\theta+1} \partial_x (r^{n-1} u) \frac{u}{r} - 2c_1(n-1)\rho^\theta \frac{u^2}{r^2} \right\} \\
&\quad \times \left[(n-1) \frac{u_t}{r\rho} + r^{n-1} u_{tx} \right] dxds + 2c_1(n-1) \int_0^t \int_0^M \left\{ (n-1) r^{n-2} u \rho^\theta \left(\frac{u}{r} \right)_x u_t \right. \\
&\quad \left. - \theta r^{n-1} \rho^{\theta+1} (r^{n-1} u)_x \left(\frac{u}{r} \right)_x u_t - r^{n-1} \rho^\theta \left(\frac{u^2}{r^2} \right)_x u_t \right\} dxds \\
&\le - C_{15} \int_0^t \int_0^M (\rho^{\theta+1} u_{xt}^2 + \rho^{\theta-1} u_t^2) dxds + C \\
&\quad + C \int_0^t \left(\|u\|_{L_x^\infty}^2 + \|\rho (r^{n-1} u)_x\|_{L_x^\infty}^2 \right) \int_0^M \left[\rho^{\theta+1} u_x^2 + \rho^{\theta-1} u^2 \right] dxds. \qquad (3.84)
\end{aligned}
$$

From (1.11), (3.17)–(3.20), and the Cauchy–Schwarz inequality, we obtain

$$
\begin{aligned}
J_2 &\le \frac{C_{15}}{8} \int_0^t \int_0^M \left(\rho^{\theta-1} u_t^2 + \rho^{\theta+1} u_{xt}^2 \right) dxds + C \\
&\quad + C \int_0^t \left(\|u\|_{L_x^\infty}^2 + \|\rho (r^{n-1} u)_x\|_{L_x^\infty}^2 \right) \int_0^M \left[\rho^{\theta+1} u_x^2 + \rho^{\theta-1} u^2 \right] dxds \qquad (3.85)
\end{aligned}
$$

and

$$
\begin{aligned}
J_5 &\le \frac{C_{15}}{8} \int_0^t \int_0^M \rho^{\theta-1} u_t^2 dxds \\
&\quad + C \int_0^t \int_0^M (G(M_0 + x) r^{-n} |u| + |\partial_r \Delta f u| + |\partial_t \Delta f|)^2 \rho^{1-\theta} dxds \\
&\le \frac{C_{15}}{8} \int_0^M r^{\alpha-2} u_t^2 dxds + C. \qquad (3.86)
\end{aligned}
$$

From (3.82)–(3.86), we have

$$\int_0^M u_t^2(x,t)dx + \int_0^t \int_0^M \left[\rho^{\theta+1}u_{xt}^2 + \rho^{\theta-1}u_t^2\right] dxds$$
$$\leq C + C\int_0^t \left(\|u\|_{L_x^\infty}^2 + \|\rho(r^{n-1}u)_x\|_{L_x^\infty}^2\right) \int_0^M \left[\rho^{\theta+1}u_x^2 + \rho^{\theta-1}u^2\right] dxds. \tag{3.87}$$

From (3.71)–(3.72) and (3.79), we have

$$\|\rho\partial_x(r^{n-1}u)\|_{L_x^\infty} \leq C + C\|u_t\|_{L_x^2}. \tag{3.88}$$

From (3.20), (3.71)–(3.72), and (3.87)–(3.88), we obtain

$$\int_0^M u_t^2(x,t)dx + \int_0^t \int_0^M \left[\rho^{\theta+1}u_{xt}^2 + \rho^{\theta-1}u_t^2\right] dxds$$
$$\leq C + C\int_0^t \int_0^M \left[\rho^{\theta+1}u_x^2 + \rho^{\theta-1}u^2\right] dx\|u_t\|_{L_x^2}^2 ds.$$

Using Gronwall's inequality and the estimate (3.20), we can immediately obtain (3.80)–(3.81). □

Proof of existence and uniqueness. If $T^* < \infty$, from Lemmas 3.9–3.13, we have, for all $t \in [0, T^*)$,

$$C^{-1}(M-x)^{1/\gamma} \leq \rho(x,t) \leq C(M-x)^{1/\gamma},\ x \in [0,M],$$

$$\int_0^M (M-x)^{1-\frac{\theta}{\gamma}}(\rho^\theta)_x^2 dx \leq C,\ \int_0^M (\rho^\gamma)_x^2 dx \leq C,$$

$$\int_0^M u^2 + (M-x)^{\frac{\theta+1}{\gamma}}u_x^2 dx \leq C,\ u(0,t) = 0$$

and

$$\int_0^M \left\{\left((2c_1+c_2)\rho^{\theta+1}(r^{n-1}u)_x\right)_x - 2c_1(n-1)\frac{u}{r}\partial_x\rho^\theta\right\}^2 dx \leq C.$$

Thus, from Theorem 3.1, there exists $T_2 > 0$ such that the free boundary problem (1.15)–(1.17) admits a unique weak solution $(\rho_2, u_2, r_2)(x,t)$ on $[0,M]\times[T^* - \frac{T_2}{2}, T^* + \frac{T_2}{2}]$, with initial data $(\rho,u,r)(x, T^* - \frac{T_2}{2})$. Using the uniqueness result in Theorem 3.1, we obtain that

$$(\tilde\rho, \tilde u, \tilde r)(x,t) = \begin{cases} (\rho,u,r)(x,t), & t \in [0, T^* - \frac{T_2}{2}], \\ (\rho_2,u_2,r_2)(x,t), & t \in [T^* - \frac{T_2}{2}, T^* + \frac{T_2}{2}], \end{cases}$$

is a solution of the system (1.15)–(1.17), which is in contradiction with the definition of T^*. Thus, we have that $T^* = \infty$. From Lemmas 3.9–3.13, we can show that the global weak solution satisfies the regularity conditions (1.25)–(1.26) and (1.28) in Theorem 1.1.

Remark 3.2. The uniqueness of the solution of Theorem 3.1 is obtained by the energy method. Let (u_i, ρ_i, r_i), $i = 1, 2$, be two solutions of the system (1.15)–(1.17)

satisfying the regularity conditions in Theorem 1.1. Using similar arguments as that in the uniqueness part in [3], we can obtain, for all $T > 0$,

$$\begin{aligned}
&\frac{d}{dt}\int_0^M \left(w^2 + \rho_1^{1-\theta}\rho_2^{2\theta-4}\varrho^2 + \rho_1^{\theta}\rho_2^{-1}\mathcal{R}^2\right) dx \\
&+ C^{-1}\int_0^M \rho_1^{1+\theta}\left(\rho_1 r_1^{2n-2}\left(\partial_x w\right)^2 + \frac{w^2}{r_1^2\rho_1}\right) dx \\
&\le C\int_0^M \left(w^2 + \rho_1^{1-\theta}\rho_2^{2\theta-4}\varrho^2 + \rho_1^{\theta}\rho_2^{-1}\mathcal{R}^2\right) dx,\ t \in [0,T],
\end{aligned}$$

where $(w, \varrho, \mathcal{R}) = (u_1 - u_2, \rho_1 - \rho_2, r_1 - r_2)$. Using Gronwall's inequality, we could obtain that $(u_1, \rho_1, r_1) = (u_2, \rho_2, r_2)$, a.e. $(x,t) \in [0,M] \times [0,T]$.

4. Further decay result.

LEMMA 4.1. *Let ν be a positive constant satisfying $\nu < \min\{1, \frac{2\gamma-2}{\gamma+\theta}\}$. Under the assumptions of Theorem* 1.1*, we obtain*

$$\left\|\rho^{\frac{\gamma+\theta}{2}}(\cdot,t) - \rho_\infty^{\frac{\gamma+\theta}{2}}(\cdot)\right\|_{L^\infty} \le C(1+t)^{-\frac{1}{4}+\frac{\epsilon_2}{2}}, \tag{4.1}$$

and

$$\|r(\cdot,t) - r_\infty(\cdot)\|_{L^\infty} \le C_\nu(1+t)^{-\frac{1}{4}\nu+\frac{\epsilon_2\nu}{2}} \tag{4.2}$$

for all $t \ge 0$, where C_ν is a positive constant depending on ν.

Proof. From (3.48), (3.56), and (3.66), we have

$$\int_0^M \left(\rho^{\frac{\gamma+\theta}{2}} - \rho_\infty^{\frac{\gamma+\theta}{2}}\right)_x^2 dx \le C(1+t)^{\epsilon_2},\ t \ge 0. \tag{4.3}$$

Combining (3.33) and the Galiardo–Nirenberg inequality $\|\phi\|_{L^\infty} \le \|\phi\|_{L^2}^{\frac{1}{2}}\|\phi'\|_{L^2}^{\frac{1}{2}}$, we obtain

$$\left\|\rho^{\frac{\gamma+\theta}{2}} - \rho_\infty^{\frac{\gamma+\theta}{2}}\right\|_{L^\infty} \le C(1+t)^{-\frac{1}{4}+\frac{\epsilon_2}{2}},\ t \ge 0.$$

From (3.17)–(3.18), (3.48), and (4.1), we have

$$\left\|\rho^{\frac{\nu(\gamma+\theta)}{2}}(\cdot,t) - \rho_\infty^{\frac{\nu(\gamma+\theta)}{2}}(\cdot)\right\|_{L^\infty} \le C\|\rho^{\frac{\gamma+\theta}{2}}(\cdot,t) - \rho_\infty^{\frac{\gamma+\theta}{2}}(\cdot)\|_{L^\infty}^{\nu} \le C(1+t)^{-\frac{\nu}{4}+\frac{\epsilon_2\nu}{2}}$$

and

$$\begin{aligned}
\|r(\cdot,t) - r_\infty(\cdot)\|_{L^\infty} &\le C\int_0^M |\rho^{-1} - \rho_\infty^{-1}|dx \\
&\le C\|\rho^{\frac{\nu(\gamma+\theta)}{2}}(\cdot,t) - \rho_\infty^{\frac{\nu(\gamma+\theta)}{2}}(\cdot)\|_{L^\infty}\int_0^M (M-x)^{-\frac{1}{\gamma}-\frac{\nu(\gamma+\theta)}{2\gamma}}dx \\
&\le C_\nu(1+t)^{-\frac{\nu}{4}+\frac{\epsilon_2\nu}{2}}
\end{aligned}$$

for all $t \ge 0$. ∎

Using similar arguments as that in Lemmas 3.6, 3.11–3.12, and 4.1 with $\nu = \frac{\gamma-1}{2\gamma}$, we can obtain the following lemma and omit the proof.

LEMMA 4.2. *Under the assumptions of Theorem* 1.1*, we have*

$$\int_0^M (M-x)^{\frac{\theta-1}{\gamma}}(g-g_\infty)^2dx+\int_0^t\int_0^M \left[\rho_\infty^{\gamma-1}(g-g_\infty)^2+(r-r_\infty)^2\right]dxds\le C, \tag{4.4}$$

$$\int_0^M \left\{u^2(x,t)+(M-x)^{\frac{\gamma-1}{\gamma}}(g-g_\infty)^2+(r-r_\infty)^2\right\}dx\le C(1+t)^{-1}, \tag{4.5}$$

$$\int_0^t(1+s)\|u(\cdot,s)\|_{L^\infty}^2ds+\int_0^t\int_0^M(1+s)\left(\rho^{\theta+1}u_x^2+\rho^{\theta-1}u^2\right)(x,s)dxds\le C, \tag{4.6}$$

$$\int_0^M (M-x)^{1-\frac{\theta}{\gamma}}(\rho^\theta-\rho_\infty^\theta)_x^2dx+\int_0^t\int_0^M(M-x)^{2-\frac{2\theta}{\gamma}}(\rho^\theta-\rho_\infty^\theta)_x^2dxds\le C, \tag{4.7}$$

and

$$(1+t)\int_0^M\left(\rho^{\theta-1}u^2+\rho^{\theta+1}u_x^2\right)(x,t)dx+\int_0^t\int_0^M(1+s)u_t^2(x,s)dxds\le C \tag{4.8}$$

for all $t\ge 0$.

Remark 4.1. The key point is as follows: Similar to (3.32), using the estimates (3.18), (3.34), (4.2), and the condition $\epsilon_2<\frac{\gamma-1}{2(3\gamma-1)}$, we have

$$\begin{aligned}
&\int_0^t\int_0^M 2c_1(n-1)\rho^\theta\left(\frac{u}{r}\left(\frac{r^n}{n}-\frac{r_\infty^n}{n}\right)\right)_x dxds\\
&\le C\int_0^t\int_0^M\left\{|r-r_\infty|\left(|\rho^\theta u_x|+|\rho^{\theta-1}u|\right)+\rho^\theta|u(\rho^{-1}-\rho_\infty^{-1})|\right\}dxds\\
&\le C\int_0^t\int_0^M(1+s)^{1-\epsilon_2}\left(\rho^{\theta+1}u_x^2+\rho^{\theta-1}u^2\right)(x,s)dxds\\
&\quad+C\int_0^t(1+s)^{\epsilon_2-1}\|r(\cdot,t)-r_\infty(\cdot)\|_{L^\infty}^2\int_0^M(M-x)^{\frac{\theta-1}{\gamma}}dxds\\
&\quad+C\int_0^t(1+s)^{\epsilon_2-1}\|\rho^{\frac{\nu(\gamma+\theta)}{2}}(\cdot,t)-\rho_\infty^{\frac{\nu(\gamma+\theta)}{2}}(\cdot)\|_{L^\infty}^2\int_0^M(M-x)^{\frac{\theta+1}{\gamma}-\frac{2}{\gamma}-\frac{\nu(\gamma+\theta)}{\gamma}}dxds\\
&\le C+C\int_0^t(1+s)^{\epsilon_2-1-\frac{\nu}{2}+\epsilon_2\nu}ds\le C,\quad t\ge 0.
\end{aligned}$$

Without loss of generality, we assume that $\eta\in(0,\frac{2(\gamma+\theta)}{\gamma+\theta+1})$. Let $\epsilon_4\in(0,\frac{\gamma+\theta-1}{3(\gamma+\theta)})$ be a constant satisfying $\frac{1-\epsilon_4}{2-\frac{3\gamma+3\theta-1}{2(\gamma+\theta)}+\frac{\epsilon_4}{2}}>\frac{2(\gamma+\theta)}{\gamma+\theta+1}-\eta$. Define $\{\kappa_j\}$ and $\{\eta_j\}$ by $\eta_{j+1}=1+\kappa_j$, $\kappa_j=\frac{3\gamma+3\theta-1}{4(\gamma+\theta)}\eta_j-\frac{\epsilon_4}{4}\eta_j-\frac{1}{2}-\frac{\epsilon_4}{2}$, and $\eta_0=1$. Let N_5 be a positive integer satisfying $\eta_{N_5}>\frac{2(\gamma+\theta)}{\gamma+\theta+1}-\eta$. It is easy to see that $\eta<2$ and $\kappa_j<1$, $j=0,1\ldots,N_5$. Using similar arguments as that in Lemma 4.2, applying the induction method, we can obtain the following lemma and omit the proof.

LEMMA 4.3. *Under the assumptions of Theorem* 1.1*, we have*

$$\int_0^M\left\{u^2(x,t)+(M-x)^{\frac{\gamma-1}{\gamma}}(g-g_\infty)^2+(r-r_\infty)^2\right\}dx\le C_{\eta,j}(1+t)^{-\eta_j}, \tag{4.9}$$

$$\int_0^t(1+s)^{\eta_j}\|u(\cdot,s)\|_{L^\infty}^2ds+\int_0^t\int_0^M(1+s)^{\eta_j}\left(\rho^{\theta+1}u_x^2+\rho^{\theta-1}u^2\right)(x,s)dxds\le C_{\eta,j}, \tag{4.10}$$

$$\left\|\rho^{\frac{\gamma+\theta}{2}}(\cdot,t)-\rho_\infty^{\frac{\gamma+\theta}{2}}(\cdot)\right\|_{L^\infty}\le C_{\eta,j}(1+t)^{-\frac{\eta_j}{4}}, \tag{4.11}$$

$$\int_0^t\int_0^M(1+s)^{\kappa_j}\left[\rho_\infty^{\gamma-1}(g-g_\infty)^2+(r-r_\infty)^2\right]dxds\le C_{\eta,j}, \tag{4.12}$$

$$(1+t)^{\eta_j}\int_0^M\left(\rho^{\theta-1}u^2+\rho^{\theta+1}u_x^2\right)(x,t)dx+\int_0^t\int_0^M(1+s)^{\eta_j}u_t^2(x,s)dxds\le C_{\eta,j}, \tag{4.13}$$

and

$$\|u(\cdot,t)\|_{L^\infty}\le C_{\eta,j}(1+t)^{-\frac{\eta_j}{2}} \tag{4.14}$$

for all $t\ge 0$, $j=0,\dots,N_5$, where $C_{\eta,j}$ is a positive constant depending on η and j.

Remark 4.2. The main difficulty is to show (4.12) with $j=k$, when (4.9)–(4.11) hold with $j=k$. From (3.27)–(3.28), we have

$$\begin{aligned}
&\int_0^T\int_0^M(1+t)^{\kappa_k}\left[A(\rho_\infty^\gamma-\rho^\gamma)(\rho^{-1}-\rho_\infty^{-1})\right.\\
&\left.+G(M_0+x)(r^{2-2n}-r_\infty^{2-2n})\left(\frac{r^n}{n}-\frac{r_\infty^n}{n}\right)\right]dxdt\\
=&-\int_0^T\int_0^M(1+t)^{\kappa_k}\frac{u_t}{r^{n-1}}\left(\frac{r^n}{n}-\frac{r_\infty^n}{n}\right)dxdt\\
&-\int_0^T\int_0^M(1+t)^{\kappa_k}\Delta f r^{1-n}\left(\frac{r^n}{n}-\frac{r_\infty^n}{n}\right)dxdt\\
&+\int_0^T\int_0^M(1+t)^{\kappa_k}(2c_1+c_2)\rho^{1+\theta}\partial_x(r^{n-1}u)(\rho_\infty^{-1}-\rho^{-1})dxdt\\
&+\int_0^T\int_0^M(1+t)^{\kappa_k}2c_1(n-1)\rho^\theta\left(\frac{u}{r}\left(\frac{r^n}{n}-\frac{r_\infty^n}{n}\right)\right)_x dxdt\\
:=&\sum_{i=1}^4 Q_i,\ T>0
\end{aligned} \tag{4.15}$$

and

$$\text{L.H.S. of (4.15)}\ge C_{12}\int_0^T\int_0^M(1+t)^{\kappa_k}\left[\rho_\infty^{\gamma-1}(g-g_\infty)^2+(r-r_\infty)^2\right]dxdt. \tag{4.16}$$

Similar to (3.29)–(3.32), applying the estimates (3.17)–(3.20), (4.4), integration by parts, the Cauchy–Schwarz inequality, and the fact that $\kappa_j=\frac{3\gamma+3\theta-1}{4(\gamma+\theta)}\eta_j-\frac{\epsilon_4}{4}\eta_j-\frac12-\frac{\epsilon_4}{2}<\eta_j$, we can estimate Q_i as follows:

$$\begin{aligned}
Q_1\le&-\int_0^M(1+t)^{\kappa_k}\frac{u}{r^{n-1}}\left(\frac{r^n}{n}-\frac{r_\infty^n}{n}\right)dx\Big|_0^T+C\int_0^T(1+t)^{\kappa_k}\|u\|_{L_x^\infty}^2dt\\
&+C\int_0^T\int_0^M(1+t)^{\kappa_k-1}|u||r-r_\infty|dxdt\\
\le&\ C,
\end{aligned} \tag{4.17}$$

$$Q_2\le\frac{C_{12}}{6}\int_0^T\int_0^M(1+t)^{\kappa_k}(r-r_\infty)^2dxdt+C\int_0^T f_1^2(1+t)^{\kappa_k}dt$$

$$\leq \frac{C_{12}}{6}\int_0^T\int_0^M (1+t)^{\kappa_k}\,(r-r_\infty)^2\,dxdt + C, \tag{4.18}$$

$$\begin{aligned} Q_3 &\leq C\int_0^T\int_0^M (1+t)^{\eta_k}\rho^{1+\theta}(r^{n-1}u_x)_x^2 dxdt \\ &\quad + C\int_0^T\int_0^M (1+t)^{2\kappa_k-\eta_k-\frac{\eta_k\nu_1}{2}}(M-x)^{\frac{\theta-1}{\gamma}-\frac{\nu_1(\gamma+\theta)}{\gamma}}dxdt \\ &\leq C, \end{aligned} \tag{4.19}$$

where $\nu_1 = \frac{\gamma+\theta-1}{\gamma+\theta} - \epsilon_4$,

$$\|r(\cdot,t) - r_\infty(\cdot)\|_{L^\infty} \leq C(1+t)^{\frac{\nu_1\eta_k}{4}},$$

and

$$\begin{aligned} Q_4 &\leq C\int_0^T\int_0^M (1+t)^{\eta_k}\left(\rho^{1+\theta}(r^{n-1}u_x)_x^2 + \rho^{\theta-1}u^2\right)dxdt \\ &\quad + C\int_0^T\int_0^M (1+t)^{2\kappa_k-\eta_k-\frac{\eta_k\nu_1}{2}}(M-x)^{\frac{\theta-1}{\gamma}}dxdt \\ &\leq C. \end{aligned} \tag{4.20}$$

From (4.15)–(4.20), we finish the proof of (4.12) with $j = k$.

From (4.12) with $j = N_5$, using similar arguments as that in Lemma 3.11, we can obtain the following lemma and omit the proof.

LEMMA 4.4. *Under the assumptions of Theorem* 1.1*, we have*

$$\int_0^M (M-x)^{2-\frac{2\theta}{\gamma}}(\rho^\theta - \rho_\infty^\theta)_x^2 dx \leq C_\eta(1+t)^{\eta-\frac{\gamma+\theta-1}{\gamma+\theta}}, \tag{4.21}$$

and

$$\|\rho^\gamma(\cdot,t) - \rho_\infty^\gamma(\cdot)\|_{L^\infty} \leq C(1+t)^{\frac{\eta}{2}-\frac{3\gamma+3\theta-1}{4(\gamma+\theta+1)}},\ t \geq 0. \tag{4.22}$$

Thus, we finish the proof of Theorem 1.1.

REFERENCES

[1] D. BRESCH AND B. DESJARDINS, *On the construction of approximate solutions for the* 2*D viscous shallow water model and for compressible Navier-Stokes models*, J. Math. Pure Appl., 86 (2006), pp. 362–368.

[2] G. Q. CHEN AND M. KRATKA, *Global solutions to the Navier-Stokes equations for compressible heat-conducting flow with symmetry and free boundary*, Comm. Partial Differential Equations, 27 (2002), pp. 907–943.

[3] P. CHEN AND T. ZHANG, *A vacuum problem for multidimensional compressible Navier-Stokes equations with degenerate viscosity coefficients*, Commun. Pure Appl. Anal., 7 (2008), pp. 987–1016.

[4] B. DUCOMET AND A. A. ZLOTNIK, *Viscous compressible barotropic symmetric flows with free boundary under general mass force.* I. *uniform-in-time bounds and stabilization*, Math. Methods Appl. Sci., 28 (2005), pp. 827–863.

[5] B. DUCOMET AND A. A. ZLOTNIK, *Lyapunov functional method for* 1*D radiative and reactive viscous gas dynamics*, Arch. Ration. Mech. Anal., 177 (2005), pp. 185–229.

[6] H. GRAD, *Asymptotic theory of the Boltzmann equation* II, in Rarefied Gas Dynamics, Vol. 1. J. Laurmann, ed., Academic Press, New York, 1963, pp. 26–59.

[7] P. Hartman, *Ordinary Differential Equations*, Wiley, New York, 1964.
[8] D. Hoff and D. Serre, *The failure of continuous dependence on initial data for the Navier–Stokes equations of compressible flow*, SIAM J. Appl. Math., 51 (1991), pp. 887–898.
[9] T. Kobayashi and Y. Shibata, *Decay estimates of solutions for the equations of motion of compressible viscous and heat-conductive gases in an exterior domain in* R^3, Comm. Math. Phys., 200 (1999), pp. 621–659.
[10] P. L. Lions, *Mathematical Topics in Fluid Mechanics*, Vol. 1-2, Oxford University Press, New York, 1996, 1998.
[11] T. P. Liu, Z. P. Xin, and T. Yang, *Vacuum states of compressible flow*, Discrete Contin. Dyn. Syst., 4 (1998), pp. 1–32.
[12] A. Matsumura and S. Yanagi, *Uniform boundedness of the solutions for a one-dimensional isentropic model system of a compressible viscous gas*, Comm. Math. Phys., 175 (1996), pp. 259–274.
[13] Š. Matušů-Nečasová, M. Okada, and T. Makoni, *Free boundary problem for the equation of spherically symmetric motion of viscous gas* (II)–(III), Japan J. Indust. Appl. Math., 12 (1995) pp. 195–203; 14 (1997), pp. 199–213.
[14] M. Okada and T. Makino, *Free boundary value problems for the equation of spherically symmetrical motion of viscous gas*, Japan J. Appl. Math., 10 (1993), pp. 219–235.
[15] M. Okada, Š. Matušů-Nečasová, and T. Makino, *Free boundary problem for the equation of one-dimensional motion of compressible gas with density-dependent viscosity*, Ann. Univ. Ferrara Sez. VII (N.S.), 48 (2002), pp. 1–20.
[16] I. Straškraba and A. A. Zlotnik, *Global behavior of* $1d$*-viscous compressible barotropic fluid with a free boundary and large data*, J. Math. Fluid Mech., 5 (2003), pp. 119–143.
[17] V. A. Vaigant and A. V. Kazhikhov, *On existence of global solutions to the two-dimensional Navier-Stokes equations for a compressible viscous fluid*, Siberian Math. J., 36 (1995), pp. 1108–1141.
[18] S. Ukai, T. Yang, and H. J. Zhao, *Convergence rate for the compressible Navier-Stokes equations with external force*, J. Hyperbolic Differ. Equ., 3 (2006), pp. 561–574.
[19] S. W. Vong, T. Yang, and C. J. Zhu, *Compressible Navier-Stokes equations with degenerate viscosity coefficient and vacuum* (II), J. Differential Equations, 192 (2003), pp. 475–501.
[20] Z. P. Xin, *Blow-up of smooth solution to the compressible Navier-Stokes equations with compact density*, Comm. Pure Appl. Math., 51 (1998), pp. 229–240.
[21] T. Yang and C. J. Zhu, *Compressible Navier-Stokes equations with degenerate viscosity coefficient and vacuum*, Comm. Math. Phys., 230 (2002), pp. 329–363.
[22] T. Zhang and D. Y. Fang, *Global behavior of compressible Navier-Stokes equations with a degenerate viscosity coefficient*, Arch. Ration. Mech. Anal., 182 (2006), pp. 223–253.
[23] T. Zhang and D. Y. Fang, *Global behavior of spherically symmetric Navier-Stokes equations with density-dependent viscosity*, J. Differential Equations, 236 (2007), pp. 293–341.
[24] A. A. Zlotnik and B. Ducomet, *The stabilization rate and stability of viscous compressible barotropic symmetric flows with a free boundary for a general mass force*, Sb. Math., 196 (2005), pp. 1745–1799.

SIAM J. MATH. ANAL.
Vol. 40, No. 3, pp. 905–938

COMPACTLY SUPPORTED SYMMETRIC C^∞ WAVELETS WITH SPECTRAL APPROXIMATION ORDER*

BIN HAN† AND ZUOWEI SHEN‡

Abstract. In this paper, we obtain symmetric C^∞ real-valued tight wavelet frames in $L_2(\mathbb{R})$ with compact support and the spectral frame approximation order. Furthermore, we present a family of symmetric compactly supported C^∞ orthonormal complex wavelets in $L_2(\mathbb{R})$. A complete analysis of nonstationary tight wavelet frames and orthonormal wavelet bases in $L_2(\mathbb{R})$ is given.

Key words. symmetric tight wavelet frames, spectral frame approximation order, symmetric orthonormal complex wavelets, nonstationary cascade algorithm, nonstationary C^∞ wavelets

AMS subject classifications. 42C40, 41A25, 41A05, 42C05

DOI. 10.1137/060675009

1. Introduction.

In this paper, we are interested in symmetric compactly supported C^∞ tight wavelet frames with the spectral frame approximation order. Since it is impossible to achieve all these properties under the framework of stationary tight wavelet frames, it is natural for us to consider nonstationary tight wavelet frames, in particular, nonstationary tight wavelet frames derived from nonstationary multiresolution analysis by the new (nonstationary) unitary extension principle.

We start with a family of 2π-periodic trigonometric polynomials $\widehat{a_j}, j \in \mathbb{N}$, and their associated nonstationary refinable functions (or tempered distributions) $\phi_{j-1}, j \in \mathbb{N}$, defined by

$$\widehat{\phi_{j-1}}(\xi) := \widehat{a_j}(\xi/2)\widehat{\phi_j}(\xi/2) = \prod_{n=1}^{\infty} \widehat{a_{n+j-1}}(2^{-n}\xi), \qquad \xi \in \mathbb{R},\ j \in \mathbb{N}, \tag{1.1}$$

where the 2π-periodic trigonometric polynomials $\widehat{a_j}$, $j \in \mathbb{N}$, are called refinement *masks*. Here, the Fourier transform $\hat{f}$ of a function $f \in L_1(\mathbb{R})$ used in this paper is defined to be $\hat{f}(\xi) := \int_{\mathbb{R}} f(t)e^{-it\xi}\,dt$ and can be naturally extended to square integrable functions and tempered distributions.

The stationary multiresolution analysis corresponds to the case that all the masks $\widehat{a_j}$ are the same; therefore, all the functions ϕ_j are the same, and, in particular, $\widehat{\phi_0}(\xi) = \widehat{a_1}(\xi/2)\widehat{\phi_1}(\xi/2)$. We say that a function $\phi : \mathbb{R} \mapsto \mathbb{C}$ is refinable with a 2π-periodic trigonometric polynomial refinement mask $\hat{a}$ if $\hat{\phi}(\xi) = \hat{a}(\xi/2)\hat{\phi}(\xi/2)$. The frame generators ψ^ℓ are generally obtained from the refinable function ϕ via $\widehat{\psi^\ell}(\xi) = \widehat{b^\ell}(\xi/2)\hat{\phi}(\xi/2)$ for some 2π-periodic trigonometric polynomials $\widehat{b^\ell}$ with some desirable properties.

*Received by the editors November 15, 2006; accepted for publication (in revised form) April 3, 2008; published electronically September 8, 2008. This research was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC Canada) under grant RGP 228051 and in part by several grants from the National University of Singapore.
http://www.siam.org/journals/sima/40-3/67500.html

†Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton T6G 2G1, AB, Canada (bhan@math.ualberta.ca, http://www.ualberta.ca/~bhan).

‡Department of Mathematics, National University of Singapore, Singapore (matzuows@nus.edu.sg, http://www.math.nus.edu.sg/~matzuows).

A tight wavelet frame in $L_2(\mathbb{R})$ (in the stationary case) is generated by the integer translates and dyadic dilates of a finite set of elements in $L_2(\mathbb{R})$. More precisely, we say that $\{\psi^1, \ldots, \psi^L\}$ generates a (normalized) *tight wavelet frame* in $L_2(\mathbb{R})$ if

$$\|f\|^2_{L_2(\mathbb{R})} = \sum_{\ell=1}^{L} \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} |\langle f, \psi^\ell_{j,k}\rangle|^2 \qquad \forall\, f \in L_2(\mathbb{R}), \tag{1.2}$$

where $\psi^\ell_{j,k} := 2^{j/2}\psi^\ell(2^j \cdot -k)$ and $\langle f, g\rangle := \int_{\mathbb{R}} f(t)\overline{g(t)}\, dt$. As a redundant wavelet system, tight frame wavelet systems are easier to design and provide more flexibilities in applications than orthonormal wavelet bases, especially in image inpainting (see [1, 2, 3, 4, 5] for details). Because of this, tight wavelet frames have been extensively studied in the literature; to mention only a few, see [6, 7, 8, 14, 16, 18, 19, 21, 25, 26, 27, 33] and the many references therein.

Tight wavelet frames obtained from refinable functions are of particular interest, due to their associated multiresolution structure (which we refer to as MRA-based) and fast frame algorithms. Constructions of tight wavelet frames from a refinable function can be done by the unitary extension principle in [33]. In fact, many tight wavelet frames have been constructed in [6, 16, 33]. Later, by using the more general oblique extension principle, which is independently developed in [7, 14], more tight wavelet frames with various desirable properties have been obtained in [7, 14, 21, 25, 26, 27] and many other references therein.

For the stationary case, it already has been pointed out in [13] that there does not exist a compactly supported refinable function ϕ with a 2π-periodic trigonometric polynomial refinement mask such that ϕ belongs to $C^\infty(\mathbb{R})$. Hence, it is impossible to obtain MRA-based compactly supported (stationary) tight wavelet frames in $L_2(\mathbb{R})$ whose generators are in $C^\infty(\mathbb{R})$. However, it is shown in [10] that, by using the class of masks for orthonormal refinable functions of [12] whose integer shifts form an orthonormal system, one can obtain a family of nonstationary refinable functions such that every nonstationary refinable function belongs to $C^\infty(\mathbb{R})$ and its integer shifts still form an orthonormal system in $L_2(\mathbb{R})$. For this family of nonstationary refinable functions, a C^∞ nonstationary orthonormal wavelet basis in $L_2(\mathbb{R})$ is derived in [10]. In fact, ideas of generating a class of nonstationary refinable functions in $C^\infty(\mathbb{R})$ from a given family of masks for stationary refinable functions have already been discussed in [15, 35]. One such example is the *up-function* [10, 15, 35] generated from the family of masks for the B-splines. Let $\widehat{a_j}(\xi) = 2^{-j}(1 + e^{-i\xi})^j, j \in \mathbb{N}$, be the mask for the B-spline of order j and define $\phi_{j-1}, j \in \mathbb{N}$, as in (1.1). Then all $\phi_{j-1}, j \in \mathbb{N}$, are compactly supported C^∞ functions. In particular, the function ϕ_0 is supported on $[0, 2]$ (see [10, 15, 35]).

Motivated by the interesting work of Cohen and Dyn [10] and equipped with the pseudosplines (a more general class of refinable functions containing B-splines, interpolatory refinable functions, and Daubechies orthonormal refinable functions in [12] as special cases), together with the idea of the unitary extension principle, we establish the analysis needed here for constructing nonstationary $C^\infty(\mathbb{R})$ tight wavelet frames in $L_2(\mathbb{R})$ with desirable properties, especially the symmetry property, which cannot be achieved by real-valued orthonormal dyadic refinable functions. As we will see, the construction more or less follows the idea of the unitary extension principle for the stationary case, while the main analysis of this paper is somehow different from that of [10]. For example, in the orthonormal wavelet case, the approximation order of the truncated wavelet series in [10] is the same as that of the (nonstationary)

multiresolution analysis, while in the tight wavelet frame case, they are different, even for the stationary case, as shown in [14].

Next, we briefly describe ideas of the construction of tight wavelet frames. Although one of our major objectives of this paper is to use the family of refinement masks for pseudosplines to construct tight wavelet frames and to provide the corresponding analysis, the construction in this paper is given for the general setting.

We start with 2π-periodic measurable functions $\widehat{a_j}$, $j \in \mathbb{N}$, as a sequence of refinement masks. To make the idea of the unitary extension principle work, it is necessary to require that for every $j \in \mathbb{N}$, the mask $\widehat{a_j}$ should satisfy

$$|\widehat{a_j}(\xi)|^2 + |\widehat{a_j}(\xi+\pi)|^2 \leqslant 1 \qquad \text{a.e. } \xi \in \mathbb{R}. \tag{1.3}$$

Since we are interested only in compactly supported tight wavelet frames, it is natural to start with compactly supported refinable functions ϕ_j, which, in turn, require that the degrees of the trigonometric polynomials $\widehat{a_j}$ do not increase too fast. For a 2π-periodic trigonometric polynomial $\hat{a}$, we denote $\deg(\hat{a})$ the smallest nonnegative integer such that its Fourier coefficients of $\hat{a}$ vanish outside $[-\deg(\hat{a}), \deg(\hat{a})]$. We note that $\deg(\hat{a})$ defined here is somewhat slightly different from the usual definition of the degree of a trigonometric polynomial; $\deg(\hat{a})$ here is the minimal integer k such that $[-k,k]$ contains the support of the Fourier coefficients of both $\hat{a}$ and $\hat{a}(-\cdot)$. For ϕ_0 in (1.1) to be compactly supported, by a simple calculation, it is very natural to require [10] that

$$\sum_{j=1}^{\infty} 2^{-j} \deg(\widehat{a_j}) < \infty. \tag{1.4}$$

With (1.3) and (1.4), under the condition that $\sum_{j=1}^{\infty} |\widehat{a_j}(0) - 1| < \infty$, it can be proved that all the corresponding refinable functions ϕ_{j-1} in (1.1) are well-defined compactly supported functions in $L_2(\mathbb{R})$.

Wavelet functions ψ_{j-1}^ℓ, $j \in \mathbb{N}$ and $\ell \in \{1, \dots, \mathcal{J}_j\}$, are obtained from ϕ_j by

$$\widehat{\psi_{j-1}^\ell}(\xi) := \widehat{b_j^\ell}(\xi/2)\widehat{\phi_j}(\xi/2), \qquad \ell = 1, \dots, \mathcal{J}_j, \tag{1.5}$$

where $\mathcal{J}_j$ are positive integers and each $\widehat{b_j^\ell}$, $\ell = 1, \dots, \mathcal{J}_j$, is called a (high-pass) wavelet mask. Denote $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. We say that $\{\phi_0\} \cup \{\psi_j^\ell \,:\, j \in \mathbb{N}_0, \ell = 1, \dots, \mathcal{J}_{j+1}\}$ generates *a nonstationary tight wavelet frame* in $L_2(\mathbb{R})$ if

$$\{\phi_0(\cdot - k) \,:\, k \in \mathbb{Z}\} \cup \{\psi_{j;j,k}^\ell := 2^{j/2}\psi_j^\ell(2^j \cdot - k) \,:\, j \in \mathbb{N}_0, k \in \mathbb{Z}, \ell = 1, \dots, \mathcal{J}_{j+1}\} \tag{1.6}$$

is a tight frame of $L_2(\mathbb{R})$; that is, the following holds:

$$\|f\|_{L_2(\mathbb{R})}^2 = \sum_{k\in\mathbb{Z}} |\langle f, \phi_0(\cdot - k)\rangle|^2 + \sum_{j=0}^{\infty} \sum_{\ell=1}^{\mathcal{J}_{j+1}} \sum_{k\in\mathbb{Z}} |\langle f, \psi_{j;j,k}^\ell\rangle|^2 \quad \forall\, f \in L_2(\mathbb{R}). \tag{1.7}$$

We say that ψ_j^ℓ has ν *vanishing moments* if $\widehat{\psi_j^\ell}^{(n)}(0) = 0$ for all $n = 0, \dots, \nu - 1$, where $\widehat{\psi_j^\ell}^{(n)}$ denotes the nth derivative of $\widehat{\psi_j^\ell}$. It is clear that (1.7) is equivalent to

$$f = \sum_{k\in\mathbb{Z}} \langle f, \phi_0(\cdot - k)\rangle \phi_0(\cdot - k) + \sum_{j=0}^{\infty} \sum_{\ell=1}^{\mathcal{J}_{j+1}} \sum_{k\in\mathbb{Z}} \langle f, \psi_{j;j,k}^\ell\rangle \psi_{j;j,k}^\ell, \qquad f \in L_2(\mathbb{R}). \tag{1.8}$$

The frame approximation operators $Q_n, n \in \mathbb{N}$, associated with the truncation of the tight wavelet frame in (1.6) at level n, are defined to be

$$Q_n(f) := \sum_{k\in\mathbb{Z}} \langle f, \phi_0(\cdot-k)\rangle \phi_0(\cdot-k) + \sum_{j=0}^{n-1} \sum_{\ell=1}^{\mathcal{J}_{j+1}} \sum_{k\in\mathbb{Z}} \langle f, \psi^\ell_{j;j,k}\rangle \psi^\ell_{j;j,k}, \qquad f \in L_2(\mathbb{R}). \tag{1.9}$$

For $\nu \geqslant 0$, we denote $W_2^\nu(\mathbb{R})$ the Sobolev space of all functions $f \in L_2(\mathbb{R})$ such that

$$\|f\|^2_{W_2^\nu(\mathbb{R})} := \int_{\mathbb{R}} (1+|\xi|^{2\nu}) |\hat{f}(\xi)|^2 \, d\xi < \infty. \tag{1.10}$$

The Sobolev seminorm $|f|_{W_2^\nu(\mathbb{R})}$ is defined to be

$$|f|^2_{W_2^\nu(\mathbb{R})} := \int_{\mathbb{R}} |\xi|^{2\nu} |\hat{f}(\xi)|^2 \, d\xi, \qquad f \in W_2^\nu(\mathbb{R}). \tag{1.11}$$

The unitary extension principle provides a sufficient condition on the wavelet masks $\widehat{a_j}$ and $\widehat{b_j^\ell}$, $\ell = 1, \ldots, \mathcal{J}_j$, so that, with ψ^ℓ_{j-1} defined in (1.5), the wavelet system in (1.6) forms a tight frame in $L_2(\mathbb{R})$. Altogether, we have the following result on nonstationary tight wavelet frames in $L_2(\mathbb{R})$.

THEOREM 1.1. *Let $\widehat{a_j}, j \in \mathbb{N}$, be 2π-periodic trigonometric polynomials with $\widehat{a_j}(0) = 1$ for all $j \in \mathbb{N}$. If (1.3) and (1.4) hold, letting ϕ_j and ψ^ℓ_{j-1}, $j \in \mathbb{N}$ and $\ell \in \{1, \ldots, \mathcal{J}_j\}$, be defined in (1.1) and (1.5), respectively, then the following hold:*

(i) *All functions ϕ_{j-1}, $j \in \mathbb{N}$, are well-defined compactly supported functions in $L_2(\mathbb{R})$.*

(ii) *If $\widehat{b_j^\ell}$, $j \in \mathbb{N}$ and $\ell \in \{1, \ldots, \mathcal{J}_j\}$, are 2π-periodic trigonometric polynomials satisfying*

$$\begin{aligned} &|\widehat{a_j}(\xi)|^2 + \sum_{\ell=1}^{\mathcal{J}_j} |\widehat{b_j^\ell}(\xi)|^2 = 1 \quad \text{and} \\ &\widehat{a_j}(\xi)\overline{\widehat{a_j}(\xi+\pi)} + \sum_{\ell=1}^{\mathcal{J}_j} \widehat{b_j^\ell}(\xi)\overline{\widehat{b_j^\ell}(\xi+\pi)} = 0, \end{aligned} \tag{1.12}$$

then the wavelet system in (1.6) is a compactly supported tight wavelet frame in $L_2(\mathbb{R})$.

(iii) *If, in addition to (1.12), we assume that*

$$\deg(\widehat{a_j}) = O(j^\alpha 2^{\beta j}) \qquad \text{as} \quad j \to \infty \quad \text{for some} \quad \alpha \geqslant 0,\ 0 \leqslant \beta < 1 \tag{1.13}$$

and assume that there exist a positive number $\nu \in \frac{1}{2}\mathbb{N}$ and a positive integer N such that $1 - |\widehat{a_j}(\xi)|^2$ has a zero of order 2ν at $\xi = 0$ for all $j \geqslant N$, that is, for $j \geqslant N$,

$$|\widehat{a_j}(\xi)|^2 = 1 + O(|\xi|^{2\nu}), \qquad \xi \to 0, \tag{1.14}$$

then there exists a positive constant C, independent of f and n, such that

$$\begin{aligned} \|f - Q_n(f)\|_{L_2(\mathbb{R})} &\leqslant C n^{\nu\alpha} 2^{-\nu(1-\beta)n} |f|_{W_2^\nu(\mathbb{R})} \\ &\forall\, f \in W_2^\nu(\mathbb{R}) \quad \text{and} \quad n \geqslant N, \end{aligned} \tag{1.15}$$

where the linear operators Q_n are defined in (1.9).

Item (ii) of Theorem 1.1 is called the unitary extension principle for the nonstationary case. Theorem 1.1 will be proved in section 4. As we shall see in section 4, the main effort there is to prove items (i) and (iii) of Theorem 1.1. To show item (ii) of Theorem 1.1, one needs to show the convergence of the frame series in the right side of (1.8) to the function f in $L_2(\mathbb{R})$. When (1.13) holds, the convergence of the frame series follows from (iii) by observing that the masks in item (ii) satisfy (1.14) for $\nu = 1/2$. Furthermore, a refined analysis establishes the convergence of the frame series even without assuming (1.13).

We further remark that (1.12) guarantees the multiresolution frame decomposition algorithm whose proof can be straightforwardly verified and is more or less known. In fact, Theorem 1.1 generalizes the unitary extension principle from the stationary case in [33] to the general nonstationary case. It is clear that, similar to the stationary case, for every fixed $j \in \mathbb{N}$, in order to construct a set of 2π-periodic trigonometric polynomials $\widehat{b_j^\ell}$, $\ell = 1, \ldots, \mathcal{J}_j$, derived from the mask $\widehat{a_j}$ so that (1.12) is satisfied, the mask $\widehat{a_j}$ must satisfy (1.3). Hence, (1.3) is a necessary and sufficient condition to make (1.12) hold, as we shall see later in this section.

Nonstationary spline tight wavelet frames using the oblique extension principle developed in [7, 14] have been systematically studied in Chui, He, and Stöckler [8] recently. There, they considered an even more general nonstationary setting; i.e., it is not even shift-invariant at each level. Since the oblique extension principle is a generalization of the unitary extension principle, the proof of [8] might be modified to prove item (ii) of Theorem 1.1. However, this at most leads to the conclusion $Q_n(f) \to f$ in $L_2(\mathbb{R})$. Our approach of item (ii) is beyond the proof of the tight frame property itself in (1.8). Instead, we analyze the approximation power of the truncated tight wavelet frame series as stated in item (iii) of Theorem 1.1. As a consequence of this analysis, we obtain the tight frame property stated in item (ii) of Theorem 1.1. Finally, we remark that a systematic study of general nonstationary wavelet frames that may not be MRA-based was given in [34]

Following [14], we say that a tight wavelet frame $\{\phi_0(\cdot - k) : k \in \mathbb{Z}\} \cup \{\psi_{j;j,k}^\ell : j \in \mathbb{N}_0, k \in \mathbb{Z}, \ell = 1, \ldots, \mathcal{J}_{j+1}\}$ provides *frame approximation order* ν if there exist a positive constant C, independent of f and n, and a positive integer N such that

$$\|f - Q_n(f)\|_{L_2(\mathbb{R})} \leqslant C2^{-\nu n}|f|_{W_2^\nu(\mathbb{R})} \quad \forall\, f \in W_2^\nu(\mathbb{R}) \quad \text{and} \quad n \geqslant N. \tag{1.16}$$

We say that a tight wavelet frame provides *the spectral frame approximation order* if it provides frame approximation order ν for any positive integer ν. Here, we point out that for the frame approximation order discussed in this paper, the constant C in (1.15) of Theorem 1.1 and the constant C in (1.16) of Theorems 1.2 and 1.4 can be explicitly obtained.

In section 4, we shall study when $Q_n(f)$ approaches f with an approximation order ν as $n \to \infty$. As a consequence, we prove item (ii) of Theorem 1.1 and (1.8) by showing that $Q_n(f) \to f$ in $L_2(\mathbb{R})$ as $n \to \infty$ for every $f \in L_2(\mathbb{R})$, provided that the conditions in Theorem 1.1 are satisfied. The approximation order of $Q_n(f)$ was not studied in [10], since it was not needed there. In fact, since only orthonormal wavelet systems were considered in [10], the associated operators Q_n become orthogonal projections and attain the approximation order provided by the nonstationary multiresolution analysis. Therefore, one only needs to understand the conditions under which $Q_n(f) \to f$ in $L_2(\mathbb{R})$ as $n \to \infty$ in the orthonormal wavelet case. Nevertheless, our approach here applies to this special case as well and simplifies the conditions given in [10]. The approximation order of $Q_n(f)$ was not studied in [8] either, since it is a

more challenging problem in its more general setting of [8]. For the stationary case, it is evident that (1.13) holds with $\alpha = \beta = 0$ and, consequently, the notion of the frame approximation power in (1.15) agrees with that of the frame approximation order in (1.16). However, we shall present an example of nonstationary tight wavelet frames derived from the up-function (see Theorem 1.3) to demonstrate that (1.15) holds with $\nu = 2$, $\alpha = 1$, and $\beta = 0$, while (1.16) fails for any $\nu > 0$; that is, this particular nonstationary tight wavelet frame has a "weak" frame approximation order 2 in the sense of (1.15), but it does not have any "strong" frame approximation order in the sense of (1.16).

Finally, we note that the 2π-periodic trigonometric polynomial wavelet masks $\widehat{b_j^\ell}$, $j \in \mathbb{N}$ and $\ell \in \{1, \ldots, \mathcal{J}_j\}$, can be constructed from the masks $\widehat{a_j}$ by many ways provided that the refinement masks $\widehat{a_j}, j \in \mathbb{N}$, satisfy (1.3). Here is one such construction modified from the stationary case of [6] (also cf. [16, 27, 25]). For every $j \in \mathbb{N}$, from the mask $\widehat{a_j}$ with real coefficients and satisfying (1.3), define

$$\begin{aligned} \widehat{b_j^1}(\xi) &:= e^{-i\xi}\overline{\widehat{a_j}(\xi+\pi)}, \\ \widehat{b_j^2}(\xi) &:= 2^{-1}[A_j(\xi) + e^{-i\xi}\overline{A_j(\xi)}], \\ \widehat{b_j^3}(\xi) &:= 2^{-1}[A_j(\xi) - e^{-i\xi}\overline{A_j(\xi)}], \end{aligned} \tag{1.17}$$

where A_j is a π-periodic trigonometric polynomial with real coefficients such that

$$|A_j(\xi)|^2 = 1 - |\widehat{a_j}(\xi)|^2 - |\widehat{a_j}(\xi+\pi)|^2.$$

Then, $\widehat{a_j}, \widehat{b_j^1}, \widehat{b_j^2}$, and $\widehat{b_j^3}$, $j \in \mathbb{N}$, satisfy (1.12) with $\mathcal{J}_j = 3$. Furthermore, the corresponding wavelets defined by (1.5) using masks in (1.17) are symmetric or antisymmetric whenever ϕ_j is symmetric.

After establishing Theorem 1.1, we focus on constructing nonstationary $C^\infty(\mathbb{R})$ wavelets derived from a family of refinement masks for pseudosplines. Pseudosplines (of type I) were first introduced in [14] and [36] to improve the approximation order of truncated tight wavelet frame series for the tight wavelet frame system obtained by the unitary extension principle. The pseudosplines in [14] are generally not symmetric. The pseudosplines of type II are symmetric and were introduced in [16]. Since we are aiming at constructing symmetric tight wavelet frames, we will use pseudosplines of type II. For positive integers $m, l \in \mathbb{N}$, throughout the paper we denote

$$P_{m,l}(x) := \sum_{j=0}^{l-1} \binom{m+j-1}{j} x^j = \sum_{j=0}^{l-1} \frac{(m+j-1)!}{j!(m-1)!} x^j, \qquad x \in \mathbb{R}. \tag{1.18}$$

The masks for pseudosplines of type II with order (m, l) [16] are given by

$$\widehat{a_{m,l}}(\xi) := \cos^{2m}(\xi/2) P_{m,l}(\sin^2(\xi/2)), \qquad m \in \mathbb{N},\ l = 1, \ldots, m. \tag{1.19}$$

Since it is evident that $\widehat{a_{m,l}}(\xi) \geqslant 0$ for all $\xi \in \mathbb{R}$, the mask $\widehat{a_{m,l}^I}$, for the pseudospline of type I with order (m, l) introduced in [14] and [36], is obtained by taking the square root of the mask $\widehat{a_{m,l}}$ in (1.19) for the pseudospline of type II with order (m, l) using the Fejér–Riesz lemma such that

$$|\widehat{a_{m,l}^I}(\xi)|^2 = \widehat{a_{m,l}}(\xi), \qquad \xi \in \mathbb{R}. \tag{1.20}$$

While the pseudosplines of type II and their masks in (1.19) are symmetric, their type I counterparts usually do not have symmetry. For the case $l = 1$, the corresponding refinable pseudosplines are B-splines for both types. For the case $l = m$, the corresponding refinable pseudospline ϕ of type I with mask $\widehat{a^I_{m,m}}$ in (1.20) has orthonormal integer shifts (i.e., $\{\phi(\cdot - k) : k \in \mathbb{Z}\}$ is an orthonormal system in $L_2(\mathbb{R})$), and the corresponding refinable pseudospline ϕ of type II with mask $\widehat{a_{m,m}}$ in (1.19) is interpolatory (i.e., $\phi(0) = 1$ and $\phi(k) = 0$ for all $k \in \mathbb{Z}\backslash\{0\}$). It is easy to verify that the condition in (1.3) is satisfied for all the masks for pseudosplines of type I and type II (e.g., see [14, 16]).

Our construction here employs masks $\widehat{a_{m,l}}$ in (1.19) for pseudosplines of type II, since we are interested in constructing symmetric tight wavelet frames. We have the following result on symmetric C^∞ tight wavelet frames in $L_2(\mathbb{R})$ with compact support and the spectral frame approximation order.

THEOREM 1.2. *Let $\widehat{a_j} := \widehat{a_{m_j,l_j}}$ be defined in* (1.19)*, where $1 \leqslant l_j \leqslant m_j$ and m_j ($j \in \mathbb{N}$) are positive integers satisfying*

$$\lim_{j\to\infty} m_j = \infty \quad \text{and} \quad \sum_{j=1}^{\infty} 2^{-j} m_j < \infty. \tag{1.21}$$

For $j \in \mathbb{N}$, define ϕ_{j-1} as in (1.1) *and ψ^1_{j-1}, ψ^2_{j-1}, and ψ^3_{j-1} as in* (1.5) *with the wavelet masks $\widehat{b^1_j}$, $\widehat{b^2_j}$, and $\widehat{b^3_j}$ being derived from $\widehat{a_j}$ in* (1.17)*. Then the following hold:*

1. *Each nonstationary refinable function ϕ_j, $j \in \mathbb{N}_0$, is a compactly supported C^∞ real-valued function that is symmetric about the origin: $\phi_j(-\cdot) = \phi_j$.*
2. *Each wavelet function ψ^ℓ_j, $\ell = 1, 2, 3$ and $j \in \mathbb{N}_0$, is a compactly supported C^∞ function with l_{j+1} vanishing moments and satisfies $\psi^\ell_j(1 - \cdot) = \psi^\ell_j$ for $\ell = 1, 2$ and $\psi^3_j(1 - \cdot) = -\psi^3_j$.*
3. *The system $\{\phi_0(\cdot - k) : k \in \mathbb{Z}\} \cup \{\psi^\ell_{j;j,k} := 2^{j/2}\psi^\ell_j(2^j \cdot -k) : j \in \mathbb{N}_0, k \in \mathbb{Z}, \ell = 1, 2, 3\}$ is a compactly supported symmetric C^∞ tight wavelet frame in $L_2(\mathbb{R})$.*
4. *If in addition $\liminf_{j\to\infty} l_j/m_j > 0$, then the tight wavelet frame in item* (3) *has the spectral frame approximation order.*

The simplest choice in Theorem 1.2 is $m_j = l_j = j$ for all $j \in \mathbb{N}$, for which the condition in (1.21) is evidently satisfied and $\liminf_{j\to\infty} l_j/m_j = 1 > 0$. Therefore, by Theorem 1.2, we have a symmetric C^∞ tight wavelet frame in $L_2(\mathbb{R})$ with compact support and the spectral frame approximation order. Of course, the claims in Theorem 1.2 also hold if one chooses $m \approx \rho_1 j$ and $l_j \approx \rho_2 j$ for all $j \in \mathbb{N}$ with some fixed positive numbers ρ_1 and ρ_2. In order to have refinable functions ϕ_j, $j \in \mathbb{N}$, in (1.1) with support as small as possible, one should choose a sequence $\{m_j\}_{j=1}^\infty$ so that m_j goes to ∞ as slowly as possible. This is one of our motivations to choose a general integer m_j instead of the standard choice $m_j = j$ for our setup. We point out that such a strategy has already been considered by Cohen [9]. We also mention that all the claims in Theorem 1.2 hold, except possibly for the symmetry property, if one chooses the masks $\widehat{a_j} := \widehat{a^I_{m_j,l_j}}$ in (1.20) for the pseudosplines of type I instead of type II in Theorem 1.2.

It is clear that the frame approximation order in (1.16) implies (1.15). For the stationary case, it is evident that (1.13) holds with $\alpha = \beta = 0$ and, consequently, the notion of the frame approximation power in (1.15) agrees with that of the frame approximation order in (1.16). However, as illustrated by the following result, they could be quite different in the case of nonstationary tight wavelet frames.

THEOREM 1.3. *Let* $\widehat{a_j}(\xi) := 2^{-j}(1+e^{-i\xi})^j$, $j \in \mathbb{N}$, *be the masks for the up-function; in other words, we take* $m_j := j$ *and* $l_j := 1$ *in Theorem* 1.2. *For* $j \in \mathbb{N}$, *define* ϕ_{j-1} *as in* (1.1) *and* ψ^1_{j-1}, ψ^2_{j-1}, *and* ψ^3_{j-1} *as in* (1.5) *with the wavelet masks* $\widehat{b^1_j}$, $\widehat{b^2_j}$, *and* $\widehat{b^3_j}$ *being derived from* $\widehat{a_j}$ *in* (1.17). *Then the following hold:*

(i) $\{\phi_0(\cdot - k) : k \in \mathbb{Z}\} \cup \{\psi^\ell_{j;j,k} : j \in \mathbb{N}_0, k \in \mathbb{Z}, \ell = 1,2,3\}$ *is a compactly supported symmetric* C^∞ *tight wavelet frame in* $L_2(\mathbb{R})$, *and each* ψ_j *has one vanishing moment.*

(ii) *There exists a positive constant* C, *independent of* f *and* n, *such that*

$$\|f - Q_n(f)\|_{L_2(\mathbb{R})} \leqslant Cn^2 2^{-2n} |f|_{W^2_2(\mathbb{R})} \quad \forall\, f \in W^2_2(\mathbb{R}) \quad \text{and} \quad n \geqslant 2, \tag{1.22}$$

where the linear operators Q_n *are defined in* (1.9).

(iii) *The nonstationary tight wavelet frame in* (i) *does not have any frame approximation order; i.e., for any given* $\nu > 0$, *there does not exist a positive constant* C *such that* (1.16) *is satisfied.*

The Daubechies orthogonal masks $\widehat{a^I_{j,j}}$ in (1.20) with real coefficients for the pseudosplines of type I with order (j,j) have been considered in [10] (also see [9] for the general case $\widehat{a^I_{m_j,m_j}}$) to obtain C^∞ compactly supported (nonstationary) orthonormal refinable functions, from which (nonstationary) orthonormal wavelets with the spectral approximation order are derived in [9, 10]. However, it is well known [13] that such Daubechies orthogonal masks $\widehat{a^I_{j,j}}$, having real coefficients and obtained from $\widehat{a_{j,j}}$ via the Fejér–Riesz lemma in (1.20), are not symmetric (except $j = 1$) and therefore, all the associated nonstationary refinable functions $\phi_j, j \in \mathbb{N}_0$, are not symmetric. One way to achieve symmetry is to split the masks $\widehat{a_{j,j}}$ into masks that are similar to $\widehat{a^I_{j,j}}$ in (1.20) but allow complex-valued coefficients (see [30]). Examples of symmetric orthonormal complex wavelets were first constructed in [30] in this way from Daubechies orthogonal masks of odd orders. Recently, symmetric orthonormal complex-valued wavelets have been systematically studied in Han [23].

Let $P_{j,j}$ be the polynomial defined in (1.18). For an odd integer j, one can always construct [23, Lemma 6 and section 2] two polynomials P^r_j and P^i_j with real coefficients such that

$$P_{j,j}(x) = [P^r_j(x)]^2 + [P^i_j(x)]^2, \qquad x \in \mathbb{R} \quad \text{with} \quad P^r_j(0) = 1, \quad P^i_j(0) = 0. \tag{1.23}$$

Now define

$$\widehat{a^S_j}(\xi) := e^{i(j-1)\xi/2} 2^{-j}(1+e^{-i\xi})^j [P^r_j(\sin^2(\xi/2)) + iP^i_j(\sin^2(\xi/2))]. \tag{1.24}$$

It is easy to check [23, Lemma 3] that $|\widehat{a^S_j}(\xi)|^2 = |\widehat{a^I_{j,j}}(\xi)|^2 = \widehat{a_{j,j}}(\xi)$ and the integer shifts of the stationary refinable function associated with the mask $\widehat{a^S_j}$ are orthonormal. Using this family of masks, one can obtain symmetric C^∞ orthonormal complex wavelets with compact support. We summarize the above discussion into the following result.

THEOREM 1.4. *Let* $m_j, j \in \mathbb{N}$, *be positive odd integers such that* (1.21) *holds. Take* $\widehat{a_j}(\xi) := \widehat{a^S_{m_j}}$, *where* $\widehat{a^S_{m_j}}$ *is defined in* (1.24). *Define*

$$\widehat{\psi_{j-1}}(2\xi) := e^{-i\xi}\overline{\widehat{a_j}(\xi+\pi)}\widehat{\phi_j}(\xi), \qquad j \in \mathbb{N},$$

where $\phi_j, j \in \mathbb{N}_0$, *are defined in* (1.1). *Then the following hold:*

(1) *Each refinable function* $\phi_j, j \in \mathbb{N}_0$, *is a compactly supported* C^∞ *complex-valued function such that* $\phi_j(1-\cdot) = \phi_j$ *and* $\{\phi_j(\cdot - k) : k \in \mathbb{Z}\}$ *is an orthonormal system in* $L_2(\mathbb{R})$.

(2) *Each wavelet function* $\psi_j, j \in \mathbb{N}_0$, *is a compactly supported* C^∞ *complex-valued function such that* $\psi_j(1-\cdot) = -\psi_j$ *and* ψ_j *has* m_{j+1} *vanishing moments.*

(3) $\{\phi_0(\cdot - k) : k \in \mathbb{Z}\} \cup \{\psi_{j;j,k} := 2^{j/2}\psi_j(2^j \cdot -k) : j \in \mathbb{N}_0, k \in \mathbb{Z}\}$ *is a compactly supported symmetric* C^∞ *orthonormal basis of* $L_2(\mathbb{R})$ *and has the spectral approximation order.*

This paper is organized as follows. In section 2, we shall discuss nonstationary cascade algorithms and some properties of nonstationary refinable functions. In particular, we study the initial functions in a nonstationary cascade algorithm and provide a sufficient condition for the convergence of a nonstationary cascade algorithm in a Sobolev space $W_2^\nu(\mathbb{R})$. As a consequence, we obtain a characterization for nonstationary orthonormal wavelet bases in $L_2(\mathbb{R})$. In section 3, we shall study the frame approximation order of a nonstationary tight wavelet frame. The proofs of Theorems 1.1–1.4 will be given in section 4.

2. Nonstationary cascade algorithms and refinable functions. In this section, we first discuss the existence of L_2-solutions of nonstationary refinable functions for a general set of masks satisfying (1.3). In fact, this follows from the following result, proved in this section:

$$[\widehat{\phi_j}, \widehat{\phi_j}](\xi) := \sum_{k\in\mathbb{Z}} |\widehat{\phi_j}(\xi + 2\pi k)|^2 \leqslant 1, \qquad \text{a.e. } \xi \in \mathbb{R} \quad \forall\, j \in \mathbb{N}_0 \tag{2.1}$$

for all masks satisfying (1.3), provided that the infinite products in (1.1) exist almost everywhere. The above inequality plays a critical role in our study of nonstationary tight wavelet frames and their frame approximation orders.

The question of when the refinable functions are in Sobolev spaces is discussed next. In fact, we prove it as a consequence of the convergence of the cascade algorithm in various Sobolev spaces when $\widehat{a_j}$, $j \in \mathbb{N}$, are masks of pseudosplines. The proof is done in the Fourier domain with the initial function whose Fourier transform is the characteristic function of $[-\pi, \pi]$. We then prove that when the cascade algorithm converges for one initial function, it converges for a large class of functions. Although a similar result is well known for the stationary case, it is not straightforward for the case of nonstationary cascade algorithms. However, this result is important in computer aided geometric design, because it results in a compactly supported function in each iteration of the cascade algorithm that generates a curve from a finitely supported sequence of points to approximate the underlying curve. Hence, it is desirable to prove the convergence of a cascade algorithm with a compactly supported initial function instead of an infinitely supported band-limited function in computer aided geometric design [17].

For the nonstationary refinable functions $\phi_j, j \in \mathbb{N}_0$, defined by masks for pseudosplines as in Theorem 1.2, one could use the same techniques developed in [10] to show that the cascade algorithm converges for the special initial function whose Fourier transform is the characteristic function of $[-\pi, \pi]$ that leads to $\phi_j \in C^\infty(\mathbb{R})$ for all $j \in \mathbb{N}_0$. But our discussion on nonstationary cascade algorithms in this section will supplement the results in [10] on nonstationary cascade algorithms. We use the results in [10] whenever they can be directly applied, e.g., Lemma 2.1, and at the same

time develop our own results to achieve our goal with a systematic and comprehensive approach. We also believe that some results (e.g., Theorem 2.4 and Lemmas 2.2 and 2.7) derived in this section have their own value in addition to being used to prove Theorem 2.8 in this section.

2.1. L_2-solutions. We start with a basic property about the pointwise convergence of the infinite product in (1.1). A sufficient condition for the convergence of the infinite product in (1.1) has been established in the following lemma by Cohen and Dyn in [10, Theorem 2.1].

LEMMA 2.1. *Let $\widehat{a_j}, j \in \mathbb{N}$, be 2π-periodic trigonometric polynomials such that $\sup_{j\in\mathbb{N}} \|\widehat{a_j}\|_{L_\infty(\mathbb{R})} < \infty$. If (1.4) holds and $\sum_{j=1}^{\infty} |\widehat{a_j}(0) - 1| < \infty$, then the infinite product in (1.1) converges uniformly on every compact set of $\mathbb{R}$ and all $\phi_j, j \in \mathbb{N}_0$, in (1.1) are well-defined compactly supported tempered distributions.*

Next, we consider when $\phi_j \in L_2(\mathbb{R})$, $j \in \mathbb{N}_0$, provided that the infinite products in (1.1) exist almost everywhere. In order to investigate the frame approximation order of a nonstationary tight wavelet frame, we establish (2.1), which is the following lemma.

LEMMA 2.2. *Let $\widehat{a_j}, j \in \mathbb{N}$, be 2π-periodic measurable functions satisfying (1.3) for each $j \in \mathbb{N}$. Assume that, for every $j \in \mathbb{N}_0$, $\widehat{\phi_j}(\xi) := \lim_{N\to\infty} \prod_{n=1}^{N} \widehat{a_{n+j}}(2^{-n}\xi)$ is well defined for almost every $\xi \in \mathbb{R}$; that is, the infinite product in (1.1) exists for almost every point in $\mathbb{R}$. Then (2.1) holds, and, consequently, $\phi_j \in L_2(\mathbb{R})$ with $\|\phi_j\|_{L_2(\mathbb{R})} \leqslant 1$ for every $j \in \mathbb{N}_0$.*

Proof. It suffices to prove the case $j = 0$, since the proof of the general case $j \in \mathbb{N}_0$ is the same. Note that $\widehat{\phi_0}(\xi) = \lim_{n\to\infty} \prod_{j=1}^{n} \widehat{a_j}(2^{-j}\xi)$ for almost every $\xi \in \mathbb{R}$. For any fixed positive integer K, we have

$$\sum_{k=-K}^{K} |\widehat{\phi_0}(\xi + 2\pi k)|^2 = \lim_{n\to\infty} \sum_{k=-K}^{K} \prod_{j=1}^{n} \big|\widehat{a_j}\big(2^{-j}(\xi + 2\pi k)\big)\big|^2 \qquad \text{a.e. } \xi \in \mathbb{R}. \tag{2.2}$$

Let N be the smallest positive integer such that $N > 1 + \log_2 K$. Then we have $[-K, K] \subseteq [-K, 2^N - 1 - K]$. Consequently, for all $n \geqslant N$, we have $[-K, K] \subseteq [-K, 2^n - 1 - K]$ and

$$\sum_{k=-K}^{K} \prod_{j=1}^{n} \big|\widehat{a_j}\big(2^{-j}(\xi + 2\pi k)\big)\big|^2 \leqslant \sum_{k=-K}^{2^n-1-K} \prod_{j=1}^{n} \big|\widehat{a_j}\big(2^{-j}(\xi + 2\pi k)\big)\big|^2. \tag{2.3}$$

Let $L_\infty(\mathbb{T}) := \{f \in L_\infty(\mathbb{R}) : f \text{ is } 2\pi\text{-periodic}\}$. The transition operator $T_j : L_\infty(\mathbb{T}) \to L_\infty(\mathbb{T})$ is defined for each $f \in L_\infty(\mathbb{T})$ as follows:

$$[T_j f](\xi) := |\widehat{a_j}(\xi/2)|^2 f(\xi/2) + |\widehat{a_j}(\xi/2 + \pi)|^2 f(\xi/2 + \pi), \qquad \xi \in \mathbb{R}.$$

Observing that $|[T_j f](\xi)| \leqslant \big[|\widehat{a_j}(\xi/2)|^2 + |\widehat{a_j}(\xi/2+\pi)|^2\big]\|f\|_{L_\infty(\mathbb{R})}$, by (1.3), we deduce that

$$\|T_j f\|_{L_\infty(\mathbb{R})} \leqslant \|f\|_{L_\infty(\mathbb{R})} \big\||\widehat{a_j}(\cdot/2)|^2 + |\widehat{a_j}(\cdot/2 + \pi)|^2\big\|_{L_\infty(\mathbb{R})} \leqslant \|f\|_{L_\infty(\mathbb{R})}. \tag{2.4}$$

By induction on n, we can verify (e.g., [21, Lemma 2.1]) that

$$\sum_{k=-K}^{2^n-1-K} \prod_{j=1}^{n} \big|\widehat{a_j}\big(2^{-j}(\xi + 2\pi k)\big)\big|^2 = [T_1 T_2 \cdots T_{n-1} T_n 1](\xi), \qquad \xi \in \mathbb{R}. \tag{2.5}$$

Now it follows from (2.3) and (2.4) that, for $n \geqslant N$ and for almost every $\xi \in \mathbb{R}$,

$$\sum_{k=-K}^{K} \prod_{j=1}^{n} \bigl|\widehat{a_j}\bigl(2^{-j}(\xi+2\pi k)\bigr)\bigr|^2 \leqslant [T_1T_2\cdots T_{n-1}T_n1](\xi) \leqslant \|T_1T_2\cdots T_{n-1}T_n1\|_{L_\infty(\mathbb{R})} \leqslant 1.$$

That is, we have

$$\sum_{k=-K}^{K} \prod_{j=1}^{n} \bigl|\widehat{a_j}\bigl(2^{-j}(\xi+2\pi k)\bigr)\bigr|^2 \leqslant 1 \qquad \text{a.e. } \xi \in \mathbb{R},\ n \geqslant N.$$

It follows from the above inequality and (2.2) that, for any fixed positive integer K,

$$\sum_{k=-K}^{K} |\widehat{\phi_0}(\xi+2\pi k)|^2 = \lim_{n\to\infty} \sum_{k=-K}^{K} \prod_{j=1}^{n} \bigl|\widehat{a_j}\bigl(2^{-j}(\xi+2\pi k)\bigr)\bigr|^2 \leqslant 1 \qquad \text{a.e. } \xi \in \mathbb{R}.$$

Taking $K \to \infty$ in the above inequality, we conclude that (2.1) is true for $j = 0$.

Since (2.1) implies that $\|\widehat{\phi_j}\|^2_{L_2(\mathbb{R})} = \int_{-\pi}^{\pi} [\widehat{\phi_j}, \widehat{\phi_j}](\xi)\, d\xi \leqslant \int_{-\pi}^{\pi} 1\, d\xi = 2\pi$, by Plancherel's theorem, it follows that $\|\phi_j\|_{L_2(\mathbb{R})} \leqslant 1$. □

As Lemma 2.1 states, the assumption that, for $j \in \mathbb{N}_0$,

$$\widehat{\phi_j}(\xi) := \lim_{N\to\infty} \prod_{n=1}^{N} \widehat{a_{n+j}}(2^{-n}\xi)$$

is well defined for almost every $\xi \in \mathbb{R}$, required in Lemma 2.2, is satisfied whenever the conditions $\widehat{a_j}(0) = 1$, $j \in \mathbb{N}$, and (1.4) hold. In other words, Lemma 2.2 says that if the masks $\widehat{a_j}$, $j \in \mathbb{N}$, satisfy (1.3), (1.4), and $\widehat{a_j}(0) = 1$, then the corresponding nonstationary refinable functions $\phi_j \in L_2(\mathbb{R})$, $j \in \mathbb{N}_0$.

Since the approximation property of $\phi_j, j \in \mathbb{N}_0$, discussed in this paper depends only on ϕ_j for large enough j, without loss of generality, throughout the paper we shall assume that the normalization condition $\widehat{a_j}(0) = 1$ holds for all $j \in \mathbb{N}$. In fact, if the conclusion in Lemma 2.1 holds, since $\prod_{n=1}^{\infty} \widehat{a_{n+j}}(0)$ converges and is nonzero for sufficiently large j, then we can replace $\widehat{\phi_j}$ and $\widehat{a_j}$ with $\widehat{\phi_j}/\widehat{\phi_j}(0)$ and $\widehat{a_j}(\xi)/\widehat{a_j}(0)$, respectively.

2.2. Cascade algorithms. A cascade algorithm is often used to study various properties of refinable functions and is closely related to a subdivision scheme in computer aided geometric design for generating smooth curves [10, 15, 20, 24]. For a given sequence of masks $\{\widehat{a_j}\}_{j=1}^{\infty}$, starting with an initial function $f \in L_2(\mathbb{R})$, one computes a sequence of cascade functions f_n by

$$\widehat{f_n}(\xi) := \hat{f}(2^{-n}\xi) \prod_{j=1}^{n} \widehat{a_j}(2^{-j}\xi), \qquad \xi \in \mathbb{R},\ n \in \mathbb{N}. \tag{2.6}$$

If $\lim_{n\to\infty} \hat{f}(2^{-n}\xi) = 1$ and $\hat{\phi}(\xi) := \lim_{n\to\infty} \prod_{j=1}^{n} \widehat{a_j}(2^{-j}\xi)$ exists for almost every $\xi \in \mathbb{R}$, then by (2.6) it is evident that $\lim_{n\to\infty} \widehat{f_n}(\xi) = \hat{\phi}(\xi)$ for almost every $\xi \in \mathbb{R}$.

The cascade algorithm is closely related to another algorithm, called a subdivision scheme, which we define next. For a sequence $u : \mathbb{Z} \mapsto \mathbb{C}$, we denote $\hat{u}$ its Fourier series as $\hat{u}(\xi) := \sum_{k\in\mathbb{Z}} u(k)e^{-ik\xi}$. In particular, by δ we denote the *Dirac sequence* on $\mathbb{Z}$ such

that $\delta(0)=1$ and $\delta(k)=0$ for $k\in\mathbb{Z}\backslash\{0\}$. That is, $\hat{\delta}=1$. For a sequence u and a mask a, the subdivision operator S_a maps the sequence u into a new sequence, $S_a u$ on $\mathbb{Z}$, which is determined by $\widehat{S_a u}(\xi)=2\hat{a}(\xi)\hat{u}(2\xi)$. In fact, the product $2^n\prod_{j=1}^n \widehat{a_j}(2^{n-j}\xi)$ is the Fourier series of the subdivision sequence $S_{a_n}S_{a_{n-1}}\cdots S_{a_2}S_{a_1}\delta$. More precisely, it follows from (2.6) that the cascade sequence $\{f_n\}_{n=1}^\infty$ and the subdivision sequence $\{S_{a_n}S_{a_{n-1}}\cdots S_{a_2}S_{a_1}\delta\}_{n=1}^\infty$ are related by

$$f_n=\sum_{k\in\mathbb{Z}}[S_{a_n}S_{a_{n-1}}\cdots S_{a_2}S_{a_1}\delta](k)f(2^n\cdot-k),\qquad n\in\mathbb{N}. \tag{2.7}$$

Recall that $f\in W_2^\nu(\mathbb{R})$ if $\|f\|^2_{W_2^\nu(\mathbb{R})}:=\int_{\mathbb{R}}(1+|\xi|^{2\nu})|\hat{f}(\xi)|^2\,d\xi<\infty$. For a sequence of masks $\{\widehat{a_j}\}_{j=1}^\infty$ and an initial function $f\in W_2^\nu(\mathbb{R})$, we say that the (nonstationary) cascade algorithm associated with masks $\{\widehat{a_j}\}_{j=1}^\infty$ and an initial function f converges in the Sobolev space $W_2^\nu(\mathbb{R})$ if $f_n\in W_2^\nu(\mathbb{R})$ for all $n\in\mathbb{N}$ and the sequence $\{f_n\}_{n=1}^\infty$ is convergent in $W_2^\nu(\mathbb{R})$. Many (but not all) functions in $W_2^\nu(\mathbb{R})$ can serve as an initial function in a cascade algorithm. One popular and natural choice of an initial function f in computer aided geometric design is from the B-spline functions, since they are compactly supported functions of piecewise polynomials. Hence, it is easy to compute the values of the underlying approximating function. However, to analyze the convergence of the cascade algorithm in the frequency domain, the sinc function $f(x)=\frac{\sin(\pi x)}{\pi x}$, that is, $\widehat{f}=\chi_{[-\pi,\pi]}$, the characteristic function of the interval $[-\pi,\pi]$, is a more natural choice (see, e.g., [9, 10, 11, 13, 22]). Our analysis will show that the cascade algorithm generated by pseudospline masks with the sinc function being the initial function converges in $W_2^\nu(\mathbb{R})$. To make sure that this cascade algorithm also converges in $W_2^\nu(\mathbb{R})$ when the initial seed is replaced by splines, we prove a more general result as follows: if a cascade algorithm converges in $W_2^\nu(\mathbb{R})$ for one initial seed with stable integer shifts, then it converges in $W_2^\nu(\mathbb{R})$ for a class of initial functions. As we will see, the proof is more technical than the stationary case, because a stationary refinable function is a fixed point of a stationary cascade algorithm, while this is no longer the case for the nonstationary case.

Before proceeding further, let us introduce the following notation. For $\nu\in\mathbb{R}$ and $f\in L_2(\mathbb{R})$, we define

$$[\hat{f},\hat{f}]_\nu(\xi):=\frac{1}{|\xi|^{2\nu}}\sum_{k\in\mathbb{Z}}|\hat{f}(\xi+2\pi k)|^2|\xi+2\pi k|^{2\nu},\qquad \xi\in\mathbb{R}, \tag{2.8}$$

and

$$\{\hat{f},\hat{f}\}_\nu(\xi):=\frac{1}{|\xi|^{2\nu}}\sum_{k\in\mathbb{Z}\backslash\{0\}}|\hat{f}(\xi+2\pi k)|^2|\xi+2\pi k|^{2\nu},\qquad \xi\in\mathbb{R}. \tag{2.9}$$

Clearly, we have $[\hat{f},\hat{f}]:=[\hat{f},\hat{f}]_0=\sum_{k\in\mathbb{Z}}|\hat{f}(\cdot+2\pi k)|^2$ and

$$[\hat{f},\hat{f}]_\nu(\xi)=\{\hat{f},\hat{f}\}_\nu(\xi)+|\hat{f}(\xi)|^2. \tag{2.10}$$

Following [20, 22], we introduce the set $\mathcal{F}_\nu$ of initial functions in a cascade algorithm as

$$\begin{aligned}\mathcal{F}_\nu:=\Big\{f\in W_2^\nu(\mathbb{R})\ :\ &\lim_{n\to\infty}\hat{f}(2^{-n}\xi)=1,\quad \lim_{n\to\infty}\{\hat{f},\hat{f}\}_\nu(2^{-n}\xi)=0,\\ &\text{a.e. }\xi\in[-\pi,\pi],\quad [\hat{f},\hat{f}]_\nu\in L_\infty([-\pi,\pi])\Big\}.\end{aligned} \tag{2.11}$$

The following result will be needed later; its proof is rather simple and therefore is omitted.

LEMMA 2.3. *Let $f \in \mathcal{F}_\nu$ for some $\nu \geqslant 0$. Then $[\hat{f}, \hat{f}](\xi) \leqslant [\hat{f}, \hat{f}]_\nu(\xi)$ for almost every $\xi \in [-\pi, \pi]$ (consequently, $[\hat{f}, \hat{f}] \in L_\infty(\mathbb{R})$), and*

$$\lim_{n\to\infty} [\hat{f}, \hat{f}](2^{-n}\xi) = \lim_{n\to\infty} [\hat{f}, \hat{f}]_\nu(2^{-n}\xi) = 1 \qquad \text{a.e. } \xi \in \mathbb{R}. \tag{2.12}$$

We say that the integer shifts of a function $f \in L_2(\mathbb{R})$ are *stable* in $L_2(\mathbb{R})$ if there exists a positive constant C such that

$$C^{-1} \leqslant [\hat{f}, \hat{f}](\xi) \leqslant C, \qquad \text{a.e. } \xi \in \mathbb{R}. \tag{2.13}$$

Now we state the following result on an initial function with stable integer shifts in a nonstationary cascade algorithm.

THEOREM 2.4. *Let $\widehat{a_j}, j \in \mathbb{N}$, be 2π-periodic measurable functions such that*

$$\widehat{f_\infty}(\xi) := \lim_{n\to\infty} \prod_{j=1}^{n} \widehat{a_j}(2^{-j}\xi)$$

exists for almost every $\xi \in \mathbb{R}$. For a function $f \in \mathcal{F}_\nu$ with $\nu \geqslant 0$ and stable integer shifts, define $f_n, n \in \mathbb{N}$, by (2.6). Assume that $\{f_n\}_{n=1}^\infty$ converges in $W_2^\nu(\mathbb{R})$. Then it converges to f_∞ in $W_2^\nu(\mathbb{R})$. Furthermore, for every $g \in \mathcal{F}_\nu$, the sequence of functions $g_n, n \in \mathbb{N}$, defined by

$$\widehat{g_n}(\xi) := \hat{g}(2^{-n}\xi) \prod_{j=1}^{n} \widehat{a_j}(2^{-j}\xi), \qquad \xi \in \mathbb{R},\ n \in \mathbb{N}, \tag{2.14}$$

converges to f_∞ in $W_2^\nu(\mathbb{R})$.

Proof. By the definition of f_n in (2.6), we deduce that

$$\begin{aligned}
&\int_{\mathbb{R}} (1 + |\xi|^{2\nu}) |\widehat{f_n}(\xi)|^2 \, d\xi \\
&\qquad = \int_{\mathbb{R}} \chi_{[-\pi,\pi]}(2^{-n}\xi) \Big([\hat{f}, \hat{f}](2^{-n}\xi) + |\xi|^{2\nu} [\hat{f}, \hat{f}]_\nu(2^{-n}\xi) \Big) \prod_{j=1}^{n} |\widehat{a_j}(2^{-j}\xi)|^2 \, d\xi.
\end{aligned}$$

That is, (2.6) implies

$$\|f_n\|^2_{W_2^\nu(\mathbb{R})} := \int_{\mathbb{R}} (1 + |\xi|^{2\nu}) |\widehat{f_n}(\xi)|^2 \, d\xi = \int_{\mathbb{R}} F_n(\xi) \, d\xi \tag{2.15}$$

with

$$F_n(\xi) := \Big([\hat{f}, \hat{f}](2^{-n}\xi) + |\xi|^{2\nu} [\hat{f}, \hat{f}]_\nu(2^{-n}\xi) \Big) \chi_{[-\pi,\pi]}(2^{-n}\xi) \prod_{j=1}^{n} |\widehat{a_j}(2^{-j}\xi)|^2, \quad \xi \in \mathbb{R}.$$

Similarly, by (2.14), we deduce that

$$\|g_n\|^2_{W_2^\nu(\mathbb{R})} := \int_{\mathbb{R}} (1 + |\xi|^{2\nu}) |\widehat{g_n}(\xi)|^2 \, d\xi = \int_{\mathbb{R}} G_n(\xi) \, d\xi \tag{2.16}$$

with

$$G_n(\xi) := \Big([\hat{g},\hat{g}](2^{-n}\xi) + |\xi|^{2\nu}[\hat{g},\hat{g}]_\nu(2^{-n}\xi)\Big)\chi_{[-\pi,\pi]}(2^{-n}\xi)\prod_{j=1}^{n}|\widehat{a_j}(2^{-j}\xi)|^2.$$

On the one hand, since $g \in \mathcal{F}_\nu$, by the definition of $\mathcal{F}_\nu$ in (2.11) and Lemma 2.3, we see that there exists a positive constant C_1 such that $[\hat{g},\hat{g}](\xi) \leqslant [\hat{g},\hat{g}]_\nu(\xi) \leqslant C_1$ for almost every $\xi \in [-\pi,\pi]$. By Lemma 2.3, it follows from (2.13) that $C^{-1} \leqslant [\hat{f},\hat{f}]_\nu(\xi)$ for almost every $\xi \in [-\pi,\pi]$ and

$$[\hat{g},\hat{g}](\xi) \leqslant C_1 \leqslant CC_1[\hat{f},\hat{f}](\xi) \quad \text{and} \quad [\hat{g},\hat{g}]_\nu(\xi) \leqslant C_1 \leqslant CC_1[\hat{f},\hat{f}]_\nu(\xi) \quad \text{a.e. } \xi \in [-\pi,\pi].$$

Now it follows from the above inequalities that

$$|G_n(\xi)| \leqslant CC_1F_n(\xi) \qquad \text{a.e. } \xi \in \mathbb{R} \quad \text{and} \quad n \in \mathbb{N}. \tag{2.17}$$

On the other hand, by $f, g \in \mathcal{F}_\nu$ and Lemma 2.3, since $\lim_{n\to\infty}\prod_{j=1}^{n}\widehat{a_j}(2^{-j}\xi) = \widehat{f_\infty}(\xi)$ for almost every $\xi \in \mathbb{R}$, we see that $\lim_{n\to\infty}F_n(\xi) = \lim_{n\to\infty}G_n(\xi) = (1+|\xi|^{2\nu})|\widehat{f_\infty}(\xi)|^2$ for almost every $\xi \in \mathbb{R}$. Since $\{f_n\}_{n=1}^\infty$ is a convergent sequence in $W_2^\nu(\mathbb{R})$ and $\lim_{n\to\infty}\widehat{f_n}(\xi) = \widehat{f_\infty}(\xi)$ for almost every $\xi \in \mathbb{R}$, we have $f_\infty \in W_2^\nu(\mathbb{R})$ and $\lim_{n\to\infty}\|f_n - f_\infty\|_{W_2^\nu(\mathbb{R})} = 0$. In particular, we have $\lim_{n\to\infty}\|f_n\|^2_{W_2^\nu(\mathbb{R})} = \|f_\infty\|^2_{W_2^\nu(\mathbb{R})}$. By (2.15), we have $\|f_n\|^2_{W_2^\nu(\mathbb{R})} = \int_{\mathbb{R}}F_n(\xi)\,d\xi$. Therefore, we conclude that $\lim_{n\to\infty}\int_{\mathbb{R}}F_n(\xi)\,d\xi = \|f_\infty\|^2_{W_2^\nu(\mathbb{R})}$. Now by (2.17) and the generalized Lebesgue dominated convergence theorem, it follows from (2.16) and $\lim_{n\to\infty}G_n(\xi) = (1+|\xi|^{2\nu})|\widehat{f_\infty}(\xi)|^2$ for almost every $\xi \in \mathbb{R}$ that

$$\lim_{n\to\infty}\int_{\mathbb{R}}(1+|\xi|^{2\nu})|\widehat{g_n}(\xi)|^2\,d\xi = \lim_{n\to\infty}\int_{\mathbb{R}}G_n(\xi)\,d\xi = \int_{\mathbb{R}}(1+|\xi|^{2\nu})|\widehat{f_\infty}(\xi)|^2\,d\xi. \tag{2.18}$$

But we also have

$$(1+|\xi|^{2\nu})|\widehat{g_n}(\xi) - \widehat{f_\infty}(\xi)|^2 \leqslant 2(1+|\xi|^{2\nu})\Big[|\widehat{g_n}(\xi)|^2 + |\widehat{f_\infty}(\xi)|^2\Big].$$

By (2.18), we have

$$\lim_{n\to\infty}\int_{\mathbb{R}}2(1+|\xi|^{2\nu})\Big[|\widehat{g_n}(\xi)|^2 + |\widehat{f_\infty}(\xi)|^2\Big]\,d\xi = 4\int_{\mathbb{R}}(1+|\xi|^{2\nu})|\widehat{f_\infty}(\xi)|^2\,d\xi < \infty.$$

Now by the generalized Lebesgue dominated convergence theorem again, we conclude that

$$\begin{aligned}\lim_{n\to\infty}\|g_n - f_\infty\|^2_{W_2^\nu(\mathbb{R})} &= \lim_{n\to\infty}\int_{\mathbb{R}}(1+|\xi|^{2\nu})|\widehat{g_n}(\xi) - \widehat{f_\infty}(\xi)|^2\,d\xi \\ &= \int_{\mathbb{R}}(1+|\xi|^{2\nu})\lim_{n\to\infty}|\widehat{g_n}(\xi) - \widehat{f_\infty}(\xi)|^2\,d\xi = 0,\end{aligned}$$

since $\lim_{n\to\infty}\widehat{g_n}(\xi) = \widehat{f_\infty}(\xi)$ for almost every $\xi \in \mathbb{R}$. This completes the proof. □

As shown in the next result, the conditions in (2.11) and (2.13) are not very restrictive. In fact, the sinc function and all B-spline functions belong to $\mathcal{F}_\nu$ for some ν.

PROPOSITION 2.5. *If $f(x) := \frac{\sin(\pi x)}{\pi x}$, then $f \in \mathcal{F}_\nu$ for all $\nu \geqslant 0$ and (2.13) holds, where $\mathcal{F}_\nu$ is defined in (2.11). For the* B*-spline B_m of order m, that is, $\widehat{B_m}(\xi) =$*

$\left(\frac{1-e^{-i\xi}}{i\xi}\right)^m$, $B_m \in \mathcal{F}_\nu$ *and* (2.13) *holds for all* $0 \leqslant \nu < m - 1/2$. *Note that* $B_m \notin W_2^{m-1/2}(\mathbb{R})$.

Proof. Letting $\nu \geqslant 0$, we note that $\hat{f} = \chi_{[-\pi,\pi]}$ and therefore, $f \in W_2^\nu(\mathbb{R})$. For $\xi \in (-\pi, \pi)$, we have $\hat{f}(\xi + 2\pi k) = 0$ for all $k \in \mathbb{Z}\backslash\{0\}$ and, hence,

$$\{\hat{f}, \hat{f}\}_\nu(\xi) := \frac{1}{|\xi|^{2\nu}} \sum_{k\in\mathbb{Z}\backslash\{0\}} |\hat{f}(\xi + 2\pi k)|^2 |\xi + 2\pi k|^{2\nu} = 0, \qquad \xi \in (\pi, \pi)\backslash\{0\}.$$

Therefore, it is evident that $\lim_{n\to\infty} \hat{f}(2^{-n}\xi) = 1$ and $\lim_{n\to\infty}\{\hat{f}, \hat{f}\}_\nu(2^{-n}\xi) = 0$ for almost every $\xi \in [-\pi, \pi]$. Moreover, by (2.10), we have

$$[\hat{f}, \hat{f}]_\nu(\xi) = |\hat{f}(\xi)|^2 + \{\hat{f}, \hat{f}\}_\nu(\xi) = |\hat{f}(\xi)|^2 = 1, \qquad \xi \in (-\pi, \pi)\backslash\{0\}.$$

Hence, $[\hat{f}, \hat{f}]_\nu \in L_\infty([-\pi, \pi])$. Thus, $f \in \mathcal{F}_\nu$. Inequality (2.13) is obviously true by $[\hat{f}, \hat{f}](\xi) = 1$ for almost every $\xi \in \mathbb{R}$.

Letting $0 \leqslant \nu < m-1/2$, from $\widehat{B_m}(\xi) = \left(\frac{1-e^{-i\xi}}{i\xi}\right)^m$, we have $\lim_{j\to\infty} \widehat{B_m}(2^{-j}\xi) = \widehat{B_m}(0) = 1$, $B_m \in W_2^\nu(\mathbb{R})$, and $|\widehat{B_m}(\xi)|^2 = \frac{\sin^{2m}(\xi/2)}{(\xi/2)^{2m}}$. Now a simple computation shows that

$$\{\widehat{B_m}, \widehat{B_m}\}_\nu(\xi) \leqslant 2^{2(m-\nu)} \sin^{2(m-\nu)}(\xi/2) \sum_{k\in\mathbb{Z}\backslash\{0\}} |\xi + 2\pi k|^{-2(m-\nu)}.$$

Since $0 \leqslant \nu < m - 1/2$, we have $2(m - \nu) > 1$. Hence, for all $\xi \in [-\pi, \pi]$, the series $\sum_{k\in\mathbb{Z}\backslash\{0\}} |\xi + 2\pi k|^{-2(m-\nu)}$ uniformly converges and therefore is uniformly bounded. Now it follows from the above inequality that $\{\widehat{B_m}, \widehat{B_m}\}_\nu(\xi)$ is uniformly bounded on $[-\pi, \pi]$ and

$$\lim_{n\to\infty} \{\widehat{B_m}, \widehat{B_m}\}_\nu(2^{-n}\xi) \leqslant \lim_{n\to\infty} \sin^{2(m-\nu)}(2^{-n-1}\xi) = 0 \qquad \forall\, \xi \in [-\pi, \pi].$$

Now it follows from (2.10) that $[\widehat{B_m}, \widehat{B_m}]_\nu \in L_\infty([-\pi, \pi])$, since $\widehat{B_m} \in L_\infty(\mathbb{R})$ and $\{\widehat{B_m}, \widehat{B_m}\}_\nu \in L_\infty([-\pi, \pi])$. It is well known that (2.13) holds for B_m. □

The following result provides us a sufficient condition on the convergence of a nonstationary cascade algorithm in a Sobolev space $W_2^\nu(\mathbb{R})$. As we will see, this result is sufficient to study the convergence of nonstationary cascade algorithms with masks for pseudosplines.

PROPOSITION 2.6. *Let* $\widehat{a_j}$ *and* $\widehat{b_j}$ $(j \in \mathbb{N})$ *be* 2π*-periodic measurable functions such that, for all* $j \in \mathbb{N}$,

$$|\widehat{a_j}(\xi)| \leqslant |\widehat{b_j}(\xi)| \qquad \text{a.e. } \xi \in \mathbb{R}. \tag{2.19}$$

Let $\eta \in W_2^\nu(\mathbb{R})$ *such that* $\lim_{j\to\infty} \hat{\eta}(2^{-j}\xi) = 1$ *for almost every* $\xi \in \mathbb{R}$. *Define*

$$\widehat{f_n}(\xi) := \hat{\eta}(2^{-n}\xi) \prod_{j=1}^n \widehat{a_j}(2^{-j}\xi) \quad \text{and} \quad \widehat{g_n}(\xi) := \hat{\eta}(2^{-n}\xi) \prod_{j=1}^n \widehat{b_j}(2^{-j}\xi), \qquad \xi \in \mathbb{R}.$$

Assume that $\widehat{f_\infty}(\xi) := \lim_{n\to\infty} \prod_{j=1}^n \widehat{a_j}(2^{-j}\xi)$ *and* $\widehat{g_\infty}(\xi) := \lim_{n\to\infty} \prod_{j=1}^n \widehat{b_j}(2^{-j}\xi)$ *are well defined for almost every* $\xi \in \mathbb{R}$. *Then,* $\lim_{n\to\infty} \|g_n - g_\infty\|_{W_2^\nu(\mathbb{R})} = 0$ *implies*

$\lim_{n\to\infty}\|f_n - f_\infty\|_{W_2^\nu(\mathbb{R})} = 0$. *In particular, suppose that there are a positive integer* J *and a* 2π*-periodic measurable function* $\hat{b}$ *such that*

$$|\widehat{a_j}(\xi)| \leqslant |\hat{b}(\xi)| \qquad \text{a.e. } \xi \in \mathbb{R},\ \forall\, j > J \quad \text{and} \quad \widehat{a_j} \in L_\infty(\mathbb{R}), \qquad 1 \leqslant j \leqslant J. \tag{2.20}$$

For $n \in \mathbb{N}$, *define* $\widehat{h_n}(\xi) := \hat{\eta}(2^{-n}\xi)\prod_{j=1}^n \hat{b}(2^{-j}\xi)$. *If* $\{h_n\}_{n=1}^\infty$ *converges in* $W_2^\nu(\mathbb{R})$, *then* f_n *converges to* f_∞ *in* $W_2^\nu(\mathbb{R})$; *i.e.,* $\lim_{n\to\infty}\|f_n - f_\infty\|_{W_2^\nu(\mathbb{R})} = 0$.

Proof. The assumption that the functions $\widehat{f_\infty}$ and $\widehat{g_\infty}$ are well defined implies that $\lim_{n\to\infty}\widehat{f_n}(\xi) = \widehat{f_\infty}(\xi)$ and $\lim_{n\to\infty}\widehat{g_n}(\xi) = \widehat{g_\infty}(\xi)$ for almost every $\xi \in \mathbb{R}$.

By assumption, $\{g_n\}_{n=1}^\infty$ converges to g_∞ in $W_2^\nu(\mathbb{R})$; together with the fact that $\lim_{n\to\infty}\widehat{g_n}(\xi) = \widehat{g_\infty}(\xi)$ for almost every $\xi \in \mathbb{R}$, we must have $g_\infty \in W_2^\nu(\mathbb{R})$ and $\lim_{n\to\infty}\|g_n - g_\infty\|_{W_2^\nu(\mathbb{R})} = 0$. In particular, we have

$$\lim_{n\to\infty}\int_{\mathbb{R}}(1+|\xi|^{2\nu})|\widehat{g_n}(\xi)|^2\,d\xi = \int_{\mathbb{R}}(1+|\xi|^{2\nu})|\widehat{g_\infty}(\xi)|^2\,d\xi < \infty.$$

Denote $\eta_n(\xi) := 2(1+|\xi|^{2\nu})\big[|\widehat{g_n}(\xi)|^2 + |\widehat{g_\infty}(\xi)|^2\big]$. It follows from the above identity that

$$\lim_{n\to\infty}\int_{\mathbb{R}}\eta_n(\xi)\,d\xi = 4\int_{\mathbb{R}}(1+|\xi|^{2\nu})|\widehat{g_\infty}(\xi)|^2\,d\xi < \infty. \tag{2.21}$$

By (2.19), it follows from the definition of $\widehat{g_n}$ and $\widehat{f_n}$ that $|\widehat{f_n}(\xi)| \leqslant |\widehat{g_n}(\xi)|$ for almost every $\xi \in \mathbb{R}$. Since we have $\lim_{n\to\infty}\widehat{f_n}(\xi) = \widehat{f_\infty}(\xi)$ for almost every $\xi \in \mathbb{R}$, we also have $|\widehat{f_\infty}(\xi)| \leqslant |\widehat{g_\infty}(\xi)|$ for almost every $\xi \in \mathbb{R}$. Consequently, by $g_\infty, g_n \in W_2^\nu(\mathbb{R})$ for all $n \in \mathbb{N}$, we have $f_\infty, f_n \in W_2^\nu(\mathbb{R})$ for all $n \in \mathbb{N}$. Moreover, we have

$$(1+|\xi|^{2\nu})|\widehat{f_n}(\xi) - \widehat{f_\infty}(\xi)|^2 \leqslant 2(1+|\xi|^{2\nu})\big[|\widehat{f_n}(\xi)|^2 + |\widehat{f_\infty}(\xi)|^2\big] \leqslant \eta_n(\xi) \qquad \text{a.e. } \xi \in \mathbb{R}.$$

By (2.21) and the generalized Lebesgue dominated convergence theorem, we conclude that

$$\begin{aligned}\lim_{n\to\infty}\|f_n - f_\infty\|_{W_2^\nu(\mathbb{R})}^2 &= \lim_{n\to\infty}\int_{\mathbb{R}}(1+|\xi|^{2\nu})|\widehat{f_n}(\xi) - \widehat{f_\infty}(\xi)|^2\,d\xi \\ &= \int_{\mathbb{R}}\lim_{n\to\infty}(1+|\xi|^{2\nu})|\widehat{f_n}(\xi) - \widehat{f_\infty}(\xi)|^2\,d\xi = 0.\end{aligned}$$

Therefore, $\lim_{n\to\infty}\|g_n - g_\infty\|_{W_2^\nu(\mathbb{R})} = 0$ implies $\lim_{n\to\infty}\|f_n - f_\infty\|_{W_2^\nu(\mathbb{R})} = 0$.

If (2.20) holds, for $n > J$ we deduce that, for almost every $\xi \in \mathbb{R}$,

$$\begin{aligned}|\widehat{f_n}(\xi)| &= \left[\prod_{j=1}^J |\widehat{a_j}(2^{-j}\xi)|\right]|\hat{\eta}(2^{-n}\xi)|\prod_{j=J+1}^n |\widehat{a_j}(2^{-j}\xi)| \\ &\leqslant C|\hat{\eta}(2^{-n}\xi)|\prod_{j=J+1}^n |\hat{b}(2^{-j}\xi)| = C|\widehat{h_{n-J}}(2^{-J}\xi)|,\end{aligned}$$

where $C := \prod_{j=1}^J \|\widehat{a_j}\|_{L_\infty(\mathbb{R})} < \infty$. Since $\{h_n\}_{n=1}^\infty$ is convergent in $W_2^\nu(\mathbb{R})$, it is evident that $\{2^J h_{n-J}(2^J\cdot)\}_{n=J+1}^\infty$ is also convergent in $W_2^\nu(\mathbb{R})$. Note that the Fourier transform of $2^J h_{n-J}(2^J\cdot)$ is $\widehat{h_{n-J}}(2^{-J}\cdot)$. Now by the generalized Lebesgue dominated

convergence theorem, we conclude that $\{f_n\}_{n=1}^\infty$ is also convergent in $W_2^\nu(\mathbb{R})$; that is, we have $\lim_{n\to\infty}\|f_n - f_\infty\|_{W_2^\nu(\mathbb{R})} = 0$. □

Let $\{\widehat{a_j}\}_{j=1}^\infty$ be a sequence of 2π-periodic measurable functions. Define $\{f_n\}_{n=1}^\infty$ by

$$\widehat{f_n}(\xi) := \chi_{[-\pi,\pi]}(2^{-n}\xi)\prod_{j=1}^{n}\widehat{a_j}(2^{-j}\xi), \qquad \xi\in\mathbb{R},\ n\in\mathbb{N}, \tag{2.22}$$

where $\chi_{[-\pi,\pi]}$ denotes the characteristic function of the interval $[-\pi,\pi]$. This can be understood as a representation of the nonstationary cascade algorithm associated with the masks $\{\widehat{a_j}\}_{n=1}^\infty$ in the frequency domain. Due to Theorem 2.4, we say that a nonstationary cascade algorithm associated with masks $\{\widehat{a_j}\}_{j=1}^\infty$ converges in $W_2^\nu(\mathbb{R})$ if the sequence $\{f_n\}_{n=1}^\infty$ in (2.22) converges in $W_2^\nu(\mathbb{R})$. Note that the initial function here in (2.22) is the sinc function $f(x) = \frac{\sin(\pi x)}{\pi x}$ since $\hat{f}(\xi) = \chi_{[-\pi,\pi]}$. Similarly, we say that a stationary cascade algorithm associated with a mask $\hat{a}$ converges in $W_2^\nu(\mathbb{R})$ if the cascade algorithm associated with $\{\widehat{a_j}\}_{j=1}^\infty$ converges in $W_2^\nu(\mathbb{R})$ with $\widehat{a_j} = \hat{a}$ for all $j\in\mathbb{N}$.

Basically, Proposition 2.6 says that if (2.19) holds and if the nonstationary cascade algorithm associated with $\{\widehat{b_j}\}_{j=1}^\infty$ converges in $W_2^\nu(\mathbb{R})$, then so does the nonstationary cascade algorithm associated with $\{\widehat{a_j}\}_{j=1}^\infty$. Similarly, if (2.20) holds and the stationary cascade algorithm associated with mask $\hat{b}$ converges in $W_2^\nu(\mathbb{R})$, then Proposition 2.6 says that the nonstationary cascade algorithm associated with masks $\{\widehat{a_j}\}_{j=1}^\infty$ must converge in $W_2^\nu(\mathbb{R})$.

The convergence of a stationary cascade algorithm associated with a finitely supported mask can be verified easily by calculating the spectrum of the transition operator. Let $\hat{a}$ be a 2π-periodic trigonometric polynomial with $\hat{a}(0) = 1$. Write $\hat{a}(\xi) = (1+e^{-i\xi})^m\hat{c}(\xi)$ for some nonnegative integer m and some 2π-periodic trigonometric polynomial $\hat{c}(\xi)$ with $\hat{c}(\pi)\neq 0$. Write $|\hat{c}(\xi)|^2 = \sum_{k=-K}^K c_k e^{-ik\xi}$, where K is some nonnegative integer. Denote $\rho(\hat{a})$ the spectral radius of the square matrix $(c_{2j-k})_{-K\leqslant j,k\leqslant K}$ and define $\nu_2(\hat{a}) := -1/2 - \log_2\sqrt{\rho(\hat{a})}$. It is known ([20, Theorem 4.3 and Proposition 7.2] and [22, Theorem 2.1]) that the stationary cascade algorithm associated with a 2π-periodic trigonometric polynomial mask $\hat{a}$ converges in $W_2^\nu(\mathbb{R})$ if and only if $\nu_2(\hat{a}) > \nu$. Moreover, $\phi\in W_2^\nu(\mathbb{R})$ for all $0\leqslant\nu<\nu_2(\hat{a})$, where ϕ is the nontrivial compactly supported refinable function associated with mask $\hat{a}$ such that $\hat{\phi}(2\xi) = \hat{a}(\xi)\hat{\phi}(\xi)$. Moreover, $\phi\notin W_2^\nu(\mathbb{R})$ for $\nu>\nu_2(\hat{a})$ whenever the integer shifts of ϕ are stable. See [17, 20, 22, 24, 32, 37] and the many references therein on the convergence of stationary cascade algorithms.

2.3. Convergence of cascade algorithms with pseudospline masks. We show that the nonstationary cascade algorithm associated with masks for pseudosplines in Theorem 1.2 converges in $W_2^\nu(\mathbb{R})$ for arbitrary $\nu\geqslant 0$. We first prove the following lemma which is not only used in the proof of the convergence of the nonstationary cascade algorithms associated with pseudospline refinement masks, but also plays an important role in our proof of Theorem 1.2 and the spectral frame approximation order.

LEMMA 2.7. *Let $\widehat{a_{m,l}}$ be the refinement mask for the pseudospline of type* II *with order (m,l) in* (1.19). *For positive integers m_1, m_2, and l_2 such that $1\leqslant m_1\leqslant m_2$*

and $1 \leqslant l_2 \leqslant m_2$*, the following inequality holds:*

$$|\widehat{a_{m_2,l_2}}(\xi)| \leqslant |\widehat{a^I_{m_2,l_2}}(\xi)| \leqslant |\widehat{b_{m_1}}(\xi)| \qquad \forall\, \xi \in \mathbb{R} \quad \textit{with}$$
$$\widehat{b_{m_1}}(\xi) := 2(1+e^{-i\xi})^{-1}\widehat{a^I_{m_1,m_1}}(\xi). \tag{2.23}$$

Note that $\widehat{b_{m_1}}$ *is uniquely determined by* $\widehat{a^I_{m_1,m_1}}(\xi) = 2^{-1}(1+e^{-i\xi})\widehat{b_{m_1}}(\xi)$.

Proof. Since $|\widehat{a_{m,l}}(\xi)| \leqslant 1$ for all $\xi \in \mathbb{R}$ and $1 \leqslant l \leqslant m$, it follows from the relation $|\widehat{a^I_{m,l}}(\xi)|^2 = \widehat{a_{m,l}}(\xi)$ and (1.18) that $|\widehat{a_{m,l}}(\xi)| \leqslant |\widehat{a^I_{m,l}}(\xi)| \leqslant |\widehat{a^I_{m,m}}(\xi)|$ for all $\xi \in \mathbb{R}$ and $1 \leqslant l \leqslant m$. Now in order to prove (2.23), it suffices to prove that $\widehat{a_{m_2,m_2}}(\xi) = |\widehat{a^I_{m_2,m_2}}(\xi)|^2 \leqslant |\widehat{b_{m_1}}(\xi)|^2$ for all $\xi \in \mathbb{R}$ and $1 \leqslant m_1 \leqslant m_2$.

Setting $x = \sin^2(\xi/2)$, by the definition of $\widehat{a_{m_2,m_2}}$ in (1.19), we see that $\widehat{a_{m_2,m_2}}(\xi) \leqslant |\widehat{b_{m_1}}(\xi)|^2$ for all $\xi \in \mathbb{R}$ and $1 \leqslant m_1 \leqslant m_2$ is equivalent to

$$(1-x)^{m_2}P_{m_2,m_2}(x) \leqslant (1-x)^{m_1-1}P_{m_1,m_1}(x) \qquad \forall\, x \in [0,1] \quad \text{and} \quad 1 \leqslant m_1 \leqslant m_2. \tag{2.24}$$

By the definition of $P_{m,l}$ in (1.18), we deduce that

$$\begin{aligned}
(1-x)^{m+1}P_{m+1,m+1}(x) &= (1-x)^{m+1}\sum_{j=0}^{m}\frac{(m+j)!}{j!m!}x^j \\
&= (1-x)^m\sum_{j=0}^{m}\frac{(m+j)!}{j!m!}(x^j - x^{j+1}) \\
&= (1-x)^m\left[\sum_{j=0}^{m}\frac{(m+j)!}{j!m!}x^j - \sum_{j=1}^{m+1}\frac{(m+j-1)!}{(j-1)!m!}x^j\right] \\
&= (1-x)^m\left[1 - \frac{(2m)!}{m!m!}x^{m+1} + \sum_{j=1}^{m}\left(\frac{(m+j)!}{j!m!} - \frac{(m+j-1)!}{(j-1)!m!}\right)x^j\right] \\
&= (1-x)^m\left[1 - \frac{(2m)!}{m!m!}x^{m+1} + \sum_{j=1}^{m}\frac{(m+j-1)!}{j!(m-1)!}x^j\right].
\end{aligned}$$

That is, we have

$$\begin{aligned}
&(1-x)^{m+1}P_{m+1,m+1}(x) \\
&\quad= (1-x)^m\left[\frac{(2m-1)!}{m!(m-1)!}x^m - \frac{(2m)!}{m!m!}x^{m+1} + \sum_{j=0}^{m-1}\frac{(m+j-1)!}{j!(m-1)!}x^j\right] \\
&\quad= (1-x)^m\left[\frac{(2m-1)!}{m!(m-1)!}x^m(1-2x) + P_{m,m}(x)\right] \\
&\quad= (1-x)^mP_{m,m}(x) - \frac{(2m-1)!}{m!(m-1)!}x^m(1-x)^m(2x-1) \qquad \forall\, m \in \mathbb{N}.
\end{aligned}$$

Consequently, for $x \in [1/2, 1]$ and $m_2 \geqslant m_1$, we have $2x - 1 \geqslant 0$ and

$$(1-x)^{m_2} P_{m_2,m_2}(x) \leqslant (1-x)^{m_2-1} P_{m_2-1,m_2-1}(x) \leqslant \cdots \leqslant (1-x)^{m_1} P_{m_1,m_1}(x)$$
$$\leqslant (1-x)^{m_1-1} P_{m_1,m_1}(x).$$

Therefore, (2.24) holds for $x \in [1/2, 1]$.

It remains to prove (2.24) for all $x \in [0, 1/2]$. Since

$$(1-x)^{m_2} P_{m_2,m_2}(x) \leqslant (1-x)^{m_2} P_{m_2,m_2}(x) + x^{m_2} P_{m_2,m_2}(1-x) = 1,$$

in order to prove (2.24) for all $x \in [0, 1/2]$, now it suffices to show that

$$1 \leqslant (1-x)^{m_1-1} P_{m_1,m_1}(x) \qquad \forall\, x \in [0, 1/2]. \tag{2.25}$$

Note that

$$1 = (1-x)(1-x)^{m_1-1} P_{m_1,m_1}(x) + x x^{m_1-1} P_{m_1,m_1}(1-x)$$
$$= (1-x)^{m_1-1} P_{m_1,m_1}(x) - x\big[(1-x)^{m_1-1} P_{m_1,m_1}(x) - x^{m_1-1} P_{m_1,m_1}(1-x)\big],$$

from which we have

$$(1-x)^{m_1-1} P_{m_1,m_1}(x) = 1 + x\big[(1-x)^{m_1-1} P_{m_1,m_1}(x) - x^{m_1-1} P_{m_1,m_1}(1-x)\big].$$

In order to prove (2.25), by the above identity, it suffices to prove that

$$(1-x)^{m_1-1} P_{m_1,m_1}(x) \geqslant x^{m_1-1} P_{m_1,m_1}(1-x) \qquad \forall\, x \in [0, 1/2]. \tag{2.26}$$

Note that for $x \in [0, 1/2]$, we have $0 \leqslant x/(1-x) \leqslant 1$. By the definition of $P_{m_1,m_1}(x) := \sum_{j=0}^{m_1-1} \binom{m_1+j-1}{j} x^j$, we have

$$1 = \frac{1}{1} \geqslant \frac{\binom{m_1}{1} x}{\binom{m_1}{1}(1-x)} \geqslant \frac{\binom{m_1+1}{2} x^2}{\binom{m_1+1}{2}(1-x)^2} \geqslant \cdots \geqslant \frac{\binom{2m_1-2}{m_1-1} x^{m_1-1}}{\binom{2m_1-2}{m_1-1}(1-x)^{m_1-1}}$$
$$= \frac{x^{m_1-1}}{(1-x)^{m_1-1}}, \quad x \in [0, 1/2].$$

But for positive numbers a, b, c, d, it is easy to see that $\frac{a}{b} \geqslant \frac{c}{d}$ implies $\frac{a}{b} \geqslant \frac{a+c}{b+d} \geqslant \frac{c}{d}$. Now it follows from the above inequalities that

$$1 \geqslant \frac{P_{m_1,m_1}(x)}{P_{m_1,m_1}(1-x)} = \frac{\sum_{j=0}^{m_1-1} \binom{m_1+j-1}{j} x^j}{\sum_{j=0}^{m_1-1} \binom{m_1+j-1}{j}(1-x)^j} \geqslant \frac{x^{m_1-1}}{(1-x)^{m_1-1}}, \qquad x \in [0, 1/2],$$

from which we see that (2.26) holds. □

Next we establish the following result on C^∞ nonstationary refinable functions. In fact, we prove that the nonstationary cascade algorithm associated with the masks for pseudosplines of type I or type II in Theorem 1.2 converges in $W_2^\nu(\mathbb{R})$ for any $\nu \geqslant 0$.

THEOREM 2.8. *Let $\widehat{a_j}$ be the mask for the pseudospline of type* I *or type* II *with order (m_j, l_j), where $1 \leqslant l_j \leqslant m_j$ and $j \in \mathbb{N}$ are positive integers such that* (1.21) *holds. Then, for every $n \in \mathbb{N}_0$, the nonstationary cascade algorithm* (2.22)

associated with $\{\widehat{a_{j+n}}\}_{j=1}^{\infty}$ *converges in* $W_2^\nu(\mathbb{R})$ *for any* $\nu \geqslant 0$. *Consequently, the nonstationary refinable functions* $\phi_j, j \in \mathbb{N}_0$, *in* (1.1) *must be well-defined compactly supported* $C^\infty(\mathbb{R})$ *functions.*

Proof. Since $\widehat{a_j}(\xi) = \widehat{a_{m_j,l_j}}(\xi)$ or $\widehat{a_j}(\xi) = \widehat{a^I_{m_j,l_j}}(\xi)$, it is easy to see that $\deg(\widehat{a_j}) \leqslant 2m_j$ and $\widehat{a_j}(0) = 1$. Therefore, by our assumption in (1.21), we see that

$$\sum_{j=1}^{\infty} 2^{-j} \deg(\widehat{a_j}) \leqslant 2 \sum_{j=1}^{\infty} 2^{-j} m_j < \infty.$$

Moreover, we have $|\widehat{a_j}(\xi)| \leqslant 1$ for all $j \in \mathbb{N}$ and $\xi \in \mathbb{R}$. Thus, the condition of Lemma 2.1 is satisfied. By Lemma 2.1, we conclude that $\phi_j, j \in \mathbb{N}_0$, are well-defined compactly supported tempered distributions.

Let $\widehat{a_{j,j}}$ be the refinement mask for the pseudospline of type II with order (j, j). It was proved by Daubechies [12, 13] that $\lim_{j\to\infty} \nu_2(\widehat{a_{j,j}}) = \lim_{j\to\infty} \nu_2(\widehat{a^I_{j,j}}) = \infty$. Hence, there exists a positive integer J such that $\nu_2(\widehat{a^I_{J,J}}) \geqslant \nu + 2$. By $\lim_{j\to\infty} m_j = \infty$, there exists a positive integer N such that

$$m_j \geqslant J \quad \text{and} \quad \nu_2(\widehat{a_j}) \geqslant \nu_2(\widehat{a^I_{J,J}}) \geqslant \nu + 2 \qquad \forall\, j \geqslant N. \tag{2.27}$$

Let $\hat{b}$ be the unique 2π-periodic trigonometric polynomial such that $\widehat{a^I_{J,J}}(\xi) = 2^{-1}(1 + e^{-i\xi})\hat{b}(\xi)$. By the definition of $\nu_2(\hat{b})$ and (2.27), it is straightforward to see that $\nu_2(\hat{b}) = \nu_2(\widehat{a^I_{J,J}}) - 1 \geqslant \nu + 1 > \nu$. Therefore, the stationary cascade algorithm associated with the mask $\hat{b}$ converges in $W_2^\nu(\mathbb{R})$ (see [20, Theorem 4.3]). On the other hand, by (2.23) of Lemma 2.7, since $m_j \geqslant J$ for $j \geqslant N$, we deduce that

$$|\widehat{a_j}(\xi)| \leqslant |\widehat{a^I_{m_j,l_j}}(\xi)| \leqslant |\hat{b}(\xi)| \qquad \forall\, \xi \in \mathbb{R},\ j \geqslant N.$$

Since the stationary cascade algorithm associated with the mask $\hat{b}$ converges in $W_2^\nu(\mathbb{R})$, by Proposition 2.6, the nonstationary cascade algorithm associated with masks $\{\widehat{a_j}\}_{j=1}^{\infty}$ converges in $W_2^\nu(\mathbb{R})$. Therefore, we have $\phi_0 \in W_2^\nu(\mathbb{R})$ for all $\nu \geqslant 0$. That is, ϕ_0 is a compactly supported C^∞ function. The same proof works for every ϕ_n and for the nonstationary cascade algorithm associated with masks $\{\widehat{a_{n+j}}\}_{j=1}^{\infty}$. ∎

In computer aided geometric design, it is of interest to consider the convergence of a subdivision scheme and a cascade algorithm in $C^\kappa(\mathbb{R})$, the space of functions with the κth continuous derivative, instead of the Sobolev space $W_2^\nu(\mathbb{R})$. When a cascade algorithm is implemented in the space domain as given in (2.7), the initial function is often chosen to be a compactly supported function such as spline functions of order ν with $\nu > \kappa + 1/2$. This is indeed true and can be proved easily by the imbedding theorems of Sobolev spaces.

COROLLARY 2.9. *Let* $\widehat{a_j}$ $(j \in \mathbb{N})$ *be the masks satisfying the conditions given in Theorem* 2.8. *Let* κ *be any nonnegative integer, and let* f *be a compactly supported initial function in* $W_2^\nu \cap \mathcal{F}_\nu$ *with* $\nu > \kappa + 1/2$. *Then the nonstationary cascade algorithm defined by* (2.7) *associated with* $\{\widehat{a_{n+j}}\}_{j=1}^{\infty}$ *converges in* $C^\kappa(\mathbb{R})$.

Proof. The convergence of the cascade algorithm in $W_2^\nu(\mathbb{R})$ defined by (2.7) follows from Proposition 2.6 and Theorem 2.8. Since all f_n $(n \in \mathbb{N}_0)$ and ϕ_0 are supported inside some compact set, it follows from the imbedding theorem that the sequence f_n also converges in $C^\kappa(\mathbb{R})$. ∎

2.4. Orthogonality. For the stationary case, it is well known that for a compactly supported refinable function ϕ with a 2π-periodic trigonometric polynomial $\hat{a}$, the integer shifts of ϕ form an orthonormal system if and only if its refinement mask $\hat{a}$ satisfies

$$|\hat{a}(\xi)|^2 + |\hat{a}(\xi+\pi)|^2 = 1 \qquad \text{a.e. } \xi \in \mathbb{R}$$

and the corresponding stationary cascade algorithm converges in $L_2(\mathbb{R})$ (that is, $\nu_2(\hat{a}) > 0$; see [20, 22]). Furthermore, if one chooses the wavelet function ψ by $\hat{\psi}(2\xi) := e^{-i\xi}\overline{\hat{a}(\xi+\pi)}\hat{\phi}(\xi)$, then the wavelet system generated by ψ forms an orthonormal basis in $L_2(\mathbb{R})$. For example, see [11, 12, 13, 20, 22, 24, 31, 32, 37] and the references therein. It turns out that this is also true for the nonstationary case, as a consequence of the proof of Theorem 2.4.

THEOREM 2.10. *Let $\widehat{a_j}, j \in \mathbb{N}$, be 2π-periodic measurable functions such that, for every $j \in \mathbb{N}_0$, $\widehat{\phi_j}(\xi) := \lim_{N\to\infty}\prod_{n=1}^{N}\widehat{a_{n+j}}(2^{-n}\xi)$ is well defined for almost every $\xi \in \mathbb{R}$. Then the integer shifts of ϕ_j form an orthonormal system in $L_2(\mathbb{R})$ for all $j \in \mathbb{N}$; i.e.,*

$$\langle \phi_j(\cdot - k), \phi_j\rangle := \int_{\mathbb{R}} \phi_j(x-k)\overline{\phi_j(x)}\,dx = \delta(k) \qquad \forall\, k \in \mathbb{Z} \quad \text{and} \quad j \in \mathbb{N}_0, \tag{2.28}$$

where $\delta(0) = 1$ and $\delta(k) = 0$ for all $k \neq 0$ if and only if

(1) *all the masks $\widehat{a_j}$ satisfy*

$$|\widehat{a_j}(\xi)|^2 + |\widehat{a_j}(\xi+\pi)|^2 = 1 \qquad \text{a.e. } \xi \in \mathbb{R}, \quad \forall\, j \in \mathbb{N}_0; \tag{2.29}$$

(2) *the nonstationary cascade algorithm associated with masks $\{a_{n+j}\}_{n=1}^{\infty}$ converges in $L_2(\mathbb{R})$ for large enough $j \in \mathbb{N}_0$.*

Moreover, if (1.4) and (2.28) hold, define $\widehat{\psi_{j-1}}(\xi) := e^{-i\xi/2}\overline{\widehat{a_j}(\xi/2+\pi)}\widehat{\phi_j}(\xi/2)$ for $j \in \mathbb{N}$; then $\{\phi_0(\cdot - k) : k \in \mathbb{Z}\} \cup \{\psi_{j;j,k} : j \in \mathbb{N}_0, k \in \mathbb{Z}\}$ is an orthonormal basis of $L_2(\mathbb{R})$.

Proof. It is known that (2.28) holds for each j if and only if $[\widehat{\phi_j}, \widehat{\phi_j}](\xi) = 1$ almost everywhere $\xi \in \mathbb{R}$.

Assume that (1) and (2) hold. Then $[\widehat{\phi_j}, \widehat{\phi_j}](\xi) = 1$ almost everywhere $x \in \mathbb{R}$ for all $j \in \mathbb{N}_0$ can be proved by an argument similar to that in the stationary case (see [11, 13, 22, 24, 32]). We omit the details here.

The necessity part is proved as follows. If (2.28) holds, then $[\widehat{\phi_j}, \widehat{\phi_j}] = 1$ for all $j \in \mathbb{N}_0$. Now (2.29) can be verified by the same argument as in the stationary case. So, item (1) holds. Next, we prove item (2); that is, the sequence $\{f_N\}_{N=1}^{\infty}$, defined by

$$\widehat{f_N}(\xi) := \chi_{[-\pi,\pi]}(2^{-N}\xi)\prod_{n=1}^{N}\widehat{a_{n+j}}(2^{-n}\xi), \qquad N \in \mathbb{N},$$

converges in $L_2(\mathbb{R})$ for every $j \in \mathbb{N}_0$. By the relation $\widehat{\phi_j}(\xi) = \widehat{a_{j+1}}(\xi/2)\widehat{\phi_{j+1}}(\xi/2)$, we deduce that

$$\widehat{\phi_j}(\xi) = \widehat{\phi_{N+j}}(2^{-N}\xi)\prod_{n=1}^{N}\widehat{a_{n+j}}(2^{-n}\xi), \qquad N \in \mathbb{N} \quad \text{and} \quad j \in \mathbb{N}_0. \tag{2.30}$$

Applying (2.30) and $[\widehat{\phi_N}, \widehat{\phi_N}] = 1$, one obtains that

$$\|\widehat{\phi_j}\|^2_{L_2(\mathbb{R})} = \int_{\mathbb{R}} |\widehat{f_N}(\xi)|^2 \, d\xi, \qquad N \in \mathbb{N}.$$

In particular, we have $\lim_{N\to\infty} \int_{\mathbb{R}} |\widehat{f_N}(\xi)|^2 \, d\xi = \|\widehat{\phi_j}\|^2_{L_2(\mathbb{R})} < \infty$. Note that

$$|\widehat{f_N}(\xi) - \widehat{\phi_j}(\xi)|^2 \leqslant 2\Big[|\widehat{f_N}(\xi)|^2 + |\widehat{\phi_j}(\xi)|^2\Big] =: \eta_N(\xi), \qquad \xi \in \mathbb{R},$$

and

$$\lim_{N\to\infty} \int_{\mathbb{R}} \eta_N(\xi) \, d\xi = 4 \int_{\mathbb{R}} |\widehat{\phi_j}(\xi)|^2 \, d\xi < \infty.$$

By $\lim_{N\to\infty} \widehat{f_N}(\xi) = \widehat{\phi_j}(\xi)$ for almost every $\xi \in \mathbb{R}$ and the generalized Lebesgue dominated convergence theorem, we conclude that $\lim_{N\to\infty} \|\widehat{f_N} - \widehat{\phi_j}\|_{L_2(\mathbb{R})} = 0$. Thus, for every $j \in \mathbb{N}_0$, the cascade algorithm associated with masks $\{\widehat{a_{n+j}}\}_{n=1}^{\infty}$ converges in $L_2(\mathbb{R})$.

If (2.28) holds, by the definition of ψ_j and (2.29), then it is easy to check that $\{\phi_0(\cdot - k) \, : \, k \in \mathbb{Z}\} \cup \{\psi_{j;j,k} \, : \, j \in \mathbb{N}_0, k \in \mathbb{Z}\}$ is an orthonormal system of $L_2(\mathbb{R})$. By Theorem 1.1, we see that $\{\phi_0(\cdot - k) \, : \, k \in \mathbb{Z}\} \cup \{\psi_{j;j,k} \, : \, j \in \mathbb{N}_0, k \in \mathbb{Z}\}$ is a tight frame in $L_2(\mathbb{R})$. Therefore, $\{\phi_0(\cdot - k) \, : \, k \in \mathbb{Z}\} \cup \{\psi_{j;j,k} \, : \, j \in \mathbb{N}_0, k \in \mathbb{Z}\}$ is an orthonormal basis of $L_2(\mathbb{R})$. □

3. The approximation order of the truncated frame series. In this section, we shall study the approximation property of a nonstationary tight wavelet frame, i.e., the frame approximation properties of the operators Q_n in (1.9). The approximation operators Q_n and their approximation order for a given stationary tight wavelet frame have been extensively studied in [14]. The approximation operators Q_n provide a simple approximation scheme for a given tight wavelet frame and have close links to the frame decomposition and reconstruction algorithms for a tight wavelet frame (see, e.g., [14]). Moreover, their approximation order determines the accuracy of the truncation operators and is not necessarily equal to the best approximation order provided by the underlying nonstationary multiresolution analysis.

Since the approximation operators Q_n provide a simple approximation scheme for a given tight wavelet frame, they are often used in various applications. For example, in [1, 2, 3, 4, 5], where the (stationary) tight wavelet frame based algorithms for high/super resolution image reconstructions, image inpainting, and deconvolutions are given, the operators Q_n are used there to approximate the underlying function from a given data set. The interested reader should consult [1, 2, 3, 4, 5] for details.

The operators Q_n are closely related to other operators. For a sequence $\{\phi_n\}_{n=0}^{\infty}$ of functions in $L_2(\mathbb{R})$, we define the linear operators $P_n(f), n \in \mathbb{N}_0$, by

$$P_n(f) := \sum_{k\in\mathbb{Z}} \langle f, \phi_{n;n,k}\rangle \phi_{n;n,k}, \qquad f \in L_2(\mathbb{R}) \quad \text{with} \tag{3.1}$$

$$\phi_{n;n,k} := 2^{n/2} \phi_n(2^n \cdot - k).$$

Similar to the stationary case, by calculation, it is easy to verify that (1.12) implies

$$\begin{aligned}(3.2)\qquad &\sum_{\ell=1}^{\mathcal{J}_j}\sum_{k\in\mathbb{Z}}\langle f,\psi^\ell_{j-1;j-1,k}\rangle\psi^\ell_{j-1;j-1,k}\\ &\qquad=\sum_{k\in\mathbb{Z}}\langle f,\phi_{j;j,k}\rangle\phi_{j;j,k}-\sum_{k\in\mathbb{Z}}\langle f,\phi_{j-1;j-1,k}\rangle\phi_{j-1;j-1,k}\end{aligned}$$

for $f\in L_2(\mathbb{R})$. Consequently, by the definition of the linear operators Q_n in (1.9), it follows from the relation in (3.2) that

$$\begin{aligned}Q_n(f)&=\sum_{k\in\mathbb{Z}}\langle f,\phi_0(\cdot-k)\rangle\phi_0(\cdot-k)+\sum_{j=0}^{n-1}\sum_{\ell=1}^{\mathcal{J}_{j+1}}\sum_{k\in\mathbb{Z}}\langle f,\psi^\ell_{j;j,k}\rangle\psi^\ell_{j;j,k}\\ &=\sum_{k\in\mathbb{Z}}\langle f,\phi_{n;n,k}\rangle\phi_{n;n,k}=P_n(f).\end{aligned}$$

That is, if (1.12) holds, then the linear operators Q_n in (1.9) and P_n in (3.1) are the same.

The approximation order of Q_n's for a stationary tight wavelet frame is investigated in [14] through that of P_n's, since $Q_n=P_n$ for a tight frame system constructed from the unitary extension principle. The relationship has been studied in [14] for stationary tight wavelet frames between the approximation order of P_n's and the (best) approximation order provided by the spaces $S_n(\phi_n)$, where $S_n(\phi_n)$ is the smallest closed subspace of $L_2(\mathbb{R})$ generated by the linear span of $\phi_n(2^n\cdot-k)$, $k\in\mathbb{Z}$; that is, $S_n(\phi_n)$ is the same as the smallest closed subspace of $L_2(\mathbb{R})$ containing the truncated tight frame system $\{\phi_0(\cdot-k)\,:\,k\in\mathbb{Z}\}\cup\{\psi^\ell_{j;j,k}\,:\,k\in\mathbb{Z},0\leqslant j<n,\ell=1,\dots,\mathcal{J}_{j+1}\}$. It is well known that the approximation order provided by the spaces $S_n(\phi_n)$ is determined by the order of the Strang–Fix conditions satisfied by ϕ_n. However, the approximation order of P_n's is determined by the order of the zero at the origin of the function $1-|\widehat{\phi}|^2$, in addition to the order of the Strang–Fix conditions satisfied by ϕ_n. Consequently, the frame approximation order can be (much) smaller than the approximation order provided by the spaces $S_n(\phi_n)$ (see [14] for details). This is also true for the nonstationary case. For example, let $\widehat{a_j}(\xi):=2^{-j}(1+e^{-i\xi})^j$, $j\in\mathbb{N}$, be the masks for the up-functions. Then, item (iii) of Theorem 1.3 says that it does not have any "strong" frame approximation order in the sense of (1.16); i.e., for any given $\nu>0$, there does not exist a positive constant C such that (1.16) is satisfied. But one can check that the corresponding spaces $S_n(\phi_n)$ provide a spectral approximation order.

If the integer shifts of ϕ_n are orthonormal, then the linear operator P_n in (3.1) becomes an orthogonal projection from $L_2(\mathbb{R})$ to $S_n(\phi_n)$. That is, for this case, $P_n(f)$ is the best approximation of $f\in L_2(\mathbb{R})$ in the closed subspace $S_n(\phi_n)$ of $L_2(\mathbb{R})$. This is the reason why the approximation order of Q_n's is identified with the best approximation order provided by the spaces $S_n(\phi_n)$ in [10], which is simpler to understand, since only orthonormal wavelets are studied in [10].

To summarize our discussion here, the understanding of the approximation order of the approximation operators Q_n for a given tight wavelet frame is necessary, since it is simple and used in applications such as image inpainting, and since, unlike orthonormal wavelets, the approximation order of a truncated tight frame series is

not necessarily the same as the best approximation order provided by the underlying nonstationary multiresolution analysis.

The main result of this section is Theorem 3.2, which is interesting in its own right and is independent of its role in our proofs of some of major parts of Theorems 1.1–1.4.

The following result can be directly obtained by applying Jetter and Zhou [28] and [29, Theorem 2.1].

PROPOSITION 3.1. *Let $\varphi \in L_2(\mathbb{R})$ and $\nu \geqslant 0$. Define a linear operator P by*

$$P(f) := \sum_{k\in\mathbb{Z}} \langle f, \varphi(\cdot - k)\rangle \varphi(\cdot - k), \qquad f \in L_2(\mathbb{R}).$$

Then $\|f - P(f)\|_{L_2(\mathbb{R})} \leqslant C_\varphi |f|_{W_2^\nu(\mathbb{R})}$ for all $f \in W_2^\nu(\mathbb{R})$ with a positive constant

$$C_\varphi := \pi^{-1/2} \sqrt{\max(c_1, c_3) + \max(2c_2, 2c_4 + 1)}, \tag{3.3}$$

provided that there exist positive constants c_1, c_2, c_3, c_4 such that, for almost every $\xi \in [-\pi, \pi]$, the following inequalities hold:

$$\left|1 - |\hat{\varphi}(\xi)|^2\right|^2 \leqslant c_1 |\xi|^{2\nu}, \tag{3.4}$$

$$\sum_{k\in\mathbb{Z}\setminus\{0\}} |\hat{\varphi}(\xi)|^2 |\hat{\varphi}(\xi + 2\pi k)|^2 \leqslant c_2 |\xi|^{2\nu}, \tag{3.5}$$

$$\sum_{k\in\mathbb{Z}\setminus\{0\}} |\xi + 2\pi k|^{-2\nu} |\hat{\varphi}(\xi)|^2 |\hat{\varphi}(\xi + 2\pi k)|^2 \leqslant c_3, \tag{3.6}$$

$$\sum_{k\in\mathbb{Z}\setminus\{0\}} |\xi + 2\pi k|^{-2\nu} \sum_{\ell\in\mathbb{Z}\setminus\{0\}} |\hat{\varphi}(\xi + 2\pi\ell)|^2 |\hat{\varphi}(\xi + 2\pi k)|^2 \leqslant c_4. \tag{3.7}$$

Next, we present the following result on the approximation properties of the operators P_n defined in (3.1).

THEOREM 3.2. *Let $\widehat{a_j}, j \in \mathbb{N}$, be 2π-periodic measurable functions such that* (1.3) *holds for all $j \in \mathbb{N}$, and, for every $n \in \mathbb{N}_0$, the function $\widehat{\phi_n}(\xi) := \lim_{J\to\infty} \prod_{j=1}^{J} \widehat{a_{j+n}}(2^{-j}\xi)$ is well defined for almost every $\xi \in \mathbb{R}$. Let $\nu \geqslant 0$. If, for $n \in \mathbb{N}$,*

$$\begin{aligned} &\left|1 - |\widehat{\phi_n}(\xi)|^2\right|^2 \leqslant C_{\phi_n} |\xi|^{2\nu} \qquad \text{a.e. } \xi \in [-\pi, \pi], \\ &\sum_{k\in\mathbb{Z}\setminus\{0\}} |\widehat{\phi_n}(\xi)|^2 |\widehat{\phi_n}(\xi + 2\pi k)|^2 \leqslant C_{\phi_n} |\xi|^{2\nu} \qquad \text{a.e. } \xi \in [-\pi, \pi], \end{aligned} \tag{3.8}$$

where C_{ϕ_n} is a constant depending only on ϕ_n, then, for the linear operators P_n in (3.1),

$$\|f - P_n(f)\|_{L_2(\mathbb{R})} \leqslant \max(2, \sqrt{C_{\phi_n}})\, 2^{-\nu n} |f|_{W_2^\nu(\mathbb{R})} \quad \forall\, f \in W_2^\nu(\mathbb{R}) \quad \text{and} \quad n \in \mathbb{N}. \tag{3.9}$$

In particular, (3.8) *is satisfied if*

$$1 - |\widehat{\phi_n}(\xi)|^2 \leqslant C_{\phi_n} |\xi|^{2\nu} \qquad \text{a.e. } \xi \in [-\pi, \pi]. \tag{3.10}$$

Proof. By (1.3) and Lemma 2.2, we have $\phi_n \in L_2(\mathbb{R})$ for all $n \in \mathbb{N}_0$. For each fixed $n \in \mathbb{N}_0$, we denote $P_{n,0}$ the following linear operator on $L_2(\mathbb{R})$:

$$P_{n,0}(f) := \sum_{k\in\mathbb{Z}} \langle f, \phi_n(\cdot - k)\rangle \phi_n(\cdot - k), \qquad f \in L_2(\mathbb{R}).$$

It is apparent [28] that the operators P_n and $P_{n,0}$ are linked through the relation $P_n(f) = [P_{n,0}(f(2^{-n}\cdot))](2^n\cdot)$ and

$$\|f - P_n(f)\|_{L_2(\mathbb{R})} = 2^{-n/2}\|f(2^{-n}\cdot) - P_{n,0}(f(2^{-n}\cdot))\|_{L_2(\mathbb{R})}. \tag{3.11}$$

Since (1.3) holds, by Lemma 2.2, we have

$$[\widehat{\phi_n}, \widehat{\phi_n}](\xi) := \sum_{k\in\mathbb{Z}} |\widehat{\phi_n}(\xi + 2\pi k)|^2 \leqslant 1 \quad \text{a.e. } \xi \in \mathbb{R}.$$

In particular, we have $|\widehat{\phi_n}(\xi)| \leqslant 1$ for almost every $\xi \in \mathbb{R}$.

Since $\nu \geqslant 0$, for $\xi \in [-\pi, \pi]$, it is evident that $|\xi + 2\pi k|^{-2\nu} \leqslant 1$ for all $k \in \mathbb{Z}\backslash\{0\}$ and

$$\sum_{k\in\mathbb{Z}\backslash\{0\}} |\xi + 2\pi k|^{-2\nu} \sum_{\ell\in\mathbb{Z}\backslash\{0\}} |\widehat{\phi_n}(\xi + 2\pi\ell)|^2 |\widehat{\phi_n}(\xi + 2\pi k)|^2$$

$$\leqslant \sum_{k\in\mathbb{Z}} |\widehat{\phi_n}(\xi + 2\pi k)|^2 \sum_{\ell\in\mathbb{Z}} |\widehat{\phi_n}(\xi + 2\pi\ell)|^2 \leqslant 1.$$

Hence, (3.7) holds with $c_4 = 1$ and $\varphi = \phi_n$. Similarly, for $\xi \in [-\pi, \pi]$, we have

$$\sum_{k\in\mathbb{Z}\backslash\{0\}} |\xi + 2\pi k|^{-2\nu} |\widehat{\phi_n}(\xi)|^2 |\widehat{\phi_n}(\xi + 2\pi k)|^2 \leqslant \sum_{k\in\mathbb{Z}} |\widehat{\phi_n}(\xi + 2\pi k)|^2 \leqslant 1.$$

Therefore, (3.6) holds with $c_3 = 1$ and $\varphi = \phi_n$.

By (3.8), we see that (3.4) and (3.5) hold with $c_1 = c_2 = C_{\phi_n}$ and $\varphi = \phi_n$.

Note that $|f(2^{-n}\cdot)|_{W_2^\nu(\mathbb{R})} = 2^{n/2}2^{-\nu n}|f|_{W_2^\nu(\mathbb{R})}$. Therefore, by Proposition 3.1 and (3.11), we conclude that, for all $f \in W_2^\nu(\mathbb{R})$,

$$\begin{aligned}\|f - P_n(f)\|_{L_2(\mathbb{R})} &= 2^{-n/2}\|f(2^{-n}\cdot) - P_{n,0}(f(2^{-n}\cdot))\|_{L_2(\mathbb{R})} \\ &\leqslant \max(2, \sqrt{C_{\phi_n}})2^{-\nu n}|f|_{W_2^\nu(\mathbb{R})},\end{aligned} \tag{3.12}$$

since

$$\max(c_1, c_3) + \max(2c_2, 2c_4 + 1) = \max(C_{\phi_n}, 1) + \max(2C_{\phi_n}, 3) \leqslant \max(3C_{\phi_n}, 6)$$

$$\leqslant \pi \max(C_{\phi_n}, 4).$$

Therefore, (3.9) is verified.

If (3.10) holds, then

$$\big|1 - |\widehat{\phi_n}(\xi)|^2\big|^2 \leqslant C_{\phi_n}^2 |\xi|^{2\nu}|\xi|^{2\nu} \leqslant C_{\phi_n}|\xi|^{2\nu}, \qquad |\xi| \leqslant C_{\phi_n}^{-\frac{1}{2\nu}},$$

and by $|\widehat{\phi_n}(\xi)| \leqslant 1$,

$$\big|1 - |\widehat{\phi_n}(\xi)|^2\big|^2 \leqslant 1 = C_{\phi_n}[C_{\phi_n}^{-\frac{1}{2\nu}}]^{2\nu} \leqslant C_{\phi_n}|\xi|^{2\nu}, \qquad |\xi| \geqslant C_{\phi_n}^{-\frac{1}{2\nu}}.$$

Also, by $|\widehat{\phi_n}(\xi)|^2 \leqslant [\widehat{\phi_n}, \widehat{\phi_n}](\xi) \leqslant 1$ and the above two inequalities, we have

$$\sum_{k\in\mathbb{Z}\backslash\{0\}} |\widehat{\phi_n}(\xi)|^2 |\widehat{\phi_n}(\xi + 2\pi k)|^2 \leqslant \sum_{k\in\mathbb{Z}\backslash\{0\}} |\widehat{\phi_n}(\xi + 2\pi k)|^2 = [\widehat{\phi_n}, \widehat{\phi_n}](\xi) - |\widehat{\phi_n}(\xi)|^2$$

$$\leqslant 1 - |\widehat{\phi_n}(\xi)|^2.$$

Consequently, (3.10) implies (3.8). □

The behavior of $1-|\widehat{\phi_n}(\xi)|^2$ near the origin $\xi=0$ in (3.10) is closely related to that of the masks $1-|\widehat{a_j}(\xi)|^2$ for all $j\in\mathbb{N}$ near the origin $\xi=0$. As we will see in the proof of the next section, when only masks are available and the nonstationary refinable functions are not explicitly given, one can use the estimate of $1-|\widehat{a_j}(\xi)|^2$ near $\xi=0$ for all $j\in\mathbb{N}$ to obtain the estimate of $1-|\widehat{\phi_n}(\xi)|^2$ near $\xi=0$. The following result provides the estimate of $1-|\widehat{a_j}(\xi)|^2$ near the origin for the masks of pseudosplines, which will be needed later in our proof of the spectral frame approximation order in Theorems 1.2 and 1.4.

LEMMA 3.3. *Let $\widehat{a^I_{m,l}}$ and $\widehat{a_{m,l}}$ be pseudospline masks of types* I *and* II *of order (m,l) defined in* (1.19) *and* (1.20)*, respectively. For any $0<\rho\leqslant 1$ and $\nu\geqslant 0$, there exist a positive integer N and a positive constant C (both of them depend only on ρ and ν), such that, for all $N\leqslant \rho m<l\leqslant m$,*

$$0\leqslant 1-|\widehat{a^I_{m,l}}(\xi)|^2\leqslant 1-|\widehat{a_{m,l}}(\xi)|^2\leqslant C|\xi|^{2\nu}\qquad \forall\,\xi\in[-\pi,\pi]. \tag{3.13}$$

Proof. Since $|\widehat{a^I_{m,l}}(\xi)|^2=\widehat{a_{m,l}}(\xi)\leqslant 1$, we have

$$0\leqslant 1-|\widehat{a^I_{m,l}}(\xi)|^2\leqslant 1-|\widehat{a_{m,l}}(\xi)|^2=[1+\widehat{a_{m,l}}(\xi)][1-\widehat{a_{m,l}}(\xi)]\leqslant 2[1-\widehat{a_{m,l}}(\xi)].$$

Setting $x=\sin^2(\xi/2)$, by the definition of $\widehat{a_{m,l}}(\xi)$ in (1.19), we have $\widehat{a_{m,l}}(\xi)=(1-x)^mP_{m,l}(x)$, where the polynomial $P_{m,l}$ is defined in (1.18). In order to prove (3.13), now it is easy to see that it is equivalent to proving that, for any $0<\rho\leqslant 1$ and any positive integer ν, there exist a positive integer N and a positive constant C, all depending only on ρ and ν, such that

$$1-(1-x)^mP_{m,l}(x)\leqslant Cx^\nu\qquad \forall\,x\in[0,1]\quad\text{and}\quad N\leqslant\rho m<l\leqslant m. \tag{3.14}$$

Since $(1-x)^mP_{m,m}(x)+x^mP_{m,m}(1-x)=1$, we deduce that

$$\begin{aligned}1-(1-x)^mP_{m,l}(x)&=(1-x)^mP_{m,m}(x)+x^mP_{m,m}(1-x)-(1-x)^mP_{m,l}(x)\\&=x^mP_{m,m}(1-x)+(1-x)^m\sum_{j=l}^{m-1}\frac{(m+j-1)!}{j!(m-1)!}x^j.\\&=x^\nu\left[x^{m-\nu}P_{m,m}(1-x)+(1-x)^m\sum_{j=l}^{m-1}\frac{(m+j-1)!}{j!(m-1)!}x^{j-\nu}\right].\end{aligned}$$

By Lemma 2.7, (2.24) holds. In particular, replacing x by $1-x$ in (2.24), we conclude that

$$x^{m-\nu}P_{m,m}(1-x)\leqslant x^{N-\nu-1}P_{N,N}(1-x)\qquad\forall\,x\in[0,1]\quad\text{and}\quad m\geqslant N\geqslant\nu+1.$$

Therefore, on the one hand, we have

$$x^{m-\nu}P_{m,m}(1-x)\leqslant C_N:=\max_{x\in[0,1]}x^{N-\nu-1}P_{N,N}(1-x)<\infty\qquad\forall\,x\in[0,1],\ m\geqslant N\geqslant\nu+1. \tag{3.15}$$

On the other hand, for $x \in [0,1]$, we have

$$\begin{aligned}(1-x)^m \sum_{j=l}^{m-1} \frac{(m+j-1)!}{j!(m-1)!} x^{j-\nu} &= (1-x)^m x^{l-\nu} \sum_{j=l}^{m-1} \frac{(m+j-1)!}{j!(m-1)!} x^{j-l} \\ &\leqslant (1-x)^m x^{l-\nu} \sum_{j=l}^{m-1} \frac{(m+j-1)!}{j!(m-1)!} \\ &= (1-x)^m x^{l-\nu} \sum_{j=l}^{m-1} \left[\frac{(m+j)!}{j!m!} - \frac{(m+j-1)!}{(j-1)!m!} \right] \\ &\leqslant (1-x)^m x^{l-\nu} \frac{(2m-1)!}{(m-1)!m!}.\end{aligned}$$

By Stirling's formula, we have $\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi} n^{n+1/2} e^{-n}} = 1$. Thus, when m is large enough and for all $x \in [0,1]$, we have

$$\frac{(2m-1)!}{(m-1)!m!} = \frac{1}{2} \frac{(2m)!}{m!m!} \leqslant \frac{1}{\sqrt{2\pi}} \frac{(2m)^{2m+1/2} e^{-2m}}{m^{2m+1} e^{-2m}} \leqslant 4^m m^{-1/2}.$$

Thus, for large enough m, we have

$$(1-x)^m \sum_{j=l}^{m-1} \frac{(m+j-1)!}{j!(m-1)!} x^{j-\nu} \leqslant (1-x)^m x^{l-\nu} 4^m m^{-1/2} = [4(1-x)x^{(l-\nu)/m}]^m m^{-1/2}.$$

Since $\rho m < l$, we have $l/m > \rho$. Since ν is fixed, when m is large enough, we have $(l-\nu)/m \geqslant \rho$ and, hence, $x^{(l-\nu)/m} \leqslant x^\rho$ for all $x \in [0,1]$. Therefore, we have

$$(1-x)^m \sum_{j=l}^{m-1} \frac{(m+j-1)!}{j!(m-1)!} x^{j-\nu} \leqslant [4(1-x)x^\rho]^m m^{-1/2} \qquad \forall\, x \in [0,1].$$

Note that $\rho > 0$. The continuous function $(1-x)x^\rho$ has only one critical point $x = \frac{\rho}{1+\rho}$ on the interval $(0,1)$, and it takes value zero at $x = 0$. Thus, we can choose $0 < \tau \leqslant 1$ such that $4(1-x)x^\rho \leqslant 1$ for all $x \in [0,\tau]$; one may prefer to choose such τ as large as possible; in particular, τ may be obtained by solving $4(1-\tau)\tau^\rho = 1$. Therefore, there exists a positive integer N such that, for $N \leqslant \rho m < l \leqslant m$,

$$(1-x)^m \sum_{j=l}^{m-1} \frac{(m+j-1)!}{j!(m-1)!} x^{j-\nu} \leqslant [4(1-x)x^\rho]^m m^{-1/2} \leqslant m^{-1/2} \leqslant 1 \qquad \forall\, x \in [0,\tau]. \tag{3.16}$$

Combining (3.15) and (3.16), we conclude that

$$1 - (1-x)^m P_{m,l}(x) \leqslant x^\nu [C_N + 1] \qquad \forall\, x \in [0,\tau],\ N \leqslant \rho m < l \leqslant m.$$

Observing that $0 \leqslant (1-x)^m P_{m,l}(x) \leqslant 1$ for all $x \in [0,1]$, we deduce that

$$1 - (1-x)^m P_{m,l}(x) \leqslant 1 = x^{-\nu} x^\nu \leqslant \tau^{-\nu} x^\nu \qquad \forall\, x \in [\tau,1],\ 1 \leqslant l \leqslant m,\ m \in \mathbb{N}.$$

Thus, (3.14) is verified with $C := \max(C_N + 1, \tau^{-\nu})$. This completes the proof. ∎

4. Proofs of Theorems 1.1–1.4. In this section, we shall prove Theorems 1.1–1.4. We start with a proof of Theorem 1.1.

Proof of Theorem 1.1. By our assumption in Theorem 1.1, item (i) follows from Lemmas 2.1 and 2.2. Since (1.3) holds, by Lemma 2.2, we have $[\widehat{\phi_n}, \widehat{\phi_n}] \leqslant 1$ and $\phi_n \in L_2(\mathbb{R})$. Therefore, the linear operators Q_n in (1.9) and P_n in (3.1) are well defined, bounded, and the same (see section 3).

Let us first prove (1.15) in item (iii). In order to do so, in the following we estimate the constants C_{ϕ_n} in (3.10). Denote $\widehat{d_j}(\xi) := |\widehat{a_j}(\xi)|^2$ for $j \in \mathbb{N}$. Since $\widehat{a_j}(0) = 1$, we have $\widehat{d_j}(0) = 1$ for all $j \in \mathbb{N}$. By our assumption and Lemma 2.1, we have $|\widehat{\phi_n}(\xi)|^2 = \prod_{j=1}^{\infty} \widehat{d_{j+n}}(2^{-j}\xi)$ for all $\xi \in \mathbb{R}$. Therefore, we have

$$\begin{aligned} 1 - |\widehat{\phi_n}(\xi)|^2 &= \prod_{j=1}^{\infty} \widehat{d_{j+n}}(0) - \prod_{j=1}^{\infty} \widehat{d_{j+n}}(2^{-j}\xi) \\ &= \sum_{\ell=1}^{\infty} \Big[\prod_{j=1}^{\ell-1} \widehat{d_{j+n}}(0) \Big] [\widehat{d_{\ell+n}}(0) - \widehat{d_{\ell+n}}(2^{-\ell}\xi)] \Big[\prod_{j=\ell+1}^{\infty} \widehat{d_{j+n}}(2^{-j}\xi) \Big]. \end{aligned}$$

Since $\widehat{d_{j+n}}(0) = 1$ and $0 \leqslant \widehat{d_{j+n}}(\xi) \leqslant 1$ by (1.3), we conclude that

$$0 \leqslant 1 - |\widehat{\phi_n}(\xi)|^2 \leqslant \sum_{\ell=1}^{\infty} |\widehat{d_{\ell+n}}(0) - \widehat{d_{\ell+n}}(2^{-\ell}\xi)|, \qquad \xi \in \mathbb{R}. \tag{4.1}$$

Since $\widehat{a_j}$ is a 2π-periodic trigonometric polynomial, by the definition of $\widehat{d_j}$, we see that $\widehat{d_j}$ is a real-valued C^∞ function. Note that, by our assumption, 2ν is a positive integer. Therefore, for $\xi \in [-\pi, \pi]$, there exists $\zeta_{\xi,j} \in [-\pi, \pi]$ such that

$$\widehat{d_j}(\xi) = \widehat{d_j}(0) + \frac{\widehat{d_j}^{(1)}(0)}{1!}\xi + \cdots + \frac{\widehat{d_j}^{(2\nu-1)}(0)}{(2\nu-1)!}\xi^{2\nu-1} + \frac{\widehat{d_j}^{(2\nu)}(\zeta_{\xi,j})}{(2\nu)!}\xi^{2\nu}. \tag{4.2}$$

By Bernstein's inequality and $0 \leqslant \widehat{d_j}(\xi) \leqslant 1$, we have

$$\|\widehat{d_j}^{(2\nu)}\|_{L_\infty(\mathbb{R})} \leqslant [\deg(\widehat{d_j})]^{2\nu} \|\widehat{d_j}\|_{L_\infty(\mathbb{R})} \leqslant [\deg(\widehat{d_j})]^{2\nu}.$$

By assumption in (1.14), for $j \geqslant N$, we have $\widehat{d_j}^{(\ell)}(0) = 0$ for all $\ell = 1, \ldots, 2\nu - 1$. Therefore, it follows from (4.2) that, for $n \geqslant N$, $\ell \in \mathbb{N}$, and $\xi \in [-\pi, \pi]$,

$$|\widehat{d_{\ell+n}}(0) - \widehat{d_{\ell+n}}(2^{-\ell}\xi)| \leqslant \frac{[\deg(\widehat{d_{\ell+n}})]^{2\nu}}{(2\nu)!} |2^{-\ell}\xi|^{2\nu} \leqslant \frac{1}{(2\nu)!} |\xi|^{2\nu} 2^{-2\nu\ell} [\deg(\widehat{d_{\ell+n}})]^{2\nu}.$$

Therefore, we have

$$\sum_{\ell=1}^{\infty} |\widehat{d_{\ell+n}}(0) - \widehat{d_{\ell+n}}(2^{-\ell}\xi)| \leqslant \frac{1}{(2\nu)!} |\xi|^{2\nu} \sum_{\ell=1}^{\infty} 2^{-2\nu\ell} [\deg(\widehat{d_{\ell+n}})]^{2\nu}.$$

That is, by (4.1) we see that, for every $n \geqslant N$, (3.10) holds with

$$C_{\phi_n} := \frac{1}{(2\nu)!} \sum_{\ell=1}^{\infty} 2^{-2\nu\ell} [\deg(\widehat{d_{\ell+n}})]^{2\nu}. \tag{4.3}$$

Now we estimate C_{ϕ_n} using the condition in (1.13). By (1.13), there exists a positive constant C_1 such that

$$\deg(\widehat{a_j}) \leqslant C_1 j^\alpha 2^{\beta j} \qquad \forall\, j \in \mathbb{N}. \tag{4.4}$$

By the definition of $\widehat{d_j}$, we have $\deg(\widehat{d_j}) \leqslant 2\deg(\widehat{a_j})$ for all $j \in \mathbb{N}$. Therefore, from (4.4), we deduce that

$$\begin{aligned} 2^{-2\nu\ell}[\deg(\widehat{d_{\ell+n}})]^{2\nu} &\leqslant 2^{2\nu} C_1^{2\nu} (\ell+n)^{2\nu\alpha} 2^{2\nu\beta(\ell+n)} 2^{-2\nu\ell} \\ &= 2^{2\nu} C_1^{2\nu} n^{2\nu\alpha} 2^{2\nu\beta n} (1+\ell/n)^{2\nu\alpha} 2^{-2\nu(1-\beta)\ell} \\ &\leqslant 2^{2\nu} C_1^{2\nu} n^{2\nu\alpha} 2^{2\nu\beta n} (1+\ell)^{2\nu\alpha} 2^{-2\nu(1-\beta)\ell}. \end{aligned}$$

Consequently, we have the following estimate for the constant C_{ϕ_n}:

$$\begin{aligned} C_{\phi_n} &= \frac{1}{(2\nu)!} \sum_{\ell=1}^{\infty} 2^{-2\nu\ell} [\deg(\widehat{d_{\ell+n}})]^{2\nu} \leqslant \frac{2^{2\nu} C_1^{2\nu}}{(2\nu)!} n^{2\nu\alpha} 2^{2\nu\beta n} \sum_{\ell=1}^{\infty} (1+\ell)^{2\nu\alpha} 2^{-2\nu(1-\beta)\ell} \\ &= C_2 n^{2\nu\alpha} 2^{2\nu\beta n}, \end{aligned}$$

where $C_2 := 2^{2\nu} C_1^{2\nu} [(2\nu)!]^{-1} \sum_{\ell=1}^{\infty} (1+\ell)^{2\nu\alpha} 2^{-2\nu(1-\beta)\ell} < \infty$, since $1-\beta > 0$ and $\nu > 0$. Since $Q_n = P_n$, by Theorem 3.2, we conclude that

$$\|f - Q_n(f)\|_{L_2(\mathbb{R})} \leqslant \max(2, \sqrt{C_2}) n^{\nu\alpha} 2^{-\nu(1-\beta)n} |f|_{W_2^\nu(\mathbb{R})} \quad \forall\, f \in W_2^\nu(\mathbb{R}) \quad \text{and} \quad n \geqslant N.$$

That is, (1.15) holds with $C := \max(2, \sqrt{C_2}) < \infty$, which is independent of f and n.

Now we prove item (ii). In order to show that $\{\phi_0(\cdot - k) : k \in \mathbb{Z}\} \cup \{\psi_{j;j,k}^\ell : j \in \mathbb{N}_0, k \in \mathbb{Z}, \ell = 1, \ldots, \mathcal{J}_{j+1}\}$ is a tight frame of $L_2(\mathbb{R})$, since $Q_n = P_n$, it now suffices to show that

$$\lim_{n\to\infty} \|f - Q_n(f)\|_{L_2(\mathbb{R})} = \lim_{n\to\infty} \|f - P_n(f)\|_{L_2(\mathbb{R})} = 0 \qquad \forall\, f \in L_2(\mathbb{R}).$$

Since $\widehat{a_j}(0) = 1$ and $\widehat{d_j}(\xi) = |\widehat{a_j}(\xi)|^2$, it is evident that $\widehat{d_j}(0) = 1$. Since $\widehat{d_j}$ is a 2π-periodic trigonometric polynomial, the condition in (1.14) is automatically satisfied with $\nu = 1/2$.

Now from the above proof of item (iii) and by Theorem 3.2, we see that (3.9) holds with the constant C_{ϕ_n} defined in (4.3) and $\nu = 1/2$. More precisely, by Theorem 3.2, for $\nu = 1/2$, we have

$$\|f - P_n(f)\|_{L_2(\mathbb{R})}^2 \leqslant C_n |f|_{W_2^{1/2}(\mathbb{R})}^2 \qquad \forall\, f \in W_2^{1/2}(\mathbb{R}) \tag{4.5}$$

with

$$C_n := \max(4, C_{\phi_n}) 2^{-n} \quad \text{and} \quad C_{\phi_n} := \sum_{\ell=1}^{\infty} 2^{-\ell} \deg(\widehat{d_{\ell+n}}). \tag{4.6}$$

Now we prove that $\lim_{n\to\infty} C_n = 0$ by showing $\lim_{n\to\infty} 2^{-n} C_{\phi_n} = 0$. Note that

$$2^{-n} C_{\phi_n} = \sum_{\ell=1}^{\infty} 2^{-(\ell+n)} \deg(\widehat{d_{\ell+n}}) = \sum_{j=n+1}^{\infty} 2^{-j} \deg(\widehat{d_j}). \tag{4.7}$$

Since $\deg(\widehat{d_j}) \leqslant 2\deg(\widehat{a_j})$, by our assumption in (1.4), we have

$$\sum_{j=1}^{\infty} 2^{-j}\deg(\widehat{d_j}) \leqslant 2\sum_{j=1}^{\infty} 2^{-j}\deg(\widehat{a_j}) < \infty.$$

Consequently, by (4.7), we conclude that

$$0 \leqslant \lim_{n\to\infty} 2^{-n}C_{\phi_n} \leqslant \lim_{n\to\infty} \sum_{j=n+1}^{\infty} 2^{-j}\deg(\widehat{d_j}) = 0.$$

That is, $\lim_{n\to\infty} 2^{-n}C_{\phi_n} = 0$. Thus, we have $\lim_{n\to\infty} C_n = \lim_{n\to\infty} \max(4, C_{\phi_n})2^{-n} = 0$. Now from (4.5), we see that

$$\lim_{n\to\infty} \|f - P_n(f)\|^2_{L_2(\mathbb{R})} = \lim_{n\to\infty} C_n|f|^2_{W_2^{1/2}(\mathbb{R})} = 0 \qquad \forall\, f \in W_2^{1/2}(\mathbb{R}).$$

Since $P_n = Q_n$ in (3.1), we conclude that

$$\|f\|^2_{L_2(\mathbb{R})} = \lim_{n\to\infty} \langle Q_n(f), f\rangle = \sum_{k\in\mathbb{Z}} |\langle f, \phi_0(\cdot - k)\rangle|^2 + \sum_{j=0}^{\infty}\sum_{\ell=1}^{\mathcal{J}_{j+1}}\sum_{k\in\mathbb{Z}} |\langle f, \psi^{\ell}_{j;j,k}\rangle|^2$$

$$\forall\, f \in W_2^{1/2}(\mathbb{R}).$$

Since $W_2^{1/2}(\mathbb{R})$ is dense in $L_2(\mathbb{R})$, (1.2) must hold for all $f \in L_2(\mathbb{R})$. Therefore, $\{\phi_0(\cdot - k) : k \in \mathbb{Z}\} \cup \{\psi^{\ell}_{j;j,k} : j \in \mathbb{N}_0, k \in \mathbb{Z}, \ell = 1, \ldots, \mathcal{J}_{j+1}\}$ is a tight frame of $L_2(\mathbb{R})$. □

Next, we prove Theorem 1.2.

Proof of Theorem 1.2. For item (1), applying Theorem 2.8, we conclude that all $\phi_j, j \in \mathbb{N}_0$, are compactly supported functions in $C^{\infty}(\mathbb{R})$. Since all the masks $\widehat{a_j}$ are 2π-periodic trigonometric polynomials with real coefficients and are symmetric about the origin, we have $\overline{\widehat{a_j}(\xi)} = \widehat{a_j}(\xi)$. Now by the definition of $\widehat{\phi_j}$ in (1.1), it is straightforward to see that all $\phi_j, j \in \mathbb{N}_0$, are real-valued and $\overline{\widehat{\phi_j}(\xi)} = \widehat{\phi_j}(\xi)$; that is, all $\phi_j, j \in \mathbb{N}_0$, are symmetric about the origin.

By the definition of $\widehat{a^I_{m,l}}$ and $\widehat{a_{m,l}}$, we have

$$1 - |\widehat{a^I_{m,l}}(\xi)|^2 = 1 + O(|\xi|^{2l}) \quad \text{and} \quad 1 - |\widehat{a_{m,l}}(\xi)|^2 = 1 + O(|\xi|^{2l}), \qquad \xi \to 0. \tag{4.8}$$

By (1.12) and (4.8), we see that $\widehat{b^{\ell}_j}(\xi) = O(|\xi|^{l_j})$ as $\xi \to 0$. Therefore, ψ^{ℓ}_j has l_{j+1} vanishing moments. Thus, item (2) holds.

For item (3), by the definition of $\widehat{b^{\ell}_j}$ in (1.17), it is straightforward to check that (1.12) holds with $\mathcal{J}_j = 3$ for all $j \in \mathbb{N}$. Now by Theorem 1.1, we see that $\{\phi_0(\cdot - k) : k \in \mathbb{Z}\} \cup \{\psi^{\ell}_{j;j,k} : j \in \mathbb{N}_0, k \in \mathbb{Z}, \ell = 1, \ldots, \mathcal{J}_{j+1}\}$ is a tight wavelet frame in $L_2(\mathbb{R})$.

Now we prove item (4) by using Theorem 3.2 and Lemma 3.3. Let ν be an arbitrary positive integer. Since $\liminf_{j\to\infty} l_j/m_j > 0$, there exist a positive integer N and $0 < \rho < \liminf_{j\to\infty} l_j/m_j$ such that $2\nu < N < \rho m_j < l_j \leqslant m_j$ for all $j \geqslant N$. Denote $\widehat{d_j}(\xi) := |\widehat{a_j}(\xi)|^2$. By Lemma 3.3, we see that (3.13) holds. That is, there exists a positive constant C, independent of j, such that

$$0 \leqslant 1 - \widehat{d_j}(\xi) \leqslant C|\xi|^{2\nu}, \qquad \xi \in [-\pi, \pi] \quad \text{and} \quad j \geqslant N. \tag{4.9}$$

We now use (4.9) to estimate the constants C_{ϕ_n} in (3.10) of Theorem 3.2. For $n \geqslant N$ and $\ell \in \mathbb{N}$, since $\widehat{d_{\ell+n}}(0) = 1$, it follows from (4.9) that

$$|\widehat{d_{\ell+n}}(0) - \widehat{d_{\ell+n}}(2^{-\ell}\xi)| = |1 - \widehat{d_{\ell+n}}(2^{-\ell}\xi)| \leqslant C 2^{-2\nu\ell}|\xi|^{2\nu} \qquad \forall\, \xi \in [-\pi, \pi].$$

Now by (4.1), we conclude that

$$1 - |\widehat{\phi_n}(\xi)|^2 \leqslant C|\xi|^{2\nu} \sum_{\ell=1}^{\infty} 2^{-2\nu\ell}, \qquad \xi \in [-\pi, \pi].$$

Therefore, (3.10) holds with

$$C_{\phi_n} := C \sum_{\ell=1}^{\infty} 2^{-2\nu\ell} = \frac{C}{1 - 2^{-2\nu}} < \infty.$$

Consequently, by $Q_n = P_n$ and Theorem 3.2, we conclude that

$$\|f - Q_n(f)\|_{L_2(\mathbb{R})} \leqslant C_1 2^{-\nu n} |f|_{W_2^\nu(\mathbb{R})} \qquad \forall\, f \in W_2^\nu(\mathbb{R}) \quad \text{and} \quad n \geqslant N,$$

where

$$C_1 := \max(2, \sqrt{C/(1 - 2^{-2\nu})}) < \infty$$

is independent of f and n. Since ν is arbitrary, the tight wavelet frame has the desired spectral frame approximation order. □

Now we prove Theorem 1.3.

Proof of Theorem 1.3. Item (i) has been proved in Theorem 1.2. It is evident that

$$\deg(\widehat{a_j}) = j \quad \text{and} \quad |\widehat{a_j}(\xi)|^2 = \cos^{2j}(\xi/2) = 1 + O(|\xi|^2), \qquad \xi \to 0.$$

Now it follows from the proof of item (iii) of Theorem 1.1 that there exists a positive constant C_1 such that

$$1 - |\widehat{\phi_n}(\xi)|^2 \leqslant C_1 n^2 |\xi|^2 \qquad \forall\, \xi \in [-\pi, \pi] \quad \text{and} \quad n \in \mathbb{N}.$$

That is, we conclude that

$$\left|1 - |\widehat{\phi_n}(\xi)|^2\right|^2 \leqslant C_1^2 n^4 |\xi|^4 \qquad \forall\, \xi \in [-\pi, \pi] \quad \text{and} \quad n \in \mathbb{N}. \tag{4.10}$$

Let B_2 be the B-spline of order 2. Then

$$|\widehat{B_2}(\xi)|^2 = \frac{\sin^4(\xi/2)}{(\xi/2)^4} \quad \text{and} \quad |\widehat{B_2}(2\xi)|^2 = \cos^4(\xi/2)|\widehat{B_2}(\xi)|^2.$$

Since $|\widehat{a_j}(\xi)|^2 = \cos^{2j}(\xi/2) \leqslant \cos^4(\xi/2)$ for all $\xi \in \mathbb{R}$ and $j \geqslant 2$, it is evident that $|\widehat{\phi_n}(\xi)|^2 \leqslant |\widehat{B_2}(\xi)|^2$ for all $\xi \in \mathbb{R}$ and $n \geqslant 2$. In particular, for $\xi \in [-\pi, \pi]$, we deduce that

$$\begin{aligned} \sum_{k \in \mathbb{Z}\setminus\{0\}} |\widehat{\phi_n}(\xi)|^2 |\widehat{\phi_n}(\xi + 2\pi k)|^2 &\leqslant \sum_{k \in \mathbb{Z}\setminus\{0\}} |\widehat{B_2}(\xi + 2\pi k)|^2 \\ &= |\xi|^4 \frac{\sin^4(\xi/2)}{(\xi/2)^4} \sum_{k \in \mathbb{Z}\setminus\{0\}} |\xi + 2\pi k|^{-4}. \end{aligned}$$

Setting $C_2 := \sup_{\xi\in[-\pi,\pi]} \sum_{k\in\mathbb{Z}\setminus\{0\}} |\xi+2\pi k|^{-4} < \infty$, we conclude that

$$\sum_{k\in\mathbb{Z}\setminus\{0\}} |\widehat{\phi_n}(\xi)|^2 |\widehat{\phi_n}(\xi+2\pi k)|^2 \leqslant C_2|\xi|^4, \qquad \xi\in[-\pi,\pi] \quad \text{and} \quad n\geqslant 2.$$

Now taking into account (4.10), we see that the two inequalities in (3.8) hold with $\nu=2$ and $C_{\phi_n} := \max(C_1,C_2)n^4$. Thus, by Theorem 3.2, we see that item (ii) holds.

Now we prove item (iii) using proof by contradiction. Suppose that (1.16) holds for some $\nu>0$. By [29, Theorem 2.2], we have $\big|1-|\widehat{\phi_n}(\xi)|^2\big|^2 \leqslant \pi C^2|\xi|^{2\nu}$ for almost every $\xi\in[-\pi,\pi]$ and $n\geqslant N$, where C is the positive constant in (1.16). That is, we must have

$$1-|\widehat{\phi_j}(\xi)|^2 \leqslant C_3|\xi|^\nu \qquad \forall\, \xi\in[-\pi,\pi] \quad \text{and} \quad j\geqslant N, \tag{4.11}$$

where $C_3 := \sqrt{\pi}C$ is a positive constant independent of j.

By the definition of $\widehat{\phi_j}$ in (1.1) and $\widehat{a_j}(\xi) = 2^{-j}(1+e^{-i\xi})^j$, it is evident that $\widehat{\phi_j}(\xi) = \widehat{B_j}(\xi)\widehat{\phi_0}(\xi)$, where B_j is the B-spline of order j. Since $|\widehat{\phi_0}(\xi)| \leqslant 1$, by $\widehat{\phi_j}(\xi) = \widehat{B_j}(\xi)\widehat{\phi_0}(\xi)$, we have $|\widehat{\phi_j}(\xi)| \leqslant |\widehat{B_j}(\xi)|$, and, therefore, it follows from (4.11) that

$$1-|\widehat{B_j}(\xi)|^2 \leqslant 1-|\widehat{\phi_n}(\xi)|^2 \leqslant C_3|\xi|^\nu \qquad \forall\, \xi\in[-\pi,\pi] \quad \text{and} \quad j\geqslant N. \tag{4.12}$$

Since $|\widehat{B_j}(\xi)|^2 = \cos^{2j}(\xi/4)|\widehat{B_j}(\xi/2)|^2$ and $|\widehat{B_j}(\xi)| \leqslant 1$, we have

$$1-|\widehat{B_j}(4\xi)|^2 = 1-\cos^{2j}(\xi)+\cos^{2j}(\xi)(1-|\widehat{B_j}(2\xi)|^2) \geqslant 1-\cos^{2j}(\xi).$$

Consequently, (4.12) implies

$$\frac{1-\cos^{2j}(\xi)}{|\xi|^\nu} \leqslant \frac{1-|\widehat{B_j}(4\xi)|^2}{|\xi|^\nu} \leqslant C_4 \qquad \forall\, \xi\in[-\pi/4,\pi/4] \quad \text{and} \quad j\geqslant N, \tag{4.13}$$

where $C_4 := 4^\nu C_3 < \infty$. Noting $1=[\cos^2(\xi)+\sin^2(\xi)]^j \geqslant \cos^{2j}(\xi)+j\cos^{2j-2}(\xi)\sin^2(\xi)$, by $4\pi^{-2}\xi^2 \leqslant \sin^2(\xi) \leqslant |\xi|^2$ for all $\xi\in[-\pi/2,\pi/2]$, we deduce that, for $\xi\in[-\pi/4,\pi/4]$,

$$\begin{aligned}\frac{1-\cos^{2j}(\xi)}{|\xi|^\nu} &\geqslant j\frac{\sin^2(\xi)\cos^{2j-2}(\xi)}{|\xi|^\nu} = j\frac{\sin^2(\xi)(1-\sin^2(\xi))^{j-1}}{|\xi|^\nu}\\ &\geqslant 4\pi^{-2}j(\xi^2)^{1-\nu/2}(1-\xi^2)^{j-1}.\end{aligned}$$

Taking $\xi_j := \sqrt{\frac{1-\nu/2}{j-\nu/2}}$, we observe that $\lim_{j\to\infty}\xi_j = 0$, and it follows from the above inequalities and (4.13) that, for $\xi\in[-\pi/4,\pi/4]$ and sufficiently large j,

$$\begin{aligned}C_4 &\geqslant \frac{1-\cos^{2j}(\xi_j)}{|\xi_j|^\nu} \geqslant 4\pi^{-2}j(\xi_j^2)^{1-\nu/2}(1-\xi_j^2)^{j-1}\\ &= 4\pi^{-2}j\frac{(1-\nu/2)^{1-\nu/2}(j-1)^{j-1}}{(j-\nu/2)^{j-\nu/2}} := C_5c_j,\end{aligned}$$

where $0<C_5 := 4\pi^{-2}(1-\nu/2)^{1-\nu/2} < \infty$ and $c_j := j\frac{(j-1)^{j-1}}{(j-\nu/2)^{j-\nu/2}}$. That is, we must have

$$c_j := j\frac{(j-1)^{j-1}}{(j-\nu/2)^{j-\nu/2}} \leqslant C_4/C_5 \tag{4.14}$$

for all sufficiently large integers j.

By calculation, we have

$$c_j = j\frac{(j-1)^{j-1}}{(j-\nu/2)^{j-\nu/2}} = j\frac{j^{j-1}(1-\frac{1}{j})^{j-1}}{j^{j-\nu/2}(1-\frac{\nu}{2j})^{j-\nu/2}} = j^{\nu/2}\Big(1-\frac{1}{j}\Big)^{j-1}\Big(1-\frac{\nu}{2j}\Big)^{-j+\nu/2}.$$

We note that $\lim_{j\to\infty}(1-\frac{1}{j})^{j-1} = e^{-1}$ and $\lim_{j\to\infty}(1-\frac{\nu}{2j})^{-j+\nu/2} = e^{\nu/2}$. Hence, by $\nu > 0$, we conclude that $\lim_{j\to\infty} c_j = \lim_{j\to\infty} j^{\nu/2}e^{\nu/2-1} = \infty$, which is a contradiction to (4.14). So, the tight wavelet frame does not have any frame approximation order. Now item (iii) is verified. □

We finish this paper by proving Theorem 1.4.

Proof of Theorem 1.4. For item (1), by a direct calculation [23, Lemma 3], we observe that

$$\begin{aligned}|\widehat{a_j}(\xi)|^2 &= \cos^{2m_j}(\xi/2)([P^r_{m_j}(\sin^2(\xi/2))]^2 + [P^i_{m_j}(\sin^2(\xi/2))]^2)\\ &= \cos^{2m_j}(\xi/2)P_{m_j,m_j}(\sin^2(\xi/2)) = |\widehat{a^I_{m_j,m_j}}(\xi)|^2.\end{aligned}$$

Hence $|\widehat{a_j}(\xi)|^2 + |\widehat{a_j}(\xi+\pi)|^2 = 1$. By Theorem 2.8 and Proposition 2.6, we see that the nonstationary cascade algorithm associated with $\{\widehat{a_j}\}_{j=1}^\infty$ converges in $W_2^\nu(\mathbb{R})$ for any $\nu \geqslant 0$ and, therefore, all ϕ_j, $j \in \mathbb{N}_0$, are well-defined compactly supported functions in $C^\infty(\mathbb{R})$. By Theorem 2.10, (2.28) holds, and $\{\phi_0(\cdot - k) \ : \ k \in \mathbb{Z}\} \cup \{\psi_{j;j,k} \ : \ j \in \mathbb{N}_0, k \in \mathbb{Z}\}$ is an orthonormal basis of $L_2(\mathbb{R})$. By the same proof as in Theorem 1.2 and Lemma 3.3 with $l_j = m_j$, this orthonormal wavelet basis has the spectral approximation order. So, item (3) is verified.

The symmetry $\phi_j(1-\cdot) = \phi_j$ follows [23, Lemma 2] from the definition of ϕ_j in (1.1) and the symmetry of the masks $\widehat{a_j}$: $\widehat{a_j}(\xi) = e^{-i\xi}\widehat{a_j}(-\xi)$. Item (2) can be easily verified by (4.8). □

Acknowledgment. This work was done and completed while BH visited the National University of Singapore in 2006. BH would like to thank the National University of Singapore for their hospitality and support during his visit.

REFERENCES

[1] J.-F. Cai, R. Chan, L. Shen, and Z. Shen, *Restoration of chopped and nodded images by framelets*, SIAM J. Sci. Comput., 30 (2008), pp. 1205–1227.

[2] J.-F. Cai, R. H. Chan, and Z. Shen, *A framelet-based image inpainting algorithm*, Appl. Comput. Harmon. Anal., 24 (2008), pp. 131–149.

[3] A. Chai and Z. Shen, *Deconvolution: A wavelet frame approach*, Numer. Math., 106 (2007), pp. 529–587.

[4] R. H. Chan, S. D. Riemenschneider, L. Shen, and Z. Shen, *Tight frame: An efficient way for high-resolution image reconstruction*, Appl. Comput. Harmon. Anal., 17 (2004), pp. 91–115.

[5] R. H. Chan, Z. Shen, and T. Xia, *A framelet algorithm for enhancing video stills*, Appl. Comput. Harmon. Anal., 23 (2007), pp. 153–170.

[6] C. K. Chui and W. He, *Compactly supported tight frames associated with refinable functions*, Appl. Comput. Harmon. Anal., 8 (2000), pp. 293–319.

[7] C. K. Chui, W. He, and J. Stöckler, *Compactly supported tight and sibling frames with maximum vanishing moments*, Appl. Comput. Harmon. Anal., 13 (2002), pp. 224–262.

[8] C. K. Chui, W. He, and J. Stöckler, *Nonstationary tight wavelet frames*, II: *Unbounded interval*, Appl. Comput. Harmon. Anal., 18 (2005), pp. 25–66.

[9] A. Cohen, *Non-stationary multiscale analysis*, in Wavelets: Theory, Algorithms, and Applications (Taormina, 1993), Wavelet Anal. Appl. 5, Academic Press, San Diego, CA, 1994, pp. 3–12.

[10] A. Cohen and N. Dyn, *Nonstationary subdivision schemes and multiresolution analysis*, SIAM J. Math. Anal., 27 (1996), pp. 1745–1769.
[11] A. Cohen and R. D. Ryan, *Wavelets and Multiscale Signal Processing*, Chapman & Hall, London, 1995.
[12] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.
[13] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Ser. Appl. Math. 61, SIAM, Philadelphia, 1992.
[14] I. Daubechies, B. Han, A. Ron, and Z. Shen, *Framelets: MRA-based constructions of wavelet frames*, Appl. Comput. Harmon. Anal., 14 (2003), pp. 1–46.
[15] G. Derfel, N. Dyn, and D. Levin, *Generalized functional equations and subdivision processes*, J. Approx. Theory, 80 (1995), pp. 272–297.
[16] B. Dong and Z. Shen, *Pseudosplines, wavelets and framelets*, Appl. Comput. Harmon. Anal., 22 (2007), pp. 78–104.
[17] N. Dyn and D. Levin, *Subdivision schemes in geometric modeling*, Acta Numer., 11 (2002), pp. 73–144.
[18] S. S. Goh, Z. Y. Lim, and Z. Shen, *Symmetric and antisymmetric tight wavelet frames*, Appl. Comput. Harmon. Anal., 20 (2006), pp. 411–421.
[19] B. Han, *On dual wavelet tight frames*, Appl. Comput. Harmon. Anal., 4 (1997), pp. 380–413.
[20] B. Han, *Vector cascade algorithms and refinable function vectors in Sobolev spaces*, J. Approx. Theory, 124 (2003), pp. 44–88.
[21] B. Han, *Compactly supported tight wavelet frames and orthonormal wavelets of exponential decay with a general dilation matrix*, J. Comput. Appl. Math., 155 (2003), pp. 43–67.
[22] B. Han, *Refinable functions and cascade algorithms in weighted spaces with Hölder continuous masks*, SIAM J. Math. Anal., 40 (2008), pp. 70–102.
[23] B. Han, *Symmetric orthonormal complex wavelets with masks of arbitrarily high linear-phase moments and sum rules*, Adv. Comput. Math., to appear.
[24] B. Han and R.-Q. Jia, *Multivariate refinement equations and convergence of subdivision schemes*, SIAM J. Math. Anal., 29 (1998), pp. 1177–1199.
[25] B. Han and Q. Mo, *Tight wavelet frames generated by three symmetric B-spline functions with high vanishing moments*, Proc. Amer. Math. Soc., 132 (2004), pp. 77–86.
[26] B. Han and Q. Mo, *Splitting a matrix of Laurent polynomials with symmetry and its application to symmetric framelet filter banks*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 97–124.
[27] B. Han and Q. Mo, *Symmetric MRA tight wavelet frames with three generators and high vanishing moments*, Appl. Comput. Harmon. Anal., 18 (2005), pp. 67–93.
[28] K. Jetter and D. X. Zhou, *Order of linear approximation from shift-invariant spaces*, Constr. Approx., 11 (1995), pp. 423–438.
[29] K. Jetter and D. X. Zhou, *Approximation Order of Linear Operators and Finitely Generated Shift-Invariant Spaces*, preprint, 1998.
[30] W. Lawton, *Applications of complex valued wavelet transforms to subband decomposition*, IEEE Trans. Signal. Process., 41 (1993), pp. 3566–3568.
[31] W. Lawton, S. L. Lee, and Z. Shen, *Stability and orthonormality of multivariate refinable functions*, SIAM J. Math. Anal., 28 (1997), pp. 999–1014.
[32] W. Lawton, S. L. Lee, and Z. Shen, *Convergence of multidimensional cascade algorithm*, Numer. Math., 78 (1998), pp. 427–438.
[33] A. Ron and Z. Shen, *Affine systems in $L_2(\mathbb{R}^d)$: The analysis of the analysis operator*, J. Funct. Anal., 148 (1997), pp. 408–447.
[34] A. Ron and Z. Shen, *Generalized shift invariant systems*, Constr. Approx., 22 (2005), pp. 1–45.
[35] V. L. Rvachev and V. A. Rvachev, *A certain finite function*, Dopov. Dokl. Akad. Nauk. Ukraini, 8 (1971), pp. 705–707.
[36] I. W. Selesnick, *Smooth wavelet tight frames with zero moments*, Appl. Comp. Harmon. Anal., 10 (2000), pp. 163–181.
[37] Z. Shen, *Refinable function vectors*, SIAM J. Math. Anal., 29 (1998), pp. 235–250.

SIAM J. MATH. ANAL.
Vol. 40, No. 3, pp. 939–951

STABILITY AND SYNCHRONISM OF CERTAIN COUPLED DYNAMICAL SYSTEMS*

JOSÉ A. BARRIONUEVO† AND JACQUES A. L. SILVA†

Abstract. We obtain sufficient conditions for the stability of the synchronized solution for certain classes of coupled dynamical systems. These discrete time systems can be used to describe population patches coupled by migration, which are density and/or time dependent. Our results follow from an analytic expression for the transverse Lyapunov exponent obtained through spectral analysis. We then indicate some applications to population dynamics.

Key words. coupled dynamical systems, synchronism

AMS subject classifications. 37A30, 37H15, 37N30

DOI. 10.1137/060658436

1. Introduction.

The study of coupled dynamical systems has received considerable attention recently for its interest from the mathematical, physical, and biological points of view; see, for instance, [22], [1], [7], [29], [26], and [3], among others. One concern about such systems is whether or not they will present synchronization phenomena and whether or not such synchronization is stable.

The system below describes the evolution of a system consisting of d identical subsystems where, on every iteration, each subsystem undergoes its common local evolution determined by f, followed by a density-dependent coupling process encoded by C and φ. The system can model a population consisting of d patches, x_j, $j = 1, \dots, d$, where, in the absence of migration, patch j is controlled by a *local dynamics* $x^j_{t+1} = f(x^j_t)$. When migration is present, $\varphi(f(x^j_t))$ individuals leave patch j and are distributed with density c_{ji} on patch i. The *global dynamics* is then

$$x^j_{t+1} = f(x^j_t) - \varphi(f(x^j_t)) + \sum_{i=1}^{d} c_{ji}\varphi(f(x^i_t)); \;\; i, j = 1, \dots, d. \tag{1}$$

Here f is a bounded C^1 map on $[0, \infty)$, $C = [c_{ij}]$ is doubly stochastic, that is, $c_{ij} \geq 0$, and for all i, j, $\sum_{i=1}^{d} c_{ij} = \sum_{j=1}^{d} c_{ij} = 1$. Furthermore we will assume φ differentiable a.e. with φ' bounded. The models in population dynamics considered in [1], [7], [8], [27], [13], [26] are all particular cases of (1) for special choices of C and φ.

The condition on C being doubly stochastic reflects that there are no losses during the migration process. It is also necessary for the invariance of the diagonal of the phase space, that is, for $x^j_t = x^i_t = x_t$ to be a solution of (1), where each x^j_t satisfies $x^j_{t+1} = f(x^j_t)$.

In this paper we obtain sufficient conditions for the stability of the aforementioned synchronized solutions. The criteria involve the Lyapunov exponents, to be defined below, of the one-dimensional map f and of the codimension one transverse dynamical system.

*Received by the editors April 28, 2006; accepted for publication (in revised form) May 5, 2008; published electronically September 17, 2008.

http://www.siam.org/journals/sima/40-3/65843.html

†Department of Mathematics, Universidade Federal RS, Av. Bento Gonçalves 9500, 91509-900 Porto Alegre, RS, Brazil (josea@mat.ufrgs.br, jaqx@mat.ufrgs.br).

The paper is organized as follows. In the next section we provide a criterion for stability for general systems. In section 3 we improve our result for the case of normal operators. In section 4 we consider a system where the coupling/migration process is time dependent. We formulate and prove the corresponding results of the previous sections in this setting. In the last section we indicate some applications to population dynamics.

Previous results treated only the cases where φ' is a constant or a 2-valued step function, as well as particular examples of matrices C, and these cases are covered by our results. Our treatment only requires φ' to be bounded and C to be doubly stochastic and irreducible. This last condition is easily seen to be necessary; see below. Moreover, we are unaware of any treatment of the systems as in section 4 in the literature. Thus we extend some of the results of [7], [8], [10], [11], [27], [15], [3], [13], and others to these more general situations.

2. Stability: General case. In order to understand the behavior of orbits starting at nearby points of the diagonal of the phase space, we first linearize (1). If $J_t = [\alpha_{ij}]$ denotes the Jacobian matrix of (1) restricted to the synchronized orbit, we have

$$\alpha_{ij} = \begin{cases} f'(x_t)\,(1-(1-c_{ii})\varphi'(f(x_t))) & \text{for } \ i=j, \\ f'(x_t)\varphi'(f(x_t))c_{ij} & \text{for } \ i\neq j. \end{cases}$$

We have $J_t = f'(x_t)H_t$, where $H_t = I - \varphi'(f(x_t))B$, and $B = I - C$.

We will assume that C is irreducible. This is almost a necessary condition, for otherwise it permits the existence of uncoupled unsynchronized subsystems that are each synchronized. In this case we can apply the Fröbenius theorem [12] to show that $\lambda = 1$ is the simple dominant eigenvalue of C, associated to the eigenvector $v = (1,\dots,1)$. This furnishes the decomposition $\mathbb{R}^d = \mathbb{R}v \oplus W$, where W is a C-invariant $(d-1)$-dimensional subspace. Under these conditions,

$$B = P^{-1}\begin{bmatrix} 0 & \\ & A \end{bmatrix} P, \tag{2}$$

where P is the matrix of the appropriate change of basis. This decomposition implies that the stability of the synchronized solution of (1) is a consequence of the stability of the trivial solution of the transversal component, w_t, which satisfies

$$w_{t+1} = f'(x_t)\,(I - \varphi'(f(x_t))A)\,w_t. \tag{3}$$

We will show that under a certain integrability condition the map above is in fact a contraction, which in turn implies the stability of $w_t \equiv 0$. The analysis of (3) will be based on the Lyapunov exponents (see [17], [18]) of (3). Define

$$K_n(x) = \left|\left|\prod_{k=0}^{n-1} f'(f^k(x))\,\big(I - \varphi'(f^{k+1}(x))A\big)\right|\right|^{1/n},$$

where $f^0(x) = x$ and $f^k(x) = f(f^{k-1}(x))$ for $k > 0$. Clearly if $\mathcal{K} = \limsup K_n$ satisfies $\mathcal{K} < 1$, we have that (3) is a contraction. Now observe

$$K_n(x) = L_n(x)\Lambda_n(x) = \left|\prod_{k=0}^{n-1} f'(f^k(x))\right|^{1/n} \left|\left|\prod_{k=0}^{n-1} I - \varphi'(f^{k+1}(x))A\right|\right|^{1/n}. \tag{4}$$

$L_n(x)$ depends only on the local dynamics f, while $\Lambda_n(x)$ reflects also the effects of φ and C. Let ρ be an invariant measure of the local system. Define for $x > 0$, $\ln^+(x) = \max(\ln(x), 0)$. By Birkhoff's ergodic theorem, if $\ln^+ |f'| \in L^1(\rho)$, there exists $\lim_n \exp(\frac{1}{n}\sum_{k=0}^{n-1} \ln |f'(f^k(x))|)$ for ρ-a.e. x. For ρ ergodic, this limit, call it L, is independent of x and is given by $\exp\left(\int_0^\infty \ln |f'(s)|\, d\rho(s)\right)$. L is the Lyapunov exponent of the local system governed by f.

Similarly the ergodic theorem of Oseledec [18] implies that if

$$\int_0^\infty \ln^+ ||I - \varphi'(s)A||\, d\rho(s) < \infty, \tag{5}$$

there exists $\lim_n \Lambda_n(x) =: \Lambda(x)$ for ρ-a.e. x, and this limit is independent of x provided ρ is ergodic. In the rest of the paper we will assume that the f-invariant measure, ρ, is ergodic. Nonergodic ρ can be treated at the cost of additional technicalities. See remark (iii) in section 2.1. Our first result is the following.

THEOREM 1. *Consider the system* (1), *where f is a C^1 map, φ' is bounded, and C is doubly stochastic and irreducible such that $\ln^+ ||I - \varphi'(s)A||$ and $\ln^+ |f'(s)|$ are in $L^1(\rho)$, where ρ is an ergodic f-invariant measure. Let $L = \lim_n L_n(x)$ and $\Lambda = \lim_n \Lambda_n(x)$ be as above. If $L\,\Lambda < 1$, there exists a set E with $\rho(E) = 1$ such that for all $x \in E$, the synchronized solution of* (1) *is asymptotically stable.*

Proof. By Birkhoff's theorem there exists a set E with $\rho(E) = 1$ such that for all x in E, $\lim_n \Lambda_n(x) = \Lambda$. We claim that

$$\Lambda \le \exp\left(\int_0^\infty \ln^+ ||I - \varphi'(s)A||\, d\rho(s)\right). \tag{6}$$

By the continuity of the norm and the function $\ln^+(\cdot)$, the dominated convergence theorem implies that we need only prove (6) for φ' simple.

Let $\varphi'(x) = \sum_{k=1}^r a_k \chi_{E_k}(x)$, where E_k are measurable and disjoint with $\rho(E_k) > 0$ and such $E \subset \bigcup_k E_k$. For $1 \le k \le r$, and x in E, define $\rho_{k,n} = \rho_{k,n}(x)$ by

$$\rho_{k,n} = \frac{\sharp\{0 \le j < n : f^j(x) \in E_k\}}{n}.$$

To keep the notation simple we omit the dependence of $\rho_{k,n}$ on x since Birkhoff's ergodic theorem applied to χ_{E_k} shows that for $1 \le k \le r$, and all x in E, we have

$$\lim_n \rho_{k,n} = \rho(E_k).$$

This gives

$$\Lambda_n(x) = \left|\left| \prod_k (I - a_k A)^{n\rho_{k,n}} \right|\right|^{1/n},$$

which imply

$$\Lambda_n(x) \le \prod_k ||I - a_k A||^{\rho_{k,n}} \text{ and}$$

$$\ln \Lambda_n(x) \le \sum_k \rho_{k,n} \ln ||I - a_k A||.$$

Therefore, given $\epsilon > 0$, there exists n_0 such that for $n > n_0$, we have $\rho_{k,n} \leq (1+\epsilon)\rho(E_k)$ and $L_n(x) \leq (1+\epsilon)L$. Since $\ln(\cdot) \leq \ln^+(\cdot)$,

$$\ln \Lambda_n(x) \leq (1+\epsilon) \sum_{k=1}^{r} \ln^+ ||I - a_k A||\, \rho(E_k) = (1+\epsilon) \int_0^\infty \ln^+ ||I - \varphi'(s)A||\, d\rho(s).$$

This proves (6). Note that the right-hand side is independent of x. Thus for $n > n_0$,

$$K_n(x) = L_n(x)\, \Lambda_n(x) \leq (1+\epsilon) L\, \Lambda^{1+\epsilon},$$

and since $\epsilon > 0$ is arbitrary we have

$$\lim_n K_n(x) \leq L\, \Lambda < 1,$$

and thus the transversal map (3) is a contraction. □

2.1. Remarks.

(i) If $E = [0, \infty)$, $L\ \Lambda \leq 1$ is necessary for the stability of the synchronized solution.

(ii) In certain cases, which include the case H_t semisimple, we will prove below that the norm $||I - \varphi'(f^k(x))A||$ is simply the spectral radius, $\sigma_{-1}(H_t)$, of the restriction of H_t to the subspace W. Therefore, with the same hypothesis on f and φ, we have

$$\Lambda \leq \Lambda_1 = \exp\left(\int_0^\infty \ln\, \sigma_{-1}(H_{\varphi'(s)})\, d\rho(s) \right), \tag{7}$$

and thus $L\Lambda_1 < 1$ is a sufficient condition for the stability of the synchronized solution of (1).

(iii) If ρ fails to be ergodic, a deep result of Choquet (Theorem 31.3 in [6]) gives the representation

$$\rho = \int_{\mathcal{E}(P_1)} \pi\, d\nu(\pi), \tag{8}$$

where P_1 is the space of f-invariant probability measures and $\mathcal{E}(P_1)$ is the set of extreme points of P_1, that is, the ergodic f-invariant probability measures, and ν is a probability measure on $\mathcal{E}(P_1)$. Also equality in (8) is, in the weak sense,

$$\int \Psi\, d\rho = \int_{\mathcal{E}(P_1)} \left(\int \Psi\, d\pi \right) d\nu(\pi)$$

for all continuous Ψ. Given $\pi \in \mathcal{E}(P_1)$, we apply Theorem 1 to obtain a set E_π of full π measure such that for all $x \in E_\pi$, the limit $\Lambda_\pi = \lim_n \Lambda_n(x)$ exists and

$$\Lambda_\pi \leq \exp\left(\int_0^\infty \ln^+ ||I - \varphi'(s)A||\, d\pi(s) \right).$$

Similarly, if $\ln^+ |f'| \in L^1(\pi)$, by Birkhoff's theorem, there exists $L_\pi = \exp(\lim_n \frac{1}{n} \sum_{k=0}^{n-1} \ln |f'(f^k(x))|)$ for $x \in E_\pi$, and it is given by $L_\pi = \exp(\int_0^\infty \ln |f'(s)|\, d\pi(s))$. From Theorem 1 it follows that if $L_\pi\, \Lambda_\pi < 1$, then the synchronized solution of (1) is stable for all initial x in $D_\pi = \{(x, x, \ldots, x) : x \in E_\pi\}$. Since any two distinct ergodic f-invariant measures are mutually singular, the sets E_π form a partition of the whole space $[0, \infty)$ modulo null sets. Therefore, if $\sup_{\pi \in \mathcal{E}(P_1)} L_\pi\, \Lambda_\pi < 1$, we have that the synchronized solution of (1) is asymptotically stable. Since the set of ergodic invariant measures can be very complicated, it might be difficult to verify the above condition in general.

3. Stability: Normal operators. In the case of normal operators, that is, $AA^* = A^*A$, one can improve the previous result with the help of the functional calculus of [16] and [23]. Let L_n be as in (4) and define the operator valued cocycle

$$\mathcal{O}_n(x) = \prod_{k=0}^{n-1} \big(I - \varphi'(f^{k+1}(x))A\big) \tag{9}$$

so that we have $K_n(x) = L_n(x)\,||\mathcal{O}_n||^{1/n}$. The proof of Oseledec's theorem in [24] shows that

$$\lim_n (\,\mathcal{O}_n^*(x)\mathcal{O}_n(x)\,)^{\frac{1}{2n}} = \mathcal{O}(x) \tag{10}$$

exists a.e. in operator norm. The ergodicity of ρ implies that $\mathcal{O}(x)$ is independent of x; call it $\mathcal{O}$. Spectral analysis of $\mathcal{O}$ determines the Lyapunov exponents and respective subspaces. Oseledec's theorem was extended to operator valued cocycles in Hilbert and Banach spaces in [25] and [21] under an additional compactness hypothesis and with the operator norm convergence replaced by strong convergence in (10). The *very special* nature of our cocycles (9) allows us to determine $\mathcal{O}$ above whenever A is a bounded, normal, and not necessarily compact operator on a Hilbert space $\mathcal{H}$. We present the result in this generality for its own sake and because it involves little extra work. Note, however, that the study of the stability of synchronized solutions of (1) for $d = \infty$ offers additional difficulties; see the remarks in section 3.1.

If A is normal, Theorem 12.23 in [23] gives

$$A = \int_{\sigma(A)} \lambda \, dP(\lambda), \tag{11}$$

where $dP(\lambda)$ are the spectral projections associated with A; that is, for all u, v in $\mathcal{H}$, $d\,\langle P(\lambda)u, v\rangle$ is a complex measure and

$$\langle Au,\, v\rangle = \int_{\sigma(A)} \lambda \, d\,\langle P(\lambda)u, v\rangle. \tag{12}$$

For a bounded measurable function $F : \sigma(A) \to \mathbb{C}$, $F(A)$ is defined by

$$F(A) = \int_{\sigma(A)} F(\lambda)\, dP(\lambda). \tag{13}$$

Observe that

$$||F(A)|| \le ||F(\lambda)||_{L^\infty(\sigma(A))}, \tag{14}$$

implying that if $F_n(\lambda) \to F(\lambda)$ in $L^\infty(\sigma(A))$, $F_n(A) \to F(A)$ in operator norm.

Setting $\ln 0 = -\infty$, $e^{-\infty} = 0$, and recalling that an *upper semicontinuous* real valued function satisfies that for all $\alpha \in \mathbb{R}$, $\{\lambda : F(\lambda) < \alpha\}$ is open, we now prove the following.

PROPOSITION 2. *The function $F : \mathbb{C} \to \mathbb{R}$ defined by*

$$F(\lambda) = \exp\left(\int_0^\infty \ln|1 - \lambda\varphi'(s)|\, d\rho(s)\right) \tag{15}$$

is upper semicontinuous and therefore bounded on compact subsets of $\mathbb{C}$.

Proof. First note that $F(\lambda) \geq 0$ and

$$F(\lambda) = 0, \text{ if and only if } \int_0^\infty \ln|1 - \lambda\varphi'(s)|\, d\rho(s) = -\infty. \tag{16}$$

Moreover, f bounded implies that the support of ρ, supp ρ, is compact and Jenssen's inequality gives

$$F(\lambda) \leq \int_{supp\,\rho} |1 - \lambda\varphi'(s)|\, d\rho(s) \leq 1 + K|\lambda| < \infty, \tag{17}$$

where $K = \sup\{\, |\varphi'(s)| : s \in \operatorname{supp}\rho\}$. Thus $F(\lambda)$ is a well-defined nonnegative real number. Defining

$$F_M(\lambda) = \exp\left(\int_0^\infty g_M(\lambda, s)\, d\rho(s)\right), \text{ where } g_M(\lambda, s) = \max\{\, -M,\, \ln|1 - \lambda\varphi'(s)|\,\},$$

we have $F_{M+1}(\lambda) \leq F_M(\lambda)$, and monotone convergence gives $F(\lambda) = \lim_M F_M(\lambda)$. Since $g_M(\lambda, s)$ is continuous with respect to λ on $\mathbb{C} \times$ supp ρ and supp ρ is compact, F_M is continuous on $\mathbb{C}$. Thus F is the pointwise limit of a nonincreasing sequence of continuous functions and is therefore upper semicontinuous. ∎

The proposition enables us to define $\mathcal{O} = F(A)$,

$$\mathcal{O} = \int_{\sigma(A)} \exp\left(\int_0^\infty \ln|1 - \lambda\varphi'(s)|\, d\rho(s)\right) dP(\lambda). \tag{18}$$

The spectral mapping theorem then implies

$$\sigma(\mathcal{O}) = F(\sigma(A)) = \{F(\lambda) : \lambda \in \sigma(A)\}. \tag{19}$$

The application of (10) to synchronization is contained in the next theorem.

THEOREM 3. *Let f, φ', C, and L be as in Theorem* 1. *Assume A is normal and let Λ be the spectral radius of* (18). *Then, if $L\,\Lambda < 1$, the synchronized solution of* (1) *is asymptotically stable.*

Proof. As in Theorem 1 the inequality $L\ \Lambda < 1$ implies that the transverse component (3) is a contraction, implying the stability of the synchronized solution. Therefore, in order to prove the theorem we need only establish (10) in the strong sense, with $\mathcal{O}$ given by (18). We first consider φ' simple. Let $\varphi' = \sum_{k=1}^N \varphi_k \chi_{E_k}$, where E_k are disjoint measurable. As before, we define

$$\rho_{k,n} = \frac{\sharp\{0 \leq j < n : f^j(x) \in E_k\}}{n}.$$

Then

$$(\mathcal{O}_n^* \mathcal{O}_n)^{\frac{1}{2n}} = \prod_{k=1}^N |I - \varphi_k A|^{\rho_{k,n}} = \int_{\sigma(A)} \left(\prod_{k=1}^N |1 - \lambda\varphi_k|^{\rho_{k,n}}\right) dP(\lambda). \tag{20}$$

For each n, $F_n(\lambda) = \prod_{k=1}^N |1 - \lambda\varphi_k|^{\rho_{k,n}}$ is continuous. Moreover, for $k = 1, \ldots, N$, Birkhoff's theorem gives $\lim_n \rho_{n,k} = \rho(E_k) > 0$. Since $\sigma(A)$ is compact, we have that

$$\lim_n F_n(\lambda) = F(\lambda) = \prod_{k=1}^N |1 - \lambda\varphi_k|^{\rho(E_k)}$$

uniformly for λ in $\sigma(A)$. We can rewrite this as

$$\lim_n F_n(\lambda) = \exp\left(\sum_{k=1}^{N} \rho(E_k) \ln|1-\lambda\varphi_k|\right) = \exp\left(\int_0^\infty \ln|1-\lambda\varphi'(s)|\, d\rho(s)\right),$$

where the last equality is just the definition of the Lebesgue integral. By (14), we have (10) for φ' simple. For an arbitrary φ', the right-hand side of (10) is already well defined by (18). We now have

$$(\mathcal{O}_n^*\mathcal{O}_n)^{\frac{1}{2n}} = \int_{\sigma(A)} F_n(\lambda)\, dP(\lambda),$$

where $F_n(\lambda) = \left(\prod_{k=1}^n |1-\lambda\varphi'(x_k)|\right)^{1/n}$. Since one can approximate φ' uniformly by simple functions, the ergodic theorem and the above result for simple functions imply that for all λ in $\sigma(A)$, $\lim_n F_n(\lambda) = F(\lambda)$-pointwise convergence.

Thus for all u, v in $\mathcal{H}$, the Lebesgue dominated convergence theorem implies

$$\lim_n \int_{\sigma(A)} F_n(\lambda)\langle dP(\lambda)u, v\rangle = \int_{\sigma(A)} F(\lambda)\,\langle dP(\lambda)u, v\rangle.$$

Since $dP(\lambda)$ is a resolution of the identity, this implies that

$$\lim_n \left|\left|\left((\mathcal{O}_n^*\mathcal{O}_n)^{\frac{1}{2n}} - F(A)\right)u\right|\right| = 0$$

for all u in $\mathcal{H}$, which establish strong convergence in (10). This finishes the proof of Theorem 3. □

3.1. Remarks.

(i) In finite dimension we have $\sigma(A) = \{\lambda_1, \ldots, \lambda_M\}$ with corresponding spaces E_j, mutually orthogonal, with $\mathbb{R}^{d-1} = E_1 \oplus \cdots \oplus E_M$, and the conclusion of the theorem is that if $u \neq 0$ is in E_j, then $\lim_n ||(\mathcal{O}_n^*\mathcal{O}_n)^{\frac{1}{2n}}u|| = F(\lambda_j)$.

(ii) Since for all $\lambda \in \sigma(A)$, $|1-\lambda\varphi'(s)| \leq ||I - \varphi'(s)A||$, we have $\Lambda \leq \exp\left(\int_0^\infty \ln^+ ||I-\varphi'(s)A||\, d\rho(s)\right)$, and therefore Theorem 3 is an improvement of Theorem 1.

(iii) By definition $\Lambda = \sup\{|F(\lambda)| : \lambda \in \sigma(A)\}$, and thus (19) gives (7). All the effect of coupling to synchronization is reflected in Λ and can therefore, in certain cases, be independent of d; see section 5.

(iv) The case $d = \infty$ is more subtle. First, the choice of a function space for $\{x_j(t)\}_j$ plays an important role. For instance, the synchronized solution belongs to l^p only for $p = \infty$. Second, the Fröbenius theorem is not valid in the same generality. Third, if the function space considered is not a Hilbert space, the functional calculus above is not available.

4. Extensions. The analysis above can be applied to more general systems. One such instance is the case below, where the coupling/migration process no longer depends on the density but instead obeys a seasonal dynamics; that is, on each cycle a time-dependent fraction μ_t of each patch will mix according to a time-dependent distribution $C_t = [c_{ij}^t]$ as follows:

$$x_{t+1}^j = f(x_t^j) - \mu_t\, f(x_t^j) + \sum_{i=1}^{d} c_{ji}^t\, \mu_t\, f(x_t^i);\ \ i,j = 1,\ldots,d. \tag{21}$$

Again f is a C^1 map on $[0,\infty)$. The evolution of the migration parameters μ_t and C_t is governed by maps g and h on $[0,1]$. The idea is to write $\mu_{t+1} = g(\mu_t)$ and $C_{t+1} = h(C_t)$, but this is not accurate. To make this precise we argue as follows. For g continuous on $[0,1]$ and μ_0 arbitrary, we assume that $\mu_t = \mu_t(\mu_0)$ is given by $\mu_{t+1} = g(\mu_t)$. Similarly if $\{C_s\}_{s\in[0,1]}$ is a family of doubly stochastic irreducible matrices, h is a continuous map on $[0,1]$, and s_0 is arbitrary, we define $s_{t+1} = h(s_t)$ and $C_t = C_t(s_0)$ by $C_{t+1} = C_{h(s_t)}$. Assume that $G(\mu, s) = \mu\, C_s$ is a measurable operator valued map with respect to the product measure $\nu \times \eta$ on $[0,1]\times[0,1]$, where ν and η are, respectively, a g-invariant and h-invariant *ergodic* measure on $[0,1]$. Under these assumptions the diagonal of the phase space, $x_t^i = x_t^j = x_t$, is a synchronized solution and we are interested in its stability. This system falls under the general theory of random dynamical systems as presented in [2] and thus obeys, under the appropriate integrability condition, a mutiplicative ergodic theorem. Our approach is direct and avoids the use of this theory. Moreover, due to the particular nature of (21), we are able, as in the previous sections, to obtain much more precise information about the limit operators than is given by the ergodic theorem alone.

The Jacobian matrix of (21), $J_t = [\alpha_{ij}^t]$, is now given by

$$\alpha_{ij} = \begin{cases} f'(x_t)\,(1-\mu_t(1-c_{ii}^t)) & \text{for } i = j, \\ f'(x_t)\,\mu_t\,c_{ij}^t & \text{for } i \neq j, \end{cases}$$

yielding $J_t = f'(x_t)(I - \mu_t B_t)$, where $B_t = I - C_t$. Our first task is to decompose the linearized system into diagonal and transversal components and then estimate the respective Lyapunov exponents by the appropriate ergodic theorem. This cannot be done in general for arbitrary families $\{C_t\}$; therefore we will consider special cases of increasing generality.

4.1. Simultaneous diagonalizable matrices. This includes the families of commuting symmetric matrices and of circulant matrices. In this case there exists a matrix P such that for all t, $B_t = P^{-1}\, M_t\, P$, where M_t is a diagonal matrix with entries $\{0, \lambda_2(t), \ldots, \lambda_d(t)\}$ in its diagonal. Each $\lambda_j(t)$ is measurable and satisfies $0 < |1 - \lambda_j(t)| < 1$ by the Fröbenius theorem (applied to C_t). In the symmetric (commuting) case the $\lambda_j(t)$ are real and lie in $(0,2)$. We obtain that the transversal component of (21) satisfies

$$w_{t+1} = f'(x_t)\,(I - \mu_t\, D_t)\, w_t, \tag{22}$$

where D_t is the $(d-1)$-dimensional $\{\lambda_2(t), \ldots, \lambda_d(t)\}$ diagonal matrix. As before, the synchronized solution of (21) is stable if and only if $w_t = 0$ is a stable solution of (22). For each x, μ_0, s_0 in $[0,1]$, define $L_n(x)$ as in (4), and for $j = 2, \ldots, d$, define

$$\Lambda_{j\,n}(\mu_0,\, s_0) = \left(\prod_{k=0}^{n-1} |1 - g^k(\mu_0)\, \lambda_j(h^k(s_0))| \right)^{1/n}.$$

The analogue of Theorem 3 is as follows.

THEOREM 4. *Consider the system* (21) *under the above conditions. Assume in addition that*

(i) $\ln^+ |f'(s)|$ *belongs to* $L^1(\rho)$, *where* ρ *is an ergodic* f*-invariant probability measure on* $[0,\infty)$.

(ii) *for all* $2 \le j \le d$, $\ln^+ |1 - \mu\, \lambda_j(s)|$ *belong to* $L^1(\nu \times \eta)$.

Then there exist sets $E \subset [0,\infty)$ *and* $F \subset [0,1]\times[0,1]$ *with* $\rho(E) = (\nu\times\eta)(F) = 1$, *such that for all* $x \in E$ *and* $(\mu_0, s_0) \in F$, *the limits* $L = \lim_n L_n(x)$ *and* $\Lambda_j = \lim_n \Lambda_{j\,n}(\mu_0, s_0)$ *exist and are independent of* x *and* (μ_0, s_0). *Moreover, if* $\Lambda = \max_j \Lambda_j$ *satisfies* $L\Lambda < 1$, *the synchronized solution of* (21) *is asymptotically stable.*

Proof. The existence of $L(x)$ for x in a set of full measure follows, as before, from Birkhoff's ergodic theorem applied to $\ln L_n(\cdot)$ and (i). In addition, the limit is independent of x since ρ is ergodic. The decomposition (22) allows us to derive the transversal component directly without making use of Oseledec's theorem. For $2 \le j \le d$, $(\mu_0, s_0) \in [0,1]\times[0,1]$, we can write

$$\Lambda_{j\,n}(\mu_0, s_0) = \exp\left(\frac{1}{n}\sum_{k=0}^{n-1} \ln|1 - g^k(\mu_0)\,\lambda_j(h^k(s_0))|\right).$$

Condition (ii) and a $(d-1)$-fold application of Birkhoff's ergodic theorem imply the existence of a set F, with $\nu\times\eta\,(F) = 1$ such that for all j and all $(\mu_0, s_0) \in F$, there exists $\Lambda_j(\mu_0, s_0) = \lim_n \Lambda_{j\,n}(\mu_0, s_0)$. Since, in addition, $\nu\times\eta$ is ergodic, the limit is independent of (μ_0, s_0) and is given by

$$\Lambda_j = \exp\left(\int_{[0,1]\times[0,1]} \ln|1 - \mu\,\lambda_j(s)|\; d(\nu\times\eta)(\mu, s)\right). \tag{23}$$

This gives $\operatorname{diag}\{\Lambda_2, \ldots, \Lambda_d\}$ as the analogue of (18). Its spectral radius is then $\Lambda = \max_j \Lambda_j$. Thus, if $L\,\Lambda < 1$, the synchronized solution is asymptotically stable. □

4.2. Symmetric (noncommuting) matrices. In this case we no longer have a decomposition yielding a diagonal transversal component, like (22), and therefore our conclusions are somewhat weaker than the previous theorem. Since each C_t is symmetric, doubly stochastic, and irreducible, $\lambda = 1$ is the dominant eigenvalue with corresponding eigenvector $v = \{1, \ldots, 1\}$. Moreover, the $(d-1)$-dimensional subspace $W = (\mathbb{R}v)^{\perp}$ is invariant for all C_t. This provides a decomposition like (2),

$$B_t = P^{-1}\begin{bmatrix} 0 & \\ & A_t \end{bmatrix} P. \tag{24}$$

Accordingly, the transversal component of (21) now satisfies

$$w_{t+1} = f'(x_t)\,(I - \mu_t\, A_t)\, w_t. \tag{25}$$

The following theorem is the analogue of Theorem 1 in the present situation.

THEOREM 5. *Consider the system* (21), *where* f *and* μ_t *are as in Theorem* 4, *and assume* $\{C_t\}$ *as above. For each* (μ_0, s_0), *define*

$$\Lambda_n(\mu_0, s_0) = \left|\left|\prod_{k=0}^{n-1}\left(I - g^k(\mu_0)A_{h^k(s_0)}\right)\right|\right|^{1/n}.$$

Assume that $\ln^+ ||I - \mu\, A_{h(s)}||$ *belongs to* $L^1(\nu\times\eta)$. *Then there exists a set* $F \subset [0,1]\times[0,1]$ *of full* $\nu\times\eta$*-measure such that for all* $(\mu_0, s_0) \in F$, *the limit* $\Lambda = \lim_n \Lambda_n(\mu_0, s_0)$ *exists and is independent of* (μ_0, s_0). *Moreover, if* Λ *satisfies* $L\,\Lambda < 1$, *the synchronized solution of* (21) *is asymptotically stable.*

Proof. Consider the map $\mathcal{Q}$ on $[0,1]\times[0,1]$, given by $\mathcal{Q}(\mu, s) = (g(\mu), h(s))$. Clearly $\nu\times\eta$ is $\mathcal{Q}$ invariant, and by assumption, $\Lambda_n(\mu, s)$ above defines a $(\nu\times\eta)$-measurable cocycle. The existence of $\Lambda(\mu_0, s_0)$ for (μ_0, s_0) on a set of full measure

then follows from Oseledec's ergodic theorem applied to $\Lambda_n(\mu, s)$. We also have that for a.e. (μ_0, s_0),

$$\Lambda(\mu_0, s_0) \leq \exp\left(\int_{[0,1]\times[0,1]} \ln^+ ||I - \mu A_{h(s)}||\, d(\nu \times \eta)(\mu, s)\right). \tag{26}$$

The proof of (26) follows the same path as the proof of (6), and thus we will omit the details. Note that since each A_t is symmetric, there exists $\lambda(t) \in \sigma(A_t)$ for which $||I - \mu\, A_t|| = |1 - \mu\lambda(t)|$. $\lambda(t)$ is the lowest eigenvalue of C_t. Moreover, since $\nu \times \eta$ is ergodic, $\Lambda(\mu_0, s_0)$, is independent of (μ_0, s_0), and in this case we have

$$\Lambda = \exp\left(\int_{[0,1]\times[0,1]} \ln\, |1 - \mu\, \lambda(h(s))|\, d(\nu \times \eta)(\mu, s)\right)$$

as a consequence of Birkhoff's theorem. As in the previous theorems, the condition $L\,\Lambda < 1$ implies the asymptotic stability of the synchronized solution. □

4.3. Remarks.

(i) The assumptions on the continuity of g and h above can be relaxed. All that is needed is that each map possess an ergodic invariant probability measure, ν and η.

(ii) Similarly to the density-dependent migration, the results above *do not* extend to the $d = \infty$ case even though, once again, the effect of d is encoded in the joint spectrum of $\{C_t\}$.

4.4. Examples. Some special cases of (23) and (26) are worth mentioning.

(i) If s_0 is a periodic point for h, that is, $h^p(s_0) = s_0$, and we take $\eta = \frac{1}{p}\sum_{k=0}^{p-1} \delta_{h^k(s_0)}$, then (23) becomes

$$\Lambda_j = \exp\left(\frac{1}{p}\sum_{k=0}^{p-1} \int_{[0,1]} \ln\, |1 - \mu\, \lambda_j(s_k)|\, d\nu(\mu)\right).$$

A similar expression corresponding to the case of a periodic point for g also holds.

(ii) If $g = h$ and $\eta = \nu$, but $\mu_0 \neq s_0$, we get for (23),

$$\Lambda_j = \exp\left(\int_{[0,1]\times[0,1]} \ln\, |1 - \mu\, \lambda_j(s)|\, d(\nu \times \nu)(\mu, s)\right).$$

(iii) If in (ii) we have, in addition, $\mu_0 = s_0$, that is, the term $\mu_t\, A_t$ in (25) is of the form $g^t(s_0)\, A_{g^t(s_0)}$, then (23) becomes

$$\Lambda_j = \exp\left(\int_{[0,1]} \ln\, |1 - s\, \lambda_j(s)|\, d\nu(s)\right).$$

Similar corresponding expressions can be obtained, without difficulty, to represent (26) in the special cases above. We leave the details, as well as the formulation of other special instances of (26), to the interested reader. These formulas are useful in cases where ν and η are known, say Lebesgue, for they permit the exact calculations of the Lyapunov numbers of the transversal dynamical system.

5. Applications. In this section we will restrict ourselves to the density-dependent system (1) and leave the corresponding formulations of the results related to the time-dependent system (21) to the interested reader.

A direct problem consists of determining stability once f, φ, and C are given. In such situations we compute the eigenvalues $\lambda_1, \dots, \lambda_d$ of C. Discarding the eigenvalue 1, we evaluate the integrals defining the spectrum of Λ above by Birkhoff's ergodic theorem

$$\int_0^\infty \ln|1-\lambda\varphi'(s)|\, d\rho(s) = \lim_n \frac{1}{n}\sum_{k=0}^{n-1} \ln|1-\lambda\varphi'(f^k(x))|, \tag{27}$$

and then the maximum of these values gives the spectral radius Λ and we can then determine stability once the Lyapunov exponent for f is known. We note that in order to use (27) we need the ergodic f-invariant measure, ρ, to be such that for x belonging to a set E of full Lebesgue measure, and all continuous functions Ψ,

$$\lim_n \frac{1}{n}\sum_{j=0}^{n-1} \Psi(f^i(x)) = \int \Psi(s)\, d\rho(s).$$

Such measures, called physical or SRB (Sinai–Ruelle–Bowen) measures, are known to exist in certain important cases, such as expansive maps, piecewise monotonic maps of the interval, as well as *axiom A* diffeomorphisms, even though a recent result of Ávila and Bochi [4] shows that they are not typical in the C^1-topology. Further information can be found in [31], [19], [20]. See also [32] for a survey.

In [5] we consider the following particular instance of (1). The local map f is taken from the Riker family, $f(x) = xe^{r(1-x)}$, $r > 0$, often considered in population dynamics. It is known that there exists an ergodic f-invariant probability SRB measure ρ which describes the asymptotic distribution of almost all orbits of f (see Theorems 25 and 29 in [30]). We let $\varphi(x) = x\mu(x)$, where the dispersal fraction, $\mu(x)$, is a sigmoidal function

$$\mu(x) = \frac{\alpha}{1+e^{\beta(\gamma-x)}}, \tag{28}$$

where the parameter $\alpha \in (0,1)$ determines the maximal dispersal, and β describes the steepness of dispersal. Moreover, the sign of β determines whether dispersal is positively or negatively density dependent. γ is the inflection point, at which dispersal is half of its maximum. The cases $\beta = \pm\infty$ correspond to threshold population size triggering migration and have been considered in [8]. Here we consider the following two extreme types of coupling:

(I) A ring configuration of d patches, each patch connected to the $2k$-nearest neighbors, with uniform dispersion, that is, $c_{ij} = 1/2k$ for $0 < |i-j| \le k$ and $c_{ij} = 0$ otherwise. In this case the eigenvalues of B are given by (see [27], [28], where the stability of equilibria is analyzed)

$$\lambda_j = 1 - \left(\frac{D_k(\frac{2\pi(j-1)}{d}) - 1}{2k}\right), \qquad j = 1, 2, \dots, d,$$

where $D_k(x) = \sin(k+\frac{1}{2})x / \sin\frac{x}{2}$.

(II) Global uniform coupling, that is, each of the d patches is connected to the others yielding $c_{ii} = 0$ and $c_{ij} = 1/(d-1)$ for $i \neq j$. Here the eigenvalues of B are $\lambda_1 = 0$ and $\lambda_2 = \cdots = \lambda_d = d/(d-1)$.

For $i = \text{I, II}$, let $\Lambda_i = \Lambda_i(d, r, \alpha, \beta, \gamma)$ be the value of Λ, obtained using (27), for the respective configuration. In [5] we observe numerically the dependence of Λ_i on the different parameters and its effects on the stability of synchronization. For instance, one can prove (see Corollary 3.1 there) that $\lim_{d\to\infty} \Lambda_I \geq 1$, independent of the other parameters, which are kept fixed. This easily implies the impossibility of chaotic synchronization for sufficiently large ring configurations such as (I). The situation is different for (II), where, given any d, one can find a local chaotic map f ($L > 1$) and migration parameters, α, β, and γ, yielding a stable synchronized dynamics ($L\,\Lambda < 1$).

One feature of Theorems 1 and 3 is their continuous dependence on φ', that is, C^1-uniform topology. This follows directly from the formulas for Λ above. On the other hand, by analyzing (27) one can see that for certain f, C, it is possible to make Λ arbitrarily large by taking $||\varphi'||_{L^\infty(\rho)}$ large. In our case, where φ is given using (28), this implies that there exist situations where threshold triggered migration, $\beta = \pm\infty$, that present chaotic synchronization, while all systems with β greater than a certain β_0 have unstable synchronized orbits. The same reasoning proves that, in general, there is no continuity of Λ in φ in the C^0-topology.

We can also make use of the results from previous sections to study an inverse problem. In this case f and φ are given and one is interested in finding a double stochastic matrix C, if possible, yielding the desired stability behavior for the synchronized solution of (1). In this regard, Theorem 7 below gives a partial result for symmetric matrices. We will make use of the following result from [14].

PROPOSITION 6. *The following are Corollaries* 7 *and* 8 *from* [14].

(i) *If* $\{\lambda_2, \ldots, \lambda_d\} \in [-1/(d-1), 1]$, *then there exists a symmetric doubly stochastic matrix* C *with* $\sigma(C) = \{1, \lambda_2, \ldots, \lambda_d\}$.

(ii) *If* $\lambda \in (-1, 1]$, *there exists a positive symmetric doubly stochastic matrix* C *such that* $\lambda \in \sigma(C)$.

The proof of Proposition 6 in [14] provides algorithms to find the matrices C in (i) and (ii) above.

THEOREM 7. *Let* f *and* φ *be as in Theorems* 1 *and* 3. *Let* L *denote the Lyapunov exponent of the one-dimensional map* f. *For* $\lambda \in [0, 2]$, *let* $F(\lambda)$ *be as in* (15), *and define* $m = \inf F(\lambda)$ *and* $M = \sup F(\lambda)$. *Then*

(i) *if* $L\,m > 1$, *the synchronized solution of* (1) *is unstable for* all *symmetric configurations* C.

(ii) *if* $L\,M < 1$, *the synchronized solution of* (1) *is stable for* all *symmetric configurations* C.

(iii) *if* $L \in (\frac{1}{M}, \frac{1}{m})$, *it is possible to find a symmetric doubly stochastic matrix* C *such that the synchronized solution of* (1) *has a prescribed stability behavior.*

Proof. Let us first note that any nonnegative symmetric doubly stochastic matrix has its spectrum contained in $(-1, 1]$. This follows from Gershgorin's theorem [9] which states that $\sigma(C)$ is contained in the set $\{\lambda \in \mathbb{C} : \text{ for all } i, |\lambda - c_{ii}| \leq \sum_{j\neq i} |c_{ij}|\}$. The hypotheses on C then imply that $\sigma(C) \subset [-1, 1]$. If C is positive, it is not difficult to show that -1 cannot be an eigenvalue of C. Since $B = I - C$, we have $\sigma(B) \subset [0, 2)$. Since $F(\lambda)$ gives the spectrum of $F(A)$, Theorem 7 now follows easily from Theorem 3 and Proposition 6. □

Acknowledgments. The authors would like thank the anonymous referees for carefully reading an earlier version of this work and for providing a long list of improvements. We also thank A. Lopes for some helpful conversations.

REFERENCES

[1] J. C. Allen, W. M. Schaffer, and D. Rosko, *Chaos reduces species extinction by amplifying local population noise*, Nature, 364 (1993), pp. 229–232.

[2] L. Arnold, *Random Dynamical Systems*, Springer-Verlag, New York, 1998.

[3] F. Atey, T. Biyikoğlu, and J. Jost, *Synchronization of Networks with Prescribed Degree Distributions*, preprint available online at http://xxx.lanl.gov/abs/nlin/0407024 (2005).

[4] A. Ávila and J. Bochi, *A generic C^1 map has no absolutely continuous invariant probability measure*, Nonlinearity, 19 (2006), pp. 2717–2725.

[5] J. A. Barrionuevo, F. Giordani, and J. A. L. Silva, *Synchronism in population network with non linear coupling models*, submitted. Preprint available upon request.

[6] G. Choquet, *Lectures on Analysis*, Vol. II: *Representation Theory*, W. A. Benjamin, Inc., New York, 1969.

[7] D. J. D. Earn, S. Levin, and P. Rohani, *Coherence and conservation*, Science, 290 (2000), pp. 1360–1364.

[8] J. A. L. Silva and F. T. Giordani, *Density-dependent migration and synchronism in metapopulations,* Bull. Math. Biol., 68 (2006), pp. 451–465.

[9] G. Hämmerlin and K. Hoffman, *Numerical Mathematics*, Springer-Verlag, New York, 1991.

[10] M. Hasler and Y. Maistrenko, *An introduction to the synchronization of chaotic systems: coupled skew tent maps*, IEEE Trans. Circuits Systems, 10 (1997), pp. 856–866.

[11] A. Hastings, *Complex interactions between dispersal and dynamics: Lessons from coupled logistic equations*, Ecology, 74 (1993), pp. 1362–1372.

[12] A. Householder, *The Theory of Matrices in Numerical Analysis*, Dover, New York, 1964.

[13] Y. Huang and O. Diekmann, *Interspecific influence on mobility and Turing instability*, Bull. Math. Biol., 65 (2003), pp. 143–156.

[14] S. Hwang and S. Pyo, *The inverse eigenvalue problem for symmetric doubly stochastic matrices*, Linear Algebra Appl., 379 (2004), pp. 77–83.

[15] S. R. J. Jang and A. K. Mitra, *Equilibrium stability of single species metapopulations*, Bull. Math. Biol., 62 (2000), pp. 155–161.

[16] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1995.

[17] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge University Press, Cambridge, UK, 1995.

[18] U. Krengel, *Ergodic Theorems*, Walter de Gruyter & Co., Berlin, 1985.

[19] A. Lasota and J. Yorke, *On the existence of invariant measures for piecewise monotonic transformations*, Trans. Amer. Math. Soc., 186 (1973), pp. 481–488.

[20] T. Y. Li and J. Yorke, *Ergodic transformations from an interval to itself*, Trans. Amer. Math. Soc., 235 (1978), pp. 183–192.

[21] R. Mañé, *Lyapunov exponents and stable manifolds for compact transformations*, in Geometric Dynamics, Lecture Notes in Math. 1007, Springer, 1983, pp. 522–577.

[22] A. Pikovsky, M. Rosenblum, and J. Kurths, *Synchronization—A Universal Concept in Nonlinear Sciences*, Cambridge University Press, Cambridge, UK, 2001.

[23] W. Rudin, *Functional Analysis*, 2nd ed., McGraw-Hill, New York, 1991.

[24] D. Ruelle, *Ergodic theory of differentiable dynamical systems*, Publ. Math. Inst. Hautes Études Sci., 50 (1979), pp. 27–58.

[25] D. Ruelle, *Characteristic exponents and invariant manifolds in Hilbert space*, Ann. of Math., 115 (1982), pp. 243–290.

[26] G. D. Ruxton, *Density-dependent migration and stability in a system of linked populations*, Bull. Math. Biol., 58 (1996), pp. 643–660.

[27] J. A. L. Silva, M. Castro, and D. Justo, *Stability in a metapopulation model with density-dependent dispersal*, Bull. Math. Biol., 63 (2001), pp. 485–506.

[28] J. A. L. Silva, M. Castro, and D. Justo, *Synchronism in a metapopulation model*, Bull. Math. Biol., 62 (2000), pp. 337–349.

[29] D. Tilman and P. Kareiva, *Spatial Ecology: The Role of Space in Population Dynamics and Interspecific Interactions*, Princeton University Press, Princeton, NJ, 1997.

[30] H. Thunberg, *Periodicity versus chaos in one-dimensional dynamics*, SIAM Rev., 43 (2001), pp. 3–30.

[31] M. Viana, *Lecture Notes on Attractors and Physical Measures*, A paper from the 12th Escuela Latinoamericana de Matemáticas, Monografías del IMCA 8, IMCA, Lima, 1999.

[32] L. S. Young, *What are SRB measures and which dynamical systems have them?*, J. Statist. Phys., 108 (2002), pp. 733–754.

SIAM J. MATH. ANAL.
Vol. 40, No. 3, pp. 952–967

STABILITY OF SINGLE-WAVE-FORM SOLUTIONS IN THE UNDERDAMPED FRENKEL–KONTOROVA MODEL*

WEN-XIN QIN†, CHUN-LAN XU†, AND XIN MA‡

Abstract. Via the monotonicity approach, we prove that the single-wave-form solution for the underdamped Frenkel–Kontorova model with dc-driving and periodic boundary conditions is globally stable, provided the driving force is large enough.

Key words. single-wave-form solution, Frenkel–Kontorova model, monotonicity

AMS subject classifications. 34C12, 34C15, 34C60, 34D45, 37C65

DOI. 10.1137/070699950

1. Introduction. The dynamics of many physical systems, including charge density waves and Josephson junctions, can be described by the standard Frenkel–Kontorova (F-K) model [5, 6], which can also be seen as a one-dimensional lattice of identical pendula oscillating in parallel planes, coupled to the nearest neighbors. The equation of the F-K model with dc-driving is

$$\ddot{u}_j + \Gamma \dot{u}_j + \sin u_j = K(u_{j-1} - 2u_j + u_{j+1}) + F, \tag{1.1}$$

where $\Gamma > 0$ denotes the damping effect, $K > 0$ measures the coupling strength, and $F \geq 0$ is the constant driving force.

In this paper we are concerned with the single-wave-form solutions of the F-K model (1.1) with periodic boundary conditions

$$u_{j+N}(t) = u_j(t) + 2\pi M, \tag{1.2}$$

where M and N are positive integers. A solution $\{u_j(t)\}$ of system (1.1)–(1.2) is called a single-wave-form solution if there exist a constant $T > 0$ and a waveform function $f\colon \mathbb{R} \to \mathbb{R}$ satisfying

$$f(t+T) = f(t) + 2\pi, \quad t \in \mathbb{R}, \tag{1.3}$$

such that

$$u_j(t) = f\left(t + jTM/N\right).$$

The positive number T is said to be the period of the waveform function f.

A running periodic solution is one for which there is a minimal $T > 0$ such that $u_j(t+T) = u_j(t) + 2\pi$ for $t \in \mathbb{R}$ and $j = 1, \ldots, N$. Note that a single-wave-form solution is a special form of running periodic solution with the same waveform and

*Received by the editors August 11, 2007; accepted for publication (in revised form) May 7, 2008; published electronically September 17, 2008. This work was supported by the National Natural Science Foundation of China (grants 10771155, 10571131) and the Natural Science Foundation of Jiangsu Province (grant BK 2006046).
http://www.siam.org/journals/sima/40-3/69995.html
†Department of Mathematics, Suzhou University, Suzhou 215006, People's Republic of China (wxqin@public1.sz.js.cn, xuchunlan50@163.com).
‡Golden Audit College, Nanjing Audit University, Nanjing 210029, People's Republic of China (amethystxin@163.com).

equal phase lags. Single-wave-form solutions are also called ponies on a merry-go-round [1], discrete rotating waves [2], or splay-phase solutions [11] in the literature. Although single-wave-form solutions in the coupled oscillators systems play an important role, it is not easy, as remarked in [16], to give mathematically rigorous analysis of their existence and stability. For the small coupling case, Levi proved by the Brouwer fixed point theorem the existence result in [9]. By reformulating the existence problem as a fixed point problem in a Banach space and applying the Schauder fixed point theorem, Katriel [7] presented a rigorous and detailed analysis on the existence of single-wave-form solutions of system (1.1)–(1.2). The method used in [7] is close in spirit to that in [12], which studied the existence of single-wave-form solutions for the Josephson junction systems. We remark that the stability question was not touched upon in [7] and [12]; see also the open problems in [16]. For the small coupling case, the local stability of the single-wave-form solutions, as remarked by Levi in [9], can be proved by applying the approach used in [8]. Meanwhile, by computing the Floquet multipliers and applying perturbation analysis, Watanabe and Swift investigated the local stability of single-wave-form solutions in series arrays of Josephson junctions [15]. Recently, Baesens and MacKay found the monotonicity for the F-K model under the overdamped condition $\Gamma > 2\sqrt{2K+1}$. It can be inferred from the results of [3, 4] that the single-wave-form solution of the overdamped F-K model (1.1)–(1.2) is globally stable if there is no equilibrium. Note that from [3] we know that there exists $F_d \geq 0$, which depends on Γ and K, such that the F-K model has equilibria for $0 \leq F \leq F_d$ and has no equilibria for $F > F_d$.

The goal of this paper is to study the global stability of the single-wave-form solutions for the underdamped F-K model, the existence of which was demonstrated by Katriel [7]. By global stability we mean that given any initial value $\{u_j(0)\}$ satisfying the periodic boundary condition $u_{j+N}(0) = u_j(0) + 2\pi M$, the solution $\{u_j(t)\}$ approaches the single-wave-form solution $\{\hat{u}_j(t)\}$ as $t \to +\infty$. Moreover, we will show that $\{\hat{u}_j(t)\}$ attracts each orbit with phase; i.e., for a solution $\{u_j(t)\}$, there exists a $\tau \in \mathbb{R}$, such that $\lim_{t\to+\infty} |u_j(t) - \hat{u}_j(t+\tau)| = 0$.

THEOREM 1.1. *Assume that $\Gamma > 2\sqrt{2K}$. Then there exists $F_0 > 1$, such that the F-K model* (1.1)–(1.2) *with $F \geq F_0$ admits a globally stable single-wave-form solution, which attracts each orbit with phase.*

The main idea of the proof is as follows. Inspired by the work of [3, 4, 13], we know that strong monotonicity generally implies stability of running periodic solutions. Therefore, as the first step, we show that the Poincaré map P^{mT} of system (1.1)–(1.2), where T is the period of the waveform function and m is some positive integer, is strongly monotone if $\Gamma > 2\sqrt{2K}$ and the driving force F is large enough. Then we prove via the strong monotonicity that the single-wave-form solution is globally stable.

We should remark that under the overdamped condition, i.e., $\Gamma > 2\sqrt{2K+1}$, the strong monotonicity found by Baesens and MacKay together with the approach presented for finite chains in [3] leads immediately to the stability of single-wave-form solutions. Here we extend the parameter range to the underdamped case, where the damping coefficient Γ could be small if the coupling strength K is small.

2. Trapping region. Let $u = (u_1, \ldots, u_N)^T \in \mathbb{R}^N$. Then system (1.1)–(1.2) can be written as

$$\ddot{u} + \Gamma\dot{u} + S(u) = -KAu + E, \tag{2.1}$$

where

$$A = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & -1 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ -1 & 0 & \cdots & 0 & -1 & 2 \end{pmatrix}$$

is an $N \times N$ matrix, $S(u) = (\sin u_1, \ldots, \sin u_N)^T$, $E = (-2\pi KM + F, F, F, \ldots, F, F, 2\pi KM + F)^T$. The eigenvalues of matrix A are

$$\lambda_j = 4 \sin^2 \frac{j\pi}{N}, \quad j = 0, 1, \ldots, N-1.$$

Let

$$\hat{e}_j = \left(1, \cos\frac{2j\pi}{N}, \ldots, \cos\frac{(N-1)2j\pi}{N}\right)^T, \ \tilde{e}_j = \left(0, \sin\frac{2j\pi}{N}, \ldots, \sin\frac{(N-1)2j\pi}{N}\right)^T,$$

$j = 1, \ldots, [\frac{N}{2}]$. Then $\hat{e}_j$ and $\tilde{e}_j$ are eigenvectors corresponding to $\lambda_j = \lambda_{N-j}$. Note that if N is even, then $\lambda_{N/2}$ is a simple eigenvalue with eigenvector $\hat{e}_{N/2}$. Let $e_j = \hat{e}/|\hat{e}_j|$ and $e_{N-j} = \tilde{e}_j/|\tilde{e}_j|$ for $j = 1, \ldots, [\frac{N}{2}]$. Then e_j is the normalized eigenvector corresponding to λ_j, $j = 1, \ldots, N-1$. The normalized eigenvector corresponding to $\lambda_0 = 0$ is $e_0 = (1/\sqrt{N}, \ldots, 1/\sqrt{N})^T$. Denoting $\mathbf{u} = (u, \dot{u})^T$, we have

$$\dot{\mathbf{u}} = C\mathbf{u} + \mathbf{S}(\mathbf{u}) + \mathbf{E}, \tag{2.2}$$

where

$$C = \begin{pmatrix} 0 & I \\ -KA & -\Gamma I \end{pmatrix}, \quad \mathbf{S}(\mathbf{u}) = \begin{pmatrix} 0 \\ -S(u) \end{pmatrix}, \quad \mathbf{E} = \begin{pmatrix} 0 \\ E \end{pmatrix}.$$

The eigenvalues of matrix C are

$$\mu_j^{\pm} = \frac{-\Gamma \pm \sqrt{\Gamma^2 - 4K\lambda_j}}{2},$$

with the corresponding eigenvector being $(e_j, \mu_j^{\pm} e_j)^T$, $j = 0, \ldots, N-1$. In particular, $\mu_0^+ = 0$, $\mu_0^- = -\Gamma$.

The conclusions of the following lemma have been proved by Katriel; see Theorem 2 in [7].

LEMMA 2.1. *System* (1.1)–(1.2) *admits a single-wave-form solution, provided* $F > 1$. *Moreover, the period* T *of the waveform function satisfies*

$$2\pi\Gamma/(F+1) \le T \le 2\pi\Gamma/(F-1).$$

The method of proof involves reformulating the problem as a fixed point problem in a Banach space and applying the Schauder fixed point theorem. The uniqueness of the single-wave-form solution needs some further assumptions; see the last section in [7].

A closed subset Ω of the phase space $\mathbb{R}^{2N}$ is called a trapping region of (2.2) if it is positively invariant for the flow generated by (2.2), and for each point $\mathbf{u}_0 \in \mathbb{R}^{2N}$,

there exists t_0 such that the orbit $\{\mathbf{u}(t)\}$ with $\mathbf{u}(0) = \mathbf{u}_0$ enters Ω and remains there for $t > t_0$.

LEMMA 2.2. *System* (2.2) *has a trapping region* Ω. *Moreover, there exist constants* $b > 0$, *independent of* F, v_-, *and* v^+, *depending upon the driving force* F, *such that*

$$|u_{j-1} - 2u_j + u_{j+1}| \le b, \qquad v_- \le \dot{u}_j \le v^+, \quad j = 1, \dots, N,\ u_0 = u_N,\ u_1 = u_{N+1},$$

for $\mathbf{u} = (u_1, \dots, u_n, \dot{u}_1, \dots, \dot{u}_N) \in \Omega$.

Proof. There exists an orthonormal matrix D, the jth column being e_{j-1}, such that $D^{-1}AD = \operatorname{diag}(\lambda_0, \lambda_1, \dots, \lambda_{N-1}) \triangleq \Lambda$. Note that $D^{-1} = D^T$, the transpose of D since D is an orthonormal matrix. Making a transformation $x = D^{-1}u$, where $x = (x_1, x_2, \dots, x_N)^T$, we convert system (2.1) into

$$\ddot{x} + \Gamma\dot{x} + D^{-1}S(Dx) = -K\Lambda x + D^{-1}E. \tag{2.3}$$

Noting that the jth element of $D^{-1}S(Dx)$ is $\langle e_{j-1}, S(Dx)\rangle$, denoted by $g_j(x,t)$, we have the fact that $|g_j(x,t)| \le \sqrt{N}$. Meanwhile, it follows that

$$D^{-1}E = D^{-1}\begin{pmatrix} F \\ F \\ \vdots \\ F \\ F \end{pmatrix} + D^{-1}\begin{pmatrix} -2\pi KM \\ 0 \\ \vdots \\ 0 \\ 2\pi KM \end{pmatrix} \triangleq \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{N-1} \\ f_N \end{pmatrix},$$

and hence

$$f_1 = \sqrt{N}F, \qquad |f_j| \le 2\sqrt{2}\pi KM, \quad j = 2, \dots, N.$$

Consequently, we derive that in an x-coordinate system,

$$\ddot{x}_j + \Gamma\dot{x}_j + K\lambda_{j-1}x_j = -g_j(x,t) + f_j. \tag{2.4}$$

There are three cases to discuss.

(1) $\Gamma^2 > 4K\lambda_{j-1}$. In this case the matrix

$$C_j = \begin{pmatrix} 0 & 1 \\ -K\lambda_{j-1} & -\Gamma \end{pmatrix}$$

has two real eigenvalues, $\mu_{j-1}^{\pm} = (-\Gamma \pm \sqrt{\Gamma^2 - 4K\lambda_{j-1}})/2$. It then follows that there exists a real matrix

$$T_j = \begin{pmatrix} 1 & 1 \\ \mu_{j-1}^+ & \mu_{j-1}^- \end{pmatrix}, \quad \text{such that} \quad T_j^{-1}C_jT_j = \begin{pmatrix} \mu_{j-1}^+ & 0 \\ 0 & \mu_{j-1}^- \end{pmatrix}.$$

Let

$$\begin{pmatrix} y_j \\ z_j \end{pmatrix} = T_j^{-1}\begin{pmatrix} x_j \\ \dot{x}_j \end{pmatrix}.$$

Then from (2.4) we have

$$\begin{pmatrix} \dot{y}_j \\ \dot{z}_j \end{pmatrix} = \begin{pmatrix} \mu_{j-1}^+ & 0 \\ 0 & \mu_{j-1}^- \end{pmatrix} \begin{pmatrix} y_j \\ z_j \end{pmatrix} + T_j^{-1} \begin{pmatrix} 0 \\ -g_j(x,t) + f_j \end{pmatrix}.$$

Denoting

$$Y_j = \begin{pmatrix} y_j \\ z_j \end{pmatrix}, \quad \Lambda_j = \begin{pmatrix} \mu_{j-1}^+ & 0 \\ 0 & \mu_{j-1}^- \end{pmatrix}, \quad \text{and} \quad h_j(x,t) = T_j^{-1} \begin{pmatrix} 0 \\ -g_j(x,t) + f_j \end{pmatrix}$$

yields $\dot{Y}_j = \Lambda_j Y_j + h_j(x,t)$, and hence

$$Y_j(t) = e^{\Lambda_j t} Y_j(0) + \int_0^t e^{\Lambda_j (t-\tau)} h_j(x(\tau),\tau) d\tau.$$

The boundedness of $g_j(x,t)$ and f_j implies that there exists $d_j > 0$, such that $\sup_t |h_j(x(t),t)| \le d_j$, and

$$|Y_j(t)| \le e^{(\mu_{j-1}^+)t} |Y_j(0)| + \frac{d_j}{|\mu_{j-1}^+|} \left(1 - e^{(\mu_{j-1}^+)t}\right), \quad t \ge 0.$$

Note that d_j $(j = 2, \ldots, N)$ are independent of F. Let $b_j = (d_j + 1)/|\mu_{j-1}^+|$. Then it follows that $|Y_j(t)| \le b_j$ for all $t \ge 0$ if $|Y_j(0)| \le b_j$. Moreover, for each solution of (2.2), there exists $t_0 > 0$ such that $|Y_j(t)| \le b_j$ for $t > t_0$.

In particular, when $j = 1$, we have the following estimates:

$$T_1 = \begin{pmatrix} 1 & 1 \\ 0 & \mu_0^- \end{pmatrix} \quad \text{and} \quad T_1^{-1} = \begin{pmatrix} 1 & -1/\mu_0^- \\ 0 & 1/\mu_0^- \end{pmatrix}.$$

As a consequence, it follows that

$$\dot{z}_1 = \mu_0^- z_1 - \frac{g_1(x,t)}{\mu_0^-} + \frac{f_1}{\mu_0^-} = -\Gamma z_1 + \frac{1}{\Gamma}(g_1(x,t) - f_1).$$

Let $b_1^- = -\Gamma^{-2}\sqrt{N}(F+1)$ and $b_1^+ = -\Gamma^{-2}\sqrt{N}(F-1)$. Then we have $z_1(t) \in [b_1^-, b_1^+]$ for all $t \ge 0$ if $z_1(0) \in [b_1^-, b_1^+]$. Moreover if $z_1(0) \notin [b_1^-, b_1^+]$, then there exists $t_0 > 0$ such that $z_1(t) \in [b_1^-, b_1^+]$ for all $t > t_0$.

(2) $\Gamma^2 = 4K\lambda_{j-1}$. In this case, taking

$$C_j = \begin{pmatrix} 0 & 1 \\ -K\lambda_{j-1} & -\Gamma \end{pmatrix} \quad \text{and} \quad T_j = \begin{pmatrix} 1 & 1 \\ -\Gamma/4 & -3\Gamma/4 \end{pmatrix},$$

we have

$$\begin{aligned} T_j^{-1} C_j T_j &= T_j^{-1} \begin{pmatrix} 0 & 1 \\ -\frac{3}{4}K\lambda_{j-1} & -\Gamma \end{pmatrix} T_j + T_j^{-1} \begin{pmatrix} 0 & 0 \\ -\frac{1}{4}K\lambda_{j-1} & 0 \end{pmatrix} T_j \\ &= \begin{pmatrix} -\frac{\Gamma}{4} & 0 \\ 0 & -\frac{3}{4}\Gamma \end{pmatrix} - \frac{\Gamma}{8} \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}. \end{aligned}$$

For each $z=(z_1,z_2)\in\mathbb{R}^2$, it follows that

$$\langle T_j^{-1}C_jT_jz,z\rangle=-\frac{\Gamma}{4}z_1^2-\frac{3}{4}\Gamma z_2^2-\frac{\Gamma}{8}(z_1^2-z_2^2)\le-\frac{\Gamma}{4}(z_1^2+z_2^2)=-\frac{\Gamma}{4}|z|^2.$$

Let

$$\begin{pmatrix} y_j \\ z_j \end{pmatrix}=T_j^{-1}\begin{pmatrix} x_j \\ \dot{x}_j \end{pmatrix}.$$

Then we have $\dot{Y}_j=\Lambda_jY_j+h_j(x,t)$, where $\Lambda_j=T_j^{-1}C_jT_j$. Since $\langle\Lambda_jz,z\rangle\le-\frac{\Gamma}{4}|z|^2$, then $\|e^{\Lambda_jt}\|\le e^{-\frac{\Gamma}{4}t}$ for $t\ge0$, and hence

$$\begin{aligned}|Y_j(t)|&\le e^{-\frac{\Gamma}{4}t}|Y_j(0)|+\int_0^t e^{-\frac{\Gamma}{4}(t-\tau)}|h_j(x(\tau),\tau)|d\tau\\&\le e^{-\frac{\Gamma}{4}t}|Y_j(0)|+\frac{4d_j}{\Gamma}\left(1-e^{-\frac{\Gamma}{4}t}\right),\quad t\ge0.\end{aligned}$$

Let $b_j=(4d_j+1)/\Gamma$. Then we obtain the same conclusion as in case (1).

(3) $\Gamma^2<4K\lambda_{j-1}$. In this case, the eigenvalues of the matrix C_j are a pair of conjugate complex numbers $\mu_{j-1}^{\pm}=\alpha_j\pm i\beta_j=(-\Gamma\pm\sqrt{\Gamma^2-4K\lambda_{j-1}})/2$, where $\alpha_j=-\Gamma/2<0$. Taking

$$T_j=\begin{pmatrix}1&0\\ \alpha_j&\beta_j\end{pmatrix}\ \text{ yields }\ T_j^{-1}C_jT_j=\begin{pmatrix}\alpha_j&\beta_j\\ -\beta_j&\alpha_j\end{pmatrix}\triangleq\Lambda_j.$$

Making a transformation

$$\begin{pmatrix} y_j \\ z_j \end{pmatrix}=T_j^{-1}\begin{pmatrix} x_j \\ \dot{x}_j \end{pmatrix},$$

we obtain from (2.4) that $\dot{Y}_j=\Lambda_jY_j+h_j(x,t)$. From the fact $\|e^{\Lambda_jt}\|\le e^{\alpha_jt}$ for $t\ge0$, it follows that

$$|Y_j(t)|\le e^{\alpha_jt}|Y_j(0)|+\frac{d_j}{|\alpha_j|}(1-e^{\alpha_jt}).$$

Taking $b_j=(d_j+1)/|\alpha_j|$ leads to the same conclusion as in case (1).

Consequently, the set

$$\Omega'=\{(y_1,z_1,\dots,y_N,z_N)\,|\,b_1^-\le z_1\le b_1^+,|(y_j,z_j)|\le b_j,j=2,\dots,N\}$$

is a trapping region. Let Ω denote this trapping region in a $\mathbf{u}$-coordinate system. For each $\mathbf{u}=(u,\dot{u})^T=(u_1,\dots,u_N,\dot{u}_1,\dots,\dot{u}_N)^T\in\Omega$, let $x=(x_1,\dots,x_N)^T=D^{-1}u$ and $\hat{x}=(0,x_2,\dots,x_N)^T$. Since $\mathbf{u}\in\Omega$, then the corresponding $(y_1,z_1,\dots,y_N,z_N)$ satisfies $|(y_j,z_j)|\le b_j$ for $j=2,\dots,N$, and

$$|(x_j,\dot{x}_j)|\le\|T_j\||(y_j,z_j)|\le b_j\|T_j\|\triangleq\hat{b}_j,\quad j=2,\dots,N,\tag{2.5}$$

implying that $|x_j|\le\hat{b}_j$, and $|\hat{x}|\le\sum_{j=2}^N\hat{b}_j$. From $ADx=AD\hat{x}$ we deduce that

$$|u_{j-1}-2u_j+u_{j+1}|=|Au|=|ADx|=|AD\hat{x}|\le\|A\|\|D\||\hat{x}|\le4|\hat{x}|\le4\sum_{j=2}^N\hat{b}_j\triangleq b.$$

Note that b_j $(j = 2, \ldots, N)$ are independent of F. So is b. From the facts $\dot{x}_1 = \mu_0^- z_1 = -\Gamma z_1$ and $z_1 \in [b_1^-, b_1^+]$ we derive that

$$\dot{x}_1 \in \left[\Gamma^{-1}\sqrt{N}(F-1), \Gamma^{-1}\sqrt{N}(F+1)\right].$$

Denoting $D = (D_{jk})$, we have from $\dot{u} = D\dot{x}$ that

$$\dot{u}_j = \frac{1}{\sqrt{N}}\dot{x}_1 + D_{j2}\dot{x}_2 + \cdots + D_{jN}\dot{x}_N,$$

implying by (2.5) that

$$\frac{F-1}{\Gamma} - \sum_{k=2}^{N} |D_{jk}|\hat{b}_k \le \dot{u}_j \le \frac{F+1}{\Gamma} + \sum_{k=2}^{N} |D_{jk}|\hat{b}_k.$$

Let $b^* = \sqrt{\sum_{k=2}^{N} \hat{b}_k^2}$. Then b^* is independent of F, and $\sum_{k=2}^{N} |D_{jk}|\hat{b}_k < b^*$ by the Cauchy inequality. Taking

$$v_- = \frac{F-1}{\Gamma} - b^* \quad \text{and} \quad v^+ = \frac{F+1}{\Gamma} + b^* \tag{2.6}$$

yields $v_- \le \dot{u}_j \le v^+,\ j = 1, \ldots, N.$ □

Let $\hat{\mathbf{u}}(t) = (\hat{u}(t), \dot{\hat{u}}(t))^T = (\hat{u}_1(t), \ldots, \hat{u}_N(t), \dot{\hat{u}}_1(t), \ldots, \dot{\hat{u}}_N(t))^T$ denote the single-wave-form solution obtained in Lemma 2.1.

LEMMA 2.3. *The single-wave-form solution* $\hat{\mathbf{u}}(t) \in \Omega$ *for* $t \in \mathbb{R}$.

Proof. Let $\hat{\mathbf{x}}(t) = (\hat{x}_1(t), \ldots, \hat{x}_N(t), \dot{\hat{x}}_1(t), \ldots, \dot{\hat{x}}_N(t))^T = (D^{-1}\hat{u}(t), D^{-1}\dot{\hat{u}}(t))^T$. Then it is easy to check that $\hat{x}_1(t+T) = \hat{x}_1(t) + 2\pi\sqrt{N}$ for $t \in \mathbb{R}$, and the other components are periodic with period $T > 0$, where T is the period of the waveform function corresponding to $\hat{\mathbf{u}}(t)$. From the proof of Lemma 2.2 we can denote this single-wave-form solution by $(\hat{y}_1(t), \hat{z}_1(t), \ldots, \hat{y}_N(t), \hat{z}_N(t))$. It follows that all the components, except $\hat{y}_1(t)$, are periodic, implying that the single-wave-form solution belongs to the trapping region Ω' for all $t \in \mathbb{R}$. □

In the remainder of this paper, we assume that $F \ge F_1 = \Gamma b^* + 1$ such that $v_- > 0$. Let $T_0 = 2\pi/b^*$. Then from Lemma 2.1 it follows that $0 < T \le T_0$ if $F \ge F_1$. Let

$$c = \frac{2}{v_-} + \frac{v^+ T_0(Kb + F + 1 + \Gamma v^+ + 2\pi KM)}{v_-^3}. \tag{2.7}$$

Note that c is a continuous function of F and $c \to 0$ as $F \to +\infty$.

LEMMA 2.4. *Assume that* $\mathbf{u}(t) = (u_1(t), \ldots, u_N(t), \dot{u}_1(t), \ldots, \dot{u}_N(t))$ *is a solution of* (2.2) *with* $\mathbf{u}(t) \in \Omega$ *for* $t \ge 0$. *Then*

$$\left| \int_{t_0}^{t_1} \cos u_j(t) dt \right| \le c,$$

where $0 \le t_0 < t_1$ *and* $t_1 - t_0 \le T_0$, $j = 1, \ldots, N$.

Proof. From Lemma 2.2 it follows that for $t \ge 0$,

$$\left|\frac{\ddot{u}_j}{\dot{u}_j}\right| = \left|\frac{K(u_{j-1} + u_{j+1} - 2u_j) - \sin u_j + F - \Gamma\dot{u}_j}{\dot{u}_j}\right|$$

$$\le \frac{Kb + F + 1 + \Gamma v^+ + 2\pi KM}{v_-},$$

and hence

$$\left|\int_{t_0}^{t_1}\cos u_j(t)dt\right| = \left|\int_{t_0}^{t_1}\frac{\cos u_j}{\dot u_j}du_j\right| = \left|\int_{t_0}^{t_1}\frac{1}{\dot u_j}d(\sin u_j)\right|$$

$$\le \left|\frac{\sin u_j}{\dot u_j}|_{t_0}^{t_1} + \int_{t_0}^{t_1}\frac{\sin u_j}{(\dot u_j)^2}d\dot u_j\right| \le \frac{2}{v_-} + \left|\int_{t_0}^{t_1}\frac{\sin u_j}{(\dot u_j)^2}\frac{\ddot u_j}{\dot u_j}du_j\right|$$

$$\le \frac{2}{v_-} + \frac{1}{v_-^2}\left(\frac{Kb+F+1+\Gamma v^+ + 2\pi KM}{v_-}\right)\left|\int_{t_0}^{t_1}du_j\right|$$

$$\le \frac{2}{v_-} + \frac{v^+T_0(Kb+F+1+\Gamma v^+ + 2\pi KM)}{v_-^3} = c. \qquad \square$$

3. Strong monotonicity. We know from [3] that there exists $F_d \ge 0$, which depends on Γ and K, such that the F-K model has equilibria for $0 \le F \le F_d$ and has no equilibria for $F > F_d$. Meanwhile, strong monotonicity generally implies stability of running periodic solutions if there is no equilibrium; see [3] and [13]. Therefore, we investigate in this section the strong monotonicity of system (1.1)–(1.2). To proceed, we make a transformation of variables:

$$\xi_j = u_j + \Gamma^{-1}\dot u_j, \quad \eta_j = \Gamma^{-1}\dot u_j, \quad j = 1, 2, \dots, N, \tag{3.1}$$

i.e., $\Xi = I_\Gamma\,\mathbf{u}$, in which $\Xi = (\xi, \eta)^T$, $\xi = (\xi_1, \dots, \xi_N)^T$, $\eta = (\eta_1, \dots, \eta_N)^T$,

$$I_\Gamma = \begin{pmatrix} I & \Gamma^{-1}I \\ 0 & \Gamma^{-1}I \end{pmatrix},$$

and I is the identity matrix of order N. Now the system (1.1)–(1.2) becomes

(3.2)

$$\begin{cases}
\dot\xi_1 = -\Gamma^{-1}\sin(\xi_1-\eta_1) + \Gamma^{-1}K[\xi_2-\eta_2+\xi_N-\eta_N-2(\xi_1-\eta_1)] \\
\qquad -2\pi K\Gamma^{-1}M + \Gamma^{-1}F, \\
\dot\xi_2 = -\Gamma^{-1}\sin(\xi_2-\eta_2) + \Gamma^{-1}K[\xi_1-\eta_1+\xi_3-\eta_3-2(\xi_2-\eta_2)] \\
\qquad +\Gamma^{-1}F, \\
\quad\vdots \\
\dot\xi_{N-1} = -\Gamma^{-1}\sin(\xi_{N-1}-\eta_{N-1}) \\
\qquad +\Gamma^{-1}K[\xi_{N-2}-\eta_{N-2}+\xi_N-\eta_N-2(\xi_{N-1}-\eta_{N-1})] + \Gamma^{-1}F, \\
\dot\xi_N = -\Gamma^{-1}\sin(\xi_N-\eta_N) + \Gamma^{-1}K[\xi_1-\eta_1+\xi_{N-1}-\eta_{N-1}-2(\xi_N-\eta_N)] \\
\qquad +2\pi K\Gamma^{-1}M + \Gamma^{-1}F, \\
\dot\eta_1 = -\Gamma\eta_1 + \dot\xi_1, \\
\quad\vdots \\
\dot\eta_N = -\Gamma\eta_N + \dot\xi_N.
\end{cases}$$

Set $\Phi = (\phi_1, \ldots, \phi_N)^T$ and $\Psi = (\psi_1, \ldots, \psi_N)^T$. The linearized equations along a solution $\Xi(t) \in I_\Gamma\Omega$ $(t \geq 0)$ are

$$\begin{cases} \dot{\phi}_j = -\Gamma^{-1}(\cos(\xi_j - \eta_j))(\phi_j - \psi_j) + \Gamma^{-1}K[\phi_{j+1} - \psi_{j+1} \\ \qquad + \phi_{j-1} - \psi_{j-1} - 2(\phi_j - \psi_j)] \triangleq L_j(\Phi, \Psi, t), \\ \dot{\psi}_j = -\Gamma\psi_j + \dot{\phi}_j \triangleq H_j(\Phi, \Psi, t), \end{cases} \tag{3.3}$$

in which $j = 1, 2, \ldots, N$, $\phi_0 = \phi_N, \phi_{N+1} = \phi_1, \psi_0 = \psi_N,\ \psi_{N+1} = \psi_1$.

First we study the properties of the auxiliary equations

$$\dot{\phi}_j = L_j(\Phi, \Psi, t) + \varepsilon, \quad \dot{\psi}_j = H_j(\Phi, \Psi, t), \tag{3.4}$$

in which $\varepsilon > 0$. Let

$$\Sigma = \left\{(\Phi, \Psi) \in \mathbb{R}^{2N} \mid -\phi_j \leq \psi_j \leq \phi_j,\ \ \phi_j \geq 0,\ j = 1, \ldots, N\right\}. \tag{3.5}$$

Let $\operatorname{int}\Sigma$ and $\partial\Sigma$ denote the interior and boundary of Σ, respectively.

Consider a first order equation

$$\dot{w} = \frac{1}{\Gamma}\left[-\Gamma^2 w - 2K(1-w)^2 - \cos u_j(t)(1-w)^2\right], \tag{3.6}$$

where $u_j(t)$ is a component of $\mathbf{u}(t)$ and $\mathbf{u}(t) \in \Omega$ $(t \geq 0)$ is a solution of (2.2) with external force F. The proof of the following lemma is postponed to the appendix.

LEMMA 3.1. *Assume $\Gamma > 2\sqrt{2K}$. Then there exists $F_0 \geq F_1$, such that each solution $w(t)$ of (3.6), in which $u_j(t)$ is a component of a solution $\mathbf{u}(t) \in \Omega$ $(t \geq 0)$ to (2.2) with $F \geq F_0$, satisfies that if $w(0) \in [-1, 1]$, then $1 - \Gamma^2/(4K) < w(t) < 1$ for all $t > 0$ and $-1 < w(t) < 1$ for $t \in [T_0/2, T_0]$.*

We remark that F_0 increases as N and M increase, as can be seen from the value F_1, which depends on the eigenvalues of matrix A and the proof in the appendix.

LEMMA 3.2. *Assume $\Gamma > 2\sqrt{2K}$ and $F \geq F_0$. Then the solution $(\Phi(\varepsilon, t), \Psi(\varepsilon, t))$ of (3.4), in which $\Xi(t) \in I_\Gamma\Omega$ $(t \geq 0)$ is a solution to (3.2), satisfies that $\phi_j(\varepsilon, t) > 0$ and $1 - \Gamma^2/(4K) < \psi_j(\varepsilon, t)/\phi_j(\varepsilon, t) < 1$ for all $t > 0$, and $-1 < \psi_j(\varepsilon, t)/\phi_j(\varepsilon, t) < 1$ for $t \in [T_0/2, T_0]$, $j = 1, \ldots, N$, if $(\Phi(\varepsilon, 0), \Psi(\varepsilon, 0)) \in \partial\Sigma\backslash\{0\}$.*

Proof. There are three cases to discuss for $(\Phi(\varepsilon, 0), \Psi(\varepsilon, 0)) \in \partial\Sigma\backslash\{0\}$.

(i) $\phi_j(\varepsilon, 0) > 0$ and $\phi_j(\varepsilon, 0) = \psi_j(\varepsilon, 0)$. Then there is a maximum $t_j^0 > 0$, or $t_j^0 = +\infty$, such that $\phi_j(\varepsilon, t) > 0$ for $t \in [0, t_j^0)$. Let $w_j(\varepsilon, t) = \psi_j(\varepsilon, t)/\phi_j(\varepsilon, t)$ for $t \in [0, t_j^0)$. Then it follows that

$$\begin{aligned} \dot{w}_j &= (\dot{\psi}_j\phi_j - \psi_j\dot{\phi}_j)/\phi_j^2 \\ &= \Gamma^{-1}\left[-\Gamma^2 w_j - 2K(1-w_j)^2 - \cos(\xi_j - \eta_j)(1-w_j)^2\right] - \varepsilon w_j/\phi_j \\ &\quad + \frac{K}{\Gamma\phi_j}(\phi_{j-1} - \psi_{j-1} + \phi_{j+1} - \psi_{j+1})(1 - w_j), \end{aligned}$$

and hence

$$\dot{w}_j|_{w_j=1, t=0} = -\Gamma - \varepsilon/\phi_j(\varepsilon, 0) < 0,$$

implying that $w_j(\varepsilon, t) < 1$ as long as $\phi_j(\varepsilon, t) > 0$, i.e., $\psi_j(\varepsilon, t) < \phi_j(\varepsilon, t)$ for $t \in (0, t_j^0)$.

(ii) $\phi_j(\varepsilon,0)=\psi_j(\varepsilon,0)=0$. In this case we have that $L_j(\Phi,\Psi,t)|_{(\Phi(\varepsilon,0),\Psi(\varepsilon,0))}\geq 0$, i.e., $\dot{\phi}_j|_{t=0}\geq\varepsilon>0$, and

$$0\leq\frac{H_j(\Phi,\Psi,t)}{L_j(\Phi,\Psi,t)+\varepsilon}\bigg|_{(\Phi(\varepsilon,0),\Psi(\varepsilon,0))}=\frac{\Gamma^{-1}K(\phi_{j-1}-\psi_{j-1}+\phi_{j+1}-\psi_{j+1})}{\Gamma^{-1}K(\phi_{j-1}-\psi_{j-1}+\phi_{j+1}-\psi_{j+1})+\varepsilon}<1.$$

Consequently, there is a maximum $t_j^0>0$, or $t_j^0=+\infty$, such that $\phi_j(\varepsilon,t)>0$ and $\psi_j(\varepsilon,t)/\phi_j(\varepsilon,t)<1$ for $t\in(0,t_j^0)$.

(iii) $\phi_j(\varepsilon,0)>0$ and $\psi_j(\varepsilon,0)=-\phi_j(\varepsilon,0)$. Then there exists a maximum $t_j^0>0$, or $t_j^0=+\infty$, such that $\phi_j(\varepsilon,t)>0$ and $\psi_j(\varepsilon,t)<\phi_j(\varepsilon,t)$ for $t\in(0,t_j^0)$. From the discussions of cases (i) and (ii) we know that

$$\frac{K}{\Gamma\phi_j}(\phi_{j-1}-\psi_{j-1}+\phi_{j+1}-\psi_{j+1})(1-w_j)\geq 0$$

for $t\in(0,t_0)$, where $t_0=\min_j\{t_j^0\}$. Furthermore, from the fact $\frac{-\varepsilon w_j}{\phi_j}|_{w_j=-1,t=0}>0$, Lemma 3.1, and the comparison principle we deduce that $\phi_j(\varepsilon,t)>0$ and $1-\Gamma^2/(4K)<w_j(\varepsilon,t)=\psi_j(\varepsilon,t)/\phi_j(\varepsilon,t)<1$ for $t\in(0,t_0)$. We claim that $t_0=+\infty$. If this is not true, then we have $\phi_j(\varepsilon,t)\to 0$ or $w_j(\varepsilon,t)\to 1$ as $t\to t_0^-$. In the former case, we have $\psi_j(\varepsilon,t)\to 0$ as $t\to t_0^-$, leading to a contradiction of the fact $\lim_{t\to t_0^-}\dot{\phi}_j(\varepsilon,t)\geq\varepsilon>0$. The latter case is also impossible. Indeed, if $\phi_j(\epsilon,t)\geq a>0$, where a is a positive number, then from part (i) we have that $w_j(\epsilon,t)<1-\delta$ for some small positive number δ since $\dot{w}|_{w_j=1,t=0}<0$. Consequently, we have $t_0=+\infty$. Another direct consequence from Lemma 3.1 and the comparison principle is that $-1<\psi_j(\varepsilon,t)/\phi_j(\varepsilon,t)<1$ for $t\in[T_0/2,T_0]$. Note that $\frac{-\varepsilon w_j}{\phi_j}|_{w_j=-1}>0$ and $\frac{-\varepsilon w_j}{\phi_j}|_{w_j=1}<0$. Then for $t\in[T_0/2,T_0]$, we have from the comparison principle that $-1<\psi_j(\varepsilon,t)/\phi_j(\varepsilon,t)<1$ holds true uniformly with respect to $\varepsilon>0$. ∎

LEMMA 3.3. *Assume that* $\Gamma>2\sqrt{2K}$ *and* $F\geq F_0$. *Then the solution* $(\Phi(t),\Psi(t))$ *of* (3.3), *the linearized equations of* (3.2) *along* $\Xi(t)\in I_\Gamma\Omega$ $(t\geq 0)$, *satisfies that* $\phi_j(t)\geq 0$ *and* $[1-\Gamma^2/(4K)]\phi_j(t)\leq\psi_j(t)\leq\phi_j(t)$ *for all* $t>0$, $j=1,\ldots,N$, *if* $(\Phi(0),\Psi(0))\in\partial\Sigma\backslash\{0\}$.

Proof. Assume that $(\Phi(\varepsilon,t),\Psi(\varepsilon,t))$ is a solution of the auxiliary equations (3.4) with the initial values $\Phi(\varepsilon,0)=\Phi(0)$ and $\Psi(\varepsilon,0)=\Psi(0)$. Then from Lemma 3.2 it follows that $[1-\Gamma^2/(4K)]\phi_j(\varepsilon,t)<\psi_j(\varepsilon,t)<\phi_j(\varepsilon,t)$ and $\phi_j(\varepsilon,t)>0$ for all $t>0$. Keeping t fixed and taking $\varepsilon\to 0$, we complete the proof. ∎

In fact, we have further conclusions on $\phi_j(t)$ and $\psi_j(t)$ under the same conditions of Lemma 3.3.

LEMMA 3.4. *Assume that* $\Gamma>2\sqrt{2K}$ *and* $F\geq F_0$. *Then the solution* $(\Phi(t),\Psi(t))$ *of* (3.3), *the linearized equations of* (3.2) *along* $\Xi(t)\in I_\Gamma\Omega$ $(t\geq 0)$, *satisfies that* $\phi_j(t)>0$ *for all* $t>0$, *and* $-\phi_j(t)<\psi_j(t)<\phi_j(t)$ *for* $t\in[T_0/2,T_0]$, $j=1,\ldots,N$, *if* $(\Phi(0),\Psi(0))\in\partial\Sigma\backslash\{0\}$.

Proof. Since $(\Phi(0),\Psi(0))\in\partial\Sigma\backslash\{0\}$, then we may assume that there is some k such that $\phi_k(0)>0$. From Lemma 3.3 we know that $[1-\Gamma^2/(4K)]\phi_k(t)\leq\psi_k(t)\leq\phi_k(t)$, $\psi_{k+1}(t)\leq\phi_{k+1}(t)$, and $\psi_{k-1}(t)\leq\phi_{k-1}(t)$ for $t\geq 0$, and thus we obtain from (3.3)

$$\dot{\phi}_k(t)\geq-\Gamma^{-1}\Gamma^2/(4K)\phi_k(t)-2\Gamma^{-1}K\Gamma^2/(4K)\phi_k(t)=-\frac{\Gamma(2K+1)}{4K}\phi_k(t),$$

implying from the Gronwall inequality that

$$\phi_k(t)\geq\phi_k(0)\,e^{\frac{-\Gamma(2K+1)\,t}{4K}}>0\quad\text{for}\quad t>0.$$

Let $w_k(t) = \psi_k(t)/\phi_k(t)$. Then

$$\dot{w}_k = \Gamma^{-1}\left[-\Gamma^2 w_k - 2K(1-w_k)^2 - \cos(\xi_k - \eta_k)(1-w_k)^2\right]$$
$$+ \frac{K}{\Gamma\phi_k}(\phi_{k-1} - \psi_{k-1} + \phi_{k+1} - \psi_{k+1})(1-w_k).$$

From the fact $\dot{w}_k|_{w_k=1} = -\Gamma < 0$, we obtain $w_k(t) < 1$, i.e., $\phi_k(t) - \psi_k(t) > 0$ for $t > 0$. Therefore from (3.3) we have that

$$\dot{\phi}_{k+1} \geq -\frac{\Gamma(2K+1)}{4K}\phi_{k+1} + \Gamma^{-1}K(\phi_k(t) - \psi_k(t))$$

and, again by the Gronwall inequality, that

$$\phi_{k+1}(t) \geq e^{\frac{-\Gamma(2K+1)\,t}{4K}}\phi_{k+1}(0) + \int_0^t e^{\frac{-\Gamma(2K+1)(t-\tau)}{4K}}\,\Gamma^{-1}K\left(\phi_k(\tau) - \psi_k(\tau)\right)d\tau > 0,\ \ t > 0.$$

With the same reasoning we have that $\phi_{k+2}(t) > 0$, $t > 0$, ..., and $\phi_{k-1}(t) > 0$, $t > 0, \ldots$. Consequently, we derive that $\phi_j(t) > 0$ for $t > 0$, $j = 1, \ldots, N$.

Keeping $t \in [T_0/2, T_0]$ fixed, we know from Lemma 3.2 that there exists $\delta > 0$ such that the solution $(\Phi(\varepsilon, t), \Psi(\varepsilon, t))$ of the auxiliary equations (3.4) with the initial values $\Phi(\varepsilon, 0) = \Phi(0)$ and $\Psi(\varepsilon, 0) = \Psi(0)$ has the property that

$$-1 + \delta \leq \frac{\psi_j(\varepsilon, t)}{\phi_j(\varepsilon, t)} \leq 1 - \delta.$$

We remark by virtue of the proof of Lemma 3.2 that δ is independent of ε. Taking $\varepsilon \to 0$, we derive that $-1 + \delta \leq \psi_j(t)/\phi_j(t) \leq 1 - \delta$. □

Now we define a partial order in the phase space $\mathbb{R}^{2N}$. Let $\Xi = (\xi_1, \ldots, \xi_N, \eta_1, \ldots, \eta_N)^T$ and $\Xi' = (\xi'_1, \ldots, \xi'_N, \eta'_1, \ldots, \eta'_N)^T$.

DEFINITION 3.5. *We define* $\Xi \leq \Xi'$ *if* $\xi_j \leq \xi'_j$ *and* $|\eta_j - \eta'_j| \leq |\xi_j - \xi'_j|$ *for* $j = 1, \ldots, N$. *We also define* $\Xi < \Xi'$ *if* $\Xi \leq \Xi'$ *and* $\Xi \neq \Xi'$, *and*

$$\Xi \ll \Xi' \ \text{ if } \ \xi_j < \xi'_j \ \text{ and } |\eta_j - \eta'_j| < |\xi_j - \xi'_j| \ \text{ for } j = 1, \ldots, N.$$

We remark that $\Xi \leq \Xi'$ if and only if $\Xi' - \Xi \in \Sigma$, and $\Xi \ll \Xi'$ if and only if $\Xi' - \Xi \in \text{int}\,\Sigma$. Let P^t denote the Poincaré map of system (3.2), i.e., $P^t\Xi = \Xi(t)$, in which $\Xi(t)$ is a solution of (3.2) with initial condition $\Xi(0) = \Xi$.

DEFINITION 3.6. *The Poincaré map* P^t $(t > 0)$ *of system* (3.2) *is said to be strongly monotone in the set* $\mathcal{B}$ *with* $P^t(\mathcal{B}) \subset \mathcal{B}$ *if*

$$\Xi_0 < \Xi'_0 \Rightarrow P^t\Xi_0 \ll P^t\Xi'_0 \ \text{ for } \ t > 0 \ \text{ and } \ \Xi_0, \Xi'_0 \in \mathcal{B}.$$

Now we can prove strong monotonicity by the above discussion and defining a homotopy as in [4, 14].

LEMMA 3.7. *Assume that* $\Gamma > 2\sqrt{2K}$ *and* $F \geq F_0$. *Then the Poincaré map* P^t *of system* (3.2) *is strongly monotone in* $I_\Gamma\Omega$ *for* $t \in [T_0/2, T_0]$.

Proof. Assume that $\Xi_0, \Xi'_0 \in I_\Gamma\Omega$, that $\Xi_0 < \Xi'_0$, and that $\Xi(t)$ and $\Xi'(t)$ are two solutions of (3.2) with initial values $\Xi(0) = \Xi_0$ and $\Xi'(0) = \Xi'_0$, respectively. Let

$$\Xi(\lambda, 0) = (1-\lambda)\Xi(0) + \lambda\Xi'(0),$$

where $\lambda \in [0,1]$. Assume that $\Xi(\lambda,t)$ is a solution of (3.2) with initial value $\Xi(\lambda,0)$. Then $\Xi(\lambda,t) \in I_\Gamma\Omega$ for $t \geq 0$ since $I_\Gamma\Omega$ is convex and positively invariant for system (3.2). Set

$$\Upsilon(\lambda,t) = \frac{\partial\, \Xi(\lambda,t)}{\partial\lambda} = (\phi_1(\lambda,t),\dots,\phi_N(\lambda,t),\psi_1(\lambda,t),\dots,\psi_N(\lambda,t))^T.$$

Then

$$\Upsilon(\lambda,0) = \frac{\partial\, \Xi(\lambda,0)}{\partial\lambda} = \Xi'(0) - \Xi(0) \quad \text{and} \quad \Xi'(t) - \Xi(t) = \int_0^1 \Upsilon(\lambda,t)d\lambda.$$

One can easily verify that $\Upsilon(\lambda,t)$ is a solution of the linearized equations (3.3) with initial value $\Upsilon(\lambda,0)$. Since $\Xi(0) < \Xi'(0)$, then

$$\phi_j(\lambda,0) \geq 0 \ \text{ and } \ |\psi_j(\lambda,0)| \leq \phi_j(\lambda,0), \quad j = 1,\dots,N,$$

i.e., $\Upsilon(\lambda,0) \in \Sigma\backslash\{0\}$. From Lemma 3.4 it follows that $\Upsilon(\lambda,t) \in \operatorname{int}\Sigma$ for $t \in [T_0/2, T_0]$, i.e.,

$$\phi_j(\lambda,t) > 0 \ \text{ and } \ |\psi_j(\lambda,t)| < \phi_j(\lambda,t), \quad j = 1,\dots,N, \quad t \in [T_0/2, T_0],$$

which implies that

$$\xi_j'(t) - \xi_j(t) = \int_0^1 \phi_j(\lambda,t)d\lambda > \int_0^1 |\psi_j(\lambda,t)|d\lambda \geq \left|\int_0^1 \psi_j(\lambda,t)d\lambda\right| = |\eta_j'(t) - \eta_j(t)|,$$

i.e., $\Xi(t) \ll \Xi'(t)$, $t \in [T_0/2, T_0]$. This completes the proof. □

4. Stability analysis. In this section we assume that $\Gamma > 2\sqrt{2K}$ and $F \geq F_0$ so that the conclusion of Lemma 3.7 holds true. We know from Lemma 2.1 that system (1.1)–(1.2) admits a single-wave-form solution $(\hat{u}_1(t),\dots,\hat{u}_N(t))$ since $F \geq F_0 > 1$. Let us denote it in the Ξ-coordinate system by $\hat{\Xi}(t) = (\hat{\xi}_1(t),\dots,\hat{\xi}_N(t),\hat{\eta}_1(t),\dots,\hat{\eta}_N(t))$, i.e.,

$$\hat{\xi}_j(t) = \hat{u}_j(t) + \Gamma^{-1}\dot{\hat{u}}_j(t), \quad \hat{\eta}_j(t) = \Gamma^{-1}\dot{\hat{u}}_j(t), \quad j = 1,\dots,N.$$

Consequently, $\hat{\Xi}(t)$ has the following property:

$$\hat{\Xi}(t+T) = \hat{\Xi}(t) + 2\pi\mathbf{e}, \tag{4.1}$$

where $\mathbf{e} = (\mathbf{1},0)^T$, and $\mathbf{1}$ denotes a vector in $\mathbb{R}^N$ with all components equal to 1. From Lemma 2.1 it follows that there exists an integer m such that $mT \in [T_0/2, T_0]$. Let

$$\ell = \{\hat{\Xi}(t) | t \in \mathbb{R}\} \quad \text{and} \quad Q(\Xi) = P^{mT}\Xi - 2m\pi\mathbf{e}.$$

The following properties will be used in the subsequent discussion.

(I) Q is strongly monotone in $I_\Gamma\Omega$, and each point in ℓ is a fixed point of the map Q.

(II) For any $\Xi \in \mathbb{R}^{2N}$, there exist $\Xi_1, \Xi_2 \in \ell$ such that $\Xi_1 \ll \Xi \ll \Xi_2$. Indeed, we can take $\Xi_1 = \hat{\Xi}(-nT)$ and $\Xi_2 = \hat{\Xi}(nT)$ with n large enough.

LEMMA 4.1. *$\hat{\Xi}(t_1) \ll \hat{\Xi}(t_2)$ if $t_1 < t_2$.*

Proof. First we show that every two points in ℓ are ordered by the partial order "$\leq$." Assume that there are two points $\hat{\Xi}(t_1)$ and $\hat{\Xi}(t_2)$ $(t_1 < t_2)$ which are unordered, i.e., $\hat{\Xi}(t_2) - \hat{\Xi}(t_1) \notin \Sigma$. From the above property (II) we know that there exists a t' such that $\hat{\Xi}(t') \ll \hat{\Xi}(t_2)$, i.e., $\hat{\Xi}(t_2) - \hat{\Xi}(t') \in \operatorname{int}\Sigma$. Increasing t', we deduce from the continuity that there exists t_0 such that $\hat{\Xi}(t_2) - \hat{\Xi}(t_0) \in \partial\Sigma$, implying by the strong monotonicity of Q that $Q\hat{\Xi}(t_2) - Q\hat{\Xi}(t_0) \in \operatorname{int}\Sigma$. This is a contradiction since $\hat{\Xi}(t_2)$ and $\hat{\Xi}(t_0)$ are fixed points of Q.

Now we show that $\hat{\Xi}(t_1) \ll \hat{\Xi}(t)$ for all $t > t_1$. The above property (II) implies that there exists t_2 such that $\hat{\Xi}(t_1) \ll \hat{\Xi}(t)$ for $t \geq t_2$. Let

$$t_0 = \sup\{t \geq t_1 \,|\, \hat{\Xi}(t) - \hat{\Xi}(t_1) \in \partial\Sigma\}.$$

It is evident that $t_0 = t_1$. Indeed, if $t_0 > t_1$, then $\hat{\Xi}(t_0) - \hat{\Xi}(t_1) \in \partial\Sigma$, which is impossible by the strong monotonicity of Q. Therefore, $\hat{\Xi}(t) - \hat{\Xi}(t_1) \in \operatorname{int}\Sigma$, i.e., $\hat{\Xi}(t_1) \ll \hat{\Xi}(t)$ for $t > t_1$. □

Proof of Theorem 1.1. Let

$$[\Xi_1, \Xi_2] = \{\Xi \,|\, \Xi_1 \leq \Xi \leq \Xi_2\} \quad \text{for } \Xi_1 \leq \Xi_2.$$

For each $\Xi_0 \in I_\Gamma\Omega$, by the above property (II), there always exist Ξ_1 and $\Xi_2 \in \ell$ such that $\Xi_1 \ll \Xi_2$ and $\Xi_0 \in [\Xi_1, \Xi_2]$. The strong monotonicity of Q leads to the conclusion that $\{Q^n\Xi_0\}_{n=1}^{+\infty} \subset [\Xi_1, \Xi_2]$ since $Q(\Xi_1) = \Xi_1$ and $Q(\Xi_2) = \Xi_2$. From the boundedness of the set $[\Xi_1, \Xi_2]$, it follows that the ω-limit set $\omega(\Xi_0)$ of $\{Q^n\Xi_0\}_{n=1}^{+\infty}$ is nonempty. Define for $n \geq 0$

$$\tau_-(n; \Xi_0) = \sup\{\tau \in \mathbb{R} \,|\, \hat{\Xi}(\tau) \leq Q^n(\Xi_0)\},$$

$$\tau_+(n; \Xi_0) = \inf\{\tau \in \mathbb{R} \,|\, Q^n(\Xi_0) \leq \hat{\Xi}(\tau)\}.$$

From Lemma 4.1 and the strong monotonicity of Q we know that $\tau_-(n; \Xi_0)$ is strictly increasing, $\tau_+(n; \Xi_0)$ is strictly decreasing with respect to n, and $\tau_-(n; \Xi_0) \leq \tau_+(n; \Xi_0)$. Moreover, $\tau_\pm(n; \Xi_0)$ are continuous with respect to Ξ_0. If $\tau_-(0; \Xi_0) = \tau_+(0; \Xi_0) = \tau$, then $\Xi_0 = \hat{\Xi}(\tau)$ and $\Xi_0(t) = \hat{\Xi}(t + \tau)$, $t \in \mathbb{R}$, where $\Xi_0(t)$ is a solution of (3.2) with initial value Ξ_0 at $t = 0$. If $\tau_-(0; \Xi_0) < \tau_+(0; \Xi_0)$, let

$$\tau_-^\infty = \sup_{n\geq 0}\{\tau_-(n; \Xi_0)\}, \quad \tau_+^\infty = \inf_{n\geq 0}\{\tau_+(n; \Xi_0)\}.$$

Then $\tau_-^\infty \leq \tau_+^\infty$. If $\tau_-^\infty < \tau_+^\infty$, denote a limit point of $\{Q^n\Xi_0\}$ by $\tilde{\Xi}_0$. Then $\tau_\pm(0; \tilde{\Xi}_0) = \tau_\pm^\infty$, and

$$\tau_-^\infty = \tau_-(0; \tilde{\Xi}_0) < \tau_-(1; \tilde{\Xi}_0) \leq \tau_+(1; \tilde{\Xi}_0) < \tau_+(0; \tilde{\Xi}_0) = \tau_+^\infty.$$

Take k large enough such that $Q^k(\Xi_0)$ is sufficiently close to $\tilde{\Xi}_0$. Then we have

$$\tau_-^\infty < \tau_-(1; Q^k\Xi_0) \leq \tau_+(1; Q^k\Xi_0) < \tau_+^\infty,$$

i.e.,

$$\tau_-^\infty < \tau_-(k+1; \Xi_0) \leq \tau_+(k+1; \Xi_0) < \tau_+^\infty,$$

implying a contradiction. Consequently, $\tau_-^\infty = \tau_+^\infty \triangleq \tau$. We deduce that $|\Xi_0(nmT) - \hat{\Xi}(nmT + \tau)| \to 0$ as $n \to +\infty$, leading to the conclusion $|\Xi_0(t) - \hat{\Xi}(t + \tau)| \to 0$ as $t \to +\infty$. □

Appendix. Proof of Lemma 3.1. Since $u_j(t)$ in the right-hand side of (3.6) depends upon the external driving force F, so does the solution $w(t)$ of (3.6). If $w(0) = 1$, then since $\dot{w}|_{w=1} = -\Gamma < 0$, it is easy to check that $w(t) < 1$ for $t > 0$. Now we assume that $w(0) = -1$. First we show by contradiction that $1 - \frac{\Gamma^2}{4K} < w(t) < 1$ for $t \in [0, T_0]$ if F is large enough.

Assume that there exists $\hat{t} \in [0, T_0]$ such that $w(\hat{t}) \le 1 - \frac{\Gamma^2}{4K} < -1$. Taking

$$t_1 = \inf\left\{ t \,\middle|\, t \in (0, \hat{t}], w(t) = 1 - \Gamma^2/(4K) \right\} \quad \text{and} \quad t_0 = \sup\{ t \,|\, t \in [0, t_1), w(t) = -1 \}$$

yields that $w(t) \in [1 - \frac{\Gamma^2}{4K}, -1]$ for $t \in [t_0, t_1]$. Note that the function $f(x) = -\Gamma^2 x - 2K(1-x)^2$ is positive on the interval $[1 - \frac{\Gamma^2}{4K}, -1]$. It follows that

$$w(t_1) = w(t_0) + \int_{t_0}^{t_1} \frac{1}{\Gamma} \left[-\Gamma^2 w(t) - 2K(1 - w(t))^2 - \cos u_j(t)(1 - w(t))^2 \right] dt$$

$$> -1 - \frac{1}{\Gamma} \left| \int_{t_0}^{t_1} \cos u_j(t)(1 - w(t))^2 dt \right|.$$

Note also that

$$\int_{t_0}^{t_1} \cos u_j(t)(1 - w(t))^2 dt = \int_{t_0}^{t_1} (1 - w(t))^2 d \left(\int_{t_0}^{t} \cos u_j(\tau) d\tau \right)$$

$$= (1 - w(t))^2 \int_{t_0}^{t} \cos u_j(\tau) d\tau \Big|_{t_0}^{t_1} + 2 \int_{t_0}^{t_1} \left[\int_{t_0}^{t} \cos u_j(\tau) d\tau \right] (1 - w(t)) \dot{w}(t) dt.$$

Since $[t_0, t_1] \in [0, T_0]$, we deduce from Lemma 2.4 that

$$\left| (1 - w(t))^2 \int_{t_0}^{t} \cos u_j(\tau) d\tau \Big|_{t_0}^{t_1} \right| \le c(\Gamma^2/4K)^2. \tag{A.1}$$

From the fact

$$\max_{x \in [1 - \frac{\Gamma^2}{4K}, -1]} f(x) = \Gamma^2 \left(\frac{\Gamma^2}{8K} - 1 \right),$$

it follows that for $t \in [t_0, t_1]$,

$$|\dot{w}| = \frac{1}{\Gamma} \left| -\Gamma^2 w - 2K(1 - w)^2 - \cos u_j (1 - w)^2 \right|$$

$$\le \frac{1}{\Gamma} \left| -\Gamma^2 w - 2K(1 - w)^2 \right| + \frac{1}{\Gamma} \left| \cos u_j (1 - w)^2 \right|$$

$$\le \Gamma \left(\frac{\Gamma^2}{8K} - 1 \right) + \frac{1}{\Gamma} \left(\frac{\Gamma^2}{4K} \right)^2,$$

and hence

$$\begin{aligned}
&\left|\int_{t_0}^{t_1}\left[\int_{t_0}^{t}\cos u_j(\tau)d\tau\right](1-w(t))\dot{w}(t)dt\right| \\
&\leq \sup_{t\in[t_0,t_1]}\left\{|1-w(t)|\cdot|\dot{w}(t)|\cdot\left|\int_{t_0}^{t}\cos u_j(\tau)d\tau\right|\right\}\cdot(t_1-t_0) \\
&\leq \frac{c\,T_0\Gamma^2}{4K}\left[\Gamma\left(\frac{\Gamma^2}{8K}-1\right)+\frac{1}{\Gamma}\left(\frac{\Gamma^2}{4K}\right)^2\right]. \qquad \text{(A.2)}
\end{aligned}$$

Combining $(A.1)$ and $(A.2)$ yields

$$\begin{aligned}
&\frac{1}{\Gamma}\left|\int_{t_0}^{t_1}\cos u_j(t)(1-w(t))^2dt\right| \\
&< \frac{c}{\Gamma}\left(\frac{\Gamma^2}{4K}\right)^2+\frac{c\,T_0\Gamma}{2K}\left[\Gamma\left(\frac{\Gamma^2}{8K}-1\right)+\frac{1}{\Gamma}\left(\frac{\Gamma^2}{4K}\right)^2\right]=c\,\Delta, \qquad \text{(A.3)}
\end{aligned}$$

where

$$\Delta=\frac{1}{\Gamma}\left(\frac{\Gamma^2}{4K}\right)^2+\frac{T_0}{2K}\left[\Gamma^2\left(\frac{\Gamma^2}{8K}-1\right)+\left(\frac{\Gamma^2}{4K}\right)^2\right].$$

From (2.6) and (2.7) it follows that $c\to 0$ as $F\to+\infty$. Choosing F large enough, we have $c\,\Delta<\frac{\Gamma^2}{4K}-2$ and hence a contradiction with $w(t_1)=1-\frac{\Gamma^2}{4K}$. Consequently, there exists $F_2\geq F_1$ such that $1-\frac{\Gamma^2}{4K}<w(t)<1$ for $t\in[0,T_0]$ if $F\geq F_2$.

Now we construct a first order equation

$$\dot{s}=\Gamma^{-1}\left[-\Gamma^2 s-2K(1-s)^2\right].$$

It is easy to check by the assumption $\Gamma>2\sqrt{2K}$ that $s(t)\in(-1,1)$ for $t>0$ if $s(0)\in[-1,1]$. Assume that $s(0)=w(0)=-1$. Then for $t\in[0,T_0]$ it follows that

$$\begin{aligned}
|w(t)-s(t)|\leq{}&\frac{1}{\Gamma}\left|\int_0^t\left(2K(w(\tau)+s(\tau)-2)+\Gamma^2\right)(w(\tau)-s(\tau))d\tau\right| \\
&+\frac{1}{\Gamma}\left|\int_0^t\cos u_j(\tau)(1-w(\tau))^2d\tau\right|.
\end{aligned}$$

Since $s(t)\in(-1,1)$ and $1-\frac{\Gamma^2}{4K}<w(t)<1$ for $t\in(0,T_0]$, then we have

$$\left|\int_0^t\left(2K(w(\tau)+s(\tau)-2)+\Gamma^2\right)(w(\tau)-s(\tau))d\tau\right|\leq\int_0^t\Gamma^2\,|w(\tau)-s(\tau)|\,d\tau, \qquad \text{(A.4)}$$

and hence by (A.3) and (A.4),

$$|w(t)-s(t)|\leq\int_0^t\Gamma\,|w(\tau)-s(\tau)|\,d\tau+c\,\Delta.$$

The Gronwall inequality implies that

$$|w(t) - s(t)| \le c\,\Delta e^{\Gamma t} \le c\,\Delta e^{\Gamma T_0}, \quad t \in [0, T_0].$$

As a consequence, there exists $F_3 \ge F_1$, such that if $F \ge F_3$, then

$$c\,\Delta e^{\Gamma T_0} < s(T_0/2) + 1,$$

and hence $w(t) > -1$ for $t \in [T_0/2, T_0]$. Meanwhile, since $w(T_0) > -1$, then it follows that $1 - \Gamma^2/(4K) < w(t) < 1$ for $t \in [T_0, 2T_0]$ if $F \ge F_2$, and thus $1 - \Gamma^2/(4K) < w(t) < 1$ for all $t > 0$. The proof is completed by taking $F_0 = \max\{F_2, F_3\}$. □

We should mention here that the above proof was highly inspired by Levi's idea on dealing with a single oscillator [10].

REFERENCES

[1] D. G. Aronson, M. Golubitsky, and J. Mallet-Paret, *Ponies on a merry-go-round in large arrays of Josephson junctions*, Nonlinearity, 4 (1991), pp. 903–910.

[2] D. G. Aronson and Y. S. Huang, *Single wave form solutions for linear arrays of Josephson junctions*, Phys. D, 101 (1997), pp. 157–177.

[3] C. Baesens and R. S. MacKay, *Gradient dynamics of tilted Frenkel-Kontorova models*, Nonlinearity, 11 (1998), pp. 949–964.

[4] C. Baesens and R. S. MacKay, *A novel preserved partial order for cooperative networks of units with overdamped second order dynamics, and application to tilted Frenkel-Kontorova chains*, Nonlinearity, 17 (2004), pp. 567–580.

[5] O. M. Braun and Y. S. Kivshar, *The Frenkel-Kontorova Model. Concepts, Methods, and Applications*, Springer-Verlag, Berlin, 2004.

[6] L. M. Floría and J. J. Mazo, *Dissipative dynamics of the Frenkel-Kontorova model*, Adv. Phys., 45 (1996), pp. 505–598.

[7] G. Katriel, *Existence of travelling waves in discrete sine-Gordon rings*, SIAM J. Math. Anal., 36 (2005), pp. 1434–1443.

[8] M. Levi, *Caterpillar solutions in coupled pendula*, Ergodic Theory Dynam. Systems, 8 (1988), pp. 153–174.

[9] M. Levi, *Dynamics of discrete Frenkel-Kontorova models*, in Analysis, Et Cetera, P. Rabinowitz and E. Zehnder, eds., Academic Press, Boston, 1990, pp. 471–494.

[10] M. Levi, *private communications*.

[11] R. E. Mirollo, *Splay-phase orbits for equivariant flows on tori*, SIAM J. Math. Anal., 25 (1994), pp. 1176–1180.

[12] R. Mirollo and N. Rosen, *Existence, uniqueness, and nonuniqueness of single-wave-form solutions to Josephson junction systems*, SIAM J. Appl. Math., 60 (2000), pp. 1471–1501.

[13] M. Qian, S. Zhu, and W.-X. Qin, *Dynamics in a chain of overdamped pendula driven by constant torques*, SIAM J. Appl. Math., 57 (1997), pp. 294–305.

[14] H. L. Smith, *Monotone Dynamical Systems*, AMS, Providence, RI, 1995.

[15] S. Watanabe and J. W. Swift, *Stability of periodic solutions in series arrays of Josephson junctions with internal capacitance*, J. Nonlinear Sci., 7 (1997), pp. 503–536.

[16] S. Watanabe, H. S. J. van der Zant, S. H. Strogatz, and T. P. Orlando, *Dynamics of circular arrays of Josephson junctions and the discrete sine-Gordon equation*, Phys. D, 97 (1996), pp. 429–470.

SIAM J. MATH. ANAL.
Vol. 40, No. 3, pp. 968–983

ERGODICITY AND RATE OF CONVERGENCE FOR A NONSECTORIAL FIBER LAY-DOWN PROCESS*

MARTIN GROTHAUS† AND AXEL KLAR†

Abstract. A stochastic model for the lay-down of fibers on a conveyor belt in the production process of nonwovens is investigated. In particular, convergence of the stochastic process to the stationary solution is proven and estimates on the speed of convergence are given. Numerical results and examples are presented and compared with the analytical estimates on the speed of convergence.

Key words. fiber dynamics, Fokker–Planck equations, Dirichlet forms, ergodicity, rate of convergence

AMS subject classifications. 37A25, 47D07, 82C31

DOI. 10.1137/070697173

1. Introduction.

The understanding of the forms generated by the lay-down of flexible fibers onto a moving conveyor belt is of great interest in the production process of nonwovens that find their applications, e.g., in composite materials (filters), textiles, and the hygiene industry. In the melt-spinning process of nonwoven materials, hundreds of individual endless fibers obtained by the continuous extrusion of a melted polymer are stretched and entangled by highly turbulent air flows to finally form a web on the conveyor belt. The quality of this web and the resulting nonwoven material—in terms of homogeneity and load capacity—depends essentially on the dynamics and the deposition of the fibers.

For the description of the interaction between fibers and turbulent flow a stochastic force model is derived and analyzed in [MW06]. Applying this concept, the fiber fabric can in principle be numerically generated and its quality investigated. However, these or similar simulations usually lead to excessively large computation times, when all physical details of the production process are considered. Thus, simplified models for the lay-down process are needed. In particular, this is true for optimization and control procedures where many different simulations are needed. In [GKM+07] a new simplified stochastic model for the fiber lay-down process, i.e., for the generation of a fiber web on a conveyor belt, has been presented. The process, which takes into account the fiber motion under the influence of turbulence, is described by a stochastic differential system; see section 2.

The solution of the associated Fokker–Planck equation gives the density of the stochastic process. An important criterion for the quality of the web and the resulting nonwoven material is how it converges to equilibrium. In particular, the speed of convergence to the stationary solution is important. The faster this convergence is, the more uniform the produced textile will be. This means that process parameters should be adjusted such that the speed of convergence to equilibrium is optimal.

The trend to equilibrium for Fokker–Planck-type equations has been investigated in many papers. In [AMTU01] a unified presentation of entropy methods for

*Received by the editors July 13, 2007; accepted for publication (in revised form) May 27, 2008; published electronically September 19, 2008. This work was supported by the Kaiserslautern Excellence Cluster *Dependable Adaptive Systems and Mathematical Modeling*.
http://www.siam.org/journals/sima/40-3/69717.html
†Mathematics Department, University of Kaiserslautern, P.O. Box 3049, 67653 Kaiserslautern, Germany (grothaus@mathematik.uni-kl.de, klar@mathematik.uni-kl.de).

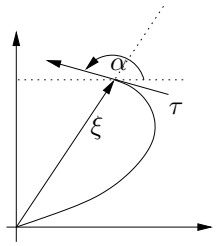


FIG. 1. *Fiber scenario on the conveyor belt.*

nondegenerate linear and nonlinear Fokker–Planck-type equations is given. In [DV01] the linear kinetic Fokker–Planck equation is discussed. In this case the degeneracy of the diffusion operator is similar to the present case. A more general theory for degenerate linear and nonlinear problems has been developed in [Vil06]. A more detailed discussion of these methods with respect to the present problem is given in the next section.

As in the above-mentioned papers, the purpose of the present paper is to investigate analytically the convergence to equilibrium. In particular, we aim at obtaining explicit estimates on the rate of convergence. However, instead of using entropy methods we use here Dirichlet form and operator semigroup techniques.

The paper is organized as follows. In section 2 we present the stochastic model for the fiber dynamics. Furthermore, we relate our convergence result for the solution of the stochastic differential equation with the convergence of the solution of the associated Fokker–Planck equation and outline our strategy for proving ergodicity. In section 3 we prove existence results for the stochastic process as well as several other properties of the process which are important for our analysis. Section 4 investigates several properties of an associated nondegenerate Ornstein–Uhlenbeck-type process. Section 5 proves the ergodicity of the original process and gives explicit bounds on the rates to convergence. Section 6 is devoted to some numerical results comparing the analytical rates of convergence with the numerical ones.

2. The stochastic model for the fiber lay-down process and its associated Fokker–Planck equation. Focusing on a single slender elastic inextensible fiber in a lay-down process, the fiber on the nonmoving conveyor belt can be described by an arc-length parameterized curve $\xi : \mathbb{R}_0^+ \to \mathbb{R}^2$, as visualized in Figure 1.

Due to its inextensibility, $\|\partial_t \xi\| = 1$ holds. The web-forming is modeled as

$$\partial_t \xi = \tau(\alpha),$$
$$\partial_t \alpha = -\nabla\phi(\xi) \cdot \tau^\perp(\alpha),$$

where $\tau(\alpha) = (\cos\alpha, \sin\alpha)^T$ denotes the normalized tangent on the fiber. Since a curved fiber tends back to its starting point, the change of the angle α is assumed to be proportional to $\nabla\phi(\xi)\cdot\tau^\perp(\alpha)$ with $\tau^\perp(\alpha) = (-\sin\alpha, \cos\alpha)^T$. The potential of this drive is prescribed by a function $\phi : \mathbb{R}^2 \to \mathbb{R}$ with the generic example $\phi(\xi) = \|\xi\|^2/2$. In general, the potential depends on the physical process and needs to be adapted to the experimental parameters.

Considering a turbulent flow in the deposition region of the fiber close to the conveyor belt, the fiber lay-down is additionally affected by a stochastic force that can be modeled by a Wiener process B_t in $\mathbb{R}$ with amplitude σ.

The resulting stochastic differential system reads

$$\mathrm{d}\xi_t = \tau(\alpha_t)\,\mathrm{d}t, \tag{2.1}$$

$$\mathrm{d}\alpha_t = -\nabla\phi(\xi_t)\cdot\tau^{\perp}(\alpha_t)\,\mathrm{d}t + \sigma\,\mathrm{d}B_t, \tag{2.2}$$

where the conservation of length $\|\partial_t\xi\| = 1$ is still valid. For details on the mathematical modeling of the process we refer to [GKM+07]. An illustration of the pathwise behavior of the process for different σ and $\phi(\xi) = \|\xi\|^2/2$ is displayed in Figure 2, where the paths of Monte Carlo simulations for different σ are shown. The slow convergence to equilibrium is clearly seen for small and large values of σ.

The density of the above stochastic process can be found by solving the associated Fokker–Planck equation:

$$\partial_t u + \cos(\alpha)\partial_{\xi^1} u + \sin(\alpha)\partial_{\xi^2} u - \partial_\alpha\left(\nabla\phi(\xi)\begin{pmatrix} -\sin(\alpha) \\ \cos(\alpha)\end{pmatrix} u\right) = \frac{\sigma^2}{2}\partial_\alpha^2 u. \tag{2.3}$$

A stationary solution of (2.3) is the function

$$(\xi,\alpha)\mapsto m(\xi) := \frac{1}{N}\exp(-\phi(\xi)), \quad N > 0.$$

For $\phi(\xi) = \|\xi\|^2/2$ we obtain the standard Gaussian as stationary distribution. By μ we denote the measure on $E = \mathbb{R}^2 \times [0, 2\pi]$ having density m with respect to (w.r.t.) the Lebesgue measure. Except for sections 3 and 4 we consider only potentials ϕ such that m is integrable and choose N as a normalizing constant to obtain a probability measure μ. In the following sections we are concerned with the approach to equilibrium of the stochastic process (2.1) or the convergence to the stationary solution of (2.3), respectively.

Remark 2.1. A formal argument for convergence to equilibrium of the solution of (2.3) is given by the following computations; cf. [DV01]. Define the relative entropy

$$H(u/m) = \int_{\mathbb{R}^2}\int_0^{2\pi} u\log\left(\frac{u}{m}\right) d\alpha d\xi.$$

An easy computation gives

$$\partial_t H(u/m) = -\int_{\mathbb{R}^2}\int_0^{2\pi} u\left(\partial_\alpha\log\left(\frac{u}{m}\right)\right)^2 d\alpha d\xi.$$

This vanishes if and only if

$$u = \rho(t,\xi)m(\xi).$$

That means u converges formally to a local equilibrium. Plugging $u = \rho(t,\xi)m(\xi)$ into the evolution equation, one obtains

$$m\partial_t\rho + \cos(\alpha)\left(\partial_{\xi^1}(\rho m) + \rho m\partial_{\xi^1}\phi\right) + \sin(\alpha)\left(\partial_{\xi^2}(\rho m) + \rho m\partial_{\xi^2}\phi\right) = 0$$

or

$$\partial_t\rho = 0, \quad \partial_{\xi^1}\rho = 0, \quad \partial_{\xi^2}\rho = 0.$$

This means $u = m(\xi)$, i.e., convergence to global equilibrium.

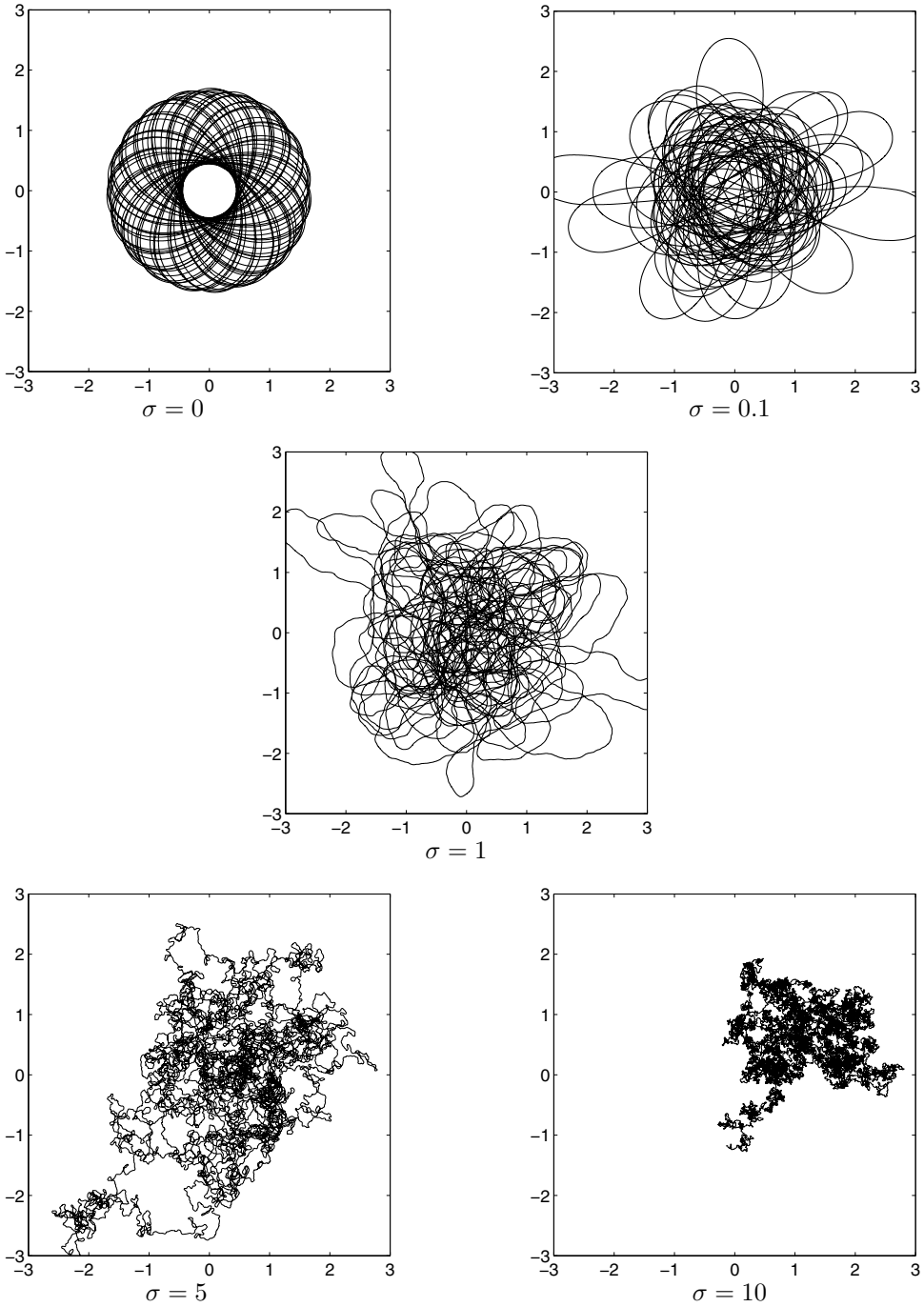


FIG. 2. *Representative path behavior for balanced ($\sigma = 1$) as well as deterministic ($\sigma < 1$) and stochastic ($\sigma > 1$) dominated (ξ, α)-systems.*

In [Vil06] the arguments described in the above remark were rigorously justified for similar equations. To apply these methods in the context of the present paper among other conditions, the existence of proper Liapunov functions would be necessary. Their existence is another open problem. In the following we choose a different route and consider the stochastic differential equation (2.1) directly.

In section 3 we construct a diffusion process solving (2.1) and having μ as an invariant measure. The law of this process we denote by $\mathbf{P}_\mu$. Then in section 5 we prove that this diffusion is ergodic with rate of convergence

$$\begin{aligned}(2.4)\quad &\Big\|\frac{1}{t}\int_0^t f(\mathbf{X}_s)\,ds - \mathbf{E}_\mu[f]\Big\|_{L^2(\mathbf{P}_\mu)}\\ &\le \frac{1}{c^{1/2}}\Big(\frac{2}{t}+\frac{1}{t^{1/2}}\Big(\frac{A(C)c^{1/2}+B(C)c^{-1/2}}{\sigma}+(1+2^{-1/2})\sigma\Big)\Big)\|f-\mathbf{E}_\mu[f]\|_{L^2(\mu)},\quad t>0;\end{aligned}$$

see Theorem 5.3 for details. The convergence in (2.4) implies mean ergodicity of the corresponding semigroup $(T_t)_{t\ge 0}$, i.e.,

$$\begin{aligned}(2.5)\quad &\Big\|\frac{1}{t}\int_0^t T_s f\,ds - \mathbf{E}_\mu[f]\Big\|_{L^2(\mu)}\\ &\le \frac{1}{c^{1/2}}\Big(\frac{2}{t}+\frac{1}{t^{1/2}}\Big(\frac{A(C)c^{1/2}+B(C)c^{-1/2}}{\sigma}+(1+2^{-1/2})\sigma\Big)\Big)\|f-\mathbf{E}_\mu[f]\|_{L^2(\mu)},\quad t>0.\end{aligned}$$

In the case of an analytic semigroup from (2.5) we even could conclude strongly mixing, i.e.,

$$\lim_{t\to\infty}\|T_t f-\mathbf{E}_\mu[f]\|_{L^2(\mu)}=0;$$

see, e.g., [Gol85, Exer. 8.24.17]. In our case, however, the corresponding generator

$$(2.6)\qquad L=S+A,\quad S=\frac{\sigma^2}{2}\partial_\alpha^2,\quad A=\cos(\alpha)\partial_{\xi^1}+\sin(\alpha)\partial_{\xi^2}-\nabla\phi(\xi)\cdot\tau^\perp(\alpha)\partial_\alpha$$

is nonsectorial. Hence, $(T_t)_{t\ge 0}$ cannot be the restriction of an analytic semigroup; see, e.g., [Gol85, Thm. 5.9].

However, since the adjoint to L w.r.t. the scalar product in $L^2(\mu)$ is given by $L^*=S-A$, then also for the adjoint process and semigroup we have ergodicity with the same rate of convergence.

Now, strong mixing of $(T_t^*)_{t\ge 0}$ would imply L^1-convergence of the solution u of the associated Fokker–Planck equation (2.3) with normalized nonnegative initial distribution $u(0)=f$, because

$$\|u(t)-m\|_{L^1(dx)}=\Big\|\frac{u(t)}{m}-1\Big\|_{L^1(\mu)}=\|T_t^*f-1\|_{L^1(\mu)}\le\|T_t^*f-1\|_{L^2(\mu)}.$$

Although $(L,D(L))$ is nonsectorial, we are convinced that for C^∞-potentials ϕ we have strong mixing by another reasoning not worked out rigorously in the present paper. Since the generator $(L,D(L))$ is hypoelliptic in the sense of Hörmander, it is not too hard to show that $(T_t)_{t\ge 0}$ is strong Feller. Showing that $(T_t)_{t\ge 0}$ additionally is irreducible, strong mixing then follows from Doob's theorem; see, e.g., [DPZ96, Prop. 4.1.1, Thm. 4.2.1]. But here we would like to stress that this does not give an explicit rate of convergence. Moreover, our approach also applies to C^3-potentials ϕ.

That $((T_t)_{t\geq 0})$ is weak mixing, i.e.,

$$\lim_{t\to\infty, t\in I} \|T_t f - \mathbf{E}_\mu[f]\|_{L^2(\mu)} = 0,$$

where $I \subset \mathbb{R}_0^+$ has relative measure 1, follows from weaker spectral properties of $(L, D(L))$ than those giving rise to a corresponding analytic semigroup. Sufficient for $(T_t)_{t\geq 0}$ being weak mixing is the following condition:

Let $f \in L^2(\mu; \mathbb{C})$ and $\lambda \in \mathbb{R}$. If

$$T_t f = \exp(i\lambda t) f \quad \text{for all} \quad t \geq 0, \tag{2.7}$$

then $\lambda = 0$ and f is constant; see, e.g., [DPZ96, Thm. 3.4.1]. Here of course we have to consider the corresponding complexified spaces.

The condition in (2.7) implies that $f \in D(L)$ is an eigenvector to the eigenvalue $i\lambda$, $\lambda \in \mathbb{R}$. If f is now sufficiently smooth, such that we can apply the operators S and A from (2.6) separately, then this condition easily can be verified (S is a symmetric and A an antisymmetric operator in $L^2(\mu)$). Hence it is left to show that eigenvectors to $(L, D(L))$ are sufficiently smooth. This is not clear a priori, since S is degenerated. But $L = S + A$ is of Hörmander form such that we expect to obtain hypoelliptic estimates.

The interplay of S and A is very crucial for our proof of ergodicity. Only the operator S without A would not give ergodic behavior. The idea is to project onto the orthogonal complement of the kernel of $(S, D(S))$. On this subspace $(S, D(S))$ has a bounded inverse. On the kernel of $(S, D(S))$ in turn, L can be associated with a nondegenerated, self-adjoint operator $(G, D(G))$ in $L^2(\gamma)$, where γ is the marginal measure of μ on $\mathbb{R}^2$. Assuming that the Dirichlet form corresponding to $(G, D(G))$ fulfills a Poincaré inequality, finally ergodicity can be shown. Such ideas have been used before in, e.g., [OT03] in the context of scaling limit for interacting particles systems.

Crucial for the rate of convergence is Lemma 4.1. There Kato perturbation techniques and a ground state transform are used to show that the functions $h = \nabla_\xi g \cdot \tau$, $g \in D(G)$, are in $D(L)$. These functions are our "key functions" for relating $(L, D(L))$ with $(G, D(G))$; see the proof of Theorem 5.3. The explicit estimates obtained in the proof of Lemma 4.1 finally provide us with the constants for our rate of convergence (2.4).

3. Stochastic differential equation and its solution. We are considering the following stochastic differential equation in $E := \mathbb{R}^2 \times [0, 2\pi]$:

$$\begin{aligned} d\xi_t &= \tau(\alpha_t)\, dt, \\ d\alpha_t &= -\nabla\phi(\xi_t) \cdot \tau^\perp(\alpha_t)\, dt + \sigma\, dB_t, \quad \sigma > 0, \quad t \geq 0. \end{aligned} \tag{3.1}$$

Using Itô's formula we find the generator of the corresponding Markov process:

$$L = S + A, \quad S = \frac{\sigma^2}{2}\partial_\alpha^2, \quad A = \cos(\alpha)\partial_{\xi^1} + \sin(\alpha)\partial_{\xi^2} - \nabla\phi(\xi) \cdot \tau^\perp(\alpha)\partial_\alpha.$$

Set

$$D := C^\infty_{0,\mathrm{pb}}(E) := \{f|_E \,|\, f \in C_0^\infty(\mathbb{R}^3),\ f(\xi, 0) = f(\xi, 2\pi) \text{ for all } \xi \in \mathbb{R}^2 \text{ and the same holds for all derivatives of } f\},$$

where $C_0^\infty(\mathbb{R}^3)$ denotes the set of compactly supported, infinitely differentiable functions on $\mathbb{R}^3$. (S, D) is symmetric and (A, D) is antisymmetric w.r.t. the scalar product in $L^2(\mu)$, where μ is the measure on E having density m w.r.t. the Lebesgue measure.

In the following theorem $H^{1,\infty}(\mathbb{R}^2)$ denotes the Sobolev space of weakly differentiable functions on $\mathbb{R}^2$, which are essentially bounded together with their derivative. The adjoint operator $(L^*, D(L^*))$ to (L, D) obviously is a closed extension of $(S - A, D)$.

LEMMA 3.1. *Let $\phi \in H^{1,\infty}(\mathbb{R}^2)$. Then (L, D) and $(L^*|_D, D)$ are essentially m-dissipative Dirichlet operators in $L^2(\mu)$. In particular, $(L^*, D(L^*))$ is the m-dissipative extension of $(L^*|_D, D)$. The m-dissipative extension (closure) of (L, D) we denote by $(L, D(L))$.*

Proof. Since (S, D) is dissipative and (A, D) is antisymmetric, obviously

$$(Lf, f)_{L^2(\mu)} \leq 0 \quad \text{for all} \quad f \in D,$$

i.e., (L, D) is dissipative. Moreover, as in, e.g., [CG08a] one can show that

$$(Lf, (f-1)^+)_{L^2(\mu)} \leq 0 \quad \text{for all} \quad f \in D,$$

i.e., (L, D) is a Dirichlet operator (here $g^+ := (|g| + g)/2$ for a mapping $g : E \to \mathbb{R}$). Analogously one obtains that $(L^*|_D, D)$ is also a dissipative Dirichlet operator in $L^2(\mu)$. Hence it is left to show that D is a core for the m-dissipative extensions of (L, D) and $(L^*|_D, D)$.

First we consider the case $\phi = 0$. Then we have

$$Lf = Sf + \cos(\alpha)\partial_{\xi^1} f + \sin(\alpha)\partial_{\xi^2} f, \quad f \in D.$$

In the following we show that in this case L is essentially m-dissipative on

$$F := S(\mathbb{R}^2) \times C^\infty_{\mathrm{pb}}([0, 2\pi]),$$

where $S(\mathbb{R}^2)$ denotes the space of infinitely differentiable, more than polynomial decreasing functions on $\mathbb{R}^2$. Since $D \subset F$ is dense in graph norm, this implies that (L, D) is essentially m-dissipative.

Consider the complexified setting, i.e., the corresponding spaces

$$F_{\mathbb{C}} := S(\mathbb{R}^2; \mathbb{C}) \times C^\infty_{\mathrm{pb}}([0, 2\pi]; \mathbb{C}) \subset L^2(\mu; \mathbb{C}),$$

of complex valued functions, and let L act componentwise. Then we apply the Fourier transform $\mathcal{F}$ in the first variable and obtain the corresponding operator

$$\hat{L} := \mathcal{F} L \mathcal{F}^{-1} = Sf + i\cos(\alpha)\xi_1 + i\sin(\alpha)\xi_2.$$

Recall that $\mathcal{F} : F_{\mathbb{C}} \to F_{\mathbb{C}}$ is a continuous isomorphism and $\mathcal{F} : L^2(\mu; \mathbb{C}) \to L^2(\mu; \mathbb{C})$ is a unitary isomorphism. Furthermore, by an elementary consideration one shows that $F_{\mathbb{C}}$ is a core for the m-dissipative extension of $(\hat{L}, F_{\mathbb{C}})$. Hence, L is essentially m-dissipative on F.

Finally the case $\phi \in H^{1,\infty}(\mathbb{R}^2)$ follows by a standard Kato perturbation technique. The adjoint operator can be treated analogously. ∎

Since $(L, D(L))$ is m-dissipative, it generates a semigroup of contractions $(T_t)_{t\geq 0}$. Let $(T_t^*)_{t\geq 0}$ denote the dual semigroup. It is well known that its generator is $(L^*, D(L^*))$. The corresponding stochastic process we can construct as a diffusion process

(i.e., Markov process with continuous sample path) only on the manifold $\mathbb{R}^2 \times S^1$ which is naturally associated with E. Here S^1 denotes the unit circle. Since the measure μ and above function spaces naturally can be lifted on $\mathbb{R}^2 \times S^1$, in what follows we simply identify

$$E = \mathbb{R}^2 \times S^1.$$

By γ we denote the marginal measure of μ on $\mathbb{R}^2$.

THEOREM 3.2. *Let m be bounded and continuous, $m > 0$ almost everywhere w.r.t. the Lebesgue measure on $\mathbb{R}^2$, and ϕ weakly differentiable on $\{m > 0\}$ with $\nabla\phi \in L^2(\gamma)$. Furthermore, let $\nabla\phi$ be bounded on the sets $\{\phi \le K\}$ for all $K \in \mathbb{R}$. Then for any probability measure ν which is absolutely continuous and has a bounded density w.r.t. μ there exists a Markov process $(C([0,\infty); E), \mathcal{B}, (\mathbf{F}_t)_{t\ge 0}, (\mathbf{X}_t)_{t\ge 0}, \mathbf{P}_\nu)$ having the following properties:*

(i) *The initial distribution of the process is given by ν. If μ is a probability measure, then it is an invariant measure for the process.*

(ii) *Finite-dimensional distributions of $\mathbf{P}_\nu$ are given in terms of a contraction semigroup $(T_t)_{t\ge 0}$ having as generator $(L, D(L))$ an m-dissipative extension of (L, D). This determines the law $\mathbf{P}_\nu$ uniquely.*

(iii) *The process solves* (3.1) *in the sense of the corresponding martingale problem, i.e., for all $f \in D$,*

$$\mathbf{M}(f)_t = f(\mathbf{X}_t) - f(\mathbf{X}_0) - \int_0^t Lf(\mathbf{X}_s)\, ds, \quad t \ge 0,$$

is an $\mathbf{F}_t$-martingale under $\mathbf{P}_\nu$.

Analogous statements hold for the adjoint operators. The associated stochastic process $(C([0,\infty); E), \mathcal{B}, (\mathbf{F}_t)_{t\ge 0}, (\mathbf{X}_t)_{t\ge 0}, \mathbf{P}^*_\nu)$ is called the adjoint process.

Proof. First we consider $\phi \in H^{1,\infty}(\mathbb{R}^2)$. Then from Lemma 3.1 we can conclude that $(L, D(L))$ is an m-dissipative Dirichlet operator with core D. Hence $(L, D(L))$ generates a quasi-regular, generalized Dirichlet form and therefore has an associated Markov process; see [Sta99].

For a general ϕ a corresponding stochastic process can be constructed as in [CG08b, sec. 2.3] by an approximation via Markov process associated to potentials from $H^{1,\infty}(\mathbb{R}^2)$.

Then as in [CG08b] and [CG08a], respectively, one can show that it is a diffusion with the properties (i)–(iii). ∎

Remark 3.3. (i) For $\phi \in H^{1,\infty}(\mathbb{R}^2)$ one can obtain much stronger existence results, as shown in [CG08a], where the concepts of generalized Dirichlet forms were applied to a similar problem.

(ii) The process $(\mathbf{X}_t)_{t\ge 0}$ in Theorem 3.2 is the coordinate process, i.e., $X_t(\omega) = \omega(t)$, $t \ge 0$, $\omega \in C([0,\infty); E)$, $\mathcal{B}$ is the Borel σ-algebra, and $(\mathbf{F}_t)_{t\ge 0}$ is the corresponding natural filtration.

COROLLARY 3.4. *Let the assumptions of Theorem* 3.2 *hold and let μ be a probability measure. Then*

$$\mathbf{M}(f)_t = f(\mathbf{X}_t) - f(\mathbf{X}_0) - \int_0^t Lf(\mathbf{X}_s)\, ds, \quad t \ge 0, \tag{3.2}$$

is a $\mathbf{F}_t$*-martingale under* $\mathbf{P}_\mu$ *for all* f *from* $\overline{D}^L$*, the closure of* D *w.r.t. the graph norm to* $(L, D(L))$*. For* $f \in \overline{D}^L$ *the quadratic variation process of* $\mathbf{M}(f)_t$ *is given by*

$$\langle \mathbf{M}(f) \rangle_t = \sigma^2 \int_0^t |\partial_\alpha f|^2(\mathbf{X}_s)\, ds, \quad t \geq 0. \tag{3.3}$$

The corresponding statement holds for the adjoint process. The corresponding martingale we denote by

$$\mathbf{M}^*(f)_t := f(\mathbf{X}_t) - f(\mathbf{X}_0) - \int_0^t L^* f(\mathbf{X}_s)\, ds, \quad t \geq 0, \tag{3.4}$$

where $f \in \overline{D}^{L^*}$*. For* $f \in \overline{D}^{L^*}$ *again we have*

$$\langle \mathbf{M}^*(f) \rangle_t = \sigma^2 \int_0^t |\partial_\alpha f|^2(\mathbf{X}_s)\, ds, \quad t \geq 0. \tag{3.5}$$

Proof. The enlargement of the class of admissible functions for the martingale problem is obvious. Formulas (3.3) and (3.5) for the corresponding quadratic variation processes can be derived as in [CG08b, sec. 3]. □

4. Associated nondegenerated generator. In $L^2(\gamma)$ we consider the pre-Dirichlet form

$$\mathcal{E}(f, g) = \frac{1}{2} \int_{\mathbb{R}^2} \nabla_\xi f \cdot (\nabla_\xi g)^T \, d\gamma, \quad f, g \in C_0^\infty(\mathbb{R}^2).$$

Its closure $(\mathcal{E}, D(\mathcal{E}))$ has a corresponding self-adjoint generator $(G, D(G))$ which on smooth, compactly supported functions is given by

$$Gf = \frac{1}{2}\Delta_\xi f - \frac{1}{2}\nabla_\xi \phi \cdot (\nabla_\xi f)^T, \quad f \in C_0^\infty(\mathbb{R}^2).$$

LEMMA 4.1. *Let* $\phi \in C^3(\mathbb{R}^2)$*,* $\nabla\phi \in L^2(\gamma)$*, and assume there exist* $0 < C < \infty$ *and* $K \subset \mathbb{R}^2$ *compact such that*

$$\begin{aligned}(\partial_{\xi^i}\partial_{\xi^j}\partial_{\xi^k}\phi(\xi))^2 + (\partial_{\xi^i}\partial_{\xi^j}\phi(\xi))^2 \\ \leq C\big(\partial_{\xi^1}\phi(\xi)\big)^2 + \big(\partial_{\xi^2}\phi(\xi)\big)^2 \quad \text{for all} \quad i, j, k \in \{1, 2\},\ \xi \in \mathbb{R}^2 \setminus K.\end{aligned}$$

Then

$$\nabla_\xi f \cdot \tau \in \overline{D}^L \subset D(L)$$

for all $f \in D(G)$*.*

Proof. First recall that $D(\mathcal{E}) \subset D(G)$; hence $\nabla_\xi f \cdot \tau \in L^2(\gamma)$ is well-defined. Our assumptions imply that $(G, C_0^\infty(\mathbb{R}^2))$ is essentially self-adjoint; see, e.g., [BKR97, Thm. 7]. Hence there exists a sequence $(f_n)_{n\in\mathbb{N}}$ in $C_0^\infty(\mathbb{R}^2)$ which converges to f in the graph norm. Moreover

$$\nabla_\xi f_n \cdot \tau \in D \subset D(L) \quad \text{for all} \quad n \in \mathbb{N}.$$

Obviously, $(\nabla_\xi f_n \cdot \tau)_{n\in\mathbb{N}}$ converges to $\nabla_\xi f \cdot \tau$ in $L^2(\mu)$. Hence it is sufficient to show that $(L(\nabla_\xi f_n \cdot \tau))_{n\in\mathbb{N}}$ is a Cauchy sequence in $L^2(\mu)$.

We fix $g \in C_0^\infty(\mathbb{R}^2)$. Then

$$\begin{aligned}(4.1)\quad \|L(\nabla_\xi g\cdot\tau)\|^2_{L^2(\mu)} &= \int_{\mathbb{R}^2}\int_{[0,2\pi]} (L(\cos(\alpha)\partial_{\xi^1}g(\xi)+\sin(\alpha)\partial_{\xi^2}g(\xi)))^2\,d\alpha\,d\gamma(\xi)\\ &= \frac{1}{8}\Big(\sigma^4\mathcal{E}(g,g)+\int_{\mathbb{R}^2}\big((2\partial_{\xi^1}\partial_{\xi^2}+\partial_{\xi^1}\phi(\xi)\partial_{\xi^2}+\partial_{\xi^2}\phi(\xi)\partial_{\xi^1})g(\xi)\big)^2\\ &\quad+3\big((\partial^2_{\xi^1}-\partial_{\xi^2}\phi(\xi)\partial_{\xi^2})g(\xi)\big)^2+3\big((\partial^2_{\xi^2}-\partial_{\xi^1}\phi(\xi)\partial_{\xi^1})g(\xi)\big)^2\\ &\quad+2(\partial^2_{\xi^1}-\partial_{\xi^2}\phi(\xi)\partial_{\xi^2})g(\xi)(\partial^2_{\xi^2}-\partial_{\xi^1}\phi(\xi)\partial_{\xi^1})g(\xi)d\gamma(\xi)\Big).\end{aligned}$$

Therefore $(L(\nabla_\xi f_n\cdot\tau))_{n\in\mathbb{N}}$ is a Cauchy sequence if the operators

$$\partial_{\xi^i}\partial_{\xi^j},\quad \partial_{\xi^i}\phi(\xi)\partial_{\xi^j},\quad i,j\in\{1,2\},$$

are Kato bounded by G on $C_0^\infty(\mathbb{R}^2)$.

To show this, it is useful to introduce the operators

$$G_i := \partial^2_{\xi^i}-\partial_{\xi^i}\phi(\xi)\partial_{\xi^i},\quad i\in\{1,2\}.$$

A straightforward integration by parts yields the following for $g\in C_0^\infty(\mathbb{R}^2)$:

$$\begin{aligned}\big(G_1g,G_1g\big)_{L^2(\gamma)}+\big(G_2g,G_2g\big)_{L^2(\gamma)} &\le C_1\big(Gg,Gg\big)_{L^2(\gamma)}+C_2\big(g,g\big)_{L^2(\gamma)}\\ &\quad+C_3\sum_{i=1}^2\big((\partial_{\xi^i}\phi)g,(\partial_{\xi^i}\phi)g\big)_{L^2(\gamma)}+C_4\sum_{i,j=1}^2\big((\partial_{\xi^i}\partial_{\xi^j}\phi)g,(\partial_{\xi^i}\partial_{\xi^j}\phi)g\big)_{L^2(\gamma)}\\ &\quad+C_5\sum_{i,j,k=1}^2\big((\partial_{\xi^i}\partial_{\xi^j}\partial_{\xi^k}\phi)g,(\partial_{\xi^i}\partial_{\xi^j}\partial_{\xi^k}\phi)g\big)_{L^2(\gamma)}\end{aligned}$$

and

$$\begin{aligned}\big(\partial^2_{\xi^i}g,\partial^2_{\xi^i}g\big)_{L^2(\gamma)} &\le C_6\big(G_ig,G_ig\big)_{L^2(\gamma)}+C_7\big(g,g\big)_{L^2(\gamma)}+C_8\big((\partial^2_{\xi^i}\phi)g,(\partial^2_{\xi^i}\phi)g\big)_{L^2(\gamma)}\\ &\quad+C_9\big((\partial^3_{\xi^i}\phi)g,(\partial^3_{\xi^i}\phi)g\big)_{L^2(\gamma)},\quad i\in\{1,2\},\end{aligned}$$

and

$$\begin{aligned}\big(\partial_{\xi^2}\partial_{\xi^1}g,\partial_{\xi^2}\partial_{\xi^1}g\big)_{L^2(\gamma)} &\le C_{10}\big(G_1g,G_1g\big)_{L^2(\gamma)}+C_{11}\big(G_2g,G_2g\big)_{L^2(\gamma)}+C_{12}\big(g,g\big)_{L^2(\gamma)}\\ &\quad+C_{13}\sum_{i,j,k=1}^2\big((\partial_{\xi^i}\partial_{\xi^j}\partial_{\xi^k}\phi)g,(\partial_{\xi^i}\partial_{\xi^j}\partial_{\xi^k}\phi)g\big)_{L^2(\gamma)}\end{aligned}$$

and

$$\begin{aligned}\big(\partial_{\xi^i}\phi\partial_{\xi^j}g,\partial_{\xi^i}\phi\partial_{\xi^j}g\big)_{L^2(\gamma)} &\le C_{14}\big(G_jg,G_jg\big)_{L^2(\gamma)}+C_{15}\big(g,g\big)_{L^2(\gamma)}+C_{16}\big((\partial_{\xi^i}\phi)g,(\partial_{\xi^i}\phi)g\big)_{L^2(\gamma)}\\ &\quad+C_{17}\sum_{k,l=1}^2\big((\partial_{\xi^k}\partial_{\xi^l}\phi)g,(\partial_{\xi^k}\partial_{\xi^l}\phi)g\big)_{L^2(\gamma)},\ i\ne j,i,j\in\{1,2\},\end{aligned}$$

where $0 < C_1, \dots, C_{17} < \infty$ are constants. Hence it is left to show that the operators given through multiplication by the functions $\partial_{\xi^i}\phi$, $\partial_{\xi^i}\partial_{\xi^j}\phi$ and $\partial_{\xi^i}\partial_{\xi^j}\partial_{\xi^k}\phi$, $i,j,k \in \{1,2\}$, are Kato bounded by G on $C_0^\infty(\mathbb{R}^2)$.

To get this, we transform the problem to $L^2(dx)$. That is, we use the isometric isomorphism

$$L^2(\gamma) \ni f \mapsto \exp(-\phi/2)f \in L^2(dx)$$

to define

$$\hat{G} := \exp(-\phi/2)G\exp(\phi/2) = -\frac{1}{2}\Delta_\xi + \frac{1}{8}\Big(\big(\partial_{\xi^1}\phi\big)^2 + \big(\partial_{\xi^2}\phi\big)^2\Big) - \frac{1}{4}\Big(\big(\partial^2_{\xi^1}\phi\big) + \big(\partial^2_{\xi^2}\phi\big)\Big)$$

on $\hat{D} := \{\exp(-\phi/2)f \mid f \in C_0^\infty(\mathbb{R}^2)\}$. Using our assumptions on ϕ one easily shows that the multiplication operators $\partial_{\xi^i}\phi$, $\partial_{\xi^i}\partial_{\xi^j}\phi$ and $\partial_{\xi^i}\partial_{\xi^j}\partial_{\xi^k}\phi$, $i,j,k \in \{1,2\}$, are Kato bounded by $\hat{G}$ on $\hat{D}$. Hence they are also Kato bounded by G on $C_0^\infty(\mathbb{R}^2)$. ∎

Remark 4.2. The assumptions in Lemma 4.1 allow a large class of potentials ϕ. For example, the potentials $\phi = \|\cdot\|^p$, $p = 2, 4$ or $p \geq 6$, are admissible. Of course, ϕ may also be nonisotropic.

Remark 4.3. It is interesting to note that the stochastic process associated to the generator G is obtained from the original process (2.1) in the large σ-limit; see [BGK+07].

5. Ergodicity and rate of convergence. From now on we assume that μ is a probability measure on E. We are interested in the convergence to zero of

$$\Big\|\frac{1}{t}\int_0^t f(\mathbf{X}_s)\,ds - \mathbf{E}_\mu[f]\Big\|_{L^2(\mathbf{P}_\mu)} = \mathbf{E}_{\mathbf{P}_\mu}\Big[\Big(\frac{1}{t}\int_0^t f(\mathbf{X}_s)\,ds - \mathbf{E}_\mu[f]\Big)^2\Big]^{1/2} \quad \text{as} \quad t \to \infty$$

for $f \in L^2(\mu)$. Denote by $\mathbf{E}_\mu[f|p^1]$ the conditional expectation of $f \in L^2(\mu)$ w.r.t. the σ-algebra in $E = \mathbb{R}^2 \times [0, 2\pi]$ generated by the projection on the first variable. Note that

$$\mathbf{E}_\mu[f|p^1](\xi,\alpha) = \frac{1}{2\pi}\int_{[0,2\pi]} f(\xi,\alpha)\,d\alpha \quad \text{for } \mu \text{ a.a. } (\xi,\alpha) \in E,$$

i.e., $\mathbf{E}_\mu[f|p^1]$ has a version which is independent of $\alpha \in [0, 2\pi]$. Each $f \in L^2(\mu)$ we can write as

$$f = (f - \mathbf{E}_\mu[f|p^1]) + \mathbf{E}_\mu[f|p^1], \tag{5.1}$$

where

$$\mathbf{E}_\mu[(f - \mathbf{E}_\mu[f|p^1])|p^1]) = 0. \tag{5.2}$$

Hence, by the triangle inequality, it is sufficient to consider the following two types of functions: (i) $f \in L^2(\mu)$ with $\mathbf{E}_\mu[f|p^1] = 0$. (ii) $\mathbf{E}_\mu[f|p^1]$ with $f \in L^2(\mu)$.

Denote by $(S, D(S))$ the closure of (S, D). Of course, $(S, D(S))$ is self-adjoint.

PROPOSITION 5.1. *Let the assumptions of Theorem* 3.2 *hold and let* $f \in D(S)$. *Then*

$$\mathbf{E}_{\mathbf{P}_\mu}\Big[\Big(\frac{1}{t}\int_0^t Sf(\mathbf{X}_s)\,ds\Big)^2\Big] \leq \frac{\sigma^2}{t}\|\partial_\alpha f\|^2_{L^2(\mu)}, \quad t > 0.$$

Proof. It is sufficient to consider $f \in D$. We fix $t > 0$ and below we canonically project the laws of the equilibrium processes $\mathbf{P}_\mu$ onto $C([0,t],E)$ without expressing this explicitly. We define the time reversal $r_t(\omega) := \omega(t-\cdot)$, $\omega \in C([0,t],E)$.

Using that

$$S = \frac{1}{2}(L + L^*) \quad \text{on} \quad D,$$

we obtain

$$-\int_0^t Sf(\mathbf{X}_s)\,ds = \frac{1}{2}\Big(\mathbf{M}(f)_t - \mathbf{R}^*(f)_0\Big),$$

where $\mathbf{M}(f)_t$ is the martingale as in (3.2) and

$$\mathbf{R}^*(f)_u := f(\mathbf{X}_t) - f(\mathbf{X}_u) + \int_u^t L^* f(\mathbf{X}_s)\,ds, \quad 0 \le u \le t.$$

Observe that

$$-\mathbf{R}^*(f)_{t-u} = \mathbf{M}^*(f)_u \circ r_t, \quad 0 \le u \le t,$$

where $\mathbf{M}^*(f)_t$ is the martingale as in (3.4). Hence

$$\mathbf{E}_{\mathbf{P}_\mu}\big[\big(\mathbf{R}^*(f)_0\big)^2\big] = \mathbf{E}_{\mathbf{P}^*_\mu}\big[\big(\mathbf{M}^*(f)_t\big)^2\big],$$

because time reversal gives the adjoint process. Now by the Burkholder–Gundy–Davis inequality and (3.3), (3.5) we obtain

$$\begin{aligned}\mathbf{E}_{\mathbf{P}_\mu}\Big[\Big(\frac{1}{t}\int_0^t Sf(\mathbf{X}_s)\,ds\Big)^2\Big]^{1/2} &\le \frac{1}{2t}\mathbf{E}_{\mathbf{P}_\mu}\big[\big(\mathbf{M}(f)_t\big)^2\big]^{1/2} + \frac{1}{2t}\mathbf{E}_{\mathbf{P}^*_\mu}\big[\big(\mathbf{M}^*(f)_t\big)^2\big]^{1/2}\\ &\le \frac{1}{2t}\mathbf{E}_{\mathbf{P}_\mu}\big[\,\langle \mathbf{M}(f)\rangle_t\,\big]^{1/2} + \frac{1}{2t}\mathbf{E}_{\mathbf{P}^*_\mu}\big[\,\langle \mathbf{M}^*(f)\rangle_t\,\big]^{1/2} \le \frac{\sigma}{t^{1/2}}\|\partial_\alpha f\|_{L^2(\mu)}. \qquad \square\end{aligned}$$

COROLLARY 5.2. *Let the assumptions of Theorem* 3.2 *hold and let* $f \in L^2(\mu)$ *with* $\mathbf{E}_\mu[f|p^1] = 0$. *Then*

$$\mathbf{E}_{\mathbf{P}_\mu}\Big[\Big(\frac{1}{t}\int_0^t f(\mathbf{X}_s)\,ds\Big)^2\Big] \le \frac{4}{\sigma^2 t}\|f\|^2_{L^2(\mu)}, \quad t > 0.$$

Proof. Since $\mathbf{E}_\mu[f|p^1] = 0$, f is in the orthogonal complement of the kernel of $(S, D(S))$. But the orthogonal complement of the kernel of $(S, D(S))$ is the range of $(S, D(S))$, because $(S, D(S))$ is self-adjoint. Hence there exists $S^{-1}f \in D(S)$. Then Proposition 5.1 yields

$$\mathbf{E}_{\mathbf{P}_\mu}\Big[\Big(\frac{1}{t}\int_0^t f(\mathbf{X}_s)\,ds\Big)^2\Big] \le \frac{\sigma^2}{t}\|\partial_\alpha S^{-1} f\|^2_{L^2(\mu)} \le \frac{4}{\sigma^2 t}\|f\|^2_{L^2(\mu)},$$

because S^{-1} is bounded by $2\sigma^{-2}$ on the range of $(S, D(S))$. $\square$

THEOREM 5.3. *Let* $\phi \in C^3(\mathbb{R}^2)$, $\nabla\phi \in L^2(\gamma)$, *and assume there exist* $0 < C < \infty$ *and* $K \subset \mathbb{R}^2$ *compact such that*

$$\begin{aligned}&(\partial_{\xi^i}\partial_{\xi^j}\partial_{\xi^k}\phi(\xi))^2 + (\partial_{\xi^i}\partial_{\xi^j}\phi(\xi))^2\\ &\qquad \le C\big(\partial_{\xi^1}\phi(\xi)\big)^2 + \big(\partial_{\xi^2}\phi(\xi)\big)^2 \quad \textit{for all} \quad i,j,k \in \{1,2\},\ \xi \in \mathbb{R}^2 \setminus K.\end{aligned}$$

Furthermore assume that the Dirichlet form $(\mathcal{E}, D(\mathcal{E}))$ corresponding to the nondegenerated generator from section 4 fulfills a Poincaré inequality, i.e., there exists $0 < c < \infty$ such that

$$\mathcal{E}(f - \mathbf{E}_\gamma[f], f - \mathbf{E}_\gamma[f]) \geq c\,\big(f - \mathbf{E}_\gamma[f], f - \mathbf{E}_\gamma[f]\big)_{L^2(\gamma)} \quad \textit{for all} \quad f \in D(\mathcal{E}).$$

Then

$$\begin{aligned}(5.3)\quad &\Big\|\frac{1}{t}\int_0^t f(\mathbf{X}_s)\,ds - \mathbf{E}_\mu[f]\Big\|_{L^2(\mathbf{P}_\mu)} \\ &\leq \frac{1}{c^{1/2}}\Big(\frac{2}{t} + \frac{1}{t^{1/2}}\Big(\frac{A(C)c^{1/2} + B(C)c^{-1/2}}{\sigma} + (1 + 2^{-1/2})\sigma\Big)\Big)\|f - \mathbf{E}_\mu[f]\|_{L^2(\mu)}\end{aligned}$$

for some constants $0 < A(C), B(C) < \infty$ independent of $c, \sigma > 0$, $f \in L^2(\mu)$, and $t > 0$.

Proof. Let $f \in L^2(\mu)$, and without lost of generality we assume $\mathbf{E}_\mu[f] = 0$. Since $(\mathcal{E}, D(\mathcal{E}))$ fulfills a Poincaré inequality, there exists $g \in D(G)$ such that

$$Gg = \mathbf{E}_\mu[f|p^1].$$

Now consider

$$h := \nabla_\xi g \cdot \tau.$$

Then $h \in \overline{D}^L$ by Lemma 4.1 and

$$\begin{aligned}Lh - \mathbf{E}_\mu[f|p^1] = Lh - Gg = &-\frac{\sigma^2}{2}\Big(\cos(\alpha)\partial_{\xi^1} g + \sin(\alpha)\partial_{\xi^2} g\Big) \\ &+ 2\sin(\alpha)\cos(\alpha)\partial_{\xi^2}\partial_{\xi^1} g + \Big(\cos(\alpha)^2 - \frac{1}{2}\Big)\partial^2_{\xi^1} g + \Big(\sin(\alpha)^2 - \frac{1}{2}\Big)\partial^2_{\xi^2} g \\ &- \partial_{\xi^1}\phi(\xi)\Big(\Big(\sin(\alpha)^2 - \frac{1}{2}\Big)\partial_{\xi^1} g - \cos(\alpha)\sin(\alpha)\partial_{\xi^2} g\Big) \\ &- \partial_{\xi^2}\phi(\xi)\Big(\Big(\cos(\alpha)^2 - \frac{1}{2}\Big)\partial_{\xi^2} g - \cos(\alpha)\sin(\alpha)\partial_{\xi^1} g\Big).\end{aligned}$$

It is easy to check that

$$\mathbf{E}_\mu[Lh - \mathbf{E}_\mu[f|p^1]|p^1] = 0.$$

Hence $Lh - \mathbf{E}_\mu[f|p^1]$ can be treated by Corollary 5.2. Thus by (5.1), (5.2) it is only left to consider Lh. Using (3.2), the Burkholder–Gundy–Davis inequality, and (3.3) we obtain

$$\begin{aligned}(5.4)\quad &\Big\|\frac{1}{t}\int_0^t Lh(\mathbf{X}_s)\,ds\Big\|_{L^2(\mathbf{P}_\mu)} \leq \frac{1}{t}\|h(\mathbf{X}_t) - h(\mathbf{X}_0)\|_{L^2(\mathbf{P}_\mu)} + \frac{1}{t}\|\mathbf{M}^h_t\|_{L^2(\mathbf{P}_\mu)} \leq \frac{2}{t}\|h\|_{L^2(\mu)} \\ &+ \frac{1}{t}\mathbf{E}_{\mathbf{P}_\mu}\big[\langle \mathbf{M}^h\rangle_t\big]^{1/2} \leq \frac{2}{t}\|\nabla_\xi g\|_{L^2(\gamma)} + \frac{\sigma}{t^{1/2}}\|\partial_\alpha h\|_{L^2(\mu)} \leq \frac{2 + \sigma t^{1/2}}{t}\|\nabla_\xi g\|_{L^2(\gamma)} \\ &= \frac{2 + \sigma t^{1/2}}{t}\big(G^{-1}\mathbf{E}_\mu[f|p^1], \mathbf{E}_\mu[f|p^1]\big)^{1/2}_{L^2(\gamma)} \leq \frac{(2 + \sigma t^{1/2})}{c^{1/2} t}\|f\|_{L^2(\mu)}, \quad t > 0,\end{aligned}$$

where $0 < c < \infty$ is the spectral gap of G.

Now we obtain by Corollary 5.2 and (5.4) for $t > 0$

$$\begin{aligned}
&\Big\|\frac{1}{t}\int_0^t f(\mathbf{X}_s)\,ds\Big\|_{L^2(\mathbf{P}_\mu)} \\
&\quad\le \Big\|\frac{1}{t}\int_0^t (f-\mathbf{E}_\mu[f|p^1])(\mathbf{X}_s)\,ds\Big\|_{L^2(\mathbf{P}_\mu)} + \Big\|\frac{1}{t}\int_0^t \mathbf{E}_\mu[f|p^1](\mathbf{X}_s)\,ds\Big\|_{L^2(\mathbf{P}_\mu)} \\
&\le \frac{2}{\sigma t^{1/2}}\|f\|_{L^2(\mu)} + \Big\|\frac{1}{t}\int_0^t (\mathbf{E}_\mu[f|p^1]-Lh)(\mathbf{X}_s)\,ds\Big\|_{L^2(\mathbf{P}_\mu)} + \Big\|\frac{1}{t}\int_0^t Lh(\mathbf{X}_s)\,ds\Big\|_{L^2(\mathbf{P}_\mu)} \\
&\qquad\le \frac{2}{\sigma t^{1/2}}\big(\|f\|_{L^2(\mu)} + \|Lh\|_{L^2(\mu)}\big) + \frac{(2+\sigma t^{1/2})}{c^{1/2}t}\|f\|_{L^2(\mu)}.
\end{aligned}$$

To obtain the last inequality we used the fact that

$$\|Lh - \mathbf{E}_\mu[f|p^1]\|^2_{L^2(\mu)} = (Lh-Gg, Lh-Gg)_{L^2(\mu)} = (Lh,Lh)_{L^2(\mu)} - (Gg,Gg)_{L^2(\mu)},$$

because

$$(Lh, Gg)_{L^2(\mu)} = (Gg,Gg)_{L^2(\mu)}.$$

Finally, using (4.1) and the Kato bound provided in Lemma 4.1, we get the estimate (5.3). □

Remark 5.4. The assumptions in Theorem 5.3 also allow potentials of the form $\phi = \|\cdot\|^p$, $p = 2, 4$, or $p \ge 6$, as in Remark 4.2, since in these cases also a Poincaré inequality holds. More generally, under our assumptions on the potential a Poincaré inequality holds if ϕ grows as fast as or faster than $\|x\|$ for large $x \in \mathbb{R}^2$; see [RW01].

6. Numerical Results. For a numerical investigation of the rate of convergence to equilibrium for (2.3) we use a semi-Lagrangian method. The method consists of two fractional steps. The first step is the Lagrangian interpretation for the advection part of (2.3) by the modified method of characteristics, while the second step uses Eulerian coordinates for the discretization of the reaction-diffusion part in (2.3). To preserve mass we use the modified method of characteristics with adjusted advection; see [DHP99]. For details on the numerical method we refer to [KRS07]. We consider the cases $\phi(\xi) = \|\xi\|$ and $\phi(\xi) = \|\xi\|^2/2$. We concentrate on the behavior of the relative entropy. One could also consider other observables. However, since convergence of the relative entropy is stronger than convergence of the distribution functions or convergence of time averages, we restrict ourselves to the investigation of the time development of the relative entropy. Figures 3 and 4 show log-plots of the time development of the relative entropy $H(u/m)(t)$ for different values of σ, and Figure 5 shows plots of the inverse decay rate

$$\lim_{T\to\infty} \int_0^T H(u/m)(t)dt$$

for various values of σ. The quantitative behavior depends on the initial values chosen; however, for other initial values qualitatively similar results are obtained. From Figures 3 and 4 an exponential decay seems possible. Moreover, one observes from Figure 5 that the optimal rate of convergence is found for a finite value of σ. On the other hand, considering the analytical estimate (5.3) one notes that the speed of convergence is dominated by the factor

$$\frac{1}{c^{1/2}t^{1/2}}\Big(\frac{A(C)c^{1/2} + B(C)c^{-1/2}}{\sigma} + (1+2^{-1/2})\sigma\Big).$$

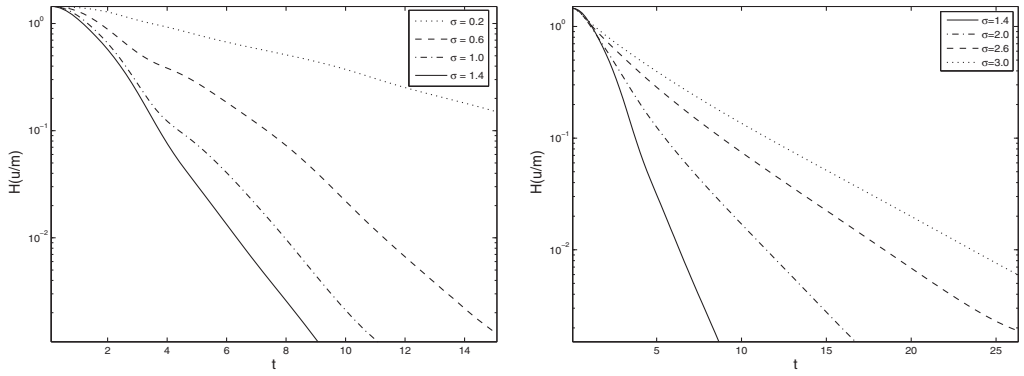


FIG. 3. $H(u/m)(t)$ *for* $\phi(\xi) = \|\xi\|$.

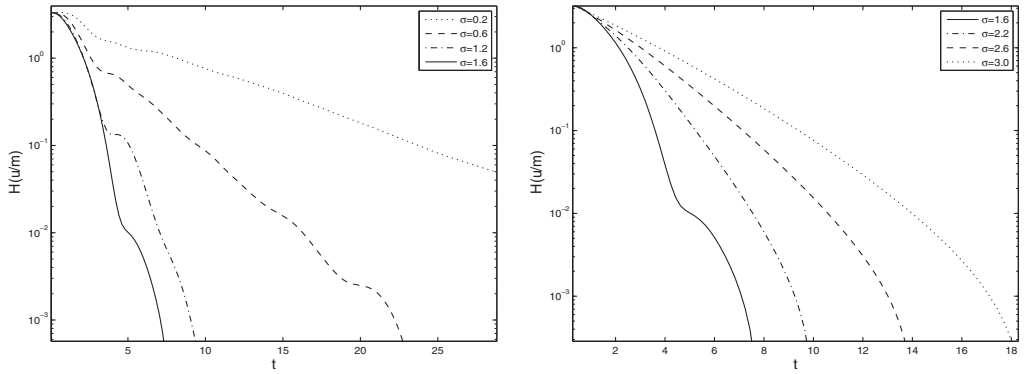


FIG. 4. $H(u/m)(t)$ *for* $\phi(\xi) = \|\xi\|^2/2$.

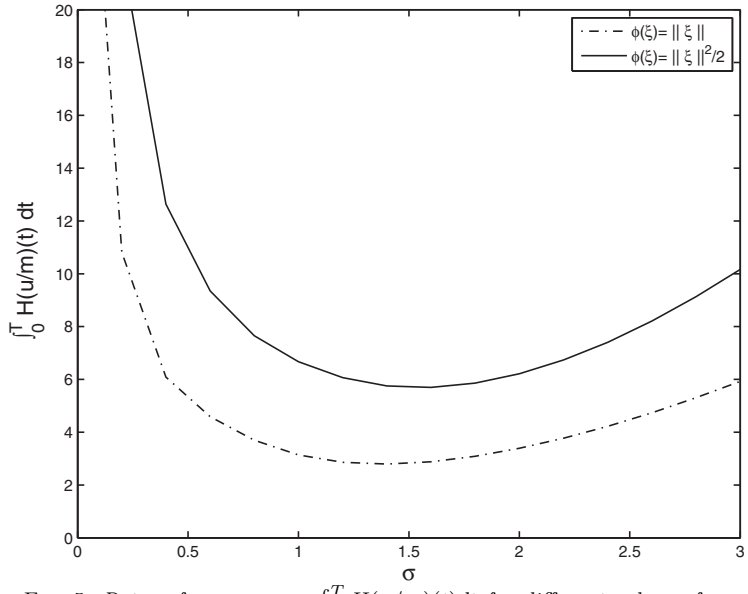


FIG. 5. *Rates of convergence* $\int_0^T H(u/m)(t)dt$ *for different values of* σ.

Obviously, a minimal rate of convergence can be determined—as from the numerical results—for a finite value of σ. Moreover, these observations fit the Monte Carlo results shown in Figure 2. From a practical point of view, this means that the process parameters have to be adapted such that σ is in an intermediate range of values to obtain the fastest possible decay to equilibrium and a fiber web which is as uniform as possible.

7. Concluding remarks. A stochastic model for fiber lay-down processes has been investigated analytically. Existence and ergodicity of the process have been shown and estimates on the rates of convergence presented. The results are supported by numerical simulations. According to these results the fastest decay to equilibrium is obtained for process parameters giving a diffusion coefficient σ in an intermediate range of values. We plan to extend this work to problems with moving conveyor belts; see [BGK+07]. In this case an additional difficulty is given by the fact that the stationary solution is not known explicitly.

Acknowledgment. We are grateful to L. Bonilla, F. Conrad, N. Marheineke, M. Seaid, and C. Villani as well as to the unknown referees for helpful discussions.

REFERENCES

[AMTU01] A. Arnold, P. Markowich, G. Toscani, and A. Unterreiter, *On convex Sobolev inequalities and the rate of convergence to equilibrium for Fokker–Planck type equations*, Comm. Partial Differential Equations, 26 (2001), pp. 43–100.

[BKR97] V. I. Bogachev, N. V. Krylov, and M. Röckner, *Elliptic regularity and essential self-adjointness of Dirichlet operators on* $\mathbb{R}^n$, Ann. Sc. Norm. Super. Pisa Cl. Sci. (4), 24 (1997), pp. 451–461.

[BGK+07] L. Bonilla, T. Götz, A. Klar, N. Marheineke, and R. Wegener, *Hydrodynamic limit of the Fokker–Planck equation describing fiber lay-down processes*, SIAM J. Appl. Math., 68 (2007), pp. 648–665.

[CG08a] F. Conrad and M. Grothaus, *Construction of N-particle Langevin dynamics for* $H^{1,\infty}$*-potentials via generalized Dirichlet forms*, Potential Anal., 28 (2008), pp. 261–282.

[CG08b] F. Conrad and M. Grothaus, *Construction, Ergodicity and Rate of Convergence of N-particle Langevin Dynamics with Singular Potentials*, preprint.

[DPZ96] G. Da Prato and J. Zabczyk, *Ergodicity for Infinite Dimensional Systems*, London Math. Soc. Lecture Notes Ser. 229, Cambridge University Press, Cambridge, UK, 1996.

[DV01] L. Desvilettes and C. Villani, *On the trend to global equilibrium for spatially inhomogeneous entropy-dissipating systems: The linear Fokker–Planck equation*, Comm. Pure Appl. Math., 54 (2001), pp. 1–42.

[DHP99] J. Douglas, C. Huang, and F. Pereira, *The modified method of characteristics with adjusted advection*, Numer. Math., 83 (1999), pp. 353–369.

[Gol85] J. A. Goldstein, *Semigroups of Linear Operators and Applications*, Oxford Math. Monogr., Oxford University Press, New York, 1985.

[GKM+07] T. Götz, A. Klar, N. Marheineke, and R. Wegener, *A stochastic model and associated Fokker–Planck equation for the fiber lay-down process in nonwoven production processes*, SIAM J. Appl. Math., 67 (2007), pp. 1704–1717.

[KRS07] A. Klar, P. Reuterswärd, and M. Seaïd, *A Semi-Lagrangian Method for the Fokker–Planck Equation of Fiber Dynamics*, preprint.

[MW06] N. Marheineke and R. Wegener, *Fiber dynamics in turbulent flows: General modeling framework*, SIAM J. Appl. Math., 66 (2006), pp. 1703–1726.

[OT03] S. Olla and C. Tremoulet, *Equilibrium fluctuations for interacting Ornstein–Uhlenbeck particles*, Comm. Math. Phys., 233 (2003), pp. 463–491.

[RW01] M. Röckner and F.-Y. Wang, *Weak Poincaré inequalities and* L^2*-convergence rates of Markov semigroups*, J. Funct. Anal., 185 (2001), pp. 564–603.

[Sta99] W. Stannat, *The theory of generalized Dirichlet forms and its applications in analysis and stochastics*, Mem. Amer. Math. Soc., 142 (678) (1999).

[Vil06] C. Villani, *Hypocoercivity*, preprint.

SIAM J. MATH. ANAL.
Vol. 40, No. 3, pp. 984–1002

DETERMINATION OF A LINEAR CRACK IN AN ELASTIC BODY FROM BOUNDARY MEASUREMENTS—LIPSCHITZ STABILITY*

ELENA BERETTA†, ELISA FRANCINI‡, AND SERGIO VESSELLA§

Abstract. We discuss the stability issue for the problem of determining a thin inclusion in a homogeneous isotropic elastic body from boundary measurements. This problem is severely ill-posed, but in this paper we prove that by restricting ourselves to the class of thin neighborhoods of line segments, a Lipschitz stability estimate holds.

Key words. inverse problems, linear elasticity, cracks

AMS subject classification. 35R30

DOI. 10.1137/070698397

1. Introduction.

Let $\Omega \subset \mathbb{R}^2$ be a bounded connected domain, with a sufficiently smooth boundary, representing the region occupied by an elastic, homogeneous, isotropic material.

Let $\sigma \subset \Omega$ be a simple smooth curve and define, for a positive small ϵ, the set

$$\omega_\epsilon = \{x \in \Omega \,:\, d(x,\sigma) < \epsilon\},$$

which represents an inclusion of small size made of a different elastic material.

Let $\mathbb{C}_0$ and $\mathbb{C}_1$ be the elastic tensor fields in $\Omega \setminus \overline{\omega}_\epsilon$ and ω_ϵ, respectively.

Given a traction field g on $\partial\Omega$, the displacement field u_ϵ, generated by this traction in the body containing the inclusion ω_ϵ, solves the following system of linearized elasticity:

$$\begin{cases} \operatorname{div}\left(\mathbb{C}_\epsilon \widehat{\nabla} u_\epsilon\right) = 0 & \text{in} \quad \Omega, \\ (\mathbb{C}_\epsilon \widehat{\nabla} u_\epsilon)\nu = g & \text{on} \quad \partial\Omega, \end{cases} \tag{1}$$

where $\mathbb{C}_\epsilon = \mathbb{C}_0 \chi_{\Omega\setminus\omega_\epsilon} + \mathbb{C}_1 \chi_{\omega_\epsilon}$, $\widehat{\nabla} u_\epsilon = \frac{1}{2}\left(\nabla u_\epsilon + (\nabla u_\epsilon)^T\right)$ is the symmetric deformation tensor and ν denotes the outward unit normal to $\partial\Omega$.

Let us also introduce the background displacement field u_0 generated by the same traction g in the absence of the inclusion, namely, the solution of

$$\begin{cases} \operatorname{div}\left(\mathbb{C}_0 \widehat{\nabla} u_0\right) = 0 & \text{in} \quad \Omega, \\ (\mathbb{C}_0 \widehat{\nabla} u_0)\nu = g & \text{on} \quad \partial\Omega. \end{cases} \tag{2}$$

*Received by the editors July 25, 2007; accepted for publication (in revised form) May 29, 2008; published electronically September 19, 2008. This work was supported by MIUR, PRIN, grant 2006014115.

http://www.siam.org/journals/sima/40-3/69839.html

†Dipartimento di Matematica "G. Castelnuovo," Università di Roma "La Sapienza," Piazzale Aldo Moro 5, 00185 Roma, Italy (beretta@mat.uniroma1.it).

‡Dipartimento di Matematica "U. Dini," Università degli Studi di Firenze, Viale Morgagni 67A, 50134 Firenze, Italy (francini@math.unifi.it).

§Dipartimento di Matematica per le Decisioni, Università degli Studi di Firenze, Via C. Lombroso, 6/17, 50134 Firenze, Italy (sergio.vessella@dmd.unifi.it).

In [BF06] the following asymptotic expansion for $(u_\epsilon - u_0)|_{\partial\Omega}$ as $\epsilon \to 0$ has been derived:

$$(u_\epsilon - u_0)(y) = 2\epsilon \int_\sigma \mathcal{M}(x)\widehat{\nabla}u_0(x) \cdot \widehat{\nabla}N(x,y)\, d\sigma(x) + O(\epsilon^{1+\theta}) \quad \text{for } y \in \partial\Omega, \tag{3}$$

where $N(x,y)$ denotes the Neumann function for the operator $\operatorname{div}\big(\mathbb{C}_0\widehat{\nabla}\cdot\big)$ in Ω, $\mathcal{M}$ is a fourth order symmetric tensor which depends on σ and on the elastic properties of the material occupied by the region Ω, and $\theta \in (0,1)$ is independent of ϵ.

In the following, the symbol "$\cdot$" (as in formula (3)) will denote the usual scalar product between matrices ($A \cdot B = \sum_{ij} a_{ij}b_{ij}$) or vectors ($u \cdot v = \sum_i u_i v_i$).

Clearly, in order to get information on ω_ϵ, we need to consider the first order term of the expansion (3). Let us consider

$$u_\sigma(y) = 2\int_\sigma \mathcal{M}(x)\widehat{\nabla}u_0(x) \cdot \widehat{\nabla}N(x,y)\, d\sigma(x) \quad \text{for } y \in \overline{\Omega}\backslash\sigma. \tag{4}$$

The inverse problem we are interested in is the following:

Given the trace of the correction term u_σ on some open subset Γ of $\partial\Omega$, determine the curve σ.

Due to the nonlinearity, the problem is severely ill-posed and the reconstruction of an arbitrary smooth curve is expected to fail. On the other hand, restricting the class of admissible curves σ, we expect to regularize the problem. In fact, in the scalar conductivity case, if σ is restricted to be linear, it is possible to show a Lipschitz continuous dependence of the segment σ from the boundary data $u_\sigma|_{\partial\Omega}$ (cf. [ABF04, ABF06]).

In this paper we show that the same type of continuous dependence holds true in the elastic case. More precisely, in Theorem 2.2 we prove a Lipschitz continuous dependence of the segment σ from boundary data of the correction term u_σ. In order to do this we use the qualitative properties of the correction term and quantitative estimates of the unique continuation property for elliptic systems with constant coefficients. In Corollary 2.3 we show that the segment σ depends Lipschitz continuously on the rescaled boundary deviation of the solution u_ϵ.

Our method allows us to detect thin elastic inclusions from knowledge of one boundary measurement. In the book [AK04], the authors deal with a similar problem for diametrically small inclusions. We would like to point out that this analysis is on one side motivated by applications (breast imaging, mine detection), and, on the other side, it is not known if an arbitrary inclusion can be detected with a finite number of boundary measurements, not even in the scalar case.

The plan of the paper is the following: In section 2 we introduce some notation and the main assumptions, and we state the main result. In section 3 we describe some properties and the asymptotic behavior near the endpoints of the crack σ of the function u_σ. Section 4 contains the proof of the main result. The final section is devoted to the proof of some auxiliary lemmas and propositions stated in section 4.

2. The main result. We introduce some notation and assumptions that will be useful in what follows. For any $r > 0$ we denote

$$\Omega_r = \{x \in \Omega \mid d(x, \partial\Omega) > r\},$$

and by $diam\,\Omega$ the diameter of Ω. We also denote by $B_r(P)$ the open disc centered in P with radius r.

For any $P, Q \in \mathbb{R}^2$ we denote by $[P, Q]$ the segment with endpoints P and Q; more precisely,

$$[P, Q] = \{P + t(Q - P) \mid t \in [0, 1]\}.$$

Given positive constants $\gamma, \rho_0, E_0, E_1, L, \gamma_0$ with $L \geq 1$ and $\gamma \in (0, 1]$, which we shall name a priori data,

(H1) we assume that Ω is a bounded simply connected domain of $\mathbb{R}^2$ such that

$$diam\, \Omega \leq E_1.$$

Concerning the regularity of $\partial\Omega$, we assume that

$$\partial\Omega \text{ is of class } C^{2,\gamma} \text{ with constants } \rho_0, E_0.$$

More precisely, for any point $\tilde{P} \in \partial\Omega$, there exists a rigid transformation of coordinates under which we have $\tilde{P} = 0$ and

$$\Omega \cap B_{\rho_0}(0) = \{x \in B_{\rho_0}(0) \quad | \quad x_2 > \psi(x_1)\},$$

where ψ is a $C^{2,\gamma}$ function on $(-\rho_0, \rho_0) \subset \mathbb{R}$ satisfying

$$\psi(0) = 0, \qquad \psi'(0) = 0, \qquad \text{and} \qquad \|\psi\|_{C^{2,\gamma}(-\rho_0,\rho_0)} \leq E_0.$$

Moreover we assume that there exists an open subset Γ of $\partial\Omega$ such that for some point $\tilde{Q} \in \Gamma$,

$$\partial\Omega \cap B_{\rho_0}(\tilde{Q}) \subset \Gamma.$$

(H2) We assume that σ is a segment of endpoints P, Q satisfying

$$L^{-1} \leq |P - Q| \leq L \qquad \text{and} \qquad d(\sigma, \mathbb{R}^2 \backslash \Omega) \geq L^{-1}.$$

(H3) We will assume Ω and ω_ϵ are both homogeneous and isotropic, i.e., the elastic tensor fields $\mathbb{C}_0$ and $\mathbb{C}_1$ are of the form

$$(\mathbb{C}_m)_{ijlk} = \lambda_m \delta_{ij}\delta_{kl} + \mu_m(\delta_{ki}\delta_{lj} + \delta_{kj}\delta_{li}) \quad \text{for } i, j, k, l = 1, 2, \;\; m = 0, 1,$$

where (λ_0, μ_0) and (λ_1, μ_1) are the Lamé coefficients corresponding to $\Omega \setminus \overline{\omega}_\epsilon$ and ω_ϵ, respectively, and satisfy the monotonicity condition

$$(\lambda_0 - \lambda_1)(\mu_0 - \mu_1) \geq \gamma_0$$

for some positive constant γ_0.

(H4) There are two positive constants α_0, β_0 such that

$$\min(\mu_0, \mu_1) \geq \alpha_0, \quad \min(\lambda_0 + \mu_0, \lambda_1 + \mu_1) \geq \beta_0/2.$$

We note that hypothesis (H4) ensures that $\mathbb{C}_\epsilon$ is strongly convex in Ω; i.e., if we set $\xi_0 = \min(2\alpha_0, \beta_0)$, then

$$\mathbb{C}_\epsilon A \cdot A \geq \xi_0 |A|^2$$

for any symmetric 2×2 matrix A.

(H5) We shall prescribe the traction field g on $\partial\Omega$ of the form

$$g = (\mathbb{C}_0 W)\nu,$$

where W is a nonzero symmetric 2×2 constant matrix.

Remark 2.1. Under assumptions (H3), (H4) there exist weak solutions u_ϵ and u_0 in $H^1(\Omega, \mathbb{R}^2)$ of (1) and of (2), respectively (see, for example, [F72]). In particular u_0 can be explicitly calculated and corresponds to a pure strain displacement

$$u_0(x) = Wx + c,$$

where c is an arbitrary constant vector. We observe that choosing g as in (H5) is not too restrictive since a pure strain can always be accomplished either by simple extensions in perpendicular directions or by a uniform dilation followed by an isochoric pure strain [G72].

In order to uniquely determine u_ϵ and u_0, we assume that they satisfy the following normalization conditions:

$$\int_{\partial\Omega} u = 0, \quad \int_\Omega \nabla u - (\nabla u)^T = 0. \tag{5}$$

For $y \in \Omega$, we will denote by $N(\cdot, y)$ the Neumann function related to Ω, i.e., the weak solution to the problem

$$\begin{cases} \operatorname{div}\left(\mathbb{C}_0 \widehat{\nabla} \mathrm{N}(\cdot, \mathrm{y})\right) = -\delta_y \,\mathrm{Id} \quad \text{in} \quad \Omega, \\ \left(\mathbb{C}_0 \widehat{\nabla} N(\cdot, y)\right) \cdot \nu = -\frac{1}{|\partial\Omega|} \,\mathrm{Id} \quad \text{on} \quad \partial\Omega, \end{cases}$$

with the normalization conditions (5). Here Id is the identity matrix in $\mathbb{R}^2$. Note that by well-known regularity results for elliptic systems (cf. [C80]) we have

$$N(x, y) = \Gamma(x - y) + w(x, y), \tag{6}$$

where w is a smooth function of x and y and $\Gamma(x, y)$ is the fundamental free space solution of $\operatorname{div}\big(\mathbb{C}_0 \widehat{\nabla} \cdot\big)$, given by

$$\Gamma_{ij}(x) = \frac{A}{2\pi}\, \delta_{ij} \log|x| - \frac{B}{2\pi} \frac{x_i x_j}{|x|^2}, \quad i, j = 1, 2, \tag{7}$$

where $A = \frac{1}{2}\big(\frac{1}{\mu_0} + \frac{1}{\lambda_0 + 2\mu_0}\big)$ and $B = \frac{1}{2}\big(\frac{1}{\mu_0} - \frac{1}{\lambda_0 + 2\mu_0}\big)$.

Let us fix an orthonormal system (n, τ) on σ such that n is a unit normal vector field to the segment and τ is a unit tangent vector field.

By Theorem 2.1 of [BF06] we have that, for any $y \in \partial\Omega$, the first order term u_σ of the expansion of $u_\epsilon - u_0$ can be computed explicitly up to a reminder term and has the form

$$u_\sigma(y) = 2 \int_\sigma \mathcal{M} W \cdot \widehat{\nabla} N(x, y)\, d\sigma(x), \tag{8}$$

where

$$\mathcal{M} W = a(\mathrm{tr} W)\mathrm{Id} + bW + c(\tau^T W \tau)\tau \otimes \tau + d(n^T W n) n \otimes n.$$

Here tr$W$ denotes the trace of the matrix W and the coefficients are given by

$$a = (\lambda_1 - \lambda_0)\frac{\lambda_0 + 2\mu_0}{\lambda_1 + 2\mu_1}, \quad b = 2(\mu_1 - \mu_0)\frac{\mu_0}{\mu_1}, \tag{9}$$

$$c = 2(\mu_1 - \mu_0)\left[\left(\frac{2\lambda_1 + 2\mu_1 - \lambda_0}{\lambda_1 + 2\mu_1} - \frac{\mu_0}{\mu_1}\right)\right], \tag{10}$$

and

$$d = 2(\mu_1 - \mu_0)\frac{\mu_1\lambda_0 - \mu_0\lambda_1}{\mu_1(\lambda_1 + 2\mu_1)}. \tag{11}$$

We are now ready to state our main result.

THEOREM 2.2. *Let σ_0 and σ_1 be two segments satisfying assumption* (H2). *Assume* (H1), (H3), (H4), *and* (H5). *Let u_{σ_0} and u_{σ_1} be the functions defined by* (8) *where σ is replaced by σ_0 and σ_1, respectively. Then there exists a constant C, depending only on the a priori data, such that*

$$d_{\mathcal{H}}(\sigma_0, \sigma_1) \leq C\|u_{\sigma_0} - u_{\sigma_1}\|_{L^2(\Gamma)}.$$

Here $d_{\mathcal{H}}$ denotes the Hausdorff distance

$$d_{\mathcal{H}}(\sigma_0, \sigma_1) = \min\{\max\{|P_0 - P_1|, |Q_0 - Q_1|\}, \max\{|P_0 - Q_1|, |Q_0 - P_1|\}\},$$

where P_0, Q_0 and P_1, Q_1 are the endpoints of σ_0 and σ_1, respectively.

The proof of Theorem 2.2 is postponed to section 4.

A straightforward consequence of Theorem 2.2 and of the asymptotic formula (3) is the following.

COROLLARY 2.3. *Let σ_0 and σ_1 be as in Theorem* 2.2. *For $i = 0, 1$, let u_ϵ^i be the solution of* (1) *corresponding to $\omega_\epsilon^i = \{x \in \Omega : d(x, \sigma_i) < \epsilon\}$. Then there exist a positive constant C and $\theta \in (0, 1)$, depending only on the a priori data, such that*

$$d_{\mathcal{H}}(\sigma_0, \sigma_1) \leq C\left(\epsilon^{-1}\|u_\epsilon^0 - u_\epsilon^1\|_{L^2(\Gamma)} + \epsilon^\theta\right).$$

Remark 2.4. A similar rescaled stability estimate has been obtained, for example, in [FV89] and in [CMV98] for the case of diametrically small inclusions.

3. Properties of the function u_σ. In this section we first establish a representation formula for the function u_σ (defined in (4)), and then we deduce some results concerning the regularity of this function and its behavior near the endpoints of σ.

LEMMA 3.1. *Let $\sigma = [P, Q]$, and fix $\tau = \frac{Q-P}{|Q-P|}$ and let $n = \tau^\perp$. For $y \in \overline{\Omega} \setminus \sigma$,*

$$u_\sigma(y) = \int_\sigma \mathbb{C}_0\widehat{\nabla}N(x, y)n \cdot \varphi_\sigma \, d\sigma(x) + (N(Q, y) \cdot \tau - N(P, y) \cdot \tau)\, f_\sigma, \tag{12}$$

where φ_σ is the vector whose components in the (n, τ) directions are given by

$$\begin{aligned} \varphi_\sigma \cdot n &= \frac{2}{\lambda_1 + 2\mu_1}((\lambda_1 - \lambda_0)\,\mathrm{tr}W + 2(\mu_1 - \mu_0)\, n^T W n), \\ \varphi_\sigma \cdot \tau &= \frac{4(\mu_1 - \mu_0)}{\mu_1}\, n^T W \tau \end{aligned} \tag{13}$$

and f_σ is the constant function

$$f_\sigma = \alpha \operatorname{tr} W - \beta\, n^T W n, \tag{14}$$

where

$$\alpha = \frac{4}{\lambda_1 + 2\mu_1}((\mu_1 - \mu_0)(\lambda_1 + 2\mu_1) + \mu_1(\lambda_1 - \lambda_0)), \quad \beta = \frac{8(\mu_1 - \mu_0)}{\lambda_1 + 2\mu_1}(\lambda_1 + \mu_1).$$

Proof. By straightforward calculations and by the symmetry of N and W, one can easily check that

$$u_\sigma(y) = \int_\sigma \mathbb{C}_0 \widehat{\nabla} N(x,y) n \cdot \varphi_\sigma \, d\sigma(x) + 2 \int_\sigma \tau^T \widehat{\nabla} N(x,y) \tau f_\sigma d\sigma(x). \tag{15}$$

Finally an integration by parts in (15) gives the thesis. ∎

Let us consider the double layer potential

$$\mathcal{D}_\sigma \varphi(y) = \int_\sigma \mathbb{C}_0 \widehat{\nabla} \Gamma(x,y) n \cdot \varphi \, d\sigma(x), \quad y \in R^2 \backslash \sigma,$$

with φ constant vector field. Let us point out the following properties.

LEMMA 3.2. *We have*

$$[\mathcal{D}_\sigma \varphi]_\sigma = \varphi, \tag{16}$$

and there exists a positive constant C such that, for any $y \in \mathbb{R}^2 \backslash \sigma$,

$$|\mathcal{D}_\sigma \varphi(y)| \le C \left(\left| \log \frac{|Q-y|}{|P-y|} \right| + 1 \right) |\varphi|,$$

$$|\nabla \mathcal{D}_\sigma \varphi(y)| \le C \left(\frac{1}{|P-y|} + \frac{1}{|Q-y|} + 1 \right) |\varphi|,$$

where $[\mathcal{D}_\sigma \varphi]_\sigma$ denotes the jump of $\mathcal{D}_\sigma \varphi$ across σ.

Proof. Formula (16) follows from standard properties of double layer potentials (see, for example, [AK04]). The behavior of $\mathcal{D}_\sigma \varphi$ at the endpoints of segment σ follows from straightforward computations. ∎

We are now ready to state and prove the following.

PROPOSITION 3.3. *Under assumptions* (H1)–(H5), *u_σ satisfies*

$$\begin{cases} \operatorname{div}\left(\mathbb{C}_0 \widehat{\nabla} u_\sigma \right) = 0 & \text{in} \quad \Omega \backslash \sigma, \\ (\mathbb{C}_0 \widehat{\nabla} u_\sigma) \nu = 0 & \text{on} \quad \partial\Omega, \end{cases} \tag{17}$$

and the normalization condition (5).

The function u_σ has a jump on σ given by

$$[u_\sigma]_\sigma = \varphi_\sigma. \tag{18}$$

Moreover there exists a positive constant C, depending only on the a priori data, such that

$$|u_\sigma(y)| \le C \left(\left| \log \frac{|Q-y|}{|P-y|} \right| + 1 \right). \tag{19}$$

Proof. From the representation formula (12) and by the properties of the function N it is easy to see that u_σ satisfies (17). Recalling that u_σ is the correction term in the expansion of $u_\epsilon - u_0$ and that u_ϵ and u_0 have the same Neumann condition on $\partial\Omega$, and observing that the expansion (3) holds true in a neighborhood of $\partial\Omega$, it follows that u_σ satisfies the homogeneous Neumann datum.

Observe now that u_σ is expressed in terms of a double layer potential and the Neumann function at the endpoints of σ. From (6) and Lemma 3.2, equations (18) and (19) follow. ∎

4. Proof of Theorem 2.2. We start by establishing a first rough estimate of $d_{\mathcal{H}}(\sigma_0, \sigma_1)$ in terms of the boundary data error.

LEMMA 4.1. *Let $\sigma_0 = [P_0, Q_0]$ and $\sigma_1 = [P_1, Q_1]$ be two segments satisfying assumption* (H2). *Assume* (H1), (H3), (H4), *and* (H5). *Let $u_{\sigma_0}, u_{\sigma_1}$ be the functions defined by* (12) *and corresponding to σ_0 and σ_1, respectively. Let $\varepsilon := \|u_{\sigma_0} - u_{\sigma_1}\|_{L^2(\Gamma)}$; then*

$$d_{\mathcal{H}}(\sigma_0, \sigma_1) \leq \omega_1(\varepsilon),$$

where $\omega_1(s) = C\left(\log|\log(\varepsilon)|\right)^{-1/4}$ for some positive constant C depending only on the a priori data.

Proof. We first describe the proof in the case $\varepsilon = 0$ (corresponding to the uniqueness).

Let $u := u_{\sigma_0} - u_{\sigma_1}$. Recalling that $(\mathbb{C}_0\widehat{\nabla}u_{\sigma_0}) \cdot \nu = (\mathbb{C}_0\widehat{\nabla}u_{\sigma_1}) \cdot \nu = 0$ on $\partial\Omega$, we have that

$$u = (\mathbb{C}_0\widehat{\nabla}u) \cdot \nu = 0 \quad \text{on} \quad \partial\Omega.$$

Since u is solution of

$$\operatorname{div}\left(\mathbb{C}_0\widehat{\nabla}u\right) = 0 \quad \text{in} \quad \Omega\backslash(\sigma_0 \cup \sigma_1),$$

by the unique continuation property of solutions to elliptic systems with constant coefficients (see [AM01] and [MR04]) we have that

$$u = 0 \quad \text{in} \quad \Omega\backslash(\sigma_0 \cup \sigma_1). \tag{20}$$

Assume that $\sigma_0 \neq \sigma_1$. Then there exists a portion Σ, for instance, of σ_0, containing an endpoint P_0 of σ_0 such that Σ is not contained in σ_1. For the sake of simplicity, let us denote $\varphi_0 := \varphi_{\sigma_0}$ and $f_0 := f_{\sigma_0}$ the functions appearing in (12) corresponding to σ_0. By (20), $[u]_\Sigma = 0$, and since $[u]_\Sigma = [u_{\sigma_0}]_\Sigma = \varphi_0$ we get

$$\varphi_0 = 0. \tag{21}$$

Furthermore, from the representation formula (12), u_{σ_0} has logarithmic singularities at the endpoints P_0, Q_0. Hence again by (20) and by (12) we have

$$f_0 = 0. \tag{22}$$

Recalling (13) and (14), conditions (21) and (22) can be written as

$$\begin{cases} n_0^T W \tau_0 = 0, \\ \alpha(\operatorname{tr}W) - \beta n_0^T W n_0 = 0, \\ a(\operatorname{tr}W) + (b+d) n_0^T W n_0 = 0, \end{cases} \tag{23}$$

where n_0 and τ_0 denote the normal and tangent unit vector to σ_0, the numbers a, b, and d are defined in (9), (10), (11), and α and β are defined in (14). By assumption (H3) it is easy to see that the conditions (23) imply $\mathrm{tr}W = 0$, $n_0^T W n_0 = 0$, from which it follows that $\tau_0^T W \tau_0 = 0$. Finally, by the symmetry of W, since $n_0^T W \tau_0 = \tau_0^T W n_0$ we get that also $\tau_0^T W n_0 = 0$. Hence $W = 0$, which contradicts assumption (H5).

The proof of the case $\varepsilon > 0$ is a quantitative version of the uniqueness part. For the sake of simplicity we postpone this part of the proof to section 5. □

Denote by σ_t the segment $[P_t, Q_t]$, where

$$P_t = (1-t)P_0 + tP_1, \quad Q_t = (1-t)Q_0 + tQ_1,$$

and for $y \in \Omega \backslash \sigma_t$ let

$$u_t(y) := u_{\sigma_t}(y) = \int_0^1 \mathbb{C}_0 \widehat{\nabla} N((1-s)P_t + sQ_t, y) n_t \cdot \varphi_t |Q_t - P_t| \, ds + (N(Q_t, y)\cdot \tau_t - N(P_t, y)\cdot \tau_t) f_t,$$

where $\tau_t = \frac{Q_t - P_t}{|Q_t - P_t|}$, $n_t = \tau_t^\perp$, $f_t := f_{\sigma_t}$, and $\varphi_t := \varphi_{\sigma_t}$. Clearly, from the above formula, u_t is differentiable for any $t \in [0,1]$ and its derivative u_t' satisfies the problem

$$\begin{cases} \mathrm{div}\left(\mathbb{C}_0 \widehat{\nabla} u_t'\right) = 0 \quad \text{in} \quad \Omega \backslash \sigma_t, \\ (\mathbb{C}_0 \widehat{\nabla} u_t')\nu = 0 \quad \text{on} \quad \partial\Omega \end{cases} \tag{24}$$

and the normalization condition (5).

The following statements, whose proof we defer until section 5, allow us to conclude the proof of Theorem 2.2.

LEMMA 4.2. *For any $y \in \partial\Omega$*

$$\begin{aligned} u_t'(y) &:= \frac{d}{dt} u_t(y) \\ &= -\int_0^1 \mathbb{C}_0 \widehat{\nabla} N((1-s)P_t + sQ_t, y) n_t \cdot \varphi_t((Q_1 - Q_0 - (P_1 - P_0))\cdot \tau_t) ds \\ &\quad + \int_0^1 \mathbb{C}_0 \widehat{\nabla} N((1-s)P_t + sQ_t, y) n_t \cdot \frac{d}{dt}(\varphi_t |Q_t - P_t|)\, ds \\ &\quad + \mathbb{C}_0 \widehat{\nabla} N(Q_t, y)\left((Q_1 - Q_0)\cdot \tau_t) n_t - (Q_1 - Q_0)\cdot n_t)\tau_t\right)\cdot \varphi_t \\ &\quad - \mathbb{C}_0 \widehat{\nabla} N(P_t, y)\left((P_1 - P_0)\cdot \tau_t) n_t - (P_1 - P_0)\cdot n_t)\tau_t\right)\cdot \varphi_t \\ &\quad + \left(\nabla N(Q_t, y)\frac{dQ_t}{dt}\cdot \tau_t - \nabla N(P_t, y)\frac{dP_t}{dt}\cdot \tau_t\right) f_t \\ &\quad + (N(Q_t, y) - N(P_t, y))\frac{d(\tau_t f_t)}{dt}. \end{aligned} \tag{25}$$

PROPOSITION 4.3. *There exist constants C_0, $m_0 > 0$, depending only on the a priori data, such that*

$$\|u_0'\|_{L^2(\Gamma)} \geq m_0 d_{\mathcal{H}}(\sigma_0, \sigma_1)$$

and

$$\|u_t' - u_0'\|_{L^2(\Gamma)} \leq C_0 (d_{\mathcal{H}}(\sigma_0, \sigma_1))^2. \tag{26}$$

Let us conclude now the proof of Theorem 2.2.

Let $\varepsilon_0 > 0$ such that

$$\omega_1(\varepsilon_0) \leq \frac{m_0}{2C_0}, \tag{27}$$

where ω_1 is defined as in Lemma 4.1.

Let us distinguish two cases:

(1) $\varepsilon \in (0, \varepsilon_0]$. By the differentiability of u_t we have that for any $x \in \Gamma$

$$u_{\sigma_1}(x) - u_{\sigma_0}(x) = \int_0^1 u_t'(x)\, dt = u_0'(x) + \int_0^1 (u_t'(x) - u_0'(x))\, dt.$$

By the triangular inequality we get

$$\varepsilon = \|u_{\sigma_1} - u_{\sigma_0}\|_{L^2(\Gamma)} \geq \|u_0'\|_{L^2(\Gamma)} - \|u_t' - u_0'\|_{L^2(\Gamma)}.$$

By Proposition 4.3 we get

$$\varepsilon \geq (m_0 - C_0 d_{\mathcal{H}}(\sigma_0, \sigma_1))\, d_{\mathcal{H}}(\sigma_0, \sigma_1). \tag{28}$$

On the other hand, from Lemma 4.1 we have that

$$d_{\mathcal{H}}(\sigma_0, \sigma_1) \leq \omega_1(\varepsilon_0).$$

Hence, by (27), we get

$$m_0 - C_0 d_{\mathcal{H}}(\sigma_0, \sigma_1) \geq m_0 - C_0 \omega_1(\varepsilon_0) \geq \frac{m_0}{2}.$$

By the last inequality and (28) we derive

$$d_{\mathcal{H}}(\sigma_0, \sigma_1) \leq \frac{2}{m_0}\, \varepsilon. \tag{29}$$

(2) Let us consider the case $\varepsilon \geq \varepsilon_0$. One easily gets

$$d_{\mathcal{H}}(\sigma_0, \sigma_1) \leq diam\, \Omega \leq E_1 \frac{\varepsilon}{\varepsilon_0}. \tag{30}$$

Hence, by (29) and (30) we have

$$d_{\mathcal{H}}(\sigma_0, \sigma_1) \leq \max\left\{ \frac{E_1}{\varepsilon_0}, \frac{2}{m_0} \right\} \varepsilon,$$

which concludes the proof of the theorem. □

5. Proofs of the preliminary results.

5.1. Proof of Lemma 4.1. Let us now consider the case $\varepsilon > 0$. Let $d := d_{\mathcal{H}}(\sigma_0, \sigma_1)$ and let us assume that $d = |P_0 - P_1|$. By classical regularity results in

potential theory and by Lemma 3.2 we have that $u_{\sigma_j} \in C^1(\overline{\Omega}\backslash\sigma_j)$ for $j = 0, 1$ and, for any $r \in (0, \rho_1)$, where $\rho_1 = \min(L^{-1}, \rho_0)$,

$$\begin{aligned} \|u_{\sigma_j}\|_{L^\infty(\Omega\backslash(B_r(P_j)\cup B_r(Q_j)))} &\le C\left(\log\tfrac{1}{r} + 1\right), \\ \|\nabla u_{\sigma_j}\|_{L^\infty(\Omega\backslash(\sigma_j\cup B_r(P_j)\cup B_r(Q_j)))} &\le C\tfrac{1}{r}, \end{aligned} \tag{31}$$

where C depends only on the a priori data. From stability estimates for the Cauchy problem for solutions of elliptic systems (cf. [MR04]) and by (31) we have

$$\|u_{\sigma_j}\|_{L^\infty\left(\Omega\cap B_{\frac{\rho_1}{2}}(\overline{P})\right)} \le C\varepsilon^\delta, \tag{32}$$

where $\overline{P}$ is a point on Γ and $C > 0$, $\delta \in (0, 1)$ depend on the a priori data only. Starting from (32) we now establish a propagation error estimate in a neighborhood of $\Omega\backslash(\sigma_0 \cup \sigma_1)$. Let $r \in (0, \frac{\rho_1}{156})$, and define

$$(\sigma_i)_{8r} = \{x \in \Omega : dist(x, \sigma_i) > 8r\} \quad \text{for} \quad i = 0, 1 \tag{33}$$

and $\Omega'_{(r)} = \Omega_{8r}\backslash((\sigma_0)_{8r}) \cup (\sigma_1)_{8r})$ and let $z_0 = \overline{P} - \frac{\nu}{16}\rho_1$, where ν is the outward unit normal vector to $\partial\Omega$ at the point $\overline{P}$. Let $x \in \Omega'_{(r)}$ and let $\gamma \subset \Omega'_{(r)}$ be an open arc joining x to z_0. With the usual procedure we construct a chain of balls with centers along γ where we apply the three-spheres inequality (cf. [AM01] and [MR04]),

$$\int_{B_{3r}} |u|^2 \le C\left(\int_{B_r} |u|^2\right)^\tau \left(\int_{B_{4r}} |u|^2\right)^{1-\tau}$$

($\tau \in (0, 1), C > 0$ absolute constants), in order to give an estimate of the propagation of the error from $B_r(z_0)$ to x. In fact, by (31) and (32) and the above inequality we get

$$\int_{B_{3r}(x)} |u|^2 \le Cr^2(\varepsilon^\delta)^{\tau^{N_r}} \left(\log\frac{1}{r}\right)^{1-\tau^{N_r}},$$

where $N_r \le \frac{|\Omega|}{\pi r^2}$. Finally by interpolation we get

$$|u(x)| \le \omega(\varepsilon, r) := C(\varepsilon^\delta)^{\tau^{N_r}} \left(\log\frac{1}{r}\right)^{1-\tau^{N_r}} \tag{34}$$

for every $x \in \Omega'_{(r)}$.

It is easy to check that there exist a constant C_0, $0 < C_0 < 1/2$, depending only on the a priori constants such that if $d > 0$ and $\rho \le C_0 d^2$, then at least one of the following conditions is satisfied:

$$B_\rho(P_0) \cap \sigma_1 = \emptyset \quad \text{or} \quad B_\rho(P_1) \cap \sigma_0 = \emptyset. \tag{35}$$

Without loss of generality we might assume that

$$B_\rho(P_0) \cap \sigma_1 = \emptyset.$$

Up to a rigid transformation we might assume that $P_0 = 0$ and that σ_0 lies along the positive x_1 axis. For $t \in (0, 1], \rho \in (0, C_0 d^2)$, let

$$x^+_{t\rho} = \overline{x}_\rho + e_2\frac{\rho t}{2}, \quad x^-_{t\rho} = \overline{x}_\rho - e_2\frac{\rho t}{2},$$

where $\overline{x}_\rho = (\rho/2, 0)$ and $e_2 = (0, 1)$. Let

$$u^{\pm}_{\sigma_0}(\overline{x}_\rho) = \lim_{s\to 0^+} u_{\sigma_0}(\overline{x}_\rho \pm s e_2)$$

and

$$u^{\pm}(\overline{x}_\rho) = \lim_{s\to 0^+} u(\overline{x}_\rho \pm s e_2).$$

We have

$$|[u_{\sigma_0}(\overline{x}_\rho)]_{\sigma_0}| = |[u(\overline{x}_\rho)]_{\sigma_0}| \le |u^+(\overline{x}_\rho) - u(x^+_{t\rho})| + |u^-(\overline{x}_\rho) - u(x^-_{t\rho})| + |u(x^+_{t\rho})| + |u(x^-_{t\rho})|.$$

By applying the second of (31) we get

$$|u^+(\overline{x}_\rho) - u(x^+_{t\rho})| \le Ct$$

and

$$|u^-(\overline{x}_\rho) - u(x^-_{t\rho})| \le Ct,$$

where $C > 0$ depends only on the a priori data. Hence using (34) and the last inequalities we derive

$$|[u_{\sigma_0}(\overline{x}_\rho)]| \le Ct + 2\omega(\varepsilon; t\rho/2).$$

Therefore, from the properties of double layer potentials, we get

$$|\varphi_0| = |[u_{\sigma_0}(\overline{x}_\rho)]| \le Ct + 2\omega(\varepsilon; t\rho/2), \tag{36}$$

where C depends only on the a priori constants.

Now we argue by contradiction. Assume

$$d \ge \eta_\varepsilon,$$

where

$$\eta_\varepsilon := \frac{\sqrt{2}}{\sqrt{C_0}} \left(\frac{8E_1^2 |\log \tau|}{\log |\log \varepsilon^\delta|} \right)^{1/8}.$$

Let

$$\rho = \left(\frac{8E_1^2 |\log \tau|}{\log |\log \varepsilon^\delta|} \right)^{1/4} \qquad \text{and} \qquad t = \left(\frac{8E_1^2 |\log \tau|}{\log |\log \varepsilon^\delta|} \right)^{1/4}.$$

By (36) there exists $\varepsilon_0 > 0$, depending only on the a priori data, such that if $\varepsilon \le \varepsilon_0$ and

$$d \ge \eta_\varepsilon,$$

then

$$|\varphi_0| \le \omega_1(\varepsilon) := C \left(\log |\log(\varepsilon)| \right)^{-1/4}, \tag{37}$$

where C depends only on the a priori constants.

By using (12) for σ_0 and σ_1 and estimating the difference, we can easily see that

$$|\Gamma(P_0,y)\tau_0 f_0| \le |\Gamma(Q_0,y)\tau_0 f_0| + \left|\int_{\sigma_0} \mathbb{C}_0\widehat{\nabla}N(x,y)n_0\cdot\varphi_0\,ds(x)\right| + \left|\int_{\sigma_1} \mathbb{C}_0\widehat{\nabla}N(x,y)n_1\cdot\varphi_1\,ds(x)\right| + |(N(P_0,y)-\Gamma(P_0,y))\tau_0 f_0| + |u(y)| + |N(Q_1,y)\tau_1 f_1| + |N(P_1,y)\tau_1 f_1|.$$

From Lemma 3.2 and (37) we get, for any $y\in\Omega\backslash\sigma_0$

$$\left|\int_{\sigma_0} \mathbb{C}_0\widehat{\nabla}N(x,y)n_0\cdot\varphi_0\,ds(x)\right| \le C\omega_1(\varepsilon)\left|\log\frac{|y-P_0|}{|y-Q_0|}\right|$$

and, for any $y\in\Omega\backslash\sigma_1$,

$$\left|\int_{\sigma_1} \mathbb{C}_0\widehat{\nabla}N(x,y)n_1\cdot\varphi_1\,ds(x)\right| \le C|\varphi_1|\left|\log\frac{|y-P_1|}{|y-Q_1|}\right|,$$

where C depends only on the a priori data. Furthermore, by regularity estimates, we have

$$|N(P_0,y)-\Gamma(P_0,y)| \le C \quad \text{for every} \quad y\in\Omega,$$

where C depends only on the a priori constants. It is now straightforward to see that there exists $\varepsilon_1>0$, depending only on the a priori data, such that if $\varepsilon\le\varepsilon_1$ and

$$d \ge \left(\log(\log|\log\varepsilon^\delta|)\right)^{-1},$$

then

$$|f_0| \le C\omega_1(\varepsilon). \tag{38}$$

Hence, for $\varepsilon\le\varepsilon_2 := \min(\varepsilon_0,\varepsilon_1)$ and $d\ge\left(\log(\log|\log\varepsilon^\delta|)\right)^{-1}$, inequalities (37) and (38) hold true and we get

$$\begin{cases} |n_0^T W\tau_0| \le C\omega_1(\varepsilon), \\ |\alpha(\mathrm{tr}W)-\beta n_0^T W n_0| \le C\omega_1(\varepsilon), \\ |a(\mathrm{tr}W)+(b+d)n_0^T W n_0| \le C\omega_1(\varepsilon), \end{cases}$$

which lead to

$$k_0 = \max_{|x|=1}|x^T W x| \le C\omega_1(\varepsilon),$$

a contradiction for ε small enough. Hence

$$d \le \left(\log(\log|\log\varepsilon^\delta|)\right)^{-1}$$

for every $\varepsilon\le\varepsilon_2$. If $\varepsilon\ge\varepsilon_2$, then

$$d \le E_1 \le E_1\frac{\varepsilon}{\varepsilon_1}$$

and the proof is complete. □

5.2. Proof of Lemma 4.2. Taking the derivative of u_t,

$$
\begin{aligned}
u_t'(y) &= \int_0^1 \left(\frac{d}{dt} \mathbb{C}_0 \widehat{\nabla} N((1-s)P_t + sQ_t, y) \right) n_t \cdot \varphi_t |Q_t - P_t| \, ds \\
&\quad + \int_0^1 \mathbb{C}_0 \widehat{\nabla} N((1-s)P_t + sQ_t, y) n_t \cdot \frac{d}{dt} (\varphi_t |Q_t - P_t|) \, ds \\
&\quad + \int_0^1 \mathbb{C}_0 \widehat{\nabla} N((1-s)P_t + sQ_t, y) \left(\frac{d}{dt} n_t \right) \cdot \varphi_t |Q_t - P_t| \, ds \\
&\quad + \left(\nabla N(Q_t, y) \frac{dQ_t}{dt} \cdot \tau_t - \nabla N(P_t, y) \frac{dP_t}{dt} \cdot \tau_t \right) f_t \qquad (39) \\
&\quad + (N(Q_t, y) - N(P_t, y)) \frac{d(\tau_t f_t)}{dt} \\
&:= I_1 + I_2 + I_3 + \left(\nabla N(Q_t, y) \frac{dQ_t}{dt} \cdot \tau_t - \nabla N(P_t, y) \frac{dP_t}{dt} \cdot \tau_t \right) f_t \\
&\quad + (N(Q_t, y) - N(P_t, y)) \frac{d(\tau_t f_t)}{dt}.
\end{aligned}
$$

Observe that

$$
\begin{aligned}
&\frac{d}{dt} \mathbb{C}_0 \widehat{\nabla} N((1-s)P_t + sQ_t, y) \\
&= \nabla_x (\mathbb{C}_0 \widehat{\nabla} N((1-s)P_t + sQ_t, y)) \bullet ((1-s)(P_1 - P_0) + s(Q_1 - Q_0)),
\end{aligned}
$$

where the symbol $\bullet$ should be intended in the following way:

$$
\nabla_x M \bullet v := \sum_i \frac{\partial M}{\partial x_i} v_i.
$$

On the other hand

$$
\begin{aligned}
\frac{d}{ds} \Big(\mathbb{C}_0 \widehat{\nabla} N((1-s)P_t + sQ_t, y) \Big) &= \nabla_x (\mathbb{C}_0 \widehat{\nabla} N((1-s)P_t + sQ_t, y)) \bullet (Q_t - P_t) \\
&= \nabla_x (\mathbb{C}_0 \widehat{\nabla} N((1-s)P_t + sQ_t, y)) \bullet \tau_t |Q_t - P_t|.
\end{aligned}
$$

Set

$$
(1-s)(P_1 - P_0) + s(Q_1 - Q_0) = A_t(s)\tau_t + B_t(s) n_t,
$$

where

$$
A_t(s) = (1-s)(P_1 - P_0) \cdot \tau_t + s(Q_1 - Q_0) \cdot \tau_t
$$

and

$$
B_t(s) = (1-s)(P_1 - P_0) \cdot n_t + s(Q_1 - Q_0) \cdot n_t.
$$

Hence

$$I_1 = \int_0^1 \left(\nabla_x(\mathbb{C}_0\widehat{\nabla}N((1-s)P_t+sQ_t,y)) \bullet (A_t(s)\tau_t + B_t(s)n_t)\right) n_t\cdot\varphi_t\, ds$$

$$= \int_0^1 A_t(s)\frac{d}{ds}\left(\mathbb{C}_0\widehat{\nabla}N((1-s)P_t+sQ_t,y)\right) n_t\cdot\varphi_t\, ds \tag{40}$$

$$+ \int_0^1 B_t(s)\left(\nabla_x(\mathbb{C}_0\widehat{\nabla}N((1-s)P_t+sQ_t,y)) \bullet n_t\right) n_t\cdot\varphi_t|Q_t-P_t|\, ds.$$

Integrating by parts the first integral appearing on the right-hand side of (40), we obtain

$$I_1 = A_t(1)\mathbb{C}_0\widehat{\nabla}N(Q_t,y)n_t\cdot\varphi_t - A_t(0)\mathbb{C}_0\widehat{\nabla}N(P_t,y)n_t\cdot\varphi_t$$

$$- \int_0^1 \mathbb{C}_0\widehat{\nabla}N((1-s)P_t+sQ_t,y)\frac{d}{ds}\left(A_t(s)n_t\cdot\varphi_t\right)\, ds \tag{41}$$

$$+ \int_0^1 B_t(s)\left(\nabla_x(\mathbb{C}_0\widehat{\nabla}N((1-s)P_t+sQ_t,y)) \bullet n_t\right) n_t\cdot\varphi_t|Q_t-P_t|\, ds.$$

Observing now that, for any $y\in\partial\Omega$,

$$\nabla_x\left(\mathbb{C}_0\widehat{\nabla}N(\cdot,y)n_t\right)\bullet n_t + \nabla_x\left(\mathbb{C}_0\widehat{\nabla}N(\cdot,y)\cdot\tau_t\right)\bullet\tau_t = 0 \quad \text{in} \quad \Omega,$$

we can rewrite

$$\int_0^1 B_t(s)\left(\nabla_x(\mathbb{C}_0\widehat{\nabla}N((1-s)P_t+sQ_t,y)) \bullet n_t\right) n_t\cdot\varphi_t|Q_t-P_t|ds$$

$$= -\int_0^1 B_t(s)\left(\nabla_x(\mathbb{C}_0\widehat{\nabla}N((1-s)P_t+sQ_t,y)) \bullet \tau_t\right) \tau_t\cdot\varphi_t|Q_t-P_t|ds$$

$$= -\int_0^1 B_t(s)\left(\frac{d}{ds}\mathbb{C}_0\widehat{\nabla}N((1-s)P_t+sQ_t,y)\right)\tau_t\cdot\varphi_t ds$$

$$= -B_t(1)\mathbb{C}_0\widehat{\nabla}N(Q_t,y)\tau_t\cdot\varphi_t + B_t(0)\mathbb{C}_0\widehat{\nabla}N(P_t,y)\tau_t\cdot\varphi_t \tag{42}$$

$$+ \int_0^1 \mathbb{C}_0\widehat{\nabla}N((1-s)P_t+sQ_t,y)\frac{d}{ds}\left(B_t(s)\tau_t\cdot\varphi_t\right) ds.$$

Inserting (42) into (41) we derive finally

$$I_1 = \mathbb{C}_0\widehat{\nabla}N(Q_t,y)(A_t(1)n_t - B_t(1)\tau_t)\cdot\varphi_t$$

$$- \mathbb{C}_0\widehat{\nabla}N(P_t,y)(A_t(0)n_t - B_t(0)\tau_t)\cdot\varphi_t \tag{43}$$

$$- \int_0^1 \mathbb{C}_0\widehat{\nabla}N((1-s)P_t+sQ_t,y)\frac{d}{ds}\left(A_t(s)n_t\cdot\varphi_t - B_t(s)\tau_t\cdot\varphi_t\right) ds,$$

where

$$\frac{d}{ds}\left(A_t(s)n_t - B_t(s)\tau_t\right)\cdot\varphi_t$$
$$= \{((Q_1-Q_0)-(P_1-P_0))\cdot\tau_t)n_t - ((Q_1-Q_0)-(P_1-P_0))\cdot n_t)\tau_t\}\cdot\varphi_t,$$
$$A_t(1) = (Q_1-Q_0)\cdot\tau_t, \qquad A_t(0) = (P_1-P_0)\cdot\tau_t$$

and

$$B_t(1) = (Q_1-Q_0)\cdot n_t, \qquad B_t(0) = (P_1-P_0)\cdot n_t.$$

It is easy to calculate

$$\frac{d}{dt}n_t = -\left(\frac{(Q_1-Q_0)-(P_1-P_0)}{|Q_t-P_t|}\cdot n_t\right)\tau_t,$$

and, with this, we get

$$I_3 = -\int_0^1 \mathbb{C}_0\widehat{\nabla}N((1-s)P_t+sQ_t,y)\tau_t\cdot\varphi_t((Q_1-Q_0)-(P_1-P_0))\cdot n_t\,ds. \tag{44}$$

Inserting (43) and (44) into (39) gives (25), which concludes the proof. ∎

5.3. Proof of Proposition 4.3. In Lemma 4.2 we established that

$$\begin{aligned}
u_0'(y) = &-\int_0^1 \mathbb{C}_0\widehat{\nabla}N((1-s)P_0+sQ_0,y)n_0\cdot\varphi_0((Q_1-Q_0)-(P_1-P_0))\cdot\tau_0\\
&+\int_0^1 \mathbb{C}_0\widehat{\nabla}N((1-s)P_0+sQ_0,y)n_0\cdot\frac{d}{dt}(\varphi_t|Q_t-P_t|)|_{t=0}\\
&+\mathbb{C}_0\widehat{\nabla}N(Q_0,y)\left(((Q_1-Q_0)\cdot\tau_0)n_0-((Q_1-Q_0)\cdot n_0)\tau_0\right)\cdot\varphi_0\\
&-\mathbb{C}_0\widehat{\nabla}N(P_0,y)\left(((P_1-P_0)\cdot\tau_0)n_0-((P_1-P_0)\cdot n_0)\tau_0\right)\cdot\varphi_0\\
&+\left(\nabla N(Q_0,y)(Q_1-Q_0)-\nabla N(P_0,y)(P_1-P_0)\right)\cdot\tau_0 f_0\\
&+(N(Q_0,y)-N(P_0,y))\,\frac{d(\tau_t f_t)}{dt}|_{t=0}.
\end{aligned} \tag{45}$$

We divide the proof of the first inequality of Proposition 4.3 into two steps.

(1) We first prove that if

$$\|u_0'\|_{L^2(\partial\Omega)} = 0, \quad \text{then} \quad d_{\mathcal{H}}(\sigma_0,\sigma_1) = 0. \tag{46}$$

Since $d_{\mathcal{H}}(\sigma_0,\sigma_1) = |P_1-P_0|$ it will be sufficient to show that $P_1 = P_0$.

By (24), (46), and the unique continuation property, we have that

$$u_0'(y) = 0 \quad \text{for} \quad y\in\overline{\Omega}\backslash\sigma_0.$$

In particular u_0' is zero in a neighborhood of P_0 contained in $\Omega\backslash\sigma_0$. From (45) it is immediate to see that u_0' has a singularity at P_0 whose leading term is

$$-\mathbb{C}_0\widehat{\nabla}\Gamma(P_0,y)\left((P_1-P_0)^{\perp}\right)\cdot\varphi_0-\nabla\Gamma(P_0,y)(P_1-P_0)\cdot\tau_0 f_0. \tag{47}$$

Without loss of generality let us now fix the coordinate system in such a way that $P_0=0$, $\tau_0=(1,0)$, and $n_0=(0,1)$, let $\varphi_0=(\varphi_0^1,\varphi_0^2)$, and let us set $h_1=(h_1^1,h_1^2):=(P_1-P_0)$. Since the leading singularity given by (47) cannot appear and by the definition of $\mathbb{C}_0$, we must have, for $i=1,2$ and for any $y\in B_r(0)\setminus\sigma_0$ for some small and positive radius r,

$$\alpha_1\,\partial_1\Gamma_{i1}(y)+\alpha_2\,\partial_1\Gamma_{i2}(y)+\alpha_3\,\partial_2\Gamma_{i1}(y)+\alpha_4\,\partial_2\Gamma_{i2}(y)=0, \tag{48}$$

where the coefficients are given by

$$\begin{aligned}&\alpha_1=\lambda_0(h_1^1\varphi_0^2-h_1^2\varphi_0^1)-2\mu_0h_1^2\varphi_0^1+h_1^1f_0, \qquad \alpha_2=-\mu_0h_1^2\varphi_0^2+\mu_0h_1^1\varphi_0^1,\\ &\alpha_3=-\mu_0h_1^2\varphi_0^2+\mu_0h_1^1\varphi_0^1+h_1^2f_0, \qquad \alpha_4=\lambda_0(-h_1^2\varphi_0^1+h_1^1\varphi_0^2)+2\mu_0h_1^1\varphi_0^2.\end{aligned} \tag{49}$$

By calculating explicitly the derivatives of Γ, equations (48) become

$$\begin{aligned}&\frac{1}{2\pi|y|^4}\left[y_1^3(A\alpha_1-B\alpha_4)+y_1^2y_2(B\alpha_2+(A+2B)\alpha_3)+y_1y_2^2((A-2B)\alpha_1+B\alpha_4)\right.\\ &\left.+y_2^3(-B\alpha_2+A\alpha_3)\right]=0 \quad \text{for every} \quad y\in B_r(0)\setminus\sigma_0\end{aligned} \tag{50}$$

for $i=1$, and

$$\begin{aligned}&\frac{1}{2\pi|y|^4}\left[y_1^3(A\alpha_2-B\alpha_3)+y_1^2y_2(B\alpha_1+(A-2B)\alpha_4)+y_1y_2^2((A+2B)\alpha_2+B\alpha_3)\right.\\ &\left.+y_2^3(-B\alpha_1+A\alpha_4)\right]=0 \quad \text{for every} \quad y\in B_r(0)\sigma_0\end{aligned} \tag{51}$$

for $i=2$.

Equations (50) and (51) can be satisfied if and only if each coefficient of the polynomials appearing in the numerators is zero. Since $A+B$ and $A-B$ are different from zero this implies that

$$\alpha_1=\alpha_2=\alpha_3=\alpha_4=0.$$

Easy computations show that these four equations can be satisfied for some vector $h_1\neq(0,0)$ only if

$$f_0=\varphi_0^1=\varphi_0^2=0,$$

which implies that W is identically zero, leading to a contradiction. Hence $h_1=P_1-P_0=0$ and the first step is proved.

(2) Since u_0' depends linearly on $h:=(h_1,h_2):=(P_1-P_0,Q_1-Q_0)$, in order to prove the estimate from below it is sufficient to show that if $|h|_{\mathbb{R}^4}=(|h_1|^2+|h_2|^2)^{1/2}=1$, then

$$\|u_0'\|_{L^2(\partial\Omega)}\geq m_0,$$

where $m_0>0$ depends only on the a priori data.

Let $h \in \mathbb{R}^4$ be such that $|h|_{\mathbb{R}^4} = 1$, where $d_{\mathcal{H}}(\sigma_0, \sigma_1) = |h_1|_{\mathbb{R}^2} \geq |h_2|_{\mathbb{R}^2}$. Denote

$$v = u_0', \qquad \text{and} \qquad \theta = \|v\|_{L^2(\partial\Omega)}.$$

Recall that v is a solution of (24) and proceed similarly as in Lemma 4.1. If we define $\rho_1 := \min\left(L^{-1}, \rho_0\right)$, we have, for any $r \in (0, \rho_1/156)$,

$$|v(x)| \leq \omega(\theta, r) := C(\theta^\delta)^{\tau^{N_r}} \left(\log \frac{1}{r}\right)^{1-\tau^{N_r}}$$

for every $x \in \Omega_r' := \Omega_{8r} \backslash (\sigma_0)_{8r}$, where $(\sigma_0)_{8r}$ is defined as in (33), $\tau, \delta \in (0,1)$ and $C > 1$ depend on the a priori data only, and $N_r \leq \frac{|\Omega|}{\pi r^2}$. Without loss of generality we might assume that $P_0 = 0$. By (44) we have that

$$v(y) = \frac{H(y)}{|y|^4} + J(y) \quad \text{for} \quad y \in \Omega \backslash \sigma_0, \tag{52}$$

where H is an homogeneous polynomial of degree three whose coefficients are equal to the expressions in (50) and (51), and J satisfies the inequality

$$|J(y)| \leq C\,|\log|y|| \quad \text{for} \quad y \in \Omega\backslash\sigma_0,\ |y| \leq \frac{1}{2L},$$

where C depends on the a priori data only. Let $\overline{y} \in \mathbb{R}^2$, $|y| = 1$ such that

$$|H(\overline{y})| = \max_{|y|=1} |H(y)|.$$

Denote

$$M = |H(\overline{y})|, \quad y_r = 8r\overline{y}.$$

Since $|H(y_r)| = |H(-y_r)|$ and y_r or $-y_r \in \Omega_r'$, it is not restrictive to assume that $y_r \in \Omega_r'$.

By (52) we have

$$|H(\overline{y})| = \frac{|H(y_r)|}{|y_r|^3} \leq |y_r||v(y_r)| + |y_r||J(y_r)|,$$

and hence

$$M \leq 8r\omega(\theta; r) + Cr \log \frac{1}{r} \tag{53}$$

for every $0 \leq r \leq \rho_2$, where $\rho_2 = \frac{1}{256} \min\left(L^{-1}, \rho_0\right)$. Denote

$$r_\theta = \left(\frac{E_1^2 |\log\tau|}{\log|\log\theta^\delta|}\right)^{1/2}.$$

There exists $\theta_0 < e^{-1/\delta}$, θ_0 depending only on the a priori data, such that if $\theta \in (0, \theta_0)$, then $r_\theta \in (0, \rho_2)$. We now choose, for $\theta \in (0, \theta_0)$, $r = r_\theta$ and by (53) we have

$$M \leq C\omega_2(\theta), \tag{54}$$

where

$$\omega_2(\theta) = \left(\log|\log\theta^\delta|\right)^{-1/4}$$

and C depends only on the a priori data. This implies that each coefficient of the polynomials appearing in the numerators of (50) and (51) are bounded by $C\omega_2(\theta)$ for some positive constant C depending only on a priori data.

Now, proceeding as in step (1) we derive easily that

$$\text{if} \quad \theta \in (0,\theta_0), \quad \text{then} \quad |h_2| \le |h_1| \le C\omega_2(\theta), \tag{55}$$

where C depends only on the a priori data. On the other hand, since $|h_1|^2+|h_2|^2 = 1$, by (55) we obtain that

$$\text{if } \theta \in (0,\theta_0), \quad \text{then} \quad 1 \le 2C_0^2(\omega_2(\theta))^2. \tag{56}$$

Denote by $\theta_1 = \min\{\theta_0, \exp(-\frac{1}{\delta}\exp(16C_0^4))\}$. By (55) we get that if $\theta \in (0,\theta_1)$, then $1 \le 2C_0^2(\omega_2(\theta))^2 \le 1/2$, a contradiction. Therefore $\theta \ge \theta_1$ and recalling that $v = u_0'$, and $\theta = \|v\|_{L^2(\partial\Omega)}$, we get the thesis for $m_0 = \theta_1$.

Let us finally prove (26). Observe that u_t' is differentiable in t and taking, for instance, the following term appearing in $u_t' - u_0'$ we get

$$\begin{aligned}
&|\nabla N(Q_t,y)(Q_1-Q_0)\cdot\tau_t f_t - \nabla N(Q_0,y)(Q_1-Q_0)\cdot\tau_0 f_0| \\
&\quad\le |(\nabla N(Q_t,y)-\nabla N(Q_0,y))||(Q_1-Q_0)\cdot\tau_t f_t| \\
&\qquad + |\nabla N(Q_0,y)(Q_1-Q_0)\cdot(\tau_t-\tau_0) f_t| \\
&\qquad + |\nabla N(Q_0,y)(Q_1-Q_0)\cdot\tau_0(f_t-f_0)|.
\end{aligned}$$

Since

$$|\nabla N(Q_t,y)-\nabla N(Q_0,y)| \le C|Q_1-Q_0| \le Cd_{\mathcal{H}}(\sigma_0,\sigma_1),$$
$$|\tau_t-\tau_0| \le C|P_1-P_0| = Cd_{\mathcal{H}}(\sigma_0,\sigma_1),$$

and, also,

$$|f_t-f_0| \le C|Q_1-Q_0| \le Cd_{\mathcal{H}}(\sigma_0,\sigma_1),$$

where C is a constant depending only on the a priori data, hence

$$|\nabla N(Q_t,y)(Q_1-Q_0)\cdot\tau_t f_t - \nabla N(Q_0,y)(Q_1-Q_0)\cdot\tau_0 f_0| \le C\left(d_{\mathcal{H}}(\sigma_0,\sigma_1)\right)^2.$$

Evaluating the other terms appearing in $u_t' - u_0'$ in a similar way, we get the desired inequality and the proof is concluded. □

REFERENCES

[ABF04] H. Ammari, E. Beretta, and E. Francini, *Reconstruction of thin conductivity imperfections*, Appl. Anal., 83 (2004), pp. 63–78.

[ABF06] H. Ammari, E. Beretta, and E. Francini, *Reconstruction of thin conductivity imperfections*, II. *The case of multiple segments*, Appl. Anal., 85 (2006), pp. 87–105.

[AK04] H. Ammari and H. Kang, *Reconstruction of Small Inhomogeneities from Boundary Measurements*, Lecture Notes in Math. 1846, Springer-Verlag, Berlin, 2004.

[AKNT02] H. Ammari, H. Kang, G. Nakamura, and K. Tanuma, *Complete asymptotic expansions of solutions of the system of elastostatics in the presence of an inclusion of small diameter and detection of an inclusion*, J. Elasticity, 67 (2002), pp. 97–129.

[AM01] G. Alessandrini and A. Morassi, *Strong unique continuation for the Lamé system of elasticity*, Comm. Partial Differential Equations, 26 (2001), pp. 1787–1810.

[BF06] E. Beretta and E. Francini, *An asymptotic formula for the displacement field in the presence of thin elastic inhomogeneities*, SIAM J. Math. Anal., 38 (2006), pp. 1249–1261.

[C80] S. Campanato, *Sistemi ellittici in forma divergenza. Regolarità all'interno*, Quaderni della Scuola Normale Superiore di Pisa, 1980.

[CMV98] D. J. Cedio-Fengya, S. Moskow, and M. Vogelius, *Identification of conductivity imperfections of small diameter by boundary measurements. Continuous dependence and computational reconstruction*, Inverse Problems, 14 (1998), pp. 553–595.

[F72] G. Fichera, *Existence theorems in elasticity*, in Handbuch der Physik, Vol. VI, Springer-Verlag, Berlin, Heidelberg, New York, 1972, pp. 347–389.

[FV89] A. Friedman and M. Vogelius, *Identification of small inhomogeneities of extreme conductivity by boundary measurements: A theorem on continuous dependence*, Arch. Ration. Mech. Anal., 105 (1989), pp. 299–326.

[G72] M. E. Gurtin, *The linear theory of elasticity*, in Handbuch der Physik, Vol. VI, Springer-Verlag, Berlin, Heidelberg, New York, 1972, pp. 1–295.

[MR04] A. Morassi and E. Rosset, *Stable determination of cavities in elastic bodies*, Inverse Problems, 20 (2004), pp. 453–480.

SIAM J. MATH. ANAL.
Vol. 40, No. 3, pp. 1003–1028

NEUTRAL FUNCTIONAL DIFFERENTIAL EQUATIONS WITH APPLICATIONS TO COMPARTMENTAL SYSTEMS*

VÍCTOR MUÑOZ-VILLARRAGUT†, SYLVIA NOVO†, AND RAFAEL OBAYA†

Abstract. We study the monotone skew-product semiflow generated by a family of neutral functional differential equations with infinite delay and stable D-operator. The stability properties of D allow us to introduce a new order and to take the neutral family to a family of functional differential equations with infinite delay. Next, we establish the 1-covering property of omega-limit sets under the componentwise separating property and uniform stability. Finally, the obtained results are applied to the study of the long-term behavior of the amount of material within the compartments of a neutral compartmental system with infinite delay.

Key words. nonautonomous dynamical systems, monotone skew-product semiflows, neutral functional differential equations, infinite delay, compartmental systems

AMS subject classifications. 37B55, 34K40, 34K14

DOI. 10.1137/070711177

1. Introduction.

After the pioneering work of Hale and Meyer [11], the theory of neutral functional differential equations (NFDE) aroused considerable interest and a fast development ensued. At present a wide collection of theoretical and practical results make up the main body of the theory of NFDEs (see Hale [10], Hale and Verduyn Lunel [12], Kolmanovskii and Myshkis [18], and Salamon [22], among many others). In particular, a substantial number of results for delayed functional differential equations (FDEs) have been generalized for NFDEs solving new and challenging problems in these extensions.

In this paper we provide a dynamical theory for nonautonomous monotone NFDEs with infinite delay and autonomous stable D-operator along the lines of the results by Jiang and Zhao [17] and Novo, Obaya, and Sanz [20]. We assume some recurrence properties on the temporal variation of the NFDE. Thus, its solutions induce a skew-product semiflow with minimal flow on the base. In particular, the uniform almost periodic and almost automorphic cases are included in this formulation. The skew-product formalism permits the analysis of the dynamical properties of the trajectories using methods of ergodic theory and topological dynamics.

Novo et al. [20] study the existence of recurrent solutions of nonautonomous FDEs with infinite delay using the phase space $BU \subset C((-\infty, 0], \mathbb{R}^m)$ of bounded and uniformly continuous functions with the supremum norm. Assuming some technical conditions on the vector field, it is shown that every bounded solution is relatively compact for the compact-open topology, and its omega-limit set admits a flow extension. An alternative method for the study of recurrent solutions of almost periodic FDEs with infinite delay makes use of a *fading memory* Banach phase space (see Hino, Murakami, and Naiko [13] for an axiomatic definition and main properties). Since this kind of space contains BU and, under natural assumptions, the restriction of the norm

*Received by the editors December 17, 2007; accepted for publication (in revised form) June 12, 2008; published electronically October 13, 2008. This work was partially supported by Junta de Castilla y León under project VA024A06 and by the M.E.C. under project MTM2005-02144.
http://www.siam.org/journals/sima/40-3/71117.html

†Departamento de Matemática Aplicada, E.T.S. de Ingenieros Industriales, Universidad de Valladolid, 47011 Valladolid, Spain (vicmun@wmatem.eis.uva.es, sylnov@wmatem.eis.uva.es, rafoba@wmatem.eis.uva.es).

topology to the closure of a bounded solution agrees with the compact-open topology, it seems that the approach considered in Novo et al. [20] becomes natural in many cases of interest.

In this paper we consider NFDEs with linear autonomous operator D defined on BU which is continuous for the norm, continuous for the compact-open topology on bounded sets, and atomic at zero. We obtain an integral representation of D by means of Riesz theorem $Dx = \int_{-\infty}^{0} [\, d\mu]\, x$, where μ is a real Borel measure with finite total variation. The convolution operator $\widehat{D}$ defined by $\widehat{D}x(s) = \int_{-\infty}^{0} [d\mu(\theta)]\, x(\theta + s)$ maps BU into BU. We prove that if D is stable in the sense of Hale [10], then $\widehat{D}$ is an isomorphism of BU which is continuous for the norm and continuous for the compact-open topology on bounded sets. Moreover, $\widehat{D}^{-1}$ inherits these same properties and is associated to a linear stable operator D^*. In fact, the mentioned behavior of $\widehat{D}^{-1}$ characterizes the stability of the operator D. The proofs are self-contained and require only quantitative estimates associated to the stability of the operator D.

Staffans [24] shows that every NFDE with finite delay and autonomous stable D-operator can be written as an FDE with infinite delay in an appropriate fading memory space. A more systematic study on the inversion of the convolution operator $\widehat{D}$ can be found in the work of Gripenberg, Londen, and Staffans [4]. The papers by Haddock et al. [8, 9], Arino and Bourad [1], among others, make a systematic use of these ideas which have a theoretical and practical interest. We give a version of the above results for infinite delay NFDEs. It is obvious that the inversion of the convolution operator $\widehat{D}$ on BU allows us to transform the original equation into a retarded nonautonomous FDE with infinite delay. In addition, we transfer the dynamical theory of Jiang and Zhao [17] and Novo et al. [20] to nonautonomous monotone NFDEs with infinite delay and autonomous stable D-operator. In an appropriate dynamical framework we assume that the trajectories are bounded, uniformly stable, and satisfy a componentwise separating property, and we show that the omega-limit sets are all copies of the base. It is important to mention that no conditions of strong monotonicity are required, which permits the application of the results under natural physical conditions.

In this paper we provide a detailed description of the long-term behavior of the dynamics in some classes of compartmental systems extensively studied in the literature. Compartmental systems have been used as mathematical models for the study of the dynamical behavior of many processes in biological and physical sciences, which depend on local mass balance conditions (see Jacquez [14], Jacquez and Simon [15, 16], and the references therein). Some initial results for models described by FDEs with finite and infinite delay can be found in Györi [5] and Györi and Eller [6]. The paper by Arino and Haourigui [2] proves the existence of almost periodic solutions for compartmental systems described by almost periodic finite delay FDEs. NFDEs represent systems where the compartments produce or swallow material. Györi and Wu [7] and Wu and Freedman [28] study autonomous NFDEs with finite and infinite delay similar to those considered in this paper. We provide a nonautonomous version, under more general assumptions, of the monotone theory for NFDEs included in Wu and Freedman [28] and Wu [27]. More precise results for the case of scalar NFDEs can be found in Arino and Bourad [1] and Krisztin and Wu [19].

We study the dynamics of monotone compartmental systems in terms of the geometrical structure of the pipes connecting the compartments. Irreducible subsets of the set of indices detect the occurrence of subsystems on the complete system which reduce the dimension of the problem being studied. When the system is closed,

the total mass is an invariant continuous function which implies the stability and boundedness of the solutions. In particular, the omega-limit set of every solution is a minimal set and a copy of the base. In a general compartmental system the existence of a bounded solution assures that every solution is bounded and uniformly stable. We first check that when there is no inflow of material then all the compartments of each irreducible subset with some outflow of material are empty on minimal subsets. On the contrary, when the solutions remain bounded and there is an inflow of material in some compartment of an irreducible subset, then for the indices of this irreducible subset all the minimal sets agree, and all the compartments contain some material. Finally, we describe natural physical conditions which ensure the existence of a unique minimal set asymptotically stable.

This paper is organized as follows. Basic notions in topological dynamics, used throughout the rest of the sections, are stated in section 2. Section 3 is devoted to the study of general and stability properties of linear autonomous operators from BU to $\mathbb{R}^m$, as well as the behavior of solutions of the corresponding homogeneous and nonhomogeneous equations given by them. In section 4, we study the monotone skew-product semiflow generated by a family of NFDEs with infinite delay and stable D-operator. In particular, we establish the 1-covering property of omega-limit sets under the componentwise separating property and uniform stability. These results are applied in section 5 to show that the solutions of a compartmental system given by a monotone NFDE with infinite delay are asymptotically of the same type as the transport functions. Finally, section 6 deals with the long-term behavior of its solutions in terms of the geometrical structure of the pipes, as explained above.

2. Some preliminaries. Let (Ω, d) be a compact metric space. A real *continuous flow* $(\Omega, \sigma, \mathbb{R})$ is defined by a continuous mapping $\sigma : \mathbb{R} \times \Omega \to \Omega,\ (t, \omega) \mapsto \sigma(t, \omega)$ satisfying

(i) $\sigma_0 = \mathrm{Id}$,

(ii) $\sigma_{t+s} = \sigma_t \circ \sigma_s$ for each $s,\ t \in \mathbb{R}$,

where $\sigma_t(\omega) = \sigma(t, \omega)$ for all $\omega \in \Omega$ and $t \in \mathbb{R}$. The set $\{\sigma_t(\omega) \mid t \in \mathbb{R}\}$ is called the *orbit* or the *trajectory* of the point ω. We say that a subset $\Omega_1 \subset \Omega$ is *σ-invariant* if $\sigma_t(\Omega_1) = \Omega_1$ for every $t \in \mathbb{R}$. A subset $\Omega_1 \subset \Omega$ is called *minimal* if it is compact, σ-invariant, and its only nonempty compact σ-invariant subset is itself. Every compact and σ-invariant set contains a minimal subset; in particular it is easy to prove that a compact σ-invariant subset is minimal if and only if every trajectory is dense. We say that the continuous flow $(\Omega, \sigma, \mathbb{R})$ is *recurrent* or *minimal* if Ω is minimal.

The flow $(\Omega, \sigma, \mathbb{R})$ is *distal* if for any two distinct points $\omega_1,\ \omega_2 \in \Omega$ the orbits keep at a positive distance, that is, $\inf_{t\in\mathbb{R}} d(\sigma(t, \omega_1), \sigma(t, \omega_2)) > 0$. The flow $(\Omega, \sigma, \mathbb{R})$ is *almost periodic* when for every $\varepsilon > 0$ there is a $\delta > 0$ such that if $\omega_1,\ \omega_2 \in \Omega$ with $d(\omega_1, \omega_2) < \delta$, then $d(\sigma(t, \omega_1), \sigma(t, \omega_2)) < \varepsilon$ for every $t \in \mathbb{R}$. If $(\Omega, \sigma, \mathbb{R})$ is almost periodic, it is distal. The converse is not true; even if $(\Omega, \sigma, \mathbb{R})$ is minimal and distal, it does not need to be almost periodic. For the basic properties of almost periodic and distal flows we refer the reader to Ellis [3] and Sacker and Sell [21].

A *flow homomorphism* from another continuous flow $(Y, \Psi, \mathbb{R})$ to $(\Omega, \sigma, \mathbb{R})$ is a continuous map $\pi\colon Y \to \Omega$ such that $\pi(\Psi(t, y)) = \sigma(t, \pi(y))$ for every $y \in Y$ and $t \in \mathbb{R}$. If π is also bijective, it is called a *flow isomorphism*. Let $\pi : Y \to \Omega$ be a surjective flow homomorphism and suppose $(Y, \Psi, \mathbb{R})$ is minimal (then so is $(\Omega, \sigma, \mathbb{R})$). $(Y, \Psi, \mathbb{R})$ is said to be an *almost automorphic extension* of $(\Omega, \sigma, \mathbb{R})$ if there is $\omega \in \Omega$ such that $\mathrm{card}(\pi^{-1}(\omega)) = 1$. Then actually $\mathrm{card}(\pi^{-1}(\omega)) = 1$ for ω in a residual subset $\Omega_0 \subseteq \Omega$; in the nontrivial case $\Omega_0 \subsetneq \Omega$ the dynamics can be very complicated. A minimal flow

$(Y, \Psi, \mathbb{R})$ is *almost automorphic* if it is an almost automorphic extension of an almost periodic minimal flow $(\Omega, \sigma, \mathbb{R})$. We refer the reader to the work of Shen and Yi [23] for a survey of almost periodic and almost automorphic dynamics.

Let E be a complete metric space and $\mathbb{R}^+ = \{t \in \mathbb{R} \,|\, t \geq 0\}$. A *semiflow* $(E, \Phi, \mathbb{R}^+)$ is determined by a continuous map $\Phi : \mathbb{R}^+ \times E \to E,\ (t, x) \mapsto \Phi(t, x)$ which satisfies

(i) $\Phi_0 = \mathrm{Id}$,

(ii) $\Phi_{t+s} = \Phi_t \circ \Phi_s$ for all $t,\, s \in \mathbb{R}^+$,

where $\Phi_t(x) = \Phi(t, x)$ for each $x \in E$ and $t \in \mathbb{R}^+$. The set $\{\Phi_t(x) \mid t \geq 0\}$ is the *semiorbit* of the point x. A subset E_1 of E is *positively invariant* (or just Φ-*invariant*) if $\Phi_t(E_1) \subset E_1$ for all $t \geq 0$. A semiflow $(E, \Phi, \mathbb{R}^+)$ admits a *flow extension* if there exists a continuous flow $(E, \widetilde{\Phi}, \mathbb{R})$ such that $\widetilde{\Phi}(t, x) = \Phi(t, x)$ for all $x \in E$ and $t \in \mathbb{R}^+$. A compact and positively invariant subset admits a flow extension if the semiflow restricted to it admits one.

Write $\mathbb{R}^- = \{t \in \mathbb{R} \,|\, t \leq 0\}$. A *backward orbit* of a point $x \in E$ in the semiflow $(E, \Phi, \mathbb{R}^+)$ is a continuous map $\psi : \mathbb{R}^- \to E$ such that $\psi(0) = x$ and for each $s \leq 0$ it holds that $\Phi(t, \psi(s)) = \psi(s + t)$ whenever $0 \leq t \leq -s$. If for $x \in E$ the semiorbit $\{\Phi(t, x) \mid t \geq 0\}$ is relatively compact, we can consider the *omega-limit set* of x,

$$\mathcal{O}(x) = \bigcap_{s \geq 0} \mathrm{closure}\{\Phi(t + s, x) \mid t \geq 0\}\,,$$

which is a nonempty compact connected and Φ-invariant set. Namely, it consists of the points $y \in E$ such that $y = \lim_{n\to\infty} \Phi(t_n, x)$ for some sequence $t_n \uparrow \infty$. It is well known that every $y \in \mathcal{O}(x)$ admits a backward orbit inside this set. Actually, a compact positively invariant set M admits a flow extension if every point in M admits a unique backward orbit which remains inside the set M (see Shen and Yi [23, part II]).

A compact positively invariant set M for the semiflow $(E, \Phi, \mathbb{R}^+)$ is *minimal* if it does not contain any nonempty compact positively invariant set other than itself. If E is minimal, we say that the semiflow is minimal.

A semiflow is of *skew-product type* when it is defined on a vector bundle and has a triangular structure; more precisely, a semiflow $(\Omega \times X, \tau, \mathbb{R}^+)$ is a *skew-product* semiflow over the product space $\Omega \times X$, for a compact metric space (Ω, d) and a complete metric space (X, d), if the continuous map τ is as follows:

$$\begin{array}{rccc} \tau\colon & \mathbb{R}^+ \times \Omega \times X & \longrightarrow & \Omega \times X \\ & (t, \omega, x) & \mapsto & (\omega{\cdot}t, u(t, \omega, x))\,, \end{array} \tag{2.1}$$

where $(\Omega, \sigma, \mathbb{R})$ is a real continuous flow $\sigma : \mathbb{R} \times \Omega \to \Omega,\ (t, \omega) \mapsto \omega{\cdot}t$, called the *base flow*. The skew-product semiflow (2.1) is *linear* if $u(t, \omega, x)$ is linear in x for each $(t, \omega) \in \mathbb{R}^+ \times \Omega$.

Now, we introduce some definitions concerning the stability of the trajectories. A forward orbit $\{\tau(t, \omega_0, x_0) \mid t \geq 0\}$ of the skew-product semiflow (2.1) is said to be *uniformly stable* if for every $\varepsilon > 0$ there is a $\delta(\varepsilon) > 0$, called the *modulus of uniform stability*, such that if $s \geq 0$ and $\mathsf{d}(u(s, \omega_0, x_0), x) \leq \delta(\varepsilon)$ for certain $x \in X$, then for each $t \geq 0$,

$$\mathsf{d}(u(t + s, \omega_0, x_0), u(t, \omega_0{\cdot}s, x)) = \mathsf{d}(u(t, \omega_0{\cdot}s, u(s, \omega_0, x_0)), u(t, \omega_0{\cdot}s, x)) \leq \varepsilon\,.$$

A forward orbit $\{\tau(t, \omega_0, x_0) \mid t \geq 0\}$ of the skew-product semiflow (2.1) is said to be *uniformly asymptotically stable* if it is uniformly stable and there is a $\delta_0 > 0$ with

the following property: for each $\varepsilon > 0$ there is a $t_0(\varepsilon) > 0$ such that if $s \geq 0$ and $\mathsf{d}(u(s, \omega_0, x_0), x) \leq \delta_0$, then

$$\mathsf{d}(u(t+s, \omega_0, x_0), u(t, \omega_0 \cdot s, x)) \leq \varepsilon \quad \text{for each } t \geq t_0(\varepsilon)\,.$$

3. Stable *D*-operators. We consider the Fréchet space $X = C((-\infty, 0], \mathbb{R}^m)$ endowed with the compact-open topology, i.e., the topology of uniform convergence over compact subsets, which is a metric space for the distance

$$\mathsf{d}(x, y) = \sum_{n=1}^{\infty} \frac{1}{2^n} \frac{\|x - y\|_n}{1 + \|x - y\|_n}\,, \quad x, y \in X\,,$$

where $\|x\|_n = \sup_{s \in [-n, 0]} \|x(s)\|$ and $\|\cdot\|$ denotes the maximum norm on $\mathbb{R}^m$.

Let $BU \subset X$ be the Banach space

$$BU = \{x \in X \mid x \text{ is bounded and uniformly continuous}\}$$

with the supremum norm $\|x\|_\infty = \sup_{s \in (-\infty, 0]} \|x(s)\|$. Given $r > 0$, we will denote

$$B_r = \{x \in BU \mid \|x\|_\infty \leq r\}\,.$$

As usual, given $I = (-\infty, a] \subset \mathbb{R}$, $t \in I$, and a continuous function $x : I \to \mathbb{R}^m$, x_t will denote the element of X defined by $x_t(s) = x(t + s)$ for $s \in (-\infty, 0]$.

This section is devoted to the study of general and stability properties of linear autonomous operators $D\colon BU \to \mathbb{R}^m$, as well as the behavior of solutions of the corresponding homogeneous equation $Dx_t = 0$, $t \geq 0$, and nonhomogeneous equations $Dx_t = h(t)$, $t \geq 0$, for $h \in C([0, \infty), \mathbb{R}^m)$. We will assume the following:

(D1) D is linear and continuous for the norm.

(D2) For each $r > 0$, $D\colon B_r \to \mathbb{R}^m$ is continuous when we take the restriction of the compact-open topology to B_r; i.e., if $x_n \overset{\mathsf{d}}{\to} x$ as $n \to \infty$ with x_n, $x \in B_r$, then $\lim_{n \to \infty} Dx_n = Dx$.

(D3) D is atomic at 0 (see the definition in Hale [10] or Hale and Verduyn Lunel [12]).

From (D1) and (D2) we obtain the following representation.

PROPOSITION 3.1. *If $D\colon BU \to \mathbb{R}^m$ satisfies* (D1) *and* (D2)*, then for each $x \in BU$*

$$Dx = \int_{-\infty}^{0} [d\mu(s)]\, x(s),$$

where $\mu = [\mu_{ij}]$ and μ_{ij} is a real regular Borel measure with finite total variation $|\mu_{ij}|(-\infty, 0] < \infty$ for all i, $j \in \{1, \ldots, m\}$.

Proof. From Riesz representation theorem we obtain the above relation for each x whose components are of compact support. Moreover, if $x \in BU$, there are an $r > 0$ and a sequence of functions of compact support $\{x_n\}_{n \in \mathbb{N}} \subset B_r$ with $\|x_n\|_\infty \leq \|x\|_\infty$ such that $x_n \overset{\mathsf{d}}{\to} x$ as $n \to \infty$ and, from hypothesis (D2), $\lim_{n \to \infty} Dx_n = Dx$. However,

$$Dx_n = \int_{-\infty}^{0} [d\mu(s)]\, x_n(s),$$

and the Lebesgue-dominated convergence theorem yields

$$\lim_{n \to \infty} Dx_n = \int_{-\infty}^{0} [d\mu(s)]\, x(s)\,,$$

which finishes the proof. □

Since in addition D is atomic at 0, $\det[\mu_{ij}(\{0\})] \neq 0$, and without loss of generality, we may assume that

$$Dx = x(0) - \int_{-\infty}^{0} [d\nu(s)]\, x(s), \tag{3.1}$$

where $\nu = [\nu_{ij}]_{i,j\in\{1,\dots,m\}}$, ν_{ij} is a real regular Borel measure with finite total variation, and $|\nu_{ij}|(\{0\}) = 0$ for all $i,\ j \in \{1,\dots,m\}$. We will denote by $|\nu|[-r,0]$ the $m \times m$ matrix $[\,|\nu_{ij}|[-r,0]\,]$ and by $\|\nu\|_\infty[-r,0]$ the corresponding matricial norm.

From now on, we will assume that the operator D satisfying (D1)–(D3) has the form (3.1). First, it is easy to check the following result, whose proof is omitted.

PROPOSITION 3.2. *For all $h \in C([0,\infty),\mathbb{R}^m)$ and $\varphi \in BU$ with $D\varphi = h(0)$, the nonhomogeneous equation*

$$\begin{cases} Dx_t = h(t)\,, \quad t \geq 0\,, \\ x_0 = \varphi \end{cases} \tag{3.2}$$

has a solution defined for all $t \geq 0$.

Next we obtain a bound for the solution in a finite interval $[0,T]$, in terms of the initial data and the independent term h, which in particular implies the uniqueness of the solution of (3.2).

LEMMA 3.3. *Given $T > 0$, there are positive constants k_T^1, k_T^2 such that if x is a solution of (3.2), then for each $t \in [0,T]$*

$$\|x_t\|_\infty \leq k_T^1 \sup_{0\leq u\leq t} \|h(u)\| + k_T^2\, \|\varphi\|_\infty\,. \tag{3.3}$$

Proof. Since $|\nu_{ij}|[-r,0] \to 0$ as $r \to 0$, for each $i,\ j \in \{1,\dots,m\}$, there is an $r > 0$ such that $\|\nu\|_\infty[-r,0] < 1/2$. Let x be a solution of (3.2). From (3.1),

$$x(t) = h(t) + \int_{-t}^{0} [d\nu(s)]\, x(t+s) + \int_{-\infty}^{-t} [d\nu(s)]\, \varphi(t+s)$$

for each $t \geq 0$. Consequently, if $t \in [0,r]$,

$$\|x(t)\| \leq \|h(t)\| + \frac{1}{2} \sup_{0\leq u\leq t} \|x(u)\| + \|\varphi\|_\infty \|\nu\|_\infty(-\infty,0]\,,$$

from which we deduce that if $t \in [0,r]$,

$$\sup_{0\leq u\leq t} \|x(u)\| \leq 2 \sup_{0\leq u\leq t} \|h(u)\| + 2\,a\,\|\varphi\|_\infty\,, \tag{3.4}$$

where $a = \|\nu\|_\infty(-\infty,0]$. Next, let $y(t) = x(t+r)$, which is a solution of

$$\begin{cases} Dy_t = h(t+r)\,, \quad t \geq 0\,, \\ y_0 = x_r\,. \end{cases}$$

As above, we conclude that if $t \in [0,r]$,

$$\sup_{0\leq u\leq t} \|y(u)\| \leq 2 \sup_{0\leq u\leq t} \|h(u+r)\| + 2\,a\,\|x_r\|_\infty\,,$$

which together with $\|x_r\|_\infty \le \|\varphi\|_\infty + \sup_{0\le u\le r}\|x(u)\|$ and (3.4) yields

$$\sup_{0\le u\le t}\|x(u)\| \le b \sup_{0\le u\le t}\|h(u)\| + c\,\|\varphi\|_\infty$$

for $t \in [r, 2\,r]$ and some positive constants b and c independent of h and φ. This way, the result is obtained in a finite number of steps. □

Following Hale [10], we introduce the concept of stability for the operator D. Although the initial definition is given for the homogeneous equation, it is easy to deduce quantitative estimates in terms of the initial data for the solution of a nonhomogeneous equation.

DEFINITION 3.4. *The linear operator D is said to be* stable *if there is a continuous function $c \in C([0,\infty), \mathbb{R}^+)$ with $\lim_{t\to\infty} c(t) = 0$ such that, for each $\varphi \in BU$ with $D\varphi = 0$, the solution of the homogeneous problem*

$$\begin{cases} Dx_t = 0\,, \quad t \ge 0, \\ x_0 = \varphi \end{cases}$$

satisfies $\|x(t)\| \le c(t)\,\|\varphi\|_\infty$ for each $t \ge 0$.

PROPOSITION 3.5. *Let us assume that D is stable. Then there is a positive constant $d > 0$ such that, for each $h \in C([0,\infty), \mathbb{R}^m)$ with $h(0) = 0$, the solution of*

$$\begin{cases} Dx_t = h(t)\,, \quad t \ge 0\,, \\ x_0 = 0 \end{cases}$$

satisfies $\|x(t)\| \le d \sup_{0\le u\le t}\|h(u)\|$ for each $t \ge 0$.

Proof. Let $\{e_1, \dots, e_m\}$ be the canonical basis of $\mathbb{R}^m$. A proof similar to that of Lemma 3.2 of Hale [10, sec. 12] shows that there are m functions $\phi_1, \dots, \phi_m \in BU$ such that $D\phi_j = e_j$ for each $j \in \{1, \dots, m\}$. We will denote by Φ the $m \times m$ matrix function $\Phi = [\phi_1, \dots, \phi_m]$ and by $\|\Phi\|_\infty$ the matricial norm corresponding to the norm $\|\cdot\|_\infty$ on BU.

Let $c \in C([0,\infty), \mathbb{R}^m)$ be the function given in Definition 3.4. Assume that c is decreasing and take $T > 0$ such that $c(T) < 1$. From Lemma 3.3, $\|x(t)\| \le k_T^1 \sup_{0\le u\le t}\|h(u)\|$, provided that $t \in [0, T]$.

If $t \ge T$, there is a $j \in \mathbb{N}$ such that $t \in [j\,T, (j+1)\,T]$ and it is easy to check that $x(t) = x^1(t - (j-1)\,T) + x^2(t - (j-1)\,T)$, where x^1 and x^2 are the solutions of

$$\begin{cases} Dx_t^1 = 0\,, \quad t \ge 0\,, \\ x_0^1 = x_{(j-1)\,T} - \Phi\,h((j-1)\,T)\,, \end{cases} \qquad \begin{cases} Dx_t^2 = h(t + (j-1)\,T)\,, \quad t \ge 0\,, \\ x_0^2 = \Phi\,h((j-1)\,T)\,, \end{cases}$$

respectively. From the stability of D and Lemma 3.3, we deduce that

$$\|x(t)\| \le c(t - (j-1)\,T)\,\|x_{(j-1)\,T} - \Phi\,h((j-1)\,T)\|_\infty + k_{2T} \sup_{(j-1)\,T\le u\le t}\|h(u)\|\,.$$

In addition, since $t - (j-1)\,T \ge T$ and c is decreasing we conclude that

$$\|x(t)\| \le c(T)\,c_j + (c(T)\,\|\Phi\|_\infty + k_{2T}) \sup_{0\le u\le t}\|h(u)\|\,, \quad t \in [j\,T, (j+1)\,T], \tag{3.5}$$

where $c_j = \|x_{jT}\|_\infty = \sup_{0\le u\le jT}\|x(u)\|$.

Let $a_T = \max\{k_T^1,\, c(T)\,\|\Phi\|_\infty + k_{2T}\}$. We have $c_1 \le a_T \sup_{0\le u\le T}\|h(u)\|$ and from (3.5), if $j \ge 2$,

$$c_j \le \max\left\{c_{j-1},\, c(T)\,c_{j-1} + a_T \sup_{0\le u\le j\,T}\|h(u)\|\right\}.$$

Hence, we check that for each $j \ge 2$

$$c_j \le a_T\left(1 + c(T) + \cdots + c(T)^{j-1}\right)\sup_{0\le u\le j\,T}\|h(u)\|,$$

and again from (3.5) we finally deduce that for $t \ge 0$ (and hence $t \in [j\,T, (j+1)\,T]$ for some $j \ge 0$)

$$\|x(t)\| \le a_T \sum_{k=0}^{j} c(T)^k \sup_{0\le u\le t}\|h(u)\| \le \frac{a_T}{1-c(T)}\sup_{0\le u\le t}\|h(u)\|,$$

which finishes the proof. □

THEOREM 3.6. *Let us assume that D is stable. Then there is a continuous function $c \in C([0,\infty), \mathbb{R}^+)$ with $\lim_{t\to\infty} c(t) = 0$ and a positive constant $k > 0$ such that the solution of* (3.2) *satisfies*

$$\|x(t)\| \le c(t)\,\|\varphi\|_\infty + k \sup_{0\le u\le t}\|h(u)\|$$

for each $t \ge 0$.

Proof. It is not hard to check that $x(t) = x^1(t) + x^2(t)$, where x^1 and x^2 are the solutions of

$$\begin{cases} Dx_t^1 = \psi(t)\,h(t), & t \ge 0, \\ x_0^1 = \varphi, \end{cases} \qquad \begin{cases} Dx_t^2 = (1-\psi(t))\,h(t), & t \ge 0, \\ x_0^2 = 0, \end{cases}$$

respectively, and

$$\begin{array}{rccl} \psi\colon & [0,\infty) & \longrightarrow & \mathbb{R} \\ & t & \mapsto & \psi(t) = \begin{cases} 1-t, & 0 \le t \le 1, \\ 0, & 1 \le t. \end{cases} \end{array}$$

Moreover, since $y(t) = x^1(t+1)$ satisfies $Dy_t = 0$, $t \ge 0$, with $y_0 = x_1^1$, the result follows from the application of Definition 3.4, Lemma 3.3, and Proposition 3.5 to y, x^1 on $[0,1]$, and x^2, respectively. □

The conclusions of Theorem 3.6 are essential in what follows. In particular, it allows us to estimate the norm of a function x in terms of the norm of the function $(-\infty, 0] \to \mathbb{R}^m$, $s \mapsto Dx_s$.

PROPOSITION 3.7. *Let us assume that D is stable. Then there is a positive constant $k > 0$ such that $\|x^h\|_\infty \le k\,\|h\|_\infty$ for all $h \in BU$ and $x^h \in BU$ satisfying $Dx_s^h = h(s)$ for $s \le 0$.*

Proof. Let $x(t)$ be the solution of

$$\begin{cases} Dx_t = h(0), & t \ge 0, \\ x_0 = x^h, \end{cases}$$

$$\widetilde{h}(t) = \begin{cases} h(t)\,, & t \le 0\,, \\ h(0)\,, & t \ge 0\,, \end{cases}$$

and for $s \le 0$ we define

$$y^s(t) = \begin{cases} x(t+s)\,, & t+s \ge 0\,, \\ x^h(t+s)\,, & t+s \le 0\,. \end{cases}$$

Then

$$\begin{cases} Dy_t^s = \widetilde{h}(t+s)\,, & t \ge 0\,, \\ y_0^s = x_s^h\,, \end{cases}$$

and Theorem 3.6 yields

$$\|y^s(t)\| \le c(t)\,\|x_s^h\|_\infty + k \sup_{0\le u\le t} \|\widetilde{h}(u+s)\|_\infty \le c(t)\,\|x^h\|_\infty + k\,\|h\|_\infty$$

for all $t \ge 0$ and $s \le 0$. Hence, $\|x^h(s)\| = \|y^{s-t}(t)\| \le c(t)\,\|x^h\|_\infty + k\,\|h\|_\infty$, and as $t \to \infty$ we prove the result. □

Let D be stable and given by (3.1). We define the linear operator

$$\begin{array}{rccccc} \widehat{D}\colon & BU & \longrightarrow & BU & & \\ & x & \mapsto & \widehat{D}x\colon (-\infty,0] & \to & \mathbb{R}^m \\ & & & s & \mapsto & Dx_s\,, \end{array} \tag{3.6}$$

that is, $\widehat{D}x(s) = x(s) - \int_{-\infty}^0 [d\nu(\theta)]\,x(\theta+s)$ for each $s \in (-\infty,0]$, which is well defined, i.e., $h = \widehat{D}x \in BU$, provided that $x \in BU$ because D is bounded and $h(s+\tau) - h(s) = D\,(x_{s+\tau} - x_s)$ for all τ, $s \le 0$. Moreover, it is easy to check that $\widehat{D}$ is bounded for the norm and uniformly continuous when we take the restriction of the compact-open topology to B_r; i.e., given $\varepsilon > 0$ there is a $\delta(r) > 0$ such that $\mathsf{d}(\widehat{D}x_1, \widehat{D}x_2) < \varepsilon$ for all x_1, $x_2 \in B_r$ with $\mathsf{d}(x_1,x_2) < \delta(r)$. The next result shows, after proving that $\widehat{D}$ is invertible, that the same happens for $\widehat{D}^{-1}$.

THEOREM 3.8. *Let us assume that D is stable. Then $\widehat{D}$ is invertible, and $\widehat{D}^{-1}$ is bounded for the norm and uniformly continuous when we take the restriction of the compact-open topology to B_r; i.e., given $\varepsilon > 0$ there is a $\delta(r) > 0$ such that $\mathsf{d}(\widehat{D}^{-1}h_1, \widehat{D}^{-1}h_2) < \varepsilon$ for all h_1, $h_2 \in B_r$ with $\mathsf{d}(h_1,h_2) < \delta(r)$.*

Proof. $\widehat{D}$ is injective because from Proposition 3.7 the only solution of $Dx_s = 0$ for $s \le 0$ is $x = 0$. To show that $\widehat{D}$ is onto, let $h \in BU$ and $\{h_n\}_{n\in\mathbb{N}} \subset B_r$, for some $r > 0$, be a sequence of continuous functions whose components are of compact support such that $h_n \xrightarrow{\mathsf{d}} h$ as $n \uparrow \infty$. Moreover, it is easy to choose them with the same modulus of uniform continuity as h. It is not hard to check that for each $n \in \mathbb{N}$ there is an $x^n \in BU$ such that $\widehat{D}x^n = h_n$, that is, $Dx_s^n = h_n(s)$ for $s \le 0$ and $n \in \mathbb{N}$. From Proposition 3.7, $x^n \in B_{kr}$ because $\|x^n\|_\infty \le k\,\|h_n\|_\infty \le k\,r$ and $\|x^n - x_\tau^n\|_\infty \le k\,\|h_n - (h_n)_\tau\|_\infty$ for each $\tau \le 0$ and $n \in \mathbb{N}$, which implies that $\{x_n\}_{n\in\mathbb{N}}$ is equicontinuous, and hence relatively compact for the compact-open topology. Hence, there is a convergent subsequence, let us assume the whole sequence; i.e., there is a continuous function x such that $x^n \xrightarrow{\mathsf{d}} x$ as $n \uparrow \infty$. From this, $x_s^n \xrightarrow{\mathsf{d}} x_s$ for each $s \le 0$ and (3.1) yields $Dx_s^n = h_n(s) \to Dx_s$, i.e., $Dx_s = h(s)$ for $s \le 0$ and $\widehat{D}x = h$. It is immediate to check that $x \in BU$ and then $\widehat{D}$ is onto, as claimed.

Since $\widehat{D}$ is linear, bounded for the norm, and bijective, the continuity of $\widehat{D}^{-1}$ for the norm is immediate. However, it also follows from Proposition 3.7 which reads as $\|\widehat{D}^{-1}h\|_\infty \leq k\,\|h\|_\infty$. Finally, since $\widehat{D}^{-1}$ is linear, to check the uniform continuity for the metric on each B_r, it is enough to prove the continuity at 0, i.e., $\widehat{D}^{-1}x_n \xrightarrow{\mathsf{d}} 0$ as $n \uparrow \infty$, whenever $x_n \xrightarrow{\mathsf{d}} 0$ as $n \uparrow \infty$ and $\{x_n\}_{n\in\mathbb{N}} \subset B_r$. Let $y^n = \widehat{D}^{-1}x_n \in B_{kr}$. We extend the definition of y^n to $t \geq 0$ as the solution of

$$\begin{cases} Dy^n_t = x_n(0)\,, \quad t \geq 0\,, \\ y^n_0 = y^n\,. \end{cases}$$

The stability of D provides

$$\|y^n(t+s)\| \leq c(t)\,\|y^n\|_\infty + k \sup_{s\leq u\leq 0} \|x_n(u)\| \tag{3.7}$$

for all $t \geq 0$ and $s \leq 0$. Now we check that $\{y^n\}_{n\in\mathbb{N}}$ converges uniformly to 0 on each compact set $K = [-a, 0]$. Given $\varepsilon > 0$, there is a $t_0 > 0$ such that $c(t_0)\,\|y^n\|_\infty < \varepsilon/2$ for each $n \in \mathbb{N}$. Moreover, $x_n \to 0$ in $\widetilde{K} = [-a - t_0, 0]$, and hence there is an n_0 such that for each $n \geq n_0$ we have $k\,\|x_n\|_{\widetilde{K}} < \varepsilon/2$. Therefore, from (3.7) we deduce that for all $u \in K = [-a, 0]$ and $n \geq n_0$,

$$\|y^n(u)\| = \|y^n(t_0 + u - t_0)\| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon\,,$$

that is, $\|y^n\|_K < \varepsilon$, and $\widehat{D}^{-1}x_n = y^n \xrightarrow{\mathsf{d}} 0$ as $n \uparrow \infty$, which finishes the proof. □

A more systematic study of the properties of the linear operator $\widehat{D}$ defined in (3.6) can be found in Staffans [25]. The next result provides a necessary and sufficient condition for a continuous operator D to be stable. In particular, if $\widehat{D}$ is invertible and $\widehat{D}^{-1}$ is continuous for the restriction of the compact-open topology to B_r, then D is stable.

THEOREM 3.9. *Let $D\colon BU \to \mathbb{R}^m$ be given by* (3.1) *and let $\widehat{D}$ be the linear operator in BU defined in* (3.6)*. The following statements are equivalent:*

(i) *D is stable.*

(ii) *For each $r > 0$ and each sequence $\{x_n\}_{n\in\mathbb{N}}$ in BU such that $\|\widehat{D}x_n\|_\infty \leq r$ and $\widehat{D}x_n \xrightarrow{\mathsf{d}} 0$ as $n \uparrow \infty$, $x_n(0) \to 0$ as $n \uparrow \infty$.*

Proof. (i) ⇒ (ii) is a consequence of Theorem 3.8.

(ii) ⇒ (i) For each $T > 0$ we define $\mathcal{L}_T\colon \{\varphi \in BU \mid D\varphi = 0\} \to \mathbb{R}^m$, $\varphi \mapsto x(T)$, where x is the solution of

$$\begin{cases} Dx_t = 0\,, \quad t \geq 0\,, \\ x_0 = \varphi\,. \end{cases}$$

It is easy to check that $\mathcal{L}_T$ is well defined and linear. In addition, from (3.3) we deduce that $\|\mathcal{L}_T(\varphi)\| \leq \|x_T\|_\infty \leq k_T^2\,\|\varphi\|_\infty$, and hence it is bounded.

Next we check that $\|\mathcal{L}_T\|_\infty \to 0$ as $T \to \infty$, which shows the stability of D because $\|x(T)\| \leq c(T)\,\|\varphi\|_\infty$ for $c(T) = \|\mathcal{L}_T\|_\infty$. Let us assume, on the contrary, that there exist $\delta > 0$, a sequence $T_n \uparrow \infty$, and a sequence $\{\varphi_n\}_{n\in\mathbb{N}}$ with $\|\varphi_n\|_\infty \leq 1$ and $D\varphi_n = 0$ such that $\|\mathcal{L}_{T_n}(\varphi_n)\| \geq \delta$ for each $n \in \mathbb{N}$. That is, $\|x^n(T_n)\| \geq \delta$, where x^n is the solution of

$$\begin{cases} Dx^n_t = 0\,, \quad t \geq 0\,, \\ x^n_0 = \varphi_n\,. \end{cases}$$

Therefore,

$$\begin{cases} D((x^n_{T_n})_s) = D(x^n_{T_n+s}) = 0 & \text{if } s \in [-T_n, 0]\,, \\ D((x^n_{T_n})_s) = D((\varphi_n)_{T_n+s}) & \text{if } s \le -T_n\,, \end{cases}$$

and taking $r = \|D\|_\infty$, the sequence $\{x^n_{T_n}\}_{n\in\mathbb{N}} \subset BU$ satisfies $\|\widehat{D}x_n\|_\infty \le r$ and $\widehat{D}x^n_{T_n} \xrightarrow{\mathsf{d}} 0$ as $n \uparrow \infty$. Consequently, $x^n_{T_n}(0) = x^n(T_n) \to 0$ as $n \uparrow \infty$, which contradicts the fact that $\|x^n(T_n)\| \ge \delta$ and finishes the proof. □

PROPOSITION 3.10. *Let $D\colon BU \to \mathbb{R}^m$ be a stable operator given by* (3.1) *and let $\widehat{D}$ be the linear operator in BU defined in* (3.6). *Then*

$$\begin{array}{rccc} D^*\colon & BU & \longrightarrow & \mathbb{R}^m \\ & x & \mapsto & \widehat{D}^{-1}x(0) \end{array}$$

is also stable and satisfies (D1)–(D3).

Proof. From Theorem 3.8, we deduce that D^* satisfies (D1)–(D2). Hence as in Proposition 3.1, there is a real regular Borel measure μ^* with finite total variation such that $D^*x = \int_{-\infty}^0 [d\mu^*(s)]\, x(s)$ for each $x \in BU$. We can write $\mu^* = A\,\delta - \nu^*$ with $A = [\mu^*_{ij}(\{0\})]$. We claim that D^* is atomic at 0, i.e., $\det A \ne 0$. Assume, on the contrary, that $\det A = 0$ and let $v \in \mathbb{R}^m$ be a unitary vector with $A\,v = 0$. For each $\varepsilon > 0$ we take $\varphi_\varepsilon\colon (-\infty, 0] \to \mathbb{R}$ with $\|\varphi_\varepsilon\|_\infty = \varphi_\varepsilon(0) = 1$ and $\varphi_\varepsilon(s) = 0$ for each $s \in (-\infty, -\varepsilon]$. Let $x^{\,\varepsilon} \in BU$ be defined by $x^{\,\varepsilon}(s) = \varphi_\varepsilon(s)\, v$. The continuity of $\widehat{D}$ yields

$$1 = \|x^{\,\varepsilon}\|_\infty \le c\, \|\widehat{D}^{-1}x^{\,\varepsilon}\|_\infty\,. \tag{3.8}$$

However, for each $s \in (-\infty, 0]$

$$\widehat{D}^{-1}x^{\,\varepsilon}(s) = D^*x^{\,\varepsilon}_s = \varphi_\varepsilon(s)A\,v - \int_{-\infty}^0 [d\nu^*(\theta)]\, \varphi_\varepsilon(\theta + s)\, v$$

and, consequently, $\|\widehat{D}^{-1}x^{\,\varepsilon}\|_\infty \le \|\nu^*\|_\infty(-\varepsilon, 0]$, which tends to 0 as $\varepsilon \to 0$, contradicts (3.8) and shows that D^* is atomic at 0. Finally, D^* is stable as a consequence of Theorem 3.9. Notice that μ^* is the inverse of the measure μ for the convolution defining the operator $\widehat{D}$. □

4. Monotone NFDEs. Throughout this section, we will study the monotone skew-product semiflow generated by a family of NFDEs with infinite delay and stable D-operator. In particular, we establish the 1-covering property of omega-limit sets under the componentwise separating property and uniform stability, as in Jiang and Zhao [17] for FDEs with finite delay, and Novo, Obaya, and Sanz [20] for infinite delay. The main tool in the proof of the result is the transformation of the initial family of NFDEs into a family of FDEs with infinite delay in whose study the results of Novo et al. [20] turn out to be useful.

Let $(\Omega, \sigma, \mathbb{R})$ be a minimal flow over a compact metric space (Ω, d) and denote $\sigma(t, \omega) = \omega{\cdot}t$ for all $\omega \in \Omega$ and $t \in \mathbb{R}$. In $\mathbb{R}^m$, we take the maximum norm $\|v\| = \max_{j=1,\dots,m} |v_j|$ and the usual partial order relation

$$\begin{aligned} v \le w \quad &\Longleftrightarrow \quad v_j \le w_j \quad \text{for } j = 1, \dots, m\,, \\ v < w \quad &\Longleftrightarrow \quad v \le w \quad \text{and} \quad v_j < w_j \quad \text{for some } j \in \{1, \dots, m\}\,. \end{aligned}$$

As in section 3, we consider the Fréchet space $X = C((-\infty, 0], \mathbb{R}^m)$ endowed with the compact-open topology, i.e., the topology of uniform convergence over compact subsets, and $BU \subset X$ the Banach space of bounded and uniformly continuous functions with the supremum norm $\|x\|_\infty = \sup_{s\in(-\infty,0]} \|x(s)\|$.

Let $D\colon BU \to \mathbb{R}^m$ be an autonomous and stable linear operator satisfying hypotheses (D1)–(D3) and given by relation (3.1). The subset

$$BU_D^+ = \{x \in BU \mid Dx_s \geq 0 \ \text{ for each } \ s \in (-\infty, 0]\}$$

is a positive cone in BU, because it is a nonempty closed subset $BU_D^+ \subset BU$ satisfying $BU_D^+ + BU_D^+ \subset BU_D^+$, $\mathbb{R}^+ BU_D^+ \subset BU_D^+$, and $BU_D^+ \cap (-BU_D^+) = \{0\}$. As usual, a partial order relation on BU is induced, given by

$$x \leq_D y \quad \Longleftrightarrow \quad Dx_s \leq Dy_s \quad \text{for each } s \in (-\infty, 0]\,,$$

$$x <_D y \quad \Longleftrightarrow \quad x \leq_D y \quad \text{and} \quad x \neq y\,.$$

Remark 4.1. Notice that if we denote the usual partial order of BU

$$x \leq y \quad \Longleftrightarrow \quad x(s) \leq y(s) \quad \text{for each } s \in (-\infty, 0]\,,$$

we have that $x \leq_D y$ if and only if $\widehat{D}\, x \leq \widehat{D}\, y$, where $\widehat{D}$ is defined by relation (3.6). Although in some cases they may coincide, this new order is different from the one given by Wu and Freedman in [28].

We consider the family of nonautonomous NFDEs with infinite delay and stable D-operator

$$\frac{d}{dt} Dz_t = F(\omega{\cdot}t, z_t)\,, \quad t \geq 0\,,\ \omega \in \Omega\,, \tag{4.1}_\omega$$

defined by a function $F\colon \Omega \times BU \to \mathbb{R}^m$, $(\omega, x) \mapsto F(\omega, x)$ satisfying the following conditions:

- (F1) F is continuous on $\Omega \times BU$ and locally Lipschitz in x for the norm $\|\cdot\|_\infty$.
- (F2) For each $r > 0$, $F(\Omega \times B_r)$ is a bounded subset of $\mathbb{R}^m$.
- (F3) For each $r > 0$, $F\colon \Omega \times B_r \to \mathbb{R}^m$ is continuous when we take the restriction of the compact-open topology to B_r; i.e., if $\omega_n \to \omega$ and $x_n \xrightarrow{\mathsf{d}} x$ as $n \to \infty$ with $x \in B_r$, then $\lim_{n\to\infty} F(\omega_n, x_n) = F(\omega, x)$.
- (F4) If $x, y \in BU$ with $x \leq_D y$ and $D_j\, x = D_j\, y$ holds for some $j \in \{1, \ldots, m\}$, then $F_j(\omega, x) \leq F_j(\omega, y)$ for each $\omega \in \Omega$.

From hypothesis (F1), the standard theory of NFDEs with infinite delay (see Wang and Wu [26] and Wu [27]) assures that for each $x \in BU$ and each $\omega \in \Omega$ the system $(4.1)_\omega$ locally admits a unique solution $z(t, \omega, x)$ with initial value x, i.e., $z(s, \omega, x) = x(s)$ for each $s \in (-\infty, 0]$. Therefore, the family $(4.1)_\omega$ induces a local skew-product semiflow

$$\begin{aligned} \tau \quad &: \mathbb{R}^+ \times \Omega \times BU \quad \longrightarrow \quad \Omega \times BU \\ &\qquad (t, \omega, x) \qquad\quad \mapsto \quad (\omega{\cdot}t, u(t, \omega, x))\,, \end{aligned} \tag{4.2}$$

where $u(t, \omega, x) \in BU$ and $u(t, \omega, x)(s) = z(t + s, \omega, x)$ for $s \in (-\infty, 0]$.

As proved in Theorem 3.8, the operator $\widehat{D}$ defined by relation (3.6) is an isomorphism of BU. Hence, the change of variable $y = \widehat{D}z$ takes $(4.1)_\omega$ to

$$y'(t) = G(\omega{\cdot}t, y_t)\,, \quad t \geq 0\,,\ \omega \in \Omega\,, \tag{4.3}_\omega$$

with $G\colon \Omega \times BU \to \mathbb{R}^m$, $(\omega, x) \mapsto G(\omega, x) = F(\omega, \widehat{D}^{-1}x)$ satisfying the following conditions:

(H1) G is continuous on $\Omega \times BU$ and locally Lipschitz in x for the norm $\|\cdot\|_\infty$.

(H2) For each $r > 0$, $G(\Omega \times B_r)$ is a bounded subset of $\mathbb{R}^m$.

(H3) For each $r > 0$, $G\colon \Omega \times B_r \to \mathbb{R}^m$ is continuous when we take the restriction of the compact-open topology to B_r, i.e., if $\omega_n \to \omega$ and $x_n \xrightarrow{\mathsf{d}} x$ as $n \to \infty$ with $x \in B_r$, then $\lim_{n\to\infty} G(\omega_n, x_n) = G(\omega, x)$.

(H4) If $x, y \in BU$ with $x \le y$ and $x_j(0) = y_j(0)$ holds for some $j \in \{1, \dots, m\}$, then $G_j(\omega, x) \le G_j(\omega, y)$ for each $\omega \in \Omega$.

From hypothesis (H1), the standard theory of infinite delay FDEs (see Hino, Murakami, and Naiko [13]) assures that for each $x \in BU$ and each $\omega \in \Omega$ the system $(4.3)_\omega$ locally admits a unique solution $y(t, \omega, x)$ with initial value x, i.e., $y(s, \omega, x) = x(s)$ for each $s \in (-\infty, 0]$. Therefore, the new family $(4.3)_\omega$ induces a local skew-product semiflow

$$\begin{array}{rccl} \widehat{\tau} & : \mathbb{R}^+ \times \Omega \times BU & \longrightarrow & \Omega \times BU \\ & (t, \omega, x) & \mapsto & (\omega{\cdot}t, \widehat{u}(t, \omega, x))\,, \end{array} \tag{4.4}$$

where $\widehat{u}(t, \omega, x) \in BU$ and $\widehat{u}(t, \omega, x)(s) = y(t + s, \omega, x)$ for $s \in (-\infty, 0]$, and it is related to the previous one, (4.2), by

$$\widehat{u}(t, \omega, x) = \widehat{D}\, u(t, \omega, \widehat{D}^{-1}\, x)\,. \tag{4.5}$$

As a consequence, most of the results obtained in Novo et al. [20] for the skew-product semiflow (4.4) can now be translated to (4.2).

From hypotheses (F1) and (F2), each bounded solution $z(t, \omega_0, x_0)$ provides a relatively compact trajectory, as deduced from Proposition 4.1 of Novo et al. [20].

PROPOSITION 4.2. *Let $z(t, \omega_0, x_0)$ be a bounded solution of $(4.1)_{\omega_0}$, that is, $r = \sup_{t\in\mathbb{R}} \|z(t, \omega_0, x_0)\| < \infty$. Then* $\operatorname{closure}_X\{u(t, \omega_0, x_0) \mid t \ge 0\}$ *is a compact subset of BU for the compact-open topology.*

From hypotheses (F1), (F2), and (F3) and Proposition 4.2 and Corollary 4.3 of Novo et al. [20] for the skew-product semiflow (4.4), we can deduce the continuity of the semiflow (4.2) restricted to some compact subsets $K \subset \Omega \times BU$ when the compact-open topology is considered on BU.

PROPOSITION 4.3. *Let $K \subset \Omega \times BU$ be a compact set for the product metric topology and assume that there is an $r > 0$ such that $\tau_t(K) \subset \Omega \times B_r$ for all $t \ge 0$. Then the map*

$$\begin{array}{rccl} \tau & : \mathbb{R}^+ \times K & \longrightarrow & \Omega \times BU \\ & (t, \omega, x) & \mapsto & (\omega{\cdot}t, u(t, \omega, x)) \end{array}$$

is continuous when the product metric topology is considered.

From Proposition 4.2, when $z(t, \omega_0, x_0)$ is bounded we can define the omega-limit set of the trajectory of the point (ω_0, x_0) as

$$\mathcal{O}(\omega_0, x_0) = \{(\omega, x) \in \Omega \times BU \mid \exists\, t_n \uparrow \infty \ \text{ with } \omega_0{\cdot}t_n \to \omega\,,\ u(t_n, \omega_0, x_0) \xrightarrow{\mathsf{d}} x\}\,.$$

Notice that the omega-limit set of a pair $(\omega_0, x_0) \in \Omega \times BU$ makes sense whenever $\operatorname{closure}_X\{u(t, \omega_0, x_0) \mid t \ge 0\}$ is a compact set, because then $\{u(t, \omega_0, x_0)(0) = z(t, \omega_0, x_0) \mid t \ge 0\}$ is a bounded set. Proposition 4.3 implies that the restriction

of the semiflow (4.2) to $\mathcal{O}(\omega_0, x_0)$ is continuous for the compact-open topology. The following result is a consequence of Proposition 4.4 of Novo et al. [20].

PROPOSITION 4.4. *Let* $(\omega_0, x_0) \in \Omega \times BU$ *be such that* $\sup_{t\geq 0} \|z(t, \omega_0, x_0)\| < \infty$. *Then* $K = \mathcal{O}(\omega_0, x_0)$ *is a positively invariant compact subset admitting a flow extension.*

From hypothesis (F4), the monotone character of the semiflow (4.2) is deduced.

PROPOSITION 4.5. *For all* $\omega \in \Omega$ *and* $x, y \in BU$ *such that* $x \leq_D y$ *it holds that*

$$u(t, \omega, x) \leq_D u(t, \omega, y)$$

whenever they are defined.

Proof. From $x \leq_D y$ we know that $\widehat{D}\, x \leq \widehat{D}\, y$, and since (F4) $\Rightarrow$ (H4), from Proposition 4.5 of Novo et al. [20] we deduce that $\widehat{u}(t, \omega, \widehat{D}\, x) \leq \widehat{u}(t, \omega, \widehat{D}\, y)$ whenever they are defined, that is,

$$u(t, \omega, x) = \widehat{D}^{-1}\, \widehat{u}(t, \omega, \widehat{D}\, x) \leq_D \widehat{D}^{-1} \widehat{u}(t, \omega, \widehat{D}\, y) = u(t, \omega, y)\,,$$

as stated. ∎

We establish the 1-covering property of omega-limit sets when in addition to hypotheses (F1)–(F4) the componentwise separating property and uniform stability are assumed:

(F5) If $x, y \in BU$ with $x \leq_D y$ and $D_i\, x < D_i\, y$ holds for some $i \in \{1, \ldots, m\}$, then $D_i\, z_t(\omega, x) < D_i\, z_t(\omega, y)$ for all $t \geq 0$ and $\omega \in \Omega$.

(F6) There is an $r > 0$ such that all the trajectories with initial data in $\widehat{D}^{-1} B_r$ are uniformly stable in $\widehat{D}^{-1} B_{r'}$ for each $r' > r$, and relatively compact for the product metric topology.

From relation (4.5) we deduce that the transformed skew-product semiflow (4.4) satisfies the following:

(H5) If $x, z \in BU$ with $x \leq z$ and $x_i(0) < z_i(0)$ holds for some $i \in \{1, \ldots, m\}$, then $y_i(t, \omega, x) < y_i(t, \omega, z)$ for all $t \geq 0$ and $\omega \in \Omega$.

(H6) There is an $r > 0$ such that all the trajectories with initial data in B_r are uniformly stable in $B_{r'}$ for each $r' > r$, and relatively compact for the product metric topology.

Finally, from Theorem 5.3 of Novo et al. [20] applied to the skew-product semiflow (4.4) satisfying hypotheses (H1)–(H6), we obtain the next result for NFDEs with infinite delay.

THEOREM 4.6. *Assume that hypotheses* (F1)–(F6) *hold and let* $(\omega_0, x_0) \in \Omega \times \widehat{D}^{-1} B_r$ *be such that* $K = \mathcal{O}(\omega_0, x_0) \subset \Omega \times \widehat{D}^{-1} B_r$. *Then* $K = \{(\omega, c(\omega)) \mid \omega \in \Omega\}$ *is a copy of the base and*

$$\lim_{t\to\infty} d(u(t, \omega_0, x_0), c(\omega_0 \cdot t)) = 0\,,$$

where $c : \Omega \to BU$ *is a continuous equilibrium, i.e.,* $c(\omega \cdot t) = u(t, \omega, c(\omega))$ *for any* $\omega \in \Omega$, $t \geq 0$, *and it is continuous for the compact-open topology on* BU.

Remark 4.7. It is easy to check that it is enough to ask for property (F5) (and for (H5) in the case of FDEs with infinite delay) for initial data in BU whose trajectories are globally defined on $\mathbb{R}$.

5. Compartmental systems. We consider compartmental models for the mathematical description of processes in which the transport of material between compartments takes a nonnegligible length of time, and each compartment produces or swallows material. We provide a nonautonomous version, without strong monotonicity assumptions, of previous autonomous results by Wu and Freedman [28] and Wu [26].

First, we introduce the model with which we are going to deal as well as some notation. Let us suppose that we have a system formed by m compartments $C_1, \dots, C_m$. Denote by C_0 the environment surrounding the system, and by $z_i(t)$ the amount of material within compartment C_i at time t for each $i \in \{1, \dots, m\}$. Material flows from compartment C_j into compartment C_i through a pipe P_{ij} having a transit time distribution given by a positive regular Borel measure μ_{ij} with finite total variation $\mu_{ij}(-\infty, 0] = 1$ for each $i,\ j \in \{1, \dots, m\}$. Let $\widetilde{g}_{ij} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be the so-called *transport function* determining the volume of material flowing from C_j to C_i given in terms of the time t and the value of $z_j(t)$ for $i \in \{0, \dots, m\}, j \in \{1, \dots, m\}$. For each $i \in \{1, \dots, m\}$, we will assume that there exists an incoming flow of material $\tilde{I}_i$ from the environment into compartment C_i which depends only on time. For each $i \in \{1, \dots, m\}$, at time time $t \geq 0$, the compartment C_i produces material itself at a rate $\sum_{j=1}^{m} \int_{-\infty}^{0} z_j'(t+s)\, d\nu_{ij}(s)$, where ν_{ij} is a positive regular Borel measure with finite total variation $\nu_{ij}(-\infty, 0] < \infty$ and $\nu_{ij}(\{0\}) = 0$ for all $i,\ j \in \{1, \dots, m\}$.

Once the destruction and creation of material is taken into account, the change of the amount of material of any compartment C_i, $1 \leq i \leq m$, equals the difference between the amount of total influx into and total outflux out of C_i, and we obtain a model governed by the following system of infinite delay NFDEs:

$$\frac{d}{dt}\left[z_i(t) - \sum_{j=1}^{m} \int_{-\infty}^{0} z_j(t+s)\, d\nu_{ij}(s) \right] = -\widetilde{g}_{0i}(t, z_i(t)) - \sum_{j=1}^{m} \widetilde{g}_{ji}(t, z_i(t)) + \sum_{j=1}^{m} \int_{-\infty}^{0} \widetilde{g}_{ij}(t+s, z_j(t+s))\, d\mu_{ij}(s) + \tilde{I}_i(t), \tag{5.1}$$

$i = 1, \dots, m$. For simplicity, we denote $\widetilde{g}_{i0} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+,\ (t, v) \mapsto \widetilde{I}_i(t)$ for $i \in \{1, \dots, m\}$ and let $\widetilde{g} = (\widetilde{g}_{ij})_{i,j} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^{m(m+2)}$. We will assume that

- (C1) $\widetilde{g}$ is C^1-*admissible*, i.e., $\widetilde{g}$ is C^1 in its second variable and $\widetilde{g}$, $\frac{\partial}{\partial v}\widetilde{g}$ are uniformly continuous and bounded on $\mathbb{R} \times \{v_0\}$ for all $v_0 \in \mathbb{R}$; all its components are monotone in the second variable, and $\widetilde{g}_{ij}(t, 0) = 0$ for each $t \in \mathbb{R}$;
- (C2) $\widetilde{g}$ is a recurrent function, i.e., its *hull* is minimal;
- (C3) $\mu_{ij}(-\infty, 0] = 1$ and $\int_{-\infty}^{0} |s|\, d\mu_{ij}(s) < \infty$;
- (C4) $\nu_{ij}(\{0\}) = 0$ and $\sum_{j=1}^{m} \nu_{ij}(-\infty, 0] < 1$, which implies that the operator $D\colon BU \to \mathbb{R}^m$, with $D_i\, x = x_i(0) - \sum_{j=1}^{m} \int_{-\infty}^{0} x_j(s)\, d\nu_{ij}(s)$, $i = 1, \dots, m$, is stable and satisfies (D1)–(D3);
- (C5) the measures $d\eta_{ij} = c_{ij}\, d\mu_{ij} - \sum_{k=0}^{m} d_{ki}\, d\nu_{ij}$ are positive, where

$$c_{ij} = \inf_{(t,\, v) \in \mathbb{R}^2} \frac{\partial \widetilde{g}_{ij}}{\partial v}(t, v) \quad \text{and} \quad d_{ij} = \sup_{(t,\, v) \in \mathbb{R}^2} \frac{\partial \widetilde{g}_{ij}}{\partial v}(t, v)\,.$$

In practical cases, in which the solutions with physical interest belong to the positive cone and the functions g_{ij} are only defined on $\mathbb{R} \times \mathbb{R}^+$, we can extend them to $\mathbb{R} \times \mathbb{R}$ by $g_{ij}(t, -v) = -g_{ij}(t, v)$ for all $v \in \mathbb{R}^+$. Note that (C5) is a condition for controlling the material produced in the compartments in terms of the material transported through the pipes.

The above formulation includes some particularly interesting cases. When the measures ν_{ij} and μ_{ij} are concentrated on a compact set, then (5.1) is an NFDE with finite delay. When the measures $\nu_{ij} \equiv 0$, then (5.1) is a family of FDEs with finite or infinite delay.

As usual, we include the nonautonomous system (5.1) in a family of nonautonomous NFDEs with infinite delay and stable D-operator of the form $(4.1)_\omega$ as follows.

Let Ω be the *hull* of $\widetilde{g}$, namely, the closure of the set of mappings $\{\widetilde{g}_t \mid t \in \mathbb{R}\}$, with $\widetilde{g}_t(s,v) = \widetilde{g}(t+s,v)$, $(s,v) \in \mathbb{R}^2$, with the topology of uniform convergence on compact sets, which from (C1) is a compact metric space (more precisely from the admissibility of $\widetilde{g}$; see Hino et al. [13]). Let $(\Omega, \sigma, \mathbb{R})$ be the continuous flow defined on Ω by translation, $\sigma : \mathbb{R} \times \Omega \to \Omega$, $(t,\omega) \mapsto \omega{\cdot}t$, with $\omega{\cdot}t(s,v) = \omega(t+s,v)$. By hypothesis (C2), the flow $(\Omega, \sigma, \mathbb{R})$ is minimal. In addition, if $\widetilde{g}$ is almost periodic (resp., almost automorphic) the flow will be almost periodic (resp., almost automorphic). Notice that these two cases are included in our formulation.

Let $g : \Omega \times \mathbb{R} \to \mathbb{R}^{m(m+2)}$, $(\omega, v) \mapsto \omega(0,v)$, continuous on $\Omega \times \mathbb{R}$ and denote $g = (g_{ij})_{i,j}$. It is easy to check that, for all $\omega = (\omega_{ij})_{i,j} \in \Omega$ and all $i \in \{1,\dots,m\}$, ω_{i0} is a function dependent only on t; thus, we can define $I_i = \omega_{i0}$, $i \in \{1,\dots,m\}$. Let $F : \Omega \times BU \to \mathbb{R}^m$ be the map defined by

$$F_i(\omega, x) = -g_{0i}(\omega, x_i(0)) - \sum_{j=1}^{m} g_{ji}(\omega, x_i(0)) + \sum_{j=1}^{m} \int_{-\infty}^{0} g_{ij}(\omega{\cdot}s, x_j(s))\, d\mu_{ij}(s) + I_i(\omega)$$

for $(\omega, x) \in \Omega \times BU$ and $i \in \{1,\dots,m\}$. Hence, the family

$$\frac{d}{dt} D z_t = F(\omega{\cdot}t, z_t)\,, \quad t \geq 0\,,\ \omega \in \Omega\,, \tag{5.2}_\omega$$

where the stable operator D is defined in (C4) and satisfies (D1)–(D3), includes system (5.1) when $\omega = \widetilde{g}$.

It is easy to check that this family satisfies hypotheses (F1)–(F3). The following lemma will be useful when proving (F4) and (F5). We omit its proof, which is analogous to the one given in Wu and Freedman [28] for the autonomous case with finite delay.

LEMMA 5.1. *For all $\omega \in \Omega$, x, $y \in BU$ with $x \leq_D y$, and $i = 1,\dots,m$*

$$F_i(\omega, y) - F_i(\omega, x) \geq -\sum_{j=0}^{m} d_{ji}\,[D_i y - D_i x] + \sum_{j=1}^{m} \int_{-\infty}^{0} (y_j(s) - x_j(s))\, d\eta_{ij}(s), \tag{5.3}$$

where the measures η_{ij} are defined in (C5).

Condition (C5) is essential to proving the monotone character of the semiflow. It can be improved in some cases (see Arino and Bourad [1] for the scalar case).

PROPOSITION 5.2. *Under assumptions* (C1)–(C5)*, the family* $(5.2)_\omega$ *satisfies hypotheses* (F4), (F5) *and $\Omega \times BU_D^+$ is positively invariant.*

Proof. Let x, $y \in BU$ with $x \leq_D y$ and $D_i\, x = D_i\, y$ for some $i \in \{1,\dots,m\}$. From (C4), apart from the stability of the operator D, it is easy to prove that the inverse operator of $\widehat{D}$ defined by (3.6) is positive. Hence, from $x \leq_D y$, that is, $\widehat{D}\, x \leq \widehat{D}\, y$, we also deduce that $x \leq y$, which, together with $D_i\, x = D_i\, y$, relation (5.3), and hypothesis (C5), yields $F_i(\omega, y) \geq F_i(\omega, x)$, that is, hypothesis (F4) holds.

Next, we check hypothesis (F5). Let x, $y \in BU$ with $x \leq_D y$ and $D_i\, x < D_i\, y$ for some $i \in \{1,\dots,m\}$. Since (F4) holds, from Proposition 4.5 $u(t,\omega,x) \leq_D u(t,\omega,y)$ and, as before, we deduce in this case that $u(t,\omega,x) \leq u(t,\omega,y)$, i.e., $z_t(\omega,x) \leq z_t(\omega,y)$ for all $t \geq 0$ and $\omega \in \Omega$. Let $h(t) = D_i z_t(\omega,y) - D_i z_t(\omega,x)$. From $(5.2)_\omega$ and

Lemma 5.1

$$h'(t) = F_i(\omega{\cdot}t, z_t(\omega, y)) - F_i(\omega{\cdot}t, z_t(\omega, x))$$
$$\geq -\sum_{j=0}^{m} d_{ji} h(t) + \sum_{j=1}^{m} \int_{-\infty}^{0} (z_j(t+s, \omega, y) - z_j(t+s, \omega, x))\, d\eta_{ij}(s)\,,$$

and again from hypothesis (C5) we deduce that $h'(t) \geq -dh(t)$ for some $d \geq 0$, which together with $h(0) > 0$ yields $h(t) = D_i z_t(\omega, y) - D_i z_t(\omega, x) > 0$ for each $t \geq 0$ and (F5) holds. Finally, since $I_i(\omega) \geq 0$ for each $\omega \in \Omega$ and $i \in \{1, \ldots, m\}$, and the semiflow is monotone, a comparison argument shows that $\Omega \times BU_D^+$ is positively invariant, as stated. □

Next we will study some cases in which hypothesis (F6) is satisfied. In order to do this, we define $M\colon \Omega \times BU \to \mathbb{R}$, the *total mass* of the system $(5.2)_\omega$, as

$$M(\omega, x) = \sum_{i=1}^{m} D_i x + \sum_{i=1}^{m} \sum_{j=1}^{m} \int_{-\infty}^{0} \left(\int_{s}^{0} g_{ji}(\omega{\cdot}\tau, x_i(\tau))\, d\tau \right) d\mu_{ji}(s) \tag{5.4}$$

for all $\omega \in \Omega$ and $x \in BU$, which is well defined from condition (C3). The next result shows the continuity properties of M and its variation along the flow.

Proposition 5.3. *The total mass M is a continuous function on all the sets of the form $\Omega \times B_r$ with $r > 0$ for the product metric topology. Moreover, for each $t \geq 0$*

$$\frac{d}{dt} M(\tau_t(\omega, x)) = \sum_{i=1}^{m} \left[I_i(\omega{\cdot}t) - g_{0i}(\omega{\cdot}t, z_i(t, \omega, x)) \right]. \tag{5.5}$$

Proof. The continuity follows from (D2), (C1), and (C3). A straightforward computation similar to that given in Wu and Freedman [28] shows that

$$M(\omega{\cdot}t, z_t(\omega, x)) = M(\omega, x) + \sum_{i=1}^{m} \int_{0}^{t} \left[I_i(\omega{\cdot}s) - g_{0i}(\omega{\cdot}s, z_i(s, \omega, x)) \right] ds\,, \tag{5.6}$$

from which (5.5) is deduced. □

The following lemma is essential in the proof of the stability of solutions.

Lemma 5.4. *Let x, $y \in BU$ with $x \leq_D y$. Then*

$$0 \leq D_i z_t(\omega, y) - D_i z_t(\omega, x) \leq M(\omega, y) - M(\omega, x)$$

for each $i = 1, \ldots, m$ and whenever $z(t, \omega, x)$ and $z(t, \omega, y)$ are defined.

Proof. From Propositions 5.2 and 4.5 the skew-product semiflow induced by $(5.2)_\omega$ is monotone. Hence, if $x \leq_D y$, then $u(t, \omega, x) \leq_D u(t, \omega, y)$ whenever they are defined. From this, as before, since $\widehat{D}^{-1}$ is positive we also deduce that $x \leq y$ and $u(t, \omega, x) \leq u(t, \omega, y)$. Therefore, $D_i z_t(\omega, x) \leq D_i z_t(\omega, y)$ and $z_i(t, \omega, x) \leq z_j(t, \omega, y)$ for each $i = 1, \ldots, m$. In addition, the monotonicity of transport functions yields $g_{ij}(\omega, z_j(t, \omega, x)) \leq g_{ij}(\omega, z_j(t, \omega, y))$ for each $\omega \in \Omega$. From all these inequalities and (5.4) and (5.6) we deduce that

$$0 \leq D_i z_t(\omega, y) - D_i z_t(\omega, x) \leq \sum_{i=1}^{m} \left[D_i z_t(\omega, y) - D_i z_t(\omega, x) \right]$$
$$\leq M(\omega{\cdot}t, z_t(\omega, y)) - M(\omega{\cdot}t, z_t(\omega, x)) \leq M(\omega, y) - M(\omega, x)\,,$$

as stated. □

PROPOSITION 5.5. *Fix $r > 0$. Then given $\varepsilon > 0$ there exists $\delta > 0$ such that if x, $y \in B_r$ with $\mathsf{d}(x, y) < \delta$, then $\|z(t, \omega, x) - z(t, \omega, y)\| \leq \varepsilon$ whenever they are defined.*

Proof. Let $c = \max_i \sum_{j=1}^{m} \nu_{ij}(-\infty, 0] < 1$. From the continuity of M, given $\varepsilon_0 = \varepsilon\,(1 - c) > 0$ there exists $0 < \delta < \varepsilon_0$, such that if x, $y \in B_r$ with $\mathsf{d}(x, y) < \delta$, then $|M(\omega, y) - M(\omega, x)| < \varepsilon_0$. Therefore, if x, $y \in B_r$ and $x \leq_D y$, from Lemma 5.4 we deduce that $0 \leq D_i z_t(\omega, y) - D_i z_t(\omega, x) < \varepsilon_0$ whenever $\mathsf{d}(x, y) < \delta$. The definition of D_i yields

$$0 \leq z_i(t, \omega, y) - z_i(t, \omega, x) < \varepsilon_0 + \sum_{j=1}^{m} \int_{-\infty}^{0} [z_j(t + s, \omega, y) - z_j(t + s, \omega, x)]\, d\nu_{ij}(s)$$

$$\leq \varepsilon_0 + \|z_t(\omega, y) - z_t(\omega, x)\|_\infty \sum_{j=1}^{m} \nu_{ij}(-\infty, 0]\,,$$

from which we deduce that $\|z_t(\omega, y) - z_t(\omega, x)\|_\infty (1 - c) < \varepsilon_0 = \varepsilon(1 - c)$, that is, $\|z(t, \omega, x) - z(t, \omega, y)\| \leq \varepsilon$ whenever they are defined. The case in which x and y are not ordered follows easily from this one. ∎

As a consequence, from the existence of a bounded solution for one of the systems of the family, the boundedness of all solutions is inferred, and this is the case in which hypothesis (F6) holds.

THEOREM 5.6. *Under assumptions* (C1)–(C5)*, if there exists $\omega_0 \in \Omega$ such that $(5.2)_{\omega_0}$ has a bounded solution, then all solutions of $(5.2)_\omega$ are bounded as well, hypothesis* (F6) *holds, and all omega-limit sets are copies of the base.*

Proof. The boundedness of all solutions is an easy consequence of the previous proposition and the continuity of the semiflow. Let $(\omega, x) \in \Omega \times BU$ and $r' > 0$ such that $z_t(\omega, x) \in B_{r'}$ for all $t \geq 0$. Then also from Proposition 5.5, we deduce that given $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\|z(t + s, \omega, x) - z(t, \omega{\cdot}s, y)\| = \|z(t, \omega{\cdot}s, z_s(\omega, x)) - z(t, \omega{\cdot}s, y)\| < \varepsilon$$

for all $t \geq 0$ whenever $y \in B_{r'}$ and $\mathsf{d}(z_s(\omega, x), y) < \delta$, which shows the uniform stability of the trajectories in $B_{r'}$ for each $r' > 0$. Moreover, for each $r > 0$ there is an $r' > 0$ such that $\widehat{D}^{-1} B_r \subset B_{r'}$. Hence, hypothesis (F6) holds for all $r > 0$ and Theorem 4.6 applies for all initial data, which finishes the proof. ∎

Concerning the solutions of the original compartmental system, we obtain the following result providing a nontrivial generalization of the autonomous case, in which the asymptotical constancy of the solutions was shown (see Wu and Freedman [28]). Although the theorem is stated in the almost periodic case, similar conclusions are obtained changing almost periodic to periodic, almost automorphic, or recurrent, that is, all solutions are asymptotically of the same type as the transport functions.

THEOREM 5.7. *Under assumptions* (C1)–(C5) *and in the almost periodic case, if there is a bounded solution of* (5.1)*, then there is at least an almost periodic solution and all the solutions are asymptotically almost periodic. For closed systems, i.e., $\widetilde{I}_i \equiv 0$ and $\widetilde{g}_{0i} \equiv 0$ for each $i = 1, \ldots, m$, there are infinitely many almost periodic solutions and the rest of them are asymptotically almost periodic.*

Proof. The first statement is an easy consequence of the previous theorem. Let $\omega_0 = \widetilde{g}$. The omega-limit of each solution $z(t, \omega_0, x_0)$ is a copy of the base $\mathcal{O}(\omega_0, x_0) = \{(\omega, x(\omega)) \mid \omega \in \Omega\}$, and hence $z(t, \omega_0, x(\omega_0)) = x(\omega_0{\cdot}t)(0)$ is an almost periodic solution of (5.1) and

$$\lim_{t \to \infty} \|z(t, \omega_0, x_0) - z(t, \omega_0, x(\omega_0))\| = 0\,.$$

The statement for closed systems follows in addition from (5.6), which implies that the mass is constant along the trajectories. Hence, there are infinitely many minimal subsets because from the definition of the mass and (C4), given $c > 0$ there is an $(\omega_0, x_0) \in \Omega \times BU_D^+$ such that $M(\omega_0, x_0) = c$ and hence $M(\omega, x) = c$ for each $(\omega, x) \in \mathcal{O}(\omega_0, x_0)$. □

6. Long-term behavior of compartmental systems. This section deals with the long-term behavior of the amount of material within the compartments of the compartmental system (5.1) satisfying hypotheses (C1)–(C5). As in the previous section, the study of the minimal sets for the corresponding skew-product semiflow (4.2) induced by the family $(5.2)_\omega$ will be essential. In addition to hypotheses (C1)–(C5) we will assume the following hypothesis:

(C6) Given $i \in \{0, \dots, m\}$ and $j \in \{1, \dots, m\}$ either $\widetilde{g}_{ij} \equiv 0$ on $\mathbb{R} \times \mathbb{R}^+$ (and hence $g_{ij} \equiv 0$ on $\Omega \times \mathbb{R}^+$), i.e., *there is not a pipe from compartment C_j to compartment C_i*, or for each $v > 0$ there is a $\delta_v > 0$ such that $\widetilde{g}_{ij}(t, v) \geq \delta_v$ for all $t \in \mathbb{R}$ (and hence $g_{ij}(\omega, v) > 0$ for all $\omega \in \Omega$ and $v > 0$). In this case we will say that the pipe P_{ij} *carries material (or that there is a pipe from compartment C_j to compartment C_i).*

Let $I = \{1, \dots, m\}$. $\mathcal{P}(I)$ denotes, as usual, the set of all subsets of I.

DEFINITION 6.1. *Let $\zeta : \mathcal{P}(I) \to \mathcal{P}(I)$, $J \mapsto \cup_{j \in J}\{i \in I \mid P_{ij}$ carries material$\}$. A subset J of I is said to be* irreducible *if $\zeta(J) \subset J$ and no proper subset of J has that property. System* (5.1) *is* irreducible *if the whole set I is irreducible.*

Note that $\zeta(I) \subset I$, so there is always some irreducible subset of I. Irreducible sets detect the occurrence of dynamically independent subsystems. Our next result gives a useful property of irreducible sets with more than one element.

PROPOSITION 6.2. *If a subset J of I is irreducible, then, for all $i, j \in J$ with $i \neq j$, there exist $p \in \mathbb{N}$ and $i_1, \dots, i_p \in J$ such that $P_{i_1 i}$, $P_{i_2 i_1}, \dots, P_{i_p i_{p-1}}$, and $P_{j i_p}$ carry material.*

Proof. Let us assume, on the contrary, that $j \notin \cup_{n=1}^{\infty} \zeta^n(\{i\}) = \widetilde{J}_i$. Then $\widetilde{J}_i \subsetneq J$ and, obviously, $\zeta(\widetilde{J}_i) \subset \widetilde{J}_i$, which contradicts the fact that J is irreducible. □

Let $J_1, \dots, J_k$ be all the irreducible subsets of I and let $J_0 = I \setminus \cup_{l=1}^{k} J_l$. These sets reflect the geometry of the compartmental system in a good enough way as to describe the long-term behavior of the solutions, as we will see below.

Let K be any minimal subset of $\Omega \times BU$ for the skew-product semiflow induced by $(5.2)_\omega$. From Theorem 5.6, K is of the form $K = \{(\omega, x(\omega)) \mid \omega \in \Omega\}$, where x is a continuous map from Ω into BU. All of the subsequent results give qualitative information about the long-term behavior of the solutions. Let us see that, provided that we are working on a minimal set K, if there is no inflow from the environment, then the total mass is constant on K, all compartments out of an irreducible subset are empty, and, in an irreducible subset, either all compartments are empty or all are never empty. In particular, in any irreducible subset *with some outflow of material*, all compartments are empty.

THEOREM 6.3. *Assume that $\tilde{I}_i \equiv 0$ for each $i \in I$ and let $K = \{(\omega, x(\omega)) \mid \omega \in \Omega\}$ be a minimal subset of $\Omega \times BU$ with $K \subset \Omega \times BU_D^+$. Then the following hold:*

(i) *There exists $c \geq 0$ such that $M|_K \equiv c$.*

(ii) *$x_i \equiv 0$ for each $i \in J_0$.*

(iii) *If, for some $l \in \{1, \dots, k\}$, there exists $j_l \in J_l$ such that $x_{j_l} \equiv 0$, then $x_i \equiv 0$ for each $i \in J_l$. In particular, this happens if there is a $j_l \in J_l$ such that there is an outflow of material from C_{j_l}.*

Proof. We first suppose that the system is closed, i.e., $\widetilde{g}_{0i} \equiv 0$, $\widetilde{I}_i \equiv 0$ for all $i \in I$, from which we deduce $g_{0i} \equiv 0$ and $I_i \equiv 0$ for all $i \in I$.

(i) From (5.6) the total mass M is constant along the trajectories, and hence $M(\omega{\cdot}t, x(\omega{\cdot}t)) = M(\omega, x(\omega))$ for all $t \geq 0$ and $\omega \in \Omega$, which together with the fact that Ω is minimal and M continuous shows the statement.

(ii) Let $i \in J_0$. The set $\widetilde{J}_i = \cup_{n=1}^{\infty} \zeta^n(\{i\})$ satisfies $\zeta(\widetilde{J}_i) \subset \widetilde{J}_i$ and hence contains an irreducible set J_l for some $l \in \{1, \dots, k\}$. Consequently, there are $i_1, \dots, i_p \in J_0$ and $j_l \in J_l$ such that $P_{j_l i_p}$ carry material.

It is easy to prove that there is an $r > 0$ such that $\|x(\omega)\|_\infty \leq r$ for each $\omega \in \Omega$. We define $M_l \colon \Omega \times BU \to \mathbb{R}$, the *mass restricted to* J_l, as

$$M_l(\omega, y) = \sum_{i \in J_l} D_i y + \sum_{i,j \in J_l} \int_{-\infty}^{0} \left(\int_s^0 g_{ji}(\omega{\cdot}\tau, y_i(\tau))\, d\tau \right) d\mu_{ji}(s)\,, \tag{6.1}$$

which is continuous on $\Omega \times B_r$. From $x(\omega) \geq_D 0$, which also implies $x(\omega) \geq 0$, and (C1), we have $0 \leq M_l(\omega, x(\omega)) \leq M(\omega, x(\omega)) = c$ for each $\omega \in \Omega$.

Since J_l is irreducible, for all $i \in J_l$ and $\omega \in \Omega$

$$\frac{d}{dt} D_i x(\omega{\cdot}t) = -\sum_{j \in J_l} g_{ji}(\omega{\cdot}t, x_i(\omega{\cdot}t)(0)) + \sum_{j \in J_l \cup J_0} \int_{-\infty}^{0} g_{ij}(\omega{\cdot}(s+t), x_j(\omega{\cdot}t)(s))\, d\mu_{ij}(s)$$

because the rest of the terms vanish. Consequently,

$$\frac{d}{dt} M_l(\omega{\cdot}t, x(\omega{\cdot}t)) = \sum_{i \in J_l} \sum_{j \in J_0} \int_{-\infty}^{0} g_{ij}(\omega{\cdot}(s+t), x_j(\omega{\cdot}t)(s))\, d\mu_{ij}(s) \geq 0 \tag{6.2}$$

for each $\omega \in \Omega$. We claim that $M_l(\omega, x(\omega))$ is constant for each $\omega \in \Omega$. Assume, on the contrary, that there are $\omega_1, \omega_2 \in \Omega$ such that $M_l(\omega_1, x(\omega_1)) < M_l(\omega_2, x(\omega_2))$, and let $t_n \uparrow \infty$ such that $\lim_{n\to\infty} \omega_2{\cdot}t_n = \omega_1$. From (6.2) we deduce that $M_l(\omega_2, x(\omega_2)) \leq M_l(\omega_2{\cdot}t_n, x(\omega_2{\cdot}t_n))$ for each $n \in \mathbb{N}$, and taking limits as $t \to \infty$ we conclude that $M_l(\omega_2, x(\omega_2)) \leq M_l(\omega_1, x(\omega_1))$, a contradiction. Hence $M_l(\omega, x(\omega))$ is constant and from (6.2)

$$\sum_{i \in J_l} \sum_{j \in J_0} \int_{-\infty}^{0} g_{ij}(\omega{\cdot}(s+t), x_j(\omega{\cdot}t)(s))\, d\mu_{ij}(s) = 0\,. \tag{6.3}$$

Next we check that $x_{i_p} \equiv 0$. From (6.3) we deduce that for each $\omega \in \Omega$

$$\int_{-\infty}^{0} g_{j_l i_p}(\omega{\cdot}s, x_{i_p}(\omega)(s))\, d\mu_{j_l i_p}(s) = 0\,. \tag{6.4}$$

Assume that there is an $\omega_0 \in \Omega$ such that $x_{i_p}(\omega_0)(0) > 0$. Hence there is an $\varepsilon > 0$ with $x_{i_p}(\omega_0)(s) > 0$ for each $s \in (-\varepsilon, 0]$, and since $P_{j_l i_p}$ carries material $g_{j_l i_p}(\omega_0{\cdot}s, x_{i_p}(\omega_0)(s)) > 0$ for $s \in (-\varepsilon, 0]$. In addition, from $\mu_{j_l i_p}(-\infty, 0] = 1$ there is a $b \leq 0$ such that $\mu_{j_l i_p}(b - \varepsilon, b] > 0$. Hence, denoting $\omega_0{\cdot}(-b) = \omega_1$ we deduce that

$$\int_{b-\varepsilon}^{b} g_{j_l i_p}(\omega_1{\cdot}s, x_{i_p}(\omega_1)(s))\, d\mu_{j_l i_p}(s) > 0\,,$$

which contradicts (6.4) and shows that $x_{i_p} \equiv 0$, as claimed. Since $x(\omega) \geq_D 0$, we have $D_{i_p} x(\omega) \geq 0$ and from the definition of D_{i_p} we deduce that $D_{i_p} x(\omega) = 0$ for each

$\omega \in \Omega$. Therefore,

$$0 = \frac{d}{dt} D_{i_p} x(\omega \cdot t) = \sum_{j=1}^{m} \int_{-\infty}^{0} g_{i_p j}(\omega \cdot (t+s), x_j(\omega \cdot t)(s))\, d\mu_{i_p j}(s)\,,$$

from which $\int_{-\infty}^{0} g_{i_p i_{p-1}}(\omega \cdot s, x_{i_{p-1}}(\omega)(s))\, d\mu_{i_p i_{p-1}}(s) = 0$, and as before $x_{i_{p-1}} \equiv 0$. In a finite number of steps we check that $x_i \equiv 0$, as stated.

(iii) From Proposition 6.2, given i, $j_l \in J_l$ there exist $p \in \mathbb{N}$ and $i_1, \ldots, i_p \in J_l$ such that $P_{i_1 i}$, $P_{i_2 i_1}, \ldots, P_{i_p i_{p-1}}$, and $P_{j_l i_p}$ carry material. If $x_{j_l} \equiv 0$, the same argument given in the last part of (ii) shows that $x_i \equiv 0$, which finishes the proof for closed systems.

Next we deal with the case when $\tilde{I}_i \equiv 0$ for each $i \in I$ but the system is not necessarily closed. We also have $I_i \equiv 0$ and from (5.5) we deduce that the total mass M is decreasing along the trajectories. In particular,

$$\frac{d}{dt} M(\omega \cdot t, x(\omega \cdot t)) = -\sum_{i=1}^{n} g_{0i}(\omega \cdot t, x_i(\omega \cdot t)(0)) \leq 0\,. \tag{6.5}$$

Assume that there are ω_1, $\omega_2 \in \Omega$ such that $M(\omega_1, x(\omega_1)) < M(\omega_2, x(\omega_2))$, and let $t_n \uparrow \infty$ such that $\lim_{n\to\infty} \omega_1 \cdot t_n = \omega_2$. From relation (6.5) we deduce that $M(\omega_1 \cdot t_n, x(\omega_1 \cdot t_n)) \leq M(\omega_1, x(\omega_1))$ for each $n \in \mathbb{N}$, and taking limits as $n \uparrow \infty$ we conclude that $M(\omega_2, x(\omega_2)) \leq M(\omega_1, x(\omega_1))$, a contradiction, which shows that M is constant on K, as stated in (i). Consequently, the derivative in (6.5) vanishes and $g_{0i}(\omega \cdot t, x_i(\omega \cdot t)(0)) = 0$ for all $i \in I$, $\omega \in \Omega$, and $t \geq 0$. This means that $z(t, \omega, x(\omega)) = x(\omega \cdot t)(0)$ is a solution of a closed system, and (ii) and the first part of (iii) follow from the previous case.

Finally, let $j_l \in J_l$ be such that there is an outflow of material from C_{j_l}, that is, $g_{0 j_l}(\omega, v) > 0$ for all $\omega \in \Omega$ and $v > 0$. Moreover, as before, $g_{0 j_l}(\omega, x_{j_l}(\omega)(0)) = 0$ for each $\omega \in \Omega$, which implies that $x_{j_l} \equiv 0$ and completes the proof. □

Remark 6.4. Notice that, concerning the solutions of the family of systems $(5.2)_\omega$ and hence the solutions of the original system (5.1) when $\omega = \widetilde{g}$, we deduce that in the case of no inflow from the environment, $\lim_{t\to\infty} z_i(t, \omega, x_0) = 0$ for all $i \in J_0$, $i \in J_l$ for compartments J_l with some outflow, and each $x_0 \geq_D 0$.

Remark 6.5. If there is no inflow from the environment and for all $l \in \{1, \ldots, k\}$ there is a $j_l \in J_l$ such that there is outflow of material from C_{j_l}, then the only minimal set in $\Omega \times BU_D^+$ is $K = \{(\omega, 0) \mid \omega \in \Omega\}$ and all the solutions $z(t, \omega, x_0)$ with initial data $x_0 \geq_D 0$ tend to 0 as $t \to \infty$.

In a nonclosed system, that is, a system which may have any inflow and any outflow of material, if there exists a bounded solution, i.e., all solutions are bounded as shown above, and an irreducible set which has *some inflow*, then, working on a minimal set, all compartments of that irreducible set are nonempty and there must be some outflow from the irreducible set.

THEOREM 6.6. *Assume that there exists a bounded solution of* (5.1) *and let* $K = \{(\omega, x(\omega)) \mid \omega \in \Omega\}$ *be a minimal subset of* $\Omega \times BU_D^+$. *If, for some* $l \in \{1, \ldots, k\}$, *there is a* $j_l \in J_l$ *such that* $\widetilde{I}_{j_l} \neq 0$, *i.e., there is some inflow into* C_{j_l}, *then*

(i) $x_i \not\equiv 0$ *for each* $i \in J_l$, *and*

(ii) *there is a* $j \in J_l$ *such that there is outflow of material from* C_j.

Proof. (i) Let us assume, on the contrary, that there is an $i \in J_l$ such that $x_i \equiv 0$. Then since $x(\omega) \geq_D 0$ we have that $0 \leq D_i x(\omega)$, and from the definition of D_i given

in (C4) we deduce that $D_i x(\omega) = 0$ for each $\omega \in \Omega$. Therefore,

$$0 = \frac{d}{dt} D_i x(\omega \cdot t) = \sum_{j=1}^{m} \int_{-\infty}^{0} g_{ij}(\omega \cdot (t+s), x_j(\omega \cdot t)(s))\, d\mu_{ij}(s) + I_i(\omega \cdot t) \tag{6.6}$$

for all $\omega \in \Omega$, $t \geq 0$, and, as in (ii) of Theorem 6.3, we check that $x_{j_l} \equiv 0$. However, since $\widetilde{I}_{j_l} \not\equiv 0$, there is an $\omega_0 \in \Omega$ such that $I_{j_l}(\omega_0) > 0$, which contradicts (6.6) for $\omega = \omega_0$, $i = j_l$ at $t = 0$.

(ii) Assume, on the contrary, that $g_{0j} \equiv 0$ for each $j \in J_l$. Then if we consider (6.1) the restriction of the mass to J_l, we check that

$$\frac{d}{dt} M_l(\omega \cdot t, x(\omega \cdot t)) = \sum_{i \in J_l} \left[I_i(\omega \cdot t) + \sum_{j \in J_0} \int_{-\infty}^{0} g_{ij}(\omega \cdot (s+t), x_j(\omega \cdot t)(s))\, d\mu_{ij}(s) \right] \geq 0$$

for all $\omega \in \Omega$ and $t \geq 0$. A similar argument to the one given in (ii) of Theorem 6.3 shows that $M_l(\omega, x(\omega))$ is constant for each $\omega \in \Omega$, which contradicts the fact that the above derivative is strictly positive for $\omega = \omega_0$ at $t = 0$ and proves the statement. ∎

Finally, we will change hypothesis (C6) to the following, slightly stronger one.

(C6)* Given $i \in \{0, \dots, m\}$ and $j \in \{1, \dots, m\}$ either $\widetilde{g}_{ij} \equiv 0$ on $\mathbb{R} \times \mathbb{R}^+$ (and hence $g_{ij} \equiv 0$ on $\Omega \times \mathbb{R}^+$), i.e., *there is not a pipe from compartment C_j to compartment C_i*, or for each $v \geq 0$ there is a $\delta_v > 0$ such that $\frac{\partial}{\partial v} \widetilde{g}_{ij}(t, v) \geq \delta_v$ for each $t \in \mathbb{R}$ (and hence $\frac{\partial}{\partial v} g_{ij}(\omega, v) > 0$ for all $\omega \in \Omega$ and $v \geq 0$). In this case we will say that the pipe P_{ij} *carries material (or that there is a pipe from compartment C_j to compartment C_i).*

In this case, we are able to prove that if there exists a bounded solution, then all the minimal sets coincide both on irreducible sets having *some outflow* and out of irreducible sets. Concerning the solutions of the initial compartmental system (5.1),

$$\lim_{t \to \infty} |z_i(t, x_0) - z_i(t, y_0)| = 0$$

for all $i \in J_0$, $i \in J_l$ for compartments J_l with some outflow, and all x_0, $y_0 \geq_D 0$.

THEOREM 6.7. *Let us assume that hypotheses* (C1)–(C5) *and* (C6)* *hold and that there exists a bounded solution of system* (5.1). *Let* $K_1 = \{(\omega, x(\omega)) \mid \omega \in \Omega\}$ *and* $K_2 = \{(\omega, y(\omega)) \mid \omega \in \Omega\}$ *be two minimal subsets of* $\Omega \times BU_D^+$. *Then*

(i) $x_i \equiv y_i$ *for each* $i \in J_0$;

(ii) *if, for some* $l \in \{1, \dots, k\}$, *there is a* $j_l \in J_l$ *such that there is outflow of material from* C_{j_l}, *then* $x_i \equiv y_i$ *for each* $i \in J_l$.

Proof. For each $i \in \{0, \dots, m\}$ and each $j \in \{1, \dots, m\}$ we define $h_{ij} \colon \Omega \to \mathbb{R}^+$ as

$$h_{ij}(\omega) = \int_0^1 \frac{\partial g_{ij}}{\partial v}(\omega, s\, x_j(\omega)(0) + (1-s)\, y_j(\omega)(0))\, ds \geq 0\,,$$

and we consider the family of monotone linear compartmental systems

$$\begin{aligned} (6.7)_\omega \quad \frac{d}{dt} D_i \hat{z}_t = {} & -h_{0i}(\omega \cdot t)\, \hat{z}_i(t) - \sum_{j=1}^{m} h_{ji}(\omega \cdot t)\, \hat{z}_i(t) \\ & + \sum_{j=1}^{m} \int_{-\infty}^{0} h_{ij}(\omega \cdot (s+t))\, \hat{z}_j(t+s)\, d\mu_{ij}(s)\,, \quad \omega \in \Omega, \end{aligned}$$

satisfying the corresponding hypotheses (C1)–(C4) and (C6). Moreover, (C5) for each of the systems $(6.7)_\omega$, follows from

$$\inf_{\omega\in\Omega} h_{ij}(\omega) \geq \inf_{v\geq 0\,,\,\omega\in\Omega} \frac{\partial g_{ij}}{\partial v}(\omega, v)\,, \quad \sup_{\omega\in\Omega} h_{ij}(\omega) \leq \sup_{v\geq 0\,,\,\omega\in\Omega} \frac{\partial g_{ij}}{\partial v}(\omega, v)\,,$$

and (C5) for (5.1). From the definition of h_{ij} and (C6)* we deduce that the irreducible sets for the families $(6.7)_\omega$ and $(5.2)_\omega$ coincide. Consequently, Theorem 6.3 (see Remark 6.4) applies to this case, and we deduce that if $z_0 \geq_D 0$ and J_l is a compartment with some outflow of material, then

$$\lim_{t\to\infty} \hat{z}_i(t, \omega, z_0) = 0 \qquad \text{for each } i \in J_0 \cup J_l\,.$$

The same happens for $z_0 \leq_D 0$ because the systems are linear.

Let $z(\omega) = x(\omega) - y(\omega)$ for each $\omega \in \Omega$. It is easy to check $\hat{z}(t, \omega, z(\omega)) = z(\omega{\cdot}t)(0)$ for all $\omega \in \Omega$ and $t \geq 0$. Moreover, we can find $z_1 \leq_D 0$ and $z_0 \geq_D 0$ such that $z_1 \leq_D z(\omega) \leq_D z_0$ for each $\omega \in \Omega$. Hence, the monotonicity of the induced skew-product semiflow and the positivity of $\widehat{D}^{-1}$ yields

$$\hat{z}(t, \omega, z_1) \leq z(\omega{\cdot}t)(0) \leq \hat{z}(t, \omega, z_0) \quad \text{for all } \omega \in \Omega\ t \geq 0\,,$$

from which we deduce that $z_i \equiv 0$ for all $i \in J_0$, $i \in J_l$ and (i) and (ii) follow. □

As a consequence, under the same assumptions of the previous theorem, when for all $l \in \{1, \ldots, k\}$ there is an outflow of material from one of the compartments in J_l, there is a unique minimal set $K = \{(\omega, x(\omega)) \mid \omega \in \Omega\}$ in $\Omega \times BU_D^+$ attracting all the solutions with initial data in BU_D^+; i.e.,

$$\lim_{t\to\infty} \|z(t, \omega, x_0) - x(\omega{\cdot}t)(0)\| = 0\,, \qquad \text{whenever } \ x_0 \geq_D 0\,.$$

Moreover, $x \not\equiv 0$ if and only if there is some $j \in \{1, \ldots, m\}$ such that $\widetilde{I}_j \neq 0$; i.e., there is some inflow into one of the compartments C_j.

For the next result, in addition to hypotheses (C1)–(C5) and (C6)* we will assume the following hypothesis:

(C7) If $K_1 = \{(\omega, x(\omega)) \mid \omega \in \Omega\}$ and $K_2 = \{(\omega, y(\omega)) \mid \omega \in \Omega\}$ are two minimal subsets of $\Omega \times BU_D^+$ such that $x(\omega) \leq_D y(\omega)$ and $D_i x(\omega_0) = D_i y(\omega_0)$ for some $\omega_0 \in \Omega$ and $i \in \{1, \ldots, m\}$, then $x(\omega) = y(\omega)$ for each $\omega \in \Omega$, i.e., $K_1 = K_2$.

Note that if $D_i x(\omega_0) = D_i y(\omega_0)$ holds for some $\omega_0 \in \Omega$ and $i \in \{1, \ldots, m\}$, then from hypothesis (F5) we deduce that it holds for each $\omega \in \Omega$.

Hypothesis (C7) is relevant when it applies to closed systems, and it holds in many cases studied in the literature. A closed system satisfying (C7) is irreducible. Systems with a unique compartment, studied by Arino and Bourad [1] and Krisztin and Wu [19], satisfy (C7). It follows from Theorem 6.3 that irreducible closed systems described by FDEs (see Arino and Haourigui [2]) satisfy (C7). Closed systems given by Wu [27], and Wu and Freedman [28] in the strongly ordered case, also satisfy (C7).

DEFINITION 6.8. *Let $K_1 = \{(\omega, x(\omega)) \mid \omega \in \Omega\}$ and $K_2 = \{(\omega, y(\omega)) \mid \omega \in \Omega\}$ be two minimal subsets. It is said that $K_1 <_D K_2$ if $x(\omega) <_D y(\omega)$ for each $\omega \in \Omega$.*

Hypothesis (C7) allows us to classify the minimal subsets in terms of the value of their total mass, as shown in the next result.

THEOREM 6.9. *Assume that system* (5.1) *is closed (i.e., $\widetilde{I}_i \equiv 0$ and $\widetilde{g}_{0i} \equiv 0$ for each $i \in \{1, \ldots, m\}$), and hypotheses* (C1)–(C5), (C6)*, *and* (C7) *hold. Then for*

each $c > 0$ there is a unique minimal subset K_c such that $M|_{K_c} = c$. Moreover, $K_c \subset \Omega \times BU_D^+$ and $K_{c_1} <_D K_{c_2}$ whenever $c_1 < c_2$.

Proof. Since the minimal subsets are copies of the base, and the total mass (5.4) is constant along the trajectories and increasing for the D-order because $\widehat{D}^{-1}$ is positive, it is easy to check that given $c > 0$ there is a minimal subset $K_c \subset \Omega \times BU_D^+$ such that $M|_{K_c} = c$.

Let $\widehat{D}$ be the isomorphism of BU defined by the relation (3.6). For each $x \in BU$ we define $x^+ = \widehat{D}^{-1} \sup(0, \widehat{D}x)$. Hence $0 \leq_D x^+$, $x \leq_D x^+$, and if $y \in BU$ with $x \leq_D y$ and $0 \leq_D y$, then $x^+ \leq_D y$.

Since the semiflow is monotone, from $x \leq_D x^+$ we deduce that $u(t,\omega,x) \leq_D u(t,\omega,x^+)$. Since the system is closed, $u(t,\omega,0) = 0$, and from $0 \leq_D x^+$ we check that $0 \leq_D u(t,\omega,x^+)$. Consequently $u(t,\omega,x)^+ \leq_D u(t,\omega,x^+)$ for each $t \geq 0$.

Next we check that if $K = \{(\omega, x(\omega)) \mid \omega \in \Omega\}$ is minimal, the same happens for $K^+ = \{(\omega, x(\omega)^+) \mid \omega \in \Omega\}$. Since $x(\omega{\cdot}t) = u(t,\omega,x)$ for each $t \geq 0$, we deduce that $x(\omega{\cdot}t)^+ = u(t,\omega,x(\omega))^+ \leq_D u(t,\omega,x(\omega)^+)$, and the fact that $\widehat{D}^{-1}$ is positive yields $x(\omega{\cdot}t)^+ \leq u(t,\omega,x(\omega)^+)$ for each $t \geq 0$. In addition, since the total mass (5.4) is constant along the trajectories and increasing for the D-order, we deduce that $M(\omega, x(\omega)^+) = M(\omega{\cdot}t, u(t,\omega,x(\omega)^+)) \geq M(\omega{\cdot}t, u(t,\omega,x(\omega))^+) = M(\omega{\cdot}t, x(\omega{\cdot}t)^+)$ for each $t \geq 0$. Moreover, since $x(\omega)^+$ is a continuous function in ω and Ω is minimal, a similar argument to the one given in (ii) of Theorem 6.3 shows that $M(\omega, x(\omega)^+)$ is constant on Ω and, consequently, $M(\omega{\cdot}t, u(t,\omega,x(\omega)^+) = M(\omega{\cdot}t, x(\omega{\cdot}t)^+)$ for each $\omega \in \Omega$ and $t \geq 0$. Hence, from (5.4) we conclude that

$$0 = \sum_{i=1}^{m} D_i(u(t,\omega,x(\omega)^+) - x(\omega{\cdot}t)^+)\,,$$

that is, $D(u(t,\omega,x(\omega)^+)) = D(x(\omega{\cdot}t)^+)$ for each $\omega \in \Omega$ and $t \geq 0$. In addition, it is easy to check that $(\varphi_s)^+ = (\varphi^+)_s$ whenever $\varphi \in BU$ and $s \leq 0$, from which we deduce that $D((u(t,\omega,x(\omega)^+))_s) = D((x(\omega{\cdot}t)^+)_s)$ for each $s \leq 0$, $t \geq 0$, and $\omega \in \Omega$. That is, $\widehat{D}(u(t,\omega,x(\omega)^+)) = \widehat{D}(x(\omega{\cdot}t)^+)$ for each $t \geq 0$ and $\omega \in \Omega$, and since $\widehat{D}$ is an isomorphism $u(t,\omega,x(\omega)^+) = x(\omega{\cdot}t)^+$ for each $t \geq 0$ and $\omega \in \Omega$, which shows that K^+ is a minimal subset, as stated. Let $K_1 = \{(\omega, x(\omega)) \mid \omega \in \Omega\}$ and $K_2 = \{(\omega, y(\omega)) \mid \omega \in \Omega\}$ be two minimal subsets such that $M|_{K_i} = c$ for $i = 1, 2$. We fix $\omega \in \Omega$. The change of variable $\widehat{z}(t) = z(t) - y(\omega{\cdot}t)$ takes $(5.2)_\omega$ to

$$\frac{d}{dt} D\widehat{z}_t = G(\omega{\cdot}t, \widehat{z}_t)\,, \quad t \geq 0\,,\ \omega \in \Omega\,,$$

where $G(\omega{\cdot}t, \widehat{z}_t) = F(\omega{\cdot}t, \widehat{z}_t + y(\omega{\cdot}t)) - F(\omega{\cdot}t, y(\omega{\cdot}t))$. It is not hard to check that this is a new family of compartmental systems satisfying the corresponding hypotheses (C1)–(C5) and (C6)*, and

$$\widehat{K} = \{(\omega, x(\omega) - y(\omega)) \mid \omega \in \Omega\}$$

is one of its minimal subsets. As before

$$\widehat{K}^+ = \{(\omega, (x(\omega) - y(\omega))^+) \mid \omega \in \Omega\}$$

is also a minimal subset, and hence

$$K^+ = \{(\omega, y(\omega) + (x(\omega) - y(\omega))^+) \mid \omega \in \Omega\} = \{(\omega, z(\omega)) \mid \omega \in \Omega\}$$

is a minimal set for the initial family. For each $\omega \in \Omega$ we have $z(\omega) \geq_D y(\omega)$.

Let us assume that $Dz(\omega) \gg Dy(\omega)$ for each $\omega \in \Omega$, which implies that $D((x(\omega) - y(\omega))^+) \gg 0$ for each $\omega \in \Omega$. Consequently, $D((x(\omega) - y(\omega))_s^+) = D(((x(\omega) - y(\omega))_s)^+) = D((x(\omega \cdot s) - y(\omega \cdot s))^+) \gg 0$ for each $s \leq 0$, and we deduce that $\widehat{D}x(\omega) > \widehat{D}y(\omega)$, i.e., $x(\omega) >_D y(\omega)$ for each $\omega \in \Omega$, and $M|_{K_1} > M|_{K_2}$, a contradiction. Hence, there are an $\omega_0 \in \Omega$ and an $i \in \{1, \ldots, m\}$ such that $D_i z(\omega_0) = D_i y(\omega_0)$, and hypothesis (C7) provides that $z(\omega) = y(\omega)$ for each $\omega \in \Omega$. That is, $(x(\omega) - y(\omega))^+ \equiv 0$ for each $\omega \in \Omega$, or equivalently $x(\omega) - y(\omega) \leq_D 0$ for each $\omega \in \Omega$. Finally, as before, from $M|_{K_1} = M|_{K_2}$ we conclude by contradiction that $x(\omega) = y(\omega)$ for each $\omega \in \Omega$, and the minimal set K_c is unique, as stated. The same argument shows that $K_{c_1} <_D K_{c_2}$ whenever $c_1 < c_2$ and finishes the proof. ∎

REFERENCES

[1] O. Arino and F. Bourad, *On the asymptotic behavior of the solutions of a class of scalar neutral equations generating a monotone semiflow*, J. Differential Equations, 87 (1990), pp. 84–95.

[2] O. Arino and E. Haourigui, *On the asymptotic behavior of solutions of some delay differential systems which have a first integral*, J. Math. Anal. Appl., 122 (1987), pp. 36–46.

[3] R. Ellis, *Lectures on Topological Dynamics*, W. A. Benjamin, New York, 1969.

[4] G. Gripenberg, S.-O. Londen, and O. Staffans, *Volterra Integral and Functional Equations*, Encyclopedia Math. Appl., Cambridge University Press, Cambridge, New York, 1990.

[5] I. Györi, *Connections between compartmental systems with pipes and integro-differential equations*, Math. Modelling, 7 (1986), pp. 1215–1238.

[6] I. Györi and J. Eller, *Compartmental systems with pipes*, Math. Biosci., 53 (1981), pp. 223–247.

[7] I. Györi and J. Wu, *A neutral equation arising from compartmental systems with pipes*, J. Dynam. Differential Equations, 3 (1991), pp. 289–311.

[8] J. R. Haddock, T. Krisztin, and J. Wu, *Asymptotic equivalence of neutral and infinite retarded differential equations*, Nonlinear Anal., 14 (1990), pp. 369–377.

[9] J. R. Haddock, T. Krisztin, J. Terjéki, and J. Wu, *An invariance principle of Lyapunov-Razumikhin type for neutral functional differential equations*, J. Differential Equations, 107 (1994), pp. 395–417.

[10] J. K. Hale, *Theory of Functional Differential Equations*, Appl. Math. Sci. 3, Springer-Verlag, Berlin, Heidelberg, New York, 1977.

[11] J. K. Hale and K. R. Meyer, *A Class of Functional Equations of Neutral Type*, Mem. Amer. Math. Soc. 76, AMS, Providence, RI, 1967.

[12] J. K. Hale and S. M. Verduyn Lunel, *Introduction to Functional Differential Equations*, Appl. Math. Sci. 99, Springer-Verlag, Berlin, Heidelberg, New York, 1993.

[13] Y. Hino, S. Murakami, and T. Naiko, *Functional-Differential Equations with Infinite Delay*, Lecture Notes in Math. 1473, Springer-Verlag, Berlin, Heidelberg, 1991.

[14] J. A. Jacquez, *Compartmental Analysis in Biology and Medicine*, 3rd ed., Thomson-Shore Inc., Ann Arbor, MI, 1996.

[15] J. A. Jacquez and C. P. Simon, *Qualitative theory of compartmental systems*, SIAM Rev., 35 (1993), pp. 43–79.

[16] J. A. Jacquez and C. P. Simon, *Qualitative theory of compartmental systems with lags*, Math. Biosci., 180 (2002), pp. 329–362.

[17] J. Jiang and X.-Q. Zhao, *Convergence in monotone and uniformly stable skew-product semiflows with applications*, J. Reine Angew. Math, 589 (2005), pp. 21–55.

[18] V. Kolmanovskii and A. Myshkis, *Introduction to the Theory and Applications of Functional Differential Equations*, Math. Appl., Kluwer Academic, Dordrecht, The Netherlands, 1999.

[19] T. Krisztin and J. Wu, *Asymptotic periodicity, monotonicity, and oscillation of solutions of scalar neutral functional-differential equations*, J. Math. Anal. Appl., 199 (1996), pp. 502–525.

[20] S. Novo, R. Obaya, and A. M. Sanz, *Stability and extensibility results for abstract skew-product semiflows*, J. Differential Equations, 235 (2007), pp. 623–646.

[21] R. J. Sacker and G. R. Sell, *Lifting Properties in Skew-Products Flows with Applications to Differential Equations*, Mem. Amer. Math. Soc. 190, AMS, Providence, RI, 1977.

[22] D. Salamon, *Control and Observation of Neutral Systems*, Pitman, London, 1984.
[23] W. Shen and Y. Yi, *Almost Automorphic and Almost Periodic Dynamics in Skew-Product Semiflows*, Mem. Amer. Math. Soc. 647, AMS, Providence, RI, 1998.
[24] O. J. Staffans, *A neutral FDE with stable D-operator is retarded*, J. Differential Equations, 49 (1983), pp. 208–217.
[25] O. J. Staffans, *On a neutral functional differential equation in a fading memory space*, J. Differential Equations, 50 (1983), pp. 183–217.
[26] Z. Wang and J. Wu, *Neutral functional differential equations with infinite delay*, Funkcial. Ekvac., 28 (1985), pp. 157–170.
[27] J. Wu, *Unified treatment of local theory of NFDEs with infinite delay*, Tamkang J. Math., 22 (1991), pp. 51–72.
[28] J. Wu and H. I. Freedman, *Monotone semiflows generated by neutral functional differential equations with application to compartmental systems*, Canad. J. Math., 43 (1991), pp. 1098–1120.

SIAM J. MATH. ANAL.
Vol. 40, No. 3, pp. 1029–1048

COARSENING IN NONLOCAL INTERFACIAL SYSTEMS*

DEJAN SLEPČEV†

Abstract. We consider coarsening in interfacial systems driven by nonlocal energies. Of particular interest are the nonlocal Cahn–Hilliard equation and models of biological aggregation. The energies considered cause the system to separate into phases. The pattern of interfaces evolves under nonlocal surface-tension-type effects. The typical length scales grow and the pattern coarsens. We prove a rigorous upper bound on the coarsening rate. The proof uses the energy-based approach to estimates on rate of coarsening introduced by Kohn and Otto [*Comm. Math. Phys.*, 229 (2002), pp. 375–395]. To show the required estimates on the flatness of the energy landscape we develop a geometric approach which is applicable to a wider class of problems, which includes ones based on local, gradient-type energies.

Key words. coarsening, interpolation inequality, nonlocal equation, phase segregation, biological aggregation, Wasserstein distance

AMS subject classifications. 45K05, 35B99, 82C24, 92D25

DOI. 10.1137/080713598

1. Introduction.

Our main focus is systems driven by nonlocal energies. Of particular interest are nonlocal Cahn–Hilliard-type equations and equations modeling biological aggregation. The nonlocal Cahn–Hilliard equations that we consider were derived by Giacomin and Lebowitz [8] as limits of the lattice-gas dynamics modeling phase segregation in binary alloys. In this setting, they represent a refinement in modeling over the standard, fourth order Cahn–Hilliard equations. The equations modeling biological aggregation were derived by Topaz, Bertozzi, and Lewis [18].

The mathematical descriptions of these systems are rather similar. Both equations are gradient flows of the same general energy:

$$E(u) := \iint (u(x) - u(y))^2 K(x-y)dxdy + \int W(u(x))dx. \tag{1}$$

Here $K \geq 0$ is the interaction kernel, and W is a double-well potential whose minima are at 0 and 1. We leave the domain of integration vague at the moment. Heuristically it is convenient to consider the domains to be $\mathbb{R}^N \times \mathbb{R}^N$ and $\mathbb{R}^N$, respectively. However, for technical reasons, when stating and proving rigorous results we consider the problem on a finite domain.

The equations we study are gradient flows of the energy, in the appropriate metrics:

$$u_t - \nabla \cdot \left(\mu(u) \nabla \left(\frac{\delta E}{\delta u} \right) \right) = 0. \tag{2}$$

That is,

$$u_t - \nabla \cdot \left(\mu(u) \nabla \left(4 \int K(y)dy\, u - 4K * u + W'(u) \right) \right) = 0. \tag{3}$$

*Received by the editors January 17, 2008; accepted for publication (in revised form) June 10, 2008; published electronically October 13, 2008. The research of this work was partially supported by NSF grant DMS-0638481 and by the Center for Nonlinear Analysis through NSF grants DMS-0405343 and DMS-0635983.

http://www.siam.org/journals/sima/40-3/71359.html

†Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 (slepcev@math.cmu.edu).

For both equations the mobility μ is a nonnegative function. More precisely, for the aggregation equation $\mu(u) = u$, while for the nonlocal Cahn–Hilliard equation $\mu > 0$ on $[0, 1]$.

The second term of the energy causes the system to separate into phases, while the first term penalizes the existence of interfaces. The energy E is a nonlocal counterpart of the energy

$$E_{loc}(u) := \int \frac{1}{2}|\nabla u(x)|^2 + W(u(x))dx. \tag{4}$$

Roughly speaking, both of the energies measure interfacial area. The longer the length scale in the system, the better the approximation to the interfacial area. More precisely, for both energies, the Γ-limit of the appropriately rescaled energy is the functional measuring perimeter of the set occupied by one of the phases

$$E \xrightarrow{\Gamma} \text{const.}\, E_{per}.$$

E_{per} is defined for BV functions with the range $\{0, 1\}$. For E_{loc} this is the result of Modica and Mortola [15] (see also [14]), while for E it was proven by Alberti et al. [2] (see also [1]). Moreover matched asymptotics arguments (by Giacomin and Lebowitz [9] and by Bertozzi and Slepčev [4]) show that the sharp-interface limits of the dynamics described by (2) are the Mullins–Sekerka (MS) equation for the nonlocal Cahn–Hilliard equation and the Hele–Shaw (HS) equation for the aggregation model.

After the interfaces have formed, the system slowly evolves, reducing the interfacial area. During this process the length scales that characterize the coarseness of a configuration grow. We are interested in the rate at which these length scales grow—the rate of coarsening. The fact that the sharp interface limits, (MS) and (HS), are both invariant under the scaling $x \to \lambda x$, $t \to \lambda^3 t$ suggests that the typical length scale grows as $t^{1/3}$. We prove a weak formulation of this statement, following the technique of Kohn and Otto [10], who proved the result for gradient flows of local energies (4). We use the energy as the measure of the coarseness of the system. In particular let $\overline{E}$ be the energy density, that is, the energy per unit volume. Note that $\overline{E}$ has units of $1/\textit{length}$. We show a weak version of the statement

$$\overline{E} \gtrsim t^{-1/3}.$$

This provides an upper bound on rate of coarsening as it shows that the interfacial area cannot decay faster than the given rate.

Outline. In the remainder of the introduction we discuss the gradient-flow structure of the equations, and the framework for obtaining rigorous result on coarsening rates introduced by Kohn and Otto. We also introduce the two applications we have in mind in more detail. In section 2 we list the assumptions needed and give the precise formulation of the main results on the rate of coarsening. In particular we present a relaxed formulation of the abstract Kohn–Otto framework that is applicable to weak solutions of gradient-flow equations. We also show that the "geodesic distance" associated to the metric of the configuration space can be utilized even when the mobility is nonlinear. In section 3 we present the proof of the bound on the rate of coarsening. The main technical ingredient, the interpolation inequality, is proved in section 4. The approach we take in proving the interpolation inequality is general; essentially the same proof covers both types of mobilities and both nonlocal and local energies. We illustrate the application to local energies in subsection 4.1.

1.1. Gradient flow structure. We now introduce the geometric structure of (2). It is based on the formal Riemannian viewpoint developed by Otto [16]. Equation (2) can be understood as a gradient flow of the energy (1) on the manifold of configurations. Since the equation is in divergence form it preserves the integral of u over the space. Thus the solution of the equation is a path on the manifold of functions with the same integral.

At each point the tangent space is the set of possible perturbations, all of which have mean zero. The local metric is defined as follows: Let s_1, s_2 be two tangent vectors at u. Then

$$g_u(s_1, s_2) = \int \mu(u) \nabla p_1 \cdot \nabla p_2, \tag{5}$$

where

$$-\nabla \cdot (\mu(u) \nabla p_i) = s_i \quad \text{for } i = 1, 2.$$

Equation (2) is the gradient flow of energy (1) with respect to the metric (5), that is, for every tangent vector s

$$g(u_t, s) = -\frac{\delta E}{\delta u}[s].$$

Considering the configuration space as a manifold enables us to measure the steepness of the energy landscape in a certain sense. This information provides bounds on the speed of the dynamics.

In particular the local metric gives rise to a global metric on the manifold. Given a regular enough path, $v(s)$ for, say, $s \in [0, 1]$, on the manifold, we can measure its length

$$\text{length}(v) = \int_0^1 \sqrt{g_{v(s)}(v', v')}\, ds.$$

We can then define the global metric on the manifold: Let the distance of u_1 and u_2 be

$$d(u_1, u_2) = \inf\{\text{length}(v) \,:\, v \text{ is a path connecting } u_1 \text{ and } u_2\}.$$

It turns out that when $\mu = \text{const.}$, then $d(u_1, u_2)$ is a multiple of the H^{-1} norm, while for $\mu(u) = u$ the distance becomes the Wasserstein metric.

1.2. Kohn–Otto framework. Kohn and Otto [10] introduced an approach for obtaining information on the flatness of the energy landscape, and consequently on the rate of coarsening. The approach is robust and has been applied to studies of coarsening in epitaxial growth, mean-field models, thin-liquid films, and other systems [6, 7, 5, 11, 12, 17]. See also [13] for a related result.

We first present it in the abstract setting used in [17], which applies to gradient flows. Consider an energy E on a Riemannian manifold $(\mathcal{M}, g)$. The metric g introduces a global distance on $\mathcal{M}$, which we denote by d.

PROPOSITION 1. *Let $h^* \in \mathcal{M}$. Let $h : \mathbb{R}_+ \to \mathcal{M}$ be a solution of*

$$h_t \;=\; -\text{grad}E(h), \tag{6}$$

and let $h(0) = h_0$.

Assume that for some $\alpha \geq 0$ the interpolation inequality

$$E(h)\,\mathrm{dist}(h,h^*)^\alpha \geq 1 \quad \text{for all } h \in \mathcal{M} \text{ with } E(h) \leq \varepsilon \tag{7}$$

holds. Then for $\sigma \in (1, 1+\frac{2}{\alpha})$

$$\int_0^T E(h(t))^\sigma \, dt \gtrsim \int_0^T (t^{-\frac{\alpha}{\alpha+2}})^\sigma \, dt, \tag{8}$$

provided $T \gg \mathrm{dist}(h_0,h^*)^{\alpha+2}$ and $E(h(0)) \leq \varepsilon$.

Remark 1. The precise meaning of $\gtrsim$ and $\gg$ is the following: For all $\sigma \in (1, 1+\frac{2}{\alpha})$ there exists a constant $C = C(\alpha,\sigma)$ such that for all $\delta > 0$ there exists $C_\delta = C(\alpha,\sigma,\delta)$:

$$\int_0^T E(h(t))^\sigma \, dt \geq (1-\delta)C \int_0^T (t^{-\frac{\alpha}{\alpha+2}})^\sigma \, dt, \tag{9}$$

provided $T \geq C_\delta \,\mathrm{dist}(h_0,h^*)^{\alpha+2}$.

Proof of the proposition is based on the ODE arguments of [10] and can be found in [17].

We adapt Proposition 1 to our setting in section 2 (Theorem 2). In particular since the Riemannian structure is only formal, the theorem states precisely which elements of the geometric structure are needed. Furthermore it applies to weak solutions of the gradient flow equations.

The main technical difficulty in applying the proposition is showing the interpolation inequality. Section 4 is devoted to proving interpolation inequalities relevant to nonlocal energies.

1.3. Nonlocal Cahn–Hilliard equation. Equation (2) is a rescaled version of the model by Giacomin and Lebowitz [8, 9]. We introduce it in original variables below and, for completeness, present the rescaling needed.

The free energy is given by

$$\mathcal{E} = \frac{1}{4}\int_\Omega \int_{\mathbb{R}^N} K(x-y)(\rho(x)-\rho(y))^2 dxdy + \int_\Omega f_c(\rho(x))dx.$$

The associated gradient flow is

$$\rho_t - \nabla \cdot \left(\sigma(\rho)\,\nabla\left(\frac{\delta \mathcal{E}}{\delta \rho}\right)\right) = 0.$$

Here $K \geq 0$ is a smooth kernel with symmetry $K(x) = K(-x)$. Giacomin and Lebowitz assume that K is compactly supported, with support contained in Ω. This assumption is physically quite reasonable, but it is not necessary from the mathematical point of view. The function f_c is a double-well potential, symmetric about $\frac{1}{2}$ with minima at $\frac{1}{2} \pm m$. The mobility function σ is assumed to be smooth, symmetric about $\frac{1}{2}$, and positive on $(0,1)$, with

$$\sigma(0) = 0 \quad \text{and} \quad \sigma(1) = 0. \tag{GL1}$$

Let

$$f(\rho) = f_c(\rho) + \frac{\int_{\mathbb{R}^N} K(x)dx}{2}\left(\rho - \frac{1}{2}\right)^2.$$

Giacomin and Lebowitz assume that for some $c > 0$ and for all $\rho \in (0,1)$

$$\frac{1}{c} \le \sigma(\rho) f''(\rho) \le c. \tag{GL2}$$

This assumption is needed for the existence/uniqueness theory they use (it makes the equation uniformly parabolic), but is not directly required for the coarsening estimates.

Under the assumptions above, Giacomin and Lebowitz show that for initial datum $0 < \rho_0 < 1$, there exists a unique weak solution $\rho \in L^2([0,T], H^1(\Omega))$ with $\rho_t \in L^2([0,T], H^{-1}(\Omega))$ for any $T > 0$. Furthermore $0 < \rho < 1$.

Rescaling. We rescale the dependent variable, the potential, and the mobility so that the wells of the new potential are 0 and 1. Let

$$\begin{aligned} u &= \frac{1}{2m}\left(\rho - \frac{1}{2}\right) + \frac{1}{2}, \\ W_m(u) &= \frac{1}{4m^2} f_c(\rho) = \frac{1}{4m^2} f_c\left(2m\left(u - \frac{1}{2}\right) + \frac{1}{2}\right), \\ \mu_m(u) &= \sigma(\rho) = \sigma\left(2m\left(u - \frac{1}{2}\right) + \frac{1}{2}\right). \end{aligned}$$

Under this rescaling u solves (2) and is hence a gradient flow of (1).

1.4. Biological aggregation. Topaz, Bertozzi, and Lewis [18] introduced a model of biological aggregation that emerges due to "social forces" between individuals. That is, the individuals are attracted to other individuals of their species, but avoid overcrowding. The population is modeled by its density u. The velocity of individuals is modeled as

$$v = v_a + v_r = \nabla(K * u) - \nabla g(u),$$

where $v_a = \nabla(K * u)$ is the term modeling attraction to other individuals which are being sensed through the kernel K. The term modeling repulsion, v_r, is given by a local operator $v_r = -\nabla g(u)$, where g is an increasing function. The continuity equation then reads

$$u_t + \nabla \cdot (u\, v) = u_t + \nabla \cdot (u \nabla (K * u - g(u))) = 0.$$

From a biological perspective it is reasonable to assume that $g'(0) = 0$ and g is strictly convex. However, it is sufficient to assume that

$$\begin{aligned} &\text{the function } g'(z) - \int_{R^N} K(x)dx \text{ has exactly one zero on } \mathbb{R}^+, \\ &g'(0) < \int_{R^N} K(x)dx \quad \text{and} \quad \liminf_{z\to\infty} g'(z) - \int_{\mathbb{R}^N} K(x)dx > 0. \end{aligned} \tag{BA}$$

Under this assumption u solves (3) for some double-well potential W. More precisely, let $G(z) := \int_0^z g(s)ds$ and $\tilde{W}(z) := G(z) - \frac{1}{2}\int_{\mathbb{R}^N} K(x)dx\, z^2$. The condition (BA) implies that $\tilde{W}$ is concave at 0, and has exactly one inflection point. Thus we can define

$$W(z) := 4\left(\tilde{W}(z) - \left(\min_{s>0} \frac{\tilde{W}(s)}{s}\right) z\right).$$

Then W is a double-well potential on $[0,\infty)$ with one well at 0. It follows that

$$u_t - \nabla \cdot \left(u \nabla \left(\int_{\mathbb{R}^N} K(y)dy\, u - K * u + \frac{1}{4} W'(u) \right) \right) = 0,$$

which after scaling the time by factor 4 is (3).

On the level of the model, the equation provides information on why herds (or other animal groups) form, why they have an almost constant density, why they have sharp boundaries, and how they evolve. Numerical simulations conducted in one dimension in [18] also observe the coarsening phenomenon. The primary driving force for coarsening in one dimension is the nonlocal interaction via kernel K, as there are no surface-tension-like effect. The rate of coarsening depends on the decay of K at ∞. Nevertheless the rigorous bounds we prove still apply and are in fact optimal for certain kernels.

2. Statement of the result. When thinking about coarsening we have in mind an infinite domain on which coarsening persists for all time. However, building the theory for such solutions poses major challenges. We instead consider domains of finite size and prove results that are independent of the domain size. In particular we consider the domain $\Omega = [0,\Lambda]^N$. We investigate the dynamics of periodic configurations on $\mathbb{R}^N$ with period cell Ω. Thus we consider Ω with the topology of the torus $\mathbb{R}^N/(\Lambda\mathbb{Z})^N$. In particular the distances on Ω are measured on the torus, and thus may be different from the ones measured in $\mathbb{R}^N$.

Throughout the paper we use the following, somewhat nonstandard notation. For $U \subseteq \Omega$ and a function u,

$$⨍_U f(x)dx := \frac{1}{|\Omega|} \int_U f(x)dx \quad \text{and} \quad \|U\| := \frac{|U|}{|\Omega|}.$$

Let P be the maximal interval containing 1 on which $\mu > 0$:

$$P = \{z \,:\, z \le 1,\ \ \mu|_{[z,1]} > 0\} \cup \{z \,:\, z \ge 1,\ \ \mu|_{[1,z]} > 0\}.$$

The configuration space is

$$\mathcal{M} := L^1(\Omega, \overline{P}).$$

While the configurations are functions defined on Ω, when convenient we also consider them as periodic functions of $\mathbb{R}^N$.

To a configuration, u, we associate the energy density

$$\overline{E}(u) := ⨍_\Omega \int_{\mathbb{R}^N} (u(x) - u(y))^2 K(x-y)dxdy + ⨍_\Omega W(u(x))dx. \tag{10}$$

If the expression is not defined, we say that the energy density is infinite. Conditions on the interaction kernel, K, and the double-well potential, W, are described below.

We make the following assumptions on the interaction kernel K:

(K1) K is nonnegative and $K \in L^1(\mathbb{R}^N) \cap C^2(\mathbb{R}^N)$.

(K2) $K(x) = K(-x)$ for all $x \in \mathbb{R}^N$. (This condition ensures the symmetry of the interaction term in (10) with respect to x and y.)

(K3) $K(0) > 0$.

(K4) $K \in W^{2,1}(\mathbb{R}^N)$ and $\|K\|_{C^2(\mathbb{R}^N)} < \infty$.

The last condition is needed only for the existence theory [4]. Condition (K3) is not essential either, but significantly simplifies parts of the presentation. In particular it enables us to associate a length scale to a kernel in the following way: For $r > 0$ let

$$\kappa(x) := \frac{1}{|B(0,1)|}\chi_{B(0,1)}(x) \quad \text{and} \quad \kappa_r(x) := \frac{1}{r^N}\,\kappa\left(\frac{x}{r}\right), \tag{11}$$

where χ_U is the characteristic function of the set U. Given $r > 0$ let $h_K(r) := \sup\{c : K \geq c\kappa_r\}$. Note that by assumption (K3), $h_K(r) > 0$ for r small. It is not hard to prove that $h_K(r) \to 0$ as $r \to 0$ and also as $r \to \infty$. Consider the location of the maximum of $h_K(r)$. If there is more than one maximum, we pick the first one. More precisely let

$$r_K := \min\{r_{max} \mid h_K(r_{max}) = \max_{r>0} h_K(r)\}. \tag{12}$$

We let

$$H_K = h_K(r_K).$$

To state the conditions on the potential W we define

$$a := \fint_\Omega u(x,0)dx.$$

To simplify the presentation, from here on we restrict our attention only to

$$0 < a \leq \frac{1}{2}. \tag{13}$$

We assume the following:

(W1) W is a nonnegative continuous function.

(W2) $W(0) = W(1) = 0$ and $W > 0$ on $\overline{P}\backslash\{0,1\}$.

(W3) At least linear growth at $\pm\infty$: There exists a constant h_W such that $W(z) \geq 8h_W(z-1)$ for all $z \in P\cap(\frac{9}{8},\infty)$ and $W(z) \geq h_W|z|$ for all $z \in P\cap(-\infty,-\frac{a}{4})$. We can furthermore require

$$h_W \leq W(z) \text{ for all } z \in \left[\frac{a}{4},\frac{7}{8}\right].$$

The condition on growth of W is needed only if P is infinite.

We now state the main results. Theorem 2 is an adaptation of Proposition 1 that applies to the configuration spaces we are investigating. In particular the paths are given as weak solutions of the continuity equation. Having the gradient-flow structure is reduced to requiring a dissipation inequality. The theorem relies on interpolation inequalities we establish in section 4. In Corollary 3 we apply the theorem to the main equations of our interest. The classes of solutions of (3) studied in [9] and [4] satisfy the conditions of the theorem, and thus the bounds on coarsening hold.

Below, by C^{weak} we mean continuous with respect to weak topology of the target space.

THEOREM 2. *Let $\Omega = [0,\Lambda]^N$. Assume that conditions* (13), (K1)–(K3), *and* (W1)–(W3) *hold and that mobility $\mu(z) \equiv z$ or that $\mu \in C(P,(0,c_\mu])$. Suppose that $u \in C^{weak}([0,\infty),L^1(\Omega,\overline{P}))$ is a weak solution of*

$$u_t + \nabla \cdot J = 0$$

for some flux $J \in L^1(\Omega, \mathbb{R}^N)$. *Assume that the energy dissipation inequality holds: For almost all* $0 \le t_1 < t_2$

$$\int_{t_1}^{t_2} \int_\Omega \frac{1}{\mu(u)} |J|^2 dx dt \le -(E(u(t_2)) - E(u(t_1))). \tag{14}$$

In the case $\mu(z) \not\equiv z$ *we also need the regularity assumption that* $u(t) \in L^2(\Omega)$ *for all* $t \ge 0$ *and that the range of* $u(\cdot, 0)$ *is contained in* P. *Assume further that* $\limsup_{t\to 0+} \overline{E}(u(\cdot, t)) \ll 1$.

Then for all $\sigma \in (1, 2)$ *there exists a constant* $C = C(\sigma, a, H_K, r_K, c_\mu)$ *such that for all* $T \gg 1$

$$\int_0^T \overline{E}(u(t))^\sigma dt \ge C \int_0^T \left(t^{-\frac{1}{3}}\right)^\sigma dt. \tag{15}$$

The precise meaning of condition $\overline{E} \ll 1$ is that $\overline{E}$ has to be small enough for the interpolation inequality to hold. Precise values can be found in the statement of Theorem 6.

By weak solution we mean that for all $\phi \in C_c^\infty(\Omega \times [0, \infty))$

$$\iint_{[0,\infty)\times\Omega} u\phi_t + J \cdot \nabla\phi \, dx dt + \int_\Omega u(x, 0)\phi(x, 0) dx = 0.$$

Recall that we consider Ω with the topology of a torus, which means that the test functions used in the definition of a weak solution are also defined on the torus; in other words they are periodic.

The form of the equation ensures (via testing against test functions with only time dependence) that

$$\int_\Omega u(x, t) dx = \text{const}.$$

Note that the dissipation inequality (14) is an equality if u is a classical solution of the gradient flow (2), when $J = -\mu(u)\nabla\frac{\delta E}{\delta u}$.

COROLLARY 3. *Let* $\Omega = [0, \Lambda]^N$. *Assume that conditions* (K1)–(K4) *and* (W1)–(W3) *hold. Assume only one of the following holds:*

(i) $\mu(z) \equiv z$, $W \in C^2([0, \infty))$ *with* $W'' > -4\int_{\mathbb{R}^N} K(y)dy$ *on* $\mathbb{R}^+$, *and* u *is the solution of* (3) *in the sense of* [4] *with* $u(\cdot, 0) \in L^\infty(\Omega)$.

(ii) $\mu(z) \le c_\mu$ *for all* z, *conditions* (GL1) *and* (GL2) *hold, and* u *is a solution of* (3) *in the sense of* [9]. *Furthermore assume that the range of* $u(\cdot, 0)$ *is contained in the interior of* P.

(iii) $0 < \mu(z) \le c_\mu$ *for all* z, (GL2) *holds on* $\mathbb{R}$, *and* u *is a solution of* (3) *in the sense of* [9]. *Furthermore assume that* $u(\cdot, 0) \in L^\infty(\Omega)$.

Assume that $\limsup_{t\to 0+} \overline{E}(u(\cdot, t)) \ll 1$ *and* (13) *holds. Then for all* $\sigma \in (1, 2)$ *there exists a constant* $C = C(\sigma, a, H_K, r_K, c_\mu)$ *such that for all* $T \gg 1$

$$\int_0^T \overline{E}(u(t))^\sigma dt \ge C \int_0^T \left(t^{-\frac{1}{3}}\right)^\sigma dt. \tag{16}$$

Proof. In the first case the conditions of W imply that associated $g(z) := \int_{\mathbb{R}^N} K(x)dx + \frac{1}{4}W'(z)$ is an increasing function. This in turn implies that the conditions under which Bertozzi and Slepčev [4] proved existence of solutions of (3) hold.

Properties of solutions ensuring that the assumptions of Theorem 2 hold were also established. This implies the claim of the corollary.

Case (ii) the existence theory needed for Theorem 2 to apply was established by Giacomin and Lebowitz [9].

The case (iii) is in principle simpler than case (ii) and includes the constant mobility case. The only technical issue is that the L^∞ bounds used in [9] follow from condition (GL1). In our case appropriate bounds can be established, for example, as in [4]. □

3. Proof of Theorem 2. We seek to apply the framework of Proposition 1. However, the configuration space, $\mathcal{M}$, is not a true Riemannian manifold and the only remnant of the gradient flow structure is the energy dissipation inequality (14). Nevertheless arguments of the proof of the proposition can be adapted to include this setting. We define the "geodesic distance" on $\mathcal{M}$ as follows: Given $u_0, u_1 \in \mathcal{M}$ let us first define a representation of admissible paths between u_0 and u_1:

$$\mathcal{A}(u_0,u_1) := \Big\{ (u,J) \ : u:[0,1]\to\mathcal{M}, J\in L^1(\Omega\times[0,1],\mathbb{R}^N) \text{ such that}$$
$$u_t + \nabla\cdot J = 0 \quad \text{on } \Omega\times[0,1] \text{ weakly},$$
$$u\in C^{weak}([0,1],L^1(\Omega)) \text{ and } u(0)=u_0, u(1)=u_1, \text{ and}$$
$$\int_0^1\int_\Omega \frac{1}{\mu(u(x,t))}\,|J(x,t)|^2 dxdt < \infty\Big\}.$$

We define

$$d^2(u_0,u_1) := \inf_{(u,J)\in\mathcal{A}} \int_0^1\int_\Omega \frac{1}{\mu(u(x,t))}\,|J(x,t)|^2 dxdt. \tag{17}$$

Here $\frac{0}{0}=0$. We note that d may, in general, be infinite. It follows from the definition that d satisfies the triangle inequality. This can be shown by concatenating the appropriate test flows (with optimally rescaled times).

Let u be as in the statement of the theorem. We define

$$L(t) := d(u(t),a), \qquad \overline{L}(t) := \frac{1}{\sqrt{|\Omega|}}\, d(u(t),a). \tag{18}$$

In the case $\mu(u)=u$ it follows from the characterization of d given below in (19) that $L(t)$ is finite for all t. In the other case, from the assumption on the range of u_0 it follows that $L(0)$ is finite. To see this it is enough to consider the test pair $(\tilde{u},\tilde{J})$ with $\tilde{u}(s)=u_0+s(a-u_0)$ for $s\in[0,1]$ and $\tilde{J}=\nabla p$, where p solves $-\Delta p = a-u_0$. The fact that $L(t)$ is finite for all t then follows from the argument for continuity of L given in Lemma 4.

Let

$$\overline{E}(t) := \overline{E}(u(t)).$$

From (14) we have that $\overline{E}$ is nonincreasing almost everywhere. We now modify $\overline{E}$ on a set of measure 0 to ensure that it is nonincreasing:

$$\overline{E}_{new}(t) = \min\{\liminf_{s\to t-} \overline{E}(s), \overline{E}(t)\}.$$

Inspecting the proof of Proposition 1 from [17] shows that in addition to the interpolation inequality, one only needs the inequality

$$\left(\frac{d\overline{L}}{d\overline{E}}\right)^2 \leq -\frac{dt}{d\overline{E}}$$

(since $\overline{E}$ is nonincreasing, $\overline{L}$ can be considered as a function of $\overline{E}$), which follows from the more familiar form of the dissipation inequality,

$$\left(\frac{d\overline{L}}{dt}\right)^2 \leq -\frac{d\overline{E}}{dt},$$

where both inequalities are to be understood as a comparison of measures (with given densities). The latter inequality in turn follows from the assumption (14). We prove these claims in Lemmas 4 and 5.

To be able to prove the interpolation inequalities we need a more workable form of d. In the case $\mu(u) = u$ the distance d is nothing else than the Wasserstein distance. This was shown by Benamou and Brenier [3] (see also section 8.1 in Villani's book [19]):

(19)
$$d_W(u_0,u_1)^2 \;=\; \inf\left\{\iint_{\Omega\times\Omega}|x-y|^2 d\pi(x,y) \;\Big|\; \int_\Omega d\pi(\cdot,y)=u_0,\ \int_\Omega d\pi(x,\cdot)=u_1\right\}.$$

The distance above, $|x-y|$, is taken on torus Ω.

In the case $\mu(u) \leq c_\mu$ first note that from the definition of the distance (17), it follows that the distance corresponding to $\mu(u)$ is greater than the distance corresponding to constant mobility c_μ. Thus

$$\overline{L}_\mu \geq \overline{L}_{c_\mu} = \frac{1}{c_\mu}\overline{L}_1.$$

Thus it is enough to establish the interpolation inequality for $\overline{L} = \overline{L}_1$. But for mobility equal to one, the distance is the H^{-1} norm. More precisely, for $u_0, u_1 \in \mathcal{M} \cap L^2(\Omega)$, $\int_\Omega u_1 - u_0 dx = 0$, and hence we can consider the following representation of the H^{-1} norm:

$$d(u_0,u_1)^2 = \|u_0-u_1\|_{H^{-1}}^2 = \int_\Omega |\nabla p|^2 dx,$$

where $p \in H^2(\Omega)$ (p is periodic by topology of Ω) is a solution of

$$-\Delta p = u_1 - u_0.$$

Proof of this claim is straightforward; it relies on convexity in J of the functional on the right-hand side of (17) and the observation that $J = \nabla p$, along with $u(t) = u_0 + t(u_1 - u_0)$, minimizes the action functional. One can also show that for $u \in \mathcal{M} \cap L^2(\Omega)$

(20)
$$\overline{L}(u) = \max_{\xi\in H^1(\Omega),\ \xi\not\equiv\text{const.}} \frac{⨍_\Omega (u-a)\xi dx}{\sqrt{⨍_\Omega |\nabla\xi|^2 dx}}$$

by using Cauchy–Schwarz inequality to show that $\xi = p$ (with $u_1 = u$ and $u_2 = a$) is the maximizing function. Given that $u(t) \in \mathcal{M} \cap L^2(\Omega)$ for all t we can use this characterization.

Thus to complete the proof of the theorem one only needs the interpolation inequalities established in section 4.

We now prove the two lemmas that were used in the arguments above.

LEMMA 4. *Assume that u satisfies the conditions of Theorem 2 and $\overline{L}$ is defined in* (18). *Then $\overline{L}$ is a continuous function and for almost all $t \geq 0$ and $h > 0$*

$$\left(\frac{\overline{L}(u(t+h)) - \overline{L}(u(t))}{h}\right)^2 \leq -\frac{\overline{E}(u(t+h)) - \overline{E}(u(t))}{h}.$$

Proof. By the assumptions of the theorem u is a distributional solution of

$$u_t + \nabla \cdot J = 0 \qquad \text{on } \Omega \times [t, t+h].$$

Moreover it follows from assumption (14) for all $t \geq 0$ and all $h > 0$ that

$$\int_t^{t+h} \int_\Omega \frac{1}{\mu(u)} |J|^2 dx dt \leq \limsup_{s \to 0+} E(u(s)) < \infty.$$

Note that it was also assumed that $u \in C^{weak}([t, t+h], L^1(\Omega))$. Thus (u, J), after appropriate rescaling in time, belongs to $\mathcal{A}(u(t), u(t+h))$. By the triangle inequality,

$$\begin{aligned}\left(\overline{L}(u(t+h)) - \overline{L}(u(t))\right)^2 &\leq \inf_{(\tilde{u},\tilde{J}) \in \mathcal{A}(u(t),u(t+h))} \int_0^1 \int_\Omega \frac{1}{\mu(\tilde{u}(x,s))} |\tilde{J}(x,s)|^2 \, dx \, ds \\ &\leq h \int_0^h \int_\Omega \frac{1}{\mu(u(x,t+s))} |J|^2 dx ds.\end{aligned}$$

Thus $\overline{L}$ is a continuous function. Dividing the above by h^2 and using (14) gives that for almost all $t \geq 0$ and $h > 0$

$$\begin{aligned}\left(\frac{\overline{L}(u(t+h)) - \overline{L}(u(t))}{h}\right)^2 &\leq \frac{1}{h} \int_0^h \int_\Omega \frac{1}{\mu(u(x,t+s))} |J|^2 dx ds \\ &\leq -\frac{\overline{E}(u(t+h)) - \overline{E}(u(t))}{h}. \qquad \square\end{aligned}$$

For a function e on $\mathbb{R}$ let us define $e(t+) := \lim_{s \to t+} e(s)$ and $e(t-) := \lim_{s \to t-} e(s)$.

LEMMA 5. *Let e be a nonnegative, nonincreasing function on $[0, \infty)$. Let l be a continuous function on $[0, \infty)$, such that*

$$\left(\frac{l(t_2) - l(t_1)}{t_2 - t_1}\right)^2 \leq -\frac{e(t_2) - e(t_1)}{t_2 - t_1} \qquad \text{for almost all } t_2 > t_1 \geq 0. \tag{21}$$

Then $l(t)$ is an absolutely continuous function on $[0, \infty)$ and for all $\tau_2 > \tau_1 \geq 0$

$$\int_{\tau_1}^{\tau_2} \left(\frac{dl}{dt}\right)^2 dt \leq e(\tau_1+) - e(\tau_2-).$$

Furthermore, consider $t(e) := \sup\{t \,:\, e(t) \geq e\}$, the "inverse" of the function e and $l(e) := l(t(e))$. Then

$$\left(\frac{l(e_2) - l(e_1)}{e_2 - e_1}\right)^2 \leq -\frac{t(e_2) - t(e_1)}{e_2 - e_1} \qquad \text{for all } e(0) \geq e_1 > e_2 \geq 0.$$

Consequently l is an absolutely continuous function of e, and for all $e(0) \geq e_1 > e_2 \geq 0$

$$\int_{e_2}^{e_1} \left(\frac{dl}{de}\right)^2 de \leq t(e_2+) - t(e_1-).$$

Proof. If $e(0) = 0$, the proof is trivial. So assume $e(0) > 0$. Continuity of l implies that

$$\left(\frac{l(t_2) - l(t_1)}{t_2 - t_1}\right)^2 \leq -\frac{e(t_2-) - e(t_1+)}{t_2 - t_1} \qquad \text{for all } t_2 > t_1 \geq 0.$$

Let $\varepsilon > 0$ and $\delta := \varepsilon^2/e(0)$. Let $[x_i, y_i]$ for $i = 1, \ldots, m$ be a family of disjoint intervals on $[0, \infty)$ of total length less than δ:

$$\sum_{i=1}^{m} y_i - x_i < \delta.$$

Then

$$\begin{aligned}\sum_{i=1}^{m} |l(y_i) - l(x_i)| &\leq \sum_{i=1}^{m} \sqrt{e(x_i) - e(y_i)}\,\sqrt{y_i - x_i} \\ &\leq \sqrt{\sum_{i=1}^{m} e(x_i) - e(y_i)}\,\sqrt{\sum_{i=1}^{m} y_i - x_i} \\ &\leq \sqrt{e(0)}\,\sqrt{\delta} = \varepsilon.\end{aligned}$$

So l is absolutely continuous.

To prove the second claim note that for any $h > 0$

$$\begin{aligned}\int_{\tau_1}^{\tau_2} \left(\frac{l(t+h) - l(t)}{h}\right)^2 dt &\leq \int_{\tau_1}^{\tau_2} \frac{e(t+) - e(t+h-)}{h} dt \\ &\leq \frac{1}{h}\left(\int_{\tau_1}^{\tau_1+h} e(t+)dt - \int_{\tau_2}^{\tau_2+h} e(t-)dt\right).\end{aligned}$$

By taking the $\liminf_{h\to 0}$ and using Fatou's lemma we obtain

$$\int_{\tau_1}^{\tau_2} \left(\frac{dl}{dt}\right)^2 dt \leq e(\tau_1+) - e(\tau_2+).$$

Now use this claim on interval $(\tau_1, \tau_2 - \varepsilon)$ and take the limit as $\varepsilon \to 0$.

To prove the remaining claims, note that

$$(l(t(e_2)) - l(t(e_1)))^2 \leq (e(t(e_2)-) - e(t(e_1)+))(t(e_2) - t(e_1)).$$

Observing that $e(t(e_2)-) \geq e_2$ and $e(t(e_1)+) \leq e_1$ yields the desired inequality. The claims then follow from the arguments presented above. ∎

4. Interpolation inequalities. In this section we prove the interpolation inequalities needed. As shown in section 3 we only need to consider the $\overline{L}$ corresponding Wasserstein distance (19) and to H^{-1} norm (20).

The proof we present is general and extends to local energies, which we discuss in subsection 4.1. It also captures the improved constants established in [5]; see Remark 2. It is based on simple geometric heuristic. Consider function $\tilde{u}$ with range $\{0,1\}$ and $\kappa_r * \tilde{u}$, its average over ball of radius r. Then for r small

$$\int |\tilde{u} - \kappa_r * \tilde{u}| dx$$

contains information about the interfacial area, while for r large it carries information on the distance $d(\tilde{u}, a)$. This allows us to interpolate between the energy and the distance. We divided the proof into steps and present the motivation at their beginning.

THEOREM 6 (interpolation inequality). *Let $0 < a \leq \frac{1}{2}$. Assume K satisfies conditions* (K1)–(K3) *and W satisfies* (W1)–(W3). *There exists a constant $C = C(a, h_W, r_K, H_K) > 0$ such that for all $\Lambda > 0$ and all configurations $u \in \mathcal{M}$ for which $\overline{E}(u) < \frac{a}{64}\left(\frac{h_W}{1+h_W}\right)H_K$, and in the H^{-1} case also $\overline{E}(u) < \frac{1}{2^{N+2}}\frac{ah_W}{20}H_K$, the following holds:*

$$\overline{E}(u)\overline{L}(u) \geq C. \tag{22}$$

The constant $C = c(N)r_K\left(\frac{h_W}{1+h_W}\right)a^{3/2}H_K$.

Proof. Step 1: Reduction. Let κ_{r_K} be as defined by (11) and (12). We use the notation $\kappa_r := \kappa_{r_K}$. To make the distinction between energies, let $\overline{E}_K$ and $\overline{E}_{\kappa_r}$ be the energy densities corresponding to kernels K and κ_r, respectively. Note that $\overline{E}_K \geq H_K\overline{E}_{\kappa_r}$. So it is enough to show the above claim for κ_r, with $r_K = r$ and $H_K = 1$. Therefore from here on we consider only $K = \kappa_r$.

Step 2: u is separated into phases. We show that any low-energy-density configuration, u, has a significant portion of the mass on the set where values of u are close to 1. More precisely:

Claim. Let $A := \{x : u(x) \geq \frac{7}{8}\}$ and $\tilde{u} := \chi_A$. For future reference let $\underline{A} := \{x : u(x) < \frac{a}{4}\}$ and let I be the interfacial region, $I := \Omega\backslash(A \cup \underline{A})$. Assume $\overline{E}(u) < \frac{3}{32}ah_W$. Then

$$\|A\| = \frac{|A|}{|\Omega|} = \fint_\Omega \tilde{u}(x)dx > \frac{1}{2}\fint_\Omega u(x)dx = \frac{a}{2} \quad \text{and} \quad \|A\| \leq 2a. \tag{23}$$

Proof. Due to assumption (W3)

$$\overline{E}(u) \geq h_W\left(\left\|\left\{\frac{a}{4} < u \leq \frac{7}{8}\right\}\right\| + \left\|\left\{\frac{9}{8} < u\right\}\right\|\right).$$

Consequently

$$\begin{aligned}
a = \fint_\Omega u dx &= \fint_{\{u \leq \frac{a}{4}\}} u dx + \fint_{\{\frac{a}{4} < u \leq \frac{7}{8}\}} u dx + \fint_{\{\frac{7}{8} < u \leq \frac{9}{8}\}} u dx + \fint_{\{\frac{9}{8} < u\}} u dx \\
&\leq \frac{a}{4} + \frac{7}{8}\left\|\left\{\frac{a}{4} < u \leq \frac{7}{8}\right\}\right\| + \frac{9}{8}\fint_\Omega \tilde{u}dx + \left\|\left\{\frac{9}{8} < u\right\}\right\| + \frac{1}{h_W}\fint_{\{\frac{9}{8} < u\}} W(u)dx \\
&\leq \frac{a}{4} + \frac{\overline{E}}{h_W} + \frac{9}{8}\fint_\Omega \tilde{u}dx + \frac{\overline{E}}{h_W} \\
&< \frac{a}{4} + \frac{3a}{16} + \frac{9}{8}\fint_\Omega \tilde{u}dx.
\end{aligned}$$

Therefore $\fint_\Omega \tilde{u}dx > \frac{a}{2}$.

To prove the second claim, note that

$$\begin{aligned} a = \fint_\Omega u(x)dx &\geq \frac{7}{8}\|A\| + \fint_{\{-\frac{a}{4}\leq a<0\}} u(x)dx + \fint_{\{u<-\frac{a}{4}\}} u(x)dx \\ &\geq \frac{7}{8}\|A\| - \frac{a}{4} - \frac{1}{h_W}\fint_\Omega W(u(x))dx \\ &\geq \frac{7}{8}\|A\| - \frac{a}{4} - \frac{\overline{E}}{h_W} \\ &\geq \frac{7}{8}\|A\| - \frac{a}{2}. \qquad \square \end{aligned}$$

Step 3*: Energy bounds a measure of interfacial area.*
Claim.

$$\fint_\Omega |\tilde{u} - \kappa_r * \tilde{u}|dx \leq \left(\frac{16}{9} + \frac{2}{h_W}\right)\overline{E}.$$

Heuristically, when r is small the expression on the left-hand side measures r times the area of the boundary of $\{\tilde{u} = 1\}$. More precisely the area is measured in such a way that features of size less than r are smoothed out, and thus neglected, to some extent.

Proof. We use the following notation for the sum of sets $X + Y := \{x + y \mid x \in X,\ y \in Y\}$. Using the fact that $\tilde{u}$ takes only values 0 and 1, we obtain

$$\begin{aligned} &\fint_\Omega \left|\tilde{u}(x) - \int_{\mathbb{R}^N} \kappa_r(x-y)\tilde{u}(y)dy\right|dx \\ &\qquad = \fint_\Omega \int_{\mathbb{R}^N} |\tilde{u}(x) - \tilde{u}(y)|\kappa_r(x-y)dydx \\ &\qquad \leq \frac{1}{|\Omega|}\left[\int_A \int_{\underline{A}+\Lambda\mathbb{Z}^N} + \int_{\underline{A}} \int_{A+\Lambda\mathbb{Z}^N} |\tilde{u}(x) - \tilde{u}(y)|^2\kappa_r(x-y)dydx \right. \\ &\qquad\qquad \left. + \int_I \int_{\mathbb{R}^N} + \int_\Omega \int_{I+\Lambda\mathbb{Z}^N} \kappa_r(x-y)dydx\right] \\ &\qquad \leq \left(\frac{3}{4}\right)^{-2} \fint_\Omega \int_{\mathbb{R}^N} |u(x) - u(y)|^2\kappa_r(x-y)dydx + 2\left\|\left\{\frac{a}{4} \leq u \leq \frac{7}{8}\right\}\right\| \\ &\qquad \leq \left(\frac{16}{9} + \frac{2}{h_W}\right)\overline{E}. \qquad \square \end{aligned}$$

Step 4. *Claim.* $\phi : (0,\infty) \to [0,\infty)$ defined by

$$\phi(s) := \fint_\Omega |\tilde{u} - \tilde{u} * \kappa_s|dx$$

is subadditive.

One should note that some other possible measures of surface area (for example, the volume of appropriate tubular neighborhood) do not have this property in general and may have a superlinear growth (for appropriate range of r).

Proof. Let $s = p + q$ for some $p, q > 0$. As in Step 3 we have

$$\phi(s) = \fint_\Omega \int_{\mathbb{R}^N} |\tilde{u}(x) - \tilde{u}(y)| \kappa_s(y) dy dx.$$

Now let $z = x - \frac{p}{s} y$. Using periodicity and the scaling properties of kernel κ_s one finds

$$\begin{aligned}\phi(s) &\le \fint_\Omega \int_{\mathbb{R}^N} (|\tilde{u}(x) - \tilde{u}(z)| + |\tilde{u}(z) - \tilde{u}(y)|) \kappa_s(y) dy dx \\ &= \fint_\Omega \int_{\mathbb{R}^N} \left|\tilde{u}(x) - \tilde{u}\left(x - \frac{p}{s} y\right)\right| \kappa_s(y) dy dx + \fint_\Omega \int_{\mathbb{R}^N} \left|\tilde{u}(z) - \tilde{u}\left(z - \frac{q}{s} y\right)\right| \kappa_s(y) dy dz;\end{aligned}$$

substitute $\tilde{y} = \frac{p}{s}\, y$ in the first integral and $\tilde{y} = \frac{q}{s}\, y$ and $x = z$ in the second to obtain

$$\begin{aligned}&= \fint_\Omega \int_{\mathbb{R}^N} |\tilde{u}(x) - \tilde{u}(\tilde{y})| \kappa_p(\tilde{y}) d\tilde{y} dx + \fint_\Omega \int_{\mathbb{R}^N} |\tilde{u}(x) - \tilde{u}(\tilde{y})| \kappa_q(\tilde{y}) d\tilde{y} dx \\ &= \phi(p) + \phi(q). \qquad \square\end{aligned}$$

Step 5. $\kappa_l * \tilde{u} \sim \tilde{u}$ for some l of size $\frac{1}{\overline{E}}$. More precisely:

Claim. Let $\mu > 2$ be a constant, which we specify later. If $\overline{E} < \frac{a}{\mu}\left(\frac{h_W}{1+h_W}\right)$, then for

$$l = \left\lfloor \frac{a}{\mu} \left(\frac{h_W}{1 + h_W}\right) \frac{1}{\overline{E}} \right\rfloor r \;=: i\, r \tag{24}$$

the following holds:

$$\phi(l) = \fint_\Omega |\tilde{u} - \kappa_l * \tilde{u}|\, dx < \frac{2}{\mu}\, a. \tag{25}$$

Proof. The assumption on $\overline{E}$ implies that $l > \frac{1}{2} \frac{a}{\mu} \left(\frac{h_W}{1+h_W}\right) \frac{1}{\overline{E}}\, r > 0$. Subadditivity of ϕ established in Step 4 and the bound of Step 3 imply

$$\fint_\Omega |\tilde{u} - \kappa_l * \tilde{u}|\, dx \le i \fint_\Omega |\tilde{u} - \kappa_r * \tilde{u}|\, dx \le \left\lfloor \frac{a}{\mu} \left(\frac{h_W}{1 + h_W}\right) \frac{1}{\overline{E}} \right\rfloor 2 \left(1 + \frac{1}{h_w}\right) \overline{E} \le \frac{2}{\mu}\, a. \qquad \square$$

Step 6. If for some $l > 1$,

$$\kappa_l * \tilde{u} \sim \tilde{u}, \quad \text{then} \quad \overline{L} \gtrsim l.$$

More precisely:

Claim. Set $\mu = 64$. There exists a constant c, depending only on dimension N and on a, such that for any $l > 1$,

$$\text{if} \quad \fint_\Omega |\tilde{u} - \kappa_l * \tilde{u}|\, dx < \frac{2}{\mu}\, a, \quad \text{then} \quad \overline{L} > cl. \tag{26}$$

We split the proof of this claim into four parts. First we establish two auxiliary claims. Then we prove claim (26) for the Wasserstein metric case and for the H^{-1} metric case separately.

Step 6a. Let $A_l := \{x \in \Omega \,:\, \tilde{u} * \kappa_l > \frac{7}{8}\}$. If

$$\fint_\Omega |\tilde{u} - \kappa_l * \tilde{u}|\, dx < \frac{2}{\mu}\, a,$$

then

$$\|A_l\| > \frac{\mu - 32}{2\mu}\, a. \tag{27}$$

By the assumption

$$\frac{2}{\mu}\, a > \fint_\Omega |\tilde{u} - \kappa_l * \tilde{u}|\, dx \geq \fint_{A\setminus A_l} \frac{1}{8}\, dx = \frac{1}{8}\|A\setminus A_l\|.$$

From (23) we have $\|A\| > \frac{a}{2}$. Combining the two inequalities gives

$$\|A_l\| \geq \|A\| - \|A\setminus A_l\| > \left(\frac{1}{2} - \frac{16}{\mu}\right)\, a.$$

Remark. From this point on the proof does not require the closeness of $\tilde{u}$ and $\kappa_l * \tilde{u}$ explicitly, but rather it uses only the fact that A_l is large, as described by (27). That is, we require only that after $\tilde{u}$ is averaged over radius l it still have well-developed interfaces.

Step 6b. A significant subset of A_l can be well approximated by balls of radius l. More precisely:

Claim. Set $\mu = 64$. There exists a finite subset, J, of A_l such that for $A_{ball} = \cup_{x\in J} B(x, l)$,

$$\frac{8}{7}\|A_{ball} \cap A\| \geq \|A_{ball}\| > \frac{1}{2^{N+2}}\, a. \tag{28}$$

This claim has its roots in [5]; see also [17]. Let J be a maximal family of points in A_l such that balls in $\{B(x, l)\}_{x\in J}$ are disjoint. Then $A_l \subset \cup_{x\in J} B(x, 2l)$, by definition. Therefore, using (27)

$$\|A_{ball}\| \geq \frac{1}{2^N}\|A_l\| > \frac{1}{2^{N+2}} a.$$

Since $J \subset A_l$, for all $x \in J$ we have that

$$\frac{7}{8} \leq k_l * \tilde{u}(x) \leq \frac{|B(x, l) \cap A|}{|B(x, l)|}.$$

Summing over $x \in J$ gives the first inequality.

Step 6c (Wasserstein). Let $\gamma = \left(\frac{9}{8}\right)^{1/N} - 1$. For set U and $\lambda \geq 0$, let

$$U^\lambda := \{x \in \Omega \,:\, \operatorname{dist}(x, U) \leq \lambda\}.$$

Let $\lambda = \gamma l$. Using Lemma 8 and the fact that $u \geq 0$,

$$\begin{aligned}\overline{L}^2 = \frac{d_{Wass}(u,a)^2}{|\Omega|} &\geq \lambda^2 \left(\fint_{A_{ball}} u(x)dx - a\|A^\lambda_{ball}\| \right) \\ &\geq \lambda^2 \left(\frac{7}{8}\|A \cap A_{ball}\| - a\left(1+\frac{\lambda}{l}\right)^N \|A_{ball}\| \right) \\ &\geq \gamma^2 l^2 \left(\frac{49}{64} - \frac{1}{2}(1+\gamma)^N \right) \|A_{ball}\| \\ &\geq \gamma^2 \frac{1}{5}\frac{1}{2^{N+2}} a\, l^2.\end{aligned}$$

Combining the conclusions of Steps 5 and 6 now proves the interpolation inequality (22). In particular

$$\overline{L} \geq c(N) r \left(\frac{h_W}{1+h_W} \right) a^{3/2} \frac{1}{\overline{E}}.$$

Step 6d (H^{-1}). Assume $\overline{E} < \frac{1}{20}\frac{1}{2^{N+2}} a h_w$. To obtain a lower bound on $\overline{L}$, given by (20), we first build a local test function. For $\gamma > 1$ to be determined, let $\eta : [0,\infty) \to [0,1]$ be defined by

$$\eta(z) := \begin{cases} 1 & \text{if } z \in [0,l], \\ l - \frac{z-l}{\gamma l - l} & \text{if } l < z < \gamma l, \\ 0 & \text{if } \gamma l \leq z. \end{cases}$$

Let $\bar{\xi}(x) := \eta(|x|)$. This is the local test function. Assume, for the moment, that $0 \in A_l$. Let $\hat{u}(x) := \max\left\{u(x), -\frac{a}{4}\right\}$. Then

$$\int_{B(0,\gamma l)} |\nabla \bar{\xi}|^2 dx = (\gamma^N - 1)|B(0,l)| \frac{1}{(\gamma-1)^2 l^2}. \tag{29}$$

Also

$$\begin{aligned}&\int_{B(0,\gamma l)} (\hat{u} - a)\bar{\xi} dx \\ &\qquad \geq \frac{7}{8}|A \cap B(0,l)| - \frac{a}{4}\left(|B(0,l)\backslash A| + (\gamma^N - 1)|B(0,l)| \right) - a\gamma^N |B(0,l)| \\ &\qquad \geq \left(\left(\frac{7}{8}\right)^2 - \frac{a}{32} - \frac{a(\gamma^N - 1)}{4} - a\gamma^N \right) |B(0,l)|.\end{aligned}$$

Let us now set $\gamma = \left(\frac{8}{7}\right)^{1/N}$. Then

$$\int_{B(0,\gamma l)} (\hat{u} - a)\bar{\xi} dx \geq \frac{3}{20}|B(0,l)|. \tag{30}$$

Now let us construct the (global) test function on Ω. Let

$$\xi(x) = \sup_{y \in J} \bar{\xi}(x - y).$$

Using (29), (23), and (28) we obtain

$$\fint_\Omega |\nabla \xi|^2 dx \le \frac{\gamma^N - 1}{(\gamma-1)^2 l^2} \|A_{ball}\| \le \frac{a}{(\gamma-1)^2 l^2}. \tag{31}$$

Using the fact that balls of radius l centered at points in J are disjoint we obtain

$$\begin{aligned}
\fint_\Omega (u-a)\xi dx &\ge \fint_\Omega (\hat{u}-a)\xi dx + \fint_{\{u<-\frac{a}{4}\}} u dx \\
&\ge \frac{3}{20}\|A_{ball}\| - \frac{1}{h_W}\fint_\Omega W(u) dx \\
&\ge \frac{3}{20}\,\frac{1}{2^{N+2}} a - \frac{\overline{E}}{h_w} \\
&\ge \frac{1}{10}\,\frac{1}{2^{N+2}} a.
\end{aligned}$$

Here we used the assumption $\overline{E} < \frac{1}{20}\,\frac{1}{2^{N+2}} a h_w$. Therefore

$$\overline{L} \ge \frac{\fint_\Omega (u-a)\xi dx}{\sqrt{\fint_\Omega |\nabla \xi|^2 dx}} \ge \tilde{c}(N)\sqrt{a} l.$$

The definition of l now implies

$$\overline{L} \ge c(N) r \left(\frac{h_W}{1+h_W}\right) a^{3/2} \frac{1}{\overline{\overline{E}}},$$

which proves the interpolation inequality. ∎

Remark 2. In the case $N = 2$, one can obtain a sharper result with respect to scaling in a (as $a \to 0^+$) by considering more carefully constructed test functions. This was done for the Mullins–Sekerka evolution by Conti, Niethammer, and Otto in [5]. In particular if γ is taken of size $a^{-1/2}$, and on $[l, cl]$, we replace the linear η by the optimal one, $\eta(z) = (\ln \gamma l - \ln z)/\ln \gamma$. By using such a test function one obtains that

$$\overline{L} \ge c(N) r \left(\frac{h_W}{1+h_W}\right) a^{3/2} |\ln a|^{1/2} \frac{1}{\overline{\overline{E}}}.$$

4.1. Interpolation inequalities for the local energy. Let us now consider the case of the local energy density:

$$\overline{E}(u) := \fint_\Omega \frac{1}{2}|\nabla u(x)|^2 + W(u(x)) dx. \tag{32}$$

The method developed in the proof of Theorem 6 applies to the local energy with minor modifications.

COROLLARY 7 (interpolation inequality). *Let $0 < a \le \frac{1}{2}$. Assume W satisfies* (W1)–(W3). *There exists a constant $C = C(a, h_W) > 0$ such that for all $\Lambda > 0$ and all configurations $u \in \mathcal{M}$ for which $\overline{E}(u) < \frac{a}{64}\left(\frac{h_W}{1+h_W}\right)$, and in the H^{-1} case also $\overline{E}(u) < \frac{1}{2^{N+2}}\frac{a h_W}{20}$, the following holds:*

$$\overline{E}(u)\overline{L}(u) \ge C. \tag{33}$$

The constant $C = c(N)\left(\frac{h_W}{1+h_W}\right) a^{3/2}$.

Proof. Note that Step 1 is not needed, while the estimate of Step 2 used only the W term, which is the same for both the local and the nonlocal energy. The main fact we need to check is the statement of Step 3, which we prove below for $r^2 = \frac{1}{2}$. Steps 4, 5, and 6 do not require any modifications.

To prove that

$$⨍_\Omega |\tilde{u} - \kappa_r * \tilde{u}| dx \le \left(\frac{16}{9} + \frac{2}{h_W}\right) \overline{E}$$

with $r^2 = \frac{1}{2}$ we begin as in Step 3 of the proof of Theorem 6:

$$\begin{aligned} ⨍_\Omega \left|\tilde{u}(x) - \int_{\mathbb{R}^N} \kappa_r(x-y)\tilde{u}(y)dy\right| dx \\ &\le \left(\frac{3}{4}\right)^{-2} ⨍_\Omega \int_{\mathbb{R}^N} |u(x)-u(y)|^2 \kappa_r(x-y)dydx + 2\left\|\left\{\frac{a}{4} \le u \le \frac{7}{8}\right\}\right\| \\ &\le \frac{16}{9} ⨍_\Omega \frac{1}{|B(x,r)|} \int_{B(x,r)} |u(x)-u(y)|^2 dydx + \frac{2}{h_W}\overline{E}. \end{aligned}$$

It remains to further estimate the first term:

$$\begin{aligned} ⨍_\Omega \frac{1}{|B(x,r)|} \int_{B(x,r)} |u(x)-u(y)|^2 dydx \\ &\le ⨍_\Omega \frac{1}{|B(0,r)|} \int_{B(x,r)} |x-y|^2 \left|\int_0^1 \nabla u(x - s(x-y))ds\right|^2 dydx \\ &\le \frac{1}{2} ⨍_\Omega \frac{1}{|B(0,r)|} \int_{B(0,r)} \int_0^1 |\nabla u(x-sz)|^2 dsdzdx \\ &\le \frac{1}{2} \frac{1}{|B(0,r)|} \int_{B(0,r)} \int_0^1 2\overline{E} dsdz = \overline{E}. \qquad \square \end{aligned}$$

5. Appendix.

5.1. A property of Wasserstein distance. The following lemma is analogous to Lemma 5 in [17]. We state it in large generality, the reason being that we want to consider Ω with metric from the torus $R^N/(\Lambda\mathbb{Z})^N$. In applications σ is the Lebesgue measure.

LEMMA 8. *Let (Ω, d, σ) be a metric space endowed with finite measure σ. Let $u \in L^1(\Omega)$ be a nonnegative function with average $a := ⨍_\Omega u(x)d\sigma(x)$. Let $A \subset \Omega$ measurable, and let $A^l := \{x \in \Omega : d(x,A) \le l\}$. Then*

$$d^2_{Wass}(u,a) \ge l^2 \left(\int_A u(x)d\sigma(x) - a\sigma(A^l)\right).$$

Proof. We use the definition of Wasserstein distance. Let π be an admissible transportation plan, that is, a measure on $\Omega \times \Omega$ with marginals $u(x)d\sigma(x)$ and $a\sigma(y)$.

Then

$$\begin{aligned}\int_{\Omega\times\Omega} |x-y|^2 d\pi(x,y) &= \int_{A\times(\Omega\setminus A^l)} |x-y|^2 d\pi(x,y)\\ &\geq l^2\,\pi(A\times(\Omega\setminus A^l))\\ &\geq l^2(\pi(A\times\Omega)-\pi(\Omega\times A^l))\\ &= l^2\left(\int_A u(x)d\sigma(x) - \int_{A^l} a d\sigma(y)\right)\\ &= l^2\left(\int_A u(x)d\sigma(x) - a\sigma(A^l)\right). \qquad \square\end{aligned}$$

Acknowledgments. The author would like to thank Andrea Bertozzi and Chad Topaz for helpful discussions, as well as the referees for carefully reading the manuscript and providing valuable suggestions. The author would also like to thank the Center for Nonlinear Analysis for its support during the preparation of this paper.

REFERENCES

[1] G. Alberti and G. Bellettini, *A non-local anisotropic model for phase transitions: Asymptotic behaviour of rescaled energies*, European J. Appl. Math., 9 (1998), pp. 261–284.

[2] G. Alberti, G. Bellettini, M. Cassandro, and E. Presutti, *Surface tension in Ising systems with Kac potentials*, J. Statist. Phys., 82 (1996), pp. 743–796.

[3] J.-D. Benamou and Y. Brenier, *A numerical method for the optimal time-continuous mass transport problem and related problems*, in Monge Ampère Equation: Applications to Geometry and Optimization (Deerfield Beach, FL, 1997), Contemp. Math. 226, AMS, Providence, RI, 1999, pp. 1–11.

[4] A. Bertozzi and D. Slepčev, *On nonlocal equations modeling biological aggregation*, in preparation (2008).

[5] S. Conti, B. Niethammer, and F. Otto, *Coarsening rates in off-critical mixtures*, SIAM J. Math. Anal., 37 (2006), pp. 1732–1741.

[6] S. Dai and R. L. Pego, *Universal bounds on coarsening rates for mean-field models of phase transitions*, SIAM J. Math. Anal., 37 (2005), pp. 347–371.

[7] S. Dai and R. L. Pego, *An upper bound on the coarsening rate for mushy zones in a phase-field model*, Interfaces Free Bound., 7 (2005), pp. 187–197.

[8] G. Giacomin and J. L. Lebowitz, *Phase segregation dynamics in particle systems with long range interactions*. I. *Macroscopic limits*, J. Statist. Phys., 87 (1997), pp. 37–61.

[9] G. Giacomin and J. L. Lebowitz, *Phase segregation dynamics in particle systems with long range interactions*. II. *Interface motion*, SIAM J. Appl. Math., 58 (1998), pp. 1707–1729.

[10] R. V. Kohn and F. Otto, *Upper bounds on coarsening rates*, Comm. Math. Phys., 229 (2002), pp. 375–395.

[11] R. V. Kohn and X. Yan, *Upper bound on the coarsening rate for an epitaxial growth model*, Comm. Pure Appl. Math., 56 (2003), pp. 1549–1564.

[12] R. V. Kohn and X. Yan, *Coarsening rates for models of multicomponent phase separation*, Interfaces Free Bound., 6 (2004), pp. 135–149.

[13] B. Li and J.-G. Liu, *Epitaxial growth without slope selection: Energetics, coarsening, and dynamic scaling*, J. Nonlinear Sci., 14 (2004), pp. 429–451.

[14] L. Modica, *The gradient theory of phase transitions and the minimal interface criterion*, Arch. Rational Mech. Anal., 98 (1987), pp. 123–142.

[15] L. Modica and S. Mortola, *Un esempio di* Γ^-*-convergenza*, Boll. Un. Mat. Ital. B (5), 14 (1977), pp. 285–299.

[16] F. Otto, *The geometry of dissipative evolution equations: The porous medium equation*, Comm. Partial Differential Equations, 26 (2001), pp. 101–174.

[17] F. Otto, T. Rump, and D. Slepčev, *Coarsening rates for a droplet model: Rigorous upper bounds*, SIAM J. Math. Anal., 38 (2006), pp. 503–529.

[18] C. M. Topaz, A. L. Bertozzi, and M. A. Lewis, *A nonlocal continuum model for biological aggregation*, Bull. Math. Biol., 68 (2006), pp. 1601–1623.

[19] C. Villani, *Topics in Optimal Transportation*, Grad. Stud. Math. 58, AMS, Providence, RI, 2003.

SIAM J. MATH. ANAL.
Vol. 40, No. 3, pp. 1049–1057

UNIQUENESS OF POSITIVE BOUND STATES TO SCHRÖDINGER SYSTEMS WITH CRITICAL EXPONENTS*

CONGMING LI† AND LI MA‡

Abstract. We prove the uniqueness of the positive solutions of the following elliptic system: (1) $-\Delta(u(x)) = u(x)^{\alpha}v(x)^{\beta}$, (2) $-\Delta(v(x)) = u(x)^{\beta}v(x)^{\alpha}$. Here $x \in R^n$, $n \geq 3$, and $1 \leq \alpha < \beta \leq \frac{n+2}{n-2}$ with $\alpha + \beta = \frac{n+2}{n-2}$. In the special case when $n = 3$ and $\alpha = 2$, $\beta = 3$, the system is closely related to the ones from the stationary Schrödinger system with critical exponents for the Bose–Einstein condensate. As the first step, we prove the radial symmetry of the positive solutions to the elliptic system above with critical exponents. We then prove that $u = v$, which is a key point for our uniqueness result.

Key words. moving plane, positive solutions, radial symmetric, uniqueness

AMS subject classifications. 35J45, 35J60, 45G05, 45G15

DOI. 10.1137/080712301

1. Introduction.

In this paper, we consider the uniqueness of positive solutions to the following stationary Schrödinger system:

$$\begin{cases} -\Delta(u(x)) = u(x)^{\alpha}v(x)^{\beta}, \\ -\Delta(v(x)) = u(x)^{\beta}v(x)^{\alpha}. \end{cases} \tag{1}$$

Here, in the special case when $n = 3$ and $\alpha = 2, \beta = 3$, and $u = v$, system (1) is reduced to the quintic Schrödinger equation considered by Bourgain [1]. We prove in this paper that there is a uniqueness result for the system above. In general, our system is related to the ones from the stationary Schrödinger system with critical exponents for the Bose–Einstein condensate (see [15], [19], [20], and [23]). In the earlier works [19], [20], and [23], more attention was paid to the elliptic system (1) with subcritical exponents. Very interestingly, Chen and Li proved that the best constant in the weighted Hardy–Littlewood–Sobolev inequality can be achieved by explicit radial symmetric functions (see [5] and [18]). As a consequence of their work, the uniqueness of positive solutions to the corresponding elliptic system (it is (1) in the case when $\alpha = 0$ and $\beta = \frac{n+2}{n-2}$) has been settled. However, when $0 < \alpha, \beta \leq \frac{n+2}{n-2}$, the uniqueness of smooth positive solutions to the stationary Schrödinger system (1) is an open question. Generally speaking, there are very few results even for the uniqueness of positive solutions to ordinary differential systems. The aim of this paper is to prove the radial symmetry and uniqueness of positive solutions to (1) with critical exponents and $1 \leq \alpha < \beta \leq \frac{n+2}{n-2}$.

*Received by the editors January 2, 2008; accepted for publication (in revised form) June 17, 2008; published electronically October 13, 2008.
http://www.siam.org/journals/sima/40-3/71230.html

†Department of Applied Mathematics, University of Colorado at Boulder, Campus Box 526, Boulder, CO 80309 (cli@colorado.edu). The work of this author was partially supported by NSF grant DMS-0401174.

‡Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China (lma@math.tsinghua.edu.cn). The work of this author was partially supported by the National Natural Science Foundation of China through grant 10631020, a key grant KZ200710025012 of the Education Committee of Beijing, and SRFDP grant 20060003002.

As one can expect, just like in the work of Weinstein [28] in the scalar case with subcritical exponent, there is a close relationship between the stationary Schrödinger system with critical exponent and the Hardy–Littlewood–Sobolev inequality. As we show below, this is true.

Since we shall use Hardy–Littlewood–Sobolev inequality to prove radial symmetry of our solutions, let's first explore the recent progress of Lieb's conjecture. We begin by recalling the well-known Hardy–Littlewood–Sobolev inequality. Let $0 < \lambda < n$, $1 < s, r < \infty$, and $\|f\|_p$ be the $L^p(R^n)$ norm of the function f. We shall write by $\|f\|_{p,\Omega}$ the L^p norm of the function f on the domain Ω. Then the classical Hardy–Littlewood–Sobolev inequality states that

$$\int_{R^n}\int_{R^n} \frac{f(x)g(y)}{|x-y|^\lambda}dxdy \le C_{s,\lambda,n}\|f\|_r\|g\|_s \tag{2}$$

for any $f \in L^r(R^n)$, $g \in L^s(R^n)$, and for $\frac{1}{r}+\frac{1}{s}+\frac{\lambda}{n}=2$. Hardy and Littlewood also introduced the double weighted inequality, which was later generalized by Stein and Weiss in [26] in the following form:

$$\left|\int_{R^n}\int_{R^n} \frac{f(x)g(y)}{|x|^{\alpha_0}|x-y|^\lambda|y|^{\beta_0}}dxdy\right| \le C_{\alpha_0,\beta_0,s,\lambda,n}\|f\|_r\|g\|_s, \tag{3}$$

where $\alpha_0+\beta_0 \ge 0$,

$$1-\frac{1}{r}-\frac{\lambda}{n} < \frac{\alpha_0}{n} < 1-\frac{1}{r} \quad \text{and} \quad \frac{1}{r}+\frac{1}{s}+\frac{\lambda+\alpha_0+\beta_0}{n}=2. \tag{4}$$

The best constant in the weighted inequality (3) can be obtained by maximizing the functional

$$J(f,g)=\int_{R^n}\int_{R^n} \frac{f(x)g(y)}{|x|^{\alpha_0}|x-y|^\lambda|y|^{\beta_0}}dxdy \tag{5}$$

under the constraints $\|f\|_r = \|g\|_s = 1$. Then the corresponding Euler–Lagrange equations are the system of integral equations

$$\begin{cases} \lambda_1 r f(x)^{r-1} = \frac{1}{|x|^{\alpha_0}}\int_{R^n}\frac{g(y)}{|y|^{\beta_0}|x-y|^\lambda}dy, \\ \lambda_2 s g(x)^{s-1} = \frac{1}{|x|^{\beta_0}}\int_{R^n}\frac{f(y)}{|y|^{\alpha_0}|x-y|^\lambda}dy, \end{cases} \tag{6}$$

where $f, g \ge 0$, $x \in R^n$, and $\lambda_1 r = \lambda_2 s = J(f,g)$.

Let $u = c_1 f^{r-1}$, $v = c_2 g^{s-1}$, $p = \frac{1}{r-1}$, $q = \frac{1}{s-1}$, $pq \ne 1$, and for a proper choice of constants c_1 and c_2, system (6) becomes

$$\begin{cases} u(x) = \frac{1}{|x|^{\alpha_0}}\int_{R^n}\frac{v(y)^q}{|y|^{\beta_0}|x-y|^\lambda}dy, \\ v(x) = \frac{1}{|x|^{\beta_0}}\int_{R^n}\frac{u(y)^p}{|y|^{\alpha_0}|x-y|^\lambda}dy, \end{cases} \tag{7}$$

where $u, v \ge 0$, $0 < p, q < \infty$, $0 < \lambda < n$, $\frac{\alpha_0}{n} < \frac{1}{p+1} < \frac{\lambda+\alpha_0}{n}$, and $\frac{1}{p+1}+\frac{1}{q+1} = \frac{\lambda+\alpha_0+\beta_0}{n}$.

Note that in the special case where $\alpha_0 = 0$ and $\beta_0 = 0$, system (7) reduces to the following system:

$$\begin{cases} u(x) = \int_{R^n} \frac{v^q(y)}{|x-y|^\lambda} dy, \\ v(x) = \int_{R^n} \frac{u^p(y)}{|x-y|^\lambda} dy \end{cases} \tag{8}$$

with

$$\frac{1}{q+1} + \frac{1}{p+1} = \frac{\lambda}{n}. \tag{9}$$

It is well known that this integral system is closely related to the system of partial differential equations

$$\begin{cases} (-\Delta)^{\gamma/2} u = v^q, \quad u > 0, \text{ in } R^n, \\ (-\Delta)^{\gamma/2} v = u^p, \quad v > 0, \text{ in } R^n, \end{cases} \tag{10}$$

where $\gamma = n - \lambda$.

When $p = q = \frac{n+\gamma}{n-\gamma}$ and $u(x) = v(x)$, system (8) becomes the single equation

$$u(x) = \int_{R^n} \frac{u(y)^{\frac{n+\gamma}{n-\gamma}}}{|x-y|^{n-\gamma}} dy, \quad u > 0, \text{ in } R^n. \tag{11}$$

The corresponding PDE is the well-known family of semilinear equations

$$(-\Delta)^{\gamma/2} u = u^{(n+\gamma)/(n-\gamma)}, \quad u > 0, \quad \text{in } R^n. \tag{12}$$

In particular, when $n \geq 3$, and $\gamma = 2$, (12) becomes

$$-\Delta u = u^{(n+2)/(n-2)}, \quad u > 0, \text{ in } R^n. \tag{13}$$

The classification of the solutions of (13) has provided an important ingredient in the study of the well-known Yamabe problem and the prescribing scalar curvature problem. Equation (13) was studied by Gidas, Ni, and Nirenberg [11], Caffarelli, Gidas, and Spruck [2], Chen and Li [3], and Li [16]. They classified all the positive solutions. Recently, Wei and Xu [27] generalized this result to the solutions of the more general equation (12) with γ being any even number between 0 and n. Some results of Chen, Li, and Ou have been improved by Hang in [12]. One may see related results in [14] and [17].

Although the systems for other real values of α, β between 0 and n are of interest to some, we shall concentrate in this paper only on the system (1) with critical exponents when $1 \leq \alpha, \beta \leq \frac{n+2}{n-2}$ and $\alpha + \beta = \frac{n+2}{n-2}$.

Our main results are the following two theorems.

THEOREM 1. *Assume that $n \geq 3$, $1 \leq \alpha, \beta \leq \frac{n+2}{n-2}$, and $\alpha + \beta = \frac{n+2}{n-2}$. Then any $L^{\frac{2n}{n-2}}(\mathbf{R}^n) \times L^{\frac{2n}{n-2}}(\mathbf{R}^n)$ positive solution pair (u, v) to system (1) is radial symmetric.*

THEOREM 2. *Assume that $n \geq 3$, $1 \leq \alpha \leq \beta \leq \frac{n+2}{n-2}$, and $\alpha + \beta = \frac{n+2}{n-2}$. Then any $L^{\frac{2n}{n-2}}(\mathbf{R}^n) \times L^{\frac{2n}{n-2}}(\mathbf{R}^n)$ radial symmetric solution pair (u, v) to system (1) is unique and $u = v$.*

In the proof of Theorem 1, we shall use the equivalent integral form for the nonnegative solutions

$$(u, v) \in L^{\frac{2n}{n-2}}(\mathbf{R}^n) \times L^{\frac{2n}{n-2}}(\mathbf{R}^n)$$

to system (1):

$$u(x) = \int_{R^n} \frac{u^\alpha v^\beta(y)}{|x-y|^{n-2}} dy$$

and

$$v(x) = \int_{R^n} \frac{v^\alpha u^\beta(y)}{|x-y|^{n-2}} dy.$$

This equivalent form can be proved in the same way as in section 4 in the work of Chen, Li, and Ou [8].

We point out that when $u = v$, the elliptic system (1) reduces to the elliptic equation (13) with critical exponent. Then $u = v$ is in a special family of functions:

$$\phi_{x_o,t}(x) = c\left(\frac{t}{t^2 + |x - x_o|^2}\right)^{(n-2)/2}, \tag{14}$$

where $t > 0$, $x_o \in \mathbf{R}^n$, with some positive constant c such that each $\phi_{x_o,t}(x)$ solves (13). This special family of functions is important in the study of (1).

Our results are motivated by the previous work [6], where Chen, Li, and Ou considered the more general system (8) and established the symmetry and monotonicity of the solutions. In [4], Chen and Li also obtained a regularity result of the solutions to (8). To establish the symmetry of the solutions to (8), Chen, Li, and Ou [8], [6], [7] introduced a new idea: an integral form of the method of moving planes. It is entirely different from the traditional method used for PDEs. Instead of relying on maximum principles, certain integral norms were estimated. The new method is a very powerful tool in studying qualitative properties of other integral equations and systems. In fact, following Chen, Li, and Ou's work, Jin and Li [13] studied the symmetry of the solutions to the more general system (7).

Using the maximum principle trick, de Figueiredo and Felmer [10] found an interesting Liouville-type result for certain type of elliptic systems. Recently Ma and Chen [21] discussed the Liouville-type theorem for the positive solutions to the elliptic system (10). They also discussed the radial symmetry property of nonnegative solutions to a higher-order differential equation [22]. Naito and Usami [24] studied the existence of nonoscillatory solutions to second-order elliptic systems of Emden–Fowler type. Some nonexistence results for the Emden–Fowler system can be found in [25]. We refer to [13], [9], and [24] for more references about elliptic systems.

We first prove the radial symmetry of the solutions to (1) with critical exponents. It is obvious that the radial symmetry of the solutions reduces (1) to a system of ODEs, which has the special solution pair $(\phi_{o,t}(x), \phi_{o,t}(x))$. To prove the uniqueness, we prove that $u(0) = v(0)$. Then by the uniqueness of the initial value problem for ODEs, we conclude that $u = v = \phi_{o,t}$. This is the key observation in establishing the uniqueness of positive solutions for (1) with critical exponents.

Theorems 1 and 2 will be proved in the next two sections.

2. Proof of the radial symmetry. We use the moving plane method introduced by Chen, Li, and Ou in [8] to study system (1). We recall the Hardy–Littlewood–Sobolev inequality:

$$|Tf|_p \le C(n,p)|f|_{\frac{np}{n+2p}}, \tag{15}$$

where $C(n,p)$ is a uniform positive constant and

$$Tf(x) = \int_{\mathbf{R}^n} |x-y|^{2-n} f(y)dy.$$

Proof of Theorem 1. For each $\lambda \in \mathbf{R}$, we denote

$$H_\lambda = \{x \in \mathbf{R}^n; x_1 < \lambda\}.$$

For each $x = (x_1, x') \in \mathbf{R}^n$, we let

$$x_\lambda = (2\lambda - x_1, x')$$

be the reflection point of x with respect to the hyperplane ∂H_λ. We let $e_1 = (1, 0, \dots, 0)$.

We define

$$u_\lambda(x) = u(x_\lambda), \quad B_\lambda^u = \{x \in H_\lambda; u_\lambda(x) > u(x)\}$$

and

$$v_\lambda(x) = v(x_\lambda), \quad B_\lambda^v = \{x \in H_\lambda; v_\lambda(x) > v(x)\}.$$

To do the moving plane method, we need the following formula, which is obtained by a change of variables.

$$u(x) = \int_{H_\lambda} \frac{u^\alpha v^\beta(y)}{|x-y|^{n-2}}dy + \int_{H_\lambda} \frac{u_\lambda^\alpha v_\lambda^\beta(y)}{|x_\lambda - y|^{n-2}}dy$$

and

$$u_\lambda(x) = \int_{H_\lambda} \frac{u^\alpha v^\beta(y)}{|x_\lambda-y|^{n-2}}dy + \int_{H_\lambda} \frac{u_\lambda^\alpha v_\lambda^\beta(y)}{|x-y|^{n-2}}dy.$$

Then we have

$$u_\lambda(x) - u(x) = \int_{H_\lambda} (u_\lambda^\alpha v_\lambda^\beta - u^\alpha v^\beta)(y) \left(\frac{1}{|x-y|^{n-2}} - \frac{1}{|x_\lambda - y|^{n-2}} \right) dy. \tag{16}$$

Note that for $x \in H_\lambda$, we have

$$\frac{1}{|x-y|^{n-2}} > \frac{1}{|x_\lambda - y|^{n-2}}.$$

Then for $x \in B_\lambda^u$, using the mean value theorem to

$$u_\lambda^\alpha v_\lambda^\beta - u^\alpha v^\beta = (u_\lambda^\alpha - u^\alpha)v_\lambda^\beta + u^\alpha(v_\lambda^\beta - v^\beta)$$

and after dropping the term of the integration on the region $v_\lambda - v \le 0$, we have

$$\begin{cases} 0 \le u_\lambda(x) - u(x) \\ \quad \le \alpha \int_{B_\lambda^u} \frac{u_\lambda^{\alpha-1} v_\lambda^\beta (u_\lambda - u)}{|x-y|^{n-2}} dy + \beta \int_{B_\lambda^v} \frac{u_\lambda^\alpha v_\lambda^{\beta-1}(v_\lambda - v)}{|x-y|^{n-2}} dy \\ \quad := I + II. \end{cases} \tag{17}$$

Let $p = \frac{2n}{n-2}$. Using the Hardy–Littlewood–Sobolev inequality (15) we can bound the first term I in (17) by

$$\begin{cases} |I|_p \le C(n,p) |u_\lambda^{\alpha-1} v_\lambda^\beta (u_\lambda - u)|_{\frac{2n}{n+2}} \\ \le C(n,p) |u_\lambda|_p^{\alpha-1} |v_\lambda|_p^\beta |u_\lambda - u|_p. \end{cases} \tag{18}$$

Here the integrations are over the set B_λ^u.

Using again the Hardy–Littlewood–Sobolev inequality (15) we can bound the first term II in (17) by

$$\begin{cases} |II|_p \le C(n,p) |u_\lambda^\alpha v_\lambda^{\beta-1} (v_\lambda - v)|_{\frac{2n}{n+2}} \\ \le C(n,p) |u_\lambda|_p^\alpha |v_\lambda|_p^{\beta-1} |v_\lambda - v|_p. \end{cases} \tag{19}$$

Here the integrations are over the domain B_λ^v. Hence, we have

$$\begin{aligned} &|u_\lambda - u|_{p, B_\lambda^u} \\ &\le C(n,p) (|u_\lambda|_{p,B_\lambda^u}^{\alpha-1} |v_\lambda|_{p,B_\lambda^u}^\beta |u_\lambda - u|_{p,B_\lambda^u} + |u_\lambda|_{p,B_\lambda^v}^\alpha |v_\lambda|_{p,B_\lambda^v}^{\beta-1} |v_\lambda - v|_{p,B_\lambda^v}). \end{aligned} \tag{20}$$

Similarly, we have the following formulae for v and v_λ:

$$v(x) = \int_{H_\lambda} \frac{v^\alpha u^\beta(y)}{|x-y|^{n-2}} dy + \int_{H_\lambda} \frac{v_\lambda^\alpha u_\lambda^\beta(y)}{|x_\lambda - y|^{n-2}} dy$$

and

$$v_\lambda(x) = \int_{H_\lambda} \frac{v^\alpha u^\beta(y)}{|x_\lambda - y|^{n-2}} dy + \int_{H_\lambda} \frac{v_\lambda^\alpha u_\lambda^\beta(y)}{|x-y|^{n-2}} dy.$$

Then we have the following estimate:

$$\begin{aligned} &|v_\lambda - v|_{p, B_\lambda^v} \\ &\le C(n,p) (|v_\lambda|_{p,B_\lambda^v}^{\alpha-1} |u_\lambda|_{p,B_\lambda^v}^\beta |v_\lambda - v|_{p,B_\lambda^v} + |v_\lambda|_{p,B_\lambda^u}^\alpha |u_\lambda|_{p,B_\lambda^u}^{\beta-1} |u_\lambda - u|_{p,B_\lambda^u}). \end{aligned} \tag{21}$$

After these preparations, we can use the moving plane method as developed in [8] to prove the radial symmetry of the solutions.

At first, let's start the plane from infinity. Indeed, for $\lambda \gg 1$ large enough, we know that the quantities

$$|v_\lambda|_{p,B_\lambda^u}, |u_\lambda|_{p,B_\lambda^u}, |v_\lambda|_{p,B_\lambda^v}$$

and

$$|u_\lambda|_{p,B_\lambda^v}$$

all are small, which give us that

$$|u_\lambda - u|_{p,B_\lambda^u} \le \frac{1}{4}(|v_\lambda - v|_{p,B_\lambda^v} + |u_\lambda - u|_{p,B_\lambda^u})$$

and

$$|v_\lambda - v|_{p,B_\lambda^v} \le \frac{1}{4}(|v_\lambda - v|_{p,B_\lambda^v} + |u_\lambda - u|_{p,B_\lambda^u}).$$

These imply that $|u_\lambda - u|_{p,B_\lambda^u} = 0$ and $|v_\lambda - v|_{p,B_\lambda^v} = 0$. In other words, $B_\lambda^u = \phi$ and $B_\lambda^v = \phi$.

Next we define

$$\lambda_0 = \inf\{\lambda \in \mathbf{R}; B_{\lambda'}^u = B_{\lambda'}^v = \phi \ \forall\ \lambda' \ge \lambda\}.$$

Then it comes from the fact that $u(x) \to 0$ as $|x| \to \infty$ and $u(x) > 0$ in $\mathbf{R}^n$ that $\lambda_0 < +\infty$. By the definition of λ_0, we have $u_{\lambda_0}(x) \le u(x)$ for $x \in H_{\lambda_0}$. If u or v is not symmetric in the x_1 direction at λ_0, then using the expression (16), we see that $u_{\lambda_0}(x) < u(x)$ and $v_{\lambda_0}(x) < v(x)$ for $x \in H_{\lambda_0}$. This implies that both of $B_{\lambda_0}^u$ and $B_{\lambda_0}^v$ are empty. Then by the absolute continuity of the integral with respect to the domain we see that $|v_\lambda|_{p,B_\lambda^u}, |u_\lambda|_{p,B_\lambda^u}, |v_\lambda|_{p,B_\lambda^v}$, and $|u_\lambda|_{p,B_\lambda^v}$ are small for λ close to λ_0 and enable us to repeat the above argument showing that both of B_λ^u and B_λ^v are empty for λ close to λ_0. This contradicts the definition of λ_0.

3. Proof of the uniqueness. In some sense, the proof of Theorem 2 is just at hand by using the integral expression of the solution pair (u, v).

Proof of Theorem 2. Let $(u, v) \in L^{\frac{2n}{n-2}}(\mathbf{R}^n) \times L^{\frac{2n}{n-2}}(\mathbf{R}^n)$ be a pair of solutions to system (1). By Theorem 1, we know that u and v are radial symmetric about some point x_0, say, $x_0 = 0$. We will consider only the case when $1 \le \alpha < \beta < \frac{n+2}{n-2}$, and the case when $\alpha = \beta$ can be done similarly.

Since $u \in L^{\frac{2n}{n-2}}(\mathbf{R}^n)$ and $v \in L^{\frac{2n}{n-2}}(\mathbf{R}^n)$, using the same method in [8], we have that $u \in C^2(\mathbf{R}^n)$ and $v \in C^2(\mathbf{R}^n)$ with

$$u(x) \to 0, \qquad v(x) \to 0,$$

as $|x| \to \infty$.

Since our solution u is radial symmetric, we can write, in polar coordinates, the first equation in (1) as

$$(r^{n-1}u'(r))' = -r^{n-1}u(r)^\alpha v(r)^\beta,$$

where $r = |x|$.

Integrating both sides of the above equation from 0 to r yields

$$r^{n-1}u'(r) = -\int_0^r s^{n-1}u^\alpha v^\beta(s)ds.$$

It follows by another integration that

$$u(r) = u(0) - \int_0^r \frac{1}{\tau^{n-1}} \int_0^\tau s^{n-1}u^\alpha v^\beta ds d\tau. \tag{22}$$

Similarly, for $v(r)$, we have

$$v(r) = v(0) - \int_0^r \frac{1}{\tau^{n-1}} \int_0^\tau s^{n-1}v^\alpha u^\beta ds d\tau. \tag{23}$$

As we mentioned in the introduction, we need to show only that $u(0) = v(0)$. Otherwise, suppose that

$$u(0) < v(0). \tag{24}$$

Then by continuity, for all small $r > 0$,

$$u(r) < v(r). \tag{25}$$

In other words, there exists an $R > 0$, such that

$$u(r) < v(r) \ \ \forall r \in (0, R). \tag{26}$$

Let R_o be the supreme value of R such that (26) holds. Then $R_o \leq \infty$ and $u(R_o) = v(R_o)$, where we have used the fact that $u(+\infty) = v(+\infty) = 0$. By the definition of R_o and $\alpha < \beta$, we have that

$$u(r)^\alpha v(r)^\beta > v(r)^\alpha u(r)^\beta \ \ \forall r \in (0, R_o). \tag{27}$$

Then we have from (22) and (23) that

$$0 > u(0) - v(0) = \int_0^{R_o} \frac{1}{\tau^{n-1}} \int_0^\tau s^{n-1}(u^\alpha v^\beta - u^\beta v^\alpha)(s) ds d\tau > 0.$$

This is impossible.

Similarly, one can show that $u(0) > v(0)$ is impossible too. Therefore, we must have

$$u(0) = v(0).$$

Finally, by the standard ODE theory, we arrive at

$$u(r) \equiv v(r).$$

Hence, our elliptic system (1) has been reduced to the elliptic equation (13) with the critical exponent. By now, it is standard that our solution pair u and v is of the form (14). This completes the proof of Theorem 2.

Acknowledgments. The second author expresses his deep appreciation to Prof. W. X. Chen for helpful discussions on the method of moving planes. This work was done while the first author was visiting the Department of Mathematical Sciences in Tsinghua University, Beijing, which he would like to thank for their hospitality. Both authors thank the referees for detailed comments and suggestions on the original manuscript.

REFERENCES

[1] J. Bourgain, *Global Solutions of Nonlinear Schrödinger Equations*, Amer. Math. Soc. Colloq. Publ. 46, AMS, Providence, RI, 1999.

[2] L. Caffarelli, B. Gidas, and J. Spruck, *Asymptotic symmetry and local behavior of semilinear elliptic equations with critical Sobolev growth*, Comm. Pure Appl. Math., 42 (1989), pp. 271–297.

[3] W. Chen and C. Li, *Classification of solutions of some nonlinear elliptic equations*, Duke Math. J., 63 (1991), pp. 615–622.

[4] W. Chen and C. Li, *Regularity of solutions for a system of integral equations*, Comm. Pure Appl. Anal., 4 (2005), pp. 1–8.
[5] W. Chen and C. Li, *The best constant in weighted Hardy-Littlewood-Sobolev inequality*, Proc. Amer. Math. Soc., 136 (2008), pp. 955–962.
[6] W. Chen, C. Li, and B. Ou, *Classification of solutions for a system of integral equations*, Comm. Partial Differential Equations, 30 (2005), pp. 59–65.
[7] W. Chen, C. Li, and B. Ou, *Qualitative properties of solutions for an integral equation*, Discrete Contin. Dyn. Syst., 12 (2005), pp. 347–354.
[8] W. Chen, C. Li, and B. Ou, *Classification of solutions for an integral equation*, Comm. Pure Appl. Math., 59 (2006), pp. 330–343.
[9] D. G. de Figueiredo and P. L. Felmer, *On superquadratic elliptic systems*, Trans. Amer. Math. Soc., 343 (1994), pp. 99–116.
[10] D. G. de Figueiredo and P. L. Felmer, *A Liouville-type theorem for elliptic systems*, Ann. Sc. Norm. Super. Pisa Cl. Sci. (4), 21 (1994), pp. 387–397.
[11] B. Gidas, W. M. Ni, and L. Nirenberg, *Symmetry of positive solutions of nonlinear elliptic equations in* $\mathbf{R}^n$, in Mathematical Analysis and Applications, Part A, Adv. Math. Suppl. Stud. 7A, Academic Press, New York, 1981, pp. 369–402.
[12] F. B. Hang, *On the integral systems related to Hardy-Littlewood-Sobolev inequality*, Math. Res. Lett., 14 (2007), pp. 373–383.
[13] C. Jin and C. Li, *Symmetry of solutions to some integral equations*, Proc. Amer. Math. Soc., 134 (2006), pp. 1661–1670.
[14] C. Jin and C. Li, *Quantitative analysis of some system of integral equations*, Calc. Var. Partial Differential Equations, 26 (2006), pp. 447–457.
[15] T. Kanna and M. Lakshmanan, *Exact soliton solutions, shape changing collisions, and partially coherent solitons in coupled nonlinear Schrödinger equations*, Phys. Rev. Lett., 86 (2001), pp. 5043–5046.
[16] C. Li, *Local asymptotic symmetry of singular solutions to nonlinear elliptic equations*, Invent. Math., 123 (1996), pp. 221–231.
[17] C. Li and J. Lim, *The singularity analysis of solutions to some integral equations*, Commun. Pure Appl. Anal., 6 (2007), pp. 453–464.
[18] E. Lieb, *Sharp constants in the Hardy-Littlewood-Sobolev and related inequalities*, Ann. of Math., 118 (1983), pp. 349–374.
[19] T. C. Lin and J. Wei, *Ground state of N coupled nonlinear Schrödinger equations in $\mathbf{R}^n$, $n \leq 3$*, Commun. Math. Phys., 255 (2005), pp. 629–653.
[20] T. C. Lin and J. Wei, *Spikes in two coupled nonlinear Schrödinger equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 22 (2005), pp. 403–439.
[21] L. Ma and D. Z. Chen, *A Liouville type theorem for an integral system*, Comm. Pure Appl. Anal., 5 (2006), pp. 855–859.
[22] L. Ma and D. Z. Chen, *Radial symmetry and monotonicity for an integral equation*, J. Math. Anal. Appl., 342 (2008), pp. 943–949.
[23] L. Ma and L. Zhao, *Sharp thresholds of blow up and global existence for the coupled nonlinear Schrödinger system*, J. Math. Phys., 49 (2008), article 062103.
[24] M. Naito and H. Usami, *Existence of nonoscillatory solutions to second-order elliptic systems of Emden-Fowler type*, Indiana Univ. Math. J., 55 (2006), pp. 317–339.
[25] J. Serrin and H. Zou, *Non-existence of positive solutions of Lane-Emden systems*, Differential Integral Equations, 9 (1996), pp. 635–653.
[26] E. M. Stein and G. Weiss, *Fractional integrals in n-dimensional Euclidean space*, J. Math. Mech., 7 (1958), pp. 503–514.
[27] J. Wei and X. Xu, *Classification of solutions of higher order conformally invariant equations*, Math. Ann., 313 (1999), pp. 207–228.
[28] M. I. Weinstein, *Nonlinear Schrödinger equations and sharp interpolate estimates*, Comm. Math. Phys., 87 (1983), pp. 567–576.

SIAM J. MATH. ANAL.
Vol. 40, No. 3, pp. 1058–1075

STABILITY OF TRAVELING WAVES IN QUASI-LINEAR HYPERBOLIC SYSTEMS WITH RELAXATION AND DIFFUSION*

TONG LI†

Abstract. We establish the existence and the stability of traveling wave solutions of a quasi-linear hyperbolic system with both relaxation and diffusion. The traveling wave solutions are shown to be asymptotically stable under small disturbances and under the subcharacteristic condition using a weighted energy method. The delicate balance between the relaxation and the diffusion that leads to the stability of the traveling waves is identified; namely, the diffusion coefficient is bounded by a constant multiple of the relaxation time. Such a result provides an important first step toward the understanding of the transition from stability to instability as parameters vary in physical problems involving both relaxation and diffusion.

Key words. stability, traveling wave, relaxation, diffusion, weighted energy method, traffic flow

AMS subject classifications. 35B30, 35B40, 35L65, 76L05, 90B20

DOI. 10.1137/070690638

1. Introduction.

The phenomenon of relaxation arises in many physical problems such as kinetic relaxation to fluid dynamics, gases not in local thermodynamic equilibrium, elasticity with memory, phase transitions, shallow water waves, and traffic flows. A remarkable development of the stability theory for various relaxation systems has appeared in past decades; see, e.g., [1, 2, 8, 10, 11, 13, 15, 16, 18, 23]. The real physical problems usually involve both relaxation and diffusion. It was found by asymptotic analysis and numerical simulations that the fine interplay between the relaxation and the diffusion may enhance physically interesting behavior such as soliton waves and oscillatory solutions [4, 3, 5, 6, 7, 9, 14, 17, 20, 22, 24]. However, the rigorous stability theory for such systems has not been well studied. In the current paper, we establish rigorously the existence and the stability of traveling wave solutions of a quasi-linear hyperbolic system with both relaxation and diffusion. The delicate balance between the relaxation and the diffusion that leads to the nonlinear stability of the traveling wave fronts is identified to occur when the diffusion coefficient is bounded by a constant multiple of the relaxation time. Such a result provides an important first step toward understanding the transition from stability to instability as the diffusion coefficient and the relaxation time vary in the physical problems.

Consider the following quasi-linear hyperbolic system with relaxation and diffusion:

$$\begin{cases} v_t - u_x = 0, \\ u_t + p(v)_x = \dfrac{1}{\tau}(u_e(v) - u) + \mu u_{xx} \end{cases} \tag{1}$$

subject to the initial data

$$(v,u)(x,0) = (v_0,u_0)(x) \to (v_\pm,u_\pm) \text{ as } x \to \pm\infty, \quad u_\pm = u_e(v_\pm), \tag{2}$$

where, in the context of traffic flows, v is specific volume and u velocity. The first equation in (1) is a conservation law, while the second describes drivers' acceleration

*Received by the editors May 6, 2007; accepted for publication (in revised form) June 20, 2008; published electronically October 17, 2008.
http://www.siam.org/journals/sima/40-3/69063.html
†Department of Mathematics, University of Iowa, Iowa City, IA 52242 (tli@math.uiowa.edu).

behavior. The first term on the right-hand side of the second equation in (1) expresses the tendency of traffic at a given specific volume v to relax to some equilibrium speed u_e satisfying

$$u_e'(v) > 0.$$

The parameter $\tau > 0$ corresponds to drivers' response time to the traffic. The second term on the left-hand side of the second equation in (1) is an anticipation factor: drivers slow down at the sight of an increase in traffic density ahead. The function p is the so-called traffic pressure satisfying

$$p'(v) < 0. \tag{3}$$

The last term on the right-hand side of the second equation in (1) models viscosity with coefficient $\mu > 0$, a presumed tendency to adjust one's speed to that of the surrounding traffic.

A strict subcharacteristic condition,

$$-\sqrt{-p'(v)} < u_e'(v) < \sqrt{-p'(v)}, \tag{4}$$

is imposed for all v under consideration. Subcharacteristic condition (4) is a necessary condition for linear stability (Whitham [23]) and for nonlinear stability (Li and Liu [10] and Liu [15]) when $\mu = 0$.

The purpose of this paper is to show the existence and stability of the traveling wave solutions $(V, U)(x - st)$ of (1) under appropriate conditions on the relaxation time τ and the diffusion coefficient μ.

A traveling wave solution is a solution of the form

$$(v, u)(x, t) = (V, U)(x - st) \equiv (V, U)(z),$$

where $z = x - st$, satisfying

$$(V, U)(z) \to (v_\pm, u_\pm) \text{ as } z \to \pm\infty, \quad u_\pm = u_e(v_\pm), \tag{5}$$

where v_+, v_-, and s satisfy the Rankine–Hugoniot condition

$$-s(v_+ - v_-) - u_e(v_+) + u_e(v_-) = 0 \tag{6}$$

and the entropy condition

$$(v_+ - v_-)(u_e(v) - u_e(v_\pm) + s(v - v_\pm)) < 0 \tag{7}$$

for all v in between v_+ and v_-. Indeed, the corresponding jump (v_-, v_+) is an admissible shock of the equilibrium equation

$$v_t - u_e(v)_x = 0. \tag{8}$$

Furthermore, the speed s of the traveling wave must be subcharacteristic; i.e., for all $z \in R$ it holds

$$-\sqrt{-p'(V(z))} < s < \sqrt{-p'(V(z))}. \tag{9}$$

For a weight function $w \geq 0$, L^2_w denotes the space of measurable functions f satisfying $\sqrt{w} f \in L^2$ with norm

$$||f||_{L^2_w} = \left(\int w(x) |f(x)|^2 dx \right)^{1/2}.$$

H_w^j, $j \geq 0$, denotes the weighted Sobolev space with norm

$$||f||_{H_w^j} = \left(\sum_{k=0}^{j} ||\partial_x^k f||_w^2 \right)^{1/2}.$$

We now state our main results.

THEOREM 1.1. *Suppose that subcharacteristic conditions* (4) *and* (9)*, the Rankine–Hugoniot condition* (6)*, and the entropy condition* (7) *hold, and suppose that the diffusion coefficient is appropriately small,*

$$0 < \mu \leq m\tau, \tag{10}$$

for some $m > 0$ critically depending on subcharacteristic conditions (4) *and* (9)*. Then there exists a traveling wave solution $(V, U)(x - st)$ of* (1) *and* (5)*, which is unique up to a shift.*

Moreover, there exists a constant $\varepsilon_0 > 0$ such that if

$$|v_- - v_+| + ||v_0(\cdot) - V(\cdot + x_0)||_2 + ||u_0(\cdot) - U(\cdot + x_0)||_2 + ||(\phi_0, \phi_{0,x}, \psi_0)||_{L_w^2} \leq \varepsilon_0,$$

where x_0 is determined by

$$\int_{-\infty}^{+\infty} (v_0 - V)(x)dx = x_0(v_+ - v_-) \tag{11}$$

and

$$(\phi_0, \psi_0)(x) = \left(\int_{-\infty}^{x} (v_0(y) - V(y + x_0))dy, u_0(x) - U(x + x_0) \right)$$

and where the weight function w is defined as

$$w(V(x + x_0)) = \frac{(V(x + x_0) - v_+)(V(x + x_0) - v_-)}{Q(V(x + x_0))}, \tag{12}$$

then the Cauchy problem (1), (2) *has a unique global solution $(v, u)(x, t)$ satisfying*

$$v(x,t) - V(x + x_0 - st), u(x,t) - U(x + x_0 - st) \in C^0(0, \infty; H^2 \cap L_w^2) \cap L^2(0, \infty; H^2 \cap L_w^2)$$

and

$$\sup_{x \in R} |(v, u)(x, t) - (V, U)(x + x_0 - st)| \to 0 \quad as \quad t \to +\infty. \tag{13}$$

Previously, Jin and Liu [4] derived that the weakly nonlinear limit of the relaxation system of Jin and Xin [2] with a viscous term is governed by the Kortweg–de Vries (KdV) equation and dispersive waves are enhanced when the diffusion coefficient dominates in the sense that

$$\tau \ll O(\tau^{\frac{1}{4}}) \ll \mu.$$

When the diffusion coefficient is not too small, Kerner and Konhäuser [6], Kurtze and Hong [7], Lee, Lee, and Kim [9], and Ou et al. [22] were able to derive a modified KdV equation and obtain soliton-like solutions and oscillatory solutions for traffic flow models similar to (1) by asymptotic analysis and numerical simulations. The

phenomena occur in other applications. Keener [5] studied the diffusion induced insulin oscillatory secretion. He found that oscillatory secretion of insulin results from an important interplay between flow rate of the reactor and insulin diffusion.

In the absence of the diffusion term,

$$\mu = 0,$$

Li and Liu [10] established the nonlinear stability of traveling waves of (1) under the subcharacteristic conditions (4) and (9). The stability result is a consequence of the nonlinearity, the relaxation, and the subcharacteristic conditions (4) and (9).

In the current paper, we prove that, in addition to assuming the subcharacteristic conditions (4) and (9), if the diffusion coefficient is appropriately small (10), then traveling wave solutions of (1) are stable. In the context of traffic flows, our results indicate that the stability of the traveling fronts is guaranteed if the diffusion mechanism, the tendency to adjust one's speed to that of the surrounding traffic, is dominated by the relaxation mechanism, the tendency of traffic to relax to some equilibrium speed.

The plan of this paper is the following. In section 2, we prove the existence of traveling wave solutions of the relaxation system with viscosity (1) by the phase plane analysis provided that the diffusion coefficient is appropriately small (10). In section 3, the stability problem is reformulated in terms of perturbations to the underlying traveling wave. Section 4 is devoted to establishing the desired a priori estimates for the nonlinear stability of the traveling wave solutions. When the equilibrium function $u_e(v)$ is nonconvex, there are traveling wave solutions that decay at an algebraic decay rate at infinity. A weighted energy method was developed to establish the stability of such degenerate traveling wave profiles [12, 10, 16, 19]. We adapt the weighted energy method developed in [12, 10, 16, 19] to a system of conservation laws with both relaxation and diffusion (1). It is under condition (10) that we prove the desired weighted energy estimates that lead to the stability of the traveling wave solutions.

2. Existence of traveling wave profiles. In this section, we prove the existence of traveling wave solutions of the relaxation system with viscosity (1). We show that, under the subcharacteristic conditions (4) and (9), and when the diffusion coefficient is appropriately small (10), there is a traveling wave profile of (1) connecting two states satisfying the Rankine–Hugoniot condition (6) and the entropy condition (7).

LEMMA 2.1. *Assume that v_+, v_-, and s satisfy the Rankine–Hugoniot condition* (6)*, the entropy condition* (7)*, and the subcharacteristic conditions* (4), (9)*, and that the diffusion coefficient is appropriately small* (10)*. There exists a traveling wave solution $(V,U)(x-st)$ of* (1) *with boundary condition* (5)*, which is unique up to a shift.*

Moreover, the profile $V = V(z)$ is monotone and

$$(v_+ - v_-)(u_e(V(z)) - u_e(v_\pm) + s(V(z) - v_\pm)) < 0 \tag{14}$$

for all $z \in R$.

Proof. Substituting

$$(v,u)(x,t) = (V,U)(z), \quad z = x - st,$$

into (1), we have

$$\begin{cases} -sV_z - U_z = 0, \\ -sU_z + p(V)_z = \frac{u_e(V)-U}{\tau} + \mu U_{zz}. \end{cases} \tag{15}$$

Integrating the first equation of (15) over $(\pm\infty, z)$ and using boundary condition (5) yield

$$-sV - U = -sv_\pm - u_\pm = -sv_\pm - u_e(v_\pm). \tag{16}$$

Therefore the Rankine–Hugoniot condition (6) for the equilibrium equation (8) is satisfied by the end states of the traveling wave of the relaxation system (1).

We derive from (15), (16), and (5) that V satisfies a second order nonlinear differential equation

$$V_{zz} + \frac{p'(V) + s^2}{s\mu} V_z - \frac{u_e(V) - u_e(v_\pm) + s(V - v_\pm)}{\tau s\mu} = 0 \tag{17}$$

and

$$V(z) \to v_\pm \ \text{ as } \ z \to \pm\infty. \tag{18}$$

We will establish the existence of solutions for (17) and (18) by phase plane analysis.

Let $W = V'$. Equation (17) is rewritten as a system of first order differential equations

$$\begin{cases} V' = W, \\ W' = -\frac{p'(V)+s^2}{s\mu} W + \frac{u_e(V)-u_e(v_\pm)+s(V-v_\pm)}{\tau s\mu}. \end{cases} \tag{19}$$

$(v_\pm, 0)$ are two equilibrium points of (19). Without loss of generality, we assume that $v_- > v_+$. Other cases can be treated similarly.

The entropy condition (7) implies that

$$Q(V) \equiv -\frac{1}{\tau}(u_e(V) - u_e(v_\pm) + s(V - v_\pm)) < 0, \ \ v_+ < V < v_-, \tag{20}$$

and consequently

$$-u_e'(v_+) \le s < -u_e'(v_-) \le 0.$$

Consider

$$-u_e'(v_+) < s < -u_e'(v_-) < 0. \tag{21}$$

The degenerate cases $-u_e'(v_+) = s$ and $s = -u_e'(v_-)$ can be treated similarly.

Linearize the system of nonlinear equations (19) at equilibrium points $(v_\pm, 0)$ to obtain

$$\begin{cases} (V - v_\pm)' = W, \\ W' = -\frac{p'(v_\pm)+s^2}{s\mu} W + \frac{u_e'(v_\pm)+s}{\tau s\mu}(V - v_\pm). \end{cases} \tag{22}$$

The eigenvalues of the Jacobian of (22) satisfy

$$\lambda^2 + \lambda \frac{p'(v_\pm) + s^2}{s\mu} - \frac{u_e'(v_\pm) + s}{\tau s\mu} = 0.$$

The entropy condition (21), the subcharacteristic condition (9), and $s < 0$ imply that

$$-\frac{u_e'(v_-) + s}{\tau s\mu} < 0, \ \ \frac{p'(v_-) + s^2}{s\mu} > 0. \tag{23}$$

Thus $(v_-, 0)$ is a saddle with eigenvalues

$$\lambda_{1-} > 0 > \lambda_{2-}$$

and the corresponding right eigenvectors

$$r_{1-} = (1, \lambda_{1-})^T, \;\; r_{2-} = (1, \lambda_{2-})^T.$$

Therefore the unstable manifold of (19) at $(v_-, 0)$ is tangent to r_{1-}.

Similarly, the entropy condition (21), the subcharacteristic condition (9), and $s < 0$ imply that

$$-\frac{u_e'(v_+) + s}{\tau s \mu} > 0, \;\; \frac{p'(v_+) + s^2}{s\mu} > 0. \tag{24}$$

Thus $(v_+, 0)$ is a stable node provided that

$$\left(\frac{p'(v_+) + s^2}{s\mu}\right)^2 + 4\frac{u_e'(v_+) + s}{\tau s \mu} > 0$$

or

$$0 < \frac{\mu}{\tau} < -\frac{(p'(v_+) + s^2)^2}{4s(u_e'(v_+) + s)}.$$

The above inequality holds provided that the diffusion coefficient is appropriately small (10).

The eigenvalues of the Jacobian of (22) satisfy

$$0 > \lambda_{1+} > \lambda_{2+},$$

and the corresponding right eigenvectors are

$$r_{1+} = (1, \lambda_{1+})^T, \;\; r_{2+} = (1, \lambda_{2+})^T.$$

The nullclines of (19) consist of

$$W = 0 \tag{25}$$

and

$$W = \frac{u_e(V) - u_e(v_\pm) + s(V - v_\pm)}{\tau(p'(V) + s^2)} \equiv g(V). \tag{26}$$

The nullclines intersect at the equilibrium points $(v_\pm, 0)$, where

$$g(v_\pm) = 0.$$

Moreover, by using the entropy condition (21) and the subcharacteristic condition (9), we have

$$g'(v_-) = \frac{u_e'(v_-) + s}{(p'(v_-) + s^2)\tau} > 0, \;\; g'(v_+) = \frac{u_e'(v_+) + s}{(p'(v_+) + s^2)\tau} < 0, \tag{27}$$

and

$$g(V) < 0, \;\; v_+ < V < v_-.$$

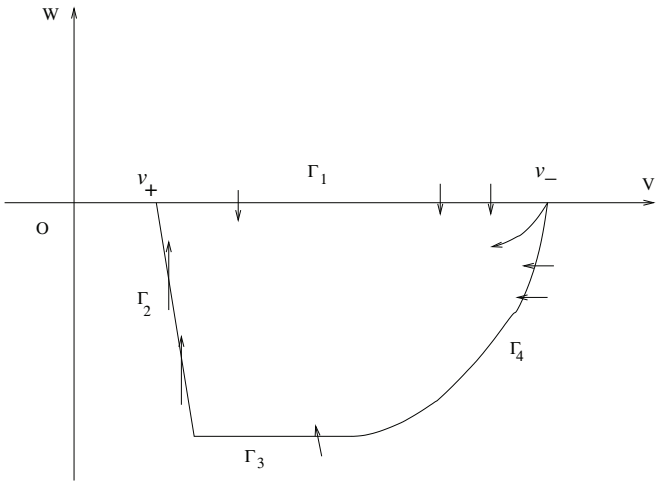


FIG. 1. *The invariant domain D.*

Let $v_0 \in (v_+, v_-)$ be the minimum point of g on $[v_+, v_-]$ and

$$W_0 = g(v_0) = \min_{v_+ < V < v_-} g(V).$$

Then

$$W_0 < 0, \;\; g'(v_0) = 0.$$

Let D be a region bounded by the following curves on the phase plane (see Figure 1):

$$\Gamma_1 : \;\; W = 0, \;\; v_+ \le V \le v_-,$$

$$\Gamma_2 : \;\; W = -a(V - v_+), \;\; v_+ \le V \le v_1,$$

$$\Gamma_3 : \;\; W = W_0 = g(v_0), \;\; v_1 \le V \le v_0,$$

$$\Gamma_4 : \;\; W = g(V), \;\; v_0 \le V \le v_-,$$

where $a > 0$ is to be determined and (v_1, W_0) is the intersection point of Γ_2 and Γ_3 satisfying

$$W_0 = -a(v_1 - v_+), \;\; v_+ \le v_1 \le v_0.$$

We will show that, under the condition that the diffusion coefficient is appropriately small (10), a can be chosen such that D is an invariant region of (19).

(i) From (19) and (21) we derive that on the top boundary of D, Γ_1, $W' < 0$, the flow points downward toward the interior of D.

(ii) On the bottom boundary of D, Γ_3, the flow points in the upper left direction, which points toward the interior of D. This is because Γ_3 lies below the nullcline $W = g(V)$ and thus $W' > 0$.

(iii) On the right boundary of D, Γ_4, which is part of the nullcline $W = g(V)$, the flow points leftward toward the interior of D. This is because $V' = W \le 0$ and $W' = 0$.

(iv) It remains to show that the flow points toward the interior of D on the left boundary of D, Γ_2.

Statement (iv) is true if we can choose $a > 0$ such that

$$\left.\frac{dW}{dV}\right|_{\Gamma_2} < -a < 0, \quad v_+ \le V \le v_1. \tag{28}$$

We show that such $a > 0$ exists if the diffusion coefficient μ is appropriately small (10). Noting (24) and (27), we have

$$\lambda_{1+} = \frac{2\frac{u_e'(v_+)+s}{\tau s\mu}}{\sqrt{\left(\frac{p'(v_+)+s^2}{s\mu}\right)^2 + 4\frac{u_e'(v_+)+s}{\tau s\mu}} + \frac{p'(v_+)+s^2}{s\mu}} < g'(v_+) < 0.$$

Thus Γ_2 is below $W = g(V)$ for V near v_+, $V \in (v_+,\ v_+ + \epsilon)$ for some $\epsilon > 0$.

Indeed, if $\lambda_{1+} < \min_{[v_+,v_0]} g'(V)$, we will choose a satisfying

$$\lambda_{1+} < -a < \min_{[v_+,v_0]} g'(V) < 0.$$

Thus Γ_2 is below $W = g(V)$ for $v_+ \le V \le v_1$.

We now show that (28) is satisfied. Since

$$\lambda_{2+} < \lambda_{1+} < -a < g'(v_+) < 0,$$

there is an $\epsilon > 0$, such that

$$\left.\frac{dW}{dV}\right|_{\Gamma_2} < -a < 0, \quad v_+ \le V \le v_+ + \epsilon. \tag{29}$$

If V is away from v_+, then there is a $\delta > 0$ such that

$$0 < \frac{g(V)}{W} \le 1 - \delta, \quad v_+ + \epsilon \le V \le v_1.$$

From (19) we have

$$\left.\frac{dW}{dV}\right|_{\Gamma_2} = \frac{W'}{V'} = -\frac{p'(V)+s^2}{s\mu}\left(1 - \frac{g(V)}{W}\right) \le -\delta\frac{p'(V)+s^2}{s\mu} < 0$$

for $v_+ + \epsilon \le V \le v_1$. Noting the definition of g and (26) and assuming that the diffusion coefficient μ is appropriately small (10), we have

$$\left.\frac{dW}{dV}\right|_{\Gamma_2} \le -\delta\frac{p'(V)+s^2}{s\mu} < g'(V) < 0, \quad v_+ + \epsilon \le V \le v_1.$$

Thus

$$\left.\frac{dW}{dV}\right|_{\Gamma_2} < -a < 0, \quad v_+ + \epsilon \le V \le v_1. \tag{30}$$

Combining (29) and (30), we prove that there is an a satisfying (28). If $\min_{[v_+,v_0]} g'(V) < \lambda_{1+}$, we consider a piecewise linear curve Γ_2 instead. The above argument can be modified to prove (28).

Therefore D is an invariant region of (19).

Now we look for a trajectory connecting the two equilibrium points $(v_\pm, 0)$. Noting (23) and (27), we have

$$0 < \lambda_{1-} = \frac{2\frac{u_e'(v_-)+s}{\tau s\mu}}{\sqrt{\left(\frac{p'(v_-)+s^2}{s\mu}\right)^2 + 4\frac{u_e'(v_-)+s}{\tau s\mu}} + \frac{p'(v_-)+s^2}{s\mu}} < g'(v_-).$$

Since the unstable manifold of (19) at $(v_-, 0)$ is tangent to $r_{1-} = (1, \lambda_{1-})^T$ with $0 < \lambda_{1-} < g'(v_-)$, there is a trajectory of (19) originating from $(v_-, 0)$ that enters the domain D. Since D is an invariant region, such a trajectory is trapped inside D. Therefore the trajectory flows into the stable node $(v_+, 0)$ as $t \to +\infty$.

We thus find a solution V of (17) and (18) which is monotone decreasing. The entropy condition (7) follows from (20) and the above monotonicity. The existence of the traveling wave profile (V, U) of (15) and (5) is proved by using (16).

Lemma 2.1 is proved. □

3. Reformulation of the stability problem. In this section, we reformulate the problem that the solution $(v, u)(x, t)$ of (1), (2) exists globally and approaches a shifted traveling wave solution $(V, U)(x+x_0-st)$ under the subcharacteristic condition (4) and under condition (10) as $t \to \infty$.

We will look for a solution of the following form:

$$(v, u)(x, t) = (V, U)(z + x_0) + (\phi_z, \psi)(z, t), \tag{31}$$

where $z = x - st$.

For initial data satisfying (11), the conservation law in (1) implies that

$$\phi(\pm\infty, t) = \int_{-\infty}^{+\infty} (v(x,t) - V(x + x_0 - st))dx = \int_{-\infty}^{+\infty} (v_0(x) - V(x + x_0))dx$$

$$= x_0(v_+ - v_-) - \int_{-\infty}^{+\infty} (V(x + x_0) - V(x))dx = 0.$$

For simplicity of notation, we assume the shift $x_0 = 0$ in the rest of the paper.

We substitute (31) into (1), by virtue of (15), and integrate the first equation once with respect to z, to obtain that the perturbation (ϕ, ψ) satisfies

$$\begin{cases} \phi_t - s\phi_z - \psi = 0, \\ \psi_t - s\psi_z + (p(V + \phi_z) - p(V))_z \\ = \dfrac{1}{\tau}(u_e(V + \phi_z) - u_e(V) - \psi) + \mu\psi_{zz}. \end{cases} \tag{32}$$

The first equation of (32) gives

$$\psi = \phi_t - s\phi_z. \tag{33}$$

Substituting (33) into the second equation of (32), we get a closed equation for ϕ,

$$L(\phi) \equiv (\phi_t - s\phi_z)_t - s(\phi_t - s\phi_z)_z + (p'(V)\phi_z)_z + \frac{1}{\tau}\phi_t + \lambda\phi_z - \mu(\phi_t - s\phi_z)_{zz}$$

$$= -F(V, \phi_z), \tag{34}$$

where $Q < 0$ is defined in (20) and

$$\lambda = Q'(V) = -\frac{1}{\tau}(u_e'(V) + s),$$

and

$$F(V, \phi_z) = F_1 + F_2$$

with

$$F_1 := -\frac{1}{\tau}(u_e(V + \phi_z) - u_e(V) - u_e'(V)\phi_z), \tag{35}$$

$$F_2 := (p(V + \phi_z) - p(V) - p'(V)\phi_z)_z = (G(V, \phi_z)\phi_z^2)_z, \tag{36}$$

and

$$G(V, \phi_z) := \int_0^1 \int_0^1 p''(V + \theta\eta\phi_z)\theta d\theta d\eta$$

which is the error term due to nonlinearity of the function p.

The corresponding initial data for (34) become

$$\phi(z, 0) = \phi_0(z), \quad \phi_t(z, 0) = s\phi_0'(z) - \psi_0 = \phi_1(z). \tag{37}$$

The asymptotic stability of the profile (V, U) means that the perturbation (ϕ_z, ψ) decays to zero as $t \to +\infty$.

First, by noting (14), we have that the weight function defined in (12) satisfies $w(V) > 0$.

Now we introduce the solution space of the problem (34), (37) as follows:

$$X(0, T) =$$
$$\{\phi(z, t) : \ \phi \in C^0([0, T); H^3 \cap L^2_w) \cap C^1(0, T; H^2 \cap L^2_w), \ \ \phi_z, \phi_t \in L^2(0, T; H^2 \cap L^2_w)\}$$

with $0 < T \le +\infty$.

By virtue of (33), we have

$$\psi \in C^0([0, T); H^2 \cap L^2_w) \cap L^2(0, T; H^2 \cap L^2_w).$$

By the Sobolev embedding theorem, if we let

$$N(t) = \sup_{0 \le s \le t} \{||\phi(s)||_3 + ||\phi_t(s)||_2 + \|\phi(s)\|_{H^1_w} + ||\phi_t(s)||_{L^2_w}\}, \tag{38}$$

then

$$\sup_{z \in R} \{|\phi|, |\phi_z|, |\phi_{zz}|, |\phi_t|, |\phi_{tz}|\} \le CN(t). \tag{39}$$

Thus Theorem 1.1 is a consequence of the following theorem.

THEOREM 3.1. *Under the conditions of Theorem* 1.1, *there exist positive constants* δ_1, C, *and* C_1 *such that if* $N(0) \le \delta_1$, *then the problem* (34), (37) *has a unique global solution* $\phi \in X(0, +\infty)$ *satisfying*

$$||\phi(t)||_3^2 + ||\phi_t||_2^2 + ||\phi||_{H^1_w}^2 + ||\phi_t||_{L^2_w}^2 + \int_0^t ||(\phi_t, \phi_z)(s)||_2^2 ds$$
$$+ \ C_1\mu \left(\int_0^t ||(\phi_t - s\phi_z)_z||_{L^2_w}^2 ds + \int_0^t ||(\phi_t - s\phi_z)_{zz}||_1^2 ds \right) \le CN^2(0) \tag{40}$$

for $t \in [0, +\infty)$. Furthermore,

$$\sup_{z \in R} |(\phi_z, \phi_t)(z,t)| \to 0 \quad \textit{as} \ \ t \to \infty. \tag{41}$$

If ϕ is the global solution in the above theorem, then (ϕ, ψ), defined in (33), becomes a global solution of the problem (32) with $(\phi, \psi)(z,0) = (\phi_0, \psi_0)(z)$, and consequently we have the desired solution of the problem (1), (2) through the relation (31). On the other hand, the solution of (1) is unique in the space $C^0(0,T; H^2 \cap L^2_w)$; therefore Theorem 1.1 follows from Theorem 3.1. The estimate (40) gives

$$\begin{aligned} \phi_t^2 + \phi_z^2 &= \int_{-\infty}^{z} (2\phi_t \phi_{tz} + 2\phi_z \phi_{zz})(y,t)dy \\ &\le \left(\int_{-\infty}^{+\infty} (\phi_t^2 + \phi_z^2) dy \right)^{1/2} \left(\int_{-\infty}^{+\infty} (\phi_{tz}^2 + \phi_{zz}^2) dy \right)^{1/2} \to 0 \ \text{ as } \ t \to \infty \end{aligned}$$

as claimed in Theorem 1.1.

Global existence for ϕ will be derived from the following local existence theorem for ϕ combined with a priori estimates.

PROPOSITION 3.2 (local existence). *For any $\delta_0 > 0$, there exists a positive constant T_0 depending on δ_0, such that if $\phi_0 \in H^3 \cap H^1_w$ and $\phi_1 \in H^2 \cap L^2_w$, with $N(0) < \delta_0/2$, then the problem* (34), (37) *has a unique solution $\phi \in X(0, T_0)$ satisfying*

$$N(t) < 2N(0) \tag{42}$$

for any $0 \le t \le T_0$.

PROPOSITION 3.3 (a priori estimates). *Let $\phi \in X(0,T)$ be a solution for a positive constant T; then there exists a positive constant δ_2 independent of T such that if*

$$N(t) < \delta_2, \ \ t \in [0,T],$$

then ϕ satisfies (40) *for any $0 \le t \le T$.*

The local existence is classical, so we omit the proof; cf. [21]. Proving Proposition 3.3 is our main task in the following section.

4. Energy estimates. In this section, we will complete the proof of our stability theorem by establishing the desired a priori estimates. The stability result is a consequence of the compressibility of the traveling wave profile (14), the subcharacteristic conditions (4) and (9), the fact that the diffusion coefficient is bounded by a constant multiple of the relaxation time (10), and the weighted energy estimates.

LEMMA 4.1. *Under the conditions of Theorem* 1.1*, there are positive constants C and C_1 such that if subcharacteristic condition* (4) *is satisfied for $v \in [v_+, v_-]$, then any solution $\phi \in X(0,T)$ of problem* (34), (37) *satisfies*

$$||\phi(t)||^2_{H^1_w} + ||\phi_t(t)||^2_{L^2_w} + \int_0^t ||(\phi_t, \phi_z)(s)||^2_{L^2_w} ds$$

$$+ \int_0^t \int_R |U_z| \phi^2 dz ds + C_1 \mu \int_0^t \int_R w(\phi_t - s\phi_z)_z^2 dz ds$$

$$\leq C\left\{||\phi_0||^2_{H^1_w}+||\phi_1||^2_{L^2_w}+\int_0^t\int_R w|F|(|\phi|+|(\phi_t,\phi_z)|)dzds\right\} \tag{43}$$

for $t\in[0,T]$.

Proof. Let $w=w(V)>0$ be the weight function defined in (12).

Multiplying (34) by $2w(V)\phi$, we obtain

$$2w(V)\phi L(\phi)=-2Fw(V)\phi. \tag{44}$$

The left-hand side of (44) is reduced to

$$\begin{aligned}
&2[(\phi_t-s\phi_z)_t-s(\phi_t-s\phi_z)_z+(p'(V)\phi_z)_z]w\phi+2\left(\frac{1}{\tau}\phi_t+\lambda\phi_z\right)w\phi\\
&-2\mu w\phi(\phi_t-s\phi_z)_{zz}\\
=&\left[\frac{1}{\tau}w\phi^2+2w\phi(\phi_t-s\phi_z)\right]_t-2w(\phi_t-s\phi_z)^2-2wp'(V)\phi_z^2+(p'(V)w_z)_z\phi^2\\
&-(\lambda w)_z\phi^2+sw_z(\phi^2)_t-s^2\{w_z(\phi^2)\}_z+s^2w_{zz}\phi^2\\
&+\{-2sw\phi(\phi_t-s\phi_z)+2p'(V)w\phi\phi_z-p'(V)w_z\phi^2+\lambda w\phi^2-2\mu w\phi(\phi_t-s\phi_z)_z\}_z\\
&+2\mu w_z\phi(\phi_t-s\phi_z)_z+2\mu w\phi_z(\phi_t-s\phi_z)_z\\
=&\left[\frac{1}{\tau}w\phi^2+2w\phi(\phi_t-s\phi_z)+sw_z\phi^2\right]_t-2w(\phi_t-s\phi_z)^2-2p'(V)w\phi_z^2+A(V)\phi^2\\
&+2\mu w_z\phi(\phi_t-s\phi_z)_z+2\mu w\phi_z(\phi_t-s\phi_z)_z+2\mu sw'(V)V_{zz}\phi\phi_z+\{\cdots\}_z,
\end{aligned} \tag{45}$$

where $\{\cdots\}_z$ denotes the terms which will disappear after integration with respect to $z\in R$ and

$$\begin{aligned}
A(V)&=-\{(-p'(V)-s^2)w'(V)V_z+\lambda w-\mu sw'(V)V_{zz}\}_z\\
&=-\{w'(V)Q(V)+Q'(V)w\}_z\\
&=-\{wQ\}''(V)V_z,
\end{aligned} \tag{46}$$

where the second equality is due to (17).

Since V is monotone decreasing and

$$(wQ)''(V)=2,$$

therefore

$$A(V)=-2V_z>0. \tag{47}$$

Then we calculate

$$2(\phi_t-s\phi_z)wL(\phi)=-2F(\phi_t-s\phi_z)w. \tag{48}$$

The left-hand side of (48) is

$$
\begin{aligned}
&2[(\phi_t - s\phi_z)_t - s(\phi_t - s\phi_z)_z + (p'(V)\phi_z)_z]w(\phi_t - s\phi_z)\\
&+ \frac{2}{\tau}w(\phi_t - s\phi_z)(\phi_t - s\phi_z - uV_e'(V)\phi_z) - 2\mu w(\phi_t - s\phi_z)(\phi_t - s\phi_z)_{zz}\\
&= [w(\phi_t - s\phi_z)^2]_t + \left(\frac{2}{\tau}w + sw_z\right)(\phi_t - s\phi_z)^2 - 2p'(V)w_z\phi_z(\phi_t - s\phi_z)\\
&- \frac{2}{\tau}wu_e'(V)\phi_z(\phi_t - s\phi_z) - [p'(V)w\phi_z^2]_t + [swp'(V)\phi_z^2]_z - s(wp'(V))_z\phi_z^2\\
&- [sw(\phi_t - s\phi_z)^2 - 2p'(V)w\phi_z(\phi_t - s\phi_z)]_z\\
&- (2\mu w(\phi_t - s\phi_z)(\phi_t - s\phi_z)_z)_z + 2\mu w_z(\phi_t - s\phi_z)(\phi_t - s\phi_z)_z + 2\mu w(\phi_t - s\phi_z)_z^2\\
&= [-wp'(V)\phi_z^2 + w(\phi_t - s\phi_z)^2]_t + \left(\frac{2}{\tau}w + sw_z\right)(\phi_t - s\phi_z)^2\\
&- s(wp'(V))_z\phi_z^2 - \frac{2}{\tau}u_e'(V)w\phi_z(\phi_t - s\phi_z) - 2p'(V)w_z\phi_z(\phi_t - s\phi_z)\\
&+ 2\mu w_z(\phi_t - s\phi_z)(\phi_t - s\phi_z)_z + 2\mu w(\phi_t - s\phi_z)_z^2 + \{\cdots\}_z.
\end{aligned}
$$

Hence, the combination (44) + (48) $\times 2\tau$ yields

$$
\begin{aligned}
&\{E_1(\phi, (\phi_t - s\phi_z)) + E_3(\phi_z)\}_t + E_2(\phi_z, (\phi_t - s\phi_z)) + E_4(\phi) + 4\mu\tau w(\phi_t - s\phi_z)_z^2\\
&\qquad + 2\mu w_z\phi(\phi_t - s\phi_z)_z + 2\mu w\phi_z(\phi_t - s\phi_z)_z + 2\mu sw'(V)V_{zz}\phi\phi_z\\
&\qquad + 4\mu\tau w_z(\phi_t - s\phi_z)(\phi_t - s\phi_z)_z + \{\cdots\}_z = -2Fw\{\phi + 2\tau(\phi_t - s\phi_z)\},
\end{aligned}
\tag{49}
$$

where

$$
\begin{aligned}
E_1(\phi, (\phi_t - s\phi_z)) &= 2\tau w(\phi_t - s\phi_z)^2 + 2w\phi(\phi_t - s\phi_z) + \left(\frac{1}{\tau}w + sw_z\right)\phi^2,\\
E_3(\phi_z) &= -2\tau wp'(V)\phi_z^2,\\
E_2(\phi_z, (\phi_t - s\phi_z)) &= 2(w + \tau sw_z)(\phi_t - s\phi_z)^2\\
&\quad + 4(-u_e'(V)w - \tau p'(V)w_z)\phi_z(\phi_t - s\phi_z)\\
&\quad + 2(-wp'(V) - \tau s(wp'(V))_z)\phi_z^2,\\
E_4(\phi) &= A(V)\phi^2,
\end{aligned}
\tag{50}
$$

where $A(V)$ is defined in (46).

The discriminants of the quadratics $E_j (j = 1, 2)$ are, respectively,

$$D_1 = -4w[w + 2\tau sw_z],$$

$$D_2 = 16\{(-u_e'(V)w - \tau p'(V)w_z)^2 + (w + \tau sw_z)(wp'(V) + \tau s(wp'(V))_z)\}.$$

We claim that

$$D_1 < 0, \quad D_2 < 0, \tag{51}$$

provided that $|v_+ - v_-|$ is suitably small and subcharacteristic condition (4) is satisfied.

Indeed, the two inequalities in (51) are equivalent to

$$1 + 2s\tau\frac{w_z}{w} > 0, \tag{52}$$

$$\left(-u_e'(V)-\tau p'(V)\frac{w_z}{w}\right)^2 < \left(1+s\tau\frac{w_z}{w}\right)\left(-p'(V)-s\tau\frac{(wp'(V))_z}{w}\right). \tag{53}$$

Noticing that $|\tau(w(V))_z|$ and $|\tau(w(V)p'(V))_z|$ are small provided that $|v_+ - v_-|$ is suitably small and subcharacteristic condition (4) is satisfied, we derive inequalities (52) and (53). Thus condition (51) is satisfied.

Therefore there exist positive constants M_0 and M such that

$$\begin{cases} M_0 w\{\phi^2+(\phi_t-s\phi_z)^2\} \;\le\; E_1 \;\le\; Mw\{\phi^2+(\phi_t-s\phi_z)^2\}, \\ \\ M_0 w\{\phi_z^2+(\phi_t-s\phi_z)^2\} \;\le\; E_2. \end{cases} \tag{54}$$

Furthermore, (3) and (47) yield

$$\begin{cases} E_3 = -2\tau w p'(V)\phi_z^2 \ge 0, \\ \\ E_4 \ge 2|V_z|\phi^2 \ge 0. \end{cases} \tag{55}$$

Now we estimate terms containing μ in (49).

First, note that $4\mu\tau w(\phi_t - s\phi_z)_z^2$ is a good term, namely, nonnegative. Other terms containing μ are estimated by using the Cauchy–Schwarz inequality, estimate (54), the fact that $|w(V)_z|$ and $|V_{zz}|$ are small, and (10). Indeed, let us estimate the second term

$$\begin{aligned} &|2\mu w\phi_z(\phi_t-s\phi_z)_z| \\ &\le \mu\tau w(\phi_t-s\phi_z)_z^2 + \frac{\mu}{4\tau}w\phi_z^2 \le \mu\tau w(\phi_t-s\phi_z)_z^2 + \frac{1}{4}M_0 w\phi_z^2 \end{aligned} \tag{56}$$

provided that $\frac{\mu}{\tau} \le m \le M_0$ (10).

Substituting estimates (54), (55), and (56) into (49) and integrating the result with respect to t and z, we arrive at the desired estimate (43).

This completes the proof of Lemma 4.1. □

Next we estimate the higher order derivatives of ϕ.

Let $\phi_z = \Phi$; then

$$\begin{aligned} \partial_z L(\phi) &= (\phi_{zt}-s\phi_{zz})_t - s(\phi_{zt}-s\phi_{zz})_z + (p'(V)\phi_z)_{zz} \\ &\quad + \phi_{zt} + \lambda\phi_{zz} + \lambda_z\phi_z - \mu(\phi_{zt}-s\phi_{zz})_{zz} \\ &= L(\phi_z) + \lambda_z\phi_z + (p'(V)_z\phi_z)_z = L(\Phi) + \lambda_z\Phi + (p'(V)_z\Phi)_z. \end{aligned} \tag{57}$$

Multiplying the derivative of (34) with respect to z by $2\phi_z$ and $2(\phi_t - s\phi_z)_z$, respectively, we have

$$\begin{aligned} 2\partial_z L(\phi)\phi_z &= -2F_z\phi_z, \\ 2\partial_z L(\phi)(\phi_t-s\phi_z)_z &= -2F_z(\phi_t-s\phi_z)_z. \end{aligned}$$

By a similar argument used in obtaining estimates for the first order derivatives with $w = 1$, we have

$$\begin{aligned} &\left[\frac{1}{\tau}\Phi^2 + 2\Phi(\Phi_t - s\Phi_z)\right]_t - 2p'(V)\Phi_z^2 - 2(\Phi_t-s\Phi_z)^2 + \lambda_z\Phi^2 \\ &\quad + 2\mu\Phi_z(\Phi_t-s\Phi_z)_z - 2p'(V)_z\Phi\Phi_z + \{\cdots\}_z = -2F_z\Phi \end{aligned} \tag{58}$$

and

$$[(\Phi_t - s\Phi_z)^2 - p'(V)\Phi_z^2]_t + \frac{2}{\tau}(\Phi_t - s\Phi_z)^2 - \frac{2}{\tau}u_e'(V)\Phi_z(\Phi_t - s\Phi_z)$$
$$- sp'(V)_z\Phi_z^2 + 2\lambda_z\Phi(\Phi_t - s\Phi_z) + 2(p'(V)\Phi)_z(\Phi_t - s\Phi_z) + \{\cdots\}_z$$
$$+ 2\mu(\Phi_t - s\Phi_z)_z^2 = -2F_z(\Phi_t - s\Phi_z). \tag{59}$$

The combination (58) + (59) $\times\, 2\tau$ yields

$$\{E_1(\Phi, (\Phi_t - s\Phi_z)) + E_3(\Phi_z)\}_t + E_2(\Phi_z, (\Phi_t - s\Phi_z)) + H(\Phi) + \{\cdots\}_z$$
$$+ 4\mu\tau(\Phi_t - s\Phi_z)_z^2 + 2\mu\Phi_z(\Phi_t - s\Phi_z)_z = -F_z\{2\Phi + 4\tau(\Phi_t - s\Phi_z)\}, \tag{60}$$

where

(61)

$$H(\Phi) = \lambda_z\Phi^2 + 4\tau\lambda_z\Phi(\Phi_t - s\Phi_z) - 2p'(V)_z\Phi\Phi_z + 4\tau(p'(V)\Phi)_z(\Phi_t - s\Phi_z),$$
$$E_1(\Phi, (\Phi_t - s\Phi_z)) = 2\tau(\Phi_t - s\Phi_z)^2 + 2\Phi(\Phi_t - s\Phi_z) + \frac{1}{\tau}\Phi^2,$$
$$E_3(\Phi_z) = -2p'(V)\Phi_z^2,$$
$$E_2(\Phi_z, (\Phi_t - s\Phi_z)) = 2(\Phi_t - s\Phi_z)^2 - 4u_e'(V)\Phi_z(\Phi_t - s\Phi_z)$$
$$-2(p'(V) + s\tau p'(V)_z)\Phi_z^2.$$

Integrating (60) with respect to t and z and taking $\frac{\mu}{\tau}$ and $|v_+ - v_-|$ suitably small, we have the following estimate:

$$||\Phi(t)||_1^2 + ||\Phi_t(t)||^2 + \int_0^t ||(\Phi_t, \Phi_z)(s)||^2 ds + C_1\mu\int_0^t ||(\Phi_t - s\Phi_z)_z(s)||^2 ds$$
$$\le C\left\{||\Phi_0||_1^2 + ||\Phi_1||^2 + \int_0^t\int |H(\Phi)|dzds + \int_0^t\int_R |F_z|(|\Phi| + |(\Phi_t, \Phi_z)|)dzds\right\}, \tag{62}$$

where $\Phi_0 = \phi_0'$ and $\Phi_1 = \phi_1'$, and C and C_1 are positive constants.

Using the estimate (43), we obtain

$$\int_0^t\int_R |H(\Phi)|dzds \le \frac{1}{2}\int_0^t ||(\Phi_t, \Phi_z)(s)||^2 ds + C\int_0^t\int_R \Phi^2 dzds$$
$$\le \frac{1}{2}\int_0^t ||(\Phi_t, \Phi_z)(s)||^2 ds$$
$$+ C\left\{||\phi_0||_{H_w^1}^2 + ||\phi_1||_{L_w^2}^2 + \int_0^t\int w|F|(|\phi| + |(\phi_t, \phi_z)|)dzds\right\}. \tag{63}$$

Substituting (63) into (62) and replacing Φ by $\partial_z\phi$, we have the following lemma.

LEMMA 4.2. *Under the conditions of Theorem* 1.1*, there are positive constants* C *and* C_1 *such that if subcharacteristic condition* (4) *is satisfied for all* $v \in [v_+, v_-]$*, then any solution* $\phi \in X(0,T)$ *of problem* (34), (37) *satisfying*

$$||\partial_z\phi||_1^2 + ||\partial_z\phi_t||^2 + \frac{1}{2}\int_0^t ||(\partial_z\phi_t, \partial_z\phi_z)(s)||^2 ds + C_1\mu\int_0^t ||(\phi_t - s\phi_z)_{zz}(s)||^2 ds$$
$$\le C\left\{||\phi_0||_2^2 + ||\phi_1||_1^2 + \int_0^t\int |F_z|(|\partial_z\phi| + |(\partial_z\phi_t, \partial_z\phi_z)|)dzds\right.$$
$$\left. + ||\phi_0||_{H_w^1}^2 + ||\phi_1||_{L_w^2}^2 + \int_0^t\int w|F|(|\phi| + |(\phi_t, \phi_z)|)dzds\right\} \tag{64}$$

holds for $t \in [0,T]$.

Next we calculate the equality

$$\partial_z^2\phi\partial_z^2 L(\phi) + 2\partial_z^2(\phi_t - s\phi_z)\partial_z^2 L(\phi) = -F_{zz}(\partial_z^2\phi + 2\partial_z^2(\phi_t - s\phi_z))$$

in the same way as in the proof of Lemma 4.2. Setting $\Psi := \partial_z^2\phi$, we have

$$\partial_z^2 L(\phi) = L(\Psi) + (p'(V)\Phi)_{zzz} - (p'(V)\Psi_z)_z + (\lambda\Phi)_{zz} - \lambda\Psi_z.$$

A straightforward calculation gives

$$\begin{aligned}
&\left[2\tau(\Psi_t - s\Psi_z)^2 - 2\tau p'(V)\Psi_z^2 + 2\tau\Psi(\Psi_t - s\Psi_z) + \frac{1}{\tau}\Psi^2\right]_t + 2(\Psi_t - s\Psi_z)^2 \\
&\quad - 4u_e'(V)\Psi_z(\Psi_t - s\Psi_z) + 2(-p'(V) - s\tau p'(V)_z)\Psi_z^2 + 8\tau\lambda_z\Psi(\Psi_t - s\Psi_z) \\
&\qquad + 3\tau\lambda_z\Psi^2 + 2\tau\lambda_{zz}\Psi\phi_z + 4\tau\lambda_{zz}\phi_z(\Psi_t - s\Psi_z) + 4\mu\tau(\Psi_t - s\Psi_z)_z^2 \\
&\qquad + 2\mu\Psi_z(\Psi_t - s\Psi_z)_z + \{\cdots\}_z = -2F_{zz}[\Psi + 2\tau(\Psi_t - s\Psi_z)] + J, \qquad (65)
\end{aligned}$$

where J satisfies

$$\begin{aligned}
|J| &= |[-(p'(V)\Phi)_{zzz} + (p'(V)\Psi_z)_z][\Psi + 2\tau(\Psi_t - s\Psi_z)]| \\
&\le \frac{1}{3}|(\Psi_t, \Psi_z)|^2 + C|(\Phi, \Psi)|^2.
\end{aligned}$$

Thus, noting $\Psi = \phi_{zz}$ and (10), we derive from (65) and $|v_+ - v_-| \ll 1$ that

$$\begin{aligned}
&||\partial_z^2\phi(t)||_1^2 + ||\partial_z^2\phi_t||^2 + \frac{2}{3}\int_0^t ||(\partial_z^2\phi_t, \partial_z^2\phi_z)(s)||^2 ds - C\int_0^t (||\partial_z^2\phi||^2 + ||\phi_z||^2) ds \\
&\qquad + C_1\mu\int_0^t ||(\partial_z^2\phi_t - s\partial_z^2\phi_z)_{zz}(s)||^2 ds \\
&\qquad \le C\left\{||\phi_0||_3^2 + ||\phi_1||_2^2 + \left|\int_0^t\int F_{zz}(\partial_z^2\phi + 2\partial_z^2(\phi_t - s\phi_z)) dz ds\right|\right\}, \qquad (66)
\end{aligned}$$

where C and C_1 are positive constants.

Combining successively estimates (43), (64), and (66), we have

$$\begin{aligned}
&||\phi(t)||_3^2 + ||\phi_t(t)||_2^2 + ||\phi(t)||_{H_w^1}^2 + ||\phi_t(t)||_{L_w^2}^2 + \int_0^t\int |\lambda_z|\phi^2 dz ds + \int_0^t ||(\phi_t, \phi_z)||_2^2 ds \\
&+ \int_0^t ||(\phi_t, \phi_z)(s)||_{L_w^2} w^2 ds + C_1\mu\left(\int_0^t ||(\phi_t - s\phi_z)_z||_{L_w^2}^2 ds + \int_0^t ||(\phi_t - s\phi_z)_{zz}||_1^2 ds\right) \\
&\qquad \le C\Big\{||\phi_0||_3^2 + ||\phi_1||_2^2 + ||\phi_0||_{H_w^1}^2 + ||\phi_1||_{L_w^2}^2 \\
&\qquad + \int_0^t\int (w|F|(|\phi| + |(\phi_t, \phi_z)|) + |F_z|(|\partial_z\phi| + |(\partial_z\phi_t, \partial_z\phi_z)|)) dz ds \\
&\qquad + \left|\int_0^t\int F_{zz}(\partial_z^2\phi + 2\partial_z^2(\phi_t - s\phi_z)) dz ds\right|\Big\}, \qquad (67)
\end{aligned}$$

where

$$F = F_1 + F_2 = O(1)(\phi_z^2 + \phi_{zz}^2)$$

as defined in (35), (36). There are terms containing fourth order derivatives of ϕ in $F_{2,zz}$. The energy estimates for the derivatives of ϕ up to fourth order can be established; see [10].

Thus by virtue of (39), the integrals on the right-hand side of (67) are majored by

$$CN(t)\left(\int_0^t ||(\phi_t,\phi_z)||_2^2 ds + \int_0^t ||(\phi_t,\phi_z)(s)||_{L_w^2}^2 ds\right);$$

then we have

$$\begin{aligned} &N^2(t) + \int_0^t ||(\phi_t,\phi_z)||_2^2 ds + \int_0^t ||(\phi_t,\phi_z)(s)||_{L_w^2}^2 ds \\ &\le CN^2(0) + CN(t)\left(\int_0^t ||(\phi_t,\phi_z)||_2^2 ds + \int_0^t ||(\phi_t,\phi_z)(s)||_{L_w^2}^2 ds\right). \end{aligned}$$

Therefore, by assuming $N(T) \le \frac{1}{2C} = \delta_2$, we obtain the desired estimate

$$N^2(t) + \int_0^t ||(\phi_t,\phi_z)||_2^2 ds + \int_0^t ||(\phi_t,\phi_z)(s)||_{L_w^2}^2 ds \le CN^2(0) \quad \text{for} \quad t \in [0,T].$$

By choosing $N(0) \le \delta_1$ small, we arrive at $N(T) \le \delta_2$ for any $T > 0$.

Thus the proof of Proposition 3.3 is completed. □

In summary, we established the weighted energy estimates by assuming the smallness of $|v_+ - v_-|$, the compressibility of the traveling wave profile (14), the subcharacteristic conditions (4) and (9), and the fact that the diffusion coefficient is bounded by a constant multiple of the relaxation time (10).

Acknowledgment. The author would like to acknowledge Professor Doochul Kim for insightful conversations which motivated the current work.

REFERENCES

[1] G.-Q. Chen, C. D. Levermore, and T. P. Liu, *Hyperbolic conservation laws with stiff relaxation terms and entropy*, Comm. Pure Appl. Math., 47 (1994), pp. 787–830.

[2] S. Jin and Z. Xin, *The relaxing schemes for systems of conservation laws in arbitrary space dimensions*, Comm. Pure Appl. Math., 48 (1995), pp. 235–276.

[3] W. L. Jin and H. M. Zhang, *The formation and structure of vehicle clusters in the Payne–Whitham traffic flow model,* Transportation Res. Part B, 37 (2003), pp. 207–223.

[4] S. Jin and J.-G. Liu, *Relaxation and diffusion enhanced dispersive waves,* Proc. R. Soc. Lond. Ser. A, 446 (1994), pp. 555–563.

[5] J. P. Keener, *Diffusion induced oscillatory insulin secretion*, Bull. Math. Biol., 63 (2001), pp. 625–641.

[6] B. S. Kerner and P. Konhäuser, *Structure and parameters of clusters in traffic flow*, Phys. Rev. E, 50 (1994), pp. 54–83.

[7] D. A. Kurtze and D. C. Hong, *Traffic jams, granular flow, and soliton selection,* Phys. Rev. E, 52 (1995), pp. 218–221.

[8] C. Lattanzio and P. Marcati, *The zero relaxation limit for the hydrodynamic Whitham traffic flow model,* J. Differential Equations, 141 (1997), pp. 150–178.

[9] H. Y. Lee, H.-W. Lee, and D. Kim, *Steady-state solutions of hydrodynamic traffic models,* Phys. Rev. E, 69 (2004), 016118.

[10] T. Li and H. Liu, *Stability of a traffic flow model with nonconvex relaxation,* Commun. Math. Sci., 3 (2005), pp. 101–118.

[11] T. Li and H. Liu, *Critical thresholds in a relaxation model for traffic flows,* Indiana Univ. Math. J., 57 (2008), pp. 1409–1431.

[12] T. Li, *Rigorous asymptotic stability of a CJ detonation wave in the limit of small resolved heat release*, Combust. Theory Model., 1 (1997), pp. 259–270.
[13] T. Li, *Global solutions of nonconcave hyperbolic conservation laws with relaxation arising from traffic flow,* J. Differential Equations, 190 (2003), pp. 131–149.
[14] T. Li, *Nonlinear dynamics of traffic jams*, Phys. D, 207 (2005), pp. 41–51.
[15] T. P. Liu, *Hyperbolic conservation laws with relaxation*, Comm. Math. Phys., 108 (1987), pp. 153–175.
[16] H. Liu, J. Wang, and T. Yang, *Stability of a relaxation model with a nonconvex flux*, SIAM J. Math. Anal., 29 (1998), pp. 18–29.
[17] M. Muramatsu and T. Nagatani, *Soliton and kink jams in traffic flow with open boundaries,* Phys. Rev. E, 60 (1999), pp. 180–187.
[18] C. Mascia and K. Zumbrun, *Stability of large-amplitude shock profiles of general relaxation systems*, SIAM J. Math. Anal., 37 (2005), pp. 889–913.
[19] A. Matsumura and K. Nishihara, *Asymptotic stability of traveling waves for scalar viscous conservation laws with non-convex nonlinearity,* Comm. Math. Phys., 165 (1994), pp. 83–96.
[20] T. Nagatani, *The physics of traffic jams*, Rep. Progr. Phys., 65 (2002), pp. 1331–1386.
[21] T. Nishida, *Nonlinear Hyperbolic Equations and Related Topics in Fluid Dynamics*, Publ. Math. d'Orsay 78-02, Département de Mathématique, Université de Paris-Sud, Orsay, France, 1978.
[22] Z.-H. Ou, S.-Q. Dai, P. Zhang, and L.-Y. Dong, *Nonlinear analysis in the Aw–Rascle anticipation model of traffic flow*, SIAM J. Appl. Math., 67 (2007), pp. 605–618.
[23] G. B. Whitham, *Linear and Nonlinear Waves*, Wiley, New York, 1974.
[24] H. M. Zhang, *Driver memory, traffic viscosity and a viscous vehicular traffic flow model,* Transportation Res. Part B, 37 (2003), pp. 27–41.

SIAM J. MATH. ANAL.
Vol. 40, No. 3, pp. 1076–1103

NONEXISTENCE OF SUPERSONIC TRAVELING WAVES FOR NONLINEAR SCHRÖDINGER EQUATIONS WITH NONZERO CONDITIONS AT INFINITY*

MIHAI MARIŞ†

Abstract. We prove that the nonexistence of supersonic finite-energy traveling waves for nonlinear Schrödinger equations with nonzero conditions at infinity is a general phenomenon which holds for a large class of equations. The same is true for sonic traveling waves in two dimensions. In higher dimensions we prove that sonic traveling waves, if they exist, must approach their limit at infinity in a very rigid way. In particular, we infer that there are no sonic traveling waves with finite energy and finite momentum.

Key words. nonlinear Schrödinger equation, nonzero conditions at infinity, traveling wave, integral identities, Gross–Pitaevskii equations and systems, cubic-quintic nonlinear Schrödinger equation

AMS subject classifications. 35Q51, 35Q55, 35Q40, 35B65, 35J15, 35J20, 35J50, 37K40, 37K05

DOI. 10.1137/070711189

1. Introduction.

The aim of this paper is to study traveling wave solutions for nonlinear Schrödinger equations

$$i\frac{\partial \Phi}{\partial t} + \Delta \Phi + F(x, |\Phi|^2)\Phi = 0 \qquad \text{in } \mathbf{R}^N, \tag{1.1}$$

where F is a real-valued function defined on $\mathbf{R}^N \times \mathbf{R}_+$, Φ is a complex-valued function on $\mathbf{R}^N$ satisfying the "boundary condition" $|\Phi| \longrightarrow r_0$ as $|x| \longrightarrow \infty$, and r_0 is a positive constant verifying $\lim_{|x|\to\infty,\ s\to r_0^2} F(x,s) = 0$.

The above equation with the considered nonzero conditions at infinity arises in a large variety of physical problems, such as superconductivity, superfluidity in helium II, phase transitions, and Bose–Einstein condensates. Two important particular cases of (1.1) have been extensively studied both by physicists and by mathematicians: the Gross–Pitaevskii equation (where $F(x,s) = 1 - s$) and the so-called cubic-quintic Schrödinger equation (where $F(x,s) = -\alpha_1 + \alpha_3 s - \alpha_5 s^2$, α_1, α_3, α_5 are positive, and $\frac{3}{16} < \frac{\alpha_1\alpha_5}{\alpha_3^2} < \frac{1}{4}$).

Equation (1.1) has a Hamiltonian structure: denoting $V(x,s) = \int_s^{r_0^2} F(x,\tau)\,d\tau$, it is easy to see that, at least formally, the "energy"

$$E(\Phi) = \int_{\mathbf{R}^N} |\nabla \Phi|^2\,dx + \int_{\mathbf{R}^N} V(x, |\Phi|^2)\,dx \tag{1.2}$$

is a conserved quantity. There is another important (vector) quantity associated to (1.1), namely, the momentum. It is given by

$$P(\Phi) = (P_1(\Phi), \dots, P_N(\Phi)), \text{where } P_k(\Phi) = \int_{\mathbf{R}^N} \left(i\frac{\partial \Phi}{\partial x_k}, \Phi \right) dx = \int_{\mathbf{R}^N} \mathrm{Re}\left(i\frac{\partial \Phi}{\partial x_k}\overline{\Phi} \right) dx. \tag{1.3}$$

*Received by the editors December 17, 2007; accepted for publication (in revised form) June 17, 2008; published electronically October 17, 2008.
http://www.siam.org/journals/sima/40-3/71118.html
†Département de Mathématiques, UMR 6623, Université de Franche-Comté, 16, Route de Gray, 25030 Besançon Cedex, France (mihai.maris@univ-fcomte.fr).

Note that, in general, the momentum is not well-defined for any function Φ of finite energy. In the case where F does not depend on the variable x_k, the momentum with respect to the x_k-direction, P_k, is conserved by those solutions of (1.1) for which it can be well-defined.

It is worth noting that (1.1) can be put into a hydrodynamical form by using Madelung's transformation $\Phi(x,t) = \sqrt{\rho(x,t)}e^{i\theta(x,t)}$ (which is singular when $\Phi = 0$). A straightforward computation shows that in the region where $\Phi \neq 0$, the functions $\rho = |\Phi|^2$ and θ satisfy the system

$$\rho_t + 2\mathrm{div}(\rho\nabla\theta) = 0, \tag{1.4}$$

$$\theta_t + |\nabla\theta|^2 - \frac{\Delta\rho}{2\rho} + \frac{|\nabla\rho|^2}{4\rho} - F(x,\rho) = 0. \tag{1.5}$$

Equation (1.4) and the derivatives with respect to $x_1, \ldots, x_N$ of (1.5) are, respectively, the equation of conservation of mass and Euler's equations for a compressible inviscid fluid of density ρ and velocity $2\nabla\theta$.

Let us assume that F admits a partial derivative with respect to the last variable (in what follows, this derivative will be denoted by $\partial_{N+1}F$ or by $\frac{\partial F}{\partial s}$) and that $\lim_{|x|\to\infty,\ \rho\to r_0^2} \partial_{N+1}F(x,\rho) = -L$, where L is a positive constant. Taking the derivative with respect to t of (1.5) and substituting ρ_t from (1.4) we obtain

$$\theta_{tt} + 2\partial_{N+1}F(x,\rho)(\rho\Delta\theta + \nabla\rho.\nabla\theta) + \frac{\partial}{\partial t}\left(|\nabla\theta|^2 - \frac{\Delta\rho}{2\rho} + \frac{|\nabla\rho|^2}{4\rho}\right) = 0. \tag{1.6}$$

For a small oscillatory motion (i.e., a sound wave), all nonlinear terms in (1.6), except $2\rho\Delta\theta$, may be neglected. In view of the behavior of ρ and $\partial_{N+1}F(x,\rho)$ for large $|x|$, we find that in a neighborhood of infinity, the velocity potential θ essentially obeys the wave equation $\theta_{tt} - 2r_0^2L\Delta\theta = 0$. It is well known that the solutions of the wave equation propagate with a finite speed; in the present situation, we infer that the velocity of sound waves at infinity is $r_0\sqrt{2L}$. In what follows we will always assume that $\partial_{N+1}F(x,\rho) \longrightarrow -L$ as $|x| \longrightarrow \infty$ and $\rho \longrightarrow r_0^2$ (the convergence being in a sense to be defined) and we will denote by $v_s = r_0\sqrt{2L}$ the sound velocity at infinity.

For a fixed $y \in S^{N-1}$, a traveling wave for (1.1) moving with velocity c in direction y is a solution of the form $\Phi(x,t) = \psi(x - cty)$. Without loss of generality we will assume that $y = (1,0,\ldots,0)$, i.e., traveling waves move in the x_1-direction. The traveling wave profile satisfies the equation

$$-ic\frac{\partial\psi}{\partial x_1} + \Delta\psi + F(x,|\psi|^2)\psi = 0 \qquad \text{in } \mathbf{R}^N. \tag{1.7}$$

In a series of papers, Grant, Jones, Putterman, and Roberts, among others, formally and numerically studied traveling waves for the Gross–Pitaevskii equation and related systems (see, e.g., [16], [19], [21], [22], [5], and the references therein). In particular, they conjectured that such solutions exist if and only if their speed c belongs to the interval $(-v_s, v_s)$. For the cubic-quintic nonlinear Schrödinger equation, the existence of subsonic traveling waves in one dimension has been proved in [1] and their stability has been studied in [2]. The nonexistence of such solutions for sonic and supersonic speeds has also been conjectured in any space dimension. In the case of the Gross–Pitaevskii equation, it has been shown in [17] that any traveling wave of finite energy and speed $c > v_s$ must be constant. It has also been proved in [18] that the same result is true if $N = 2$ and $c^2 = v_s^2$. The proofs in

[17], [18] strongly depend on the special algebraic structure of the nonlinearity in the Gross–Pitaevskii equation. In the present paper we show that the nonexistence of finite-energy traveling waves moving faster than the sound velocity is a general phenomenon which holds for a large class of equations and systems of the form (1.1). We also prove that there are no finite-energy sonic traveling waves in space dimension two. In higher dimensions we show that any finite-energy sonic traveling wave ψ must satisfy $|\psi|^2 - r_0^2 \in L^p(\mathbf{R}^N)$ for any $p > \frac{2N-1}{2N-3}$. On the other hand, if a sonic traveling wave satisfies $|\psi|^2 - r_0^2 \in L^{\frac{2N-1}{2N-3}}(\mathbf{R}^N)$, then it must be constant.

This article is organized as follows: In the next section we prove that traveling waves, whenever they exist, are smooth functions. If their speed is supersonic (or sonic, provided they converge sufficiently fast at infinity), then they must satisfy a special integral identity. This will be proved in section 3. In section 4 we show how this identity implies, under general assumptions, the nonexistence of traveling waves with finite energy. We apply our results to the Gross–Pitaevskii equation, to the cubic-quintic Schrödinger equation, and to a Gross–Pitaevskii–Schrödinger system, which describes the motion of an uncharged impurity in a Bose condensate. In the last section we describe all supersonic and sonic traveling waves (with finite or infinite energy) for one-dimensional equations with nonlinearities independent on the space variable.

2. Basic properties of traveling waves. We keep the previous notation and consider the following set of assumptions:

- **(H1)** $F : \mathbf{R}^N \times [0,\infty) \longrightarrow \mathbf{R}$ is a measurable function which has the following properties:
  (a) for any $s \in [0,\infty)$, $F(\cdot, s)$ is measurable;
  (b) for any $x \in \mathbf{R}^N$, $F(x,\cdot)$ is continuous;
  (c) F is bounded on bounded subsets of $\mathbf{R}^N \times [0,\infty)$.
- **(H2)** There exist $\alpha > 0$, $C > 0$, and $r_* > 0$ such that for any $x \in \mathbf{R}^N$ and for any $s \geq r_*$ we have $F(x,s) \leq -Cs^\alpha$.
- **(H3)** $\lim_{|x|\to\infty} F(x, r_0^2) = 0$ and $F(\cdot, r_0^2) \in L^1(\mathbf{R}^N)$.
- **(H4)** F admits a partial derivative with respect to the last variable and $\partial_{N+1}F$ is bounded on bounded subsets of $\mathbf{R}^N \times [0,\infty)$. Moreover, $\lim_{|x|\to\infty} \partial_{N+1}F(x, r_0^2) = -L$, where $L > 0$ and $\partial_{N+1}F(\cdot, r_0^2) + L \in L^{p_0}(\mathbf{R}^N)$ for some $p_0 \in [1,2]$.
- **(H5)** There are some positive constants R_0, η, M such that $\partial_{N+1}^2 F$ exists on $(\mathbf{R}^N \setminus \overline{B}(0,R_0)) \times (r_0^2 - \eta, r_0^2 + \eta)$ and

$$|\partial_{N+1}^2 F(x,s)| \leq M \qquad \text{for all } (x,s) \in (\mathbf{R}^N \setminus \overline{B}(0,R_0)) \times (r_0^2 - \eta, r_0^2 + \eta).$$

Definition 2.1. *A traveling wave (of speed c) for* (1.1) *is a function $\psi \in L^1_{loc}(\mathbf{R}^N)$ that satisfies* (1.7) *in $\mathcal{D}'(\mathbf{R}^N)$ together with the "boundary condition" $|\psi| \longrightarrow r_0$ as $|x| \longrightarrow \infty$.*

In view of (1.2)*, we say that a traveling wave ψ has finite energy if $\nabla\psi \in L^2(\mathbf{R}^N)$ and $V(\cdot, |\psi|^2) \in L^1(\mathbf{R}^N)$.*

We have the following result concerning the regularity of traveling waves.

Proposition 2.2. *Let ψ be a finite-energy traveling wave for* (1.1).

(i) *Assume that $F : \mathbf{R}^N \times \mathbf{R}_+ \longrightarrow \mathbf{R}$ is measurable and satisfies* (H1a), (H1b), (H2)*, the function $x \longmapsto \int_{r_0^2}^{r_*} F(x,\tau)\, d\tau$ belongs to $L^1_{loc}(\mathbf{R}^N)$ (where r_* is given by* (H2)*), and $F(\cdot, |\psi|^2)\psi \in L^1_{loc}(\mathbf{R}^N)$. Then $\psi \in L^\infty(\mathbf{R}^N)$.*

If, in addition, F satisfies (H1c), *then $\psi \in W^{2,p}_{loc}(\mathbf{R}^N)$ for any $p \in [1,\infty)$. In particular, $\psi \in C^{1,\alpha}(\mathbf{R}^N)$ for any $\alpha \in [0,1)$.*

(ii) *Suppose that $F \in C^k(\mathbf{R}^N \times [0,\infty))$ for some $k \in \mathbf{N}^*$,* (H2) *holds, and $F(\cdot,|\psi|^2)\psi \in L^1_{loc}(\mathbf{R}^N)$. Then $\psi \in W^{k+2,p}_{loc}(\mathbf{R}^N)$ for any $p \in [1,\infty)$. In particular, if F is C^∞, then $\psi \in C^\infty(\mathbf{R}^N)$.*

Proof. (i) The proof relies upon the ideas and methods developed by Farina in [13], [14]. By (H2) we have

$$V(x,s) = -\int_{r_0^2}^{s} F(x,\tau)\,d\tau \geq -\int_{r_0^2}^{r_*} F(x,\tau)\,d\tau + \int_{r_*}^{s} C\tau^\alpha\,d\tau$$
$$= -\int_{r_0^2}^{r_*} F(x,\tau)\,d\tau + \frac{C}{\alpha+1}(s^{\alpha+1} - r_*^{\alpha+1}).$$

Consequently, for any $s \geq r_*$ we get $s^{\alpha+1} \leq r_*^{\alpha+1} + \frac{\alpha+1}{C}(V(x,s) + \int_{r_0^2}^{r_*} F(x,\tau)\,d\tau)$, so that

$$|\psi|^{2\alpha+2}(x) \leq \max\Big(r_*^{\alpha+1}, r_*^{\alpha+1} + \frac{\alpha+1}{C}\Big(V(x,|\psi|^2(x)) + \int_{r_0^2}^{r_*} F(x,\tau)\,d\tau\Big)\Big).$$

Since $V(\cdot,|\psi|^2)$ and $\int_{r_0^2}^{r_*} F(\cdot,\tau)\,d\tau$ belong to $L^1_{loc}(\mathbf{R}^N)$, we infer that $\psi \in L^{2\alpha+2}_{loc}(\mathbf{R}^N)$.

We will use a well-known inequality of Kato (see Lemma A, p. 138, in [23]):

If $u \in L^1_{loc}(\mathbf{R}^N)$ is a real-valued function and $\Delta u \in L^1_{loc}(\mathbf{R}^N)$, then

$$\Delta(u^+) \geq \operatorname{sgn}^+(u)\Delta u \qquad \text{in } \mathcal{D}'(\mathbf{R}^N). \tag{2.1}$$

Let $\varphi(x) = e^{-\frac{icx_1}{2}}\psi(x)$. Then $\varphi \in L^{2\alpha+2}_{loc}(\mathbf{R}^N) \subset L^1_{loc}(\mathbf{R}^N)$ and an easy computation shows that φ satisfies

$$\Delta\varphi + \Big(F(x,|\varphi|^2) + \frac{c^2}{4}\Big)\varphi = 0 \qquad \text{in } \mathcal{D}'(\mathbf{R}^N). \tag{2.2}$$

It is clear that $F(\cdot,|\varphi|^2)\varphi \in L^1_{loc}(\mathbf{R}^N)$ (because $F(x,|\psi|^2)\psi \in L^1_{loc}(\mathbf{R}^N)$ by hypothesis) and it follows from (2.2) that $\Delta\varphi \in L^1_{loc}(\mathbf{R}^N)$. Choose $\tilde r \geq r_*$ and $C_1 > 0$ such that $Cs^{2\alpha} - \frac{c^2}{4} \geq C_1(s-\tilde r)^{2\alpha}$ for any $s \geq \tilde r$. Denoting $\varphi_1 = \operatorname{Re}(\varphi)$, $\varphi_2 = \operatorname{Im}(\varphi)$ and using Kato's inequality for $\varphi_i - \tilde r$, $i = 1,2$, then using (2.2) and (H2) we get

$$\begin{aligned}\Delta(\varphi_i - \tilde r)^+ &\geq \operatorname{sgn}^+(\varphi_i - \tilde r)\Delta(\varphi_i - \tilde r) = \operatorname{sgn}^+(\varphi_i - \tilde r)\left[-\left(F(x,|\varphi|^2) + \frac{c^2}{4}\right)\varphi_i\right]\\ &\geq \operatorname{sgn}^+(\varphi_i - \tilde r)\left[C|\varphi|^{2\alpha} - \frac{c^2}{4}\right]\varphi_i \geq \operatorname{sgn}^+(\varphi_i - \tilde r)\left[C|\varphi_i|^{2\alpha} - \frac{c^2}{4}\right]\varphi_i\\ &\geq C_1 \operatorname{sgn}^+(\varphi_i - \tilde r)(\varphi_i - \tilde r)^{2\alpha+1} = C_1[(\varphi_i - \tilde r)^+]^{2\alpha+1}.\end{aligned} \tag{2.3}$$

Next we use the following result of Brézis (Lemma 2, p. 273, in [9]).

LEMMA 2.3 (see [9]). *Let $p \in (1,\infty)$. Assume that $u \in L^p_{loc}(\mathbf{R}^N)$ satisfies*

$$-\Delta u + |u|^{p-1}u \leq 0 \qquad \text{in } \mathcal{D}'(\mathbf{R}^N).$$

Then $u \leq 0$ a.e. on $\mathbf{R}^N$.

It follows from (2.3) that the function $u_i = (C_1)^{\frac{1}{2\alpha}}(\varphi_i - \tilde r)^+$ satisfies $-\Delta u_i + |u_i|^{2\alpha}u_i \leq 0$ in $\mathcal{D}'(\mathbf{R}^N)$. Since $u_i \in L^{2\alpha+1}_{loc}(\mathbf{R}^N)$, we may use Lemma 2.3 and thus get $u_i \leq 0$ a.e. in $\mathbf{R}^N$, that is, $\varphi_i \leq \tilde r$ a.e. in $\mathbf{R}^N$.

It is obvious that both φ and $-\varphi$ satisfy (2.2). Repeating the above argument for $-\varphi$, we infer that $-\varphi_i \le \tilde{r}$ a.e. on $\mathbf{R}^N$. Therefore we have $|\varphi_i| \le \tilde{r}$ a.e. on $\mathbf{R}^N$, $i = 1, 2$, which implies that $\varphi \in L^\infty(\mathbf{R}^N)$. Since $|\varphi| = |\psi|$, we have proved that $\psi \in L^\infty(\mathbf{R}^N)$.

Using (H1c) and (2.2) we infer that $\Delta\varphi \in L^\infty(B(x,2R)) \subset L^p(B(x,2R))$ for any $x \in \mathbf{R}^N$, $R > 0$, and $p \ge 1$. By standard elliptic estimates we obtain $\varphi \in W^{2,p}(B(x,R))$ for any $x \in \mathbf{R}^N$, $R > 0$, and $p \in (1,\infty)$. Thus $\psi = e^{\frac{icx_1}{2}}\varphi \in W^{2,p}_{loc}(\mathbf{R}^N)$ for any $p \in (1,\infty)$, and consequently ψ belongs to $C^{1,\alpha}_{loc}(\mathbf{R}^N)$ for any $\alpha \in [0,1)$ by the Sobolev embedding theorem.

(ii) Assume $F \in C^1(\mathbf{R}^N \times [0,\infty))$. Differentiating (1.7) with respect to x_k we get

$$-ic\psi_{x_1x_k} + \Delta\psi_{x_k} + \frac{\partial F}{\partial x_k}(x,|\psi|^2)\psi + 2\partial_{N+1}F(x,|\psi|^2)\left(\psi.\frac{\partial\psi}{\partial x_k}\right)\psi + F(x,|\psi|^2)\frac{\partial\psi}{\partial x_k} = 0 \tag{2.4}$$

in $\mathcal{D}'(\mathbf{R}^N)$. Hence $\Delta\psi_{x_k} \in L^p_{loc}(\mathbf{R}^N)$ for $1 \le p < \infty$. By standard elliptic regularity theory we get $\psi_{x_k} \in W^{2,p}_{loc}(\mathbf{R}^N)$ for $1 < p < \infty$, $1 \le k \le N$; therefore $\psi \in W^{3,p}_{loc}(\mathbf{R}^N)$ for $1 \le p < \infty$. If $F \in C^k(\mathbf{R}^N \times [0,\infty))$, we may differentiate (2.4) further and repeat the above arguments. After an easy induction, we get $\psi \in W^{k+2,p}_{loc}(\mathbf{R}^N)$ for any $p \in (1,\infty)$. □

LEMMA 2.4. *Assume that* (H1), (H3), (H4), (H5) *hold and* $u \in L^4_{loc}(\mathbf{R}^N,\mathbf{C})$ *satisfies* $|u(x)| \longrightarrow r_0$ *as* $|x| \longrightarrow \infty$ *and* $V(\cdot,|u|^2) \in L^1(\mathbf{R}^N)$.

Then $|u|^2 - r_0^2 \in L^2(\mathbf{R}^N)$.

Proof. Let R_0, η, M be as in (H5). From (H4) and the fact that $|u(x)| \longrightarrow r_0$ as $|x| \longrightarrow \infty$ it follows that there exists $R_1 > R_0$ such that

$$\partial_{N+1}F(x,r_0^2) < -\frac{L}{2} \quad \text{and} \quad |u(x)|^2 \in (r_0^2-\eta, r_0^2+\eta) \quad \text{for any } x \text{ satisfying } |x| \ge R_1.$$

For $(x,s) \in (\mathbf{R}^N \setminus B(0,R_1)) \times (r_0^2-\eta, r_0^2+\eta)$ we get, by Taylor's formula with respect to the $(N+1)$th variable,

$$V(x,s) = -(s-r_0^2)F(x,r_0^2) - \frac{1}{2}(s-r_0^2)^2\partial_{N+1}F(x,r_0^2) - \frac{1}{2}\int_{r_0^2}^{s}(s-\tau)^2\partial^2_{N+1}F(x,\tau)\,d\tau.$$

In particular, for $s = |u(x)|^2$ we obtain

$$-\frac{1}{2}(|u(x)|^2 - r_0^2)^2\partial_{N+1}F(x,r_0^2) = V(x,|u(x)|^2) + (|u(x)|^2 - r_0^2)F(x,r_0^2) + \frac{1}{2}\int_{r_0^2}^{|u(x)|^2}(|u(x)|^2-\tau)^2\partial^2_{N+1}F(x,\tau)\,d\tau. \tag{2.5}$$

For $x \in \mathbf{R}^N \setminus B(0,R_1)$ we get by (H5)

$$\left|\int_{r_0^2}^{|u(x)|^2}(|u(x)|^2-\tau)^2\partial^2_{N+1}F(x,\tau)\,d\tau\right| \le M\left|\int_{r_0^2}^{|u(x)|^2}(|u(x)|^2-\tau)^2\,d\tau\right| = \frac{M}{3}\big|(|u(x)|^2 - r_0^2)\big|^3.$$

It is clear that there exists $R_2 \geq R_1$ such that $\frac{M}{3}\big||u(x)|^2 - r_0^2\big| \leq \frac{L}{4}$ on $\mathbf{R}^N \setminus B(0,R_2)$. Using (H4) and (2.5) we infer that

$$\begin{aligned}
&\frac{L}{4}(|u(x)|^2 - r_0^2)^2 \leq -\frac{1}{2}(|u(x)|^2 - r_0^2)^2 \partial_{N+1}F(x,r_0^2)\\
&\leq V(x,|u(x)|^2) + (|u(x)|^2 - r_0^2)F(x,r_0^2) + \frac{1}{2}\cdot\frac{M}{3}\Big||u(x)|^2 - r_0^2\Big|^3\\
&\leq V(x,|u(x)|^2) + (|u(x)|^2 - r_0^2)F(x,r_0^2) + \frac{L}{8}\Big||u(x)|^2 - r_0^2\Big|^2 \qquad \text{on } \mathbf{R}^N \setminus B(0,R_2).
\end{aligned}$$

Consequently

$$\frac{L}{8}(|u(x)|^2 - r_0^2)^2 \leq V(x,|u(x)|^2) + (|u(x)|^2 - r_0^2)F(x,r_0^2) \text{ on } \mathbf{R}^N \setminus B(0,R_2). \tag{2.6}$$

Since $F(\cdot,r_0^2) \in L^1(\mathbf{R}^N)$ by (H3), $V(\cdot,|u|^2) \in L^1(\mathbf{R}^N)$, and $\big||u(x)|^2 - r_0^2\big| \leq \frac{3L}{4M}$ on $\mathbf{R}^N \setminus B(0,R_2)$, using (2.6) we get $(|u|^2 - r_0^2)^2 \in L^1(\mathbf{R}^N \setminus B(0,R_2))$. It is obvious that $(|u|^2 - r_0^2)^2 \in L^1(B(0,R_2))$ because $u \in L^4_{loc}(\mathbf{R}^N)$. Hence $(|u|^2 - r_0^2)^2 \in L^1(\mathbf{R}^N)$ and Lemma 2.4 is proved. ∎

PROPOSITION 2.5. *Assume that* (H1)–(H5) *hold and let ψ be a finite-energy traveling wave for* (1.1) *(in the sense of Definition* 2.1*) such that $F(\cdot,|\psi|^2)\psi \in L^1_{loc}(\mathbf{R}^N)$. Then the following hold:*

(i) *$\nabla\psi \in W^{1,p}(\mathbf{R}^N)$ for any $p \in [2,\infty)$.*

(ii) *Let $R_* \geq 0$ be such that $|\psi(x)| \geq \frac{r_0}{2}$ for $|x| \geq R_*$. There exists a real-valued function θ such that $\theta \in W^{2,p}_{loc}(\mathbf{R}^N \setminus \overline{B}(0,R_*))$ for any $p < \infty$, $\nabla\theta \in W^{1,p}(\mathbf{R}^N \setminus \overline{B}(0,R_*))$ for any $p \in [2,\infty)$, and*

$$\psi(x) = |\psi(x)|e^{i\theta(x)} \qquad \text{on } \mathbf{R}^N \setminus B(0,R_*).$$

Proof. (i) We already know by Proposition 2.2(i) and Lemma 2.4 that ψ is bounded, $\psi \in W^{2,p}_{loc}(\mathbf{R}^N)$ for any $p \in [1,\infty)$, and $|\psi|^2 - r_0^2 \in L^2(\mathbf{R}^N)$.

Let R_0, η, M be as in (H5). Choose $R_1 > R_0$ such that $|\psi|^2(x) \in (r_0^2 - \eta, r_0^2 + \eta)$ for $x \in \mathbf{R}^N \setminus B(0,R_1)$.

By using Taylor's formula with respect to the last variable for the function F we get

$$F(x,s) = F(x,r_0^2) + (s - r_0^2)\partial_{N+1}F(x,r_0^2) + \int_{r_0^2}^{s}(s-\tau)\partial^2_{N+1}F(x,\tau)\,d\tau \tag{2.7}$$

if $(x,s) \in (\mathbf{R}^N \setminus \overline{B}(0,R_0)) \times (r_0^2 - \eta, r_0^2 + \eta)$; hence

$$\begin{aligned}
F(x,|\psi|^2(x))\psi(x) &= F(x,r_0^2)\psi(x) + (|\psi|^2(x) - r_0^2)\partial_{N+1}F(x,r_0^2)\psi(x)\\
&\quad + \psi(x)\int_{r_0^2}^{|\psi|^2(x)}(|\psi|^2(x) - \tau)\partial^2_{N+1}F(x,\tau)\,d\tau \quad \text{for any } |x| \geq R_1.
\end{aligned} \tag{2.8}$$

We analyze the three terms on the right-hand side of (2.8). Assumptions (H1) and (H3) imply $F(\cdot,r_0^2) \in L^1 \cap L^\infty(\mathbf{R}^N)$. Since $\psi \in L^\infty(\mathbf{R}^N)$, it follows that $F(\cdot,r_0^2)\psi \in L^1 \cap L^\infty(\mathbf{R}^N)$.

We may write $(|\psi|^2 - r_0^2)\partial_{N+1}F(\cdot,r_0^2)\psi = -L(|\psi|^2 - r_0^2)\psi + (|\psi|^2 - r_0^2)(L + \partial_{N+1}F(\cdot,r_0^2))\psi$. We know that $\psi \in L^\infty(\mathbf{R}^N)$, $|\psi|^2 - r_0^2 \in L^2 \cap L^\infty(\mathbf{R}^N)$ and by (H4)

we have $L + \partial_{N+1}F(\cdot, r_0^2) \in L^{p_0} \cap L^\infty(\mathbf{R}^N)$ for some $p_0 \in [1,2]$, so we infer that $(|\psi|^2 - r_0^2)\partial_{N+1}F(\cdot, r_0^2)\psi \in L^2 \cap L^\infty(\mathbf{R}^N)$.

As in the proof of Lemma 2.4, for $x \in \mathbf{R}^N \setminus B(0,R_1)$ we have

$$\left| \int_{r_0^2}^{|\psi|^2(x)} (|\psi|^2(x) - \tau)\partial_{N+1}^2 F(x,\tau)\, d\tau \right| \leq M \left| \int_{r_0^2}^{|\psi|^2(x)} \Big| |\psi|^2(x) - \tau \Big| d\tau \right| = \frac{M}{2}(|\psi|^2(x) - r_0^2)^2. \tag{2.9}$$

Consequently the function $x \longmapsto \int_{r_0^2}^{|\psi|^2(x)} (|\psi|^2(x) - \tau)\partial_{N+1}^2 F(x,\tau)\, d\tau$ belongs to $L^1 \cap L^\infty(\mathbf{R}^N \setminus B(0,R_1))$.

Summing up, we have proved that $F(\cdot, |\psi|^2)\psi \in L^2 \cap L^\infty(\mathbf{R}^N \setminus B(0,R_1))$. From (H1) and the fact that ψ is bounded on $\mathbf{R}^N$ it follows that $F(\cdot, |\psi|^2)\psi$ is bounded on $B(0,R_1)$, and hence $F(\cdot, |\psi|^2)\psi \in L^2 \cap L^\infty(\mathbf{R}^N)$.

We have $\frac{\partial \psi}{\partial x_k} \in L^2(\mathbf{R}^N)$ because ψ has finite energy. Coming back to (1.7), we get

$$\Delta\psi = ic\frac{\partial \psi}{\partial x_1} - F(\cdot, |\psi|^2)\psi \in L^2(\mathbf{R}^N).$$

It is well known that $\Delta\psi \in L^p(\mathbf{R}^N)$ with $1 < p < \infty$ implies $\frac{\partial^2 \psi}{\partial x_j \partial x_k} \in L^p(\mathbf{R}^N)$ for any $j,k \in \{1,\dots,N\}$ (this follows, e.g., from the fact that $\frac{\xi_j \xi_k}{|\xi|^2}$ is a Fourier multiplier on $L^p(\mathbf{R}^N)$ if $1 < p < \infty$; see Theorem 3, p. 96, in [27]). Therefore all second derivatives of ψ are in $L^2(\mathbf{R}^N)$, so that $\frac{\partial \psi}{\partial x_k} \in H^1(\mathbf{R}^N) = W^{1,2}(\mathbf{R}^N)$ for $k = 1,\dots,N$.

The rest of the proof is an easy bootstrap argument. Assume that $\nabla\psi \in W^{1,p}(\mathbf{R}^N)$ for some $p \geq 2$. In case $p < N$, it follows from the Sobolev embedding theorem that $\nabla\psi \in L^{p^*}(\mathbf{R}^N)$, where $\frac{1}{p^*} = \frac{1}{p} - \frac{1}{N}$. From (1.7) we have $\Delta\psi = ic\frac{\partial \psi}{\partial x_1} - F(\cdot, |\psi|^2)\psi \in L^{p^*}(\mathbf{R}^N)$ and infer as previously that $\nabla\psi \in W^{1,p^*}(\mathbf{R}^N)$. Repeating this argument if necessary, after a finite number of steps we get $\nabla\psi \in W^{1,q}(\mathbf{R}^N)$ for some $q \geq N$. Then by Sobolev embedding we get $\nabla\psi \in L^r(\mathbf{R}^N)$ for any $r \in [q,\infty)$. From (1.7) we obtain $\Delta\psi \in L^p(\mathbf{R}^N)$ for $p \in [2,\infty)$ and infer that $\nabla\psi \in W^{1,p}(\mathbf{R}^N)$ for any $p \in [2,\infty)$.

(ii) Take $R_* > 0$ such that $|\psi(x)| \geq \frac{r_0}{2}$ on $\mathbf{R}^N \setminus B(0,R_*)$ and denote $\tilde{\psi}(x) = \frac{\psi(x)}{|\psi(x)|}$. It is then standard to prove that $\tilde{\psi} \in W^{2,p}_{loc}(\mathbf{R}^N \setminus \overline{B}(0,R_*))$ for $p \in [1,\infty)$ and $\nabla\tilde{\psi} \in W^{1,p}(\mathbf{R}^N \setminus \overline{B}(0,R_*))$ for any $p \in [2,\infty)$ (see, e.g., Lemma C1, p. 66, in [11]).

Let us consider first the case $N \geq 3$. For $R_* \leq R_1 < R_2$, the domain $\Omega_{R_1,R_2} = B(0,R_2) \setminus \overline{B}(0,R_1)$ is simply connected in $\mathbf{R}^N$. It follows from Theorem 3, p. 38, in [11] that there exists a real-valued function $\theta_{R_1,R_2} \in W^{2,p}(\Omega_{R_1,R_2})$ $(1 < p < \infty)$ such that $\tilde{\psi} = e^{i\theta_{R_1,R_2}}$ on Ω_{R_1,R_2}. If $R_* \leq R_1 < R_2$, $R_* \leq R_3 < R_4$, and $(R_1,R_2) \cap (R_3,R_4) \neq \emptyset$, then $\tilde{\psi} = e^{i\theta_{R_1,R_2}} = e^{i\theta_{R_3,R_4}}$ on $\Omega_{R_1,R_2} \cap \Omega_{R_3,R_4}$, and thus $\theta_{R_3,R_4} - \theta_{R_1,R_2} \in 2\pi\mathbf{Z}$ on $\Omega_{R_1,R_2} \cap \Omega_{R_3,R_4}$. Since functions in $W^{s,p}(\Omega_{R_1,R_2} \cap \Omega_{R_3,R_4})$ with values in $\mathbf{Z}$ are constant when $sp \geq 1$ (see Theorem B1, p. 65, in [11]), there exists $k \in \mathbf{Z}$ such that $\theta_{R_3,R_4} - \theta_{R_1,R_2} = 2\pi k$ on $\Omega_{R_1,R_2} \cap \Omega_{R_3,R_4}$. Let $(R_n)_{n \geq 1}$ be an increasing sequence such that $R_* < R_1$ and $R_n \longrightarrow \infty$. Let $k_n \in \mathbf{Z}$ be such that $\theta_{R_*,R_n} = \theta_{R_*,R_1} + 2\pi k_n$ on Ω_{R_*,R_1}. Define $\theta(x) = \theta_{R_*,R_n}(x) - 2\pi k_n$ for $x \in \Omega_{R_*,R_n}$. It is clear that θ is well-defined on $\mathbf{R}^N \setminus \overline{B}(0,R_*)$, $\tilde{\psi} = e^{i\theta}$, and $\theta \in W^{2,p}_{loc}(\mathbf{R}^N \setminus \overline{B}(0,R_*))$ for any $p \in [1,\infty)$.

Next we consider the case $N = 2$. Since ψ is C^1 and $|\psi| \geq \frac{r_0}{2}$ on $\mathbf{R}^2 \setminus \overline{B}(0,R_*)$, the topological degree $deg(\psi, \partial B(0,R))$ is well-defined for any $R \geq R_*$ and does not depend on R. It is well known that ψ admits a C^1 lifting θ (i.e., $\psi = |\psi|e^{i\theta}$)

on $\mathbf{R}^2 \setminus \overline{B}(0,R_*)$ if and only if $deg(\psi, \partial B(0,R)) = 0$ for $R \geq R_*$. Denoting by $\tau = (-\sin\zeta, \cos\zeta)$ the unit tangent vector at $\partial B(0,R)$ at a point $Re^{i\zeta}$, we get

$$\begin{aligned} |deg(\psi, \partial B(0,R))| &= \left| \frac{1}{2i\pi} \int_0^{2\pi} \frac{\frac{\partial}{\partial\zeta}(\psi(Re^{i\zeta}))}{\psi(Re^{i\zeta})}\, d\zeta \right| = \left| \frac{R}{2i\pi} \int_0^{2\pi} \frac{\frac{\partial\psi}{\partial\tau}(Re^{i\zeta})}{\psi(Re^{i\zeta})}\, d\zeta \right| \\ &\leq \frac{R}{2\pi} \int_0^{2\pi} \frac{2}{r_0} |\nabla\psi(Re^{i\zeta})|\, d\zeta \leq \frac{R}{\pi r_0}\sqrt{2\pi} \left(\int_0^{2\pi} |\nabla\psi(Re^{i\zeta})|^2\, d\zeta \right)^{\frac{1}{2}}. \end{aligned} \tag{2.10}$$

On the other hand,

$$\int_{\mathbf{R}^2\setminus\overline{B}(0,R_*)} |\nabla\psi(x)|^2\, dx = \int_{R_*}^{\infty} R \int_0^{2\pi} |\nabla\psi(Re^{i\zeta})|^2\, d\zeta\, dR.$$

We have $\int_{\mathbf{R}^2\setminus\overline{B}(0,R_*)} |\nabla\psi(x)|^2\, dx < \infty$ (because ψ has finite energy) and infer that there exists $R_1 > R_*$ such that $R_1 \int_0^{2\pi} |\nabla\psi(R_1 e^{i\zeta})|^2\, d\zeta < \frac{\pi r_0^2}{8} \frac{1}{R_1}$. From (2.10) we get

$$|deg(\psi, \partial B(0,R_1))| < \frac{R_1}{\pi r_0}\sqrt{2\pi} \left(\frac{\pi r_0^2}{8} \frac{1}{R_1^2} \right)^{\frac{1}{2}} = \frac{1}{2}.$$

Since the topological degree is an integer, we have necessarily $deg(\psi, \partial B(0,R_1)) = 0$. Consequently $deg(\psi, \partial B(0,R)) = 0$ for any $R \geq R_*$ and ψ admits a C^1 lifting θ. In fact, $\theta \in W^{2,p}_{loc}(\mathbf{R}^2 \setminus \overline{B}(0,R_*))$ because $\psi \in W^{2,p}_{loc}(\mathbf{R}^2 \setminus \overline{B}(0,R_*))$ (see Theorem 3, p. 38, in [11]).

If $N = 1$, the existence of a lifting $\psi = |\psi| e^{i\theta}$ follows immediately from Theorem 1, p. 27, in [11].

Finally, it is easy to see that $|\frac{\partial\tilde{\psi}}{\partial x_j}| = |\frac{\partial\theta}{\partial x_j}|$ and $|\frac{\partial^2\tilde{\psi}}{\partial x_j \partial x_k}|^2 = |\frac{\partial^2\theta}{\partial x_j \partial x_k}|^2 + |\frac{\partial\theta}{\partial x_j}|^2 |\frac{\partial\theta}{\partial x_k}|^2 \geq |\frac{\partial^2\theta}{\partial x_j \partial x_k}|^2$, and (i) implies $\nabla\theta \in W^{1,p}(\mathbf{R}^N \setminus \overline{B}(0,R_*))$ for any $p \in [2,\infty)$. ∎

3. An integral identity. The main result of this section is given by the next theorem.

THEOREM 3.1. *Assume that* (H1)–(H5) *hold. Let* $\psi = \psi_1 + i\psi_2$ *be a finite-energy traveling wave for* (1.1) *such that* $F(\cdot, |\psi|^2) \in L^1_{loc}(\mathbf{R}^N)$. *Let* R_* *be sufficiently big, so that* $|\psi| \geq \frac{r_0}{2}$ *on* $\mathbf{R}^N \setminus B(0,R_*)$, *and let* θ *be the lifting given by Proposition* 2.5 (ii). *Let* $\chi \in C^\infty(\mathbf{R}^N)$ *be a cut-off function such that* $\chi = 0$ *on* $B(0,2R_*)$ *and* $\chi = 1$ *on* $\mathbf{R}^N \setminus B(0,3R_*)$. *Then the following hold:*

(i) *The functions* $F(\cdot,|\psi|^2)|\psi|^2 + \frac{v_s^2}{2}(|\psi|^2 - r_0^2)$ *and* $G_j = \psi_1 \frac{\partial\psi_2}{\partial x_j} - \psi_2 \frac{\partial\psi_1}{\partial x_j} - r_0^2 \frac{\partial}{\partial x_j}(\chi\theta)$, $j = 1,\dots,N$, *belong to* $L^1 \cap L^\infty(\mathbf{R}^N)$. *(We always extend* $\chi\theta$ *by zero on* $\overline{B}(0,R_*)$.*)*

(ii) *If* $N \geq 2$ *and* $c^2 > v_s^2$, *we have the identity*

$$\begin{aligned} &\int_{\mathbf{R}^N} |\nabla\psi|^2 - F(x,|\psi|^2)|\psi|^2 - \frac{v_s^2}{2}(|\psi|^2 - r_0^2)\, dx \\ &= c\left(1 - \frac{v_s^2}{c^2}\right) \int_{\mathbf{R}^N} \psi_1 \frac{\partial\psi_2}{\partial x_1} - \psi_2 \frac{\partial\psi_1}{\partial x_1} - r_0^2 \frac{\partial}{\partial x_1}(\chi\theta)\, dx. \end{aligned} \tag{3.1}$$

(iii) *Identity* (3.1) *holds if* $c^2 = v_s^2$ *and either*

- $N = 2$ *or*
- $N \geq 3$, *and we assume in addition that* $\psi_1 \frac{\partial\psi_2}{\partial x_1} - \psi_2 \frac{\partial\psi_1}{\partial x_1} \in L^{\frac{2N-1}{2N-3}}(\mathbf{R}^N)$.

Proof. (i) Let R_0, η, M be as in (H5) and take $R_1 > R_0$ such that $|\psi|^2(x) \in (r_0^2 - \eta, r_0^2 + \eta)$ for $x \in \mathbf{R}^N \setminus B(0, R_1)$. Using (2.7) and the fact that $v_s^2 = 2Lr_0^2$ we get

$$\begin{aligned} &F(x, |\psi|^2(x))|\psi|^2(x) + \frac{v_s^2}{2}(|\psi|^2(x) - r_0^2) = F(x, r_0^2)|\psi|^2(x) \\ &+ (|\psi|^2(x) - r_0^2)[\partial_{N+1}F(x, r_0^2) + L]|\psi|^2(x) - L\left(|\psi|^2(x) - r_0^2\right)^2 \\ &+ |\psi|^2(x)\int_{r_0^2}^{|\psi|^2(x)} (|\psi|^2(x) - \tau)\partial_{N+1}^2 F(x, \tau)\, d\tau \qquad \text{for any } |x| \geq R_1. \end{aligned} \tag{3.2}$$

Since $\psi \in L^\infty(\mathbf{R}^N)$ by Proposition 2.2 (i) and $F(\cdot, r_0^2) \in L^1 \cap L^\infty(\mathbf{R}^N)$ by (H1) and (H3), we infer that $F(\cdot, r_0^2)|\psi|^2 \in L^1 \cap L^\infty(\mathbf{R}^N)$.

We have $\psi \in L^\infty(\mathbf{R}^N)$, $\partial_{N+1}F(\cdot, r_0^2) + L \in L^{p_0} \cap L^\infty(\mathbf{R}^N)$ by (H4) and $|\psi|^2 - r_0^2 \in L^2 \cap L^\infty(\mathbf{R}^N)$ by Lemma 2.4; hence $(|\psi|^2 - r_0^2)[\partial_{N+1}F(\cdot, r_0^2) + L]|\psi|^2 \in L^1 \cap L^\infty(\mathbf{R}^N)$.

From Proposition 2.2 (i), Lemma 2.4, and (2.9) it follows that the last two terms on the right-hand side of (3.2) are in $L^1 \cap L^\infty(\mathbf{R}^N \setminus \overline{B}(0, R_1))$. Hence $F(\cdot, |\psi|^2)|\psi|^2 + \frac{v_s^2}{2}(|\psi|^2 - r_0^2) \in L^1 \cap L^\infty(\mathbf{R}^N \setminus \overline{B}(0, R_1))$. Clearly, the function $F(\cdot, |\psi|^2)|\psi|^2 + \frac{v_s^2}{2}(|\psi|^2 - r_0^2)$ is bounded on $\overline{B}(0, R_1)$; therefore this function belongs to $L^1 \cap L^\infty(\mathbf{R}^N)$.

Since $\psi_1 = |\psi| \cos\theta$ and $\psi_2 = |\psi| \sin\theta$, a straightforward computation gives

$$\psi_1 \frac{\partial \psi_2}{\partial x_j} - \psi_2 \frac{\partial \psi_1}{\partial x_j} = (\psi_1^2 + \psi_2^2)\frac{\partial \theta}{\partial x_j} \qquad \text{on } \mathbf{R}^N \setminus \overline{B}(0, R_*). \tag{3.3}$$

Therefore

$$\psi_1 \frac{\partial \psi_2}{\partial x_j} - \psi_2 \frac{\partial \psi_1}{\partial x_j} - r_0^2 \frac{\partial}{\partial x_j}(\chi\theta) = (|\psi|^2 - r_0^2)\frac{\partial \theta}{\partial x_j} \qquad \text{on } \mathbf{R}^N \setminus \overline{B}(0, 3R_*). \tag{3.4}$$

From Lemma 2.4, Proposition 2.5 (ii), and the Sobolev embedding theorem we have $|\psi|^2 - r_0^2 \in L^2 \cap L^\infty(\mathbf{R}^N)$ and $\frac{\partial \theta}{\partial x_j} \in L^2 \cap L^\infty(\mathbf{R}^N \setminus \overline{B}(0, R_*))$, respectively. Identity (3.4) implies $G_j \in L^1 \cap L^\infty(\mathbf{R}^N \setminus \overline{B}(0, 3R_*))$. Since G_j is continuous on $\mathbf{R}^N$, we conclude that $G_j \in L^1 \cap L^\infty(\mathbf{R}^N)$.

(ii) Equation (1.7) is equivalent to the system

$$c\frac{\partial \psi_2}{\partial x_1} + \Delta\psi_1 + F(x, |\psi|^2)\psi_1 = 0 \qquad \text{in } \mathcal{D}'(\mathbf{R}^N), \tag{3.5}$$

$$-c\frac{\partial \psi_1}{\partial x_1} + \Delta\psi_2 + F(x, |\psi|^2)\psi_2 = 0 \qquad \text{in } \mathcal{D}'(\mathbf{R}^N). \tag{3.6}$$

In view of Proposition 2.2 (i), equalities (3.5) and (3.6) hold in $L^p_{loc}(\mathbf{R}^N)$ for $1 \leq p < \infty$. Multiplying (3.5) by ψ_2 and (3.6) by ψ_1 and then subtracting the resulting equalities gives us

$$\frac{c}{2}\frac{\partial}{\partial x_1}(|\psi|^2 - r_0^2) = \operatorname{div}(\psi_1 \nabla\psi_2 - \psi_2 \nabla\psi_1). \tag{3.7}$$

We multiply (3.5) by ψ_1 and (3.6) by ψ_2, then add the corresponding equalities to obtain

$$|\nabla\psi_1|^2 + |\nabla\psi_2|^2 - F(x, |\psi|^2)|\psi|^2 - c\left(\psi_1 \frac{\partial \psi_2}{\partial x_1} - \psi_2 \frac{\partial \psi_1}{\partial x_1}\right) = \frac{1}{2}\Delta(|\psi|^2 - r_0^2). \tag{3.8}$$

From (3.7) and (3.8) we get

$$\frac{c}{2}\frac{\partial}{\partial x_1}(|\psi|^2-r_0^2)=\operatorname{div}(\psi_1\nabla\psi_2-\psi_2\nabla\psi_1-r_0^2\nabla(\chi\theta))+r_0^2\Delta(\chi\theta), \tag{3.9}$$

respectively,

(3.10)

$$\begin{aligned}\frac{1}{2}\Delta(|\psi|^2-r_0^2)-\frac{v_s^2}{2}(|\psi|^2-r_0^2)&=|\nabla\psi_1|^2+|\nabla\psi_2|^2-F(x,|\psi|^2)|\psi|^2-\frac{v_s^2}{2}(|\psi|^2-r_0^2)\\&\quad-c\left(\psi_1\frac{\partial\psi_2}{\partial x_1}-\psi_2\frac{\partial\psi_1}{\partial x_1}-r_0^2\frac{\partial}{\partial x_1}(\chi\theta)\right)-cr_0^2\frac{\partial}{\partial x_1}(\chi\theta).\end{aligned}$$

Since $\psi\in W^{2,p}_{loc}(\mathbf{R}^N)$, equalities (3.7)–(3.10) hold in $L^p_{loc}(\mathbf{R}^N)$ for $1\le p<\infty$. We denote

$$H=|\nabla\psi_1|^2+|\nabla\psi_2|^2-F(x,|\psi|^2)|\psi|^2-\frac{v_s^2}{2}(|\psi|^2-r_0^2)-c\left(\psi_1\frac{\partial\psi_2}{\partial x_1}-\psi_2\frac{\partial\psi_1}{\partial x_1}-r_0^2\frac{\partial}{\partial x_1}(\chi\theta)\right).$$

We take the derivative of (3.9) with respect to x_1 (in $\mathcal{D}'(\mathbf{R}^N)$), multiply it by c, and then take the Laplacian of (3.10) (in $\mathcal{D}'(\mathbf{R}^N)$). Summing up the resulting equalities, we obtain

$$\frac{1}{2}\Big(\Delta^2-v_s^2\Delta+c^2\frac{\partial^2}{\partial x_1^2}\Big)(|\psi|^2-r_0^2)=\Delta H+c\frac{\partial}{\partial x_1}(\operatorname{div}(G))\qquad\text{in }\mathcal{D}'(\mathbf{R}^N). \tag{3.11}$$

From (i) we have $H,G_1,\dots,G_N\in L^1\cap L^\infty(\mathbf{R}^N)$, and we know from Lemma 2.4 that $|\psi|^2-r_0^2\in L^2\cap L^\infty(\mathbf{R}^N)$. Therefore $H,G_1,\dots,G_N,|\psi|^2-r_0^2\in\mathcal{S}'(\mathbf{R}^N)$, and we infer that, in fact, equality (3.11) holds in $\mathcal{S}'(\mathbf{R}^N)$. Taking the Fourier transform of (3.11) we get

$$\frac{1}{2}(|\xi|^4+v_s^2|\xi|^2-c^2\xi_1^2)\mathcal{F}(|\psi|^2-r_0^2)=-|\xi|^2\widehat{H}-c\sum_{k=1}^N\xi_1\xi_k\widehat{G_k}\quad\text{in }\mathcal{S}'(\mathbf{R}^N). \tag{3.12}$$

We have $\widehat{H}$, $\widehat{G_k}\in L^\infty\cap C^0(\mathbf{R}^N)$ because H, $G_k\in L^1(\mathbf{R}^N)$. Thus the right-hand side of (3.12) is a continuous function on $\mathbf{R}^N$. Since $|\psi|^2-r_0^2\in L^2(\mathbf{R}^N)$, we have $\mathcal{F}(|\psi|^2-r_0^2)\in L^2(\mathbf{R}^N)$ and infer that the left-hand side of (3.12) belongs to $L^2_{loc}(\mathbf{R}^N)$ and (3.12) holds a.e. on $\mathbf{R}^N$.

We denote

$$\Gamma=\{\xi\in\mathbf{R}^N\mid|\xi|^4+v_s^2|\xi|^2-c^2\xi_1^2=0\}.$$

If $c^2\le v_s^2$, we have $\Gamma=\{0\}$. If $c^2>v_s^2$, it is easy to see that Γ is a nontrivial submanifold of $\mathbf{R}^N$. In the latter case, we claim that

$$|\xi|^2\widehat{H}(\xi)+c\sum_{k=1}^N\xi_1\xi_k\widehat{G_k}(\xi)=0\qquad\text{for any }\xi\in\Gamma. \tag{3.13}$$

To prove this claim, we argue by contradiction and suppose that there exists $\xi^0\in\Gamma$ such that $|\xi^0|^2\widehat{H}(\xi^0)+c\sum_{k=1}^N\xi_1^0\xi_k^0\widehat{G_k}(\xi^0)\neq0$. By continuity, there exist

$m > 0$ and a neighborhood U of ξ_0 such that $\big|\,|\xi|^2\widehat{H} + c\sum_{k=1}^N \xi_1\xi_k\widehat{G_k}\big| \geq m$ on U. From (3.12) we infer that

$$|\mathcal{F}(|\psi|^2 - r_0^2)(\xi)| \geq \frac{2m}{|\,|\xi|^4 + v_s^2|\xi|^2 - c^2\xi_1^2|} \qquad \text{a.e. on } U \setminus \Gamma.$$

Since 0 and $(\sqrt{c^2 - v_s^2}, 0, \dots, 0)$ are not isolated points of Γ, we may assume that $\xi^0 \neq 0$ and $\xi^0 \neq (\sqrt{c^2 - v_s^2}, 0, \dots, 0)$. A straightforward computation (details can be found in [17, p. 98] in the case $v_s^2 = 2$; the general case is similar) shows that

$$\int_{U\setminus\Gamma} \frac{1}{|\,|\xi|^4 + v_s^2|\xi|^2 - c^2\xi_1^2|^2}\, d\xi = \infty;$$

consequently $\int_{U\setminus\Gamma} |\mathcal{F}(|\psi|^2 - r_0^2)(\xi)|^2\, d\xi = \infty$. But this is in contradiction with $\mathcal{F}(|\psi|^2 - r_0^2) \in L^2(\mathbf{R}^N)$ and the claim is proved.

It is not difficult to see that $\Gamma = \{(\xi_1, \xi') \in \mathbf{R}\times\mathbf{R}^{N-1} \mid |\xi'|^2 = \frac{1}{2}(-v_s^2 - 2\xi_1^2 + \sqrt{v_s^4 + 4c^2\xi_1^2})\}$. Let $f(t) = \sqrt{\frac{1}{2}\big(-v_s^2 - 2t^2 + \sqrt{v_s^4 + 4c^2t^2}\big)}$. The function f is well-defined for $t \in [-\sqrt{c^2 - v_s^2}, \sqrt{c^2 - v_s^2}]$, $f(0) = 0$, and $\lim_{t\to 0}\frac{f^2(t)}{t^2} = -1 + \frac{c^2}{v_s^2}$. Fix $j \in \{2, \dots, N\}$. For $t \in (0, \sqrt{c^2 - v_s^2}]$, let $\xi(t) = (t, 0, \dots, 0, f(t), 0, \dots, 0)$ and $\tilde{\xi}(t) = (t, 0, \dots, 0, -f(t), 0, \dots, 0)$, where $f(t)$, respectively, $-f(t)$, stand at the jth place. It is obvious that $\xi(t), \tilde{\xi}(t) \in \Gamma$. From (3.13) we obtain, respectively,

$$(t^2 + f^2(t))\widehat{H}(\xi(t)) + ct^2\widehat{G_1}(\xi(t)) + ctf(t)\widehat{G_j}(\xi(t)) = 0, \tag{3.14}$$

$$(t^2 + f^2(t))\widehat{H}(\tilde{\xi}(t)) + ct^2\widehat{G_1}(\tilde{\xi}(t)) - ctf(t)\widehat{G_j}(\tilde{\xi}(t)) = 0. \tag{3.15}$$

We multiply (3.14) and (3.15) by $\frac{1}{t^2}$, then pass to the limit as $t \downarrow 0$ to obtain, respectively,

$$\frac{c^2}{v_s^2}\widehat{H}(0) + c\widehat{G_1}(0) + c\sqrt{-1 + \frac{c^2}{v_s^2}}\widehat{G_j}(0) = 0, \tag{3.16}$$

$$\frac{c^2}{v_s^2}\widehat{H}(0) + c\widehat{G_1}(0) - c\sqrt{-1 + \frac{c^2}{v_s^2}}\widehat{G_j}(0) = 0. \tag{3.17}$$

From (3.16) and (3.17) we infer that $\frac{c^2}{v_s^2}\widehat{H}(0) + c\widehat{G_1}(0) = 0$ and $\widehat{G_j}(0) = 0$, that is, $\int_{\mathbf{R}^N} H(x) + \frac{v_s^2}{c}G_1(x)\, dx = 0$ and $\int_{\mathbf{R}^N} G_j(x)\, dx = 0$. The first of these integral identities is exactly (3.1) and the latter can be written as

$$\int_{\mathbf{R}^N} \psi_1\frac{\partial\psi_2}{\partial x_j} - \psi_2\frac{\partial\psi_1}{\partial x_j} - r_0^2\frac{\partial}{\partial x_j}(\chi\theta)\, dx = 0 \qquad \text{for } j = 2, \dots, N. \tag{3.18}$$

(iii) Assume that $c^2 = v_s^2$. Then (3.1) is equivalent to $\widehat{H}(0) + c\widehat{G_1}(0) = 0$. Denoting $\xi = (\xi_1, \xi')$, where $\xi' = (\xi_2, \dots, \xi_N)$, identity (3.12) implies

$$\begin{aligned}\mathcal{F}(|\psi|^2 - r_0^2)(\xi) &= -2\frac{\xi_1^2}{|\xi|^4 + c^2|\xi'|^2}(\widehat{H}(\xi) + c\widehat{G_1}(\xi)) \\ &- 2c\sum_{k=2}^N \frac{\xi_1\xi_k}{|\xi|^4 + c^2|\xi'|^2}\widehat{G_k}(\xi) - 2\frac{|\xi'|^2}{|\xi|^4 + c^2|\xi'|^2}\widehat{H}(\xi) \qquad \text{a.e. } \xi \in \mathbf{R}^N.\end{aligned} \tag{3.19}$$

For $\varepsilon \in (0,1]$, we denote $\Omega_\varepsilon = \{(\xi_1,\xi') \in \mathbf{R}\times\mathbf{R}^{N-1} \mid \xi_1 \in [0,\varepsilon],\ 0 \le |\xi'| \le \xi_1\}$. We will use the following lemma.

LEMMA 3.2. *Let $N \ge 2$ and $k \in \{2,\dots,N\}$.*

(i) *The function $\xi \longmapsto \frac{\xi_1^2}{\xi_1^4+c^2|\xi'|^2}$ belongs to $L^p(\Omega_\varepsilon)$ if and only if $p < N-\frac{1}{2}$.*

(ii) *The function $\xi \longmapsto \frac{\xi_1\xi_k}{\xi_1^4+c^2|\xi'|^2}$ belongs to $L^p(\Omega_\varepsilon)$ for any $p \in [1,2N-1)$.*

Proof of Lemma 3.2. (i) Using Fubini's theorem for positive functions, then passing to spherical coordinates in $\mathbf{R}^{N-1}$ and making the change of variables $r = \xi_1^2 t$ we get

(3.20)
$$\begin{aligned}
&\int_{\Omega_\varepsilon}\left(\frac{\xi_1^2}{\xi_1^4+c^2|\xi'|^2}\right)^p d\xi = \int_0^\varepsilon \xi_1^{2p}\int_{\{|\xi'|\le\xi_1\}}\frac{1}{(\xi_1^4+c^2|\xi'|^2)^p}\,d\xi'\,d\xi_1\\
&= \int_0^\varepsilon \xi_1^{2p}|S^{N-2}|\int_0^{\xi_1}\frac{r^{N-2}}{(\xi_1^4+c^2r^2)^p}\,dr\,d\xi_1\\
&= |S^{N-2}|\int_0^\varepsilon \xi_1^{2p}\int_0^{\frac{1}{\xi_1}}\frac{(\xi_1^2t)^{N-2}}{(\xi_1^4+c^2\xi_1^4t^2)^p}\xi_1^2\,dt\,d\xi_1 \quad \text{(change of variables } r=\xi_1^2t)\\
&= |S^{N-2}|\int_0^\varepsilon \xi_1^{2(N-1-p)}\int_0^{\frac{1}{\xi_1}}\frac{t^{N-2}}{(1+c^2t^2)^p}\,dt\,d\xi_1.
\end{aligned}$$

Assume that $p < N-\frac{1}{2}$. Obviously $\frac{t^{N-2}}{(1+c^2t^2)^p} \le 1$ for $t\in[0,1]$ and $\frac{t^2}{1+c^2t^2} \le \frac{1}{c^2}$, and thus we have

$$\int_0^{\frac{1}{\xi_1}}\frac{t^{N-2}}{(1+c^2t^2)^p}\,dt \le 1+\frac{1}{c^{2p}}\int_1^{\frac{1}{\xi_1}} t^{N-2p-2}\,dt = \begin{cases} C_1+\frac{C_2}{\xi_1^{N-2p-1}} & \text{if } p\ne\frac{N-1}{2},\\ C_3+C_4\ln\xi_1 & \text{if } p=\frac{N-1}{2},\end{cases}$$

where C_j are some positive constants. This estimate implies that the right-hand side of (3.20) is finite if $p < N-\frac{1}{2}$.

If $p \ge N-\frac{1}{2}$, denote $c_p = \int_0^1 \frac{t^{N-2}}{(1+c^2t^2)^p}\,dt > 0$. Since $\frac{1}{\xi_1} > 1$ for $\xi_1 \in (0,\varepsilon)$, the right-hand side of (3.20) is greater than $|S^{N-2}|c_p\int_0^\varepsilon \xi_1^{2(N-1-p)}\,d\xi_1 = \infty$.

(ii) Proceeding as above, we have

(3.21)
$$\begin{aligned}
&\int_{\Omega_\varepsilon}\left|\frac{\xi_1\xi_k}{\xi_1^4+c^2|\xi'|^2}\right|^p d\xi \le \int_{\Omega_\varepsilon}\frac{\xi_1^p|\xi'|^p}{(\xi_1^4+c^2|\xi'|^2)^p}\,d\xi = \int_0^\varepsilon \xi_1^p|S^{N-2}|\int_0^{\xi_1}\frac{r^{p+N-2}}{(\xi_1^4+c^2r^2)^p}\,dr\,d\xi_1\\
&= |S^{N-2}|\int_0^\varepsilon \xi_1^p\int_0^{\frac{1}{\xi_1}}\frac{(\xi_1^2t)^{p+N-2}}{(\xi_1^4+c^2\xi_1^4t^2)^p}\xi_1^2\,dt\,d\xi_1 \qquad \text{(change of variables } r=\xi_1^2t)\\
&= |S^{N-2}|\int_0^\varepsilon \xi_1^{2N-p-2}\int_0^{\frac{1}{\xi_1}}\frac{t^{p+N-2}}{(1+c^2t^2)^p}\,dt\,d\xi_1.
\end{aligned}$$

As previously,

$$\int_0^{\frac{1}{\xi_1}}\frac{t^{p+N-2}}{(1+c^2t^2)^p}\,dt < \frac{1}{c^{2p}}\int_0^{\frac{1}{\xi_1}} t^{N-p-2}\,dt = \frac{1}{c^{2p}(N-p-1)}\frac{1}{\xi_1^{N-p-1}} \quad \text{if } N-p-1>0.$$

Therefore in the case $p < N-1$, the right-hand side of (3.21) is less than $C\int_0^\varepsilon \xi_1^{N-1}d\xi_1 < \infty$. If $p > N-1$, the integral $\int_0^\infty \frac{t^{p+N-2}}{(1+c^2t^2)^p}\,dt$ converges. Let a_p be its value. If

$N-1<p<2N-1$, we get $\int_{\Omega_\varepsilon}\left|\frac{\xi_1\xi_k}{\xi_1^4+c^2|\xi'|^2}\right|^p d\xi \le |S^{N-2}|a_p\int_0^\varepsilon \xi_1^{2N-2-p}d\xi_1<\infty$ by (3.21). □

Remark. It can be proved that the function $\xi\longmapsto\frac{\xi_1\xi_k}{\xi_1^4+c^2|\xi'|^2}$ does not belong to $L^p(\Omega_\varepsilon)$ if $p\ge 2N-1$, but we will not make use of this fact here.

Now we come back to the proof of Theorem 3.1. All we have to do is to show that $\widehat{H}(0)+c\widehat{G_1}(0)=0$. We argue by contradiction and assume that $\widehat{H}(0)+c\widehat{G_1}(0)\neq 0$. Since the functions $\widehat{H}$ and $\widehat{G_j}$ are continuous, there exists $\varepsilon\in(0,1)$ such that $|\widehat{H}(\xi)+c\widehat{G_1}(\xi)|\ge\frac{1}{2}|\widehat{H}(0)+c\widehat{G_1}(0)|$ for any $\xi\in\Omega_\varepsilon$. Taking a smaller ε if necessary, we may also assume that $|\xi|^4+c^2|\xi'|^2\le 2(\xi_1^4+c^2|\xi'|^2)$ for any $\xi\in\Omega_\varepsilon$. By (3.19) we have

(3.22)
$$\begin{aligned}&\frac{1}{2}\frac{\xi_1^2}{\xi_1^4+c^2|\xi'|^2}|\widehat{H}(0)+c\widehat{G_1}(0)|\le 2\frac{\xi_1^2}{|\xi|^4+c^2|\xi'|^2}|\widehat{H}(\xi)+c\widehat{G_1}(\xi)|\\&\le|\mathcal{F}(|\psi|^2-r_0^2)(\xi)|+2|c|\sum_{k=2}^N\frac{|\xi_1\xi_k|}{\xi_1^4+c^2|\xi'|^2}|\widehat{G_k}(\xi)|+2\frac{|\xi'|^2}{|\xi|^4+c^2|\xi'|^2}|\widehat{H}(\xi)|\quad\text{a.e. on }\Omega_\varepsilon.\end{aligned}$$

Consider first the case $N=2$. We know that $\mathcal{F}(|\psi|^2-r_0^2)\in L^2(\mathbf{R}^2)$, and consequently $\mathcal{F}(|\psi|^2-r_0^2)\in L^p(\Omega_\varepsilon)$ for any $p\in[1,2]$. Since $\widehat{G_k}$ are continuous and bounded, by Lemma 3.2 (ii) we infer that the functions $\xi\longmapsto\frac{\xi_1\xi_k}{\xi_1^4+c^2|\xi'|^2}\widehat{G_k}(\xi)$ belong to $L^p(\Omega_\varepsilon)$ for any $p\in[1,3)$. It is obvious that $\frac{|\xi'|^2}{|\xi|^4+c^2|\xi'|^2}|\widehat{H}(\xi)|\le\frac{1}{c^2}|\widehat{H}(\xi)|$ and $\widehat{H}$ is continuous and bounded on $\mathbf{R}^N$. We conclude that the right-hand side of (3.22) belongs to $L^p(\Omega_\varepsilon)$ for any $p\in[1,2]$. Then (3.22) implies that $\xi\longmapsto\frac{\xi_1^2}{\xi_1^4+c^2|\xi'|^2}$ belongs to $L^2(\Omega_\varepsilon)$, which contradicts Lemma 3.2 (i). This contradiction proves that $\widehat{H}(0)+c\widehat{G_1}(0)=0$.

Next we assume that $N\ge 3$ and $\psi_1\frac{\partial\psi_2}{\partial x_1}-\psi_2\frac{\partial\psi_1}{\partial x_1}\in L^{\frac{2N-1}{2N-3}}(\mathbf{R}^N)$. Equation (3.8) can be written as

$$\begin{aligned}&-\frac{1}{2}\Delta(|\psi|^2-r_0^2)+\frac{v_s^2}{2}(|\psi|^2-r_0^2)\\&=-|\nabla\psi_1|^2-|\nabla\psi_2|^2+F(x,|\psi|^2)|\psi|^2+\frac{v_s^2}{2}(|\psi|^2-r_0^2)+c\left(\psi_1\frac{\partial\psi_2}{\partial x_1}-\psi_2\frac{\partial\psi_1}{\partial x_1}\right).\end{aligned}$$

(3.23)

We have already proved that $F(\cdot,|\psi|^2)|\psi|^2+\frac{v_s^2}{2}(|\psi|^2-r_0^2)\in L^1\cap L^\infty(\mathbf{R}^N)$. From Proposition 2.5 (i) we have $|\nabla\psi|^2\in L^p(\mathbf{R}^N)$ for any $p\in[1,\infty]$. Using the assumption $\psi_1\frac{\partial\psi_2}{\partial x_1}-\psi_2\frac{\partial\psi_1}{\partial x_1}\in L^{\frac{2N-1}{2N-3}}(\mathbf{R}^N)$, we infer that the right-hand side of (3.23) belongs to $L^{\frac{2N-1}{2N-3}}(\mathbf{R}^N)$. By the Hausdorff–Young inequality, for any function $f\in L^p(\mathbf{R}^N)$ with $1\le p\le 2$ we have $\mathcal{F}(f)\in L^{p'}(\mathbf{R}^N)$, where $\frac{1}{p}+\frac{1}{p'}=1$ (see, e.g., Theorem 1.2.1, p. 6, in [4]). Passing to Fourier transforms in (3.23) we get

$$\begin{aligned}\mathcal{F}(|\psi|^2-r_0^2)(\xi)=\frac{2}{|\xi|^2+v_s^2}\mathcal{F}\Bigg[&-|\nabla\psi|^2+\left(F(\cdot,|\psi|^2)|\psi|^2+\frac{v_s^2}{2}(|\psi|^2-r_0^2)\right)\\&+c\left(\psi_1\frac{\partial\psi_2}{\partial x_1}-\psi_2\frac{\partial\psi_1}{\partial x_1}\right)\Bigg](\xi)\qquad\text{a.e. }\xi\in\mathbf{R}^N.\end{aligned}$$

(3.24)

We obtain from (3.24) that $\mathcal{F}(|\psi|^2-r_0^2)\in L^{N-\frac{1}{2}}(\mathbf{R}^N)$. Combined with Lemma 3.2 (ii) and the fact that $\widehat{H}$, $\widehat{G_j}$, and $\xi\longmapsto\frac{|\xi'|^2}{|\xi|^4+c^2|\xi'|^2}$ are bounded, this implies that the last

expression in (3.22) is in $L^{N-\frac{1}{2}}(\Omega_\varepsilon)$. We infer that the function $\xi \longmapsto \frac{\xi_1^2}{\xi_1^4+c^2|\xi'|^2}|\widehat{H}(0)+c\widehat{G_1}(0)|$ must be in $L^{N-\frac{1}{2}}(\Omega_\varepsilon)$ for any sufficiently small ε. If $\widehat{H}(0)+c\widehat{G_1}(0)\neq 0$, this contradicts Lemma 3.2 (i). Thus necessarily $\widehat{H}(0)+c\widehat{G_1}(0)=0$ and the proof of Theorem 3.1 is complete. ∎

It is an open problem whether any finite-energy traveling wave ψ of (1.1) moving with speed $c=\pm v_s$ satisfies $\psi_1\frac{\partial\psi_2}{\partial x_1}-\psi_2\frac{\partial\psi_1}{\partial x_1}\in L^{\frac{2N-1}{2N-3}}(\mathbf{R}^N)$. Even for very particular cases of (1.1), such as the Gross–Pitaevskii equation, the answer to this question is not known. However, we have the following.

PROPOSITION 3.3. *Assume that* (H1)–(H5) *hold and let* $\psi=\psi_1+i\psi_2$ *be a finite-energy traveling wave for* (1.1) *such that* $F(\cdot,|\psi|^2)\in L^1_{loc}(\mathbf{R}^N)$. *Let* R_* *be sufficiently big so that* $|\psi|\geq\frac{r_0}{2}$ *on* $\mathbf{R}^N\setminus B(0,R_*)$, *let* θ *be the lifting given by Proposition* 2.5 (ii), *and let* $\chi\in C^\infty(\mathbf{R}^N)$ *be a cut-off function as in Theorem* 3.1. *Then the following hold:*

(i) *Let* $p\in(1,\infty)$. *The following assertions are equivalent:*
  (a) $\nabla(\chi\theta)\in L^p(\mathbf{R}^N)$;
  (b) $\psi_1\frac{\partial\psi_2}{\partial x_j}-\psi_2\frac{\partial\psi_1}{\partial x_j}\in L^p(\mathbf{R}^N)$ *for any* $j\in\{1,\dots,N\}$;
  (c) $\psi_1\frac{\partial\psi_2}{\partial x_1}-\psi_2\frac{\partial\psi_1}{\partial x_1}\in L^p(\mathbf{R}^N)$;
  (d) $|\psi|^2-r_0^2\in W^{2,p}(\mathbf{R}^N)$;
  (e) $|\psi|^2-r_0^2\in L^p(\mathbf{R}^N)$.

(ii) *If* $N\geq 3$, *there exists* $\theta_0\in\mathbf{R}$ *such that* $\chi\theta-\theta_0\in W^{2,q}(\mathbf{R}^N)$ *for any* $q\in[\frac{2N}{N-2},\infty)$.

Moreover, if $c^2=v_s^2$, *we have the following:*

(iii) $|\psi|^2-r_0^2\in L^p(\mathbf{R}^N)$ *and* $\psi_1\frac{\partial\psi_2}{\partial x_j}-\psi_2\frac{\partial\psi_1}{\partial x_j}\in L^p(\mathbf{R}^N)$ *for any* $p>\frac{2N-1}{2N-3}$ *and* $j\in\{1,\dots,N\}$.

(iv) $\nabla(|\psi|^2-r_0^2)\in L^p(\mathbf{R}^N)$ *for any* $p>\frac{2N-1}{2N-2}$.

(v) $\partial_{j,k}^2(|\psi|^2-r_0^2)\in L^p(\mathbf{R}^N)$ *for any* $p\in(1,\infty)$.

Proof. (i) Since $\psi\in L^\infty(\mathbf{R}^N)$ and (3.3) holds, the equivalence (a) ⇔ (b) is clear. It is also obvious that (b) ⇒ (c).

From the classical Marcinkiewicz theorem (see Theorem 3, p. 96, in [27]) it follows that the functions $\frac{1}{|\xi|^2+v_s^2}$, $\frac{\xi_j}{|\xi|^2+v_s^2}$, and $\frac{\xi_j\xi_k}{|\xi|^2+v_s^2}$ are L^p-multipliers for $1<p<\infty$. Assume that $\psi_1\frac{\partial\psi_2}{\partial x_1}-\psi_2\frac{\partial\psi_1}{\partial x_1}\in L^p(\mathbf{R}^N)$. Since $|\nabla\psi|^2\in L^1\cap L^\infty(\mathbf{R}^N)$ and $F(\cdot,|\psi|^2)|\psi|^2+\frac{v_s^2}{2}(|\psi|^2-r_0^2)\in L^1\cap L^\infty(\mathbf{R}^N)$ by Theorem 3.1 (i), we have $-|\nabla\psi|^2+(F(\cdot,|\psi|^2)|\psi|^2+\frac{v_s^2}{2}(|\psi|^2-r_0^2))+c(\psi_1\frac{\partial\psi_2}{\partial x_1}-\psi_2\frac{\partial\psi_1}{\partial x_1})\in L^p(\mathbf{R}^N)$ and infer from (3.24) that $|\psi|^2-r_0^2\in W^{2,p}(\mathbf{R}^N)$. Hence (c) ⇒ (d). It is obvious that (d) ⇒ (e).

It follows from Proposition 2.5 (ii) that $\partial_k(\chi\theta)\in\mathcal{S}'(\mathbf{R}^N)$. It is then clear that all terms appearing in (3.9) belong to $\mathcal{S}'(\mathbf{R}^N)$. We take the derivative of (3.9) with respect to x_k (in $\mathcal{S}'(\mathbf{R}^N)$), then take the Fourier transform of the resulting equality to obtain

$$r_0^2\mathcal{F}\left(\frac{\partial}{\partial x_k}(\chi\theta)\right)=-\sum_{j=1}^N\frac{\xi_j\xi_k}{|\xi|^2}\widehat{G_j}+\frac{c}{2}\frac{\xi_1\xi_k}{|\xi|^2}\mathcal{F}(|\psi|^2-r_0^2)$$

or, equivalently,

$$r_0^2\frac{\partial}{\partial x_k}(\chi\theta)=\sum_{j=1}^N R_jR_k(G_j)-\frac{c}{2}R_1R_k(|\psi|^2-r_0^2), \tag{3.25}$$

where R_j is the Riesz transform, $R_j\phi = \mathcal{F}^{-1}(i\frac{\xi_j}{|\xi|}\widehat{\phi})$. It is well known that the Riesz transform maps continuously $L^p(\mathbf{R}^N)$ into $L^p(\mathbf{R}^N)$ for $1 < p < \infty$ (see, e.g., Theorem 3, p. 96, and Example (iii), p. 95, in [27]). From Theorem 3.1 (i) we have $G_j \in L^1 \cap L^\infty(\mathbf{R}^N)$; therefore $R_jR_k(G_j) \in L^q(\mathbf{R}^N)$ for any $q \in (1,\infty)$. Assume that $|\psi|^2 - r_0^2 \in L^p(\mathbf{R}^N)$ for some $p \in (1,\infty)$. Then $R_1R_k(|\psi|^2 - r_0^2) \in L^p(\mathbf{R}^N)$ and from (3.25) we infer that $\frac{\partial}{\partial x_k}(\chi\theta) \in L^p(\mathbf{R}^N)$ for any $k \in \{1,\dots,N\}$. Thus (e) $\Rightarrow$ (a) and (i) is proved.

(ii) It is well known that for any function ϕ satisfying $\nabla\phi \in L^p(\mathbf{R}^N)$ with $p < N$, there exists a constant λ such that $\phi - \lambda \in L^{p^*}(\mathbf{R}^N)$, where $\frac{1}{p^*} = \frac{1}{p} - \frac{1}{N}$ (see Theorem 4.5.9 in [20] or Lemma 7 and Remark 4.2 in [15, pp. 774–775] for a different proof). From Proposition 2.5 (ii) we have $\nabla(\chi\theta) \in W^{1,p}(\mathbf{R}^N)$ for any $p \in [2,\infty)$. If $N \geq 3$, we infer that there exists $\theta_0 \in \mathbf{R}$ such that $\chi\theta - \theta_0 \in L^q(\mathbf{R}^N)$ for $q \in [\frac{2N}{N-2},\infty)$. Therefore $\chi\theta - \theta_0 \in W^{2,q}(\mathbf{R}^N)$ for any $q \in [\frac{2N}{N-2},\infty)$ and, in particular, $\chi\theta - \theta_0 \longrightarrow 0$ as $|x| \longrightarrow \infty$.

(iii) We will use the following result due to Lizorkin (see Theorem 8, p. 288, in [24]).

THEOREM 3.4 (see [24]). *Let $\beta \in [0,1)$ and let $K \in L^\infty(\mathbf{R}^N) \cap C^N(\mathbf{R}^N \setminus \{0\})$. Assume that*

$$\Big(\prod_{j=1}^N |\xi_j|^{k_j+\beta}\Big)\partial_1^{k_1}\dots\partial_N^{k_N}K \in L^\infty(\mathbf{R}^N) \qquad \text{for any } k_1,\dots,k_N \in \{0,1\}.$$

Then K is a Fourier multiplier from $L^p(\mathbf{R}^N)$ to $L^{\frac{p}{1-\beta p}}(\mathbf{R}^N)$ for any $p \in (1,\frac{1}{\beta})$.

Let $K(\xi) = \frac{|\xi|^2}{|\xi|^4 + c^2|\xi'|^2}$, where $\xi' = (\xi_2,\dots,\xi_N)$. A straightforward but tedious computation shows that K satisfies the assumptions of Lizorkin's theorem for $\beta = \frac{1}{2N-1}$. From (3.19) we obtain

$$\begin{aligned}(3.26)\qquad |\psi|^2 - r_0^2 &= 2R_1^2\left(\mathcal{F}^{-1}\left(K(\widehat{H} + c\widehat{G_1})\right)\right)\\ &\quad + 2c\sum_{j=2}^N R_1R_j\left(\mathcal{F}^{-1}(K\widehat{G_j})\right) + 2\sum_{j=2}^N R_j^2\left(\mathcal{F}^{-1}(K\widehat{H})\right),\end{aligned}$$

where R_j's denote Riesz transforms. Since $H, G_1,\dots,G_N \in L^1 \cap L^\infty(\mathbf{R}^N)$, by (3.26) and Lizorkin's theorem we infer that $|\psi|^2 - r_0^2 \in L^p(\mathbf{R}^N)$ for any $p \in (\frac{2N-1}{2N-3},\infty)$. The rest of (iii) follows from part (i), (b) $\Leftrightarrow$ (e).

(iv) and (v) From (iii) and (i), (d) $\Leftrightarrow$ (e) it follows immediately that $|\psi|^2 - r_0^2 \in W^{2,p}(\mathbf{R}^N)$ for any $p \in (\frac{2N-1}{2N-3},\infty)$. Using (3.19) we obtain

$$\begin{aligned}(3.27)\qquad &\partial_{k\ell}^2\left(|\psi|^2 - r_0^2\right) = 2R_kR_\ell R_1^2\left(\mathcal{F}^{-1}\left(|\xi|^2K(\widehat{H} + c\widehat{G_1})\right)\right)\\ &+2c\sum_{j=2}^N R_kR_\ell R_1R_j\left(\mathcal{F}^{-1}(|\xi|^2K\widehat{G_j})\right)\\ &+2\sum_{j=2}^N R_kR_\ell R_j^2\left(\mathcal{F}^{-1}(|\xi|^2K\widehat{H})\right) \qquad \text{in } \mathcal{S}'(\mathbf{R}^N).\end{aligned}$$

It can be proved by direct computation that the function $|\xi|^2K$ satisfies the assumptions of Lizorkin's theorem for $\beta = 0$. Consequently $|\xi|^2K$ is an L^p-multiplier for

$1 < p < \infty$. Since $H, G_j \in L^1 \cap L^\infty(\mathbf{R}^N)$, it follows from (3.27) that $\partial^2_{k\ell}\left(|\psi|^2 - r_0^2\right) \in L^p(\mathbf{R}^N)$ for $1 < p < \infty$.

By using the Gagliardo–Nirenberg inequality

$$||\nabla\phi||^2_{L^p} \le C||\phi||_{L^q}||\nabla^2\phi||_{L^r} \qquad \text{if} \quad \frac{1}{p} = \frac{1}{2}\left(\frac{1}{q} + \frac{1}{r}\right),$$

we infer that $\nabla(|\psi|^2 - r_0^2) \in L^p(\mathbf{R}^N)$ for any $p > \frac{2N-1}{2N-2}$. □

COROLLARY 3.5. *Under the assumptions of Theorem* 3.1, *assume that* $N \ge 3$, $c^2 = v_s^2$, *and the momentum of* ψ *with respect to the* x_1*-direction is well-defined, that is,* $\psi_1 \frac{\partial\psi_2}{\partial x_1} - \psi_2 \frac{\partial\psi_1}{\partial x_1} \in L^1(\mathbf{R}^N)$. *Then* ψ *satisfies* (3.1).

Proof. From Proposition 3.3 (iii) and (i) we have $\psi_1 \frac{\partial\psi_2}{\partial x_1} - \psi_2 \frac{\partial\psi_1}{\partial x_1} \in L^p(\mathbf{R}^N)$ for $p \in (\frac{2N-1}{2N-3}, \infty)$. Then the assumption $\psi_1 \frac{\partial\psi_2}{\partial x_1} - \psi_2 \frac{\partial\psi_1}{\partial x_1} \in L^1(\mathbf{R}^N)$ implies $\psi_1 \frac{\partial\psi_2}{\partial x_1} - \psi_2 \frac{\partial\psi_1}{\partial x_1} \in L^p(\mathbf{R}^N)$ for any $p \in [1, \infty)$. Now the conclusion follows from Theorem 3.1 (iii). □

4. Nonexistence results. In this section we show how Theorem 3.1 may be used to prove nonexistence of supersonic and sonic traveling waves with finite energy for some equations of type (1.1).

4.1. Equations invariant by translations. We consider the equation

$$i\frac{\partial\Phi}{\partial t} + \Delta\Phi + G(|\Phi|^2)\Phi = 0 \qquad \text{in } \mathbf{R}^N. \tag{4.1}$$

We assume that the function $G : [0, \infty) \longrightarrow \mathbf{R}$ satisfies the following assumptions:

- **(A1)** $G \in C^2([0, \infty), \mathbf{R})$ and there exists $r_0 > 0$ such that $G(r_0^2) = 0$ and $G'(r_0^2) < 0$.
- **(A2)** There exists $\alpha > 0$ such that $\limsup_{s\to\infty} \frac{G(s)}{s^\alpha} < 0$.

Obviously, (4.1) is of the form (1.1). As previously, we associate to (4.1) the "boundary condition" $|\Phi| \longrightarrow r_0^2$ as $|x| \longrightarrow \infty$. In this context, the sound velocity at infinity is $v_s = r_0\sqrt{-2G'(r_0^2)}$. The energy corresponding to (4.1) is $E(\Phi) = \int_{\mathbf{R}^N} |\nabla\Phi|^2\, dx + \int_{\mathbf{R}^N} V(|\Phi|^2)\, dx$, where $V(s) = \int_s^{r_0^2} G(\tau)\, d\tau$. Let ψ be a finite-energy traveling wave for (4.1) (in the sense of Definition 2.1) moving with speed c. Then ψ satisfies the equation

$$-ic\frac{\partial\psi}{\partial x_1} + \Delta\psi + G(|\psi|^2)\psi = 0 \qquad \text{in } \mathcal{D}'(\mathbf{R}^N), \qquad |\psi| \longrightarrow r_0 \quad \text{as } |x| \longrightarrow \infty. \tag{4.2}$$

If G satisfies (A1)–(A2), it is easy to see that $F(x, s) := G(s)$ satisfies the assumptions (H1)–(H5) in section 2 (with $L = -G'(r_0^2)$). It is then clear that the conclusions of Propositions 2.2 and 2.5 and Theorem 3.1 (i) are valid for ψ. Moreover, we have the following.

PROPOSITION 4.1 (Pohozaev identities). *Let* ψ *be as above. Choose* $R_* > 0$ *such that* $|\psi| \ge \frac{r_0}{2}$ *on* $\mathbf{R}^N \setminus B(0, R_*)$. *Let* θ *be the lifting of* $\frac{\psi}{|\psi|}$ *on* $\mathbf{R}^N \setminus B(0, R_*)$ *(as given by Proposition* 2.5 (ii)*) and let* χ *be a cut-off function as in Theorem* 3.1. *The following identities hold:*

$$-\int_{\mathbf{R}^N} \left|\frac{\partial\psi}{\partial x_1}\right|^2 dx + \int_{\mathbf{R}^N} \sum_{j=2}^N \left|\frac{\partial\psi}{\partial x_j}\right|^2 dx + \int_{\mathbf{R}^N} V(|\psi|^2)\, dx = 0, \tag{4.3}$$

$$
\begin{aligned}
&-\int_{\mathbf{R}^N}\left|\frac{\partial\psi}{\partial x_k}\right|^2 dx+\int_{\mathbf{R}^N}\sum_{j=1,\ j\neq k}^{N}\left|\frac{\partial\psi}{\partial x_j}\right|^2 dx+\int_{\mathbf{R}^N}V(|\psi|^2)\,dx\\
&\quad -c\int_{\mathbf{R}^N}\psi_1\frac{\partial\psi_2}{\partial x_1}-\psi_2\frac{\partial\psi_1}{\partial x_1}-r_0^2\frac{\partial}{\partial x_1}(\chi\theta)\,dx=0\qquad \textit{for } k=2,\dots,N.
\end{aligned}
\tag{4.4}
$$

It is worth noting that Proposition 4.1 is valid for any speed $c\in\mathbf{R}$.

Proof. Since the arguments are rather classical, we only sketch the proof.

Formally, traveling waves are critical points of the functional $E_c=E+cP_1$, where E is the energy and P_1 is the momentum with respect to the x_1-direction (see (1.3)). Identities (4.3) and (4.4) are simple consequences of the behavior of E_c with respect to dilations in $\mathbf{R}^N$. To be more precise, define $\psi_{k,t}(x)=\psi(x_1,\dots,x_{k-1},tx_k,x_{k+1},\dots,x_N)$ and $g_k(t)=E_c(\psi_{k,t})$. If ψ is a critical point of E_c, one would expect that $g_k'(1)=\frac{d}{dt}(E_c(\psi_{k,t}))\big|_{t=1}=0$, and this is precisely (4.3) if $k=1$, respectively, (4.4) if $k\geq 2$. However, this argument is not rigorous for at least two reasons. First, it is not clear what function space one should consider to define E_c (and this could not be a vector space because of the boundary conditions at infinity). Second, even if an appropriate function space is found, we do not know whether $\frac{d}{dt}(\psi_{k,t})\big|_{t=1}=x_k\frac{\partial\psi}{\partial x_k}$ belong to the tangent space at ψ of the considered function space.

The most convenient way to prove Pohozaev identities is to use a truncation argument. Fix a function $\eta\in C_c^\infty(\mathbf{R}^N)$ such that $\eta=1$ on $B(0,1)$ and $\eta=0$ on $\mathbf{R}^N\setminus B(0,2)$. For $n\geq 1$, define $\eta_n(x)=\eta(\frac{x}{n})$. We take the scalar product of (4.2) by $x_k\eta_n(x)\frac{\partial\psi}{\partial x_k}$ and integrate by parts the resulting equality. It is standard (see, e.g., Proposition 1, p. 320, in [3] or Lemma 2.4, p. 104, in [10]) to prove that

$$
\begin{aligned}
&\lim_{n\to\infty}\int_{\mathbf{R}^N}\left(\Delta\psi,x_k\eta_n(x)\frac{\partial\psi}{\partial x_k}\right)dx\\
&=-\int_{\mathbf{R}^N}\left|\frac{\partial\psi}{\partial x_k}\right|^2 dx+\frac{1}{2}\int_{\mathbf{R}^N}|\nabla\psi|^2\,dx
\end{aligned}
\tag{4.5}
$$

and

$$
\lim_{n\to\infty}\int_{\mathbf{R}^N}\left(G(|\psi|^2)\psi,x_k\eta_n(x)\frac{\partial\psi}{\partial x_k}\right)dx=\frac{1}{2}\int_{\mathbf{R}^N}V(|\psi|^2)\,dx.
\tag{4.6}
$$

It is obvious that $(ic\frac{\partial\psi}{\partial x_1},\eta_n(x)x_1\frac{\partial\psi}{\partial x_1})=c\eta_n(x)x_1(i\frac{\partial\psi}{\partial x_1},\frac{\partial\psi}{\partial x_1})=0$. Thus taking the scalar product of (4.2) by $x_1\eta_n(x)\frac{\partial\psi}{\partial x_1}$, integrating, and using (4.5) and (4.6) we get (4.3).

By (3.3) we have $(-i\frac{\partial\psi}{\partial x_j},\psi)=\psi_1\frac{\partial\psi_2}{\partial x_j}-\psi_2\frac{\partial\psi_1}{\partial x_j}=|\psi|^2\frac{\partial\theta}{\partial x_j}$ on $\mathbf{R}^N\setminus\overline{B}(0,R_*)$. Using the convention $\partial^\alpha(\chi\theta)=0$, $(\partial^\alpha\chi)\theta=0$ on $B(0,2R_*)$, we have

$$
\begin{aligned}
&\left(-i\frac{\partial\psi}{\partial x_j},\psi\right)=(1-\chi)\left(-i\frac{\partial\psi}{\partial x_j},\psi\right)+\chi|\psi|^2\frac{\partial\theta}{\partial x_j}\\
&=(1-\chi)\left(-i\frac{\partial\psi}{\partial x_j},\psi\right)+|\psi|^2\frac{\partial(\chi\theta)}{\partial x_j}-|\psi|^2\theta\frac{\partial\chi}{\partial x_j}\qquad\text{on }\mathbf{R}^N.
\end{aligned}
\tag{4.7}
$$

Therefore we get for $k = 2, \dots, N$

$$\begin{aligned}
&\int_{\mathbf{R}^N} \left(-ic\frac{\partial \psi}{\partial x_1}, x_k \eta_n(x) \frac{\partial \psi}{\partial x_k} \right) dx \\
&= \frac{c}{2} \int_{\mathbf{R}^N} x_k \eta_n(x) \left[\frac{\partial}{\partial x_1} \left(-i\psi, \frac{\partial \psi}{\partial x_k} \right) + \frac{\partial}{\partial x_k} \left(-i\frac{\partial \psi}{\partial x_1}, \psi \right) \right] dx \\
&= -\frac{c}{2} \int_{\mathbf{R}^N} x_k \frac{\partial \eta_n}{\partial x_1}(x) \left(-i\psi, \frac{\partial \psi}{\partial x_k} \right) + \left(\eta_n(x) + x_k \frac{\partial \eta_n}{\partial x_k}(x) \right) \left(-i\frac{\partial \psi}{\partial x_1}, \psi \right) dx \\
&= \frac{c}{2} \int_{\mathbf{R}^N} x_k \frac{\partial \eta_n}{\partial x_1}(x) \left[(1-\chi) \left(-i\frac{\partial \psi}{\partial x_k}, \psi \right) + |\psi|^2 \frac{\partial(\chi\theta)}{\partial x_k} - |\psi|^2 \theta \frac{\partial \chi}{\partial x_k} \right] dx \\
&\quad -\frac{c}{2} \int_{\mathbf{R}^N} \eta_n(x) \left(-i\frac{\partial \psi}{\partial x_1}, \psi \right) dx \\
&\quad -\frac{c}{2} \int_{\mathbf{R}^N} x_k \frac{\partial \eta_n}{\partial x_k}(x) \left[(1-\chi) \left(-i\frac{\partial \psi}{\partial x_1}, \psi \right) + |\psi|^2 \frac{\partial(\chi\theta)}{\partial x_1} - |\psi|^2 \theta \frac{\partial \chi}{\partial x_1} \right] dx \\
&= \frac{c}{2} \int_{\mathbf{R}^N} x_k |\psi|^2 \left(\frac{\partial \eta_n}{\partial x_1} \frac{\partial(\chi\theta)}{\partial x_k} - \frac{\partial \eta_n}{\partial x_k} \frac{\partial(\chi\theta)}{\partial x_1} \right) - \eta_n(x) \left(-i\frac{\partial \psi}{\partial x_1}, \psi \right) dx \quad \text{if } n > 3R_*
\end{aligned} \tag{4.8}$$

because $\operatorname{supp}(1-\chi) \subset \overline{B}(0, 3R_*)$ and $\operatorname{supp}\nabla\eta_n \subset \overline{B}(0, 2n) \setminus B(0, n)$, and consequently $(1-\chi)\frac{\partial \eta_n}{\partial x_j} = 0$ and $\frac{\partial \chi}{\partial x_\ell}\frac{\partial \eta_n}{\partial x_j} = 0$ on $\mathbf{R}^N$ for $n > 3R_*$.

It is obvious that

$$\begin{aligned}
&\int_{\mathbf{R}^N} x_k \Big(\frac{\partial \eta_n}{\partial x_1} \frac{\partial(\chi\theta)}{\partial x_k} - \frac{\partial \eta_n}{\partial x_k} \frac{\partial(\chi\theta)}{\partial x_1} \Big) dx \\
&= \int_{\mathbf{R}^N} x_k \Big[\frac{\partial}{\partial x_1} \Big(\eta_n \frac{\partial(\chi\theta)}{\partial x_k} \Big) - \frac{\partial}{\partial x_k} \Big(\eta_n \frac{\partial(\chi\theta)}{\partial x_1} \Big) \Big] dx = \int_{\mathbf{R}^N} \eta_n \frac{\partial(\chi\theta)}{\partial x_1} dx.
\end{aligned} \tag{4.9}$$

Since $|\psi|^2 - r_0^2$ and $\nabla(\chi\theta)$ belong to $L^2(\mathbf{R}^N)$, using the dominated convergence theorem we obtain

$$\begin{aligned}
&\left| \int_{\mathbf{R}^N} x_k (|\psi|^2 - r_0^2) \left(\frac{\partial \eta_n}{\partial x_1} \frac{\partial(\chi\theta)}{\partial x_k} - \frac{\partial \eta_n}{\partial x_k} \frac{\partial(\chi\theta)}{\partial x_1} \right) dx \right| \\
&\quad \le 4 ||\nabla \eta||_{L^\infty(\mathbf{R}^N)} \int_{B(0,2n)\setminus B(0,n)} |\,|\psi|^2 - r_0^2| \cdot |\nabla(\chi\theta)|\, dx \longrightarrow 0 \qquad \text{as } n \longrightarrow \infty.
\end{aligned} \tag{4.10}$$

Recall that $\psi_1 \frac{\partial \psi_2}{\partial x_1} - \psi_2 \frac{\partial \psi_1}{\partial x_1} - r_0^2 \frac{\partial(\chi\theta)}{\partial x_1} \in L^1(\mathbf{R}^N)$ by Theorem 3.1 (i), and by dominated convergence we get

$$\begin{aligned}
&\int_{\mathbf{R}^N} \eta_n \Big[\Big(-i\frac{\partial \psi}{\partial x_1}, \psi \Big) - r_0^2 \frac{\partial(\chi\theta)}{\partial x_1} \Big] dx = \int_{\mathbf{R}^N} \eta_n \Big[\psi_1 \frac{\partial \psi_2}{\partial x_1} - \psi_2 \frac{\partial \psi_1}{\partial x_1} - r_0^2 \frac{\partial(\chi\theta)}{\partial x_1} \Big] dx \\
&\longrightarrow \int_{\mathbf{R}^N} \psi_1 \frac{\partial \psi_2}{\partial x_1} - \psi_2 \frac{\partial \psi_1}{\partial x_1} - r_0^2 \frac{\partial(\chi\theta)}{\partial x_1} dx \qquad \text{as } n \longrightarrow \infty.
\end{aligned} \tag{4.11}$$

Combining (4.8)–(4.11) we find

$$\lim_{n\to\infty}\int_{\mathbf{R}^N}\left(-ic\frac{\partial\psi}{\partial x_1},x_k\eta_n(x)\frac{\partial\psi}{\partial x_k}\right)dx = -\frac{c}{2}\int_{\mathbf{R}^N}\psi_1\frac{\partial\psi_2}{\partial x_1}-\psi_2\frac{\partial\psi_1}{\partial x_1}-r_0^2\frac{\partial(\chi\theta)}{\partial x_1}\,dx. \tag{4.12}$$

Taking the scalar product of (4.2) by $\eta_n(x)x_k\frac{\partial\psi}{\partial x_k}$, integrating over $\mathbf{R}^N$, and using (4.5), (4.6), and (4.12) we obtain (4.4). ∎

THEOREM 4.2. *Assume that $N\geq 2$,* (A1), (A2) *hold and let ψ be a finite-energy traveling wave for* (3.1) *such that $G(|\psi|^2)\psi\in L^1_{loc}(\mathbf{R}^N)$. Suppose that either*

- *$c^2>v_s^2$, where $v_s=r_0\sqrt{-2G'(r_0^2)}$ is the sound velocity at infinity, or*
- *$N=2$ and $c^2=v_s^2$, or*
- *$N\geq 3$, $c^2=v_s^2$, and $\psi_1\frac{\partial\psi_2}{\partial x_1}-\psi_2\frac{\partial\psi_1}{\partial x_1}\in L^{\frac{2N-1}{2N-3}}(\mathbf{R}^N)$.*

Moreover, assume that G satisfies

- **(A3)** *there exists $\alpha\in[-1+\frac{N-3}{N-1}(1-\frac{v_s^2}{c^2}),\frac{v_s^2}{c^2}]$ such that*

$$sG(s)+\frac{v_s^2}{2}(s-r_0^2)+\left(1-\alpha-\frac{v_s^2}{c^2}\right)V(s)\leq 0 \qquad \text{for any } s\geq 0.$$

Then ψ is constant.

Proof. It follows from Propositions 2.2 and 2.5 that ψ is smooth, and Proposition 4.1 implies that ψ satisfies (4.3) and (4.4). Summing up the identities (4.4) for $k=2,\dots,N$ we get

$$\begin{aligned}&\int_{\mathbf{R}^N}\left|\frac{\partial\psi}{\partial x_1}\right|^2+\frac{N-3}{N-1}\sum_{k=2}^N\left|\frac{\partial\psi}{\partial x_k}\right|^2dx+\int_{\mathbf{R}^N}V(|\psi|^2)\,dx\\ &-c\int_{\mathbf{R}^N}\psi_1\frac{\partial\psi_2}{\partial x_1}-\psi_2\frac{\partial\psi_1}{\partial x_1}-r_0^2\frac{\partial(\chi\theta)}{\partial x_1}\,dx=0.\end{aligned} \tag{4.13}$$

On the other hand, from Theorem 3.1 we have

$$\begin{aligned}&\int_{\mathbf{R}^N}|\nabla\psi|^2-G(|\psi|^2)|\psi|^2-\frac{v_s^2}{2}(|\psi|^2-r_0^2)\,dx\\ &-c\left(1-\frac{v_s^2}{c^2}\right)\int_{\mathbf{R}^N}\psi_1\frac{\partial\psi_2}{\partial x_1}-\psi_2\frac{\partial\psi_1}{\partial x_1}-r_0^2\frac{\partial}{\partial x_1}(\chi\theta)\,dx=0.\end{aligned} \tag{4.14}$$

We multiply (4.13) by $-1+\frac{v_s^2}{c^2}$ and add the resulting equality to (4.14) to get

$$\begin{aligned}&\int_{\mathbf{R}^N}\frac{v_s^2}{c^2}\left|\frac{\partial\psi}{\partial x_1}\right|^2+\left(1-\left(1-\frac{v_s^2}{c^2}\right)\frac{N-3}{N-1}\right)\sum_{k=2}^N\left|\frac{\partial\psi}{\partial x_k}\right|^2dx\\ &-\int_{\mathbf{R}^N}G(|\psi|^2)|\psi|^2+\frac{v_s^2}{2}(|\psi|^2-r_0^2)+\left(1-\frac{v_s^2}{c^2}\right)V(|\psi|^2)\,dx=0.\end{aligned} \tag{4.15}$$

Let α satisfy (A3). Multiplying (4.3) by α and adding it to (4.15) we obtain

$$\begin{aligned}&\int_{\mathbf{R}^N}\left(\frac{v_s^2}{c^2}-\alpha\right)\left|\frac{\partial\psi}{\partial x_1}\right|^2+\left(\alpha+1-\left(1-\frac{v_s^2}{c^2}\right)\frac{N-3}{N-1}\right)\sum_{k=2}^N\left|\frac{\partial\psi}{\partial x_k}\right|^2dx\\ &=\int_{\mathbf{R}^N}G(|\psi|^2)|\psi|^2+\frac{v_s^2}{2}(|\psi|^2-r_0^2)+\left(1-\alpha-\frac{v_s^2}{c^2}\right)V(|\psi|^2)\,dx.\end{aligned} \tag{4.16}$$

By (A3), the right-hand side of (4.16) is less than or equal to zero. If $\alpha \in (-1+(1-\frac{v_s^2}{c^2})\frac{N-3}{N-1}, \frac{v_s^2}{c^2})$, it follows from (4.16) that $\int_{\mathbf{R}^N} \left|\frac{\partial\psi}{\partial x_k}\right|^2 dx = 0$ for $k = 1, \dots, N$, which implies $\nabla\psi = 0$ on $\mathbf{R}^N$, i.e., ψ is constant. If $\alpha = -1 + (1 - \frac{v_s^2}{c^2})\frac{N-3}{N-1}$, we infer from (4.16) that $\int_{\mathbf{R}^N} \left|\frac{\partial\psi}{\partial x_1}\right|^2 dx = 0$; consequently $\frac{\partial\psi}{\partial x_1} = 0$ on $\mathbf{R}^N$, which implies that ψ does not depend on x_1. Since $\int_{\mathbf{R}^N} |\nabla\psi|^2\, dx$ is finite, we have necessarily $\nabla\psi = 0$ on $\mathbf{R}^N$, which means that is ψ is constant. A similar argument shows that ψ is constant in the case $\alpha = \frac{v_s^2}{c^2}$. ∎

Remark. Let α, C_1, and $\tilde{r}$ be positive constants satisfying $G(s^2) + \frac{c^2}{4} \le -C_1(s - \tilde{r})^{2\alpha}$ for any $s \ge \tilde{r}$ (such constants exist by assumption (A2)). Let ψ be as in Theorem 4.2. It follows from the proof of Proposition 2.2 (i) that $|\psi(x)| \le \tilde{r}\sqrt{2}$ for any x. Therefore the proof of Theorem 4.2 is still valid if the inequality in (A3) holds only for all $s \in [0, 2\tilde{r}^2]$.

If $c^2 = v_s^2$, $N \ge 3$, and ψ is as above, we already know from Proposition 3.3 (iii) that $\psi_1\frac{\partial\psi_2}{\partial x_1} - \psi_2\frac{\partial\psi_1}{\partial x_1} \in L^p(\mathbf{R}^N)$ for any $p \in (\frac{2N-1}{2N-3}, \infty)$. Therefore we have the following.

COROLLARY 4.3. *Assume that* (A1), (A2), (A3) *hold,* $N \ge 3$, *and* $c^2 = v_s^2$. *Let* ψ *be a traveling wave for* (4.1) *having finite energy, finite momentum with respect to the* x_1*-direction (i.e.,* $\psi_1\frac{\partial\psi_2}{\partial x_1} - \psi_2\frac{\partial\psi_1}{\partial x_1} \in L^1(\mathbf{R}^N)$*), and such that* $G(|\psi|^2)\psi \in L^1_{loc}(\mathbf{R}^N)$. *Then* ψ *is constant.*

Example 4.4. The Gross–Pitaevskii equation is of type (4.1) with $G(s) = 1 - s$. In this case we have $r_0 = 1$, $V(s) = \frac{1}{2}(s-1)^2$, and $v_s = \sqrt{2}$. For any finite-energy function ψ we have $\int_{\mathbf{R}^N}(|\psi|^2 - 1)^2\, dx < \infty$; hence $\psi \in L^4_{loc}(\mathbf{R}^N)$ and consequently $G(|\psi|^2)\psi \in L^1_{loc}(\mathbf{R}^N)$. Assumptions (A1) and (A2) are clearly satisfied. We find $sG(s) + \frac{v_s^2}{2}(s - r_0^2) + (1 - \alpha - \frac{v_s^2}{c^2})V(s) = -(\frac{1}{2} + \alpha + \frac{v_s^2}{c^2})(1-s)^2$. The last expression is nonpositive for any s if $\alpha \ge -\frac{1}{2} - \frac{v_s^2}{c^2}$, and thus assumption (A3) is also satisfied. Hence the conclusion of Theorem 4.2 holds for the Gross–Pitaevskii equation. In particular, we recover the nonexistence results in [17], [18].

Example 4.5. The cubic-quintic Schrödinger equation is of the form (4.1) with $G(s) = -\alpha_1 + \alpha_3 s - \alpha_5 s^2$, where α_1, α_3, α_5 are positive and $\frac{3}{16} < \frac{\alpha_1\alpha_5}{\alpha_3^2} < \frac{1}{4}$. The nonlinearity G can be written as $G(s) = -\alpha_5(s - r_1^2)(s - r_0^2)$, where $0 < r_1 < r_0$. In this case we have $v_s^2 = -2r_0^2G'(r_0) = 2\alpha_5 r_0^2(r_0^2 - r_1^2)$ and $V(s) = \frac{\alpha_5}{3}(s - r_0^2)^2(s + \frac{1}{2}r_0^2 - \frac{3}{2}r_1^2)$. For any function ψ with finite energy we have $V(|\psi|^2) \in L^1(\mathbf{R}^N)$, which implies $\psi \in L^6_{loc}(\mathbf{R}^N)$ and consequently $G(|\psi|^2)\psi \in L^1_{loc}(\mathbf{R}^N)$. It is obvious that G satisfies (A1) and (A2). If $c^2 \ge v_s^2$, we have $-\frac{v_s^2}{c^2} \in [-1 + \frac{N-3}{N-1}(1 - \frac{v_s^2}{c^2}), \frac{v_s^2}{c^2}]$ and an easy computation shows that $sG(s) + \frac{v_s^2}{2}(s - r_0^2) + V(s) = -\frac{\alpha_5}{6}(4s + 5r_0^2 - 3r_1^2) \le 0$ for any $s \ge 0$. Hence assumption (A3) holds for $\alpha = -\frac{v_s^2}{c^2}$; therefore the conclusion of Theorem 4.2 is valid for the cubic-quintic Schrödinger equation.

Remark. The proof of nonexistence of supersonic and sonic traveling waves for equations of type (1.1) relies on identity (3.1), combined with Pohozaev identities. We have proved (3.1) in an "indirect" way, starting from (3.11), using the Fourier transform and analyzing the behavior near the origin of the symbols of the differential operators involved. A natural question is whether (3.1) could be proved "directly" by multiplying the equations by appropriate functions and integrating by parts (and it is very tempting to try to do so because of the form of (3.7) and (3.8)!). We suspect that it is not possible to find such a proof, a heuristical reason being the following: If a "direct" proof of (3.1) could be found, it should be valid for any value of c.

Since Pohozaev identities are also valid for any c, one could infer that, for any c, equation (4.1) and system (4.17)–(4.18) below do not admit nontrivial finite-energy traveling waves. However, in the case of the Gross–Pitaevskii equation the existence of nontrivial, finite-energy traveling waves moving with sufficiently small speed c has been proved in [5] in dimension $N = 2$, respectively, in [7] and [12] in dimension $N = 3$. In a recent work [6], existence of traveling waves has been proved in space dimensions $N = 2$ and $N = 3$ for a wider range of speeds, including speeds c close to (and less than) v_s if $N = 2$. For Schrödinger equations of cubic-quintic type, the existence of small velocity traveling waves has been proved in [25] in any space dimension $N \geq 4$. Even for these particular cases, the question of whether such solutions exist for any speed $c \in (-v_s, v_s)$ is, to our knowledge, still open.

4.2. A Gross–Pitaevskii–Schrödinger system. Our second application concerns the system

$$i\frac{\partial \Psi}{\partial t} + \Delta \Psi - \frac{1}{\varepsilon^2}\left(|\Psi|^2 + \frac{1}{\varepsilon^2}|\Phi|^2 - 1\right)\Psi = 0 \qquad \text{in } \mathbf{R}^N, \tag{4.17}$$

$$i\delta\frac{\partial \Phi}{\partial t} + \Delta \Phi - \frac{1}{\varepsilon^2}(q^2|\Psi|^2 - \varepsilon^2 k^2)\Phi = 0 \qquad \text{in } \mathbf{R}^N, \tag{4.18}$$

which describes the motion of an uncharged impurity in a Bose condensate (see [16]). Here Ψ and Φ are the wavefunctions for bosons, respectively, for the impurity, and ε, δ, q, k are dimensionless physical constants. Assuming that the condensate is at rest at infinity, the functions Ψ and Φ must satisfy the "boundary conditions" $|\Psi| \longrightarrow 1$ and $|\Phi| \longrightarrow 0$ as $|x| \longrightarrow \infty$.

System (4.17)–(4.18) has a Hamiltonian structure, the associated energy being

$$E(\Psi, \Phi) = \int_{\mathbf{R}^N} |\nabla \Psi|^2 + \frac{1}{\varepsilon^2 q^2}|\nabla \Phi|^2 + \frac{1}{2\varepsilon^2}(|\Psi|^2 - 1)^2 + \frac{1}{\varepsilon^4}|\Psi|^2|\Phi|^2 - \frac{k^2}{\varepsilon^2 q^2}|\Phi|^2\, dx. \tag{4.19}$$

We are interested in traveling wave solutions for (4.17)–(4.18), i.e., solutions of the form $\Psi(x,t) = \psi(x_1 - ct, x_2, \ldots, x_N)$, $\Phi(x,t) = \varphi(x_1 - ct, x_2, \ldots, x_N)$. Such solutions must satisfy the equations

$$-ic\frac{\partial \psi}{\partial x_1} + \Delta \psi - \frac{1}{\varepsilon^2}\left(|\psi|^2 + \frac{1}{\varepsilon^2}|\varphi|^2 - 1\right)\psi = 0, \tag{4.20}$$

$$-ic\delta\frac{\partial \varphi}{\partial x_1} + \Delta \varphi - \frac{1}{\varepsilon^2}(q^2|\psi|^2 - \varepsilon^2 k^2)\varphi = 0, \tag{4.21}$$

together with the boundary conditions $|\psi| \longrightarrow 1$ and $|\varphi| \longrightarrow 0$ as $|x| \longrightarrow \infty$.

Equation (4.17) is of type (1.1). In view of the analysis in the introduction, the associated sound velocity at infinity is $\frac{\sqrt{2}}{\varepsilon}$.

In one space dimension, system (4.20)–(4.21) with the considered boundary conditions has been studied in [26]. It was proved that it admits nontrivial solutions if c is less than the sound velocity at infinity; in this case the structure of the set of traveling waves has been investigated, and it was proved that it contains global subcontinua in appropriate (weighted) Sobolev spaces.

Here we study the finite-energy traveling waves for (4.17)–(4.18) in dimension $N \geq 2$. In view of (4.19), by *finite-energy traveling wave* we mean a couple of functions

$(\psi,\varphi)\in L^1_{loc}(\mathbf{R}^N)\times L^1_{loc}(\mathbf{R}^N)$ which satisfy (4.20)–(4.21) in $\mathcal{D}'(\mathbf{R}^N)$, the boundary conditions $|\psi|\longrightarrow 1$, $\varphi\longrightarrow 0$ as $|x|\longrightarrow\infty$, and such that $\nabla\psi$, $\nabla\varphi$, $\varphi\in L^2(\mathbf{R}^N)$, $(|\psi|^2-1)^2+\frac{2}{\varepsilon^2}|\psi|^2|\varphi|^2\in L^1(\mathbf{R}^N)$. As before, we denote $\psi_1=\mathrm{Re}(\psi)$, $\psi_2=\mathrm{Im}(\psi)$, $\varphi_1=\mathrm{Re}(\varphi)$, $\varphi_2=\mathrm{Im}(\varphi)$. We have the following.

PROPOSITION 4.6. *Let $c\in\mathbf{R}$ and let (ψ,φ) be a finite-energy traveling wave for* (4.17)–(4.18). *Then the following hold.*

(i) *The function ψ is bounded and C^∞ and $\varphi,\nabla\psi\in W^{k,p}(\mathbf{R}^N)$ for any $k\in\mathbf{N}$ and $p\geq 2$.*

(ii) *There exist $R_*\geq 0$ and a real-valued function θ such that $\psi=|\psi|e^{i\theta}$ on $\mathbf{R}^N\setminus B(0,R_*)$ and $\nabla\theta\in W^{k,p}(\mathbf{R}^N\setminus B(0,R_*))$ for any $k\in\mathbf{N}$ and $p\geq 2$.*

(iii) *Let $\chi\in C^\infty(\mathbf{R}^N)$ be a cut-off function such that $\chi=0$ on $B(0,2R_*)$ and $\chi=1$ on $\mathbf{R}^N\setminus B(0,3R_*)$. We have $\psi_1\frac{\partial\psi_2}{\partial x_1}-\psi_2\frac{\partial\psi_1}{\partial x_1}-\frac{\partial}{\partial x_1}(\chi\theta)\in L^1(\mathbf{R}^N)$ and the following Pohozaev-type identities hold:*

$$\begin{aligned}&\int_{\mathbf{R}^N}-\left|\frac{\partial\psi}{\partial x_1}\right|^2-\frac{1}{\varepsilon^2q^2}\left|\frac{\partial\varphi}{\partial x_1}\right|^2+\sum_{j=2}^N\left(\left|\frac{\partial\psi}{\partial x_j}\right|^2+\frac{1}{\varepsilon^2q^2}\left|\frac{\partial\varphi}{\partial x_j}\right|^2\right)dx\\&+\int_{\mathbf{R}^N}\frac{1}{2\varepsilon^2}(|\psi|^2-1)^2+\frac{1}{\varepsilon^4}|\psi|^2|\varphi|^2-\frac{k^2}{\varepsilon^2q^2}|\varphi|^2\,dx=0,\end{aligned}\tag{4.22}$$

and for any $k\in\{2,\dots,N\}$,

$$\begin{aligned}&\int_{\mathbf{R}^N}-\left|\frac{\partial\psi}{\partial x_k}\right|^2-\frac{1}{\varepsilon^2q^2}\left|\frac{\partial\varphi}{\partial x_k}\right|^2+\sum_{j=1,\,j\neq k}^N\left(\left|\frac{\partial\psi}{\partial x_j}\right|^2+\frac{1}{\varepsilon^2q^2}\left|\frac{\partial\varphi}{\partial x_j}\right|^2\right)dx\\&+\int_{\mathbf{R}^N}\frac{1}{2\varepsilon^2}(|\psi|^2-1)^2+\frac{1}{\varepsilon^4}|\psi|^2|\varphi|^2-\frac{k^2}{\varepsilon^2q^2}|\varphi|^2\,dx\\&-c\int_{\mathbf{R}^N}\psi_1\frac{\partial\psi_2}{\partial x_1}-\psi_2\frac{\partial\psi_1}{\partial x_1}-\frac{\partial}{\partial x_1}(\chi\theta)\,dx-\frac{2c\delta}{\varepsilon^2q^2}\int_{\mathbf{R}^N}\varphi_1\frac{\partial\varphi_2}{\partial x_1}\,dx=0.\end{aligned}\tag{4.23}$$

Proof. Putting $F(x,s)=-\frac{1}{\varepsilon^2}(s+\frac{1}{\varepsilon^2}|\varphi(x)|^2-1)$, equation (4.20) is a particular case of (1.7). Clearly, in this case we have $r_0=1$.

It is obvious that F satisfies the assumptions (H1a) and (H1b) in section 2. Clearly, $F(x,s)\leq-\frac{1}{\varepsilon^2}(s-1)\leq-\frac{1}{2\varepsilon^2}s$ for any $s\geq 2$ and $x\in\mathbf{R}^N$, and hence F satisfies (H2) for $r_*=2$. Moreover, $\int_{r_0^2}^{r_*}F(x,\tau)\,d\tau=-\frac{1}{\varepsilon^2}(\frac{1}{2}+\frac{1}{\varepsilon^2}|\varphi(x)|^2)$ is a locally integrable function of x. We have $|\psi|^4\leq 2(|\psi|^2-1)^2+2$ and $(|\psi|^2-1)^2\in L^1(\mathbf{R})$ because (ψ,φ) has finite energy, and hence $\psi\in L^4_{loc}(\mathbf{R}^N)$. We also have $||\varphi|^2\psi|\leq\frac{1}{2}(|\varphi|^2+|\varphi|^2|\psi|^2)$ and $|\varphi|^2$, $|\varphi|^2|\psi|^2\in L^1(\mathbf{R})$. It is then clear that $F(\cdot,|\psi|^2)\psi=-\frac{1}{\varepsilon^2}|\psi|^2\psi-\frac{1}{\varepsilon^4}|\varphi|^2\psi+\frac{1}{\varepsilon^2}\psi$ belongs to $L^1_{loc}(\mathbf{R}^N)$. Hence we may use Proposition 2.2 (i) and infer that $\psi\in L^\infty(\mathbf{R}^N)$.

By hypothesis we have $\varphi\in L^2(\mathbf{R}^N)$ and $\nabla\varphi\in L^2(\mathbf{R}^N)$, that is, $\varphi\in W^{1,2}(\mathbf{R}^N)$. Assume that $\varphi\in W^{1,p}(\mathbf{R}^N)$ for some $p\in(1,\infty)$. Since ψ is bounded, by (4.21) we find $\Delta\varphi\in L^p(\mathbf{R}^N)$, and we infer that $\varphi\in W^{2,p}(\mathbf{R}^N)$. If $p<N$, by the Sobolev embedding we have $\varphi\in L^{p^*}(\mathbf{R}^N)$ and $\nabla\varphi\in L^{p^*}(\mathbf{R}^N)$ (where $\frac{1}{p^*}=\frac{1}{p}-\frac{1}{N}$), and hence $\varphi\in W^{1,p^*}(\mathbf{R}^N)$. Repeating the above argument if necessary, after a finite number of steps we find $\varphi\in W^{2,q}(\mathbf{R}^N)$ for some $q\geq N$ and the Sobolev embedding implies $\varphi\in L^r(\mathbf{R}^N)$ and $\nabla\varphi\in L^r(\mathbf{R}^N)$ for any $r\in[q,\infty)$. Using (4.21) again, we conclude that $\Delta\varphi\in L^r(\mathbf{R}^N)$, and hence $\varphi\in W^{2,r}(\mathbf{R}^N)$ for any $r\in[2,\infty)$.

It follows that $\varphi \in C^1(\mathbf{R}^N)$, which implies $F \in C^1(\mathbf{R}^N)$ (and consequently F satisfies (H1c)). By Proposition 2.2 (ii) we get $\psi \in W^{3,p}_{loc}(\mathbf{R}^N)$ for any $p \in [1,\infty)$. In particular, $\psi \in C^2(\mathbf{R}^N)$.

We have $F(x,1) = -\frac{1}{\varepsilon^4}|\varphi(x)|^2$, and F clearly satisfies assumption (H3). It is obvious that $\partial_{N+1}F(x,s) = -\frac{1}{\varepsilon^2}$ and $\partial^2_{N+1}F(x,s) = 0$ on $\mathbf{R}^N \times \mathbf{R}_+$, and therefore F satisfies (H4) and (H5). Thus we may use Proposition 2.5 (i) and infer that $\nabla\psi \in W^{1,p}(\mathbf{R}^N)$ for any $p \in [2,\infty)$.

The rest of the proof is a very easy induction. For $k \in \mathbf{N}^*$, assume that $\nabla\psi \in W^{k,p}(\mathbf{R}^N)$ and $\varphi \in W^{k+1,p}(\mathbf{R}^N)$ for any $p \in [2,\infty)$. Consider $\alpha \in \mathbf{N}^N$ such that $|\alpha| = k$. Differentiating (4.20) and (4.21) we obtain, respectively,

$$\Delta(\partial^\alpha\psi) = ic\partial^\alpha\frac{\partial\psi}{\partial x_1} + \frac{1}{\varepsilon^2}\partial^\alpha\Big(\Big(|\psi|^2 + \frac{1}{\varepsilon^2}|\varphi|^2 - 1\Big)\psi\Big),$$
$$\Delta(\partial^\alpha\varphi) = ic\delta\partial^\alpha\frac{\partial\varphi}{\partial x_1} + \frac{1}{\varepsilon^2}\partial^\alpha\Big((q^2|\psi|^2 - \varepsilon^2k^2)\varphi\Big).$$

We infer that $\Delta(\partial^\alpha\psi),\ \Delta(\partial^\alpha\varphi) \in L^p(\mathbf{R}^N)$ for any $p \in [2,\infty)$. By hypothesis we have $\partial^\alpha\psi,\ \partial^\alpha\varphi \in L^p(\mathbf{R}^N)$, and therefore $\partial^\alpha\psi,\ \partial^\alpha\varphi \in W^{2,p}(\mathbf{R}^N)$ for any $p \in [2,\infty)$. Since this is true for any α with $|\alpha| = k$, we have $\nabla\psi \in W^{k+1,p}(\mathbf{R}^N)$ and $\varphi \in W^{k+2,p}(\mathbf{R}^N)$. We conclude that $\nabla\psi$ and φ belong to $W^{k,p}(\mathbf{R}^N)$ for any $k \in \mathbf{N}$ and $p \in [2,\infty)$.

(ii) is an immediate corollary of Proposition 2.5 (ii).

(iii) It follows directly from Theorem 3.1 (i) that $\psi_1\frac{\partial\psi_2}{\partial x_1} - \psi_2\frac{\partial\psi_1}{\partial x_1} - \frac{\partial}{\partial x_1}(\chi\theta) \in L^1(\mathbf{R}^N)$. The proof of (4.22) and (4.23) is similar to that of (4.3) and (4.4) (multiply (4.20) by $x_j\eta_n\frac{\partial\psi}{\partial x_j}$ and (4.21) by $\frac{1}{\varepsilon^2q^2}x_j\eta_n\frac{\partial\varphi}{\partial x_j}$, where $\eta_n(x) = \eta(\frac{x}{n})$ is a cut-off function, add the resulting equalities, integrate by parts, and pass to the limit as $n \longrightarrow \infty$). We omit the details. □

We have the following result concerning the nonexistence of supersonic traveling waves for (4.17)–(4.18).

THEOREM 4.7. *Let $N \geq 2$ and let (ψ,φ) be a finite-energy traveling wave for the system* (4.17)–(4.18)*, moving with velocity c. Assume that either*

- $c^2 > \frac{2}{\varepsilon^2}$, *or*
- $N = 2$ *and* $c^2 = \frac{2}{\varepsilon^2}$, *or*
- $N \geq 3$, $c^2 = \frac{2}{\varepsilon^2}$, *and* $\psi_1\frac{\partial\psi_2}{\partial x_1} - \psi_2\frac{\partial\psi_1}{\partial x_1} \in L^{\frac{2N-1}{2N-3}}(\mathbf{R}^N)$.

Then $\varphi = 0$ and ψ is constant on $\mathbf{R}^N$.

Proof. Let θ, χ be as in Proposition 4.6 and let $F(x,s) = -\frac{1}{\varepsilon^2}(s + \frac{1}{\varepsilon^2}|\varphi(x)|^2 - 1)$. We have already seen that F satisfies assumptions (H1)–(H5), and it follows that identity (3.1) holds. Taking into account the particular form of F, this identity can be written as

$$\begin{aligned}&\int_{\mathbf{R}^N}|\nabla\psi|^2 + \frac{1}{\varepsilon^2}(|\psi|^2-1)^2 + \frac{1}{\varepsilon^4}|\varphi|^2|\psi|^2\,dx\\ &= c\left(1 - \frac{2}{\varepsilon^2c^2}\right)\int_{\mathbf{R}^N}\psi_1\frac{\partial\psi_2}{\partial x_1} - \psi_2\frac{\partial\psi_1}{\partial x_1} - \frac{\partial}{\partial x_1}(\chi\theta)\,dx.\end{aligned} \tag{4.24}$$

We take the scalar product of (4.21) by φ, then integrate the resulting equality to get

$$\int_{\mathbf{R}^N}|\nabla\varphi|^2\,dx + \frac{q^2}{\varepsilon^2}\int_{\mathbf{R}^N}|\varphi|^2|\psi|^2\,dx - k^2\int_{\mathbf{R}^N}|\varphi|^2\,dx - 2c\delta\int_{\mathbf{R}^N}\varphi_1\frac{\partial\varphi_2}{\partial x_1}\,dx = 0. \tag{4.25}$$

Summing up the identities (4.23) for $k = 2, 3, \dots, N$, we find

$$\begin{aligned}
&\int_{\mathbf{R}^N} \left|\frac{\partial\psi}{\partial x_1}\right|^2 + \frac{1}{\varepsilon^2 q^2}\left|\frac{\partial\varphi}{\partial x_1}\right|^2 + \frac{N-3}{N-1}\sum_{j=2}^{N}\left(\left|\frac{\partial\psi}{\partial x_j}\right|^2 + \frac{1}{\varepsilon^2 q^2}\left|\frac{\partial\varphi}{\partial x_j}\right|^2\right) dx \\
&+ \int_{\mathbf{R}^N} \frac{1}{2\varepsilon^2}(|\psi|^2-1)^2 + \frac{1}{\varepsilon^4}|\psi|^2|\varphi|^2 - \frac{k^2}{\varepsilon^2 q^2}|\varphi|^2\, dx \\
&- c\int_{\mathbf{R}^N} \psi_1\frac{\partial\psi_2}{\partial x_1} - \psi_2\frac{\partial\psi_1}{\partial x_1} - \frac{\partial}{\partial x_1}(\chi\theta)\, dx - \frac{2c\delta}{\varepsilon^2 q^2}\int_{\mathbf{R}^N}\varphi_1\frac{\partial\varphi_2}{\partial x_1}\, dx = 0.
\end{aligned} \tag{4.26}$$

Next we combine equalities (4.24)–(4.26) in order to eliminate the terms $\int_{\mathbf{R}^N}\varphi_1\frac{\partial\varphi_2}{\partial x_1}dx$ and $\int_{\mathbf{R}^N}\psi_1\frac{\partial\psi_2}{\partial x_1} - \psi_2\frac{\partial\psi_1}{\partial x_1} - \frac{\partial}{\partial x_1}(\chi\theta)\, dx$. We find

$$\begin{aligned}
&\frac{2}{\varepsilon^2 c^2}\int_{\mathbf{R}^N}\left|\frac{\partial\psi}{\partial x_1}\right|^2 dx + \Big(1 - \Big(1 - \frac{2}{\varepsilon^2 c^2}\Big)\frac{N-3}{N-1}\Big)\int_{\mathbf{R}^N}\sum_{j=2}^{N}\left|\frac{\partial\psi}{\partial x_j}\right|^2 dx \\
&+ \frac{2}{(N-1)\varepsilon^2 q^2}\Big(1 - \frac{2}{\varepsilon^2 c^2}\Big)\int_{\mathbf{R}^N}\sum_{j=2}^{N}\left|\frac{\partial\varphi}{\partial x_j}\right|^2 dx \\
&+ \frac{1}{2\varepsilon^2}\Big(1 + \frac{2}{\varepsilon^2 c^2}\Big)\int_{\mathbf{R}^N}(|\psi|^2-1)^2\, dx + \frac{1}{\varepsilon^4}\int_{\mathbf{R}^N}|\varphi|^2|\psi|^2\, dx = 0.
\end{aligned} \tag{4.27}$$

Obviously, all integrals in (4.27) are nonnegative. If $c^2 \geq \frac{2}{\varepsilon^2}$, all coefficients are also nonnegative, and therefore each term in (4.27) must be zero. In particular, $\int_{\mathbf{R}^N}\left|\frac{\partial\psi}{\partial x_k}\right|^2 dx = 0$ for any $k \in \{1, \dots, N\}$, which implies $\nabla\psi = 0$ on $\mathbf{R}^N$, i.e., ψ is constant. Since $\int_{\mathbf{R}^N}(|\psi|^2-1)^2\, dx = 0$, necessarily $|\psi| = 1$. We also have $0 = \int_{\mathbf{R}^N}|\varphi|^2|\psi|^2\, dx = \int_{\mathbf{R}^N}|\varphi|^2\, dx$, and hence $\varphi = 0$ on $\mathbf{R}^N$. ∎

5. The one-dimensional case. Since most of the proofs in the preceding section are not valid in space dimension $N = 1$ (in particular, we do not have identities analogous to (4.4) and (4.23)), we treat separately the one-dimensional case. It turns out that some integrations can be performed explicitly and some of the results are stronger than in higher dimensions.

Let $G : [0, \infty) \longrightarrow \mathbf{R}$ be a function satisfying the following assumption:

- **(A)** $G \in C([0,\infty))$ and there exists $r_0 > 0$ such that $G(r_0^2) = 0$. Moreover, $G \in C^1([r_0^2 - \eta, r_0^2 + \eta])$ for some $\eta > 0$ and $G'(r_0^2) = -L < 0$.

We consider the Schrödinger equation

$$i\frac{\partial\Psi}{\partial t} + \Psi_{xx} + G(|\Psi|^2)\Psi = 0 \qquad \text{in } \mathbf{R}, \tag{5.1}$$

together with the "boundary condition" $|\Psi| \longrightarrow r_0$ as $x \longrightarrow \pm\infty$. We have seen in the introduction that the sound velocity at infinity associated to (5.1) and to the considered boundary condition is $v_s = r_0\sqrt{2L}$. As usually, a traveling wave moving with velocity c is a solution of the form $\Psi(x,t) = \psi(x - ct)$. It must satisfy

$$-ic\psi' + \psi'' + G(|\psi|^2)\psi = 0 \qquad \text{in } \mathbf{R}, \qquad |\psi(x)| \longrightarrow r_0 \qquad \text{as } x \longrightarrow \pm\infty. \tag{5.2}$$

We have the following result concerning supersonic and sonic traveling waves.

THEOREM 5.1. *Let $\psi \in L^1_{loc}(\mathbf{R})$ be a solution of (5.2) in $\mathcal{D}'(\mathbf{R})$ such that $G(|\psi|^2)\psi \in L^1_{loc}(\mathbf{R})$. Assume that G satisfies* (A) *and either*

(i) $c^2 > v_s^2$, *or*

(ii) $c^2 = v_s^2$ *and, denoting* $V(s) = \int_s^{r_0^2} G(\tau)\,d\tau$ *and* $W(s) = v_s^2 s^2 - 4(s + r_0^2)V(s + r_0^2)$, *there exists* $\varepsilon > 0$ *such that one of the following conditions is verified:*

(a) $W(s) > 0$ *on* $(-\varepsilon, 0) \cup (0, \varepsilon)$;

(b) $W(s) > 0$ *on* $(-\varepsilon, 0)$ *and* $W(s) < 0$ *on* $(0, \infty)$;

(c) $W(s) > 0$ *on* $(0, \varepsilon)$ *and* $W(s) < 0$ *on* $[-r_0^2, 0)$.

Then either ψ *is constant or* $\psi(x) = r_0 e^{i(cx+\theta_0)}$, *where* θ_0 *is a real constant.*

Remark. Theorem 5.1 gives *all* supersonic and sonic traveling waves for (5.1), no matter whether their energy is finite or not (and we see that finite-energy traveling waves must be constant).

It is easy to see that W is C^2 near 0 and $W(0) = W'(0) = W''(0) = 0$. Condition (ii)(a) is satisfied, for instance, if G is C^3 near r_0^2 (this clearly implies that W is C^4 near 0) and $W'''(0) = 0$, $W^{(iv)}(0) > 0$, or equivalently $r_0^2 G''(r_0^2) = 3L$ and $4G''(r_0^2) + r_0^2 G'''(r_0^2) > 0$. The condition $W(s) > 0$ on $(-\varepsilon, 0)$ in (ii)(b), respectively, $W(s) > 0$ on $(0, \varepsilon)$ in (ii)(c), is satisfied if G is C^3 near r_0^2 and $W'''(0) < 0$ (respectively, $W'''(0) > 0$); however, in these cases only the information on the behavior of G in a neighborhood of r_0^2 is not sufficient to get the conclusion of Theorem 5.1.

Proof of Theorem 5.1. Let $\varphi(x) = e^{-\frac{icx}{2}}\psi(x)$. Then $\varphi \in L^1_{loc}(\mathbf{R})$ and it is easy to see that

$$\varphi'' + \Big(G(|\varphi|^2) + \frac{c^2}{4}\Big)\varphi = 0 \qquad \text{in } \mathcal{D}'(\mathbf{R}). \tag{5.3}$$

From (5.3) we get $\varphi'' \in L^1_{loc}(\mathbf{R})$. This implies that φ' is a continuous function on $\mathbf{R}$ (see, e.g., Lemma VIII.2, p. 123, in [8]). Thus $\varphi \in C^1(\mathbf{R})$. Since $|\varphi| \longrightarrow r_0$ as $x \longrightarrow \pm\infty$, we infer that φ is bounded on $\mathbf{R}$. Coming back to (5.3) we see that φ'' is continuous and bounded on $\mathbf{R}$. In particular $\varphi \in C^2(\mathbf{R})$, and this implies $\psi \in C^2(\mathbf{R})$.

Denoting $\psi_1 = \mathrm{Re}(\psi)$, $\psi_2 = \mathrm{Im}(\psi)$, equation (5.2) is equivalent to the system

$$c\psi_2' + \psi_1'' + G(|\psi|^2)\psi_1 = 0, \tag{5.4}$$

$$-c\psi_1' + \psi_2'' + G(|\psi|^2)\psi_2 = 0 \qquad \text{in } \mathbf{R}. \tag{5.5}$$

We multiply (5.4) by $2\psi_1'$ and (5.5) by $2\psi_2'$, then add the resulting equalities to get $[(\psi_1')^2 + (\psi_2')^2]' - (V(|\psi|^2))' = 0$. Hence there exists $k_1 \in \mathbf{R}$ such that

$$|\psi'|^2(x) - V(|\psi|^2)(x) = k_1 \qquad \text{for any } x \in \mathbf{R}. \tag{5.6}$$

Multiplying (5.4) by ψ_2 and (5.5) by $-\psi_1$, then summing up the corresponding equations we obtain $\frac{c}{2}(|\psi|^2 - r_0^2)' - (\psi_1\psi_2' - \psi_2\psi_1')' = 0$. Consequently there is some $k_2 \in \mathbf{R}$ such that

$$\frac{c}{2}(|\psi|^2 - r_0^2) - (\psi_1\psi_2' - \psi_2\psi_1') = k_2 \qquad \text{in } \mathbf{R}. \tag{5.7}$$

Next we multiply (5.4) by $2\psi_1$ and (5.5) by $2\psi_2$, then add the resulting equalities to find

$$2c(\psi_1\psi_2' - \psi_2\psi_1') + (|\psi|^2 - r_0^2)'' - 2|\psi'|^2 + 2G(|\psi|^2)|\psi|^2 = 0. \tag{5.8}$$

Taking into account (5.6) and (5.7), equation (5.8) can be written as

$$(|\psi|^2 - r_0^2)'' + c^2(|\psi|^2 - r_0^2) - 2V(|\psi|^2) + 2G(|\psi|^2)|\psi|^2 = 2k_1 + 2ck_2. \tag{5.9}$$

Denote $v(x) = |\psi|^2(x) - r_0^2$. Then v is real-valued, C^2, and tends to zero as $x \longrightarrow \pm\infty$, and hence there exists a sequence $x_n \longrightarrow \infty$ such that $v''(x_n) \longrightarrow 0$. Writing (5.9) for x_n, then passing to the limit as $n \longrightarrow \infty$ we see that necessarily $k_1 + ck_2 = 0$ and v satisfies the equation

$$v'' + c^2 v - 2V(v + r_0^2) + 2(v + r_0^2)G(v + r_0^2) = 0 \qquad \text{in } \mathbf{R}. \tag{5.10}$$

Next we multiply (5.10) by $2v'$, then integrate the resulting equation and obtain $(v')^2 + c^2v^2 - 4(v + r_0^2)V(v + r_0^2) = k_3$ in $\mathbf{R}$, where k_3 is a constant. It is clear that there exists a sequence $y_n \longrightarrow \infty$ such that $v'(y_n) \longrightarrow 0$; consequently $k_3 = 0$ and we have

$$(v')^2(x) + c^2v^2(x) - 4(v + r_0^2)V(v + r_0^2)(x) = 0 \qquad \text{for any } x \in \mathbf{R}. \tag{5.11}$$

Our aim is to prove that, under the assumptions of Theorem 5.1, we have $v = 0$ on $\mathbf{R}$.

Suppose first that $c^2 > v_s^2 = 2Lr_0^2$. Since G satisfies (A), it follows that $V \in C^2([r_0^2 - \eta, r_0^2 + \eta])$ and we have by Taylor's formula

$$V(r_0^2+s) = V(r_0^2)+sV'(r_0^2)+\frac{1}{2}s^2V''(r_0^2)+s^2h(s) = \frac{1}{2}Ls^2+s^2h(s) \qquad \text{for } s \in [-\eta, \eta],$$

where $h(s) \longrightarrow 0$ as $s \longrightarrow 0$. Take $\varepsilon_1 \in (0, \eta]$ such that $c^2 - v_s^2 - 2Ls - 4(s+r_0^2)h(s) > 0$ for any $s \in [-\varepsilon_1, \varepsilon_1]$. Suppose that $v(x_0) \in [-\varepsilon_1, 0) \cup (0, \varepsilon_1]$ for some $x_0 \in \mathbf{R}$. By (5.11) we obtain

$$0 = (v')^2(x_0) + v^2(x_0)[c^2 - v_s^2 - 2Lv(x_0) - 4(v(x_0) + r_0^2)h(v(x_0))] > 0,$$

a contradiction. Consequently we cannot have $v(x) \in [-\varepsilon_1, 0) \cup (0, \varepsilon_1]$. Since v is continuous and $v(x) \longrightarrow 0$ as $x \longrightarrow \pm\infty$, we infer that necessarily $v(x) = 0$ for any $x \in \mathbf{R}$.

Next assume that $c^2 = v_s^2$. Equation (5.11) can be written as

$$(v')^2(x) + W(v(x)) = 0 \qquad \text{on } \mathbf{R}. \tag{5.12}$$

If assumption (iia) is verified, we cannot have $v(x) \in (-\varepsilon, 0) \cup (0, \varepsilon)$ and we infer, as above, that $v = 0$ on $\mathbf{R}$. In case (iib), we cannot have $v(x) \in (-\varepsilon, 0)$ and we infer that $v(x) \geq 0$ for any $x \in \mathbf{R}$. Since $v(x) \longrightarrow 0$ as $x \longrightarrow \infty$, there is some x_0 such that v achieves a nonnegative maximum at x_0. Then $v'(x_0) = 0$ and from (5.12) we get $W(v(x_0)) = 0$. But $W(s) < 0$ for $s > 0$ by (iib); hence $v(x_0) = 0$ and consequently $v = 0$ on $\mathbf{R}$. Similarly we have $v = 0$ in the case (iic) (note that $v = |\psi|^2 - r_0^2 \geq -r_0^2$ and it suffices to know that $W < 0$ on $[-r_0^2, 0)$).

Thus we always have $v = 0$, that is, $|\psi|^2 = r_0^2$ on $\mathbf{R}$. Consequently there exists a lifting $\theta \in C^2(\mathbf{R}, \mathbf{R})$ such that $\psi(x) = r_0 e^{i\theta(x)}$ for any $x \in \mathbf{R}$. It is clear that $\psi_1\psi_2' - \psi_2\psi_1' = |\psi|^2\theta' = r_0^2\theta'$ (see (3.3)). On the other hand we have $\psi_1\psi_2' - \psi_2\psi_1' = -k_2$ by (5.7), and hence $\theta' = -\frac{k_2}{r_0^2}$ is constant; therefore $\theta(x) = -\frac{k_2}{r_0^2}x + \theta_0$, where θ_0 is a real constant. Since $\psi = r_0 e^{i(-(k_2/r_0^2)x + \theta_0)}$ satisfies (5.2), we find $-c\frac{k_2}{r_0^2} - \left(\frac{k_2}{r_0^2}\right)^2 = 0$, and thus either $\frac{k_2}{r_0^2} = 0$ or $\frac{k_2}{r_0^2} = -c$. Finally we have either $\psi(x) = e^{i\theta_0}$ or $\psi(x) = e^{i(cx+\theta_0)}$ and the proof is complete. □

Example 5.2. In the case of the Gross–Pitaevskii equation we have $G(s) = 1 - s$ and obtain $W(s) = -2s^3$ (see Example 4.4). In the case of the cubic-quintic nonlinearity we have $G(s) = -\alpha_5(s - r_1^2)(s - r_0^2)$, where $\alpha_5 > 0$, $0 < r_1 < r_0$ (see Example 4.5)

and a simple computation gives $W(s) = -2\alpha_5 s^3(\frac{4}{3}r_0^2 - r_1^2 + \frac{1}{3}s)$. Therefore both the Gross–Pitaevskii and the cubic-quintic nonlinearities satisfy assumption (iib) and Theorem 5.1 gives all sonic and supersonic traveling waves for these equations.

Remark. The proof of Theorem 5.1 provides a method to find subsonic traveling waves for (5.1). With the above notation, it follows from (5.11) that on any interval where $v' \neq 0$ we have $v'(x) = \pm\sqrt{4(v + r_0^2)V(v + r_0^2)(x) - c^2v^2(x)}$. In many interesting applications this equation can be integrated, and we obtain explicitly $v = |\psi|^2 - r_0^2$. Then it is not hard to find (up to a constant) the corresponding phase θ.

Remark. Assume that $N = 1$ and let (ψ, φ) be a finite-energy traveling wave for system (4.17)–(4.18). It follows from the proof of Proposition 4.6 that ψ and φ are C^∞ functions and $\psi', \varphi \in W^{k,p}(\mathbf{R})$ for any $k \in \mathbf{N}$ and $p \geq 2$. If $c^2 \geq \frac{2}{\varepsilon^2}$ (recall that $\frac{\sqrt{2}}{\varepsilon}$ is the sound velocity at infinity associated to (3.21)–(3.22)) and if there is a lifting $\psi(x) = v(x)e^{i\alpha(x)}$, $\varphi(x) = u(x)e^{i\beta(x)}$, where v, u, α, β are real-valued functions of class C^2, [26, Proposition 3.1, p. 1545] implies that $v = 1$, α is constant, and $\varphi = 0$ on $\mathbf{R}$.

REFERENCES

[1] I. V. Barashenkov and V. G. Makhankov, *Soliton-like "bubbles" in a system of interacting bosons*, Phys. Lett. A, 128 (1988), pp. 52–56.

[2] I. V. Barashenkov, A. D. Gocheva, V. G. Makhankov, and I. V. Puzynin, *Stability of soliton-like bubbles*, Phys. D, 34 (1989), pp. 240–254.

[3] H. Berestycki and P.-L. Lions, *Nonlinear scalar field equations*, I. *Existence of a ground state*, Arch. Rational Mech. Anal., 82 (1983), pp. 313–345.

[4] J. Bergh and J. Löfström, *Interpolation Spaces. An Introduction*, Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[5] F. Béthuel and J.-C. Saut, *Travelling-waves for the Gross-Pitaevskii equation* I, Ann. Inst. H. Poincaré Phys. Théor., 70 (1999), pp. 147–238.

[6] F. Béthuel, P. Gravejat, and J.-C. Saut, *Travelling-Waves for the Gross-Pitaevskii Equation* II, preprint; available online from http://arxiv.org/abs/0711.2408v1.

[7] F. Béthuel, G. Orlandi, and D. Smets, *Vortex rings for the Gross-Pitaevskii equation*, J. Eur. Math. Soc. (JEMS), 6 (2004), pp. 17–94.

[8] H. Brézis, *Analyse Fonctionnelle*, Masson, Paris, 1983.

[9] H. Brézis, *Semilinear equations in* $\mathbf{R}^N$ *without condition at infinity*, Appl. Math. Optim., 12 (1984), pp. 271–282.

[10] H. Brézis and E. H. Lieb, *Minimum action solutions for some vector field equations*, Comm. Math. Phys., 96 (1984), pp. 97–113.

[11] H. Brézis, J. Bourgain, and P. Mironescu, *Lifting in Sobolev spaces*, J. Anal. Math., 80 (2000), pp. 37–86.

[12] D. Chiron, *Travelling-waves for the Gross-Pitaevskii equation in dimension larger than two*, Nonlinear Anal., 58 (2004), pp. 175–204.

[13] A. Farina, *Finite-energy solutions, quantization effects and Liouville-type results for a variant of the Ginzburg-Landau systems in* $\mathbf{R}^k$, Differential Integral Equations, 11 (1998), pp. 875–893.

[14] A. Farina, *From Ginzburg-Landau to Gross-Pitaevskii*, Monatsh. Math., 139 (2003), pp. 265–269.

[15] P. Gérard, *The Cauchy problem for the Gross-Pitaevskii equation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 23 (2006), pp. 765–779.

[16] J. Grant and P. H. Roberts, *Motions in a Bose condensate* III. *The structure and effective masses of charged and uncharged impurities*, J. Phys. A, 7 (1974), pp. 260–279.

[17] P. Gravejat, *A non-existence result for supersonic travelling-waves in the Gross-Pitaevskii equation*, Comm. Math. Phys., 243 (2003), pp. 93–103.

[18] P. Gravejat, *Limit at infinity and non-existence results for sonic travelling-waves in the Gross-Pitaevskii equation*, Differential Integral Equations, 17 (2004), pp. 1213–1232.

[19] E. P. Gross, *Hydrodynamics of a superfluid condensate*, J. Math. Phys., 4 (1963), pp. 195–207.

[20] L. Hörmander, *The Analysis of Linear Partial Differential Operators, Vol.* 1, Springer-Verlag, Berlin, 1983.

[21] C. A. Jones and P. H. Roberts, *Motions in a Bose condensate* IV. *Axisymmetric solitary waves*, J. Phys. A, 15 (1982), pp. 2599–2619.

[22] C. A. Jones, S. J. Putterman, and P. H. Roberts, *Motions in a Bose condensate* V. *Stability of wave solutions of nonlinear Schrödinger equations in two and three dimensions*, J. Phys. A, 19 (1986), pp. 2991–3011.

[23] T. Kato, *Schrödinger operators with singular potentials*, Israel J. Math., 13 (1972), pp. 135–148.

[24] P. I. Lizorkin, *On multipliers of Fourier integrals in the spaces $L_{p,\theta}$*, Proc. Steklov Inst. Math., 89 (1967), pp. 269–290.

[25] M. Mariş, *Existence of nonstationary bubbles in higher dimensions*, J. Math. Pures Appl., 81 (2002), pp. 1207–1239.

[26] M. Mariş, *Global branches of travelling-waves to a Gross–Pitaevskii–Schrödinger system in one dimension*, SIAM J. Math. Anal., 37 (2006), pp. 1535–1559.

[27] E. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.

SIAM J. MATH. ANAL.
Vol. 40, No. 3, pp. 1104–1122

EULERIAN CALCULUS FOR THE DISPLACEMENT CONVEXITY IN THE WASSERSTEIN DISTANCE*

SARA DANERI† AND GIUSEPPE SAVARÉ‡

Abstract. In this paper we give a new proof of the (strong) displacement convexity of a class of integral functionals defined on a compact Riemannian manifold satisfying a lower Ricci curvature bound. Our approach does not rely on existence and regularity results for optimal transport maps on Riemannian manifolds, but it is based on the Eulerian point of view recently introduced by Otto and Westdickenberg [*SIAM J. Math. Anal.*, 37 (2005), pp. 1227–1255] and on the metric characterization of the gradient flows generated by the functionals in the Wasserstein space.

Key words. gradient flows, displacement convexity, heat and porous medium equation, nonlinear diffusion, optimal transport, Kantorovich-Rubinstein-Wasserstein distance, Riemannian manifolds with a lower Ricci curvature bound

AMS subject classifications. 58J35, 58C21, 49J40

DOI. 10.1137/08071346X

1. Introduction.

In this paper we give a new proof, based on a gradient flow approach and on the Eulerian point of view introduced by [19], of the so-called "displacement convexity" for integral functionals as

$$\mathcal{E}(\mu) := \int_{\mathbb{M}} e(\rho)\,\mathrm{dV} + e'(\infty)\,\mu^{\perp}(\mathbb{M}), \quad \rho = \frac{\mathrm{d}\mu}{\mathrm{dV}}, \tag{1.1}$$

where μ is a Borel probability measure on a compact, connected Riemannian manifold without boundary $(\mathbb{M}, \mathsf{g})$, V is the volume measure on $\mathbb{M}$ induced by the metric tensor g, $\mu^{\perp}$ is the singular part of μ with respect to V, $e : [0, +\infty) \to \mathbb{R}$ is a smooth convex function satisfying the so-called McCann conditions (see (1.7) below), and $e'(\infty) = \lim_{r\to+\infty} \frac{e(r)}{r}$. When e has a superlinear growth, $e'(\infty) = +\infty$ so that μ should be absolutely continuous with respect to V when $\mathcal{E}(\mu)$ is finite.

Displacement convexity for integral functionals. The notion of *displacement convexity* has been introduced by [15] to study the behavior of integral functionals like (1.1) along optimal transportation paths, i.e., geodesics in the space of Borel probability measures $\mathcal{P}(\mathbb{M})$ endowed with the L^2-Kantorovich-Rubinstein-Wasserstein distance.

Recall that (the square of) this distance can be defined by the following optimal transport problem

$$\begin{aligned} W_2^2(\mu^0, \mu^1) := \min \Bigg\{ & \int_{\mathbb{M}\times\mathbb{M}} \mathsf{d}^2(x, y)\,\mathrm{d}\boldsymbol{\sigma}(x, y) : \boldsymbol{\sigma} \in \mathcal{P}(\mathbb{M}\times\mathbb{M}), \\ & \boldsymbol{\sigma}(B\times\mathbb{M}) = \mu^0(B),\ \boldsymbol{\sigma}(\mathbb{M}\times B) = \mu^1(B) \quad \forall\, B \text{ Borel set in } \mathbb{M} \Bigg\} \end{aligned} \tag{1.2}$$

*Received by the editors January 15, 2008; accepted for publication (in revised form) May 1, 2008; published electronically October 17, 2008.
http://www.siam.org/journals/sima/40-3/71346.html
†S.I.S.S.A., Via Beirut 2-4, 34014, Trieste, Italy (daneri@sissa.it).
‡Università di Pavia, Department of Mathematics, Via Ferrata 1, 27100, Pavia, Italy (giuseppe.savare@unipv.it). Partially supported by grants of M.I.U.R., PRIN '06.

for the cost function induced by the Riemannian distance d on the manifold $\mathbb{M}$. We keep the usual notation to denote by $\mathcal{P}_2(\mathbb{M})$ the metric space $(\mathcal{P}(\mathbb{M}), W_2)$, which is called Wasserstein space; being $\mathbb{M}$ compact, W_2 induces the topology of the weak convergence of probability measures (i.e., the weak* topology associated with the duality of $\mathcal{P}(\mathbb{M})$ with $C^0(\mathbb{M})$).

As in any metric space (minimal, constant speed) *geodesics* can be defined as curves $\mu : s \in [0,1] \mapsto \mu^s \in \mathcal{P}_2(\mathbb{M})$ between μ^0 and μ^1, satisfying

$$W_2(\mu^r, \mu^s) = |s-r|\, W_2\left(\mu^0, \mu^1\right) \quad \forall\ 0 \le r \le s \le 1. \tag{1.3}$$

A functional $\mathcal{E} : \mathcal{P}(\mathbb{M}) \to (-\infty, +\infty]$ is then *(strongly) displacement convex* (or, more generally, *(strongly) displacement λ-convex* for some $\lambda \in \mathbb{R}$) if, *for all* Wasserstein geodesics $\{\mu^s\}_{0\le s\le 1} \subset \mathcal{P}_2(\mathbb{M})$, we have

$$\mathcal{E}(\mu^s) \le (1-s)\mathcal{E}\left(\mu^0\right) + s\mathcal{E}\left(\mu^1\right) - \frac{\lambda}{2} s(1-s) W_2^2\left(\mu^0, \mu^1\right), \quad \forall\, s \in [0,1]. \tag{1.4}$$

A weaker notion is also often considered: one can ask that there exists *at least one* geodesic connecting μ^0 to μ^1 along which (1.4) holds. Figalli and Villani [12] has recently proved that for the class of integral functionals (1.1) considered in the present paper these two notions are equivalent.

The term "displacement convexity" arises from the strictly related concept of "displacement interpolation" introduced by [15] in the Euclidean case $\mathbb{M} = \mathbb{R}^d$; in a general metric setting, property (1.4) is simply called, as in the Riemannian case, "$\lambda-$geodesic convexity" (or "geodesic convexity" if $\lambda = 0$).

It is possible to show [4] that, in the Euclidean case, the measures μ^s can also be defined through the formula

$$\mu^s(B) := \boldsymbol{\sigma}\left(\left\{(x,y) \in \mathbb{R}^d \times \mathbb{R}^d : (1-s)x + sy \in B\right\}\right), \quad \text{where } \boldsymbol{\sigma} \text{ is a minimizer of (1.2).} \tag{1.5}$$

A similar construction can also be performed in a Riemannian manifold [14, 20, 13]: the segments $s \mapsto (1-s)x + sy$ should be substituted by a Borel map $\boldsymbol{\gamma} : \mathbb{M} \times \mathbb{M} \to C^0([0,1]; \mathbb{M})$ that at each couple $(x,y) \in \mathbb{M} \times \mathbb{M}$ associate a (minimal, constant speed) geodesic $s \mapsto \boldsymbol{\gamma}^s(x,y)$ in $\mathbb{M}$ connecting x to y. We have the representation formula

$$\mu^s(B) := \boldsymbol{\sigma}\left(\{(x,y) \in \mathbb{M} \times \mathbb{M} : \boldsymbol{\gamma}^s(x,y) \in B\}\right), \quad \text{where } \boldsymbol{\sigma} \text{ is a minimizer of (1.2).} \tag{1.6}$$

After the pioneering paper [15], the notion of displacement convexity for integral functionals found applications in many different fields, as functional inequalities [18, 2, 9], generation, contraction, and asymptotic properties of diffusion equations and gradient flows [17, 1, 19, 4, 8, 5], Riemannian geometry, and synthetic study of metric-measure spaces [20, 21, 14].

In the context of Riemannian manifolds it turns out that displacement λ-convexity of certain classes of entropy functionals is equivalent to a lower bound for the Ricci curvature of the manifold. The connection between displacement convexity and Ricci curvature, introduced by [18], was then further deeply studied by [9, 10, 20]; the equivalence has been proved by Sturm and Von Renesse in [24], who considered the case in which the domain of the functional consists only of measures that are absolutely continuous with respect to the volume measure, and then completed by Lott and Villani [14] (with the remarks made in [12], where convexity in the strong form has

been proved), who extended the previous results to the functionals defined by (1.1) on all $\mathcal{P}(\mathbb{M})$. We refer to the forthcoming monograph [23] for further references, details, and discussions.

The strategy followed by the authors of [9] (and by all the following contributions) in order to characterize the displacement convexity of entropy functionals relies on a characterization of optimal transportation and Wasserstein geodesics [16] and on a careful study of the Jacobian properties of the exponential function which are crucial to estimate the integral functionals along this class of curves. The lack of regularity of Wasserstein geodesics and the lack of global smoothness of the squared distance function d^2 on the manifold $\mathbb{M}$ (due to the existence of the cut-locus) require a careful use of nonsmooth analysis arguments and nontrivial approximation processes to extend the results to geodesics between arbitrary measures (see [14, 12]).

The main result is the following.

THEOREM 1.1. (I) *If $e \in C^\infty(0,+\infty)$ satisfies the McCann conditions*

(1.7)
$$U(\rho) := \rho e'(\rho) - \big(e(\rho) - e(0_+)\big) \geq 0, \qquad \rho U'(\rho) - \left(1 - \frac{1}{n}\right) U(\rho) \geq 0, \quad n := \dim(\mathbb{M}) > 1$$

and $\mathbb{M}$ has nonnegative *Ricci curvature, then the functional $\mathcal{E}$ defined by* (1.1) *is* (strongly) displacement convex.

(II) *If $\mathcal{E}$ is the relative entropy functional, corresponding to $e(\rho) = \rho \log \rho$ (which satisfies* (1.7) *in any dimension) in* (1.1)*, and there exists $\lambda \in \mathbb{R}$ such that*

$$\mathrm{Ric}_{\mathsf{g}_x}(\xi,\xi) \geq \lambda \langle \xi, \xi \rangle_{\mathsf{g}_x} \quad \forall\, x \in \mathbb{M}, \quad \forall\, \xi \in T_x\mathbb{M}, \tag{1.8}$$

then the functional $\mathcal{E}$ defined by (1.1) *is* (strongly) displacement λ-convex.

Remark 1.2. Besides the logarithmic entropy corresponding to $e(\rho) = \rho \log \rho$ (and $U(\rho) = \rho$), typical examples of functionals that satisfy properties (1.7) are

$$e(\rho) = \frac{1}{m-1}\rho^m, \quad U(\rho) = \rho^m, \quad m \geq 1 - \frac{1}{n}. \tag{1.9}$$

We recall that assumptions (1.7) imply the convexity of the function $\rho \mapsto e(\rho)$ (since the dimension n is greater than 1, they are in fact more restrictive).

Aim of the paper: an Eulerian approach to displacement convexity. In this paper we present an alternative proof of Theorem 1.1, which does not rely on the existence and smoothness of optimal transport maps and geodesics for the Wasserstein distance.

Our strategy can be described in three steps:

1. Following the approach suggested by Otto and Westdickenberg in [19], we work in the subspace $\mathcal{P}_2^{ar}(\mathbb{M})$ of measures with smooth and positive densities, and we use the "Riemannian" formula for the Wasserstein distance, originally introduced in the Euclidean framework by Benamou and Brenier [6]: if $\mu^i = \rho^i \,\mathrm{V} \in \mathcal{P}_2^{ar}(\mathbb{M})$, $i = 0, 1$, then [19, Proposition 4.3]

$$W_2^2\left(\mu^0, \mu^1\right) = \inf_{\mathcal{C}(\mu^0,\mu^1)} \left\{ \int_0^1 \int_M |\nabla \phi^s|^2 \rho^s \,\mathrm{dV}\,\mathrm{d}s \right\} \qquad \forall\, \mu^0, \mu^1 \in \mathcal{P}_2^{ar}(\mathbb{M}) \tag{1.10}$$

where

$$\begin{aligned} \mathcal{C}\left(\mu^0, \mu^1\right) = \Big\{ &(\rho, \phi) : \rho \in C^\infty([0,1] \times \mathbb{M}; \mathbb{R}_+), \quad \phi \in C^\infty([0,1] \times \mathbb{M}) \\ &\partial_s \rho^s + \nabla \cdot (\rho^s \nabla \phi^s) = 0 \ \text{ in } (0,1) \times \mathbb{M}, \quad \mu^i = \rho^i \,\mathrm{V} \Big\}. \end{aligned} \tag{1.11}$$

Even though the Wasserstein space can't be endowed with a smooth Riemannian structure, (1.11) still shows a "Riemannian" characterization of the Wasserstein distance on $\mathcal{P}_2^{ar}(\mathbb{M})$.

2. The second important fact, originally showed by the so-called "Otto calculus" in [17], is that the nonlinear diffusion equation

$$\partial_t \rho_t - \Delta_{\mathbf{g}}\, U(\rho_t) = 0 \quad \text{in } [0,+\infty)\times\mathbb{M}, \qquad \rho_{|_{t=0}} = \rho_0, \tag{1.12}$$

where $U : \mathbb{R}^+ \to \mathbb{R}$ is the function defined in (1.7) and $\Delta_{\mathbf{g}}$ is the Laplace–Beltrami operator on $\mathbb{M}$, is the gradient flow of the functional (1.1) in $\mathcal{P}_2(\mathbb{M})$. Indeed, (1.12) corresponds to the heat equation if U is the logarithmic entropy and to the porous medium equation if U is defined by (1.9).

Starting directly from (1.10) and owing to the fact that the flow generated by (1.12) preserves smooth and positive densities, when $\mathrm{Ric}_{\mathbf{g}}(\mathbb{M}) \geq 0$ we shall show that the measures $\mu_t = \rho_t \mathrm{V} \in \mathcal{P}_2^{ar}(\mathbb{M})$ associated to the solutions of (1.12) also solve the Evolution Variational Inequality (E.V.I.)

$$\frac{1}{2}\frac{\mathrm{d}^+}{\mathrm{d}t} W_2^2(\nu,\mu_t) \leq \mathcal{E}(\nu) - \mathcal{E}(\mu_t) \quad \forall\, t \geq 0,\ \nu \in \mathcal{P}_2^{ar}(\mathbb{M}), \tag{1.13}$$

which has been introduced in [4] as a purely metric characterization of the gradient flows of geodesically convex functionals in metric spaces (and, in particular, in $\mathcal{P}_2(\mathbb{R}^d)$); here

$$\frac{\mathrm{d}^+}{\mathrm{d}t}\zeta(t) = \limsup_{h\downarrow 0} \frac{\zeta(t+h)-\zeta(t)}{h} \tag{1.14}$$

for every real function $\zeta : [0,+\infty) \to \mathbb{R}$.

When $\mathrm{Ric}\,(\mathbb{M}) \geq \lambda$ (a shorthand for (1.8)), we also show that the solutions of the heat equation satisfy the modified inequality

$$\frac{1}{2}\frac{\mathrm{d}^+}{\mathrm{d}t} W_2^2(\nu,\mu_t) + \frac{\lambda}{2} W_2^2(\nu,\mu_t) \leq \mathcal{E}(\nu) - \mathcal{E}(\mu_t) \quad \forall\, t \geq 0,\ \nu \in \mathcal{P}_2^{ar}(\mathbb{M}), \tag{1.15}$$

where $\mathcal{E}$ is the relative entropy functional whose integrand function is $e(\rho) = \rho\log\rho$. Note that (1.15) reduces to (1.13) when $\lambda = 0$. In order to prove (1.13) and (1.15), we propose an "Eulerian" strategy which could be adapted to more general situations.

3. The third crucial fact is the following: whenever a functional $\mathcal{E}$ satisfies (1.13) (or, more generally, (1.15)) for a given semigroup $\mathcal{S}_t : \mu_0 = \rho_0\mathrm{V} \mapsto \mu_t = \rho_t\mathrm{V}$ in $\mathcal{P}_2^{ar}(\mathbb{M})$, $\mathcal{E}$ is displacement convex (resp., displacement λ-convex). Thus, the question of the behavior of $\mathcal{E}$ along geodesics can be reduced to a differential estimate of $\mathcal{E}$ along the smooth and positive solutions of its gradient flow.

Plan of the paper. In section 2 we present the main ideas of our approach in the simplified (finite-dimensional and smooth) setting of geodesically convex functions on Riemannian manifolds. We think that these ideas are sufficiently general to be useful in other circumstances, at least for distances which admits a Riemannian characterization as (1.10), see, e.g., [11, 7].

After a brief review of the definition of (gradient) λ-flows in arbitrary metric spaces (basically following the ideas of [4]), we present in section 3 our first result, showing that the existence of a flow satisfying the E.V.I. (1.15) (even on a dense subset of initial data, such as $\mathcal{P}_2^{ar}(\mathbb{M})$) entails the (strong) displacement λ-convexity of the functional $\mathcal{E}$.

Following the strategy explained in the second section, in the last two sections we prove the differential estimates showing that (1.12) satisfies (1.13) (in section 4) or, in the case of the Heat equation, (1.15) (in section 5).

2. Gradient flows and geodesic convexity in a smooth setting.

Contraction semigroups and action integrals. In order to explain the main point of our strategy, let us first consider the simple setting of a smooth function $F : \mathrm{X} \to \mathbb{R}$ on a complete Riemannian manifold X with metric $\langle \cdot, \cdot \rangle_g$, (squared) norm $|\xi|_g^2 = \langle \xi, \xi \rangle_g$, and the endowed Riemannian distance

$$d^2(u,v) := \min \left\{ \int_0^1 \left|\dot\gamma^s\right|_g^2 \,\mathrm{d}s, \quad \gamma : [0,1] \to \mathrm{X},\ \gamma^0 = v,\ \gamma^1 = u \right\}. \tag{2.1}$$

In a smooth setting, the geodesic λ-convexity of F can be expressed through the differential condition

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2} F(\gamma^s) \geq \lambda \left|\dot\gamma^s\right|_g^2 \tag{2.2}$$

along any geodesic curve γ minimizing (2.1). As we discussed in the introduction, the direct computation of (2.2) could be difficult in a nonsmooth, infinite dimensional setting; it is therefore important to find equivalent conditions which avoid twofold differentiation along geodesics. One possibility, suggested in [19], is to find equivalent conditions to geodesic λ-convexity in terms of the gradient flow generated by F.

Let us recall that the gradient flow of F is a continuous semigroup of (time-dependent) maps $\mathsf{S}_t : \mathrm{X} \to \mathrm{X}$, $t \in [0, +\infty)$, which at every initial datum u associate the curve $u_t := \mathsf{S}_t(u)$ solution of the differential equation

$$\dot u_t = -\nabla F(u_t) \quad \forall\, t \geq 0, \quad u_0 = u. \tag{2.3}$$

It is well known that, when F is geodesically λ-convex, S_t is λ-contracting; i.e.,

$$d^2(\mathsf{S}_t(u), \mathsf{S}_t(v)) \leq e^{-2\lambda t} d^2(u,v), \quad \forall\, t \geq 0, \qquad \forall\, u, v \in \mathrm{X}. \tag{2.4}$$

By the semigroup property, (2.4) is also equivalent to the differential inequality (see (1.14))

$$\frac{\mathrm{d}^+}{\mathrm{d}t} d^2(\mathsf{S}_t(u), \mathsf{S}_t(v))\Big|_{t=0} \leq -2\lambda\, d^2(u,v) \quad \forall\, u, v \in \mathrm{X}. \tag{2.5}$$

Otto and Westdickenberg [19] revert this argument and observe that it could be easier to directly prove (2.5) by a differential estimate involving only the action of the semigroup along smooth curves; as a byproduct, one should obtain the convexity of F. To this aim, they consider a smooth curve γ^s, $s \in [0,1]$, connecting v to u, and the action integral $\mathcal{A}_t$ associated with its smooth perturbation

$$\gamma_t^s := \mathsf{S}_t(\gamma^s), \quad A_t^s := \left|\partial_s \gamma_t^s\right|_g^2, \quad \mathcal{A}_t := \int_0^1 A_t^s \,\mathrm{d}s, \tag{2.6}$$

where $\partial_s \gamma$ denotes the tangent vector in $T_\gamma \mathrm{X}$ obtained by differentiating w.r.t. s. Since, by the very definition of d,

$$d^2(\mathsf{S}_t(v), \mathsf{S}_t(u)) \leq \mathcal{A}_t \tag{2.7}$$

and for every $\varepsilon > 0$ one can always find a curve γ^s so that $\mathcal{A}_0 \le d^2(u,v) + \varepsilon$ (in a smooth setting one can take $\varepsilon = 0$), (2.5) surely holds if one can prove that

$$\frac{\mathrm{d}^+}{\mathrm{d}t}\mathcal{A}_t\Big|_{t=0} \le -2\lambda\,\mathcal{A}_0, \quad \text{or its pointwise version} \quad \frac{\partial^+}{\partial t}\Big|_{t=0} A_t^s \le -2\lambda\, A_0^s. \tag{2.8}$$

Having obtained the contraction property from (2.8), it still remains open how to deduce that F is geodesically convex. Notice that along an arbitrary curve η^s

$$\frac{\partial}{\partial s}F(\eta^s) = \langle \nabla F(\eta^s), \partial_s \eta^s\rangle_g = -\langle \partial_r \mathsf{S}_r(\eta^s)_{|r=0}, \partial_s\eta^s\rangle_g; \tag{2.9}$$

applied to $\eta^s := \gamma_t^s$, (2.9) and the semigroup property $\mathsf{S}_r(\gamma_t^s) = \gamma_{t+r}^s$ yield

$$\frac{\partial}{\partial s}F(\gamma_t^s) = -\langle \partial_t\gamma_t^s, \partial_s\gamma_t^s\rangle_g. \tag{2.10}$$

In a smooth setting we can assume that γ^s is a minimal geodesic; operating a further differentiation with respect to s, we obtain

$$\begin{aligned}
\frac{\partial^2}{\partial s^2}F(\gamma^s) &\overset{(2.10)}{=} -\frac{\partial}{\partial s}\langle \partial_t\gamma_t^s, \partial_s\gamma_t^s\rangle_g\Big|_{t=0} = -\langle D_{\partial_s}\partial_t\gamma_t^s, \partial_s\gamma_t^s\rangle_g\Big|_{t=0} \\
&\quad - \langle \partial_t\gamma_t^s, D_{\partial_s}\partial_s\gamma_t^s\rangle_g\Big|_{t=0} \\
&= -\langle D_{\partial_s}\partial_t\gamma_t^s, \partial_s\gamma_t^s\rangle_g\Big|_{t=0} = -\langle D_{\partial_t}\partial_s\gamma_t^s, \partial_s\gamma_t^s\rangle_g\Big|_{t=0} \\
&= -\frac{1}{2}\frac{\partial}{\partial t}\langle \partial_s\gamma_t^s, \partial_s\gamma_t^s\rangle_g\Big|_{t=0} \\
&\overset{(2.6)}{=} -\frac{1}{2}\frac{\partial}{\partial t}\Big|_{t=0} A_t^s \overset{(2.8)}{\ge} \lambda\,\big|\partial_s\gamma^s\big|_g^2,
\end{aligned} \tag{2.11, 2.12}$$

where we used the standard properties of the covariant differentiations $D_{\partial_s}, D_{\partial_t}$ and, in (2.11), the fact that at $t = 0$ $D_{\partial_s}\partial_s\gamma_t^s = 0$, being $\gamma_t^s = \gamma^s$ a geodesic.

A metric derivation of convexity. Even if the previous differential argument shows that (2.8) implies geodesic λ-convexity, it still requires nice smooth properties on geodesics and covariant differentiation, which could be hard to extend to a nonsmooth setting.

This is not at all surprising, since the contraction property (2.5) and its action-differential characterization (2.8) do not carry all the information linking the semigroup S to F: in order to conclude the argument in (2.11), we had therefore to insert the information coming from (2.9).

To overcome these difficulties, we shall deal with a more precise metric characterization of S than (2.4). As it has been proposed and studied in [4], gradient flows of geodesically λ-convex functionals in "almost" Euclidean settings should satisfy a purely metric formulation in terms of the E.V.I.

$$\frac{1}{2}\frac{\mathrm{d}^+}{\mathrm{d}t}d^2(\mathsf{S}_t(u), v) + \frac{\lambda}{2}d^2(\mathsf{S}_t(u), v) + F(\mathsf{S}_t(u)) \le F(v), \quad \forall\, v \in \mathrm{X},\ t > 0. \tag{2.13}$$

It can be proved (see [5]) that (2.13) characterizes S and implies the contractivity property (2.4).

As we discussed before, here we invert the usual procedure (starting from a convex functional, construct its gradient flow) and we suppose that there exists a smooth flow

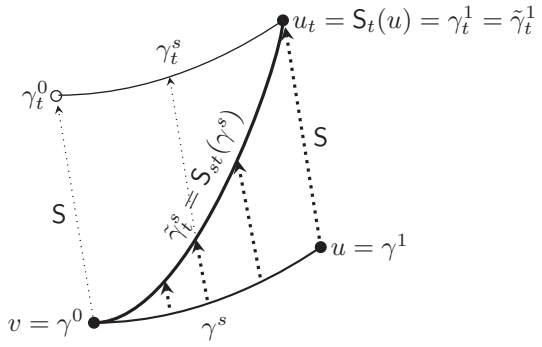


FIG. 1. *Variation of the curve γ^s under the action of the semigroup S.*

S_t satisfying (2.13). The following result, whose proof will be postponed (in a more general form) to Theorem 3.2 in the next section, shows that F is geodesically λ-convex.

THEOREM 2.1. *Suppose that there exists a continuous semigroup of maps $\mathsf{S}_t \in C^0(\mathrm{X};\mathrm{X})$, $t \ge 0$, satisfying* (2.13). *Then for every (minimal, constant speed) geodesic $\gamma : [0,1] \to \mathrm{X}$*

$$F(\gamma^s) \le (1-s)F\left(\gamma^0\right) + sF\left(\gamma^1\right) - \frac{\lambda}{2}s(1-s)d^2\left(\gamma^0,\gamma^1\right), \quad \forall\, s \in [0,1]; \tag{2.14}$$

i.e., F is (strongly) geodesically λ-convex.

E.V.I. through action-differential estimates. Thanks to Theorem 2.1, it is possible to prove the geodesic λ-convexity of F by exhibiting a flow S satisfying the E.V.I. (2.13). According to the general strategy suggested by [19], we want to reduce (2.13) to a suitable family of differential inequalities satisfied by the action A_t^s of (2.6).

The idea here is to consider a different family of perturbations of a given smooth curve $\gamma : [0,1] \to \mathrm{X}$, still induced by the semigroup S. In fact, differently from the contraction estimate (2.5) where we are flowing both the points u, v through S_t, in (2.13) we want to keep the point $v := \gamma^0$ fixed and to vary only $u := \gamma^1$. If γ^s is a smooth curve connecting them, it is then natural to consider the new families (see Figure 1)

$$\tilde\gamma_t^s := \mathsf{S}_{st}(\gamma^s) = \gamma_{st}^s, \quad \tilde F_t^s := F(\tilde\gamma_t^s) \qquad s \in [0,1],\ t \ge 0. \tag{2.15}$$

Notice that $\tilde\gamma_0^s = \gamma^s$, $\tilde\gamma_t^0 = \gamma^0 = v$, and $\tilde\gamma_t^1 = \mathsf{S}_t(\gamma^1) = \mathsf{S}_t(u)$. As before, we introduce the quantities

$$\tilde A_t^s := \left|\partial_s \tilde\gamma_t^s\right|_g^2, \quad \tilde{\mathcal{A}}_t := \int_0^1 \tilde A_t^s \,\mathrm{d}s. \tag{2.16}$$

THEOREM 2.2 (A differential inequality linking action and flow). *Suppose that for every smooth curve $\gamma : [0,1] \to \mathrm{X}$, the quantities $\tilde A_t^s, \tilde F_t^s$ induced by the flow S through* (2.15),(2.16) *satisfy*

$$\frac{1}{2}\frac{\partial}{\partial t}\tilde A_t^s + \frac{\partial}{\partial s}\tilde F_t^s \le -\lambda\, s\, \tilde A_t^s, \quad \forall\, t \ge 0. \tag{2.17}$$

Then S satisfies (2.13), *it is the gradient flow of F, and F is geodesically λ-convex. Moreover, it is sufficient to check* (2.17) *at $t = 0$.*

Proof. Let us first observe that (2.17) yields, after an integration with respect to s in $[0,1]$,

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\mathcal{A}}_t + \tilde{F}_t^1 - \tilde{F}_t^0 \le -\lambda \int_0^1 s\tilde{A}_t^s \,\mathrm{d}s. \tag{2.18}$$

By the semigroup property, it is sufficient to prove (2.13) at $t = 0$. We choose a geodesic γ^s connecting v to u, and we consider the curves given by (2.15). Since

$$d^2(\mathsf{S}_t(u), v) \le \int_0^1 \tilde{A}_t^s \,\mathrm{d}s = \tilde{\mathcal{A}}_t, \tag{2.19}$$

$$d^2(v,u) = \int_0^1 \tilde{A}_0^s \,\mathrm{d}s = \tilde{\mathcal{A}}_0, \quad \tilde{F}_t^1 = F(\mathsf{S}_t(u)), \quad \tilde{F}_t^0 = F(v),$$

by (2.18) at $t = 0$ we obtain

$$\frac{1}{2}\frac{\mathrm{d}^+}{\mathrm{d}t} d^2(\mathsf{S}_t(u), v)\Big|_{t=0} + F(u) - F(v) \le -\lambda \int_0^1 s\, \tilde{A}_0^s \,\mathrm{d}s = -\frac{\lambda}{2} d^2(u,v), \tag{2.20}$$

where in the last identity we used the fact that γ^s is a geodesic, and therefore $\tilde{A}_0^s = |\partial_s \gamma^s|_g^2$ is constant in $[0,1]$ and takes the value $d^2(\gamma^0, \gamma^1) = d^2(v,u)$.

Since $\tilde{\gamma}^s_{t_0+t} = \mathsf{S}_{st}\tilde{\gamma}^s_{t_0}$ by the semigroup property, if S satisfies (2.17) at the initial time $t = 0$ for an arbitrary smooth curve γ, then it also satisfies (2.17) for $t > 0$. ∎

Our last result provides a simple criterion to check (2.17):

THEOREM 2.3. *Suppose that $\mathsf{S} : [0,+\infty) \times \mathrm{X} \to \mathrm{X}$ is the "differential" gradient flow of F satisfying* (2.9) *for any smooth curve γ^s, let $\gamma_t^s, \tilde{\gamma}_t^s, A_t^s, \tilde{A}_t^s, \tilde{F}_t^s$ be defined as in* (2.6), (2.15), *and* (2.16), *and let us set*

$$\tilde{D}_r^s := \frac{1}{2}\lim_{h\downarrow 0} h^{-1}\Big(\big|\partial_s \gamma^s_{sr+h}\big|_g^2 - \big|\partial_s \gamma^s_{sr}\big|_g^2\Big). \tag{2.21}$$

Then

$$\frac{1}{2}\frac{\partial}{\partial t}\tilde{A}_t^s + \frac{\partial}{\partial s}\tilde{F}_t^s = s\tilde{D}_t^s. \tag{2.22}$$

Furthermore, if (2.8) *holds, then*

$$\tilde{D}_t^s \le -\lambda\, \tilde{A}_t^s, \tag{2.23}$$

and (2.17) *holds, too, so that F is geodesically λ-convex, and S is also its "metric" gradient flow, characterized by the E.V.I.* (2.13).

Proof. Let us set

$$\tilde{\gamma}^s_{t,\tau} := \mathsf{S}_\tau \tilde{\gamma}_t^s = \gamma^s_{st+\tau}, \quad \tilde{A}^s_{t,\tau} := \big|\partial_s \tilde{\gamma}^s_{t,\tau}\big|_g^2, \tag{2.24}$$

so that

$$\tilde{\gamma}^s_{t+h} = \tilde{\gamma}^s_{t,sh}, \quad \partial_s \tilde{\gamma}^s_{t+h} = \partial_s \tilde{\gamma}^s_{t,\tau} + h\partial_\tau \tilde{\gamma}^s_{t,\tau}\Big|_{\tau=sh}, \quad \tilde{D}_t^s = \frac{1}{2}\frac{\partial}{\partial\tau}\tilde{A}^s_{t,\tau}\Big|_{\tau=0}. \tag{2.25}$$

Observe that the identity

$$|x+y|_g^2 = 2\langle x+y, y\rangle_g + |x|_g^2 - |y|_g^2, \qquad \forall\, x, y \in T_\gamma X \tag{2.26}$$

yields

$$\begin{aligned}\tilde A^s_{t+h} &= \big|\partial_s\tilde\gamma^s_{t+h}\big|^2_g \overset{(2.25)}{=} \big|\partial_s\tilde\gamma^s_{t,\tau}+h\partial_\tau\tilde\gamma^s_{t,\tau}\big|^2_g\Big|_{\tau=sh}\\ &\overset{(2.26)}{=} \Big[2h\langle\partial_s\tilde\gamma^s_{t,\tau}+h\partial_\tau\tilde\gamma^s_{t,\tau},\partial_\tau\tilde\gamma^s_{t,\tau}\rangle+\big|\partial_s\tilde\gamma^s_{t,\tau}\big|^2_g-h^2\big|\partial_\tau\tilde\gamma^s_{t,\tau}\big|^2_g\Big]\Big|_{\tau=sh}\\ &= 2h\,\langle\partial_s\tilde\gamma^s_{t+h},\partial_\theta \mathsf{S}_\theta(\tilde\gamma^s_{t+h})\rangle\Big|_{\theta=0}+\tilde A^s_{t,sh}-o(h)\overset{(2.9)}{=}-2h\,\frac{\partial}{\partial s}F(\tilde\gamma^s_{t+h})+\tilde A^s_{t,sh}-o(h).\end{aligned}$$

We thus get

$$\frac{1}{2h}\big(\tilde A^s_{t+h}-\tilde A^s_t\big)+\frac{\partial}{\partial s}F(\tilde\gamma^s_{t+h})=\frac{1}{2h}\big(\tilde A^s_{t,sh}-\tilde A^s_t\big)-o(1), \tag{2.27}$$

so that, passing to the limit as $h\downarrow 0$ we get (2.22). □

Remark 2.4. Notice that the remainder term $o(1)$ in (2.27) is nonnegative, so it can be simply neglected, if one is just interested in the inequality (2.17).

3. Gradient flows and geodesic convexity in a metric setting. In this section we will briefly recall some basic definitions and properties of gradient flows in a metric setting, and we will prove Theorem 2.1 in a slightly more general framework.

Let (X,d) be a metric space (not necessarily complete) and let $F:\mathrm{X}\to(-\infty,+\infty]$ be a lower semicontinuous functional, whose proper domain $D(F):=\big\{w\in\mathrm{X}: F(w)<+\infty\big\}$ is dense in X (otherwise we can always restrict all the next statements to the closure of $D(F)$ in X). We also assume that F is bounded from below, i.e., $F_{\inf}:=\inf_{u\in\mathrm{X}}F(u)>-\infty$.

A C^0-semigroup S in $C^0(\mathrm{X};\mathrm{X})$ is a family S_t, $t\ge0$, of continuous maps in X such that

$$\mathsf{S}_{t+h}(u)=\mathsf{S}_h\big(\mathsf{S}_t(u)\big),\quad \lim_{t\downarrow0}\mathsf{S}_t(u)=\mathsf{S}_0(u)=u\quad\forall\,u\in\mathrm{X},\ t,h\ge0. \tag{3.1}$$

Given a real number $\lambda\in\mathbb{R}$, we say that S is the λ-(gradient) flow of F if it satisfies

$$\mathsf{S}_t(\mathrm{X})\subset D(F)\text{ for every }t>0; \tag{3.2a}$$

$$\text{the map } t\mapsto F(\mathsf{S}_t(u))\text{ is not increasing in }(0,+\infty); \tag{3.2b}$$

$$\frac12\frac{\mathrm{d}^+}{\mathrm{d}t}d^2(\mathsf{S}_t(u),v)+\frac{\lambda}{2}d^2(\mathsf{S}_t(u),v)+F(\mathsf{S}_t(u))\le F(v),\quad\forall\,u\in\mathrm{X},\,v\in D(F),\,t\ge0. \tag{3.2c}$$

Clearly, if S is a λ-flow for F, then it is also a λ'-flow for every $\lambda'\le\lambda$. The next proposition collects some useful properties of λ-flows.

Proposition 3.1 (Integral characterization of flows and contraction). *A C^0-semigroup S satisfies* (3.2a,b,c) *if and only if it satisfies the following integrated form*

$$\frac{e^{\lambda(t_1-t_0)}}{2}d^2(\mathsf{S}_{t_1}(u),v)-\frac12 d^2(\mathsf{S}_{t_0}(u),v)\le \mathsf{E}_\lambda(t_1-t_0)\,(F(v)-F(\mathsf{S}_{t_1}(u)))\quad\forall\,0\le t_0<t_1, \tag{3.3}$$

for every $u\in\mathrm{X}$, $v\in D(F)$, where

$$\mathsf{E}_\lambda(t):=\int_0^t e^{\lambda r}\,\mathrm{d}r=\begin{cases}\frac{e^{\lambda t}-1}{\lambda} & \text{if }\lambda\neq0,\\ t & \text{if }\lambda=0.\end{cases}$$

In particular, S *satisfies the uniform regularization bound*

$$F(\mathsf{S}_t(u)) \le F(v) + \frac{1}{2\,\mathsf{E}_\lambda(t)} d^2(u,v) \quad \forall\, u \in \mathrm{X},\ v \in D(F),\ t > 0, \tag{3.4}$$

the uniform continuity estimate

$$d^2(\mathsf{S}_{t_1}(u), \mathsf{S}_{t_0}(u)) \le 2\mathsf{E}_{-\lambda}(t_1 - t_0)\,(F(\mathsf{S}_{t_0}u) - F_{\inf}) \quad \forall\, u \in D(F),\ 0 \le t_0 \le t_1, \tag{3.5}$$

and the λ*-contraction property, i.e.,*

$$d(\mathsf{S}_t(u), \mathsf{S}_t(v)) \le e^{-\lambda t} d(u,v) \quad \forall\, u, v \in \mathrm{X},\ t \ge 0. \tag{3.6}$$

Proof. Clearly (3.3) yields (3.2a), being $D(F) \neq \emptyset$; (3.2b) and (3.5) follow by taking $v := \mathsf{S}_{t_0}(u)$, and (3.2c) can be proved by dividing both sides of (3.3) by $t_1 - t_0$ and passing to the limit as $t_1 \downarrow t_0$. In order to prove the converse implication, let us first observe that for a continuous real function $\zeta : [0, +\infty) \to \mathbb{R}$

$$\liminf_{h \downarrow 0} \frac{\zeta(t+h) - \zeta(t)}{h} \le 0 \quad \forall\, t > 0 \qquad \Longrightarrow \qquad \zeta \text{ is not increasing.} \tag{3.7}$$

In fact, if $0 \le t_0 < t_0 + \tau$ existed with $\delta := \tau^{-1}\big(\zeta(t_0+\tau) - \zeta(t_0)\big) > 0$, then a minimum point $\bar t \in [t_0, t_0 + \tau)$ of $t \mapsto \zeta(t) - \zeta(t_0) - \delta(t - t_0)$ would satisfy

$$\liminf_{h \downarrow 0} \frac{\zeta(\bar t + h) - \zeta(\bar t)}{h} - \delta \ge 0, \quad \text{which contradicts (3.7).}$$

(3.3) then follows by (3.2c), after a multiplication by $e^{\lambda t}$ and choosing

$$\zeta(t) := \frac{e^{\lambda t}}{2} d^2(\mathsf{S}_t(u), v) + \int_{\bar t}^{t} e^{\lambda r} \big(F(\mathsf{S}_r(u)) - F(v)\big)\,\mathrm{d}r, \quad \bar t > 0,$$

and recalling the monotonicity property (3.2b). A similar argument shows that

$$\frac{1}{2} d^2(\mathsf{S}_{t_1}(u), v) - \frac{1}{2} d^2(\mathsf{S}_{t_0}(u), v) + \frac{\lambda}{2} \int_{t_0}^{t_1} d^2(\mathsf{S}_r(u), v)\,\mathrm{d}r \le (t_1 - t_0)\,(F(v) - F(\mathsf{S}_{t_1}(u)))\,, \tag{3.8}$$

for every $0 \le t_0 < t_1$, $u \in \mathrm{X}$, and $v \in D(F)$. In order to prove the λ-contracting property, we apply (3.8) obtaining

$$\begin{aligned} &d^2(\mathsf{S}_h(u), \mathsf{S}_h(v)) - d^2(u,v) = d^2(\mathsf{S}_h(u), \mathsf{S}_h(v)) - d^2(\mathsf{S}_h(u), v) + d^2(\mathsf{S}_h(u), v) - d^2(u,v) \\ &\le -\lambda \int_0^h \big(d^2(\mathsf{S}_h(u), \mathsf{S}_r(v)) + d^2(\mathsf{S}_r(u), v)\big)\,\mathrm{d}r + 2h\,(F(v) - F(\mathsf{S}_h(v)))\,. \end{aligned}$$

We divide this inequality by h, and we pass to the limit as $h \downarrow 0$; the continuity of S_t, the lower semicontinuity of F, and the semigroup property of S yield

$$\frac{\mathrm{d}^+}{\mathrm{d}t} d^2(\mathsf{S}_t(u), \mathsf{S}_t(v)) \le -2\lambda\, d^2(u,v) \qquad \forall\, u, v \in \mathrm{X},\ t > 0, \tag{3.9}$$

which yields (3.6) thanks to (3.7). ∎

We can now prove the main result of this section: if a functional F admits a λ-flow, then F is geodesically λ-convex.

THEOREM 3.2 (Geodesic convexity via E.V.I.). *Let us suppose that* S *is a* λ*-flow for the functional* F*, according to* (3.2a,b,c)*, and let* $\gamma : [0,1] \to \mathrm{X}$ *be a Lipschitz curve satisfying*

$$d(\gamma^r,\gamma^s) \le L\,|r-s|, \quad L^2 \le d^2\left(\gamma^0,\gamma^1\right) + \varepsilon^2 \quad \forall\, r,s \in [0,1], \tag{3.10}$$

for some constant $\varepsilon \ge 0$*. Then for every* $t > 0$ *and* $s \in [0,1]$

$$F(\mathsf{S}_t(\gamma^s)) \le (1-s)F\left(\gamma^0\right) + sF\left(\gamma^1\right) - \frac{\lambda}{2}s(1-s)d^2\left(\gamma^0,\gamma^1\right) + \frac{\varepsilon^2}{2\mathsf{E}_\lambda(t)}s(1-s). \tag{3.11}$$

In particular, when γ *is a geodesic (i.e.,* γ *satisfies* (3.10) *with* $L = d(\gamma^0,\gamma^1)$*,* $\varepsilon = 0$*), we have*

$$F(\gamma^s) \le (1-s)F\left(\gamma^0\right) + sF\left(\gamma^1\right) - \frac{\lambda}{2}s(1-s)d^2\left(\gamma^0,\gamma^1\right); \tag{3.12}$$

i.e., F is (strongly) geodesically λ*-convex.*

Proof. Let γ be satisfying (3.10) and let us set $\gamma_t^s := \mathsf{S}_t(\gamma^s)$. Choosing $t_0 = 0$, $t_1 = t$, $u := \gamma^s$, and taking a convex combination of (3.3) written for $v := \gamma^0$, and $v := \gamma^1$, we get

$$\begin{aligned}\frac{e^{\lambda t}}{2}\left((1-s)\,d^2(\gamma_t^s,\gamma^0) + s\,d^2(\gamma_t^s,\gamma^1)\right) - \frac{1}{2}\left((1-s)\,d^2(\gamma^s,\gamma^0) + s\,d^2(\gamma^s,\gamma^1)\right)\\ \le \mathsf{E}_\lambda(t)\left((1-s)F(\gamma^0) + sF(\gamma^1) - F(\gamma_t^s)\right).\end{aligned} \tag{3.13}$$

We now observe that the elementary inequality

$$(1-s)a^2 + sb^2 \ge s(1-s)(a+b)^2 \quad \forall\, a,b \in \mathbb{R}, \quad s \in [0,1], \tag{3.14}$$

and the triangular inequality yield

$$\begin{aligned}(1-s)d^2\left(\gamma_t^s,\gamma^0\right) + sd^2\left(\gamma_t^s,\gamma^1\right) &\overset{(3.14)}{\ge} s(1-s)\left(d\left(\gamma_t^s,\gamma^0\right) + d\left(\gamma_t^s,\gamma^1\right)\right)^2\\ &\ge s(1-s)d^2\left(\gamma^0,\gamma^1\right).\end{aligned} \tag{3.15}$$

On the other hand, (3.10) yields

$$(1-s)\,d^2\left(\gamma^s,\gamma^0\right) + s\,d^2\left(\gamma^s,\gamma^1\right) \le L^2 s(1-s). \tag{3.16}$$

Inserting (3.16) and (3.15) in (3.13) we obtain

$$\frac{e^{\lambda t}-1}{2}s(1-s)d^2\left(\gamma^0,\gamma^1\right) - \frac{\varepsilon^2}{2}s(1-s) \le \mathsf{E}_\lambda(t)\left((1-s)F\left(\gamma^0\right) + sF\left(\gamma^1\right) - F(\gamma_t^s)\right). \tag{3.17}$$

Dividing then both sides of (3.17) by $\mathsf{E}_\lambda(t)$ we get (3.11); when $\varepsilon = 0$ we can pass to the limit as $t \downarrow 0$ obtaining (3.12). □

We conclude this section by considering the case when the flow S is defined only on a *dense* subset X_0 of $D(F)$ (which is dense in X). In order to prove the geodesic convexity of F in X by Theorem 3.2 we first have to extend S to the whole space X. This can be achieved by a density argument, if X is complete and the lower semicontinuous functional F satisfies the following approximation property:

$$\forall\, u \in D(F) \quad \exists\, u_n \in \mathrm{X}_0: \qquad \lim_{n\to\infty} d(u_n, u) = 0, \quad \lim_{n\to\infty} F(u_n) = F(u). \tag{3.18}$$

We state the precise extension result in the next theorem.

THEOREM 3.3. *Suppose that the functional F and the subset $\mathrm{X}_0 \subset D(F)$ satisfy* (3.18) *and let S be a λ-flow for F in X_0. If* X *is complete, S can be extended to a unique λ-flow $\bar{\mathsf{S}}$ in* X*, and therefore F is (strongly) geodesically λ-convex in* X.

Proof. Given $u \in \mathrm{X}$ and a sequence $u_n \in \mathrm{X}_0$ converging to u (X_0 is also dense in X), we can define

$$\bar{\mathsf{S}}_t(u) := \lim_{n\to\infty} \mathsf{S}_t(u_n) \quad \forall\, t > 0, \tag{3.19}$$

where it is clear that the limit in (3.19) exists (being X complete and S_t Lipschitz by (3.6)) and does not depend on the particular sequence u_n we used to approximate u. Moreover, $\bar{\mathsf{S}}_t$ is a semigroup and satisfies the estimate (3.5) and the λ-contracting property (3.6); being X_0 dense in X, it is not difficult to combine (3.5), (3.6), and (3.18) to show that $\lim_{t\downarrow 0} \mathsf{S}_t(u) = u$ for every $u \in \mathrm{X}$.

In order to prove that $\bar{\mathsf{S}}$ is still a λ-flow for F in X we have to check (3.3) in X: we fix $v \in D(F)$ and a sequence $v_n \in \mathrm{X}_0$ converging to v with $F(v_n) \to F(v)$, and we pass to the limit as $n \to \infty$ in the inequalities

$$\frac{e^{\lambda(t_1-t_0)}}{2} d^2(\mathsf{S}_{t_1}(u_n), v_n) - \frac{1}{2} d^2(\mathsf{S}_{t_0}(u_n), v_n) \le \mathsf{E}_\lambda(t_1 - t_0)\big(F(v_n) - F(\mathsf{S}_{t_1}(u_n)\big), \tag{3.20}$$

using the lower semicontinuity of F. ∎

4. Nonlinear diffusion equations as gradient flows of entropy functionals in $\mathcal{P}_2(\mathbb{M})$. We apply the strategy described in the section 2 to prove the geodesic convexity of the integral functional (1.1) in the case of a Riemannian manifold of nonnegative Ricci curvature. We therefore exhibit a smooth flow (induced by the nonlinear diffusion equation (1.12) on the dense subset $\mathcal{P}_2^{ar}(\mathbb{M})$) which satisfies the E.V.I. (1.13).

Before stating the main theorem of this section, let us recall a fundamental result on this kind of evolution equations, which can be found in [22, 19]:

THEOREM 4.1 (Classical solutions of nonlinear diffusion equations). *Let $e \in C^\infty(\mathbb{R}^+)$ and U be functions that satisfy the assumptions* (1.7) *of Theorem* 1.1. *For every $\rho_0 \in C^\infty(\mathbb{M})$ with $\rho_0 > 0$, there exists a unique smooth positive solution $\rho \in C^\infty([0,+\infty) \times \mathrm{X})$ to the Cauchy problem*

$$\partial_t \rho_t = \Delta_{\mathsf{g}} U(\rho_t), \qquad \rho_{|_{t=0}} = \lim_{t\downarrow 0} \rho_t = \rho_0. \tag{4.1}$$

Moreover, given a one-parameter family of positive initial data $s \mapsto \rho_0^s \in C^\infty([0,1] \times \mathbb{M})$, the corresponding solutions ρ_t^s of the equation (4.1) *depend smoothly on s, t.*

For every $\mu_0 = \rho_0 \mathrm{V} \in \mathcal{P}_2^{ar}(\mathbb{M})$ we denote by $\mathcal{S}_t(\mu_0) \in \mathcal{P}_2^{ar}(\mathbb{M})$ the measure $\mu_t = \rho_t \mathrm{V}$. The main result that we show in this section is the following.

THEOREM 4.2. *Let $e \in C^\infty(\mathbb{R}^+)$ and U be functions that satisfy the assumptions* (1.7) *of Theorem* 1.1 *and let us suppose that*

$$\mathrm{Ric}_{\mathsf{g}}(x) \ge 0 \quad \forall\, x \in \mathbb{M}. \tag{4.2}$$

The semigroup $\mathcal{S}$ induced by (4.1) *in $\mathcal{P}_2^{ar}(\mathbb{M})$ is a 0-flow in $\mathcal{P}_2^{ar}(\mathbb{M})$ for the functional*

$$\mathcal{E}(\mu) = \int_{\mathbb{M}} e(\rho)\, \mathrm{dV}, \quad \forall\, \mu = \rho \mathrm{V} \in \mathcal{P}_2^{ar}(\mathbb{M}). \tag{4.3}$$

In particular, for every $\mu_0 = \rho_0 \mathrm{V}, \nu \in \mathcal{P}_2^{ar}(\mathbb{M})$*, the measures* $\mu_t = \mathcal{S}_t(\mu_0) = \rho_t \mathrm{V} \in \mathcal{P}_2^{ar}(\mathbb{M})$ *solving* (4.1) *satisfy the E.V.I.*

$$\frac{1}{2}\frac{\mathrm{d}^+}{\mathrm{d}t} W_2^2(\nu, \mu_t) \le \mathcal{E}(\nu) - \mathcal{E}(\mu_t) \quad \forall\, t \in [0, +\infty). \tag{4.4}$$

In order to prove Theorem 4.2, thanks to the "Riemannian-like" characterization of the Wasserstein distance provided by (1.10), we can follow the strategy presented in section 2; in particular we want to prove the differential inequality of Theorem 2.2. Following Otto's formalism [17], we collect in the next table the formal correspondences between the various objects:

X, Riemannian manifold, with distance d	$\mathcal{P}_2^{ar}(\mathbb{M})$ with distance W_2
a smooth curve γ^s in X	a smooth family $\mu^s = \rho^s \mathrm{V} \in \mathcal{P}_2^{ar}(\mathbb{M})$
the tangent vector $\partial_s \gamma^s$ in $T_{\gamma^s}\mathrm{X}$	the vector field $\nabla\phi^s$ where $-\nabla\cdot(\rho^s \nabla\phi^s) = \frac{\partial}{\partial s}\rho^s$
$\lvert\partial_s \gamma^s\rvert_g^2$	$\int_{\mathbb{M}} \lvert\nabla\phi^s(x)\rvert_{\mathsf{g}}^2 \rho^s(x)\,\mathrm{dV}(x)$
$\gamma_t^s := \mathsf{S}_t(\gamma^s)$, $\tilde\gamma_t^s := \gamma_{st}^s = \mathsf{S}_{st}(\gamma^s)$	$\mu_t^s = \rho_t^s \mathrm{V} := \mathcal{S}_t(\mu^s)$, $\tilde\mu_t^s = \tilde\rho_t^s \mathrm{V} := \mu_{st}^s = \mathcal{S}_{st}(\mu^s)$
$\tilde A_t^s = \lvert\partial_s \tilde\gamma_t^s\rvert_g^2$	$\int_{\mathbb{M}} \lvert\nabla\tilde\phi_t^s(x)\rvert_{\mathsf{g}}^2 \tilde\rho_t^s(x)\,\mathrm{dV}(x)$
$F(\gamma^s)$	$\mathcal{E}(\mu^s) = \int_{\mathbb{M}} e(\rho^s)\,\mathrm{dV}$
$(\partial_\theta \mathsf{S}_\theta \gamma^s)_{\lvert\theta=0} = -\nabla F(\gamma^s)$	$-\nabla U(\rho^s)/\rho^s = -\nabla e'(\rho^s)$.

The core of the proof of Theorem 4.2 lies in the following lemma:

LEMMA 4.3. *Let* $\mu^s = \rho^s \mathrm{V}$, $s \in [0,1]$*, be a smooth family of measures in* $\mathcal{P}_2^{ar}(\mathbb{M})$ *and let* $\tilde\mu_t^s = \tilde\rho_t^s \mathrm{V} = \mathcal{S}_{st}(\mu^s)$ *be obtained by flowing* ρ^s *along the flow* (4.1)*; i.e.,* $\tilde\rho_t^s = \rho_{st}^s$ *where* ρ_t^s *satisfies*

$$\frac{\partial}{\partial t}\rho_t^s - \Delta_{\mathsf{g}}\, U(\rho_t^s) = 0 \ \text{in } \mathbb{M}, \quad \forall\, s \in [0,1],\ t > 0; \qquad \rho_{t=0}^s = \rho^s. \tag{4.5}$$

Let $\tilde\phi_t^s \in C^\infty([0,1] \times [0,+\infty) \times \mathbb{M})$ *be the functions defined by the equation*

$$-\nabla\cdot\big(\tilde\rho_t^s \nabla\tilde\phi_t^s\big) = \partial_s \tilde\rho_t^s \quad \text{in } \mathbb{M}, \quad \int_{\mathbb{M}} \tilde\phi_t^s(x)\,\mathrm{dV}(x) = 0 \quad \forall\, s \in [0,1],\ t \in [0,+\infty), \tag{4.6}$$

and let us set

$$\tilde A_t^s := \int_{\mathbb{M}} |\nabla\tilde\phi_t^s(x)|_{\mathsf{g}}^2\, \tilde\rho_t^s(x)\,\mathrm{dV}(x), \tag{4.7}$$

$$\tilde D_t^s := -\int_{\mathbb{M}} \left[\Big(|\mathrm{Hess}\,\tilde\phi_t^s|_{\mathsf{g}}^2 + \mathrm{Ric}_{\mathsf{g}}\Big(\nabla\tilde\phi_t^s, \nabla\tilde\phi_t^s\Big)\Big)\, U\left(\tilde\rho_t^s\right) + \left(\Delta_{\mathsf{g}}\,\tilde\phi_t^s\right)^2 \left(\tilde\rho_t^s U'(\tilde\rho_t^s) - U\left(\tilde\rho_t^s\right)\right) \right] \mathrm{dV}.$$

Then, we have the formula

$$\frac{\partial}{\partial t}\frac{1}{2}\tilde A_t^s + \frac{\partial}{\partial s}\mathcal{E}(\tilde\rho_t^s \mathrm{V}) = s\tilde D_t^s, \quad \forall\, t \in [0,+\infty),\ \forall\, s \in [0,1]. \tag{4.8}$$

In particular, if $\mathbb{M}$ *has nonnegative Ricci curvature, then* $\tilde D_t^s \le 0$ *and therefore*

$$\frac{\partial}{\partial t}\frac{1}{2}\tilde A_t^s + \frac{\partial}{\partial s}\mathcal{E}(\tilde\rho_t^s \mathrm{V}) \le 0. \tag{4.9}$$

Proof. Being $\tilde\rho_t^s := \rho_\tau^\sigma|_{\sigma=s,\tau=st}$ we get

$$\frac{\partial}{\partial s}\tilde\rho_t^s = \left(\frac{\partial}{\partial\sigma}\rho_\tau^\sigma + t\frac{\partial}{\partial\tau}\rho_\tau^\sigma\right)_{\sigma=s,\tau=st}, \qquad \frac{\partial}{\partial t}\tilde\rho_t^s = s\partial_\tau\rho_\tau^s|_{\tau=st} = s\Delta_{\mathsf g}\,U\,(\tilde\rho_t^s), \tag{4.10}$$

$$\frac{\partial^2}{\partial t\,\partial s}\tilde\rho_t^s \overset{(4.6)}{=} -\nabla\cdot\left(\frac{\partial}{\partial t}\tilde\rho_t^s\,\nabla\tilde\phi_t^s\right) - \nabla\cdot\left(\tilde\rho_t^s\,\frac{\partial}{\partial t}\nabla\tilde\phi_t^s\right), \tag{4.11}$$

$$\frac{\partial^2}{\partial s\,\partial t}\tilde\rho_t^s \overset{(4.10)}{=} s\Delta_{\mathsf g}\left(U'\,(\tilde\rho_t^s)\;\frac{\partial}{\partial s}\tilde\rho_t^s\right) + \Delta_{\mathsf g}\,U\,(\tilde\rho_t^s)$$

$$\overset{(4.6)}{=} -s\Delta_{\mathsf g}\left(U'\,(\tilde\rho_t^s)\;\nabla\cdot\left(\tilde\rho_t^s\nabla\tilde\phi_t^s\right)\right) + \Delta_{\mathsf g}\,U\,(\tilde\rho_t^s). \tag{4.12}$$

Differentiation and integration by parts yield

$$\begin{aligned}
&\frac{\partial}{\partial t}\int_{\mathbb M}\frac12|\nabla\tilde\phi_t^s|_{\mathsf g}^2\,\tilde\rho_t^s\,\mathrm{dV} = \int_{\mathbb M}\left\langle\frac{\partial}{\partial t}\nabla\tilde\phi_t^s,\nabla\tilde\phi_t^s\right\rangle_{\mathsf g}\tilde\rho_t^s\,\mathrm{dV} + \frac12\int_{\mathbb M}|\nabla\tilde\phi_t^s|_{\mathsf g}^2\,\frac{\partial}{\partial t}\tilde\rho_t^s\,\mathrm{dV}\\
&= -\int_{\mathbb M}\nabla\cdot\left(\tilde\rho_t^s\frac{\partial}{\partial t}\nabla\tilde\phi_t^s\right)\tilde\phi_t^s\,\mathrm{dV} \overset{(4.10)}{+} \frac12 s\int_{\mathbb M}\Delta_{\mathsf g}\left(|\nabla\tilde\phi_t^s|_{\mathsf g}^2\right)U(\tilde\rho_t^s)\,\mathrm{dV}\\
&\overset{(4.11)}{=}\int_{\mathbb M}\frac{\partial^2}{\partial t\partial s}\tilde\rho_t^s\,\tilde\phi_t^s\,\mathrm{dV} + \int_{\mathbb M}\left(\nabla\cdot\left(\frac{\partial}{\partial t}\tilde\rho_t^s\nabla\tilde\phi_t^s\right)\right)\tilde\phi_t^s\,\mathrm{dV} + \frac12 s\int_{\mathbb M}\Delta_{\mathsf g}\left(|\nabla\tilde\phi_t^s|_{\mathsf g}^2\right)U(\tilde\rho_t^s)\,\mathrm{dV}
\end{aligned} \tag{4.13}$$

$$\begin{aligned}
&\overset{(4.12)}{=}\int_{\mathbb M}\left(\Delta_{\mathsf g}\,U\,(\tilde\rho_t^s) - s\Delta_{\mathsf g}\left(U'(\tilde\rho_t^s)\,\nabla\cdot\left(\tilde\rho_t^s\nabla\tilde\phi_t^s\right)\right)\right)\tilde\phi_t^s\,\mathrm{dV}\\
&\quad - s\int_{\mathbb M}\Delta_{\mathsf g}U(\tilde\rho_t^s)\,|\nabla\tilde\phi_t^s|_{\mathsf g}^2\,\mathrm{dV} + \frac{s}{2}\int_{\mathbb M}\Delta_{\mathsf g}\left(|\nabla\tilde\phi_t^s|_{\mathsf g}^2\right)U(\tilde\rho_t^s)\,\mathrm{dV}\\
&= \int_{\mathbb M}U(\tilde\rho_t^s)\,\Delta_{\mathsf g}\,\tilde\phi_t^s\,\mathrm{dV} - s\int_{\mathbb M}\Big(\left\langle\nabla U(\tilde\rho_t^s),\nabla\tilde\phi_t^s\right\rangle_{\mathsf g}\Delta_{\mathsf g}\,\tilde\phi_t^s + \tilde\rho_t^s\,U'\,(\tilde\rho_t^s)\left(\Delta_{\mathsf g}\,\tilde\phi_t^s\right)^2\Big)\,\mathrm{dV}\\
&\quad - \frac{s}{2}\int_{\mathbb M}\Delta_{\mathsf g}\left(|\nabla\tilde\phi_t^s|_{\mathsf g}^2\right)U(\tilde\rho_t^s)\,\mathrm{dV}\\
&= -\int_{\mathbb M}\left\langle\nabla U\,(\tilde\rho_t^s),\nabla\tilde\phi_t^s\right\rangle_{\mathsf g}\mathrm{dV} + s\int_{\mathbb M}\left[-\frac12\Delta_{\mathsf g}\left(|\nabla\tilde\phi_t^s|_{\mathsf g}^2\right) + \left\langle\nabla\tilde\phi_t^s,\nabla\Delta_{\mathsf g}\,\tilde\phi_t^s\right\rangle_{\mathsf g}\right]U\,(\tilde\rho_t^s)\,\mathrm{dV}
\end{aligned} \tag{4.14}$$

$$+\;s\int_{\mathbb M}\left(\Delta_{\mathsf g}\,\tilde\phi_t^s\right)^2\left(U\,(\tilde\rho_t^s) - \tilde\rho_t^s U'(\tilde\rho_t^s)\right)\mathrm{dV}$$

Applying Bochner formula:

$$\langle\nabla\phi,\nabla\Delta_{\mathsf g}\,\phi\rangle_{\mathsf g} - \frac12\Delta_{\mathsf g}\left(|\nabla\phi|_{\mathsf g}^2\right) = -|\mathrm{Hess}\,\phi|_{\mathsf g}^2 - \mathrm{Ric}_{\mathsf g}(\nabla\phi,\nabla\phi), \tag{4.15}$$

we get

$$\frac{\partial}{\partial t}\frac12\int_{\mathbb M}|\nabla\tilde\phi_t^s|_{\mathsf g}^2\,\tilde\rho_t^s\,\mathrm{dV} + \int_{\mathbb M}\left\langle\nabla U(\tilde\rho_t^s),\nabla\tilde\phi_t^s\right\rangle_{\mathsf g}\mathrm{dV} = s\tilde D_t^s. \tag{4.16}$$

Now we observe that the second term in the right-hand side of (4.16) is the derivative of the functional (4.3) along the curve $s\mapsto\tilde\rho_t^sV\in\mathcal P_2^{ar}(\mathbb M)$:

$$\frac{\partial}{\partial s}\mathcal E(\tilde\mu_t^s) = \int_{\mathbb M}e'(\tilde\rho_t^s)\,\frac{\partial}{\partial s}\tilde\rho_t^s\,\mathrm{dV} = -\int_{\mathbb M}e'(\tilde\rho_t^s)\nabla\cdot\left(\tilde\rho_t^s\nabla\tilde\phi_t^s\right)\mathrm{dV} = \int_{\mathbb M}\nabla U\,(\tilde\rho_t^s)\cdot\nabla\tilde\phi_t^s\,\mathrm{dV}, \tag{4.17}$$

and we eventually obtain (4.8).

Finally, when $\mathrm{Ric}_{\mathsf{g}}(\mathbb{M}) \geq 0$, using the inequality $(\Delta_{\mathsf{g}}\,\phi)^2 \leq n|\mathrm{Hess}\,\phi|^2_{\mathsf{g}}$ and (1.7) we easily get $\tilde{D}^s_t \leq 0$ and (4.9). ∎

Proof of Theorem 4.2. We argue as in the proof of Theorem 2.2: we fix $\varepsilon > 0$ and we choose a smooth curve $(\rho, \phi) \in \mathcal{C}(\nu, \mu)$ such that

$$\int_0^1 \tilde{A}^s_0 \,\mathrm{d}s = \int_0^1 \int_{\mathbb{M}} |\nabla \phi^s|^2_{\mathsf{g}}\, \rho^s \,\mathrm{dV}\mathrm{d}s \leq W_2^2(\nu,\mu) + \varepsilon. \tag{4.18}$$

Let $(\tilde{\rho}, \tilde{\phi})$ be a smooth variation defined as in Lemma 4.3; since $\tilde{\rho}^0_t \mathrm{V} = \rho^0 \mathrm{V} = \nu$ and $\tilde{\rho}^1_t \mathrm{V} = \mu_t$, for every $t > 0$ we have $(\tilde{\rho}^s_t, \tilde{\phi}^s_t) \in \mathcal{C}(\nu, \mu_t)$, and therefore

$$W_2^2(\nu,\mu_t) \leq \int_0^1 \int_{\mathbb{M}} |\nabla \tilde{\phi}^s_t|^2_{\mathsf{g}}\, \tilde{\rho}^s_t \,\mathrm{dV}\,\mathrm{d}s = \int_0^1 \tilde{A}^s_t \,\mathrm{d}s. \tag{4.19}$$

Integrating (4.9) for $s \in [0,1]$ and $t \in [0,\tau]$ and recalling that $t \mapsto \mathcal{E}(\mu_t)$ is not increasing, we get

$$\frac{1}{2}\int_0^1 \tilde{A}^s_\tau \,\mathrm{d}s - \frac{1}{2}\int_0^1 \tilde{A}^s_0 \,\mathrm{d}s \leq \tau\left(\mathcal{E}(\nu) - \mathcal{E}(\mu_\tau)\right). \tag{4.20}$$

Combining (4.20) with (4.19) and (4.18) we get

$$\frac{1}{2}W_2^2(\nu,\mu_\tau) - \frac{1}{2}W_2^2(\nu,\mu) \leq \tau\left(\mathcal{E}(\nu) - \mathcal{E}(\mu_\tau)\right) + \varepsilon, \tag{4.21}$$

and, as ε is arbitrary,

$$\frac{1}{2}W_2^2(\nu,\mu_\tau) - \frac{1}{2}W_2^2(\nu,\mu) \leq \tau\left(\mathcal{E}(\nu) - \mathcal{E}(\mu_\tau)\right). \tag{4.22}$$

Since the semigroup associated with (4.1) is translation invariant, (4.22) is the integral formulation (3.3) of (4.4). ∎

Remark 4.4. Taking into account Theorem 2.3, (4.8) perfectly fits with the calculation performed by [19, Lemma 4.4], which provides the same expression for $\tilde{D}^s_t$. Applying now Theorem 3.3, with the choices $\mathrm{X} := \mathcal{P}_2(\mathbb{M})$, $\mathrm{X}_0 := \mathcal{P}_2^{ar}(\mathbb{M})$, and $F := \mathcal{E}$ (which satisfies the approximation condition (3.18); see [3]), we can prove the first part of Theorem 1.1.

Corollary 4.5. *Let $\mathcal{E} : \mathcal{P}_2(\mathbb{M}) \to (-\infty, +\infty]$ be the functional defined in* (1.1). *If $\mathcal{E}$ satisfies McCann conditions* (1.7) *and $Ric_{\mathsf{g}}(\mathbb{M}) \geq 0$, then $\mathcal{E}$ is (strongly) displacement convex along* every *geodesic $\mu : s \in [0,1] \mapsto \mu^s \in \mathcal{P}_2(\mathbb{M})$, i.e.,*

$$\mathcal{E}(\mu^s) \leq (1-s)\mathcal{E}(\mu^0) + s\mathcal{E}(\mu^1) \quad \forall\, s \in [0,1]. \tag{4.23}$$

5. The heat equation and the displacement λ-convexity of the logarithmic entropy. In this last section we prove the second part of Theorem 1.1: we thus assume that the Riemannian manifold $\mathbb{M}$ satisfies the lower Ricci curvature bound

$$\mathrm{Ric}_{\mathsf{g}}(\mathbb{M}) \geq \lambda \quad \text{i.e.,} \quad \mathrm{Ric}_{\mathsf{g}_x}(\xi,\xi) \geq \lambda\,|\xi|^2_{\mathsf{g}} \quad \forall\, \xi \in T_x\,\mathbb{M}, \tag{5.1}$$

and we consider the logarithmic entropy functional

$$\mathcal{E}(\mu) = \int_{\mathbb{M}} \rho \log \rho \,\mathrm{dV}, \quad \rho = \frac{d\mu}{\mathrm{dV}}, \tag{5.2}$$

corresponding to $e(\rho) := \rho\log\rho$. Since $U(\rho) = \rho$, the Wasserstein gradient flow associated to $\mathcal{E}$ is the Heat equation

$$\frac{\partial}{\partial t}\rho_t - \Delta_{\mathbf{g}}\,\rho_t = 0 \quad \text{in } \mathbb{M}, \qquad \rho_{|_{t=0}} = \rho_0. \tag{5.3}$$

The main result of this section is the following.

THEOREM 5.1. *The semigroup $\mathcal{S}_t : \mu_0 = \rho_0\mathrm{V} \mapsto \mu_t = \rho_t\mathrm{V}$, generated by the solution of the Heat equation* (5.3)*, is a λ-flow in $\mathcal{P}_2^{ar}(\mathbb{M})$ for the logarithmic entropy functional; i.e., μ_t satisfies the inequality*

$$\frac{1}{2}\frac{\mathrm{d}^+}{\mathrm{d}t}W_2^2(\nu,\mu_t) + \frac{\lambda}{2}W_2^2(\nu,\mu_t) \le \mathcal{E}(\nu) - \mathcal{E}(\mu_t) \quad \forall\, t\in[0,+\infty),\ \nu\in\mathcal{P}_2^{ar}(\mathbb{M}). \tag{5.4}$$

In particular, the logarithmic entropy functional (5.2) *is (strongly) displacement λ-convex; i.e., for every geodesic $\mu^s : [0,1] \to \mathcal{P}_2(\mathbb{M})$ between μ^0 and μ^1, we have*

$$\mathcal{E}(\mu^s) \le (1-s)\mathcal{E}(\mu^0) + s\mathcal{E}(\mu^1) - \frac{\lambda}{2}s(1-s)W_2^2(\mu^0,\mu^1), \quad \forall\, s\in[0,1]. \tag{5.5}$$

Proof. By Theorem 3.3, if $\mathcal{S}$ is a λ-flow for the functional (5.2) in $\mathcal{P}_2^{ar}(\mathbb{M})$, then $\mathcal{E}$ is (strongly) displacement λ-convex. In order to prove that $\mathcal{S}$ is a λ-flow, since (3.2a,b) are immediate, we check that $\mathcal{S}$ satisfies the E.V.I. (3.2c), and we argue as in the proof of Theorem 4.2 and Theorem 2.2. We thus fix $\varepsilon > 0$, and we choose a smooth curve $(\rho,\phi) \in \mathcal{C}(\nu,\mu)$

$$\int_0^1 \tilde{A}_0^s\,\mathrm{d}s = \int_0^1\int_{\mathbb{M}} |\nabla\phi^s|_{\mathbf{g}}^2\,\rho^s\,\mathrm{dV}\mathrm{d}s \le W_2^2(\nu,\mu) + \varepsilon^2. \tag{5.6}$$

By a standard reparametrization technique (see next Lemma 5.1), we can also assume that

$$W_2(\mu^{s_0},\mu^{s_1}) \le L|s_0 - s_1|, \quad L^2 := W_2^2(\nu,\mu) + \varepsilon^2 \qquad \forall\, s_0,s_1\in[0,1]; \qquad \mu^s := \rho^s\,\mathrm{V}. \tag{5.7}$$

We keep the same notation of Theorem 4.2 and Lemma 4.3; i.e.,

$$\tilde{\mu}_t^s = \tilde{\rho}_t^s\,\mathrm{V} := \mathcal{S}_{st}(\mu^s), \quad \tilde{A}_t^s := \int_{\mathbb{M}} |\nabla\tilde{\phi}_t^s|_{\mathbf{g}}^2\,\tilde{\rho}_t^s\,\mathrm{dV}, \quad \tilde{F}_t^s = \mathcal{E}(\tilde{\mu}_t^s), \tag{5.8}$$

where $\tilde{\phi}_t^s$ is a family of potentials associated with $\tilde{\rho}_t^s$ as in (4.6). Since $U(\rho) = \rho$, the term $\rho U'(\rho) - U(\rho)$ in the definition of $\tilde{D}_t^s$ vanishes, so that in the present case

$$\tilde{D}_t^s = -\int_{\mathbb{M}} \left(|\mathrm{Hess}\,\tilde{\phi}_t^s|_{\mathbf{g}}^2 + \mathrm{Ric}_{\mathbf{g}}\left(\nabla\tilde{\phi}_t^s,\nabla\tilde{\phi}_t^s\right)\right)\tilde{\rho}_t^s\,\mathrm{dV} \overset{(5.1)}{\le} -\lambda\int_{\mathbb{M}} |\nabla\tilde{\phi}_t^s|_{\mathbf{g}}^2\tilde{\rho}_t^s\,\mathrm{dV} = -\lambda\tilde{A}_t^s, \tag{5.9}$$

and (4.8) yields the differential inequality

$$\frac{1}{2}\frac{\partial}{\partial t}\tilde{A}_t^s + \lambda s\tilde{A}_t^s + \frac{\partial}{\partial s}\tilde{F}_t^s \le 0 \qquad \forall\, s\in[0,1], \quad \forall\, t>0. \tag{5.10}$$

Multiplying inequality (5.10) by $e^{2\lambda st} > 0$ we obtain

$$\frac{1}{2}\frac{\partial}{\partial t}\left(e^{2\lambda st}\tilde{A}_t^s\right) + \frac{\partial}{\partial s}\left(e^{2\lambda st}\tilde{F}_t^s\right) \le 2\lambda t\,e^{2\lambda st}\,\tilde{F}_t^s. \tag{5.11}$$

Integrating with respect to s from 0 to 1 we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac12\int_0^1 e^{2\lambda st}\tilde A_t^s\,\mathrm{d}s\right)+e^{2\lambda t}\tilde F_t^1-\tilde F_t^0\le\int_0^1 2\lambda\,t\,e^{2\lambda st}\tilde F_t^s\,\mathrm{d}s, \tag{5.12}$$

and a further integration with respect to t yields

$$\frac12\int_0^1 e^{2\lambda st}\tilde A_t^s\,\mathrm{d}s-\frac12\int_0^1 A_0^s\,\mathrm{d}s+\mathsf{E}_{2\lambda}(t)\mathcal{E}(\mu_t)-t\mathcal{E}(\nu)\le\int_0^t\int_0^1 2\lambda\,r\,e^{2\lambda sr}\,\tilde F_r^s\,\mathrm{d}s\,\mathrm{d}r. \tag{5.13}$$

Applying the next Lemma 5.1, since for $\lambda\neq0$ $\int_0^1\frac{1}{e^{2\lambda st}}\,\mathrm{d}s=\frac{1-e^{-2\lambda t}}{2\lambda t}=\frac{1}{e^{\lambda t}\mathsf{s}(\lambda t)}$, $\mathsf{s}(t):=\frac{t}{\sinh(t)}$, we get

$$\begin{aligned}&\frac{e^{\lambda t}\mathsf{s}(\lambda t)}{2}W_2^2(\mu_t,\nu)-\frac12W_2^2(\mu,\nu)+\mathsf{E}_{2\lambda}(t)\mathcal{E}(\mu_t)-t\mathcal{E}(\nu)\\&\qquad\le\int_0^t\int_0^1 2\lambda r e^{2\lambda sr}\tilde F_r^s\,\mathrm{d}s\mathrm{d}r+\varepsilon^2/2.\end{aligned} \tag{5.14}$$

Let us first consider the case $\lambda\le0$: being $\mathcal{E}$ nonnegative, the right-hand side in (5.14) is less or equal than ε; since $\varepsilon>0$ is arbitrary, we obtain the same inequality with 0 in the right-hand side. Since $t^{-1}\mathsf{E}_{2\lambda}(t)\to1$ as $t\downarrow0$ and $\mathsf{s}(0)=1$, we thus obtain

$$\frac12\frac{\mathrm{d}^+}{\mathrm{d}t}\Big(e^{\lambda t}\mathsf{s}(\lambda t)W_2^2(\mu_t,\nu)\Big)\Big|_{t=0}+\mathcal{E}(\mu)\le\mathcal{E}(\nu). \tag{5.15}$$

Being $\mathsf{s}'(0)=0$ it is then easy to check that

$$\frac{\mathrm{d}^+}{\mathrm{d}t}\Big(e^{\lambda t}\mathsf{s}(\lambda t)W_2^2(\mu_t,\nu)\Big)\Big|_{t=0}=\frac{\mathrm{d}^+}{\mathrm{d}t}\Big(W_2^2(\mu_t,\nu)\Big)\Big|_{t=0}+\lambda\,W_2^2(\mu,\nu),$$

which yields (5.4).

Let us now consider the case $\lambda>0$. Notice that we already know that $\mathcal{S}$ is a 0-flow; by (5.7) we can apply the estimate (3.11) with $\lambda=0$ obtaining

$$\begin{aligned}r\tilde F_r^s=r\mathcal{E}(\mathcal{S}_{rs}(\mu^s))&\overset{(3.11)}{\le}r\left((1-s)\mathcal{E}\left(\mu^0\right)+s\mathcal{E}\left(\mu^1\right)+\frac{\varepsilon^2}{2\,rs}s(1-s)\right)\\&\le r\left(\mathcal{E}\left(\mu^0\right)+\mathcal{E}\left(\mu^1\right)\right)+\varepsilon^2/2,\end{aligned}$$

since $s\in[0,1]$. We thus get

$$\int_0^t\int_0^1 2\lambda\,r\,e^{2\lambda sr}\tilde F_r^s\,\mathrm{d}s\,\mathrm{d}r\le\lambda te^{2\lambda t}\Big(t\big(\mathcal{E}(\mu_0)+\mathcal{E}(\mu_1)\big)+\varepsilon^2\Big); \tag{5.16}$$

inserting this bound in (5.14) and passing to the limit as $\varepsilon\downarrow0$ we find

$$\frac{e^{\lambda t}\mathsf{s}(\lambda t)}{2}W_2^2(\mu_t,\nu)-\frac12W_2^2(\mu,\nu)+\mathsf{E}_{2\lambda}(t)\mathcal{E}(\mu_t)-t\mathcal{E}(\nu)\le\lambda t^2e^{2\lambda t}\left(\mathcal{E}(\mu_0)+\mathcal{E}(\mu_1)\right). \tag{5.17}$$

Dividing by t and letting t tend to 0, the second term vanishes, so we obtain the E.V.I. also in the case in which $\lambda>0$. ∎

LEMMA 5.1. *Let $\nu,\mu\in\mathcal{P}_2^{ar}(\mathbb{M})$ and let $(\rho,\phi)\in\mathcal{C}(\nu,\mu)$ be a smooth solution of the continuity equation*

$$\frac{\partial}{\partial s}\rho^s+\nabla\cdot(\rho^s\,\nabla\phi^s)=0\quad \textit{in } [0,1]\times\mathbb{M}\quad \textit{with}$$
$$\rho^0\mathrm{V}=\nu,\ \rho^1\mathrm{V}=\mu,\quad \textit{and}\quad A^s:=\int_{\mathbb{M}}|\nabla\phi^s|^2_{\mathsf{g}}\,\rho^s\,\mathrm{dV}.$$

For every positive function $f\in C^\infty[0,1]$

$$W_2^2(\nu,\mu)\le L_f\int_0^1 f(s)A^s\,ds,\quad \textit{where}\quad L_f:=\int_0^1\frac{1}{f(s)}\,ds. \tag{5.18}$$

Moreover, for every $\varepsilon>0$ there exists a smooth rescaling $\mathsf{s}_\varepsilon:[0,1]\to[0,1]$ so that the reparametrized families

$$\bar\rho^r:=\rho^{\mathsf{s}_\varepsilon(r)},\quad \bar\phi^r:=\mathsf{s}_\varepsilon'(r)\phi^{\mathsf{s}_\varepsilon(r)},\quad \bar\mu^r:=\bar\rho^r\,\mathrm{V} \tag{5.19}$$

satisfy

$$\big(\bar\rho,\bar\phi\big)\in\mathcal{C}(\nu,\mu),\quad W_2(\bar\mu^{r_0},\bar\mu^{r_1})\le L|r_0-r_1|,\quad L^2\le\int_0^1 A^s\,\mathrm{d}s+\varepsilon^2. \tag{5.20}$$

Proof. Let us consider the smooth increasing map $\mathsf{r}:[0,1]\to[0,1]$

$$\mathsf{r}(s):=L_f^{-1}\int_0^s\frac{1}{f(s)}\,ds\quad\text{and its inverse } \mathsf{s}:=\mathsf{r}^{-1}\quad\text{with}\quad \mathsf{s}'(\mathsf{r}(s))=L_f f(s).$$

It is immediate to check that the smooth (reparametrized) curve

$$\bar\rho^r(x):=\rho^{\mathsf{s}(r)}(x),\quad \bar\phi^r(x):=\mathsf{s}'(r)\phi^{\mathsf{s}(r)}(x) \tag{5.21}$$

belongs to $\mathcal{C}(\nu,\mu)$. It follows that

$$W_2^2(\nu,\mu)\le\int_0^1\bar A^r\,\mathrm{d}r,\quad\text{where}\quad \bar A^r:=\int_{\mathbb{M}}|\nabla\bar\phi^r|^2_{\mathsf{g}}\,\bar\rho^r\,\mathrm{dV}\overset{(5.21)}{=}\big(\mathsf{s}'(r)\big)^2A^{\mathsf{s}(r)},$$

so that

$$\int_0^1\bar A^r\mathrm{d}r=\int_0^1 A^{\mathsf{s}(r)}\big(\mathsf{s}'(r)\big)^2\,\mathrm{d}r=\int_0^1 A^s\mathsf{s}'(\mathsf{r}(s))\,\mathrm{d}s=L_f\int_0^1 f(s)A^s\,\mathrm{d}s.$$

Choosing now the reparametrization s_ε corresponding to the choice

$$f_\varepsilon(s):=\frac{1}{\sqrt{\varepsilon^2+A^s}},\quad L_{f_\varepsilon}:=\int_0^1\sqrt{\varepsilon^2+A^s}\,\mathrm{d}s,\quad L_{f_\varepsilon}^2\le\varepsilon^2+\int_0^1 A^s\,\mathrm{d}s, \tag{5.22}$$

we get

$$W^2(\bar\mu^{r_0},\bar\mu^{r_1})\le|r_1-r_0|\textstyle\int_{r_0}^{r_1}\bar A^r\,\mathrm{d}r=|r_1-r_0|L_{f_\varepsilon}^2\int_{r_0}^{r_1}A^{\mathsf{s}(r)}f_\varepsilon^2(\mathsf{s}(r))\,\mathrm{d}r$$
$$\le(r_1-r_0)^2L_{f_\varepsilon}^2,$$

which yields (5.20). ∎

REFERENCES

[1] M. Agueh, *Existence of solutions to degenerate parabolic equations via the Monge-Kantorovich theory*, Adv. Differential Equations, 10 (2005), pp. 309–360.

[2] M. Agueh, N. Ghoussoub, and X. Kang, *Geometric inequalities via a general comparison principle for interacting gases*, Geom. Funct. Anal., 14 (2004), pp. 215–244.

[3] L. Ambrosio and G. Buttazzo, *Weak lower semicontinuous envelope of functionals defined on a space of measures*, Ann. Mat. Pura Appl., 150 (1988), pp. 311–339.

[4] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient flows in metric spaces and in the space of probability measures*, Lectures Math. ETH Zürich, Birkhäuser Verlag, Basel, 2005.

[5] L. Ambrosio and G. Savaré, *Gradient flows of probability measures*, in Handbook of Evolution Equations (III), Elsevier, New York, 2006.

[6] J.-D. Benamou and Y. Brenier, *A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem*, Numer. Math., 84 (2000), pp. 375–393.

[7] J. A. Carrillo, S. Lisini, and G. Savaré, *The porous medium flow and generalized displacement convexity*, Technical report, in preparation (2008).

[8] J. A. Carrillo, R. J. McCann, and C. Villani, *Contractions in the 2-Wasserstein length space and thermalization of granular media*, Arch. Ration. Mech. Anal., 179 (2006), pp. 217–263.

[9] D. Cordero-Erausquin, R. J. McCann, and M. Schmuckenschläger, *A Riemannian interpolation inequality à la Borell, Brascamp and Lieb*, Invent. Math., 146 (2001), pp. 219–257.

[10] D. Cordero-Erausquin, R. J. McCann, and M. Schmuckenschläger, *Prékopa-Leindler type inequalities on Riemannian manifolds, Jacobi fields, and optimal transport*, Ann. Fac. Sci. Toulouse Math. (6), 15 (2006), pp. 613–635.

[11] J. Dolbeault, B. Nazaret, and G. Savaré, *A new class of "dynamic" transport distances between measures*, Calc. Var. Partial Differential Equations, to appear.

[12] A. Figalli and C. Villani, *Strong displacement convexity on Riemannian manifolds*, Math. Z., 257 (2007), pp. 251–259.

[13] S. Lisini, *Characterization of absolutely continuous curves in Wasserstein spaces*, Calc. Var. Partial Differential Equations, 28 (2007), pp. 85–120.

[14] J. Lott and C. Villani, *Ricci curvature for metric-measure spaces via optimal transport*, Ann. of Math. (2), to appear.

[15] R. J. McCann, *A convexity principle for interacting gases*, Adv. Math., 128 (1997), pp. 153–179.

[16] R. J. McCann, *Polar factorization of maps on Riemannian manifolds*, Geom. Funct. Anal., 11 (2001), pp. 589–608.

[17] F. Otto, *The geometry of dissipative evolution equations: The porous medium equation*, Comm. Partial Differential Equations, 26 (2001), pp. 101–174.

[18] F. Otto and C. Villani, *Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality*, J. Funct. Anal., 173 (2000), pp. 361–400.

[19] F. Otto and M. Westdickenberg, *Eulerian calculus for the contraction in the Wasserstein distance*, SIAM J. Math. Anal., 37 (2005), pp. 1227–1255.

[20] K.-T. Sturm, *On the geometry of metric measure spaces.* I, Acta Math., 196 (2006), pp. 65–131.

[21] K.-T. Sturm, *On the geometry of metric measure spaces.* II, Acta Math., 196 (2006), pp. 133–177.

[22] J. L. Vázquez, *The Porous Medium Equation*, Oxford Mathematical Monographs, The Clarendon Press Oxford University Press, Oxford, 2007. Mathematical theory.

[23] C. Villani, *Optimal Transport, Old and New*, Springer-Verlag, 2009, to appear.

[24] M.-K. von Renesse and K.-T. Sturm, *Transport inequalities, gradient estimates, entropy, and Ricci curvature*, Comm. Pure Appl. Math., 58 (2005), pp. 923–940.

SIAM J. MATH. ANAL.
Vol. 40, No. 3, pp. 1123–1151

POSITIVITY PROPERTIES OF THE FOURIER TRANSFORM AND THE STABILITY OF PERIODIC TRAVELLING-WAVE SOLUTIONS*

JAIME ANGULO PAVA† AND FÁBIO M. A. NATALI‡

Abstract. In this paper we establish a method to obtain the stability of periodic travelling-wave solutions for equations of Korteweg–de Vries-type $u_t + u^p u_x - Mu_x = 0$, with M being a general pseudodifferential operator and where $p \geq 1$ is an integer. Our approach uses the theory of totally positive operators, the Poisson summation theorem, and the theory of Jacobi elliptic functions. In particular we obtain the stability of a family of periodic travelling waves solutions for the Benjamin–Ono equation. The present technique gives a new way to obtain the existence and stability of cnoidal and dnoidal waves solutions associated with the Korteweg–de Vries and modified Korteweg–de Vries equations, respectively. The theory has prospects for the study of periodic travelling-wave solutions of other partial differential equations.

Key words. dispersive equations, Korteweg–de Vries-type equations, periodic travelling waves, Jacobi elliptic functions, nonlinear stability

AMS subject classifications. 76B25, 35Q51, 35Q53

DOI. 10.1137/080718450

1. Introduction.

One of the main properties of dispersive nonlinear evolution equations is that usually they sustain steadily translating waves called travelling waves. These solutions imply a balance between the effects of nonlinearity and of frequency dispersion. By depending on specific boundary conditions on the wave's shape, for instance, in the case of water waves, these special states of motion can arise either solitary or periodic waves. The study of these special steady waveforms is essential to the explanation of many wave phenomena observed in the practice, for instance, in surface water waves propagating in a canal, in propagation of internal waves, or in shallow-water ocean surface waves (see Benjamin [12], [13], [14] and Osborne et al. [44]).

In the water wave context, Constantin in [21] and Constantin and Escher in [22] analyzed a free boundary problem for harmonic functions and showed that periodic or solitary travelling waves possess stability properties within the shallow-water regime (see also Toland [47] and Constantin and Strauss [23] and the citations therein). Moreover, various nonlinear dispersive model equations are an accurate approximation to the governing equations for water waves (see [5]). From these considerations, questions about the stability of travelling waves and their existence as exact solutions of the dynamical equations are very important.

The solitary waves are in general single crested, symmetric, localized travelling waves, whose hyperbolic sech profiles are well known (see Ono [43] and Benjamin [16] for the existence of solitary waves of algebraic type or with a finite number of oscillations). The study of the nonlinear stability or instability in the form of solitary waves has had a terrific development and refinement in recent years. The proofs have

*Received by the editors March 13, 2008; accepted for publication (in revised form) June 17, 2008; published electronically October 22, 2008.

http://www.siam.org/journals/sima/40-3/71845.html

†Department of Mathematics, IME-USP Rua do Matão 1010, CEP 05508-090, São Paulo - SP - Brazil (angulo@ime.usp.br). This author's research was partially supported by CNPq/Brazil.

‡Department of Mathematics, Universidade Estadual de Maringá Avenida Colombo, 5790, CEP 87020-900, Maringá - PR - Brazil (fmanatali@uem.br). This author's research was supported by FAPESP/Brazil under grant 06/61310-3.

been simplified and sufficient conditions were obtained to ensure the stability to small localized perturbations in the waveform. Those conditions have shown to be effective in a variety of circumstances; see, for example, [2], [3], [4], [14], [17], [27], [28], [49], and [48].

The situation regarding periodic travelling waves is very different. The stability and the existence of explicit formulas of these progressive wavetrains have received comparatively little attention. A first study of these waveforms was detemined by Benjamin in [16] with regard to the periodic steady solutions called *cnoidal waves*, which were found initially by Korteweg and de Vries in [34] for the equation currently called the Korteweg–de Vries equation (KdV henceforth):

$$u_t + uu_x + u_{xxx} = 0, \tag{1.1}$$

where $u = u(x,t)$ is a real-valued function of the two variables $x, t \in \mathbb{R}$. Benjamin put forward an approach to the stability of cnoidal waves in the form

$$\varphi(\xi) = \beta_2 + (\beta_3 - \beta_2)\mathrm{cn}^2\left(\sqrt{\frac{\beta_3 - \beta_1}{12}}\xi; k\right),$$

but did not provide a detailed justification of his assertions, and several aspects seem problematic. The first result of stability for periodic solutions of the KdV was obtained by McKean in [37], who considered the orbital stability of all periodic finite-genus solutions with respect to perturbations of the same period. McKean's approach was based on the integrable structure of the KdV. More recently Angulo, Bona, and Scialom in [10] returned to Benjamin's original question and gave a complete theory of stability of cnoidal waves for (1.1) with respect to perturbations of the same period (see also [7]). The approach for obtaining this result was based on the ideas of Bona, Weinstein, and Grillakis, Shatah, and Strauss (see [17], [27], [48]) but adapted to the periodic context. So new theories of stability for other dispersive equations such as the focusing nonlinear Schrödinger equation and the modified KdV has been obtained (see Angulo [9], [8]). It is remarkable to see that in all these works the use of an elaborated spectral theory for the periodic eigenvalue problem was necessary (see [29], [36]),

$$\begin{cases} \dfrac{d^2}{dx^2}\Psi + [\rho - n(n+1)k^2 sn^2(x;k)]\Psi = 0, \\ \Psi(0) = \Psi(2K(k)), \quad \Psi'(0) = \Psi'(2K(k)), \end{cases} \tag{1.2}$$

with specific values of $n \in \mathbb{N}$. In (1.2), $sn(\cdot;k)$ represents the Jacobi elliptic function of type snoidal with *modulus* k, $k \in (0,1)$, and K represents the complete elliptic integral of the first kind defined by

$$K(k) = \int_0^1 \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}}.$$

We recall that the second order differential equation in (1.2) is known as the *Jacobi form of the Lamé equation*.

We note also that Gardner in [26] provided a theory for determining that the large wavelength periodic waves are linearly unstable whenever the limiting homoclinic wave (solitary wave) is unstable. He applied it for diverse types of nonlinear evolution equations in one space variable, in the case of the generalized KdV equations

$$u_t + u^p u_x + u_{xxx} = 0,$$

$p \in \mathbb{N}$, and assuming that this equation admits a family of large wavelength periodic waves U^α such that the period T_α tends to infinity as α tends to zero, then they are unstable whenever $p > 4$ and $\alpha > 0$ is sufficiently small.

Then, the main focus in this paper will be the study of the existence and stability of periodic travelling-wave solutions for equations of the form

$$u_t + u^p u_x - M u_x = 0, \tag{1.3}$$

where $p \geq 1$ is an integer and M is a differential or pseudodifferential operator in the framework of periodic functions. M is defined as a Fourier multiplier operator by

$$\widehat{Mg}(k) = \alpha(k)\widehat{g}(k), \quad k \in \mathbb{Z}, \tag{1.4}$$

where the symbol α of M is assumed to be a mensurable, locally bounded, even function on $\mathbb{R}$, satisfying the conditions

$$A_1|k|^{m_1} \leqq \alpha(k) \leqq A_2(1+|k|)^{m_2} \tag{1.5}$$

for $m_1 \leqq m_2$, $|k| \geq k_0$, $\alpha(k) > b$ for all $k \in \mathbb{Z}$ and $A_i > 0$. The travelling-wave solutions for (1.3) will have the form

$$u(x,t) = \varphi_c(x - ct),$$

where the profile φ_c is a smooth periodic function with an a priori fundamental period $2L$, $L > 0$. Hence substituting this form of u into (1.3) and integrating once (with the integration constant being considered zero throughout our theory), one obtains that $\varphi = \varphi_c$ is the solution of the equation

$$(M + c)\varphi - \frac{1}{p+1}\varphi^{p+1} = 0. \tag{1.6}$$

Associated with (1.6) we consider the linear, closed, unbounded, self-adjoint operator $\mathcal{L} : D(\mathcal{L}) \to L^2_{per}([-L, L])$ defined on a dense subspace of $L^2_{per}([-L, L])$ by

$$\mathcal{L}u = (M + c)u - \varphi^p u. \tag{1.7}$$

From the theory of compact symmetric operators applied to the periodic eigenvalue problem

$$\begin{cases} \mathcal{L}\psi \quad\quad = \lambda\psi, \\ \psi(-L) = \psi(L), \quad \psi'(-L) = \psi'(L), \end{cases} \tag{1.8}$$

it is possible to see that the spectrum of $\mathcal{L}$ is a countable infinite set of eigenvalues, $\{\lambda_n\}$, with

$$\lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \cdots, \tag{1.9}$$

where $\lambda_n \to \infty$ as $n \to \infty$ (see Proposition 3.1 below for a proof of this assertion). In particular, from (1.6) we obtain that $\mathcal{L}$ has zero as an eigenvalue with eigenfunction

$d\varphi/dx$. As is well known this property of $\mathcal{L}$ is deduced from the invariance of the solutions of (1.3) by translations.

A set of sharp conditions is available in the literature to imply the stability of the orbit generated by φ_c, namely, $\Omega_{\varphi_c} = \{\varphi_c(\cdot + y) : y \in \mathbb{R}\}$. So, we say that Ω_{φ_c} (or φ_c) is stable in $H^{m_2}_{per}([-L, L])$ by the periodic flow generated by (1.3) if, for any $\epsilon > 0$, there exists a $\delta > 0$ such that, for $u_0 \in H^{m_2}_{per}([-L, L])$ with $d(u_0, \Omega_{\varphi_c}) \equiv inf_{y\in\mathbb{R}}$ $\|u_0 - \varphi_c(\cdot + y)\|_{H^{m_2}_{per}} < \delta$, the solution u of (1.3) with $u(x, 0) = u_0$ is global in time and satisfies $d(u(t), \Omega_{\varphi_c}) < \epsilon$ for all $t \in \mathbb{R}$. Thus, from [14], [17], [49], [27] the conditions that imply stability are the following:

(1.10)

(P_0) there is a nontrivial smooth curve of periodic solutions for (1.6) of the form

$c \in I \subseteq \mathbb{R} \to \varphi_c \in H^{m_2}_{per}([-L, L])$;

(P_1) $\mathcal{L}$ has an unique negative eigenvalue λ, and it is simple;

(P_2) the eigenvalue 0 is simple;

(P_3) $\dfrac{d}{dc}\displaystyle\int_{-L}^{L} \varphi_c^2(x)dx > 0.$

The problem about the existence of a nontrivial smooth curve of periodic solutions in the form required by (P_0) above presents new and delicate aspects that need to be handled. The possibility of finding explicit solutions for (1.6) will depend naturally on the form of M. If it is a differential operator of the form $M = -\partial_x^2$, then we use the quadrature method (it means writing (1.6) in the form $[\varphi_c']^2 = F(\varphi_c)$), and the theory of elliptic functions has shown to be a main tool. So, the solutions will depend on the Jacobi elliptic functions of *snoidal, cnoidal*, and *dnoidal* types (see [9], [8], [10], [15]). Now, since the period of these functions depends on the complete elliptic integral $K(k)$, we have that the elliptic modulus k will depend on the velocity c, and therefore we have that a priori the period of φ_c will depend on c. Hence, by using the implicit function theorem, the wanted smooth branch of periodic solutions with a fixed minimal period has been obtained in many cases. We note that the procedure of the quadrature method in general does not work if M is a pseudodifferential operator such as the nonlocal operator $\mathcal{H}\partial_x$, with $\mathcal{H}$ being the Hilbert transform. In this paper we will make a different approach to obtain explicit solutions of (1.6) for a specific form of M and values of p. This approach will be based on the classical *Poisson summation theorem* (see [38], [45], [46]). At least two important advantages of this approach can be obtained: The first one is that it can be used for obtaining solutions when M is a pseudodifferential operator, for example, in the case of $\mathcal{H}\partial_x$. The other one is that related with computing the integral in (1.10). In general obtaining property (P_3) can be very difficult in the periodic case, as the results that have appeared in the literature have shown, since the use of nontrivial identities for the complete elliptic integrals of the first and the second kinds sometimes come on the scene as a fundamental piece in the analysis. As we will see our method to obtain property (P_3) can be very easy as a combination of the *Poisson summation theorem* and the *Parseval theorem*.

With regard to conditions (P_1) and (P_2), the problem is very delicate. One of the most remarkable results in the theory of stability of solitary wave solutions was given by Albert [2] and Albert and Bona [3], where sufficient conditions to obtain properties (P_1) and (P_2) were given. The advantage of that approach is that it does not require an explicit computation of the spectrum of the linear operator (1.7), since (P_1) and

(P_2) are obtained exclusively from positivity properties of the Fourier transform of the solitary wave in question. The present paper establishes an extension of the theory in [2] and [3] in the case of periodic travelling-wave solutions. The periodic problem has new points not encountered when considering issues related to the solitary waves. Our analysis also relies upon the theory of totally positive operators, and so the class $PF(2)$ defined by Karlin in [32] (see also [2]) is basic in our study.

Our theory leads to a significant simplification of some recent proofs of stability of periodic travelling-wave solutions of KdV-type equations (see [9], [7], [10]) such as in the case of the KdV and the modified KdV equations, since in those cases the verification of properties (P_1) and (P_2) required the determination of the instability intervals associated with the Lamé equation in (1.2) and of an explicit formula of at least the first three eigenvalues ρ (see [36]). *Our analysis does not need this information.*

Such as will be shown in section 4, this method establishes the first result about the stability of periodic travelling-wave solutions found by Benjamin in [15] for the Benjamin–Ono equation

$$u_t + uu_x - \mathcal{H}u_{xx} = 0, \tag{1.11}$$

where $\mathcal{H}$ denotes the periodic Hilbert transform defined by

$$\mathcal{H}f(x) = \frac{1}{2L}\, p.v. \int_{-L}^{L} cotg\Big[\frac{\pi(x-y)}{2L}\Big] f(y)\, dy. \tag{1.12}$$

The associated periodic waves for (1.11) with a minimal period $2L$ are given for $c > \frac{\pi}{L}$ as

$$\varphi_c(x) = \frac{2\pi}{L}\left(\frac{\operatorname{senh}(\gamma)}{\cosh(\gamma) - \cos\left(\frac{\pi x}{L}\right)}\right), \tag{1.13}$$

such that $\gamma > 0$ satisfies $\tanh(\gamma) = \frac{\pi}{cL}$.

It is important to note that the stability results now presented are obtained by periodic initial disturbances having *the same minimal period* of our periodic solutions. This method cannot be extended for obtaining stability results with more general periodic perturbations, for instance, by periodic disturbances of two times the minimal period of our periodic solutions. In section 6 we give an explanation of this fact.

The plan of this paper is as follows. The next section is devoted to describing briefly the notation that will be used and making a few preliminary remarks regarding periodic Sobolev spaces, the Poisson summation theorem, and some results of global well-posedness in the periodic case of the KdV, modified KdV, and the Benjamin–Ono equations. Sections 3 and 4 contain our full theory which relates positivity properties of periodic travelling-wave solutions to the stability theory of [27]. Applications of section 4 to specific periodic travelling waves are presented in section 5. In section 6 some comments about the $PF(2)$ property in the periodic case are established. Finally, the appendix contains some basic properties of the elliptic integrals which are relevant to the theory of section 5, and the proof of an inequality of Poincaré–Wintinger type for nonlocal operators is established.

2. Notation and preliminaries.

2.1. Function classes. Let Ω be an open set of the real line $\mathbb{R}$ and $1 \leq p \leq \infty$; then $L^p(\Omega)$ is the usual Banach space of (equivalence classes of) real- or

complex-valued Lebesgue measurable functions defined on Ω provided with the norm

$$||f||_p = \left(\int_\Omega |f(x)|^p dx \right)^{\frac{1}{p}}$$

if $1 \leq p < \infty$. When $p = \infty$ we have $||f||_\infty = supess_{x\in\Omega}|f(x)|$. When $p = 2$ the Banach space $L^p(\Omega)$ is a Hilbert space with inner product defined by $(f,g)_2 = \int_\Omega f(x)\overline{g(x)}dx$, where $f, g \in L^2(\Omega)$. The L^2-based Sobolev spaces of periodic functions are defined as follows [30]. If $\mathcal{P} = C^\infty_{per}$ denotes the collection of all of the functions $f : \mathbb{R} \to \mathbb{C}$ which are C^∞ and periodic with period $2l > 0$, the collection $\mathcal{P}'$ of all continuous linear functionals from $\mathcal{P}$ into $\mathbb{C}$ is the set of periodic distributions. If $\Psi \in \mathcal{P}'$, we denote the evaluation of Ψ at φ by $\Psi(\varphi) = \langle \Psi, \varphi\rangle$ for $\varphi \in \mathcal{P}$. For $k \in \mathbb{Z}$, let $\Theta_k(x) = e^{\frac{ik\pi x}{l}}$, $x \in \mathbb{R}$. The Fourier transform of $\Psi \in \mathcal{P}'$ is a function $\widehat{\Psi} : \mathbb{Z} \to \mathbb{C}$ defined by $\widehat{\Psi}(k) = \frac{1}{2l}\langle \Psi, \Theta_k\rangle$, $k \in \mathbb{Z}$. $\widehat{\Psi}(k)$ are called the Fourier coefficients of Ψ. As usual, a function $\psi \in L^p([-l,l])$, $p \geq 1$, is an element of $\mathcal{P}'$ by defining

$$\langle \psi, \varphi\rangle = \frac{1}{2l}\int_{-l}^{l} \psi(x)\varphi(x)dx, \qquad \varphi \in \mathcal{P}.$$

If $\psi \in L^p([-l,l])$ for some $p \geq 1$, then, for $k \in \mathbb{Z}$,

$$\widehat{\psi}(k) = \frac{1}{2l}\int_{-l}^{l} \psi(x)e^{-\frac{ik\pi x}{l}}dx.$$

The Fourier inverse transform of a sequence $\alpha = (\alpha_k)_{k\in\mathbb{Z}} \in \mathcal{S}(\mathbb{Z})$, where $\mathcal{S}(\mathbb{Z})$ denotes the space of the rapidly decreasing sequences, is the function $\check{\alpha} \in \mathcal{P}$ defined by $\check{\alpha}(x) = \sum_{k\in\mathbb{Z}} \alpha_k\Theta_k(x)$. We consider the space $\mathcal{P}'$ provided with the usual weak-star topology, but it will not be needed here. We denote by C_{2l} the space of the continuous and $2l$-periodic functions. Let $\alpha = (\alpha_k)_{k\in\mathbb{Z}}$ be a sequence of complex value. The Hilbert space $\ell^2 := \ell^2(\mathbb{Z})$ is defined by

$$\ell^2 = \left\{ \alpha;\ ||\alpha||_{\ell^2} := \left(\sum_{k=-\infty}^{+\infty} |\alpha_k|^2 \right)^{\frac{1}{2}} < \infty \right\}.$$

For $s \in \mathbb{R}$, the Sobolev space $H^s_{per}([-l,l]) := H^s_{2l}$ is the set of all $f \in \mathcal{P}'$ such that

$$||f||^2_{H^s_{2l}} \equiv 2l \sum_{k=-\infty}^{+\infty} (1+|k|^2)^s |\widehat{f}(k)|^2 < \infty.$$

The collection H^s_{2l} is a Hilbert space with inner product

$$(f,g)_{H^s_{2l}} = 2l \sum_{k=-\infty}^{+\infty} (1+|k|^2)^s \widehat{f}(k)\overline{\widehat{g}(k)}.$$

When $s = 0$, H^s_{2l} is a Hilbert space that is isometrically isomorphic to a subspace of $L^2([-l,l])$ and $(f,g)_{H^0_{2l}} = (f,g) = \int_{-l}^{l} f(x)\overline{g(x)}dx$. The space H^0_{2l} will be denoted by L^2_{2l}, and its norm will be $||.||_{L^2_{2l}}$. Of course, $H^s_{2l} \subseteq L^2_{2l}$ for all $s \geq 0$, and we have

for $s > 1/2$ the Sobolev embedding $H^s_{2l} \hookrightarrow C_{2l}$ (see [30]). The space $\ell^2_{s,2l} := \ell^2_{s,2l}(\mathbb{Z})$, $s \in \mathbb{R}$, is defined by

$$\ell^2_{s,2l}(\mathbb{Z}) := \left\{ \alpha = (\alpha_k)_{k\in\mathbb{Z}};\ ||\alpha||_{\ell^2_s} := \left(2l \sum_{k=-\infty}^{+\infty} (1+|k|^2)^s |\alpha_k|^2 \right)^{\frac{1}{2}} < +\infty \right\}.$$

$\ell^2_{s,2l}$ is a Hilbert space with inner product

$$(\alpha,\beta)_{\ell^2_{s,2l}} = 2l \sum_{k=-\infty}^{+\infty} (1+|k|^2)^s \alpha_k \overline{\beta_k},$$

where $\alpha = (\alpha_k)_{k\in\mathbb{Z}}$ and $\beta = (\beta_k)_{k\in\mathbb{Z}}$. Then we have that $f \in H^s_{2l}$ if and only if $(\widehat{f}(k))_{k\in\mathbb{Z}} \in \ell^2_{s,2l}$, and so from the Parseval theorem (see [30]), $||\widehat{f}\ ||^2_{\ell^2} = \frac{1}{2l}||f||^2_{L^2_\ell}$, it follows that $||f||_{H^s_{2l}} = ||\widehat{f}||_{\ell^2_{s,2l}}$. The convolution of two sequences α and β is the sequence $\alpha * \beta$ defined, for all $k \in \mathbb{Z}$, by $(\alpha*\beta)_k = \sum_{j=-\infty}^{+\infty} \alpha_{k-j}\beta_j$, whenever the right-hand side of the identity above makes sense. Next, we present some results that we will need throughout this work. We start with the Young inequality (see [38]).

PROPOSITION 2.1. *Let $\alpha \in \ell^1(\mathbb{Z})$ and $\beta \in \ell^2(\mathbb{Z})$. Then $\alpha * \beta \in \ell^2(\mathbb{Z})$. Moreover,*

$$||\alpha*\beta||_{\ell^2} \le ||\alpha||_{\ell^1}||\beta||_{\ell^2}.$$

In particular, for every $\alpha \in \ell^1$ fixed, the linear operator $\beta \in \ell^2 \mapsto \alpha*\beta \in \ell^2$ is continuous.

Now, we present the Poisson summation theorem. It will be used to find the explicit form of the periodic travelling-wave solutions for some equations.

THEOREM 2.1. *Let $\widehat{f}(x) = \int_{-\infty}^{+\infty} f(y)e^{-ixy}dy$ and $f(y) = \int_{-\infty}^{\infty} \widehat{f}(x)e^{ixy}dx$ satisfying*

$$|f(y)| \le \frac{A}{(1+|y|)^{1+\delta}} \qquad \text{and} \qquad |\widehat{f}(x)| \le \frac{A}{(1+|x|)^{1+\delta}},$$

where $\delta > 0$ and $A > 0$ (then $\widehat{f}$ and f can be assumed to be continuous functions). Thus, for $L > 0$,

$$\sum_{n=-\infty}^{+\infty} f(x+2Ln) = \frac{1}{2L} \sum_{n=-\infty}^{+\infty} \widehat{f}\left(\frac{n}{2L}\right) e^{\frac{\pi i n x}{L}}.$$

The two series above converge absolutely.

Proof. See [38], [45], and [46]. □

2.2. Results of global well-posedness. Some results about local and global well-posedness associated with (1.3) in the periodic case were initially established in [1]. Here we establish two results that we will use in our theory.

THEOREM 2.2. *Let $s \ge 1$ be given. For each $u_0 \in H^s_{2l}$ there is a unique solution of (1.3), for the cases $p = 1, 2$ and $M = -\partial_x^2$, that for each $T > 0$ lies in $C(0,T;H^s_{2l})$. Moreover, the correspondence $u_0 \mapsto u$ is an analytic function of the relevance function spaces.*

Proof. See [20]. □

THEOREM 2.3. *Let $s \ge \frac{1}{2}$ be given. For each $u_0 \in H^s_{2l}$ there is a unique solution of (1.3), for the cases $p = 1$ and $M = \mathcal{H}\partial_x$, that for each $T > 0$ lies in $C(0,T;H^s_{2l})$.*

Moreover, the correspondence $u_0 \mapsto u$ is a continuous function of the relevance function spaces.

Proof. See [39], [41], and [40]. □

3. Basic functional spaces. In this section we establish the main spaces in our study of the stability of periodic wave solutions associated with (1.6) for the case of the Benjamin–Ono (BO), Korteweg–de Vries (KdV), and modified Korteweg–de Vries (mKdV) equations. For the KdV and mKdV cases we have in (1.6), $M = -\partial_x^2$ with symbol $\left(\frac{\pi}{L}\right)^2 n^2$, $p = 1$, and $p = 2$, respectively. Next, the BO equation is obtained with $M = \mathcal{H}\partial_x$, with the symbol being $\frac{\pi}{L}|n|$ and $p = 1$. Here, $\mathcal{H}$ denotes the periodic Hilbert transform defined in (1.12) and such that via the Fourier transform satisfies $\widehat{\mathcal{H}f}(n) = -isgn(n)\widehat{f}(n)$ for all $n \in \mathbb{Z}$. Then, our three periodic travelling-wave equations are

$$\begin{aligned} &\mathcal{H}\varphi_c' + c\varphi_c - \tfrac{1}{2}\varphi_c^2 = 0 \qquad \text{(BO)},\\ &\varphi_c'' + \tfrac{1}{2}\varphi_c^2 - c\varphi_c = 0 \qquad \text{(KdV)},\\ &\varphi_c'' + \tfrac{1}{3}\varphi_c^3 - c\varphi_c = 0 \qquad \text{(mKdV)}. \end{aligned} \tag{3.1}$$

Remark 3.1. (a) Here, we shall consider a more convenient form for the mKdV travelling-wave equation, namely,

$$\varphi_c'' + \varphi_c^3 - c\varphi_c = 0. \tag{3.2}$$

(b) The periodic travelling-wave solutions for the KdV and mKdV will be considered of period L, but the periodic travelling-wave solution for the BO equation will be considered of period $2L$ only by convenience.

(c) After a "bootstrap" argument, we can conclude that every φ_c belongs to H^s_{2L} for all $s \in \mathbb{R}$. Thus, φ is infinitely differentiable, with all derivatives in L^2_{2L}.

We will suppose that $c > -b$, where b satisfies $\alpha(k) > b$ for all $k \in \mathbb{Z}$. With this condition $M + c$ represents a positive operator. Then, by using the spectral theorem for compact and self-adjoint operators we have the following characterization of the spectrum of $\mathcal{L}$ defined in (1.7).

PROPOSITION 3.1. *The operator $\mathcal{L}$ in* (1.7) *is a closed, unbounded, self-adjoint operator on L^2_{2L} whose spectrum consists of an enumerable (infinite) set of eigenvalues $\{\lambda_k\}_{k=0}^{\infty}$ satisfying $\lambda_0 \le \lambda_1 \le \lambda_2 \le \cdots$, and $\lambda_k \to \infty$ as $k \to \infty$. In particular, $\mathcal{L}$ has 0 as an eigenvalue with eigenfunction $\frac{d}{dx}\varphi_c$.*

Proof. We suppose that our periodic functions have period L. Clearly $\mathcal{L}$ defined on $H^{m_2}_L$ is a closed, unbounded, self-adjoint operator on $L^2_{per}([0, L])$. Let us proof that the spectrum of $\mathcal{T} := M + c$ is a countable infinite set of eigenvalues, $\{\gamma_n\}$, with

$$\gamma_0 \le \gamma_1 \le \gamma_2 \le \gamma_3 \le \cdots, \tag{3.3}$$

where $\gamma_n \to \infty$ as $n \to \infty$. In fact, let $R_c = (M + c)^{-1}$, whose symbol is $\frac{1}{c+\alpha(k)}$ for $k \in \mathbb{Z}$. Since $\frac{1}{c+\alpha(k)} \in \ell^2(\mathbb{Z})$ we have that there is a unique $G_c \in L^2_{per}([0, L])$ such that $\widehat{G_c}(k) = \frac{1}{c+\alpha(k)}$, and, because of this, we have the action

$$R_c f(x) = \frac{1}{L}\int_0^L G_c(x - y)f(y)dy,$$

defined for $f \in L^2_{per}([0, L])$. Since $[0, L]$ is a bounded set we have that the kernel $\widetilde{G_c}(x, y) := G_c(x - y) \in L^2([0, L] \times [0, L])$. So, R_c is a Hilbert–Schmidt operator on

$L^2_{per}([0,L])$ (see [33]), and therefore R_c is a compact operator on $L^2_{per}([0,L])$ for all $c>0$ (here we supposed without loss of generality that $b=0$), and so we obtain (3.3).

Next, we will show that there is a μ_1 (large enough) such that $\mathcal{M}=(\mathcal{L}+\mu_1)^{-1}$ exists and is a bounded, positive, compact, and self-adjoint operator. In fact, first of all, it is easy to see that $\mathcal{L}$ is limited below; that is, if $f\in D(\mathcal{L})$, we have $\langle \mathcal{L}f,f\rangle \geq -\beta\langle f,f\rangle$, where $\beta=||\varphi_c||_{L^\infty_{per}}+c$. Then, we can choose a μ_1 such that $\mathcal{L}+\mu_1>0$; that is, $\mathcal{M}$ is positive. We denote $\mu_1:=\mu$ only by convenience. Let ν be a positive number such that $\nu+\varphi_c-c>0$ and $\nu+\mu>0$. Thus, for $\mu>0$ we have $f=(\mathcal{L}+\mu)g \Leftrightarrow (I-M)g=\Upsilon f$, where $Mg=R_{\mu+\nu}[(\nu+\varphi_c-c)g]$, $\Upsilon=R_{\nu+\mu}$, and we denote $h=\nu+\varphi_c-c$. Next, from the Parseval theorem, it follows that

$$||Mg||_{L^2_{per}} \leq \sup_{k\in\mathbb{Z}}\left\{\frac{1}{\alpha(k)+\nu+\mu}\right\}||h||_{L^\infty_{per}}||g||_{L^2_{per}}.$$

Thus, we can choose μ such that $||M||_{B(L^2_{per})}<1$ and $\mathcal{L}+\mu>0$. Then, $I-M$ is invertible, and we have $g=(I-M)^{-1}\Upsilon f$ and write $\mathcal{M}=(\mathcal{L}+\mu)^{-1}=(I-M)^{-1}\Upsilon$. Υ being a compact operator and $(I-M)^{-1}\in B(L^2_{per})$ we have that $\mathcal{M}$ is a compact operator. Then, there is an orthonormal basis $\{\varphi_k\}_{k=0}^\infty$ of L^2_{per} consisting of eigenfunctions of $\mathcal{M}$ with nonzero eigenvalues $\{\mu_k\}_{k=0}^\infty$ satisfying $\mu_1\geq\mu_2\geq\mu_3\geq\cdots>0$, and $\mu_k\to 0$ as $k\to\infty$. Since $\mathcal{M}\varphi_k=\mu_k\varphi_k\in D(\mathcal{L}+\mu)$ we have that

$$\mathcal{L}\varphi_k=\left(\frac{1}{\mu_k}-\mu\right)\varphi_k:=\lambda_k\varphi_k.$$

Thus, there is a sequence of eigenvalues of $\mathcal{L}$, $\{\lambda_k\}_{k=0}^\infty$, satisfying $\lambda_0\leq\lambda_1\leq\lambda_2\leq\cdots$, and $\lambda_k\to\infty$ as $k\to\infty$. This argument shows what is desired. □

The next step will be centralized in the study of some specific spectral properties of the operator $\mathcal{L}$. For this, let us define two families of linear operators. They will be related with $\mathcal{L}$ but with the advantage that both of them are compacts. The results that will be present below are extensions of the results about stability of solitary waves solutions in Albert [2] and Albert and Bona [3] to the periodic case.

For every $\theta\geq 0$ define the operator $S_\theta:\ell^2(\mathbb{Z})\to\ell^2(\mathbb{Z})$ by considering

$$S_\theta\alpha(n)=\frac{1}{\omega_\theta(n)}\sum_{j=-\infty}^{\infty}K(n-j)\alpha_j=\frac{1}{\omega_\theta(n)}(K*\alpha)_n,$$

where $\omega_\theta(n)=\alpha(n)+\theta+c$, $K(n)=\widehat{\varphi_c^p}(n)$, $n\in\mathbb{Z}$. Since $\omega_\theta(n)>0$ for all $n\in\mathbb{Z}$, it follows that the space X defined by

$$X=\left\{\alpha\in\ell^2(\mathbb{Z});||\alpha||_{X,\theta}:=\left(\sum_{n=-\infty}^{\infty}|\alpha_n|^2\omega_\theta(n)\right)^{\frac{1}{2}}<\infty\right\}$$

is a Hilbert space with norm $||\alpha||_{X,\theta}$ and corresponding inner product $\langle\alpha,\beta\rangle_{X,\theta}=\sum_{n=-\infty}^{\infty}\alpha_n\overline{\beta_n}\omega_\theta(n)$.

PROPOSITION 3.2. (a) *If $\alpha\in\ell^2$ is an eigensequence of S_θ for a nonzero eigenvalue, then $\alpha\in X$.*

(b) *The restriction of S_θ to X is a compact, self-adjoint operator on X with respect to the norm $||\cdot||_{X,\theta}$.*

Proof. In fact, we denote the norm in X simply by $||\cdot||_X$, and the operator S_θ by S, and $\mu:=\mu_\theta=\frac{1}{\omega_\theta}$. It will be shown that $S^2g=S(Sg)\in X$, since $g=\frac{1}{\gamma}Sg=\frac{1}{\gamma^2}S^2g$;

this will prove the proposition. By the Minkowski, Hölder, and Young inequalities we obtain

(3.4)

$$
\begin{aligned}
||S^2\alpha||_X &= \left\| \sum_j K(\cdot - j)\mu(\cdot) S\alpha(j) \right\|_X = \left\| \sum_j \sum_m K(\cdot - j)\mu(\cdot) K(j-m)\mu(j)\alpha(m) \right\|_X \\
&\leq \sum_j \sum_m \left[\sum_n K(n-j)^2 \mu(n) \right]^{\frac{1}{2}} K(j-m)\mu(j)|\alpha(m)| \\
&\leq ||K^2||_{\ell^1} ||\mu||_{\ell^2}^{\frac{1}{2}} ||\mu^{\frac{3}{2}}||_{\ell^1}^{\frac{1}{4}} ||\mu^{\frac{5}{4}}||_{\ell^1}^{\frac{1}{2}} ||\alpha||_{\ell^2}.
\end{aligned}
$$

Since $|\mu(x)| \leq B(1+|x|)^{-1}$ for every $x \in \mathbb{R}$ and some $B > 0$, all of the quantities in the last expression on the right-hand side in (3.4) are finite. The proof of (a) is complete.

Next, we prove that $S\big|_X$ determines a compact and self-adjoint operator. Let

$$
\widetilde{K}(n,j) = \frac{K(n-j)}{\omega_\theta(n)\omega_\theta(j)};
$$

then, by the Cauchy–Schwarz inequality we have

$$
\begin{aligned}
&\sum_n \sum_j \widetilde{K}^2(n,j)\omega_\theta(n)\omega_\theta(j) = \sum_n \sum_j \frac{K^2(n-j)}{\omega_\theta(n)\omega_\theta(j)} \\
= \quad &\sum_n (K^2 * \mu)(n)\mu(n) \leq ||K^2 * \mu||_{\ell^2} ||\mu||_{\ell^2} \leq ||K^2||_{\ell^1} ||\mu||_{\ell^2}^2 < \infty.
\end{aligned}
$$

That is, $S_\theta\alpha(n) = \sum_{j=-\infty}^{\infty} \widetilde{K}(n,j)\alpha_n\omega_\theta(j)$ is a Hilbert–Schmidt operator, and thus $S\big|_X$ is compact. To prove that S_θ is self-adjoint it is necessary to observe that ω_θ and $\widehat{\varphi_c}$ being even, the real kernel K is symmetric. □

The next two results are immediate consequences of Proposition 3.2 and the spectral theorem for compact, self-adjoint operators on a Hilbert space.

COROLLARY 3.1. *Suppose $\theta \geq 0$. Then, 1 is an eigenvalue of S_θ (as an operator of X) if and only if $-\theta$ is an eigenvalue of $\mathcal{L}$ (as an operator of L^2_{2L}). Furthermore, both eigenvalues have the same multiplicity.* □

COROLLARY 3.2. *For every $\theta \geq 0$, S_θ has a family of eigensequences $\{\psi_{i,\theta}\}_{i=0}^{\infty}$ forming an orthonormal basis of X with respect to the norm $||\cdot||_{X,\theta}$. The eigensequences correspond to real eigenvalues $\{\lambda_i(\theta)\}_{i=0}^{\infty}$ whose only possible accumulation point is zero.* □

In this way, the eigenvalues can be enumerated in order of decreasing absolute value, that is, $|\lambda_0(\theta)| \geq |\lambda_1(\theta)| \geq |\lambda_2(\theta)| \geq \cdots$.

4. Positivity properties for periodic travelling-wave solutions. In this section we give sufficient conditions to obtain properties (P_1) and (P_2) in (1.10).

DEFINITION 4.1. *We say that a sequence $\alpha = (\alpha_n)_{n\in\mathbb{Z}} \subseteq \mathbb{R}$ is in the class $PF(2)$ discrete if*

(i) $\alpha_n > 0$ *for all* $n \in \mathbb{Z}$,

(ii) $\alpha_{n_1-m_1}\alpha_{n_2-m_2} - \alpha_{n_1-m_2}\alpha_{n_2-m_1} > 0$ *for* $n_1 < n_2$ *and* $m_1 < m_2$.

The definition above is a particular case of the continuous ones which appear in [2] and [32]; namely, we say that a function $g : \mathbb{R} \to \mathbb{R}$ is in the class $PF(2)$ if

(i) $g(x) > 0$ for all $x \in \mathbb{R}$,

(ii) $g(x_1 - y_1)g(x_2 - y_2) - g(x_1 - y_2)g(x_2 - y_1) > 0$ for $x_1 < x_2$ and $y_1 < y_2$.

As an example, consider $g(x) = \operatorname{sech}^2(x)$.

The next result gives us a sufficient condition for a function g belonging to the $PF(2)$ continuous class. This result is very useful for our purpose (see [3]).

LEMMA 4.1. *Suppose g is a positive, twice-differentiable function on $\mathbb{R}$ satisfying*

$$\frac{d^2}{dx^2}(\log g(x)) < 0 \quad \textit{for} \quad x \neq 0.$$

Then $g \in PF(2)$. □

The main result of this paper is now presented.

THEOREM 4.1. *Suppose that φ_c is a positive even solution of* (1.6) *such that $\widehat{\varphi_c} > 0$ and $K = \widehat{\varphi_c^p} \in PF(2)$ discrete. Then (P_1) and (P_2) in* (1.10) *hold for the operator $\mathcal{L}$ in* (1.7).

Proof. First, we noticed that, S_θ being a compact operator on X, we get a set of eigenvalues $\{\lambda_i(\theta)\}_{i=0}^{\infty}$ and the corresponding set of eigenfunctions $\{\psi_{i,\theta}\}_{i=0}^{\infty}$, which form an orthonormal basis for X. Moreover, we have $|\lambda_0(\theta)| \geq |\lambda_1(\theta)| \geq |\lambda_2(\theta)| \geq \cdots$. It will show that the eigenvalues $\lambda_0(\theta)$ and $\lambda_1(\theta)$ *are positive, distinct, and simple.* In fact, since $S_\theta\big|_X$ is a compact, self-adjoint operator it follows that

$$\lambda_0(\theta) = \pm \sup_{||\alpha||_X = 1} |\langle S_\theta \alpha, \alpha \rangle_X|. \tag{4.1}$$

Let $\psi(\theta) := \psi$ be an eigensequence of S_θ corresponding to $\lambda_0(\theta) := \lambda_0$. We will show that ψ is one-signed; that is, either $\psi(n) \leq 0$ or $\psi(n) \geq 0$. By contradiction, suppose ψ takes both negative and positive values. Since by hypotheses the kernel K *is positive* we have

$$\begin{aligned} S_\theta|\psi|(n) &= \frac{1}{\omega_\theta(n)} \sum_{j=-\infty}^{\infty} K(n-j)\psi^+(j) + \frac{1}{\omega_\theta(n)} \sum_{j=-\infty}^{\infty} K(n-j)\psi^-(j) \\ &> \left| \frac{1}{\omega_\theta(n)} \sum_{j=-\infty}^{\infty} K(n-j)\psi^+(j) - \frac{1}{\omega_\theta(n)} \sum_{j=-\infty}^{\infty} K(n-j)\psi^-(j) \right|, \end{aligned}$$

where ψ^+ e ψ^- are the positive and negative parts of ψ, respectively. Then,

$$S_\theta|\psi|(n) > \left| \frac{1}{\omega_\theta(n)} \sum_{j=-\infty}^{\infty} K(n-j)\psi(j) \right| = |S_\theta\psi(n)| = |\lambda_0||\psi(n)|,$$

where "$>$" holds because ψ, by supposition, takes both positive and negative values. From the last inequality we conclude that

$$\begin{aligned} \langle S_\theta(|\psi|), |\psi| \rangle_{X,\theta} &= \sum_{n=-\infty}^{\infty} S_\theta|\psi|(n)|\psi(n)|\omega_\theta(n) \\ &> \sum_{n=-\infty}^{\infty} |\lambda_0||\psi(n)|^2\omega_\theta(n) = |\lambda_0|\|\psi\|_{X,\theta}^2. \end{aligned}$$

Hence, if we assume that $||\psi||_X = 1$, we obtain $\langle S_\theta(|\psi|), |\psi|\rangle_X\rangle|\lambda_0|$, which contradicts (4.1). Then, there is an eigensequence ψ_0 which is nonnegative. Since K *is a positive sequence* and $S_\theta(\psi_0) = \lambda_0\psi_0$, we have $\psi_0(n) > 0$ for all $n \in \mathbb{Z}$. Now, such a ψ_0 cannot be orthogonal to any nontrivial one-signed eigensequence in X, and so λ_0 is a simple eigenvalue. Notice that the preceding argument also shows that $-\lambda_0$ cannot be an eigenvalue of S_θ; therefore it follows that $|\lambda_1| < \lambda_0$.

Next, we will study the eigenvalue $\lambda_1(\theta) = \lambda_1$. But, first, we need some definitions and results. We consider the following set of indices:

$$\triangle = \{(n_1, n_2) \in \mathbb{Z} \times \mathbb{Z};\ n_1 < n_2\}.$$

Denoting $\overline{n} = (n_1, n_2)$ and $\overline{m} = (m_1, m_2)$, we define for $\overline{n}$, $\overline{m} \in \triangle$ the following sequence:

$$K_2(\overline{n}, \overline{m}) := K(n_1 - m_1)K(n_2 - m_2) - K(n_1 - m_2)K(n_2 - m_1).$$

By hypothesis $K \in PF(2)$ discrete it follows that $K_2 > 0$. Next, let $\ell^2(\triangle)$ be defined as

$$\ell^2(\triangle) = \left\{ \alpha = (\alpha_{\overline{n}})_{\overline{n}\in\triangle}; \sum\sum_{\triangle} |\alpha_{\overline{n}}|^2 := \sum_{n_1\in\mathbb{N}} \sum_{\substack{n_1<n_2 \\ n_2\in\mathbb{Z}}} |\alpha(n_1, n_2)|^2 < +\infty \right\},$$

and define the operator $S_{2,\theta} : \ell^2(\triangle) \to \ell^2(\triangle)$ by

$$S_{2,\theta}g(\overline{n}) = \sum\sum_{\triangle} G_{2,\theta}(\overline{n}, \overline{m})g(\overline{m}),$$

where $G_{2,\theta}(\overline{n}, \overline{m}) = \frac{K_2(\overline{n},\overline{m})}{\omega_\theta(n_1)\omega_\theta(n_2)}$. We also consider the space

$$W = \left\{ \alpha \in \ell^2(\triangle); ||\alpha||_{W,\theta} := \left(\sum\sum_{\triangle} |\alpha(\overline{n})|^2 \omega_\theta(n_1)\omega_\theta(n_2) \right)^{\frac{1}{2}} < \infty \right\}.$$

Then W is a Hilbert space with norm $||\cdot||_{W,\theta}$ given above and with inner product

$$\langle \alpha, \beta\rangle_{W,\theta} = \sum\sum_{\triangle} \alpha(\overline{n})\overline{\beta(\overline{n})}\omega_\theta(n_1)\omega_\theta(n_2).$$

Remark 4.1. (1) We can show, in an analogous way to Proposition 3.2, that $S_{2,\theta}\big|_W$ is a self-adjoint, compact operator. Therefore, the associated eigenvalues can be enumerated in order of decreasing absolute value, that is, $|\mu_0(\theta)| \geq |\mu_1(\theta)| \geq |\mu_2(\theta)| \geq \cdots$.

(2) A similar argument can be used to show that $\mu_0(\theta) := \mu_0$ is positive and simple and $|\mu_1| < \mu_0$.

Definition 4.2. *Let α^1, $\alpha^2 \in \ell^2(\mathbb{Z})$; we define the wedge product $\alpha^1 \wedge \alpha^2$ in $\triangle$ by*

$$(\alpha^1 \wedge \alpha^2)(n_1, n_2) = \alpha^1(n_1)\alpha^2(n_2) - \alpha^1(n_2)\alpha^2(n_1).$$

We have the following results from Definition 4.2.

Lemma 4.2. *Let $A = \{\alpha^1 \wedge \alpha^2;$ for $\alpha^1,\ \alpha^2 \in X,\ \alpha^1 \wedge \alpha^2 \in \ell^2(\triangle)\}$. Then A is dense in W.*

Proof. See [31] and [32]. □

LEMMA 4.3. *Let* $\alpha^1,\ \alpha^2 \in \ell^2(\mathbb{Z})$. *Then* $S_{2,\theta}(\alpha^1 \wedge \alpha^2) = S_\theta\alpha^1 \wedge S_\theta\alpha^2$.

Proof. See [31] and [32]. □

In what follows, we will represent by S_θ^X the restriction of S_θ on the Hilbert space X. We shall use some spectral results in Kato [33]. In fact, we decompose S_θ^X into the form $S_\theta^X = \lambda_0 P_\theta + Q_\theta$, where P_θ is the orthogonal projection on $M_0 = [\psi_0]$, $P_\theta Q_\theta = Q_\theta P_\theta = 0$, and $spectrum(Q_\theta) = spectrum(P_\theta) \setminus \{\lambda_0\}$. Moreover, in this case we have that the spectral radius of Q_θ when restricted to the subspace $N = \mathrm{Ker}P_\theta$ is exactly $|\lambda_1|$. Furthermore, we know that the eigenvalue $\lambda_0 = \lambda_0(\theta)$ is differentiable with respect to $\theta \geq 0$. The same argument can be applied to μ_0. Then, as a consequence ψ_0 and $\tau_0 =$ eigensequence associated with μ are differentiable eigensequences with respect to θ.

LEMMA 4.4. (a) *In the notation given above we have*

$$\frac{(S_\theta^X)^m}{\lambda_0^m} \longrightarrow P_\theta,$$

where $m \to +\infty$ *on the strong topology of* $B(X, X)$.

(b) *The statement in part* (a) *is valid if* S_θ^X *is replaced by* $S_{2,\theta}$, *the eigenvalues* λ_0 *and* λ_1 *are replaced by* μ_0 *and* μ_1, *and* P_θ, M_0, *and* Q_θ *are replaced by appropriate operators and subspaces of* X.

Proof. The proof is analogous as viewed in [2], [3] in the case of sequences. □

LEMMA 4.5. (a) $\mu_0(\theta) = \lambda_0(\theta)\lambda_1(\theta)$. *Then, we can conclude that* $\lambda_1 > 0$.

(b) λ_1 *is simple.*

Proof. (a) In fact, from Lemma 4.3 we have that $\lambda_0\lambda_1$ is an eigenvalue of $S_{2,\theta}$ whose eigensequence is $\psi_0 \wedge \psi_1$, where $\psi_1(\theta) := \psi_1$ is the eigensequence associated with λ_1. Hence $\mu_0 \geq \lambda_0|\lambda_1|$. Then, since $-\mu_0$ cannot be an eigenvalue of $S_{2,\theta}$, we will show that $\mu_0 \leq \lambda_0|\lambda_1|$. If $|\lambda_1| < \frac{\mu_0}{\lambda_0}$, let P_θ be as in Lemma 4.4 and write $W = M_0 \oplus N$. Let $\alpha^1 = r_1\psi_0 + \omega\gamma^1$ and $\alpha^2 = r_2\psi_0 + \omega\gamma^2$, where $r_1,\ r_2 \in \mathbb{R}$ and $\gamma^1,\ \gamma^2 \in N$. For instance, from the induction principle we have

(4.2)

$$\begin{aligned}\left(\frac{S_{2,\theta}}{\mu_0}\right)^m \left(\alpha^1 \wedge \alpha^2\right)(n_1, n_2) &= r_1\left[\psi_0(n_1)\left(\frac{S_\theta}{\beta}\right)^m \gamma^2(n_2) - \psi_0(n_2)\left(\frac{S_\theta}{\beta}\right)^m \gamma^2(n_1)\right] \\ &+ r_2\left[\psi_0(n_2)\left(\frac{S_\theta}{\beta}\right)^m \gamma^1(n_1) - \psi_0(n_1)\left(\frac{S_\theta}{\beta}\right)^m \gamma^1(n_2)\right] \\ &+ \left(\frac{S_\theta}{\lambda_0}\right)^m \gamma^1(n_1)\left(\frac{S_\theta}{\beta}\right)^m \gamma^2(n_2) \\ &- \left(\frac{S_\theta}{\lambda_0}\right)^m \gamma^1(n_2)\left(\frac{S_\theta}{\beta}\right)^m \gamma^2(n_1),\end{aligned}$$

where $\beta = \frac{\mu_0}{\lambda_0} > |\lambda_1|$. Since $(\frac{S_\theta^X}{\lambda_0})^m \to P_\theta$ with $P_\theta \equiv 0$ on N and β is strictly greater than the spectral radius of the restriction of S_θ^X to N, each term on the right-hand side of the preceding equality tends to zero as $m \to \infty$. But the set A defined in Lemma 4.2 is dense in W. Then, we can conclude from (4.2), after a computation, that $(\frac{S_{2,\theta}}{\mu_0})^m g \to 0$ strongly. But this contradicts Lemma 4.4. The proof of item (a) is completed.

(b) The next step is to show that λ_1 is simple. We write $\psi_1 = \psi_1^P + \psi_1^I$, where ψ_1^P and ψ_1^I denote, respectively, the even and odd parts of ψ_1. We begin by showing that $\psi_1^P \equiv 0$. Since the kernel K of S_θ is symmetric and ω_θ is even, we have S_θ map even sequences into even sequences and odd sequences into odd sequences. Then, we get that $\psi_0 \wedge \psi_1^P$ satisfies $S_{2,\theta}(\psi_0 \wedge \psi_1^P) = \lambda_0\lambda_1(\psi_0 \wedge \psi_1^P)$. Since $\mu_0 > 0$ and simple we obtain $\psi_0 \wedge \psi_1^P \in [\tau_0]$. Therefore, either $\psi_0 \wedge \psi_1^P \equiv 0$ or $\psi_0 \wedge \psi_1^P \neq 0$; that is, $\psi_0 \wedge \psi_1^P$ is one-signed. With this fact, if ψ_1^P does not vanish, then it will have at most one zero. Then, ψ_1^P being an even sequence, either $\psi_1^P \equiv 0$ or else ψ_1^P is one-signed, except possibly in $n = 0$. If the second case holds, we have to consider three cases: $\psi_1^P(0) = 0$, $\psi_1^P(0) > 0$, and $\psi_1^P(0) < 0$. If $\psi_1^P(0) = 0$, then we should have, from the definition of wedge product above, that $0 < (\psi_0 \wedge \psi_1^P)(0, n) = \psi_0(0)\psi_1^P(n)$ for $n > 0$ (where we are assuming that $\psi_0 > 0$ and $\psi_0 \wedge \psi_1^P > 0$). Then, $\psi_1^P(n) > 0$ for all $n \in \mathbb{Z}$ because $\psi_1^P(n)$ is an even sequence. Next, if $\psi_1^P(0) > 0$, then for $n > 0$ we get $0 < (\psi_0 \wedge \psi_1^P)(0, n) = \psi_0(0)\psi_1^P(n) - \psi_0(n)\psi_1^P(0) < \psi_0(0)\psi_1^P(n)$ and therefore $\psi_1^P > 0$. The last case is similar to the second one. These considerations make ψ_1^P a one-signed eigensequence of S_θ (except possibly in $n \neq 0$), and so the inner product $\langle \psi_1^P, \psi_0 \rangle_{X,\theta}$ is also one-signed. But this a contradiction because we have two eigensequences of a self-adjoint operator associated with distinct eigenvalues whose inner product is nonzero. Therefore, $\psi_1^P \equiv 0$, and so ψ_1 is odd. A similar argument shows that ψ_1 can have at most one zero. Of course, this one must be in $n = 0$.

The sequence ψ_1 shown previously was an arbitrary sequence, and this is associated with eigenvalue λ_1. Then, we show that any eigensequence ψ associated with λ_1 must be odd and $\psi(n) = 0 \Leftrightarrow n = 0$. But, two eigensequences of this kind cannot be orthogonal since the product of them is even, and thus λ_1 is simple. This fact completes the proof of the lemma. □

We turn back to the proof of Theorem 4.1. Let us consider $||\psi_i(\theta)||_{X,\theta} = 1$ for $i = 0, 1$. From Lemma 4.5, $\mu_0(\theta) = \lambda_0(\theta)\lambda_1(\theta)$ with μ_0 and λ_0 differentiable with respect to θ, and we have that λ_1 is also differentiable with respect to this parameter. Hence, the associated eigensequence ψ_1 is also differentiable since $S_\theta\psi_1(\theta) = \lambda_1(\theta)\psi_1(\theta)$. Next, we will show that

$$\frac{d}{d\theta}\lambda_i(\theta) < 0, \quad i = 0, 1, \quad \theta \geq 0. \tag{4.3}$$

In fact, writing $\psi_i(\theta)$, $i = 0, 1$, instead of $\psi_{i,\theta}(n)$, we have

$$\frac{d}{d\theta}\lambda_i(\theta) = \frac{d}{d\theta}\sum_{n=-\infty}^{\infty} S_\theta\psi_i(\theta)\psi_i(\theta)\omega_\theta(n).$$

Thus,

$$\begin{aligned}
\frac{d}{d\theta}\lambda_i(\theta) &= 2\sum_{n=-\infty}^{\infty}\frac{d}{d\theta}\psi_i(\theta)S_\theta\psi_i(\theta)\omega_\theta(n) = 2\lambda_i(\theta)\sum_{n=-\infty}^{\infty}\left(\frac{d}{d\theta}\psi_i(\theta)\right)\psi_i(\theta)\omega_\theta(n) \\
&= 2\lambda_i(\theta)\left\{\frac{d}{d\theta}\frac{1}{2}\left(\sum_{n=-\infty}^{\infty}\psi_i(\theta)^2\omega_\theta(n)\right) - \frac{1}{2}\sum_{n=-\infty}^{\infty}\psi_i(\theta)^2\right\} \\
&= -\lambda_i(\theta)\sum_{n=-\infty}^{\infty}\psi_i(\theta)^2 < 0,
\end{aligned}$$

which shows the affirmation. Next, for $\theta \geq 0$,

$$\begin{aligned}\lambda_0(\theta) &= r(S_\theta^X) = ||S_\theta^X||_{B(X,X)} \leq \left(\sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \left\{ \frac{K(n-m)}{\omega_\theta(n)} \right\}^2 \right)^{\frac{1}{2}} \\ &= \left\| K^2 * \frac{1}{\omega_\theta^2} \right\|_{\ell^1(\mathbb{Z})}^{\frac{1}{2}} \leq \|K\|_{\ell^2(\mathbb{Z})} \left\| \frac{1}{\omega_\theta} \right\|_{\ell^2(\mathbb{Z})}.\end{aligned}$$

Since $\frac{1}{\omega_\theta} \to 0$ as $\theta \to +\infty$ and $(\frac{1}{\omega_\theta})^2 \in \ell^1$ with $|\frac{1}{\omega_\theta}|^2 \leq (\frac{B}{1+|n|})^2$, for some $B > 0$, $\|\frac{1}{\omega_\theta}\|_{\ell^2} \to 0$ as $\theta \to +\infty$. Therefore,

$$\lim_{\theta \to +\infty} \lambda_0(\theta) = 0. \tag{4.4}$$

The next step is to show that

$$\lambda_1(0) = 1. \tag{4.5}$$

In fact, since $\frac{d}{dx}\varphi_c$ is an eigenfunction of $\mathcal{L}$ with eigenvalue $\theta = 0$, we can conclude from Corollary 3.1 that $\widehat{\frac{d}{dx}\varphi_c}$ is an eigensequence of S_θ with eigenvalue 1. On the other hand, $\widehat{\frac{d}{dx}\varphi_c}(n) = -in\frac{\pi}{L}\widehat{\varphi}_c(n)$ is odd and vanishes only at $n = 0$. Since ψ_1 is also odd and vanishes at $n = 0$ we must have $\langle \psi_1, \widehat{\frac{d}{dx}\varphi_c} \rangle_{X,\theta} \neq 0$. It follows that ψ_1 and $\widehat{\frac{d}{dx}\varphi_c}$ cannot be eigensequences of S_θ for distinct eigenvalues. Then, ψ_1 and $\widehat{\frac{d}{dx}\varphi_c}$ are associated with the same eigenvalue, and therefore $\lambda_1(0) = 1$. With this fact, since $\lambda_0(0) > \lambda_1(0) = 1$, from (4.3) and (4.4) it follows that there is a unique $\theta_0 \in (0, +\infty)$ such that $\lambda_0(\theta_0) = 1$. Then, from Corollary 3.1, if we consider $\kappa = -\theta_0 < 0$, then $\mathcal{L}$ has a negative eigenvalue which is simple. Now, for $i \geq 2$ and $\theta > 0$ we have from (4.5) that

$$\lambda_i(\theta) \leq \lambda_1(\theta) < \lambda_1(0) = 1.$$

It is straightforward to see that 1 cannot be an eigenvalue of S_θ for all $\theta \in (0, +\infty) \setminus \{\theta_0\}$, since 1 is an eigenvalue only for $\theta = 0$ and $\theta = \theta_0$. Then we obtain (P_1).

Because $\lambda_1(0) = 1$ and λ_1 is a simple eigenvalue it follows that $\theta = 0$ is a simple eigenvalue of $\mathcal{L}$ by the Corollary 3.1. This fact shows (P_2) and as a consequence the theorem. □

Remark 4.2. The Fourier transform in Theorem 4.1 needs to be evaluated in the minimal period of φ_c (see section 6).

5. Stability of periodic travelling-wave solutions. In this section we are interested in applying the theory in section 4 to obtain the stability of specific periodic travelling waves associated with the KdV, mKdV, and BO equations. Our approach to obtain condition (P_0) in (1.10) will be based on the Poisson summation theorem and the implicit function theorem. This new approach, in the periodic context, will give a simple way to calculate condition (P_3) in (1.10). We start with the definition of stability.

DEFINITION 5.1. *Let φ be a periodic travelling-wave solution with period $2L$ of (1.6), and consider $\tau_r\varphi(x) = \varphi(x + r)$, $x \in \mathbb{R}$, and $r \in \mathbb{R}$. We define the set $\Omega_\varphi \subset H_{2L}^{\frac{m_2}{2}}$, the orbit generated by φ, as*

$$\Omega_\varphi = \{g;\ g = \tau_r\varphi \text{ for some } r \in \mathbb{R}\}.$$

And, for any $\eta > 0$, define the set $U_\eta \subset H_{2L}^{\frac{m_2}{2}}$ by

$$U_\eta = \left\{ f;\ \inf_{g\in\Omega_\varphi} ||f-g||_{H_{2L}^{\frac{m_2}{2}}} < \eta \right\}.$$

With this terminology, we say that φ is (orbitally) stable in $H_{2L}^{\frac{m_2}{2}}$ by the flow generated by (1.3) *if the following hold:*

(i) *There is s_0 such that $H_{2L}^{s_0} \subseteq H_{2L}^{\frac{m_2}{2}}$ and the initial value problem associated with* (1.3) *is globally well-posed in $H_{2L}^{s_0}$ (see Theorems* 2.2 *and* 2.3*).*

(ii) *For every $\varepsilon > 0$, there is $\delta > 0$ such that, for all $u_0 \in U_\delta \cap H_{2L}^{s_0}$, the solution u of* (1.3) *with $u(0,x) = u_0(x)$ satisfies $u(t) \in U_\varepsilon$ for all $t > 0$.*

The proof of the following general stability theorem can be shown by using the techniques in Grillakis, Shatah, and Strauss [27] (see also Angulo [6]).

THEOREM 5.1. *Let φ_c be a periodic travelling-wave solution of* (1.6)*, and suppose that part* (i) *of the definition of stability holds. Suppose also that the operator $\mathcal{L}$ defined previously in* (1.7) *has properties* (P_1) *and* (P_2) *in* (1.10)*. Choose $\chi \in L_{2L}^2$ such that $\mathcal{L}\chi = \varphi_c$, and define $I = (\chi, \varphi_c)_{L_{2L}^2}$. If $I < 0$, then φ_c is stable.*

Remark 5.1. In our cases the function χ in Theorem 5.1 satisfies that $\chi = -\frac{d}{dc}\varphi_c$. Then, we need to verify properties (P_0) and (P_3) in (1.10).

5.1. Stability of periodic travelling-wave solutions for the BO equation. This section is concerned with the stability theory of periodic travelling-wave solutions to the BO equation found initially by Benjamin in [13]. Next, we will present an interesting method for obtaining an explicit solution to the BO equation in (3.1), by using the Poisson summation theorem. In fact, consider the following equation:

$$\mathcal{H}\phi_\omega' + \omega\phi_\omega - \frac{1}{2}\phi_\omega^2 = 0.$$

This equation determines solitary travelling-wave solutions to the BO equation on $\mathbb{R}$ in the form

$$\phi_\omega(x) = \frac{4\omega}{1+\omega^2x^2}, \quad \omega > 0.$$

Its Fourier transform is given by

$$\widehat{\phi_\omega}^{\mathbb{R}}(x) = 4\pi e^{\frac{-2\pi}{\omega}|x|}.$$

Then, by the Poisson summation theorem, we obtain that

$$\begin{aligned}
\psi_\omega(x) &\equiv \sum_{n=-\infty}^{+\infty} \phi_\omega(x+2Ln) = \frac{2\pi}{L}\sum_{n=-\infty}^{+\infty} e^{\frac{-\pi|n|}{\omega L}} e^{\frac{\pi i n x}{L}} \\
&= \frac{2\pi}{L}\sum_{n=0}^{+\infty} \varepsilon_n e^{\frac{-\pi n}{\omega L}} \cos\left(\frac{n\pi x}{L}\right) = \frac{2\pi}{L}\mathrm{Re}\left[\coth\left(\frac{\pi}{2\omega L} + \frac{i\pi x}{2L}\right)\right] \qquad (5.1)\\
&= \frac{2\pi}{L}\left(\frac{\operatorname{senh}\left(\frac{\pi}{\omega L}\right)}{\cosh\left(\frac{\pi}{\omega L}\right) - \cos\left(\frac{\pi x}{L}\right)}\right),
\end{aligned}$$

where

$$\varepsilon_n = \begin{cases} 1 & \text{if} \quad n = 0, \\ 2 & \text{if} \quad n = 1,2,3,\ldots. \end{cases}$$

Let φ_c, $c>0$, be a smooth periodic solution of the first equation in (3.1). Thus, φ_c can be expressed as a Fourier series

$$\varphi_c(x)=\sum_{n=-\infty}^{+\infty} a_n e^{\frac{in\pi x}{L}}. \tag{5.2}$$

Substituting the expression above into the BO equation in (3.1), we get

$$\left[\frac{\pi|n|}{L}+c\right]a_n=\frac{1}{2}\sum_{m=-\infty}^{+\infty}a_{n-m}a_m.$$

Next, from (5.1) we consider $a_n\equiv\frac{2\pi}{L}e^{-\gamma|n|}$, $n\in\mathbb{Z}$, $\gamma\in\mathbb{R}$. Substituting a_n into the last identity we have

$$\sum_{m=-\infty}^{+\infty}a_{n-m}a_m=\frac{4\pi^2}{L^2}e^{-\gamma|n|}\left[|n|+1+2\sum_{k=1}^{+\infty}e^{-2\gamma k}\right]=\frac{4\pi^2}{L^2}e^{-\gamma|n|}(|n|+\coth\gamma).$$

Then, we conclude that

$$c+\frac{\pi|n|}{L}=\frac{2\pi}{L}\cdot\frac{1}{2}(|n|+\coth\gamma). \tag{5.3}$$

We denote $\gamma=\frac{\pi}{\omega L}$ and consider $c>\frac{\pi}{L}$. Then, if we choose $\omega=\omega(c)>0$ such that $\tanh(\gamma)=\frac{\pi}{cL}$, we obtain from (5.3) that $\psi_{\omega(c)}=\varphi_c$ (hence, φ_c is given by (5.1)). Therefore, we obtain that φ_c has the form in (1.13) with $\gamma>0$ satisfying $\tanh(\gamma)=\frac{\pi}{cL}$.

Thus, from (5.1) we have that $\varphi_c>0$, and $\gamma:=\gamma(c)=\tanh^{-1}\left(\frac{\pi}{cL}\right)$ being a differentiable function for $c>\frac{\pi}{L}$, it follows that

$$c\in\left(\frac{\pi}{L},+\infty\right)\mapsto\varphi_c\in H^n_{2L}$$

is a smooth curve of periodic travelling-wave solutions for the BO equation for all $n\in\mathbb{N}$. Then, by defining χ in Theorem 5.1 as $\chi=-\frac{d}{dc}\varphi_c$, we obtain from the first equation in (3.1) that $\mathcal{L}\chi=\varphi_c$. Then $I=(\chi,\varphi_c)_{L^2_{2L}}$ becomes

$$I=-\frac{1}{2}\frac{d}{dc}||\varphi_c||^2_{L^2_{2L}}. \tag{5.4}$$

We will show that $I<0$. Indeed, from (5.2) and (5.1) we get

$$\varphi_c(x)=\frac{2\pi}{L}\sum_{n=-\infty}^{+\infty}e^{-\gamma|n|}e^{\frac{\pi inx}{L}}; \tag{5.5}$$

then, from the Parseval theorem we conclude that

(5.6)

$$\begin{aligned} I &= -\frac{1}{2}\frac{d}{dc}||\varphi_c||^2_{L^2_{2L}}=-\frac{1}{2}\frac{d}{dc}||\widehat{\varphi_c}||^2_{\ell^2}\cdot 2L=-\frac{1}{2}\frac{d}{dc}\left(\frac{4\pi^2}{L^2}\sum_{n=-\infty}^{\infty}e^{-2\gamma|n|}\right)\cdot 2L\\ &= -\frac{4\pi^3}{c^2L^3}\left(\frac{1}{1-\left(\frac{\pi}{cL}\right)^2}\right)\left(\sum_{n=-\infty}^{\infty}|n|e^{-2\gamma|n|}\right)\cdot 2L<0. \end{aligned}$$

Finally, we will verify that conditions (P_1) and (P_2) are true for the operator

$$\mathcal{L}_{BO} = \mathcal{H}\partial_x + c - \varphi_c.$$

Let $\widehat{\varphi_c}(n) = \frac{2\pi}{L} e^{-\gamma|n|}$ be the Fourier coefficients of φ_c. In Albert [2] it has already been seen that the function $f(x) = e^{-\gamma|x|}$ belongs to the $PF(2)$ class in the continuous case, and so $\widehat{\varphi_c}$ is in the $PF(2)$ class in the sense of Definition 4.1. Hence, we obtain from Theorem 2.3 the stability of the periodic solutions (1.13) in $H^{\frac{1}{2}}_{2L}$ by the periodic flow of the BO equation.

5.1.1. Stability of the constant solutions. To complete the investigation about periodic travelling-wave solutions for the BO equation, we will study the stability of the constant solutions. Hence, if $\varphi_c(x) \equiv \tau$ is a constant solution to the BO equation, we have $\varphi_c \equiv 2c$ and $\varphi_c \equiv 0$ as solutions. We consider only the first case. The next result will resume our purpose.

PROPOSITION 5.1. *Let $L > 0$ and $c > 0$ be given. Consider $\psi_0 \equiv 2c$ to be a nontrivial constant solution of* (3.1). *Then, ψ_0 is stable in $H^{\frac{1}{2}}_{2L}([-L, L])$, provided $c < \frac{\pi}{L}$.*

Proof. The proof of this proposition follows from standard ideas (see [10], [15]) and from the following nonlocal Poincaré–Wirtinger-type inequality: for $f \in H^{\frac{1}{2}}_{2L}$ such that $\int_{-L}^{L} f(x)dx = 0$, we have $\int_{-L}^{L} [D^{\frac{1}{2}} f(x)]^2 dx \geq \frac{\pi}{L} \int_{-L}^{L} f^2(x)dx$, where $D = \mathcal{H}\partial_x$ (see the appendix for a proof of this inequality). □

5.2. Stability of periodic travelling-wave solutions for the mKdV equation. Next, we will establish the existence of a smooth curve of periodic travelling-wave solutions for the mKdV equation

$$u_t + 3u^2 u_x + u_{xxx} = 0, \tag{5.7}$$

of the form $u(x,t) = \varphi(x - ct) := \varphi_c(\xi)$, where $\xi = x - ct$, $c \in \mathbb{R}$, and is of period L. The equation which determines the periodic travelling-wave solutions is

$$\varphi_c{}'' + \varphi_c^3 - c\varphi_c = 0. \tag{5.8}$$

Next, we obtain an explicit solution for (5.8) using the Poisson summation theorem. It considers for $\omega > 0$ the solitary wave solutions for the mKdV equation on $\mathbb{R}$:

$$\phi_\omega(x) = \sqrt{2\omega}\, \mathrm{sech}(\sqrt{\omega}x), \quad x \in \mathbb{R}.$$

Its Fourier transform is $\widehat{\phi}_\omega(x) = \sqrt{2}\pi \mathrm{sech}(\frac{\pi x}{2\sqrt{\omega}})$, where $\omega > 0$, and it will be chosen later. From the Poisson summation theorem we obtain the following periodic function of period L:

$$\psi_\omega(\xi) = \frac{\sqrt{2}\pi}{L} \sum_{n=0}^{\infty} \varepsilon_n \mathrm{sech}\left(\frac{\pi n}{2\sqrt{\omega}L}\right) \cos\left(\frac{2\pi n \xi}{L}\right), \tag{5.9}$$

where

$$\varepsilon_n = \begin{cases} 1, & n = 0, \\ 2, & n = 1, 2, 3 \ldots. \end{cases}$$

On the other hand, it considers the Fourier expansion of the Jacobi elliptic function *dnoidal*, dn, of period L (see [19], [35], [42]),

$$\frac{2K}{L}\mathrm{dn}\left(\frac{2K\xi}{L};k\right)=\frac{\pi}{L}+\frac{4\pi}{L}\sum_{n=1}^{+\infty}\frac{q^n}{1+q^{2n}}\cos\left(\frac{2n\pi\xi}{L}\right),$$

where $K=K(k)$ is the complete elliptic integral of the first kind and $q=e^{(-\frac{\pi K'}{K})}$, which is called the "nome." Here, $K'(k)=K(\sqrt{1-k^2})$. We can conclude that

$$\frac{q^n}{1+q^{2n}}=\frac{1}{2}\mathrm{sech}\left(\frac{n\pi K'}{K}\right).$$

Therefore,

$$\frac{2K}{L}\mathrm{dn}\left(\frac{2K\xi}{L};k\right)=\frac{\pi}{L}+\frac{2\pi}{L}\sum_{n=1}^{+\infty}\mathrm{sech}\left(\frac{n\pi K'}{K}\right)\cos\left(\frac{2n\pi\xi}{L}\right).$$

Because of the shape of the series that determines ψ_ω given above (see [35]), let $\varphi_c(\xi)=\eta\mathrm{dn}(\frac{\eta\xi}{\sqrt{2}};k)$ be a periodic solution of period L for (5.8), with $\eta>0$ and $k\in(0,1)$ fixed. Then, the following identities should be satisfied:

$$c=\frac{\eta^2}{2}(1+k'^2)\ \text{ and }\ \eta=\frac{2\sqrt{2}K(k)}{L} \tag{5.10}$$

with $k'^2=1-k^2$. Thus, for $k\in(0,1)$ we should have that $\eta\in(\sqrt{c},\sqrt{2c})$ and from the asymptotic properties of K that $c>\frac{2\pi^2}{L^2}$. Next, for $k\in(0,1)$ fixed, η is immediately defined from (5.10), and so

$$c=\frac{4K^2(k)}{L^2}(2-k^2)>\frac{2\pi^2}{L^2} \tag{5.11}$$

since $k\to K^2(k)(2-k^2)$ is a strictly increasing function. Therefore, with $k\in(0,1)$ and c defined in (5.11) we define $\omega=\omega(c)$ as

$$\omega=\frac{c}{16(2-k^2)K'^2(k)}.$$

Therefore, from (5.9) it follows that $\psi_{\omega(c)}=\varphi_c$ is a solution of (5.8).

Next, it is necessary to build a smooth curve, $c\to\varphi_c$, of dnoidal wave solutions. Initially, we obtain the following a priori estimate for the fundamental period of φ_c. Namely,

$$T_{\varphi_c}(\eta)=\frac{2K(k(\eta))}{\sqrt{c}}\sqrt{2-k^2}>\frac{\sqrt{2}\pi}{\sqrt{c}}, \tag{5.12}$$

where $k^2(\eta)=2-\frac{2c}{\eta^2}$. In fact, for $\eta\to\sqrt{c}$ it follows that $k\to0^+$ and thus $T_{\varphi_c}(\eta)\to\frac{\pi\sqrt{2}}{\sqrt{c}}$. For $\eta\to\sqrt{2c}$ we have $k\to1^-$ and therefore $T_{\varphi_c}(\eta)\to+\infty$. But, $\eta\mapsto T_{\varphi_c}(\eta)$ being a strictly increasing function (see Theorem 2.1 in Angulo [9]) we get (5.12).

Remark 5.2. The function φ_c obtained previously, that is,

$$\varphi_c(\xi)=\eta\mathrm{dn}\left(\frac{\eta}{\sqrt{2}}\xi;k\right), \tag{5.13}$$

is a positive function and has been built by the periodization of the solitary wave solution associated with (5.7). Thus, it is natural to ask if we can again obtain this solitary wave. Indeed, this fact can be determined by (5.13), since for $\eta \to \sqrt{2c}$ we have $k \to 1^-$ and then $\mathrm{dn}(u;1^-) = \mathrm{sech}(u)$. Hence we have formally that $\varphi_c(\xi) = \sqrt{2c}\ \mathrm{sech}(\sqrt{c}\xi)$. The other limit case, that is, $\eta \to \sqrt{c}$, we have $k \to 0^+$ and so $\mathrm{dn}(u;0^+) = 1$; then we get $\varphi_c(\xi) = \sqrt{c}$, the nontrivial constant solutions for the mKdV.

Now, we construct a family of dnoidal waves solutions with period L. Let $c > 0$ such that $\sqrt{c} > \frac{\pi\sqrt{2}}{L}$. Since $\eta \in (\sqrt{c}, \sqrt{2c}) \mapsto T_{\varphi_c}(\eta)$ is a strictly increasing mapping, it follows from (5.12) and from Theorem 2.1 in Angulo [9] that there is a unique $\eta \equiv \eta(c) \in (\sqrt{c}, \sqrt{2c})$ such that the fundamental period of the dnoidal wave φ_c will be $T_{\varphi_c}(\eta(c)) = L$. Moreover, we have $c \in \left(2\pi^2/L^2, +\infty\right) \to \varphi_c \in H^n_{per}([0,L])$ is a smooth curve, $c \in \left(2\pi^2/L^2, +\infty\right) \to \eta = \eta(c)$ is a strictly increasing function, and its derivative with respect to the velocity c is given by

$$\frac{d\eta}{dc} = \frac{\eta}{2c} + \frac{k^2k'^2\eta^3(2-k^2)K}{\sqrt{c^3}((2-k^2)E - 2(1-k^2)K)}.$$

The next result gives us a relation between the velocity of the solitary and periodic waves associated with the mKdV, and it will be useful later.

THEOREM 5.2. *We consider the mapping* $\omega : \left(2\pi^2/L^2, +\infty\right) \mapsto \mathbb{R}$, *given by*

$$\omega(c) = c/(16(2-k^2)K'^2(k)); \tag{5.14}$$

then $\frac{d\omega}{dc} > 0$.

Proof. Indeed, $\frac{d\omega}{dc} = [4(2-k^2)K'^2 + 8K'c\frac{dk}{dc}(kK' - (2-k^2)\frac{dK'}{dk})]/[16(2-k^2)^2K'^4]$. Since $\frac{dK'}{dk} = -(\frac{E'-k^2K'}{kk'^2}) < 0$, it suffices to show that $\frac{dk}{dc} > 0$. So, since

$$\frac{dk}{dc} = \frac{2}{\eta^3}\left(2c\frac{d\eta}{dc} - \eta\right),$$

we need to verify the sign of the expression $2c\frac{d\eta}{dc} - \eta$. Next, we calculate the exact value of $\frac{d\eta}{dc}$. In fact,

$$\frac{d\eta}{dc} = \frac{\eta}{2c} + \underbrace{\frac{k^2k'^2\eta^3(2-k^2)K}{\sqrt{c^3}((2-k^2)E - 2(1-k^2)K)}}_{A}.$$

Thus, $2c\frac{d\eta}{dc} - \eta = 2cA$ and therefore $\frac{dk}{dc} > 0$. This fact completes the proof of the theorem. □

Next, we will show that $\widehat{\varphi_c} > 0$ and $K = \widehat{\varphi_c^2}$ belongs to $PF(2)$ discrete. It is easy to see that $\widehat{\varphi_c} > 0$ because of the form of the Fourier coefficients of φ_c given by (5.9). Moreover, $\widehat{\varphi_c} \in PF(2)$ discrete because the function $f(x) = \mu\mathrm{sech}(\nu x)$ belongs to $PF(2)$ continuous (see [2]). So, since the convolution of even sequences in $PF(2)$ discrete is a sequence in $PF(2)$ discrete (see [32]) we can conclude that $K = \widehat{\varphi_c^2} \in PF(2)$ discrete. Now, by choosing $\chi = -\frac{d}{dc}\varphi_c$ we have that $\mathcal{L}\chi = \varphi_c$. Then, by the Parseval theorem, $I = -\frac{1}{2}\frac{d}{dc}\|\varphi_c\|^2_{L^2_{per}} = -\frac{L}{2}\frac{d}{dc}\|\widehat{\varphi_c}\|^2_{\ell^2}$. Since $\|\widehat{\varphi_c}\|^2_{\ell^2} =$

$2\frac{\pi^2}{L^2}\sum_{n=-\infty}^{+\infty}\operatorname{sech}^2(\frac{\pi n}{\sqrt{\omega(c)}L})$, we have that

$$\frac{d}{dc}\|\widehat{\varphi_c}\|_{\ell^2}^2 = \frac{C_1(L)}{\sqrt{\omega(c)^3}}\frac{d\omega}{dc}\sum_{n=-\infty}^{+\infty}\operatorname{sech}^2\left(\frac{\pi n}{\sqrt{\omega(c)}L}\right)n\operatorname{tgh}\left(\frac{\pi n}{\sqrt{\omega(c)}L}\right).$$

Since $(n\ \operatorname{tgh}(\frac{\pi n}{\sqrt{\omega(c)}L}))_{n\in\mathbb{Z}}$ is a positive sequence we have from Theorem 5.2 that $\frac{d}{dc}\|\widehat{\varphi_c}\|_{\ell^2}^2 > 0$. Thus, we obtain that the dnoidal wave φ_c in (5.14) is stable in $H^1_{per}([0,L])$ by the flow of the mKdV.

5.3. Stability of periodic travelling-wave solutions for the KdV equation. Now, we apply the results obtained previously to the proof of the stability of periodic travelling-wave solutions of cnoidal type associated with the KdV equation and satisfying

$$\varphi_c'' + \frac{1}{2}\varphi_c^2 - c\varphi_c = 0. \tag{5.15}$$

We consider the solitary wave solutions $\phi_\omega(x) = 3\omega\operatorname{sech}^2(\frac{\sqrt{\omega}x}{2})$, whose Fourier transform is given by $\widehat{\phi}_\omega(x) = \frac{12\pi x}{\operatorname{senh}\left(\frac{\pi x}{\sqrt{\omega}}\right)}$; then from the Poisson summation theorem we consider

$$\psi_\omega(\xi) = \frac{12\sqrt{\omega}}{L} + \frac{12\pi}{L^2}\sum_{n\neq 0} n\operatorname{csch}\left(\frac{\pi n}{\sqrt{\omega}L}\right)e^{\frac{2\pi i n\xi}{L}}. \tag{5.16}$$

Since ω is arbitrary, consider $\omega := \omega(k)$ such that $\sqrt{\omega(k)} = \frac{K(k)}{K(k')L}$, $k\in(0,1)$, $k'^2 = 1-k^2$. Then, we obtain

$$\psi_{\omega(k)}(\xi) = \frac{12\sqrt{\omega}}{L} + \frac{24\pi}{L^2}\sum_{n=1}^{+\infty} n\operatorname{csch}\left(\frac{\pi nK'}{K}\right)\cos\left(\frac{2\pi n\xi}{L}\right). \tag{5.17}$$

Now, we invoke the Fourier expansion of dn^2 (see [19], [42]), that is,

$$K^2\left(\operatorname{dn}^2\left(\frac{2K\xi}{L};k\right) - \frac{E}{K}\right) = 2\pi\sum_{n=1}^{+\infty}\frac{nq^n}{1-q^{2n}}\cos\left(\frac{2\pi n\xi}{L}\right),$$

where $q = e^{-(\frac{\pi K'}{K})}$. We can conclude that

$$\frac{q^n}{1-q^{2n}} = \frac{1}{2}\operatorname{csch}\left(\frac{n\pi K'}{K}\right).$$

Then, we get from (5.17)

$$\psi_{\omega(k)}(\xi) = \frac{12\sqrt{\omega(k)}}{L} + \frac{24K^2}{L^2}\left(\operatorname{dn}^2\left(\frac{2K\xi}{L};k\right) - \frac{E}{K}\right) \tag{5.18}$$

for $k\in(0,1)$.

Next, because of the equality (5.18), we consider $\varphi_c(\xi) = a + b\left(\mathrm{dn}^2(dx;k) - \frac{E}{K}\right)$ a periodic travelling-wave solution for (5.15) of period L. Then, the following nonlinear system is obtained:

$$\left\{\begin{array}{l} \dfrac{b^2}{2} - 6d^2 b = 0, \quad 4bd^2(1+k'^2) + ab - b^2\dfrac{E}{K} - cb = 0, \\ \dfrac{a^2}{2} - \dfrac{abE}{K} + \dfrac{b^2}{2}\left(\dfrac{E}{K}\right)^2 - ac - cb\dfrac{E}{K} - 2bd^2k'^2 = 0. \end{array}\right. \tag{5.19}$$

Since φ_c is periodic of period L it follows that $d = \frac{2K(k)}{L}$. Then, from the first equation of the system above we have that $b = \frac{48K^2}{L^2}$. Substituting those values at the second equation we get

$$c = \frac{16K}{L^2}\left[(1+k'^2)K - 3E\right] + a. \tag{5.20}$$

From the third equation in (5.19) and the value of c in (5.20) we have the quadratic equation in terms of a,

(5.21)

$$a^2 + \frac{32K}{L^2}\left[(1+k'^2)K - 3E\right]a - \frac{(1+k'^2)1536K^3E}{L^4} + \frac{768K^4k'^2}{L^4} + \frac{2304K^2E^2}{L^4} = 0,$$

whose positive solution is

$$a = -\frac{16K}{L^2}\left[(1+k'^2)K - 3E\right] + \frac{16K^2}{L^2}\sqrt{1-k^2+k^4}.$$

Thus, the value of c is $c = \frac{16K^2}{L^2}\sqrt{1-k^2+k^4}$. Hence, for $k \in (0,1)$ we have that $c \in (\frac{4\pi^2}{L^2}, +\infty)$. Therefore, writing φ_c in a convenient form, in terms of cn^2, we obtain

$$\varphi_c(\xi) = \frac{16K^2}{L^2}\left[\sqrt{1-k^2+k^4} + 1 - 2k^2\right] + \frac{48K^2k^2}{L^2}\mathrm{cn}^2\left(\frac{2K}{L}\xi;k\right).$$

We can see that this formula is the same that was obtained by Angulo in [7] and it can be rewritten as

$$\varphi_c(\xi) = \beta_2 + (\beta_3 - \beta_2)\mathrm{cn}^2\left(\sqrt{\frac{\beta_3-\beta_1}{12}}\xi;k\right), \tag{5.22}$$

where

$$\beta_2 = \frac{16K^2}{L^2}\left[\sqrt{1-k^2+k^4} + 1 - 2k^2\right], \quad \beta_3 = \frac{16K^2}{L^2}\left[\sqrt{1-k^2+k^4} + 1 + k^2\right],$$

and β_1 is such that

$$\beta_3 - \beta_1 = \frac{48K^2}{L^2}.$$

By making a similar analysis such as in the case of the mKdV equation, we can obtain a smooth curve of positive cnoidal waves with the form in (5.22), $c \in$

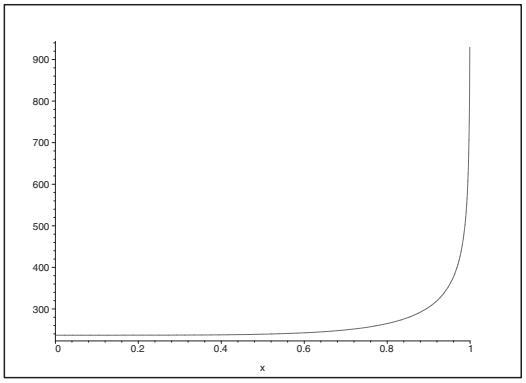


FIG. 1. *Graphic of the function $a(k)$ with period $L = 1$.*

$(\frac{4\pi^2}{L^2}, +\infty) \mapsto \varphi_c \in H^n_{per}([0, L])$, such that $k := k(c)$ is a strictly increasing smooth function (see [7]) for all $n \in \mathbb{N}$. Moreover, we can determine that for $k \in (0, 1)$ there is a unique $c \in (\frac{4\pi^2}{L^2}, +\infty)$ such that $k(c) = k$. Therefore, the function $\omega(k)$ defined above can be expressed as a function of c, $\omega = \omega(k(c))$, and it is a strictly increasing function (it will be seen later). Then, since $\frac{K(k)}{K(k')} \in (0, +\infty)$ it follows that for $c \in (\frac{4\pi^2}{L^2}, +\infty)$ we obtain $\omega(k(c)) \in (0, +\infty)$. Therefore, the mapping $c \in (\frac{4\pi^2}{L^2}, +\infty) \mapsto \psi_{\omega(k(c))} \in H^n_{per}([0, L])$ is a smooth curve for all $n \in \mathbb{N}$. Next, we note that $2\psi_{\omega(k(c))}(\xi) - \varphi_c(\xi) = \frac{24\sqrt{\omega(k(c))}}{L} - a(k(c))$, where for $k = k(c)$

$$a(k) = \frac{16K^2}{L^2}\left[\sqrt{1 - k^2 + k^4} + 2 - k^2 + 3\frac{E}{K}\right].$$

Thus, for $s(k(c)) \equiv a(k(c)) - \frac{24\sqrt{\omega(k(c))}}{L}$, we can write $\varphi_c(\xi) \equiv s(k(c)) + \psi_{\omega(k(c))}(\xi)$. Therefore, we obtain immediately that the Fourier coefficients of φ_c are

$$\widehat{\varphi_c}(n) = \begin{cases} a(k), & n = 0, \\ \dfrac{12\pi}{L^2} n \operatorname{csch}\left(\dfrac{\pi n}{\sqrt{\omega(k)}L}\right), & n \neq 0. \end{cases}$$

Remark 5.3. After some calculations, we can obtain that $s(k)$ is a positive function defined in $(0, 1)$. Making use of Maple, we can also determine that $s(k)$ does not have any root on the extremes of the interval $(0, 1)$. We can also determine that the function $a(k)$ is a positive strictly increasing function (see Figure 1).

Since $s(k) > 0$ and the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = \frac{12\pi}{L^2} x \operatorname{csch}(\frac{\pi x}{\sqrt{\omega}L})$, belongs to $PF(2)$ in the continuous case, we can use Lemma 4.1 for obtaining that $\widehat{\varphi_c}$ belongs to the $PF(2)$ discrete case. Indeed, since

$$a(k) > \frac{24\sqrt{\omega(k)}}{L} > \frac{12\sqrt{\omega(k)}}{L} = f(0) > f(x), \quad x \neq 0,$$

we can redefine f by a smooth function $h : \mathbb{R} \to \mathbb{R}$ such that $h(0) = a(k)$, $h(x) \equiv f(x)$ on $(-\infty, -1] \cup [1, +\infty)$ and on the interval $(-1, 1)$ we "complete" f in a differentiable

way, such that h belongs to $PF(2)$ continuous. Therefore, the sequence to be obtained (if we look only at the set of integers numbers) will be $h(n) = \widehat{\varphi_c}(n)$.

Next, let $\chi = -\frac{d}{dc}\varphi_c$ such that $\mathcal{L}\chi = \varphi_c$. Then by the Parseval theorem, it follows that $I = -\frac{L}{2}\frac{d}{dc}\left(\|\widehat{\varphi_c}\|^2_{\ell^2}\right)$. Hence,

$$\begin{aligned}\frac{d}{dc}\|\widehat{\varphi_c}\|^2_{L^2_{per}} &= C_1 a(k)\frac{da}{dk}\frac{dk}{dc} \\ &+ \frac{C_2}{\sqrt{\omega(k)^3}}\frac{d\omega}{dk}\frac{dk}{dc}\sum_{\substack{n=-\infty\\ n\neq 0}}^{+\infty}\underbrace{n^3\operatorname{csch}^2\left(\frac{\pi n}{L\sqrt{\omega(k)}}\right)\coth\left(\frac{\pi n}{L\sqrt{\omega(k)}}\right)}_{b_n},\end{aligned}$$

where $C_1 := C_1(L)$, $C_2 := C_2(L) > 0$. Next, we need to show only that the quantities $\frac{da}{dk}$ and $\frac{d\omega}{dk}$ are positive because $k := k(c)$ is a strictly increasing function and $(b_n)_{n\in\mathbb{Z}}$ is obviously a positive sequence. Hence, we have

$$\frac{d\omega}{dk} = 2\frac{K\left(\frac{dK}{dk}K' - K\frac{dK'}{dk}\right)}{K'^3}.$$

Since $\frac{dK}{dk} > 0$ and $\frac{dK'}{dk} < 0$ we get that $\frac{d\omega}{dk} > 0$. By making use of a similar argument, we can also show that $\frac{da}{dk} > 0$ because we have that

$$a(k) = \frac{16K(k)^2}{L^2}\left[\sqrt{1-k^2+k^4}+2-k^2\right] + 48\frac{E(k)K(k)}{L^2}.$$

Therefore, $I < 0$ and the positive cnoidal waves φ_c are stable in $H^1_{per}([0,L])$ by the periodic flow of the KdV equation.

6. Comments. In this section we make some basic remarks about the results contained in the body of this paper.

6.1. General perturbations. In contrast to the case of solitary waves for which the natural class of disturbance in the stability problem is that of localized disturbances, for periodic waves there are several classes of disturbance for which stability needs to be addressed. Here we consider periodic perturbations with the same fundamental period of the periodic travelling waves. For disturbance, for example, with a double period, stability results remain open in the context of KdV-type equations. If we consider the KdV equation, our stability result in $H^1_{per}([0,L])$ was based on the spectral structure of the operator $\mathcal{L}_{\rm cn} = -\frac{d^2}{dx^2} + c - \varphi_c$, with φ_c defined in (5.22). Now, if we consider this operator with domain $H^2_{per}([0,2L])$, then the number of negative eigenvalues will be exactly 3 (see [10]). Moreover, since the function $\frac{d}{dc}\int_{-L}^{L}\varphi_c^2(x)dx$ is even positive, the abstract setting of Grillakis, Shatah, and Strauss in [27] and [28] cannot be applied. We note that in the case of the focusing Schrödinger equation

$$iu_t + u_{xx} + |u|^2 u = 0,$$

Angulo in [9] showed an instability result for dnoidal waves solutions when the class of periodic disturbance is two times the minimal period of the periodic travelling wave in question (see also Gallay and Hărăguş [24], [25] for recent new results of stability for periodic travelling waves of Schrödinger equations).

6.2. Property $PF(2)$ discrete. We consider the BO equation (but we can do an analogous analysis with the other two equations); then we have seen throughout this paper that the spectral properties of the operator $\mathcal{L} = \mathcal{H}\partial_x + c - \varphi_c$ were obtained from the fact that the Fourier coefficients of the periodic travelling-wave solution given by (1.13) are in $PF(2)$. The Poisson summation theorem was used to find such a wave with a minimal period $2L$. Moreover, the Fourier coefficients were calculated with this period. If we double the period, that is, if we consider $\mathcal{L}$ with domain $D(\mathcal{L}) = H^1_{4L}$, *the property* $PF(2)$ *is not satisfied* to $\widetilde{K} = \widehat{\varphi_c}^{(4L)}$, where $\widehat{\varphi_c}^{(4L)}$ denotes the periodic Fourier transform of φ_c but with period $4L$. Indeed, we consider the Fourier expansion in the form $\varphi_c(x) = \sum_{n=0}^{+\infty} \widehat{\varphi_c}^{(4L)}(n)\cos\left(\frac{n\pi x}{2L}\right)$. Since $\varphi_c(0) = \varphi_c(2L)$, $\sum_{n=0}^{+\infty} \widehat{\varphi_c}^{(4L)}(n)(\cos(n\pi) - 1) = 0$. Thus, $-2\sum_{\substack{n=2k+1\\ k\in\mathbb{N}}} \widehat{\varphi_c}^{(4L)}(n) = 0$. Therefore, $\widetilde{K} = \widehat{\varphi_c}^{(4L)}$ cannot belong to $PF(2)$. Then, we cannot affirm anything about the stability of the wave φ_c when this case is considered. So, Theorem 4.1 cannot be applied here.

In Theorem 4.1, the Fourier transform needs to be evaluated in the minimal period for the solution φ_c. In fact, let L be this minimal period, and we evaluate the Fourier transform of φ_c as being of period $2L$; then

$$\begin{aligned}\widehat{\varphi_c}(k) &= \frac{1}{2L}\int_{-L}^{L} \varphi_c(x)e^{-\frac{ik\pi x}{L}}dx = \frac{1}{2L}\int_{-L}^{L} \varphi_c(x+L)e^{-\frac{ik\pi x}{L}}dx \\ &= \frac{1}{2L}\int_{0}^{2L} \varphi_c(y)e^{-\frac{ik\pi y}{L}}e^{i\pi k}dx = (-1)^k\widehat{\varphi_c}(k).\end{aligned}$$

Then, for k being odd we have $\widehat{\varphi_c}(k) = 0$. Thus, we cannot apply our theory if the minimal period is not fixed.

6.3. Positivity of the periodic travelling waves. Another fact that deserves a special mention is about the condition that the solution φ_c in Theorem 4.1 needs to be a positive solution. Indeed, the classical Fourier theorem (see [30]) told us that this is necessary. Suppose without loss of generality that $\varphi_c(0) = 0$. Then, φ_c being smooth it follows that $\varphi_c^p(0) = \sum_{n=-\infty}^{+\infty} \widehat{\varphi_c^p}(n) = 0$. In other words some Fourier coefficient of φ_c^p must be negative.

6.4. Stability and instability of periodic travelling waves for the critical KdV and NLS. In a forthcoming paper [11] we apply the theory in section 4 to obtain the stability/instability of a special family of periodic travelling-wave solutions for the critical KdV equation

$$u_t + 5u^4u_x + u_{xxx} = 0$$

and for the critical nonlinear Schrödinger (NLS) equation

$$iu_t + u_{xx} + |u|^4u = 0.$$

7. Appendix.

7.1. Jacobi elliptic functions. We establish some basic properties of Jacobian elliptic integrals (see [18] and [19]). The normal elliptic integral of the first kind is

$$\int_0^y \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}} = \int_0^\varphi \frac{d\theta}{\sqrt{1-k^2\sin^2\theta}} \equiv F(\varphi, k),$$

where $y=\sin\varphi$, whereas the normal elliptic integral of the second kind is

$$\int_0^y \sqrt{\frac{1-k^2t^2}{1-t^2}}dt=\int_0^{\varphi}\sqrt{1-k^2\sin^2\theta}d\theta\equiv E(\varphi,k).$$

The number k is called the modulus and belongs to the interval $(0,1)$. The number $k'=\sqrt{1-k^2}$ is called the complementary modulus. The parameter φ is called the argument of the normal elliptic integrals. It is usually understood that $0\leq y\leq 1$ or $0\leq\varphi\leq\frac{\pi}{2}$. For $y=1$, the integrals above are said to be complete. In this case, one writes

$$\int_0^1\frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}}=\int_0^{\frac{\pi}{2}}\frac{d\theta}{\sqrt{1-k^2\sin^2\theta}}\equiv F\left(\frac{\pi}{2},k\right)\equiv K(k)\equiv K$$

and

$$\int_0^1\sqrt{\frac{1-k^2t^2}{1-t^2}}dt=\int_0^{\frac{\pi}{2}}\sqrt{1-k^2\sin^2\theta}d\theta\equiv E\left(\frac{\pi}{2},k\right)\equiv E(k)\equiv E.$$

Clearly, we have $K(0)=E(0)=\frac{\pi}{2}$, while $E(1)=1$ and $K(1)=+\infty$. For $k\in(0,1)$, $\frac{dK}{dk}>0$, $\frac{d^2K}{dk^2}>0$, $\frac{dE}{dk}<0$, $\frac{d^2E}{dk^2}<0$, and $E(k)<K(k)$. Moreover, $E(k)+K(k)$ and $E(k)K(k)$ are strictly increasing functions for every $k\in(0,1)$. Next, we have some derivatives of the complete elliptical integrals K and E used in this work,

$$\frac{dK}{dk}=\frac{E-k'^2K}{kk'^2},\quad \frac{dE}{dk}=\frac{E-K}{k}.$$

The Jacobian elliptic functions are usually defined as follows. It considers the elliptic integral

$$u(y_1;k)\equiv u=\int_0^{y_1}\frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}}=\int_0^{\varphi}\frac{d\theta}{\sqrt{1-k^2\sin^2\theta}}\equiv F(\varphi,k),$$

which is a strictly increasing function of the variable y_1. Its inverse function is written $y_1=\sin\varphi\equiv\mathrm{sn}(u;k)$, or briefly $y_1=\mathrm{sn}u$ when it is not necessary to emphasize the modulus k. The other two basic elliptic functions, the cnoidal and dnoidal functions, are defined in terms of sn by

$$\mathrm{cn}(u;k)=\sqrt{1-y_1^2}=\sqrt{1-\mathrm{sn}^2(u;k)},\quad \mathrm{dn}(u;k)=\sqrt{1-k^2y_1^2}=\sqrt{1-k^2\mathrm{sn}^2(u;k)}.$$

Note that these functions are normalized by the requirements $\mathrm{sn}(0;k)=0$, $\mathrm{cn}(0;k)=1$, and $\mathrm{dn}(0;k)=1$. The functions $\mathrm{cn}(\cdot;k)$ and $\mathrm{dn}(\cdot;k)$ are even functions. These functions are periodic with $\mathrm{sn}(u+4K(k);k)=\mathrm{sn}(u;k)$, $\mathrm{cn}(u+4K(k));k)=\mathrm{cn}(u;k)$, $\mathrm{dn}(u+2K(k);k)=\mathrm{dn}(u;k)$. Moreover, we have the relations $\mathrm{sn}^2u+\mathrm{cn}^2u=1$, $k^2\mathrm{sn}^2u+\mathrm{dn}^2u=1$, $k'^2\mathrm{sn}^2u+\mathrm{cn}^2u=\mathrm{dn}^2u$, $\mathrm{sn}(u+2K;k)=-\mathrm{sn}(u;k)$, $\mathrm{cn}(u+4K;k)=-\mathrm{cn}(u;k)$. We also have the following explicit values: $\mathrm{sn}(0)=0$, $\mathrm{cn}(0)=1$, $\mathrm{sn}(K)=0$, $\mathrm{cn}(K)=0$ and the asymptotic behaviors $\mathrm{sn}(u;0)=\sin u$, $\mathrm{cn}(u;0)=\cos u$, $\mathrm{sn}(u;1)=\tanh u$, $\mathrm{cn}(u;1)=\mathrm{sech}u$. Finally, the formulas

$$\frac{\partial}{\partial u}\mathrm{sn}u=\mathrm{cn}u\mathrm{dn}u,\ \frac{\partial}{\partial u}\mathrm{cn}u=-\mathrm{sn}u\mathrm{dn}u,\ \frac{\partial}{\partial u}\mathrm{dn}u=-k^2\mathrm{cn}u\mathrm{sn}u$$

are straightforwardly deduced from the foregoing material.

7.2. A nonlocal Poincaré–Wirtinger inequality. The following inequality was used in Proposition 5.1. Suppose $f \in H^{\frac{1}{2}}_{2L}$ such that $\int_{-L}^{L} f(x)dx = 0$; then for $D = \mathcal{H}\partial_x$

$$\int_{-L}^{L} [D^{\frac{1}{2}} f]^2 dx \geq \frac{\pi}{L} \int_{-L}^{L} f^2 dx.$$

Proof. Let f be in $\mathcal{P}$, and consider the Fourier expansion of f given by

$$f(x) = \sum_{n=-\infty}^{+\infty} a_n e^{\frac{in\pi x}{L}} = \sum_{n=-\infty}^{+\infty} \widehat{f}(n) e^{\frac{in\pi x}{L}},$$

where $\widehat{f}(n) = \frac{1}{2L}\int_{-L}^{L} f(x)e^{-\frac{in\pi x}{L}}dx$, with $\widehat{f}(0) = 0$. Then, $\mathcal{H}\partial_x f(x) = \frac{\pi}{L}\sum_{n=-\infty}^{+\infty} a_n|n|e^{\frac{in\pi x}{L}}$. From the Parseval theorem we obtain

$$\int_{-L}^{L} f\mathcal{H}\partial_x f dx = 2L \sum_{n=-\infty}^{+\infty} \frac{\pi}{L}|a_n|^2|n| = 2L \sum_{n\neq 0} \frac{\pi}{L}|a_n|^2|n|$$

and

$$\int_{-L}^{L} f^2 dx = 2L \sum_{n\neq 0} |a_n|^2.$$

Therefore,

$$\int_{-L}^{L} f\mathcal{H}\partial_x f dx = 2L \sum_{n\neq 0} \frac{\pi}{L}|a_n|^2|n| \geq \frac{\pi}{L} \int_{-L}^{L} f^2 dx.$$

So, using density arguments, we can show that the inequality above occurs for $f \in H^{\frac{1}{2}}_{2L}$. This completes the proof of the inequality. ∎

Acknowledgments. The second author would like to express his thanks to the Institute of Mathematics and Statistic (IME) of the University of São Paulo/SP-Brazil for its hospitality. The authors also thank the referees for some helpful comments.

REFERENCES

[1] L. Abdelouhab, J. Bona, M. Felland, and J. C. Saut, *Nonlocal models for nonlinear, dispersive wave*, Phys. D, 40 (1989), pp. 360–392.
[2] J. P. Albert, *Positivity properties and stability of solitary-wave solutions of model equations for long waves*, Comm. Partial Differential Equations, 17 (1992), pp. 1–22.
[3] J. P. Albert and J. L. Bona, *Total positivity and the stability of internal waves in fluids of finite depth*, IMA J. Appl. Math., 46 (1991), pp. 1–19.
[4] J. P. Albert, J. L. Bona, and D. Henry, *Sufficient conditions for stability of solitary-wave equation of model equations for long waves*, Phys. D, 24 (1987), pp. 343–366.
[5] B. Alvarez-Samaniego and D. Lannes, *Large time existence for 3D water-waves and asymptotics*, Invent. Math., 171 (2008), pp. 485–541.
[6] J. Angulo, *Existence and Stability of Solitary Wave Solutions to Non-linear Dispersive Evolution Equations*, Publicações do 24º Colóquio Brasileiro de Matemática, IMPA, Rio de Janeiro, Brazil, 2003.
[7] J. Angulo, *Stability of cnoidal waves to Hirota-Satsuma systems*, Mat. Contemp., 27 (2004), pp. 189–223.

[8] J. Angulo, *Stability of dnoidal waves to Hirota-Satsuma system*, Differential Integral Equations, 18 (2005), pp. 611–645.
[9] J. Angulo, *Non-linear stability of periodic travelling-wave equation for the Schrödinger and modified Korteweg-de Vries equation*, J. Differential Equations, 235 (2007), pp. 1–30.
[10] J. Angulo, J. L. Bona, and M. Scialom, *Stability of cnoidal waves*, Adv. Differential Equations, 11 (2006), pp. 1321–1374.
[11] J. Angulo and F. Natali, *Stability and Instability of Periodic Travelling Waves Wave Solutions for the Critical Korteweg-de Vries and Non-linear Schrödinger Equations*, preprint, 2007.
[12] T. B. Benjamin, *Instability of periodic wavetrain in nonlinear dispersive systems*, Proc. Roy. Soc. London Ser. A, 299 (1967), pp. 59–75.
[13] T. B. Benjamin, *Internal waves of permanent form in fluids of great depth*, J. Fluid Mech., 29 (1967), pp. 559–592.
[14] T. B. Benjamin, *The stability of solitary waves*, Proc. Roy. Soc. London Ser. A, 338 (1972), pp. 153–183.
[15] T. B. Benjamin, *Lectures on linear wave motion*, in Nonlinear Wave Motion, A. C. Newell, ed., AMS, Providence, RI, 1974, pp. 3–47.
[16] T. B. Benjamin, *Solitary and periodic waves of a new kind*, Philos. Trans. Roy. Soc. London Ser. A, 354 (1996), pp. 1775–1806.
[17] J. L. Bona, *On the stability theory of solitary waves*, Proc Roy. Soc. London Ser. A, 344 (1975), pp. 363–374.
[18] F. Bowman, *Introduction to Elliptic Functions with Applications*, Dover, New York, 1961.
[19] P. F. Byrd and M. D. Friedman, *Handbook of Elliptic Integrals for Engineers and Scientists*, 2nd ed., Springer, New York, 1971.
[20] J. Colliander, M. Keel, G. Stafilani, H. Takaoka, and T. Tao, *Sharp global well-posedness for the KDV and modified KDV on* $\mathbb{R}$ *and* $\mathbb{T}$, J. Amer. Math. Soc., 16 (2003), pp. 705–749.
[21] A. Constantin, *The trajectories of particles in Stokes waves*, Invent. Math., 166 (2006), pp. 523–535.
[22] A. Constantin and J. Escher, *Particle trajectories in solitary water waves*, Bull. Amer. Math. Soc. (N.S.), 44 (2007), pp. 423–431.
[23] A. Constantin and W. Strauss, *Stability properties of steady water waves with vorticity*, Comm. Pure Appl. Math., 60 (2007), pp. 911–950.
[24] T. Gallay and M. Hărăguş, *Stability of small periodic waves for the nonlinear Schrödinger equation*, J. Differential Equations, 234 (2007), pp. 544–581.
[25] T. Gallay and M. Hărăguş, *Orbital stability of periodic waves for the nonlinear Schrödinger equation*, J. Dynam. Differential Equations, 19 (2007), pp. 825–865.
[26] R. A. Gardner, *Spectral analysis of long wavelength periodic waves and applications*, J. Reine Angew. Math., 491 (1997), pp. 149–181.
[27] M. Grillakis, J. Shatah, and W. Strauss, *Stability theory of solitary waves in the presence of symmetry* I, J. Funct. Anal., 74 (1987), pp. 160–197.
[28] M. Grillakis, J. Shatah, and W. Strauss, *Stability theory of solitary waves in the presence of symmetry* II, J. Funct. Anal., 94 (1990), pp. 308–348.
[29] E. L. Ince, *The periodic Lamé functions*, Proc. Roy. Soc. Edinburgh, 60 (1940), pp. 47–63.
[30] R. J. Iorio, Jr., and V. M. V. Iorio, *Fourier Analysis and Partial Differential Equations*, Cambridge Stud. Adv. Math. 70, Cambridge University Press, Cambridge, UK, 2001.
[31] S. Karlin, *The existence of eigenvalues for integral operators*, Trans. Amer. Math. Soc., 113 (1964), pp. 1–17.
[32] S. Karlin *Total Positivity*, Stanford University Press, Stanford, CA, 1968.
[33] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer, Berlin, 1976.
[34] D. J. Korteweg and G. de Vries, *On the change of form of long wave advancing in a rectangular canal, and on a nem type of long stationary waves*, Philos. Mag., 39 (1895), pp. 422–443.
[35] W. Magnus and E. Oberhettinger, *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer, New York, 1986.
[36] W. Magnus and S. Winkler, *Hill's Equation*, Tracts in Pure and Appl. Math. 20, Wiley, New York, 1976.
[37] H. P. McKean, *Stability for the Korteweg-de Vries equation*, Comm. Pure Appl. Math., 30 (1997), pp. 347–353.
[38] H. P. McKean and H. Dym, *Fourier Series and Integrals*, Academic Press, New York, London, 1972.
[39] L. Molinet, *Global well-posedness in* L^2 *for the periodic Benjamin-Ono equation*, Amer. J. Math., 130 (2008), pp. 635–683.

[40] L. Molinet, *Global well-posedness in the energy space for the Benjamin-Ono equation on the circle*, Math. Ann., 337 (2007), pp. 353–383.
[41] L. Molinet and F. Ribaud, *Well-Posedness in H^1 for the (Generalized) Benjamin-Ono Equation on the Circle*, preprint, 2006.
[42] F. Oberhettinger, *Fourier Expansions: A Collection of Formulas*, Academic Press, New York, London, 1973.
[43] H. Ono, *Algebraic solitary waves in stratified fluids*, J. Phys. Soc. Japan, 39 (1975), pp. 1082–1091.
[44] A. R. Osborne, M. Serio, L. Bergamasco, and L. Cavaleri, *Solitons, cnoidal waves and nonlinear interactions in shallow-water ocean surface waves*, Phys. D, 123 (1998), pp. 64–81.
[45] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.
[46] E. M. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, NJ, 1970.
[47] J. F. Toland, *Stokes waves*, Topol. Methods Nonlinear Anal., 7 (1996), pp. 1–48.
[48] M. I. Weinstein, *Existence and dynamic stability of solitary wave solutions of equations arising in long wave propagation*, Comm. Partial Differential Equations, 12 (1987), pp. 1133–1173.
[49] M. I. Weinstein, *Lyapunov stability of ground states of nonlinear dispersive evolution equations*, Comm. Pure Appl. Math., 39 (1986), pp. 51–67.

SIAM J. MATH. ANAL.
Vol. 40, No. 3, pp. 1152–1200

INHOMOGENEOUS BOUNDARY VALUE PROBLEMS FOR COMPRESSIBLE NAVIER–STOKES EQUATIONS: WELL-POSEDNESS AND SENSITIVITY ANALYSIS*

P. I. PLOTNIKOV†, E. V. RUBAN†, AND J. SOKOLOWSKI‡

Abstract. In this paper compressible, stationary Navier–Stokes equations are considered. A framework for analysis of such equations is established. In particular, the well-posedness for inhomogeneous boundary value problems of elliptic-hyperbolic type is shown. Analysis is performed for small perturbations of the so-called *approximate solutions* that take form (1.12). The approximate solutions are determined from Stokes problem (1.11). The small perturbations are given by solutions to (1.13). The uniqueness of solutions for problem (1.13) is proved, and, in addition, the differentiability of solutions with respect to the coefficients of differential operators is shown. The results on the well-posedness of nonlinear problems are interesting on their own and are used to obtain the shape differentiability of the drag functional for incompressible Navier–Stokes equations. The shape gradient of the drag functional is derived in the classical and useful for computations form; an appropriate adjoint state is introduced to this end. The *material* derivatives of solutions to the Navier–Stokes equations are given by smooth functions; however, the shape differentiability is shown in a weak norm. The method of analysis proposed in this paper is general and can be used to establish the well-posedness for distributed and boundary control problems as well as for inverse problems in the case of the state equations in the form of compressible Navier–Stokes equations. The differentiability of solutions to the Navier–Stokes equations with respect to the data leads to the first order necessary conditions for a broad class of optimization problems.

Key words. Navier–Stokes equations, compressible fluids, shape optimization

AMS subject classifications. 35Q30, 49J20, 76N10

DOI. 10.1137/070694272

1. Introduction.

Shape optimization for compressible Navier–Stokes equations is important for applications [27] and is investigated from a numerical point of view; however, the mathematical analysis of such problems is not available in the existing literature. One of the reasons is the lack of existence results for inhomogeneous boundary value problems for such equations.

The results established in this paper lead in particular to the first order optimality conditions for a class of shape optimization problems for compressible Navier–Stokes equations.

1.1. Problem formulation.

In this paper we prove the well-posedness and perform the sensitivity analysis for inhomogeneous boundary value problems for the compressible Navier–Stokes equations. We restrict ourselves to the case of a specific shape optimization problem for stationary motion of viscous compressible non–heat-conducting isentropic gas. However, the technique of modelling and analysis presented here is general and can be used for a broad class of optimization problems for nonlinear elliptic-hyperbolic equations. The sensitivity analysis is the necessary

*Received by the editors June 11, 2007; accepted for publication (in revised form) June 20, 2008; published electronically October 22, 2008. This paper was prepared in the fall of 2006 during a visit of Pavel I. Plotnikov and Evgenya V. Ruban to the Institute Elie Cartan of the University Henri Poincaré Nancy 1.

http://www.siam.org/journals/sima/40-3/69427.html

†Lavryentyev Institute of Hydrodynamics, Siberian Division of Russian Academy of Sciences, Lavryentyev pr. 15, Novosibirsk 630090, Russia (plotnikov@hydro.nsc.ru, zhenya_ruban@mail.ru).

‡Institut Elie Cartan, Laboratoire de Mathématiques, Université Henri Poincaré Nancy 1, B.P. 239, 54506 Vandoeuvre lés Nancy Cedex, France (Jan.Sokolowski@iecn.u-nancy.fr).

step for numerical methods of solution for optimization problems. In general the mathematical analysis of optimization problems includes the following steps, with the mathematical proofs of the required facts:

- existence of solutions,
- uniqueness and optimality conditions,
- numerical method of solution.

The existence of an optimal shape for the drag minimization is shown in [40] under assumptions that are compared to the assumptions in the present paper. Here we present the necessary mathematical tools required for the second step of analysis, i.e., the derivation of an optimality system. In particular, we prove the shape differentiability of solutions to (1.9) and provide the classical representation of the shape derivatives of integral shape functionals in terms of an appropriate adjoint state.

We consider in detail all questions on the existence, uniqueness, and shape differentiability of solutions to stationary boundary value problems for compressible Navier–Stokes equations. Such boundary value problems can be regarded as the mathematical models of viscous gas flow around a body tested in the wind tunnel. We assume that the viscous gas occupies the double-connected domain $\Omega = B\backslash S$, where $B \subset \mathbb{R}^3$ is a hold-all domain with the smooth boundary $\Sigma = \partial B$, and $S \subset B$ is a compact obstacle. Furthermore, we assume that the velocity of the gas coincides with a given vector field $\mathbf{U} \in C^\infty(\mathbb{R}^3)^3$ on the surface Σ. In this framework, the boundary of the flow domain Ω is divided into three subsets, inlet $\Sigma_{\rm in}$, outgoing set $\Sigma_{\rm out}$, and characteristic set Σ_0, which are defined by the equalities

$$\Sigma_{\rm in} = \{x \in \Sigma : \mathbf{U} \cdot \mathbf{n} < 0\}, \quad \Sigma_{\rm out} = \{x \in \Sigma : \mathbf{U} \cdot \mathbf{n} > 0\}, \tag{1.1}$$

$\Sigma_0 = \{x \in \partial\Omega : \mathbf{U} \cdot \mathbf{n} = 0\}$, where $\mathbf{n}$ stands for the unit outward normal to $\partial\Omega = \Sigma \cup \partial S$. In turn the compact $\Gamma = \Sigma_0 \cap \Sigma$ splits the surface Σ into three disjoint parts $\Sigma = \Sigma_{\rm in} \cup \Sigma_{\rm out} \cup \Gamma$. The problem is to find the velocity field $\mathbf{u}$ and the gas density ϱ satisfying the following equations along with the boundary conditions:

$$\Delta\mathbf{u} + \lambda\nabla \operatorname{div}\mathbf{u} = R\varrho\mathbf{u} \cdot \nabla\mathbf{u} + \frac{R}{\epsilon^2}\nabla p(\varrho) \ \text{ in } \ \Omega, \tag{1.2a}$$

$$\operatorname{div}(\varrho\mathbf{u}) = 0 \ \text{ in } \ \Omega, \tag{1.2b}$$

$$\mathbf{u} = \mathbf{U} \ \text{ on } \ \Sigma, \quad \mathbf{u} = 0 \ \text{ on } \ \partial S, \tag{1.2c}$$

$$\varrho = \varrho_0 \ \text{ on } \ \Sigma_{\rm in}, \tag{1.2d}$$

where the pressure $p = p(\varrho)$ is a smooth, strictly monotone function of the density, ϵ is the Mach number, R is the Reynolds number, λ is the viscosity ratio, and ϱ_0 is a positive constant.

For the derivation of equations (1.2) we refer to [22]. The general theory of compressible Navier–Stokes equations is covered by the monographs [11], [25], and [33]. In particular, the main results on the existence of global weak solutions for stationary problems with the zero velocity boundary conditions were established in [25] and sharpened in [33]. See also [14], [38], [39] for generalizations.

There are numerous papers dealing with the zero velocity boundary value problem for steady compressible Navier–Stokes equations in the context of small data. We recall only that there are three different approaches to this problem proposed in [2], [35], and [30], [32], respectively. The basic results on the local existence and uniqueness of strong solutions are assembled in [33]. For an interesting overview see [36].

The inhomogeneous boundary problems were studied in papers [20] and [21], where the local existence and uniqueness results were obtained in the two-dimensional

case under the assumption that the velocity $\mathbf{u}$ is close to a given constant vector. The question of the existence of strong solutions to boundary value problems in three spatial dimensions with nonzero velocity boundary data in smooth domains is still an open problem. There are difficulties including the problems of the total mass control and the singularities developed by solutions at the manifold $\overline{\Sigma}_{in} \cap \overline{\Sigma_0 \cup \Sigma}_{out}$. In this paper we prove the local existence and uniqueness of strong solutions to problem (1.2) in fractional Sobolev spaces, under the assumption that the given vector field $\mathbf{U}$ satisfies the emergent vector field conditions (**H1**)–(**H3**) on Γ. It seems that a condition of this type is necessary for the continuity of mass density ϱ.

Shape optimization problems. Among many shape optimization problems for Navier–Stokes equations, we could list the drag minimization problem, which is investigated in this paper and in [37], [38], [39], [40]. It is important to note that the drag from one side depends on the shape of the obstacle, and from the other side on the design of its surface structure as well as on the quality of its surface. The problem of the drag reduction includes the optimal design of the shape of the obstacle and control of the flow near its surface. The last question was investigated in paper [17], which also contains an overview of the problem.

Another problem of practical interest concerns the optimal shape of tunnels [27]. In the specific problem the required mass distribution on the outlet of the tunnel is given. The associated shape optimization problem can be formulated as follows. Determine an admissible domain such that the mass distribution at the inlet is given, the velocity field is prescribed on the boundary of the domain, and the mass distribution at the outlet is as close as possible to a given function. Inlet and outlet subsets are defined by the vector field $\mathbf{U}$ which serves as the inhomogeneous boundary condition for the law of momentum conservation in the form of a Navier–Stokes stationary system. The shape optimization problem as it is formulated in [27] enters into our framework, and the results on shape sensitivity analysis can be applied to solve the problem. Another class of problems which can be investigated using the tools proposed in the paper are optimal control problems, e.g., with the boundary controls. These topics are, however, beyond the scope of the paper, and we present the drag minimization problem as an example to the general theory.

Drag minimization. One of the main applications of the theory of compressible viscous flows is the optimal shape design in aerodynamics. The classical sample is the problem of the minimization of the drag of airfoil travelling in atmosphere with uniform speed $\mathbf{U}_\infty$. Recall that in our framework the hydrodynamical force acting on the body S is defined by the formula [41]

$$\mathbf{J}(S) = -\int_{\partial S} \left(\nabla \mathbf{u} + (\nabla \mathbf{u})^* + (\lambda - 1) \operatorname{div} \mathbf{u} \mathbf{I} - \frac{R}{\epsilon^2} p \mathbf{I} \right) \cdot \mathbf{n} dS.$$

In a frame attached to the moving body the drag is the component of $\mathbf{J}$ parallel to $\mathbf{U}_\infty$,

$$J_D(S) = \mathbf{U}_\infty \cdot \mathbf{J}(S), \tag{1.3}$$

and the lift is the component of $\mathbf{J}$ in the direction orthogonal to $\mathbf{U}_\infty$. For the fixed data, the drag can be regarded as a functional depending on the shape of the obstacle S. The drag minimization and the lift maximization are between shape optimization problems of some practical importance. The questions of the domain dependence of solutions to nonstationary compressible Navier–Stokes equations and of the solvability of the drag optimization problem were considered in papers [12], [13]. The solvability

of the drag minimization problem for stationary equations (1.2) is shown in [37], [40]. For incompressible Navier–Stokes equations, the existence of material derivatives of solutions and the formula for the shape derivative of the drag functional and adjoint state were obtained in [4], [5], and [42]; see also [43] and [44] for some generalizations. The growing literature on numerical and applied aspects of the problem is nicely surveyed in [18] and [27]. To the best of our knowledge, the mathematical sensitivity analysis for the compressible Navier–Stokes equations has not been studied yet. We derive the formula for the shape derivatives of the drag functional which can be used, in particular, for the explicit formulation of optimality conditions. In order to define the shape derivatives of the shape functional we combine the material derivatives of the solutions to the governing PDEs with an appropriate adjoint state according to the same scheme as is proposed, e.g., in [42] for steady incompressible equations.

We start with a description of our framework for shape sensitivity analysis, or more generally, for well-posedness of compressible Navier–Stokes equations. To this end we choose the vector field $\mathbf{T} \in C^2(\mathbb{R}^3)^3$ vanishing in the vicinity of Σ and define the mapping

$$y = x + \varepsilon \mathbf{T}(x), \tag{1.4}$$

which describes the perturbation of the shape of the obstacle. We refer the reader to [45] for more general framework and results in shape optimization and to [28] for shape calculus in the framework of fluid-structure interaction. For small ε, the mapping $x \to y$ diffeomorphically takes the flow region Ω onto $\Omega_\varepsilon = B \setminus S_\varepsilon$, where the perturbed obstacle $S_\varepsilon = y(S)$. Let $(\bar{\mathbf{u}}_\varepsilon, \bar{\varrho}_\varepsilon)$ be solutions to problem (1.2) in Ω_ε. After substituting $(\bar{\mathbf{u}}_\varepsilon, \bar{\varrho}_\varepsilon)$ into the formula for $\mathbf{J}$, the drag becomes the function of the parameter ε. Our aim is, in fact, to prove that this function is well defined and differentiable at $\varepsilon = 0$. This leads to the first order shape sensitivity analysis for solutions to compressible Navier–Stokes equations. It is convenient to reduce such an analysis to the analysis of dependence of solutions with respect to the coefficients of the governing equations. To this end, we introduce the functions $\mathbf{u}_\varepsilon(x)$ and $\varrho_\varepsilon(x)$ defined in the unperturbed domain Ω by the formulae

$$\mathbf{u}_\varepsilon(x) = \mathbf{N}\bar{\mathbf{u}}_\varepsilon(x + \varepsilon\mathbf{T}(x)), \quad \varrho_\varepsilon(x) = \bar{\varrho}_\varepsilon(x + \varepsilon\mathbf{T}(x)),$$

where

$$\mathbf{N}(x) = \det\,(\mathbf{I} + \varepsilon\mathbf{T}'(x))(\mathbf{I} + \varepsilon\mathbf{T}'(x))^{-1} \tag{1.5}$$

is the adjugate matrix of the Jacobi matrix $\mathbf{I} + \varepsilon\mathbf{T}'$. Furthermore, we also use the notation $\mathfrak{g}(x) = \sqrt{\det\,\mathbf{N}}$. It is easy to see that the matrices $\mathbf{N}(x)$ depend analytically upon the small parameter ε and

$$\mathbf{N} = \mathbf{I} + \varepsilon\mathbf{D}(x) + \varepsilon^2\mathbf{D}_1(\varepsilon, x), \tag{1.6}$$

where $\mathbf{D} = \operatorname{div}\mathbf{T}\mathbf{I} - \mathbf{T}'$. Calculations show that, for $\mathbf{u}_\varepsilon, \varrho_\varepsilon$, the following boundary value problem is obtained:

$$\Delta\mathbf{u}_\varepsilon + \nabla\left(\lambda\mathfrak{g}^{-1}\operatorname{div}\mathbf{u}_\varepsilon - \frac{R}{\epsilon^2}p(\varrho_\varepsilon)\right) = \mathscr{A}(\mathbf{u}_\varepsilon) + R\mathscr{B}(\varrho_\varepsilon, \mathbf{u}_\varepsilon, \mathbf{u}_\varepsilon) \ \text{ in } \ \Omega, \tag{1.7a}$$

$$\operatorname{div}\left(\varrho_\varepsilon\mathbf{u}_\varepsilon\right) = 0 \ \text{ in } \ \Omega, \tag{1.7b}$$

$$\mathbf{u}_\varepsilon = \mathbf{U} \ \text{ on } \ \Sigma, \quad \mathbf{u}_\varepsilon = 0 \ \text{ on } \ \partial S, \tag{1.7c}$$

$$\varrho_\varepsilon = \varrho_0 \ \text{ on } \ \Sigma_{\text{in}}. \tag{1.7d}$$

Here the linear operator $\mathscr{A}$ and the nonlinear mapping $\mathscr{B}$ are defined in terms of $\mathbf{N}$:

$$\begin{aligned}\mathscr{A}(\mathbf{u}) &= \Delta\mathbf{u} - (\mathbf{N}^*)^{-1}\operatorname{div}\big(\mathfrak{g}^{-1}\mathbf{N}\mathbf{N}^*\nabla(\mathbf{N}^{-1}\mathbf{u})\big),\\ \mathscr{B}(\varrho,\mathbf{u},\mathbf{w}) &= \varrho(\mathbf{N}^*)^{-1}\Big(\mathbf{u}\,\nabla\big(\mathbf{N}^{-1}\mathbf{w}\big)\Big).\end{aligned} \tag{1.8}$$

For the derivation of these equations, see Appendix C.

The specific structure of the matrix $\mathbf{N}$ does not play any particular role in the further analysis. Therefore, we consider a general problem of the existence, uniqueness, and dependence on coefficients of the solutions to equations (1.7) under the assumption that $\mathbf{N}$ is a given matrix-valued function which is close, in an appropriate norm, to the identity mapping $\mathbf{I}$ and coincides with $\mathbf{I}$ in the vicinity of Σ. By abuse of notation, we write simply $\mathbf{u}$ and ϱ, instead of $\mathbf{u}_\varepsilon$ and ϱ_ε, when studying the well-posedness and dependence on $\mathbf{N}$. Before formulation of main results we write the governing equation in more transparent form using the change of unknown functions proposed in [35], [15]. To do so we introduce *the effective viscous pressure*

$$q = \frac{R}{\epsilon^2}\,p(\varrho) - \lambda\mathfrak{g}^{-1}\operatorname{div}\mathbf{u}$$

and rewrite equations (1.7) in the equivalent form

$$\Delta\mathbf{u} - \nabla q = \mathscr{A}(\mathbf{u}) + R\mathscr{B}(\varrho,\mathbf{u},\mathbf{u}) \ \text{ in } \ \Omega, \tag{1.9a}$$

$$\operatorname{div}\mathbf{u} = \mathfrak{g}\sigma_0 p(\varrho) - \frac{\mathfrak{g}q}{\lambda} \ \text{ in } \ \Omega, \tag{1.9b}$$

$$\mathbf{u}\cdot\nabla\varrho + \mathfrak{g}\sigma_0 p(\varrho)\,\varrho = \frac{\mathfrak{g}q}{\lambda}\varrho \ \text{ in } \ \Omega, \tag{1.9c}$$

$$\mathbf{u} = \mathbf{U} \ \text{ on } \ \Sigma, \quad \mathbf{u} = 0 \ \text{ on } \ \partial S, \tag{1.9d}$$

$$\varrho = \varrho_0 \ \text{ on } \ \Sigma_{\text{in}}, \tag{1.9e}$$

where $\sigma_0 = R/(\lambda\epsilon^2)$. In the new variables $(\mathbf{u}, q, \varrho)$ the expression for the force $\mathbf{J}$ reads

$$\mathbf{J} = -\int_\Omega \big[\mathfrak{g}^{-1}\big(\mathbf{N}^*\nabla(\mathbf{N}^{-1}\mathbf{u}) + \nabla(\mathbf{N}^{-1}\mathbf{u})^*\mathbf{N} - \operatorname{div}\mathbf{u}\big) - q - R\varrho\mathbf{u}\otimes\mathbf{u}\big]\mathbf{N}^*\nabla\eta\,dx, \tag{1.10}$$

where $\eta \in C^\infty(\Omega)$ is an arbitrary function, which is equal to 1 in an open neighborhood of the obstacle S and 0 in a vicinity of Σ. The value of $\mathbf{J}$ is independent of the choice of the function η.

We assume that $\lambda \gg 1$ and $R \ll 1$, which corresponds to almost incompressible flow with low Reynolds number. In such a case, the *approximate solutions* to problem (1.9) can be chosen in the form $(\varrho_0, \mathbf{u}_0, q_0)$, where ϱ_0 is a constant in boundary condition (1.9e), and $(\mathbf{u}_0, q_0)$ is a solution to the boundary value problem for the Stokes equations,

$$\begin{gathered}\Delta\mathbf{u}_0 - \nabla q_0 = 0, \quad \operatorname{div}\mathbf{u}_0 = 0 \ \text{ in } \ \Omega,\\ \mathbf{u}_0 = \mathbf{U} \ \text{ on } \ \Sigma, \quad \mathbf{u}_0 = 0 \ \text{ on } \ \partial S, \quad \Pi q_0 = q_0.\end{gathered} \tag{1.11}$$

In our notation Π is the projector:

$$\Pi u = u - \frac{1}{\operatorname{meas}\Omega}\int_\Omega u\,dx.$$

Equations (1.11) can be obtained as the limit of equations (1.9) for the passage $\lambda \to \infty$, $R \to 0$. It follows from the standard elliptic theory that, for the boundary $\partial\Omega \in C^\infty$, we have $(\mathbf{u}_0, q_0) \in C^\infty(\Omega)$. We look for solutions to problem (1.9) in the form

$$\mathbf{u} = \mathbf{u}_0 + \mathbf{u}, \quad \varrho = \varrho_0 + \varphi, \quad q = q_0 + \lambda\sigma_0 p(\varrho_0) + \pi + \lambda m, \tag{1.12}$$

with the unknowns functions $\vartheta = (\mathbf{u}, \pi, \varphi)$ and the unknown constant m. Substituting (1.12) into (1.9) we obtain the following boundary problem for ϑ:

$$\begin{aligned} \Delta\mathbf{u} - \nabla\pi &= \mathscr{A}(\mathbf{u}) + R\mathscr{B}(\varrho, \mathbf{u}, \mathbf{u}) \ \text{ in } \ \Omega, \\ \operatorname{div}\mathbf{u} &= \mathfrak{g}\left(\frac{\sigma}{\varrho_0}\varphi - \Psi[\vartheta] - m\right) \ \text{ in } \ \Omega, \\ \mathbf{u}\cdot\nabla\varphi + \sigma\varphi &= \Psi_1[\vartheta] + m\mathfrak{g}\varrho \ \text{ in } \ \Omega, \\ \mathbf{u} = 0 \ \text{ on } \ \partial\Omega, \quad \varphi = 0 \ \text{ on } \ \Sigma_{\text{in}}, &\quad \Pi\pi = \pi, \end{aligned} \tag{1.13a}$$

where

$$\Psi_1[\vartheta] = \mathfrak{g}\left(\varrho\Psi[\vartheta] - \frac{\sigma}{\varrho_0}\varphi^2\right) + \sigma\varphi(1-\mathfrak{g}), \quad \Psi[\vartheta] = \frac{q_0+\pi}{\lambda} - \frac{\sigma}{p'(\varrho_0)\varrho_0}H(\varphi),$$
$$\sigma = \sigma_0 p'(\varrho_0)\varrho_0, \quad H(\varphi) = p(\varrho_0+\varphi) - p(\varrho_0) - p\prime(\varrho_0)\varphi,$$

and the vector field $\mathbf{u}$ and the function ϱ are given by (1.12). Finally, we specify the constant m. In our framework, in contrast to the case of the homogeneous boundary problem, the solution to such a problem is not trivial. Note that, since $\operatorname{div}\mathbf{u}$ is of the null mean value, the right-hand side of $(1.13\text{a})_3$ must satisfy the compatibility condition

$$m\int_\Omega \mathfrak{g}\,dx = \int_\Omega \mathfrak{g}\left(\frac{\sigma}{\varrho_0}\varphi - \Psi[\vartheta]\right)\,dx,$$

which formally determines m. This choice of m leads to essential mathematical difficulties. To make this issue clear note that in the simplest case $\mathfrak{g} = 1$ we have $m = \varrho_0^{-1}\sigma(\mathbf{I}-\Pi)\varphi + O(|\vartheta|^2, \lambda^{-1})$, and the principal linear part of the governing equations (1.13a) becomes

$$\begin{pmatrix} \Delta & -\nabla & 0 \\ \operatorname{div} & 0 & -\frac{\sigma}{\varrho_0} \\ 0 & 0 & \mathbf{u}\nabla + \sigma \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \pi \\ \varphi \end{pmatrix} + \begin{pmatrix} 0 \\ m \\ -m\varrho_0 \end{pmatrix} \sim \begin{pmatrix} \Delta\mathbf{u} - \nabla\pi \\ \operatorname{div}\mathbf{u} - \frac{\sigma}{\varrho_0}\Pi\varphi \\ \mathbf{u}\nabla\varphi + \sigma\Pi\varphi \end{pmatrix}.$$

Hence, the question of solvability of the linearized equations derived for (1.13) can be reduced to the question of solvability of the boundary value problem for the nonlocal transport equation

$$\mathbf{u}\nabla\varphi + \sigma\Pi\varphi = f,$$

which is very difficult because of the loss of maximum principle. In fact, this question is concerned with the problem of the control of the total gas mass in compressible flows. Recall that the absence of the mass control is the main obstacle for proving the global solvability of inhomogeneous boundary problems for compressible Navier–Stokes equations; we refer to [25] for discussion. In order to cope with this difficulty we write the compatibility condition in a sophisticated form, which allows us to control

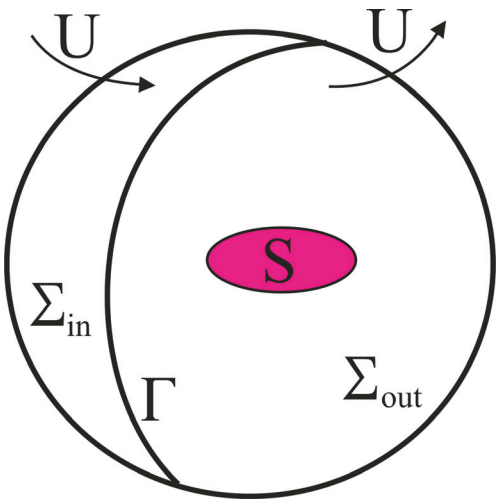


FIG. 1. *Flow domain Ω with an obstacle S.*

the total mass of the gas. To this end we introduce the auxiliary function ζ satisfying the equations

$$-\operatorname{div}(\mathbf{u}\zeta) + \sigma\zeta = \sigma\mathfrak{g} \ \text{ in } \ \Omega, \quad \zeta = 0 \ \text{ on } \ \Sigma_{\text{out}}, \tag{1.13b}$$

and fix the constant m as follows:

$$m = \varkappa \int_\Omega (\varrho_0^{-1}\Psi_1[\vartheta]\zeta - \mathfrak{g}\Psi[\vartheta])\,dx, \quad \varkappa = \left(\int_\Omega \mathfrak{g}(1 - \zeta - \varrho_0^{-1}\zeta\varphi)\,dx\right)^{-1}. \tag{1.13c}$$

In this way the auxiliary function ζ becomes an integral part of the solution to problem (1.13). Now our aim is to prove the existence and uniqueness of solutions to problem (1.13) and investigate the dependence of the solutions on matrices $\mathbf{N}$. Before the presentation of the main results we introduce some notation and formulate preliminary results.

Geometrical conditions on the flow region. We assume that a surface $\Sigma = \Sigma_{\text{in}} \cup \Sigma_{\text{out}} \cup \Gamma$ and a given vector field $\mathbf{U}$ satisfy the following conditions, referred to as the *emergent vector field conditions*.

CONDITION 1.1. *The set Γ is a closed $C^{2+\alpha}$ one-dimensional manifold. Moreover, there is a positive constant c such that*

$$\mathbf{U} \cdot \nabla(\mathbf{U} \cdot \mathbf{n}) > c > 0 \quad \textit{on} \quad \Gamma. \tag{1.14}$$

These conditions have a simple geometric interpretation, that $\mathbf{U} \cdot \mathbf{n}$ vanishes only up to the first order at Γ, and $\mathbf{U}$ is transversal to Γ; furthermore, for each point $P \in \Gamma$, the vector $\mathbf{U}(P)$ points to the part of Σ where $\mathbf{U}$ is an exterior vector field (see Figure 1). Note that the emergent vector field condition plays an important role in the theory of the classical oblique derivative problem; see [16]. In the context of our problem it is equivalent to the following conditions:

(**H1**) The boundary of Ω belongs to class $C^{2+\alpha}$, $\alpha \in (0,1)$. For each point $P \in \Gamma$ there exist the local Cartesian coordinates (x_1, x_2, x_3) with the origin at P such that in the new coordinates $\mathbf{U}(P) = (U, 0, 0)$ with $U = |\mathbf{U}(P)|$, and

$\mathbf{n}(P) = (0, 0, -1)$. Moreover, there is a neighborhood $\mathcal{O} = [-k, k]^2 \times [-t, t]$ of P such that the intersections $\Sigma \cap \mathcal{O}$ and $\Gamma \cap \mathcal{O}$ are defined by the equations

$$F_0(x) \equiv x_3 - F(x_1, x_2) = 0, \quad \nabla F_0(x) \cdot \mathbf{U}(x) = 0,$$

and $\Omega \cap \mathcal{O}$ is the epigraph $\{F_0 > 0\} \cap \mathcal{O}$. The function F satisfies the conditions

$$\|F\|_{C^2([-k,k]^2)} \le K, \quad F(0,0) = 0, \quad \nabla F(0,0) = 0, \tag{1.15}$$

where the constants $k, t < 1$ and $K > 1$ depend only on the curvature of Σ and are independent of the point P.

(**H2**) For a suitable choice of the constant k, with k independent of $P \in \Gamma$, the manifold $\Gamma \cap \mathcal{O}$ admits the parameterization

$$x = \mathbf{x}^0(x_2) := \big(\Upsilon(x_2),\, x_2,\, F(\Upsilon(x_2), x_2)\big) \tag{1.16}$$

such that $\Upsilon(0) = 0$ and $\|\Upsilon\|_{C^2([-k,k])} \le C_\Gamma$, where the constant $C_\Gamma > 1$ depends only on Σ and $\mathbf{U}$.

(**H3**) There are positive constants $N^\pm$ independent of P such that for $x \in \Sigma$ given by the condition $F_0(x_1, x_2, x_3) = x_3 - F(x_1, x_2) = 0$ we have

(1.17)
$$N^-\big(x_1 - \Upsilon(x_2)\big) \le -\nabla F_0(x) \cdot \mathbf{U}(x) \le N^+\big(x_1 - \Upsilon(x_2)\big) \quad \text{for} \quad x_1 > \Upsilon(x_2),$$
$$-N^-\big(x_1 - \Upsilon(x_2)\big) \le \nabla F_0(x) \cdot \mathbf{U}(x) \le -N^+\big(x_1 - \Upsilon(x_2)\big) \quad \text{for} \quad x_1 < \Upsilon(x_2).$$

Function spaces. In this paragraph we assemble some technical results which are used throughout the paper. Function spaces play a central role, and we recall some notation, fundamental definitions, and properties, which can be found in [1] and [6]. For the convenience of the reader we collect in Appendix B the basic facts from the theory of interpolation spaces. For our applications we need the results in three spatial dimensions; however, the results are presented for the dimension $d \ge 2$.

Let Ω be the whole space $\mathbb{R}^d$ or a bounded domain in $\mathbb{R}^d$ with the boundary $\partial\Omega$ of class C^1. For an integer $l \ge 0$ and for an exponent $r \in [1, \infty)$, we denote by $W^{l,r}(\Omega)$ the Sobolev space endowed with the norm $\|u\|_{W^{l,r}(\Omega)} = \sup_{|\alpha| \le l} \|\partial^\alpha u\|_{L^r(\Omega)}$. For real $0 < s < 1$, the fractional Sobolev space $W^{s,r}(\Omega)$ is obtained by the real interpolation method (see [46] for the proofs) between $L^r(\Omega)$ and $W^{1,r}(\Omega)$ and consists of all measurable functions with the finite norm

$$\|u\|_{W^{s,r}(\Omega)} = \|u\|_{L^r(\Omega)} + |u|_{s,r,\Omega},$$

where

$$|u|^r_{s,r,\Omega} = \int_{\Omega \times \Omega} |x - y|^{-d-rs} |u(x) - u(y)|^r \, dxdy. \tag{1.18}$$

In the general case, the Sobolev space $W^{l+s,r}(\Omega)$ is defined as the space of measurable functions with the finite norm $\|u\|_{W^{l+s,r}(\Omega)} = \sup_{|\alpha| \le l} \|\partial^\alpha u\|_{W^{s,r}(\Omega)}$. For $0 < s < 1$, the Sobolev space $W^{s,r}(\Omega)$ is, in fact [6], the interpolation space $[L^r(\Omega), W^{1,r}(\Omega)]_{s,r}$.

Furthermore, the notation $W^{l,r}_0(\Omega)$, with an integer l, stands for the closed subspace of the space $W^{l,r}(\Omega)$ of all functions $u \in L^r(\Omega)$ which being extended by zero outside of Ω belong to $W^{l,r}(\mathbb{R}^d)$.

Denote by $\mathcal{W}_0^{0,r}(\Omega)$ and $\mathcal{W}_0^{1,r}(\Omega)$ the subspaces of $L^r(\mathbb{R}^d)$ and $W^{1,r}(\mathbb{R}^d)$, respectively, of all functions vanishing outside of Ω. Obviously $\mathcal{W}_0^{1,r}(\Omega)$ and $W_0^{1,r}(\Omega)$ are isomorphic topologically and algebraically and we can identify them. However, we need the interpolation spaces $\mathcal{W}_0^{s,r}(\Omega)$ for nonintegers, in particular, for $s = 1/r$.

DEFINITION 1.2. *For all $0 < s \le 1$ and $1 < r < \infty$, we denote by $\mathcal{W}_0^{s,r}(\Omega)$ the interpolation space $[\mathcal{W}_0^{0,r}(\Omega), \mathcal{W}_0^{1,r}(\Omega)]_{s,r}$ endowed with one of the equivalent norms* (6.1) *and* (6.3) *defined by the interpolation method.*

It follows from the definition of interpolation spaces (see Appendix B) that $\mathcal{W}_0^{s,r}(\Omega) \subset W^{s,r}(\mathbb{R}^d)$ and, for all $u \in \mathcal{W}_0^{s,r}(\Omega)$,

$$\|u\|_{W^{s,r}(\mathbb{R}^d)} \le c(r,s)\|u\|_{\mathcal{W}_0^{s,r}(\Omega)}, \quad u = 0 \quad \text{outside} \ \ \Omega. \tag{1.19}$$

In other words, $\mathcal{W}_0^{s,r}(\Omega)$ consists of all elements $u \in W^{s,r}(\Omega)$ such that the extension $\overline{u}$ of u by 0 outside of Ω has the finite $[\mathcal{W}_0^{0,r}(\Omega), \mathcal{W}_0^{1,r}(\Omega)]_{s,r}$-norm. We identify u and $\overline{u}$ for the elements $u \in \mathcal{W}_0^{s,r}(\Omega)$. With this identification it follows that $W_0^{1,r}(\Omega) \subset \mathcal{W}_0^{s,r}(\Omega)$ and the space $C_0^\infty(\Omega)$ is dense in $\mathcal{W}_0^{s,r}(\Omega)$. It is worthy to note that for $0 < s < 1$ and for $1 < r < \infty$, the function $\overline{u}$ belongs to the space $W^{s,r}(\mathbb{R}^d)$ if and only if $u \in W^{s,r}(\Omega)$ and $\operatorname{dist}(x, \partial\Omega)^{-s} u \in L^r(\Omega)$. We also point out that the interpolation space $\mathcal{W}_0^{s,r}(\Omega)$ coincides with the Sobolev space $W_0^{s,r}(\Omega)$ for $s \ne 1/r$. Recall that the standard space $W_0^{s,r}(\Omega)$ is the completion of $C_0^\infty(\Omega)$ in the $W^{s,r}(\Omega)$-norm.

Embedding theorems. For $sr > d$ and $0 \le \alpha < s - r/d$, the embedding $W^{s,r}(\Omega) \hookrightarrow C^\alpha(\Omega)$ is continuous and compact. In particular, for $sr > d$, the Sobolev space $W^{s,r}(\Omega)$ is a commutative Banach algebra; i.e., for all $u, v \in W^{s,r}(\Omega)$,

$$\|uv\|_{W^{s,r}(\Omega)} \le c(r,s)\|u\|_{W^{s,r}(\Omega)}\|v\|_{W^{s,r}(\Omega)}. \tag{1.20}$$

If $sr < d$ and $t^{-1} = r^{-1} - d^{-1}s$, then the embedding $W^{s,r}(\Omega) \hookrightarrow L^t(\Omega)$ is continuous, [1, Thm. 7.57]. We also have [1, Thm. 7.58], for $\alpha < s$, $(s - \alpha)r < d$ and $\beta^{-1} = r^{-1} - d^{-1}(s - \alpha)$,

$$\|u\|_{W^{\alpha,\beta}(\Omega)} \le c(r,s,\alpha,\beta,\Omega)\|u\|_{W^{s,r}(\Omega)}. \tag{1.21}$$

It follows from (1.19) that all the embedding inequalities remain true for the elements of the interpolation space $\mathcal{W}_0^{s,r}(\Omega)$.

Duality. We define

$$\langle u, v\rangle = \int_\Omega u\, v\, dx \tag{1.22}$$

for any functions such that the right-hand side makes sense. For $r \in (1, \infty)$, each element $v \in L^{r'}(\Omega)$, $r' = r/(r-1)$, determines the functional L_v of $(\mathcal{W}_0^{s,r}(\Omega))'$ by the identity $L_v(u) = \langle u, v\rangle$. We introduce the $(-s, r')$-norm of an element $v \in L^{r'}(\Omega)$ to be by definition the norm of the functional L_v, that is

$$\|v\|_{\mathcal{W}^{-s,r'}(\Omega)} = \sup_{\substack{u \in \mathcal{W}_0^{s,r}(\Omega) \\ \|u\|_{\mathcal{W}_0^{s,r}(\Omega)} = 1}} |\langle u, v\rangle|. \tag{1.23}$$

We let $\mathcal{W}^{-s,r'}(\Omega)$ denote the completion of the space $L^{r'}(\Omega)$ with respect to the $(-s, r')$-norm. For an integer s, $\mathcal{W}^{-s,r'}(\Omega)$ is topologically and algebraically isomorphic to $(W_0^{s,r}(\Omega))'$. Moreover, we can identify $\mathcal{W}^{-s,r'}(\Omega)$ with the interpolation space

$[L^{r'}(\Omega), W_0^{-1,r'}(\Omega)]_{s,r}$; see [6] and Appendix B. With this denotation we have the duality principle

$$\|u\|_{\mathcal{W}_0^{s,r}(\Omega)} = \sup_{\substack{v\in C_0^\infty(\Omega)\\ \|v\|_{\mathcal{W}^{-s,r'}(\Omega)}=1}} |\langle u, v\rangle|. \tag{1.24}$$

With applications to the theory of Navier–Stokes equations in mind, we introduce the smaller dual space defined as follows. We identify the function $v \in L^{r'}(\Omega)$ with the functional $L_v \in (W^{s,r}(\Omega))'$ and denote by $\mathbb{W}^{-s,r'}(\Omega)$ the completion of $L^{r'}(\Omega)$ in the norm

$$\|v\|_{\mathbb{W}^{-s,r'}(\Omega)} := \sup_{\substack{u\in W^{s,r}(\Omega)\\ \|u\|_{W^{s,r}(\Omega)}=1}} |\langle u, v\rangle|. \tag{1.25}$$

In the sense of this identification the space $C_0^\infty(\Omega)$ is dense in the interpolation space $\mathbb{W}^{-s,r}(\Omega)$. It follows immediately from the definition that

$$\mathbb{W}^{-s,r'}(\Omega) \subset (W^{s,r}(\Omega))' \subset \mathcal{W}^{-s,r'}(\Omega).$$

For an arbitrary bounded domain $\Omega \subset \mathbb{R}^3$ with a Lipschitz boundary, we introduce the Banach spaces

$$X^{s,r} = W^{s,r}(\Omega) \cap W^{1,2}(\Omega), \quad Y^{s,r} = W^{s+1,r}(\Omega) \cap W^{2,2}(\Omega),$$
$$Z^{s,r} = \mathcal{W}^{s-1,r}(\Omega) \cap L^2(\Omega)$$

equipped with the norms

$$\|u\|_{X^{s,r}} = \|u\|_{W^{s,r}(\Omega)} + \|u\|_{W^{1,2}(\Omega)}, \quad \|u\|_{Y^{s,r}} = \|u\|_{W^{1+s,r}(\Omega)} + \|u\|_{W^{2,2}(\Omega)},$$
$$\|u\|_{Z^{s,r}} = \|u\|_{\mathcal{W}^{s-1,r}(\Omega)} + \|u\|_{L^2(\Omega)}.$$

It can be easily seen that the embeddings $Y^{s,r} \hookrightarrow X^{s,r} \hookrightarrow Z^{s,r}$ are compact and, for $sr > 3$, each of the spaces $X^{s,r}$ and $Y^{s,r}$ is a commutative Banach algebra.

Stokes equations. The following lemma is a straightforward consequence of classical results on solvability of the first boundary value problem for Stokes equations (see [9]) and the interpolation theory.

LEMMA 1.3. *Let $\Omega \subset \mathbb{R}^3$ be a bounded domain with $\partial\Omega \in C^2$ and $(F, G) \in \mathcal{W}^{s-1,r}(\Omega) \times W^{s,r}(\Omega)$ $(0 \le s \le 1, 1 < r < \infty)$. Then the boundary value problem*

$$\begin{gathered}\Delta\mathbf{u} - \nabla\pi = F, \quad \operatorname{div}\mathbf{u} = \Pi G \quad \text{in } \Omega,\\ \mathbf{u} = 0 \quad \text{on } \partial\Omega, \quad \Pi\pi = \pi,\end{gathered} \tag{1.26}$$

has a unique solution $(\mathbf{u}, \pi) \in W^{s+1,r}(\Omega) \times W^{s,r}(\Omega)$ such that

$$\|\mathbf{u}\|_{W^{s+1,r}(\Omega)} + \|\pi\|_{W^{s,r}(\Omega)} \le c(\Omega, r, s)(\|F\|_{\mathcal{W}^{s-1,r}(\Omega)} + \|G\|_{W^{s,r}(\Omega)}). \tag{1.27}$$

In particular, we have

$$\|\mathbf{u}\|_{Y^{s,r}} + \|\pi\|_{X^{s,r}} \le c(\Omega, r, s)(\|F\|_{Z^{s,r}} + \|G\|_{X^{s,r}}).$$

Proof. The proof is in Appendix B. □

Note that the lemma works for singular values $s = 1/r +$ integer. But in this case the traces of solutions at the boundary are not defined.

1.2. Results. *Transport equations.* Today there exists a complete theory of generalized solutions to the class of hyperbolic-elliptic equations developed in [10] and [34] under the assumptions that the equations have C^1 coefficients and satisfy the maximum principle. The questions on smoothness properties of solutions are more difficult. We recall the classical results of [19], [34], related to the case of $\overline{\Sigma}_{\rm in}\cap\overline{\Sigma}_{\rm out}=\emptyset$. The particular case, with $\Sigma_{\rm in}=\Sigma_{\rm out}=\emptyset$, in the Sobolev spaces is covered in the papers [3] and [30], [31]. The case of the nonempty interface $\Gamma=\overline{\Sigma}_{\rm in}\cap\overline{\Sigma}_{\rm out}$ is still weakly investigated. The theory of boundary value problems for transport equations is an integral part of the theory of multicomponents and inhomogeneous flows. In this context this problem has been considered by many authors; see [24] and [29] for discussion.

In general the existence of strong solutions to inhomogeneous boundary value problems for transport equations is still an open problem. The following theorem, which is used throughout this paper, partially fills this gap. Let us consider the following boundary value problems for linear transport equations:

$$\mathscr{L}\varphi:=\mathbf{u}\nabla\varphi+\sigma\varphi=f \quad \text{in } \Omega, \quad \varphi=0 \quad \text{on } \Sigma_{\rm in}, \tag{1.28}$$

$$\mathscr{L}^*\varphi^*:=-\operatorname{div}(\varphi^*\mathbf{u})+\sigma\varphi^*=f \quad \text{in } \Omega, \quad \varphi^*=0 \quad \text{on } \Sigma_{\rm out}. \tag{1.29}$$

The bounded functions φ, φ^* are called the generalized solutions to problems (1.28), (1.29), respectively, if the integral identities

$$\int_\Omega(\varphi\mathscr{L}^*\zeta^*-f\zeta^*)\,dx=0, \quad \int_\Omega(\varphi^*\mathscr{L}\zeta-f\zeta)\,dx=0 \tag{1.30}$$

hold true for all test functions $\zeta^*,\zeta\in C(\Omega)\cap W^{1,1}(\Omega)$, respectively, such that $\zeta^*=0$ on $\Sigma_{\rm out}$ and $\zeta=0$ on $\Sigma_{\rm in}$.

THEOREM 1.4. *Assume that* Σ *and* $\mathbf{U}$ *satisfy conditions* (**H1**)–(**H3**)*, the exponents* s,r *satisfy the inequalities*

$$1/2<s\le 1, \quad 1<r<3/(2s-1), \tag{1.31}$$

and the vector field $\mathbf{u}$ *belongs to the class* $C^1(\Omega)$ *and satisfies the boundary condition*

$$\mathbf{u}=\mathbf{U} \ \text{ on } \ \Sigma, \quad \mathbf{u}=0 \ \text{ on } \ \partial S. \tag{1.32}$$

Then there are positive constants σ^* *and* δ^* *depending only on* Σ, $\mathbf{U}$, s, r, *and* $\|\mathbf{u}\|_{C^1(\Omega)}$ *such that the following hold:*

(i) *For any* $\sigma>\sigma^*$ *and* $f\in W^{s,r}(\Omega)\cap L^\infty(\Omega)$, *problem* (1.28) *has a unique solution* $\varphi\in W^{s,r}(\Omega)\cap L^\infty(\Omega)$ *satisfying the inequalities*

$$\|\varphi\|_{W^{s,r}(\Omega)}\le C\|f\|_{W^{s,r}(\Omega)}, \quad \|\varphi\|_{L^\infty(\Omega)}\le\sigma^{-1}\|f\|_{L^\infty(\Omega)}. \tag{1.33}$$

(ii) *If, in addition,* $\|\operatorname{div}\mathbf{u}\|_{W^{s,r}(\Omega)}+\|\operatorname{div}\mathbf{u}\|_{L^\infty(\Omega)}\le\delta^*$, *problem* (1.29) *has a unique solution* $\varphi^*\in W^{s,r}(\Omega)\cap L^\infty(\Omega)$, *which admits the estimates*

$$\|\varphi^*\|_{W^{s,r}(\Omega)}\le C\|f\|_{W^{s,r}(\Omega)}, \quad \|\varphi^*\|_{L^\infty(\Omega)}\le(\sigma-\delta^*)^{-1}\|f\|_{L^\infty(\Omega)}. \tag{1.34}$$

The constant C *depends only on* $\|\mathbf{u}\|_{C^1(\Omega)}$, r,s σ, $\mathbf{U}$, *and* Ω.

Restriction $s\ge 1/2$ in condition (1.31) is not essential and can be removed. But it plays an important role in the proof of Lemma 1.7 and the main theorem, Theorem 1.9. On the other hand, the inequality $r<3/(2s-1)$ is crucial for the

control of singularities of solutions at the characteristic manifold Γ. It is connected with the behavior of $\mathbf{U}$ in the vicinity of Γ and does not depend on the dimension of the ambient space. Note also that solution is strong for $s = 1$, but not continuous since $W^{1,3}(\Omega)$ is not embedded in $C(\Omega)$. On the other hand, $sr \to \infty$ as $s \to 1/2$ and solution is continuous for s close to $1/2$. In order to obtain continuous strong solutions we have to consider the problem in the scale of spaces $X^{s,r}$

Since for $sr > 3$, the embeddings $X^{s,r} \hookrightarrow C(\Omega)$, $Y^{s,r} \hookrightarrow C^1(\Omega)$ are bounded, we have the following result on solvability of problems (1.28), (1.29) in space $X^{s,r}$.

COROLLARY 1.5. *Assume that $sr > 3$ and the vector field $\mathbf{u}$ has the representation $\mathbf{u} = \mathbf{u}_0 + \mathbf{u}$, where $\mathbf{u}_0 \in C^\infty(\Omega)^3$ is a solution to problem* (1.11). *Then there exist $\tau^* \in (0, 1]$ and σ^*, depending only on $\Sigma, \mathbf{u}_0$, and s, r, such that, for all $\mathbf{u}$ with $\|\mathbf{u}\|_{Y^{s,r}} \le \tau^*$, $\sigma > \sigma^*$, and $f \in X^{s,r}$, each of problems* (1.28) *and* (1.29) *has a unique solution satisfying the inequalities*

$$\|\varphi\|_{X^{s,r}} \le C\|f\|_{X^{s,r}}, \quad \|\varphi^*\|_{X^{s,r}} \le C\|f\|_{X^{s,r}}. \tag{1.35}$$

Existence and uniqueness theory. The second main result of this paper concerns the existence and local uniqueness of solutions to problem (1.13). Denote by E the closed subspace of the Banach space $Y^{s,r}(\Omega)^3 \times X^{s,r}(\Omega)^2$ in the form

$$E = \{\vartheta = (\mathbf{u}, \pi, \varphi) \ : \ \mathbf{u} = 0 \ \text{ on } \ \partial\Omega, \quad \varphi = 0 \ \text{ on } \ \Sigma_{\text{in}}, \quad \Pi\pi = \pi\}, \tag{1.36}$$

and denote by $\mathcal{B}_\tau \subset E$ the closed ball of radius τ centered at 0. Next, note that, for $sr > 3$, elements of the ball $\mathcal{B}_\tau$ satisfy the inequality

$$\|\mathbf{u}\|_{C^1(\Omega)} + \|\pi\|_{C(\Omega)} + \|\varphi\|_{C(\Omega)} \le c_e(r, s, \Omega)\|\vartheta\|_E \le c_e\tau, \tag{1.37}$$

where the norm in E is defined by

$$\|\vartheta\|_E = \|\mathbf{u}\|_{Y^{s,r}(\Omega)} + \|\pi\|_{X^{s,r}(\Omega)} + \|\varphi\|_{X^{s,r}(\Omega)}.$$

THEOREM 1.6. *Assume that the surface Σ and given vector field $\mathbf{U}$ satisfy conditions* (**H**1)–(**H**3). *Furthermore, let σ^*, τ^* be constants given by Corollary* 1.5, *and let positive numbers r, s, σ satisfy the inequalities*

$$1/2 < s \le 1, \quad 1 < r < 3/(2s - 1), \quad sr > 3, \quad \sigma > \sigma^*. \tag{1.38}$$

Then there exists $\tau_0 \in (0, \tau^*]$, depending only on $\mathbf{U}, \Omega, r, s, \sigma$, such that, for all

$$\tau \in (0, \tau_0], \quad \lambda^{-1}, R \in (0, \tau^2], \quad \|\mathbf{N} - \mathbf{I}\|_{C^2(\Omega)} \le \tau^2, \tag{1.39}$$

problem (1.13), *with $\mathbf{u}_0$ given by* (1.11), *has a unique solution $\vartheta \in B_\tau$. Moreover, the auxiliary function ζ and the constants $\varkappa, m$ admit the estimates*

$$\|\zeta\|_{X^{s,r}} + |\varkappa| \le c, \quad |m| \le c\tau < 1, \tag{1.40}$$

where the constant c depends only on $\mathbf{U}, \Omega, r, s$, and σ.

Material derivatives of solutions. Theorem 1.6 guarantees the existence and uniqueness of solutions to problem (1.13) for all $\mathbf{N}$ close to the identity matrix $\mathbf{I}$. The totality of such solutions can be regarded as the mapping from $\mathbf{N}$ to the solution to Navier–Stokes equations. The natural question is the smoothness properties of this mapping, in particular, its differentiability. With application to shape optimization problems in mind, we consider the particular case where the matrices $\mathbf{N}$ depend on

the small parameter ε and have representation (1.6). We assume that C^1-norms of the matrix-valued functions $\mathbf{D}$ and $\mathbf{D}_1(\varepsilon)$ in (1.6) have a majorant independent of ε. By virtue of Theorem 1.6, there are the positive constants ε_0 and τ such that, for all sufficiently small R, λ^{-1} and $\varepsilon \in [0, \varepsilon_0]$, problem (1.13) with $\mathbf{N} = \mathbf{N}(\varepsilon)$ has a unique solution $\vartheta(\varepsilon) = (\mathbf{u}(\varepsilon), \pi(\varepsilon), \varphi(\varepsilon))$, $\zeta(\varepsilon), m(\varepsilon)$, which admits the estimate

$$\|\vartheta(\varepsilon)\|_E + |m(\varepsilon)| \le c\tau, \quad \|\zeta(\varepsilon)\|_{X^{s,r}} \le c, \tag{1.41}$$

where the constant c is independent of ε, and the Banach space E is defined by (1.36). Denote the solution $(\vartheta(0), m(0), \zeta(0))$ for $\varepsilon = 0$ by (ϑ, m, ζ), and define the finite differences with respect to ε,

$$(\mathbf{w}_\varepsilon, \omega_\varepsilon, \psi_\varepsilon) = \varepsilon^{-1}(\vartheta - \vartheta(\varepsilon)), \quad \xi_\varepsilon = \varepsilon^{-1}(\zeta - \zeta(\varepsilon)), \quad n_\varepsilon = \varepsilon^{-1}(m - m(\varepsilon)).$$

Formal calculations show that the limit $(\mathbf{w}, \omega, \psi, \xi, n) = \lim_{\varepsilon \to 0}(\mathbf{w}_\varepsilon, \omega_\varepsilon, \psi_\varepsilon, \xi_\varepsilon, n_\varepsilon)$ is a solution to linearized equations

$$\begin{aligned}
\Delta \mathbf{w} - \nabla \omega &= R\mathscr{C}_0(\mathbf{w}, \psi) + \mathscr{D}_0(\mathbf{D}) \ \text{ in } \ \Omega,\\
\operatorname{div} \mathbf{w} &= b^0_{21}\,\psi - b^0_{22}\,\omega + b^0_{23}\,n + b^0_{20}\,\mathfrak{d} \ \text{ in } \ \Omega,\\
\mathbf{u}\nabla\psi + \sigma\psi &= -\mathbf{w}\cdot\nabla\varphi + b^0_{11}\,\psi + b^0_{12}\,\omega + b^0_{13}\,n + b^0_{10}\,\mathfrak{d} \ \text{ in } \ \Omega,\\
-\operatorname{div}(\mathbf{u}\xi) + \sigma\xi &= \operatorname{div}(\zeta\mathbf{w}) + \sigma\mathfrak{d} \ \text{ in } \ \Omega,\\
\mathbf{w} = 0 \ \text{ on } \ \partial\Omega, &\quad \psi = 0 \ \text{ on } \ \Sigma_{\text{in}}, \quad \xi = 0 \ \text{ on } \ \Sigma_{\text{out}},\\
\omega - \Pi\omega = 0, &\quad n = \varkappa \int_\Omega \left(b^0_{31}\,\psi + b^0_{32}\,\omega + b^0_{34}\,\xi + b^0_{30}\,\mathfrak{d}\right) dx,
\end{aligned} \tag{1.42}$$

where $\mathfrak{d} = 1/2\,\operatorname{Tr}\,\mathbf{D}$ and the variable coefficients b^0_{ij} and the operators $\mathscr{C}_0$, $\mathscr{D}_0$, are defined by the formulae

$$\begin{aligned}
b^0_{11} &= \Psi[\vartheta] - \varrho H'(\varphi) + m - \frac{2\sigma}{\varrho_0}\varphi, \ \ b^0_{12} = \lambda^{-1}\varrho, \ \ b^0_{13} = \varrho,\\
b^0_{10} &= \varrho\Psi[\vartheta] - \frac{\sigma}{\varrho_0}\varphi^2 - \sigma\varphi + m\varrho, \ \ b^0_{21} = \frac{\sigma}{\varrho_0}\psi_0 + H'(\varphi),\\
b^0_{22} &= -\lambda^{-1}, \ \ b^0_{23} = -1, \ \ b^0_{20} = \sigma\varphi\varrho_0^{-1} - \Psi[\vartheta] - m,\\
b^0_{31} &= \varrho_0^{-1}\zeta\left(\Psi[\vartheta] - \varrho H'(\varphi) - \frac{2\sigma}{\varrho_0}\varphi\right) - H'(\varphi) + m\varrho_0^{-1}\zeta,\\
b^0_{32} &= (\lambda\varrho_0)^{-1}\varrho\zeta b^0_{12} + \lambda^{-1}, \ \ b^0_{34} = \varrho_0^{-1}\Psi_1[\vartheta] + m(1 + \varrho_0^{-1}\varphi),\\
b^0_{30} &= \varrho_0^{-1}\zeta(\Psi_1[\vartheta] - m\varrho) + \Psi[\vartheta] - m(1 - \zeta - \varrho_0^{-1}\zeta\varphi),
\end{aligned} \tag{1.43}$$

$$\mathscr{C}_0(\psi, \mathbf{w}) = R\psi\mathbf{u}\nabla\mathbf{u} + R\varrho\mathbf{w}\nabla\mathbf{u} + R\varrho\mathbf{u}\nabla\mathbf{w}, \tag{1.44}$$

$$\begin{aligned}
\mathscr{D}_0(\mathbf{D}) &= R\mathbf{u}\nabla(\mathbf{D}\mathbf{u}) + R\mathbf{D}^*(\mathbf{u}\nabla\mathbf{u})\\
&+ \operatorname{div}\left((\mathbf{D} + \mathbf{D}^*)\nabla\mathbf{u} - \frac{1}{2}\operatorname{Tr}\,\mathbf{D}\nabla\mathbf{u}\right) - \mathbf{D}^*\Delta\mathbf{u} - \Delta(\mathbf{D}\mathbf{u}).
\end{aligned} \tag{1.45}$$

The justification of the formal procedure meets the serious problems, since the smoothness of solutions to problem (1.13) is not sufficient for the well-posedness of problem (1.42) in the standard weak formulation. In order to cope with this difficulty we define *very weak solutions* to problem (1.42). The construction of such solutions is based on

the following lemma, and the proof is given in Appendix A. The lemma is given in $\mathbb{R}^d$, for our application $d = 3$.

LEMMA 1.7. *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with the Lipschitz boundary, let exponents s and r satisfy the inequalities $sr > d$, $1/2 \le s \le 1$, and $\varphi, \varsigma \in W^{s,r}(\Omega) \cap W^{1,2}(\Omega)$, $\mathbf{w} \in \mathcal{W}_0^{1-s,r'}(\Omega) \cap W_0^{1,2}(\Omega)$. Then there is a constant c depending only on s, r, and Ω, such that the trilinear form*

$$\mathfrak{B}(\mathbf{w}, \varphi, \varsigma) = -\int_\Omega \varsigma \mathbf{w} \cdot \nabla \varphi \, dx$$

satisfies the inequality

$$|\mathfrak{B}(\mathbf{w}, \varphi, \varsigma)| \le c \|\mathbf{w}\|_{\mathcal{W}_0^{1-s,r'}(\Omega)} \|\varphi\|_{W^{s,r}(\Omega)} \|\varsigma\|_{W^{s,r}(\Omega)} \tag{1.46}$$

and can be continuously extended to $\mathfrak{B} : \mathcal{W}_0^{1-s,r'}(\Omega)^d \times W^{s,r}(\Omega)^2 \mapsto \mathbb{R}$. In particular, we have $\varsigma \nabla \varphi \in \mathcal{W}^{s-1,r}(\Omega)$ and $\|\varsigma \nabla \varphi\|_{W^{1-s,r}(\Omega)} \le c \|\varphi\|_{W^{s,r}(\Omega)} \|\varsigma\|_{W^{s,r}(\Omega)}$.

DEFINITION 1.8. *The vector field $\mathbf{w} \in \mathcal{W}_0^{1-s,r'}(\Omega)^3$, functionals $(\omega, \psi, \xi) \in \mathbb{W}^{-s,r'}(\Omega)^3$, and constant n are said to be a weak solution to problem (1.42) if $\langle \omega, 1 \rangle = 0$ and the identity*

$$\begin{aligned} &\int_\Omega \mathbf{w}\Big(\mathbf{H} - R\varrho \nabla \mathbf{u} \cdot \mathbf{h} + R\varrho \nabla \mathbf{h}^* \mathbf{u}\Big) \, dx - \mathfrak{B}(\mathbf{w}, \varphi, \varsigma) - \mathfrak{B}(\mathbf{w}, \upsilon, \zeta) \\ &+ \big\langle \omega, G - b_{12}^0 \varsigma - b_{22}^0 g - \varkappa b_{32}^0 \big\rangle + \big\langle \psi, F - b_{11}^0 \varsigma - b_{21}^0 g - \varkappa b_{31}^0 - R\mathbf{u} \cdot \nabla \mathbf{u} \cdot \mathbf{h} \big\rangle \\ &+ \big\langle \xi, M - \varkappa b_{34}^0 \big\rangle + n\big(1 - \big\langle 1, b_{13}^0 \varsigma \big\rangle\big) \\ &= \big\langle \mathfrak{d}, b_{10}^0 \varsigma + b_{20}^0 g + \varkappa b_{30}^0 + \sigma \upsilon \big\rangle + \big\langle \mathscr{D}_0, \mathbf{h} \big\rangle \end{aligned} \tag{1.47}$$

holds true for all $(\mathbf{H}, G, F, M) \in (C^\infty(\Omega))^6$ such that $G = \Pi G$. Here $\mathfrak{d} = 1/2 \ Tr \, \mathbf{D}$, and the test functions $\mathbf{h}$, g, ς, υ are defined by the solutions to adjoint problems

$$\Delta \mathbf{h} - \nabla g = \mathbf{H}, \quad \operatorname{div} \mathbf{h} = G, \quad \mathscr{L}^* \varsigma = F, \quad \mathscr{L} \upsilon = M \ \text{ in } \ \Omega, \tag{1.48}$$

$$\mathbf{h} = 0 \ \text{ on } \ \partial\Omega, \quad \Pi g = g, \quad \varsigma = 0 \ \text{ on } \ \Sigma_{\text{out}}, \quad \upsilon = 0 \ \text{ on } \ \Sigma_{\text{in}}. \tag{1.49}$$

We are now in a position to formulate the third main result of this paper.

THEOREM 1.9. *Under the above assumptions,*

$$\begin{aligned} &\mathbf{w}_\varepsilon \to \mathbf{w} \ \text{ weakly in } \ \mathcal{W}_0^{1-s,r'}(\Omega), \quad n_\varepsilon \to n \ \text{ in } \ \mathbb{R}, \\ &\psi_\varepsilon \to \psi, \quad \omega_\varepsilon \to \omega, \quad \xi_\varepsilon \to \xi \ (*)\text{-weakly in } \ \mathbb{W}^{-s,r'}(\Omega) \ \text{ as } \ \varepsilon \to 0, \end{aligned} \tag{1.50}$$

where the limits, vector field $\mathbf{w}$, functionals ψ, ω, ξ, and the constant n are given by the weak solution to problem (1.42).

Note that the matrix $\mathbf{N}(\varepsilon)$ defined by equalities (1.5) meets all requirements of Theorem 1.9, and in the special case we have in representation (1.6)

$$\mathbf{D}(x) = \operatorname{div} \mathbf{T}(x) \, \mathbf{I} - \mathbf{T}'(x). \tag{1.51}$$

Therefore, Theorem 1.9, combined with formulae (1.3) and (1.10), implies the existence of the shape derivative for the drag functional at $\varepsilon = 0$. Straightforward calculations lead to the following result.

THEOREM 1.10. *Under the assumptions of Theorem* 1.9*, there exists the shape derivative*

$$\frac{d}{d\varepsilon} J_D(S_\varepsilon)\Big|_{\varepsilon=0} = L_e(\mathbf{T}) + L_u(\mathbf{w}, \omega, \psi),$$

where the linear forms L_e *and* L_u *are defined by the equalities*

$$L_e(\mathbf{T}) = \int_\Omega \operatorname{div}\mathbf{T}(\nabla\mathbf{u} + \nabla\mathbf{u}^* - \operatorname{div}\mathbf{u}\mathbf{I})\nabla\eta\mathbf{U}_\infty\,dx$$
$$- \int_\Omega \big[\nabla\mathbf{u} + \nabla\mathbf{u}^* - \operatorname{div}\mathbf{u} - q\mathbf{I} - R\varrho\mathbf{u}\otimes\mathbf{u}\big]\mathbf{D}\nabla\eta\cdot\mathbf{U}_\infty\,dx$$
$$- \int_\Omega \big[\mathbf{D}^*\nabla\mathbf{u} + \nabla\mathbf{u}^*\mathbf{D} - \nabla(\mathbf{D}\mathbf{u}) - \nabla(\mathbf{D}\mathbf{u})^*\big]\nabla\eta\cdot\mathbf{U}_\infty\,dx$$

and

$$L_u(\mathbf{w}, \omega, \psi) = \int_\Omega \mathbf{w}\big[\Delta\eta\mathbf{U}_\infty + R\varrho(\mathbf{u}\cdot\nabla\eta)\mathbf{U}_\infty + R\varrho(\mathbf{u}\cdot\mathbf{U}_\infty)\nabla\eta\big]\,dx$$
$$+ \big\langle\omega, \nabla\eta\cdot\mathbf{U}_\infty\big\rangle + R\big\langle\psi, (\mathbf{u}\cdot\nabla\eta)(\mathbf{u}\cdot\mathbf{U}_\infty)\big\rangle.$$

While L_e depends directly on the vector field $\mathbf{T}$, the linear form L_u depends on the weak solution $(\mathbf{w},\ \psi, \omega)$ to problem (1.42), and thus depends on the *direction* $\mathbf{T}$ in a very implicit manner, which is inconvenient for applications. In order to cope with this difficulty, we define the *adjoint state* $\mathbf{Y} = (\mathbf{h}, g, \varsigma, \upsilon, l)^\top$ given as a solution to the linear equation

$$\mathfrak{L}\mathbf{Y} - \mathfrak{U}\mathbf{Y} - \mathfrak{V}\mathbf{Y} = \Theta, \tag{1.52}$$

supplemented with boundary conditions (1.49). Here the operators $\mathfrak{L}$, $\mathfrak{U}$, $\mathfrak{V}$ and the vector field Θ are defined by

$$\mathfrak{L} = \begin{pmatrix} \Delta & -\nabla & 0 & 0 & 0 \\ \operatorname{div} & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathscr{L}^* & 0 & 0 \\ 0 & 0 & 0 & \mathscr{L} & 0 \\ 0 & 0 & -\mathbb{B}_{13} & 0 & 1 \end{pmatrix}, \quad \mathfrak{U} = \begin{pmatrix} 0 & 0 & -\nabla\varphi & -\zeta\nabla & 0 \\ 0 & 0 & \Pi_{12} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathfrak{V} = \begin{pmatrix} R\varrho(\nabla\mathbf{u} - \mathbf{u}\nabla) & 0 & 0 & 0 & 0 \\ 0 & -\lambda^{-1}\Pi & 0 & 0 & \varkappa\Pi b^0_{32} \\ R\mathbf{u}\cdot\nabla\mathbf{u} & b^0_{21} & b^0_{11} & 0 & \varkappa b^0_{31} \\ 0 & 0 & 0 & 0 & \varkappa b^0_{34} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\Theta = (\Delta\eta\mathbf{U}_\infty + R\varrho(\nabla\eta\otimes\mathbf{U}_\infty + \mathbf{U}_\infty\otimes\nabla\eta)\mathbf{u},\quad \Pi(\nabla\eta\cdot\mathbf{U}_\infty),\quad R(\mathbf{u}\nabla\eta)(\mathbf{u}\mathbf{U}_\infty), 0, 0)\,,$$
$$\Pi_{2i}(\cdot) = \Pi(b^0_{2i}(\cdot)),\quad \mathbb{B}_{13}(\cdot) = \big\langle 1, b^0_{13}(\cdot)\big\rangle.$$

The following theorem guarantees the existence of the adjoint state and gives the expression of the shape derivative for the drag functional in terms of the vector field $\mathbf{T}$.

THEOREM 1.11. *Let a given solution* $\vartheta \in \mathcal{B}_\tau$, $(\zeta, m) \in X^{s,r}\times\mathbb{R}$, *to problem* (1.13) *meet all requirements of Theorem* 1.6*. Then there exists positive constant* τ_1 *(depending only on* $\mathbf{U}$, Ω, *and* r, s*) such that if* $\tau \in (0, \tau_1]$ *and* R, $\lambda^{-1} \le \tau_1^2$, *then*

there exists a unique solution $\mathbf{Y} \in (Y^{s,r})^3 \times (X^{s,r})^3 \times \mathbb{R}$ *to problem* (1.52), (1.49). *The form* L_u *has the representation*

$$L_u(\mathbf{w},\psi,\omega) = \int_\Omega \big[\operatorname{div} \mathbf{T}\big(b^0_{10}\varsigma + b^0_{20} g + \sigma \upsilon + \varkappa b^0_{30}\big) + \mathscr{D}_0(\operatorname{div}\mathbf{T} - \mathbf{T}')\mathbf{h}\big]\, dx, \tag{1.53}$$

where the coefficients b^0_{ij} *and the operator* $\mathscr{D}_0$ *are defined by formulae* (1.43) *and* (1.45).

Method and structure of the paper. The following aspects of our method deserve a brief description:

- extended form (1.13) of the governing equations which allows us to cope with the mass control problem;
- the splitting of the boundary value problem for the transport equation into two parts: the local problem in the vicinity of inlet, and the global problem with the modified vector field $\tilde{\mathbf{u}}$ and the empty inlet $\tilde{\Sigma}_{\mathrm{in}}$;
- the estimates of solutions to the model problem (4.19) in the fractional Sobolev spaces, which cannot be obtained by the interpolation method;
- the very weak formulation of linearized equations introduced to assure the existence of material derivatives.

Now we can explain the organization of this paper. Section 2 is devoted to the proof of Theorem 1.6. First of all, we establish the existence of solutions to problem (1.13) using the Schauder fixed point theorem. Next we consider the linear equations for the difference of two solutions $(\mathbf{u}_i, \varphi_i)$, $i = 1, 2$, corresponding to arbitrary matrix-valued functions $\mathbf{N}_i$. Using Theorem 1.4 we deduce the weak formulation of the boundary value problem for linearized equations. The main result of this section is Theorem 2.3, which shows that solutions of the linearized problem are stable with respect to perturbations of data in the dual Sobolev space. This result implies the local uniqueness of solutions to problem (1.13). In section 3 we exploit Theorem 2.3 to prove the existence of the material derivative of solutions. The last section is devoted to the proof of Theorem 1.4.

2. Existence and uniqueness of local solutions. Proof of Theorem 1.6.

2.1. Existence theory. In this paragraph we establish the local solvability of problem (1.13) and prove the first part of Theorem 1.6. In our notation, c denotes generic constants, which are different in different places and depend only on Ω, $\mathbf{U}$, σ, and r, s. The proof is based on the following lemma which furnishes the regularity properties of composed functions. Let us consider functions $u, v : \Omega \mapsto B_K$, where $B_K = \{\mathbf{x} : |\mathbf{x}| \le K\} \subset \mathbb{R}^3$ is the ball of radius K centered at 0.

LEMMA 2.1. *Assume that* $u, v \in X^{s,r}$, $s \in (0, 1]$, $sr > 3$, *and* $\mathfrak{f} \in C^3(\Omega \times B_K)$. *Then we have*

$$\|\mathfrak{f}(\cdot, u)\|_{X^{s,r}} \le c(r,s)\|\mathfrak{f}\|_{C^1(\Omega\times B_K(0))}(1 + \|u\|_{X^{s,r}}), \tag{2.1}$$

$$\|\mathfrak{f}(\cdot, v) - \mathfrak{f}(\cdot, u)\|_{X^{s,r}} \le c(r,s)\|\mathfrak{f}\|_{C^2(\Omega\times B_K)}(1 + \|u\|_{X^{s,r}} + \|v\|_{X^{s,r}})\|u - v\|_{X^{s,r}}. \tag{2.2}$$

Proof. In order to prove (2.1) it suffices to note that

$$|\mathfrak{f}(x, u(x)) - \mathfrak{f}(y, u(y))|^r \le c(r)\|\mathfrak{f}\|^r_{C^1(\Omega\times B_K)}(|x - y|^r + |u(x) - u(y)|^r),$$

which, in view of the inequality

$$\int_{\Omega\times\Omega} |x - y|^{r-3-rs}\, dxdy \le c(r,s),$$

yields

$$|\mathfrak{f}(\cdot,u)|_{s,r,\Omega}\le c(r,s)\|\mathfrak{f}\|_{C^1(\Omega\times B_K(0))}(1+|u|_{s,r,\Omega}).$$

On the other hand, we have

$$\|\nabla\mathfrak{f}(\cdot,u)\|_{L^2(\Omega)}\le\|\mathfrak{f}\|_{C^1(\Omega\times B_K(0))}\|\nabla u\|_{L^2(\Omega)}.$$

Combining obtained inequalities we get (2.1). It remains to note that (2.2) follows from (2.1) and the Hadamard formula for the first order expansion of $\mathfrak{f}$. □

Fix sufficiently small positive τ such that

$$c_e\tau<\delta^*, \tag{2.3}$$

where δ^* is the constant determined in Corollary 1.5 and c_e is the constant from inequality (1.37). By virtue of Corollary 1.5, there is σ^*, depending only on Ω, $\mathbf{U}$, and r,s, such that, for all $\vartheta\in\mathcal{B}_\tau$ and $\sigma>\sigma^*$, problems (1.28) and (1.29) have solutions satisfying inequalities (1.35). Finally, fix an arbitrary $\sigma>\sigma^*$.

We solve problem (1.13) by an application of the Schauder fixed point theorem in the following framework. Choose an arbitrary element $\vartheta\in\mathcal{B}_\tau$. As mentioned above, the problem

$$\mathbf{u}\cdot\nabla\varphi_1+\sigma\varphi_1=\Psi_1[\vartheta]+m\mathfrak{g}\varrho\ \text{ in }\ \Omega,\ \ \varphi_1=0\ \text{ on }\ \Sigma_{\text{in}}, \tag{2.4}$$

has a unique solution satisfying the inequality

$$\|\varphi_1\|_{X^{s,r}}\le c(\Omega,U,\sigma,r,s)(\|\Psi_1[\vartheta]\|_{X^{s,r}}+|m|). \tag{2.5}$$

Next, define $\mathbf{u}_1$ and π_1 to be the solutions of the boundary value problem for the Stokes equations

$$\begin{aligned}\Delta\mathbf{u}_1-\nabla\pi_1&=\mathscr{A}(\mathbf{u})+R\mathscr{B}(\varrho,\mathbf{u},\mathbf{u})\equiv F[\vartheta]\ \text{ in }\ \Omega,\\ \varrho_0\operatorname{div}\mathbf{u}_1&=\Pi(\mathfrak{g}\sigma\varphi_1-\mathfrak{g}\varrho_0\Psi[\vartheta]-\mathfrak{g}m\varrho_0)\ \text{ in }\ \Omega,\\ \mathbf{u}_1&=0\ \text{ on }\ \partial\Omega,\quad \pi_1-\Pi\pi_1=0,\end{aligned} \tag{2.6}$$

where m is given by (1.13c). By Lemma 1.3, this problem admits a unique solution such that

$$\|\mathbf{u}_1\|_{Y^{s,r}}+\|\pi_1\|_{X^{s,r}}\le c(\|F[\vartheta]\|_{Z^{s,r}}+|\Psi[\vartheta]|_{X^{s,r}}+\|\varphi_1\|_{X^{s,r}}+|m|). \tag{2.7}$$

Equations (2.4), (2.6), (1.13c) define the mapping $\Xi:\vartheta\to\vartheta_1=(\mathbf{u}_1,\pi_1,\varphi_1)$. We claim that for a suitable choice of the constant τ, Ξ is a weakly continuous automorphism of the ball $\mathcal{B}_\tau$. We begin with the estimates for nonlinear operators present in (2.4). Fix an arbitrary $\vartheta\in\mathcal{B}_\tau$. Applying inequality (2.2) from Lemma 2.1 to the function H which is a part of $\Psi[\vartheta]$, we obtain $\|H(\varphi)\|_{X^{s,r}}\le c\tau^2$, which leads to the estimate

$$\|\Psi[\vartheta]\|_{X^{s,r}}\le\frac{c}{\lambda}(\|q_0\|_{C^1(\Omega)}+\|\pi\|_{X^{s,r}})+c\tau^2\le\frac{c}{\lambda}+c\tau^2\le c\tau^2. \tag{2.8}$$

Since, under assumptions of Theorem 1.6, $X^{s,r}(\Omega)$ is a Banach algebra and $\|\varrho\|_{X^{s,r}}\le c+\|\varphi\|_{X^{s,r}}\le const$, we conclude from this and (2.5) that

$$\|\varphi_1\|_{X^{s,r}}\le c/\lambda+c\tau^2+c|m|\le c\tau^2+c|m|. \tag{2.9}$$

In order to estimate the right-hand side of the first equation in (2.6) we introduce the vector function $\mathbf{z} = (\mathbf{u}, \nabla\mathbf{u}, \pi, \varphi)$ and proceed as follows. It can be easily seen that $\|\mathbf{z}\|_{X^{s,r}} \le \|\vartheta\|_E \le \tau$ and $|\mathbf{z}| \le c\tau$. Recall that the operator $\mathscr{B}$ constitutes a cubic polynomial of $\mathbf{u}$ and ϱ. By Lemma 2.1, we have

$$R\|\mathscr{B}(\varrho, \mathbf{u}, \mathbf{u})\|_{X^{s,r}} \le cR(1 + \|\varrho\|_{X^{s,r}} + \|\mathbf{z}\|_{X^{s,r}}) \le c\tau^2(1+\tau) \ \text{ in } \ \mathcal{B}_\tau. \tag{2.10}$$

Next note that

$$\|\mathscr{A}(\mathbf{u})\|_{Z^{s,r}} \le c(\|\mathfrak{g} - 1\|_{C^2(\Omega)} + \|\mathbf{N} - \mathbf{I}\|_{C^2(\Omega)})(1 + \|\mathbf{u}\|_{Y^{s,r}}) \le c\tau^2\|\mathbf{u}\|_{Y^{s,r}},$$

which along with (1.39) and (2.10) implies

$$\|F[\vartheta]\|_{Z^{s,r}} \le c\tau^2(1+\tau) \ \text{ in } \ \mathcal{B}_\tau. \tag{2.11}$$

Combining inequalities (2.8) and (2.9) we get the estimate

$$\|\sigma\varphi_1 + \Psi[\vartheta]\|_{X^{s,r}} \le c\tau^2.$$

From this, (2.11), (2.7), and Lemma 1.3 we finally obtain

$$\|\mathbf{u}_1\|_{Y^{s,r}} + |\pi_1|_{X^{s,r}} \le c\tau^2 + c|m|. \tag{2.12}$$

It remains to estimate m. Recall that the vector field $\mathbf{u}$ and parameter σ meet all requirements of Corollary 1.5. Therefore, problem (1.13b) has a unique solution $\zeta \in W^{s,r}(\Omega)$ for all s, r satisfying (1.38). In particular, inequalities (1.35) yield the estimate $\|\zeta\|_{X^{s,r}} \le c$. Since, by virtue of (1.38), the pair $s = 2/3$, $r = 6$ is admissible and the embedding $W^{2/3,6}(\Omega) \hookrightarrow C^{1/6}(\Omega)$ is bounded, estimates (1.33) and (1.34) for $rs > 3$ yield

$$\|\zeta\|_{C^{1/6}(\Omega)} + \|\zeta\|_{W^{1,2}(\Omega)} \le C(\mathbf{U}, \Omega, \sigma). \tag{2.13}$$

Recalling that $\operatorname{div}\mathbf{u} = \operatorname{div}\mathbf{u}$, we obtain $|\operatorname{div}\mathbf{u}| \le c_e\tau$. From this, the inequality $|\mathfrak{g}| \le 1 + c\tau^2$, and the maximum principle (1.34), we conclude that

$$\|\zeta\|_{C(\Omega)} \le (1 + c\tau^2)(1 - \sigma^{-1}c\tau)^{-1} \le (1 - c\tau)^{-1}, \tag{2.14}$$

which leads to the following estimate:

$$|1 - \zeta| \le c\tau(1 - c\tau)^{-1}.$$

Now we can estimate the right-hand side of (1.13c). Rewrite the first integral in the form

$$\int_\Omega \mathfrak{g}(1 - \zeta - \varrho_0^{-1}\zeta\varphi)\,dx = \int_\Omega (1-\zeta)^+\,dx + \int_\Omega (\mathfrak{g} - 1)(1 - \zeta - \varrho_0^{-1}\zeta\varphi)\,dx - \int_\Omega ((1-\zeta)^- + \varrho_0^{-1}\zeta\varphi)\,dx.$$

We have

$$|(\mathfrak{g} - 1)(1 - \zeta - \varrho_0^{-1}\zeta\varphi)| \le c\tau^2, \quad |(1-\zeta)^- + \varrho_0^{-1}\zeta\varphi)| \le c_e\tau + c\tau(1 - c\tau)^{-1}.$$

On the other hand, we have $\|(1-\zeta)^+\|_{C^{1/6}(\Omega)} \le c(\mathbf{U},\Omega,\sigma)$ and $(1-\zeta)^+ = 1$ on Σ_{out}. Hence

$$\int_\Omega (1-\zeta)^+\,dx > \kappa(\mathbf{U},\Omega,\sigma) > 0.$$

Thus, we get

$$\varkappa^{-1} \ge \kappa\bigl(1 - c\kappa^{-1}\tau(1-c\tau)^{-1}\bigr).$$

In particular, there is a positive τ_0 depending only on $\mathbf{U}$, Ω, and σ, such that

$$|\varkappa| \le c \quad \text{for all} \quad \tau \le \tau_0.$$

Repeating these arguments and using inequalities (2.8), (1.13c), we arrive at $|m| \le c\tau^2$. Combining this estimate with (2.9) and (2.12), we finally obtain $\|\vartheta_1\|_{X^{s,r}} \le c\tau^2$. Choose sufficiently small $\tau_0 = \tau_0(\mathbf{U},\Omega,\sigma)$ such that $c\tau_0^2 < \tau_0$. Thus, for all $\tau \le \tau_0$, Ξ maps the ball $\mathcal{B}_\tau$ into itself. Let us show that Ξ is weakly continuous. Choose an arbitrary sequence $\vartheta_n \in \mathcal{B}_\tau$ such that $\vartheta_n = (\mathbf{u}_n, \pi_n, \varphi_n)$ converges weakly in E to some ϑ. Since the ball $\mathcal{B}_\tau$ is closed and convex, ϑ belongs to $\mathcal{B}_\tau$. Let us consider the corresponding sequences of the elements $\vartheta_{1,n} = \Xi(\vartheta_n) \in \mathcal{B}_\tau$ and functions ζ_n. There are subsequences $\{\vartheta_{1,j}\} \subset \{\vartheta_{1,n}\}$ and $\{\zeta_j\} \subset \{\zeta_n\}$ such that $\vartheta_{1,j}$ converges weakly in E to some element $\vartheta_1 \in \mathcal{B}_\tau$ and ζ_j converges weakly in $X^{s,r}$ to some function $\zeta \in X^{s,r}$. Since the embedding $E \hookrightarrow C(\Omega)^5$ is compact, we have $\vartheta_n \to \vartheta$, $\vartheta_{1j} \to \vartheta_1$ in $C(\Omega)^5$, and

$$\nabla\zeta_j \rightharpoonup \nabla\zeta \ \text{ weakly in } \ L^2(\Omega), \ \ \zeta_j \to \zeta \ \text{ in } \ C(\Omega).$$

Substituting ϑ_j and $\vartheta_{1,j}$ into equations (2.4), (2.6), (1.13c) and letting $j \to \infty$, we obtain that the limits ϑ and ϑ_1 also satisfy (2.4), (2.6), (1.13c). Thus, we get $\vartheta_1 = \Xi(\vartheta)$. Since for given ϑ, a solution to equations (2.4), (2.6) is unique, we conclude from this that all weakly convergent subsequences of $\vartheta_{1,n}$ have the unique limit ϑ_1. Therefore, the whole sequence $\vartheta_{1,n} = \Xi(\vartheta_n)$ converges weakly to $\Xi(\vartheta)$. Hence the mapping $\Xi : \mathcal{B}_\tau \mapsto \mathcal{B}_\tau$ is weakly continuous, and, by virtue of the Schauder fixed point theory, there is $\vartheta \in \mathcal{B}(\tau)$ such that $\vartheta = \Xi(\vartheta)$.

It remains to prove that ϑ is given by a solution to problem (1.13a). For $\vartheta_1 = \vartheta$, the only difference between problems (1.13a) and (2.6), (1.13c) is the presence of the projector Π in the right-hand side of (2.6). Hence, it suffices to show that

$$\Pi(\varrho_0^{-1}\mathfrak{g}\sigma\varphi - \mathfrak{g}\Psi[\vartheta] - \mathfrak{g}m) = \varrho_0^{-1}\mathfrak{g}\sigma\varphi - \mathfrak{g}\Psi[\vartheta] - \mathfrak{g}m. \tag{2.15}$$

To this end we note that φ is a generalized solution to the transport equation

$$\mathbf{u}\cdot\nabla\varphi + \sigma\varphi = \Psi_1[\vartheta] + m\mathfrak{g}\varrho.$$

Using ζ as a test function and recalling the integral identity (1.30) we obtain

$$\sigma\int_\Omega \varphi\mathfrak{g}\,dx = \int_\Omega \zeta(\Psi_1[\vartheta] + m\mathfrak{g}\varrho)\,dx.$$

On the other hand, equality (1.13c) reads

$$\int_\Omega \zeta(\varrho_0^{-1}\Psi_1[\vartheta] + m\mathfrak{g}(1+\varphi\varrho_0^{-1}))\,dx = \int_\Omega (\mathfrak{g}\Psi[\vartheta] + \mathfrak{g}m)\,dx.$$

Combining these equalities and noting that $1+\varrho_0^{-1}\varphi=\varrho/\varrho_0$, we obtain

$$\int_\Omega \Big(\frac{\mathfrak{g}\sigma}{\varrho_0}\varphi-\mathfrak{g}\Psi[\vartheta]-m\mathfrak{g}\Big)\,dx=0$$

which yields (2.15), and the proof of Theorem 1.6 is completed. □

2.2. Uniqueness and stability. In this paragraph we prove that, under the assumptions of Theorem 1.6, a solution to problem (1.13) is unique, and we investigate in detail the dependence of the solution on the matrix function $\mathbf{N}$.

Weak formulation of linearized equations. Assume that matrices $\mathbf{N}_i$, $i=1,2$, satisfy conditions of Theorem 1.6, and denote by $(\vartheta_i,\zeta_i,m_i)\in E\times X^{s,r}\times\mathbb{R}$, $i=1,2$, the corresponding solutions to problem (1.13). Recall that the solutions $(\vartheta_i,\zeta_i,m_i)$, together with the constants $\varkappa_i$, satisfy the inequalities

$$|m_i|+\|\vartheta_i\|_E\le c\tau,\quad |\varkappa_i|+\|\zeta_i\|_{X^{s,r}}\le c, \tag{2.16}$$

where the constant c depends only on $\mathbf{U},\Omega$, r,s, and σ. We denote $\mathbf{u}_i=\mathbf{u}_0+\mathbf{u}_i$, $i=1,2$, the solutions to (1.9) for $\mathbf{N}_i$, $i=1,2$. Now set

$$\mathfrak{d}=\mathfrak{g}_1-\mathfrak{g}_2\equiv\sqrt{\det\,\mathbf{N}_1}-\sqrt{\det\,\mathbf{N}_2},$$
$$\mathbf{w}=\mathbf{u}_1-\mathbf{u}_2,\quad \omega=\pi_1-\pi_2,\quad \psi=\varphi_1-\varphi_2,,\ \ \xi=\zeta_1-\zeta_2,\ \ n=m_1-m_2.$$

It follows from (1.13) that

$$\begin{aligned}
\mathbf{u}_1\nabla\psi+\sigma\psi&=-\mathbf{w}\cdot\nabla\varphi_2+b_{11}\psi+b_{12}\omega+b_{13}n+b_{10}\mathfrak{d}\ \ \text{in}\ \ \Omega,\\
\Delta\mathbf{w}-\nabla\omega&=\mathscr{A}_1(\mathbf{w})+R\mathscr{C}_1(\psi,\mathbf{w})+\mathscr{D}\ \ \text{in}\ \ \Omega,\\
\operatorname{div}\mathbf{w}&=b_{21}\psi+b_{22}\omega+b_{23}n+b_{20}\mathfrak{d}\ \ \text{in}\ \ \Omega,\\
-\operatorname{div}(\mathbf{u}_1\xi)+\sigma\xi&=\operatorname{div}(\zeta_2\mathbf{w})+\sigma\mathfrak{d}\ \ \text{in}\ \ \Omega,\\
\mathbf{w}&=0\ \ \text{on}\ \ \partial\Omega,\ \ \psi=0\ \ \text{on}\ \ \Sigma_{\text{in}},\ \ \xi=0\ \ \text{on}\ \ \Sigma_{\text{out}},\\
\omega-\Pi\omega&=0,\quad n=\varkappa\int_\Omega\Big(b_{31}\psi+b_{32}\omega+b_{34}\xi+b_{30}\mathfrak{d}\Big)\,dx.
\end{aligned}\tag{2.17}$$

Here the coefficients are given by the formula

$$\begin{aligned}
b_{11}&=\sigma(1-\mathfrak{g}_2)+\mathfrak{g}_2\Psi[\vartheta_1]-\mathfrak{g}_2\varrho_2\Phi_1(\varphi_1,\varphi_2)+\mathfrak{g}_2m_2-\frac{\sigma\mathfrak{g}_2}{\varrho_0}(\varphi_1+\varphi_2),\\
b_{12}&=\lambda^{-1}\varrho_2\mathfrak{g}_2,\quad b_{13}=\mathfrak{g}_2\varrho_1,\quad b_{10}=\varrho_1\Psi[\vartheta_1]-\frac{\sigma}{\varrho_0}\varphi_1^2-\sigma\varphi_1+m_1\varrho_1,\\
b_{21}&=\mathfrak{g}_2\left(\frac{\sigma}{\varrho_0}+\Phi_1(\varphi_1,\varphi_2)\right),\quad b_{22}=-\mathfrak{g}_2/\lambda,\\
b_{23}&=-\mathfrak{g}_2,\quad b_{20}=\sigma\varphi_1\varrho_0^{-1}-\Psi[\vartheta_1]-m_2,\\
b_{31}&=\varrho_0^{-1}\zeta_1\Big(\sigma(1-\mathfrak{g}_2)+\mathfrak{g}_2\Psi[\vartheta_1]-\mathfrak{g}_2\varrho_2\Phi_1(\varphi_1,\varphi_2)\\
&\quad-\frac{\mathfrak{g}_2\sigma}{\varrho_0}(\varphi_1+\varphi_2)\Big)-\mathfrak{g}_2\Phi_1(\varphi_1,\varphi_2)+m_2\mathfrak{g}_1\varrho_0^{-1}\zeta_2,\\
b_{32}&=\varrho_0^{-1}\zeta_1b_{12}-b_{22},\quad b_{34}=\varrho_0^{-1}\Psi_1[\vartheta_2]+m_2\mathfrak{g}_1(1+\varrho_0^{-1}\varphi_1),\\
b_{30}&=\varrho_0^{-1}\zeta_1(b_{10}-m_1\varrho_1)+\Psi[\vartheta_1]-m_2(1-\zeta_2-\varrho_0^{-1}\zeta_2\varphi_2),\\
\Phi_1(\varphi_1,\varphi_2)&=(p'(\varrho_0)\varrho_0)^{-1}\sigma\int_0^1H'(\varphi_1s+\varphi_2(1-s))\,ds,
\end{aligned}\tag{2.18}$$

and the operators $\mathscr{C}_1$ and $\mathscr{D}$ are defined by the equalities

$$\mathscr{C}(\mathbf{w}) = \mathscr{B}_1(\psi, \mathbf{u}_1, \mathbf{u}_1) + \mathscr{B}_1(\varrho_2, \mathbf{w}, \mathbf{u}_1) + \mathscr{B}_1(\varrho_2, \mathbf{u}_2, \mathbf{w}),$$
$$\mathscr{D} = \mathscr{A}_1(\mathbf{u}_2) - \mathscr{A}_2(\mathbf{u}_2) + R(\mathscr{B}_1(\varrho_2, \mathbf{u}_2, \mathbf{u}_2) - \mathscr{B}_2(\varrho_2, \mathbf{u}_2, \mathbf{u}_2)),$$

where $\mathscr{A}_i$ and $\mathscr{B}_i$ are given by (1.8) with $\mathbf{N}_i$ instead of $\mathbf{N}$.

We consider $\mathscr{D}$ and $\mathfrak{d}$ as *given functions* and equality (2.17) as the system of equations and boundary conditions for unknowns $\mathbf{w}$, ψ, ξ, and n. The next step is crucial for further analysis. We replace equations (2.17) by an integral identity, which leads to the notion of a very weak solution to problem (2.17). To this end choose an arbitrary function $(\mathbf{H}, G, F, M) \in C^\infty(\Omega)^6$ such that $G - \Pi G = 0$, and consider the auxiliary boundary value problems

$$\mathscr{L}^*\varsigma = F, \quad \mathscr{L}\upsilon = M \ \text{ in } \ \Omega, \quad \varsigma = 0 \ \text{ on } \ \Sigma_{\text{out}}, \quad \upsilon = 0 \ \text{ on } \ \Sigma_{\text{in}}. \tag{2.19}$$

$$\Delta\mathbf{h} - \nabla g = \mathbf{H}, \quad \operatorname{div}\mathbf{h} = \Pi G \ \text{ in } \ \Omega, \quad \mathbf{h} = 0 \ \text{ on } \ \partial\Omega, \quad \Pi g = g. \tag{2.20}$$

Since, under the assumptions of Theorem 1.6, $\mathbf{u}$ and σ meet all requirements of Corollary 1.5, each of problems (2.19) has a unique solution such that

$$\|\varsigma\|_{W^{s,r}(\Omega)} \le c\|F\|_{W^{s,r}(\Omega)}, \quad \|\upsilon\|_{W^{s,r}(\Omega)} \le c\|M\|_{W^{s,r}(\Omega)}, \tag{2.21}$$

where c depends only on $\mathbf{U}$, Ω, r, s, and σ. On the other hand, by virtue of Lemma 1.3, problem (2.20) has a unique solution satisfying the inequality

$$\|\mathbf{h}\|_{W^{1+s,r}(\Omega)} + \|g\|_{W^{s,r}(\Omega)} \le c\|\mathbf{H}\|_{W^{1+s,r}(\Omega)} + c\|G\|_{W^{s,r}(\Omega)}. \tag{2.22}$$

Recall that $\mathbf{w} \in W^{2,2}(\Omega)^3 \cap C^1(\Omega)^3$ vanishes on $\partial\Omega$, and $(\omega, \psi, \xi) \in W^{1,2}(\Omega)^3 \cap C(\Omega)^3$. Multiplying both sides of the first equation in system (2.17) by ς, both sides of the fourth equation in (2.17) by υ, integrating the results over Ω, and using the Green formula for the Stokes equations, we obtain the system of integral equalities

$$\begin{aligned}
&\int_\Omega \psi F\,dx = \int_\Omega (-\mathbf{w}\cdot\nabla\varphi_2 + b_{11}\psi + b_{12}\omega + b_{13}n + b_{10}\mathfrak{d})\varsigma\,dx,\\
&\int_\Omega \mathbf{w}\mathbf{H}\,dx + \int_\Omega \omega G\,dx = \int_\Omega \big(b_{21}\psi + b_{22}\omega + b_{23}n + b_{20}\mathfrak{d}\big)g\,dx\\
&+ \int_\Omega (\mathscr{A}_1(\mathbf{w}) + R\mathscr{C}(\mathbf{w},\psi) + \mathscr{D})\mathbf{h}\,dx, \quad \int_\Omega \xi M\,dx = \int_\Omega (\operatorname{div}(\zeta_2\mathbf{w}) + \sigma\mathfrak{d})\upsilon\,dx.
\end{aligned} \tag{2.23}$$

Next, since $\operatorname{div}(\varrho_2\mathbf{u}_2) = 0$, we have

$$\begin{aligned}
&\int_\Omega (\mathscr{B}_1(\varrho_2, \mathbf{w}, \mathbf{u}_1) + \mathscr{B}_1(\varrho_2, \mathbf{u}_2, \mathbf{w}))\cdot\mathbf{h}\,dx\\
&\qquad = \int_\Omega \varrho_2\mathbf{w}\cdot\Big(\nabla(\mathbf{N}_1^{-1}\mathbf{u}_1)\cdot(\mathbf{N}_1^{-1}\mathbf{h}) - (\mathbf{N}_1^{-1})^*\nabla(\mathbf{N}_1^{-1}\mathbf{h})^*\mathbf{u}_2\Big)\,dx.
\end{aligned}$$

On the other hand, integration by parts gives

$$\int_\Omega \operatorname{div}(\zeta_2\mathbf{w})\upsilon\,dx = -\int_\Omega \zeta_2\mathbf{w}\nabla\upsilon\,dx.$$

Using these identities and recalling the duality pairing, we can collect relations (2.23), together with the expression for n, in one integral identity

$$\begin{aligned}&\int_\Omega \mathbf{w}\Big(\mathbf{H} - R\varrho_2\nabla(\mathbf{N}_1^{-1}\mathbf{u}_1)\cdot(\mathbf{N}_1^{-1}\mathbf{h}) + R\varrho_2(\mathbf{N}_1^{-1})^*\nabla(\mathbf{N}_1^{-1}\mathbf{h})^*\mathbf{u}_2\Big)\,dx\\ &\quad - \mathfrak{B}(\mathbf{w},\varphi_2,\varsigma) - \mathfrak{B}(\mathbf{w},\upsilon,\zeta_2) - \mathfrak{A}_1(\mathbf{w},\mathbf{h}) + \big\langle\omega, G - b_{12}\varsigma - b_{22}g - \varkappa b_{32}\big\rangle\\ &\quad + \big\langle\psi, F - b_{11}\varsigma - b_{21}g - \varkappa b_{31} - R\mathbf{u}_1\cdot\nabla(\mathbf{N}_1^{-1}\mathbf{u}_1)\cdot\mathbf{N}_1^{-1}\mathbf{h}\big\rangle\\ &+ \big\langle\xi, M - \varkappa b_{34}\big\rangle + n - n\big\langle 1, b_{13}\varsigma + b_{23}g\big\rangle = \big\langle\mathfrak{d}, b_{10}\varsigma + b_{20}g + \varkappa b_{30} + \sigma\upsilon\big\rangle + \big\langle\mathscr{D},\mathbf{W}\big\rangle.\end{aligned}\tag{2.24}$$

Here, the trilinear form $\mathfrak{B}$ and the bilinear form $\mathfrak{A}_1$ are defined by the equalities

$$\mathfrak{B}(\mathbf{w},\varphi_2,\varsigma) = -\int_\Omega \varsigma\mathbf{w}\cdot\nabla\varphi_2\,dx,\quad \mathfrak{A}_1(\mathbf{w},\mathbf{h}) = \int_\Omega \mathscr{A}_1(\mathbf{w})\cdot\mathbf{h}\,dx.$$

Note that relations (2.24) are well defined for all $\mathbf{w}\in\mathcal{W}_0^{1-s,r'}(\Omega)$ and $\psi,\xi\in\mathbb{W}^{-s,r'}(\Omega)$. It is obviously true for all terms, with the possible exception of $\mathfrak{A}_1$ and $\mathfrak{B}$. Well-posedness of the form $\mathfrak{B}$ follows from Lemma 1.7. The well-posedness of the form $\mathfrak{A}_1$ results from the following lemma; the proof is given in Appendix A.

LEMMA 2.2. *Let $sr>3$, $1/2\le s\le 1$, and $\mathbf{w}\in\mathcal{W}_0^{1-s,r'}(\Omega)\cap W_0^{1,2}(\Omega)$, $\mathbf{h}\in W^{1+s,r}(\Omega)$, and $\mathbf{N}_1$ satisfy* (1.39). *Then there is a constant c depending only on s,r, and Ω such that*

$$|\mathfrak{A}_1(\mathbf{w},\mathbf{h})|\le c\tau^2\|\mathbf{w}\|_{\mathcal{W}_0^{1-s,r'}(\Omega)}\|\mathbf{h}\|_{W^{1+s,r}(\Omega)}.\tag{2.25}$$

Consequently, the form $\mathfrak{A}_1$ can be continuously extended to $\mathfrak{A}_1 : \mathcal{W}_0^{1-s,r'}(\Omega)^3\times W^{1+s,r}(\Omega)^3\mapsto\mathbb{R}$.

Thus, relations (2.24) are well defined for all $(\mathbf{w},\psi,\omega,\xi)\in\mathcal{W}_0^{1-s,r'}(\Omega)^3\times\mathbb{W}^{-s,r'}(\Omega)^3$. Equalities (2.24) along with (2.19), (2.20) are called *the very weak formulation* of problem (2.17). The natural question is the uniqueness of solutions to such weak formulation. The following theorem, which is the main result of this section, guarantees the uniqueness of very weak solutions for sufficiently small τ.

THEOREM 2.3. *Let s,r, and σ satisfy condition* (1.38), *parameters λ,R and matrices $\mathbf{N}_i$, $i=1,2$, satisfy conditions* (1.39), *and constants τ meet all requirements of Theorem* 1.6. *Also let the solutions $(\vartheta_i,\zeta_i,m_i)$, $i=1,2$, to problem* (1.13) *with the matrices $\mathbf{N}_i$, $i=1,2$, belong to $\mathcal{B}_\tau\times X^{s,r}\times\mathbb{R}$. Furthermore, assume that, for any $(\mathbf{H},G,F,M)\in C^\infty(\Omega)^6$ and for $(\varsigma,\upsilon,\mathbf{h},g)$ satisfying* (2.19)–(2.20), *the elements $(\mathbf{w},\omega,\psi,\xi)\in\mathcal{W}_0^{1-s,r'}(\Omega)^3\times\mathbb{W}^{-s,r'}(\Omega)^3$ and the constant n satisfy identity* (2.24).

Then there are constants c, τ_1 depending only on s,r,σ, and Ω, $\mathbf{U}$ such that, for $\tau\in(0,\tau_1]$, we have

$$\begin{aligned}&\|\mathbf{w}\|_{\mathcal{W}_0^{1-s,r'}(\Omega)} + \|\omega\|_{\mathbb{W}^{-s,r'}(\Omega)} + \|\psi\|_{\mathbb{W}^{-s,r'}(\Omega)} + \|\xi\|_{\mathbb{W}^{-s,r'}(\Omega)} + |n|\\ &\qquad\le c(\|\mathscr{D}\|_{L^1(\Omega)} + \|\mathfrak{d}\|_{\mathbb{W}^{-s,r'}(\Omega)}).\end{aligned}\tag{2.26}$$

Proof. The proof is based upon two auxiliary lemmas; the first lemma establishes the bounds for coefficients of problem (2.17).

LEMMA 2.4. *Under the assumptions of Theorem* 2.3, *all the coefficients of identity* (2.24) *satisfy the inequalities $\|b_{ij}\|_{X^{s,r}}\le c$; furthermore*

$$\begin{aligned}\|b_{12}\|_{X^{s,r}} + \|b_{22}\|_{X^{s,r}} + \|b_{11}\|_{X^{s,r}} + \|b_{10}\|_{X^{s,r}} + \|b_{20}\|_{X^{s,r}}\le c\tau,\\ \|b_{31}\|_{X^{s,r}} + \|b_{32}\|_{X^{s,r}} + \|b_{34}\|_{X^{s,r}}\le c\tau.\end{aligned}\tag{2.27}$$

Proof. The proof follows from Lemma 2.1 combined with formula (2.18). ∎

In order to formulate the second auxiliary result we introduce the following denotations:

$$\mathfrak{I}_1 = \langle \psi, b_{11}\varsigma\rangle + \langle \omega, b_{12}\varsigma\rangle + \langle \mathfrak{d}, b_{10}\varsigma\rangle, \quad \mathfrak{I}_2 = \langle \psi, b_{21}g\rangle + \langle \omega, b_{22}g\rangle + \langle \mathfrak{d}, b_{20}g\rangle,$$
$$\mathfrak{I}_3 = \varkappa\big(\langle \psi, b_{31}\rangle + \langle \omega, b_{32}\rangle + \langle \xi, b_{34}\rangle + \langle \mathfrak{d}, b_{30}\rangle\big), \quad \mathfrak{I}_4 = \Big\langle \psi, \mathbf{u}_2\nabla(\mathbf{N}_1\mathbf{u}_1)\cdot\mathbf{N}_1^{-1}\mathbf{h}\Big\rangle,$$
$$\mathfrak{I}_5 = \int_\Omega \varrho_2\mathbf{w}\cdot\Big(\nabla(\mathbf{N}_1^{-1}\mathbf{u}_1)\cdot(\mathbf{N}_1^{-1}\mathbf{h}) - (\mathbf{N}_1^{-1})^*\nabla(\mathbf{N}_1^{-1}\mathbf{h})^*\mathbf{u}_2\Big)\,dx,$$
$$\mathfrak{G} = \|\mathbf{w}\|_{\mathcal{W}_0^{1-s,r'}(\Omega)} + \|\psi\|_{\mathbb{W}^{-s,r'}(\Omega)} + \|\omega\|_{\mathbb{W}^{-s,r'}(\Omega)} + \|\xi\|_{\mathbb{W}^{-s,r'}(\Omega)},$$
$$\mathfrak{Q} = \|\mathbf{H}\|_{\mathcal{W}^{s-1,r}(\Omega)} + \|G\|_{W^{s,r}(\Omega)} + \|F\|_{W^{s,r}(\Omega)} + \|M\|_{W^{s,r}(\Omega)}.$$

LEMMA 2.5. *Under the assumptions of Theorem* 2.3*, there is a constant c, depending only on $\mathbf{U}$, Ω, s, r, and σ, such that*

$$\mathfrak{I}_1 \le c\tau\mathfrak{Q}\big[\mathfrak{G} + \|\mathfrak{d}\|_{\mathbb{W}^{-s,r'}(\Omega)}\big], \tag{2.28}$$
$$\mathfrak{I}_2 \le c\mathfrak{Q}\big[\tau\mathfrak{G} + \|\psi\|_{\mathbb{W}^{-s,r'}(\Omega)} + \|\mathfrak{d}\|_{\mathbb{W}^{-s,r'}(\Omega)}\big], \tag{2.29}$$
$$\mathfrak{I}_3 \le c\tau\big[\mathfrak{G} + \|\mathfrak{d}\|_{\mathbb{W}^{-s,r'}(\Omega)}\big], \quad \mathfrak{I}_4 + \mathfrak{I}_5 \le c\mathfrak{Q}\mathfrak{G}. \tag{2.30}$$

Proof. We have

$$\langle \psi, b_{11}\varsigma\rangle + \langle \omega, b_{12}\varsigma\rangle + \langle \mathfrak{d}, b_{10}\varsigma\rangle \le \|b_{11}\varsigma\|_{W^{s,r}(\Omega)}\|\psi\|_{\mathbb{W}^{-s,r'}(\Omega)} + \|b_{12}\varsigma\|_{W^{s,r}(\Omega)}\|\omega\|_{\mathbb{W}^{-s,r'}(\Omega)} + \|b_{10}\varsigma\|_{W^{s,r}(\Omega)}\|\mathfrak{d}\|_{\mathbb{W}^{-s,r'}(\Omega)}.$$

Recall that, for $rs > 3$, $W^{s,r}(\Omega)$ is a Banach algebra. From this, estimate (2.21), and inequalities (2.27), we obtain

$$\begin{aligned}
&\|b_{11}\varsigma\|_{W^{s,r}(\Omega)}\|\psi\|_{\mathbb{W}^{-s,r'}(\Omega)} + \|b_{12}\varsigma\|_{W^{s,r}(\Omega)}\|\omega\|_{\mathbb{W}^{-s,r'}(\Omega)} + \|b_{10}\varsigma\|_{W^{s,r}(\Omega)}\|\mathfrak{d}\|_{\mathbb{W}^{-s,r'}(\Omega)}\\
&\quad\le c\|\varsigma\|_{W^{s,r}(\Omega)}(\|b_{11}\|_{W^{s,r}(\Omega)}\|\psi\|_{\mathbb{W}^{-s,r'}(\Omega)} + \|b_{12}\|_{W^{s,r}(\Omega)}\|\omega\|_{\mathbb{W}^{-s,r'}(\Omega)}\\
&\qquad\qquad + \|b_{10}\|_{W^{s,r}(\Omega)}\|\mathfrak{d}\|_{\mathbb{W}^{-s,r'}(\Omega)})\\
&\quad\le c\tau\|F\|_{W^{s,r}(\Omega)}(\|\psi\|_{\mathbb{W}^{-s,r'}(\Omega)} + \|\omega\|_{\mathbb{W}^{-s,r'}(\Omega)} + \|\mathfrak{d}\|_{\mathbb{W}^{-s,r'}(\Omega)}),
\end{aligned}$$

which gives (2.28). Repeating these arguments and using inequality (2.21), we obtain the estimates for $\mathfrak{I}_2$ and $\mathfrak{I}_3$. Next we have

$$\begin{aligned}
\|\mathbf{u}_2\cdot\nabla(\mathbf{N}_1^{-1}\mathbf{u}_1)\cdot\mathbf{N}_1^{-1}\mathbf{h}\|_{W^{s,r}(\Omega)} &\le c\|\mathbf{u}_2\|_{W^{s,r}(\Omega)}\|\mathbf{u}_1\|_{W^{1+s,r}(\Omega)}\|\mathbf{h}\|_{W^{1+s,r}(\Omega)}\\
&\le c\|\mathbf{H}\|_{W^{1+s,r}(\Omega)},
\end{aligned}$$

which gives the estimate for $\mathfrak{I}_4$. Since the embeddings $W^{s,r}(\Omega)\hookrightarrow C(\Omega)$, $W^{1+s,r}(\Omega)\hookrightarrow C^1(\Omega)$ are bounded and $\|\mathbf{N}^{\pm1}\|_{C^1(\Omega)} \le c$, we have

$$\varrho_2|\nabla(\mathbf{N}_1^{-1}\mathbf{u}_1)||\mathbf{N}_1^{-1}\mathbf{h}| + \varrho_2|(\mathbf{N}_1^{-1}\mathbf{u}_2)||\nabla(\mathbf{N}_1^{-1}\mathbf{h})| \le c\|\mathbf{h}\|_{W^{1+s,r}(\Omega)},$$

which leads to the inequality

$$\mathfrak{I}_5 \le c\|\mathbf{h}\|_{W^{1+s,r}(\Omega)}\|\mathbf{w}\|_{L^1(\Omega)} \le c(\|\mathbf{H}\|_{\mathcal{W}^{s-1,r}(\Omega)} + \|G\|_{W^{s,r}(\Omega)})\|\mathbf{w}\|_{\mathcal{W}^{1-s,r'}(\Omega)},$$

and the proof of Lemma 2.5 is completed. ∎

Let us return to the proof of Theorem 2.3. It follows from the duality principle that the theorem is proved provided we show that, under the assumptions of Theorem 1.6, the following inequality holds:

$$\begin{aligned}&\sup_{\mathfrak{Q}(\mathbf{H},G,F,M)=1}\big(\big\langle\mathbf{w},\mathbf{H}\big\rangle+\big\langle\omega,G\big\rangle+\big\langle\psi,F\big\rangle+\big\langle\xi,M\big\rangle\big)+|n|\\&\le c\tau\left(\mathfrak{G}\big(\mathbf{w},\omega,\psi,\xi\big)+|n|\right)+c\big(\|\mathscr{D}\|_{L^1(\Omega)}+\|\mathfrak{d}\|_{\mathbb{W}^{-s,r'}(\Omega)}\big),\end{aligned}\tag{2.31}$$

where the constant c depends only on $\Omega,\mathbf{U}$ and r,s,σ. Therefore, our task is to estimate step by step all terms in the left-hand side of (2.31). We begin with an estimate for the term $\langle\psi,F\rangle$. To this end, take $\mathbf{H}=\mathbf{h}=0$, $G=g=0$, $M=\upsilon=0$, and rewrite identity (2.24) in the form

$$\big\langle\psi,F\big\rangle=\mathfrak{B}(\mathbf{w},\varphi_2,\varsigma)+\mathfrak{I}_1+\mathfrak{I}_3+n\big\langle1,b_{13}\varsigma\big\rangle-n.$$

By virtue of Lemma 1.7 and estimate (2.22), we have

$$\mathfrak{B}(\mathbf{w},\varphi_2,\varsigma)\le c\tau\|\mathbf{w}\|_{\mathcal{W}_0^{1-s,r'}(\Omega)}\|\varsigma\|_{W^{s,r}(\Omega)}\le c\tau\|\mathbf{w}\|_{W^{1-s,r'}(\Omega)}\|F\|_{W^{s,r}(\Omega)}.\tag{2.32}$$

On the other hand, Lemma 2.4 and inequality (2.21) yield $|\big\langle1,b_{13}\varsigma\big\rangle|\le c\|F\|_{W^{s,r}(\Omega)}$. From this and (2.28), (2.30) we finally obtain

$$\big\langle\psi,F\big\rangle\le|n|+c\|F\|_{W^{s,r}(\Omega)}\left[\,\tau\mathfrak{G}+\|\mathfrak{d}\|_{\mathbb{W}^{-s,r'}(\Omega)}+|n|\,\right].\tag{2.33}$$

Moreover, by virtue of the duality principle

$$\|\psi\|_{\mathbb{W}^{-s,r'}(\Omega)}=\sup_{\|F\|_{W^{s,r}(\Omega)}=1}\big|\big\langle\psi,F\big\rangle\big|,$$

we have the following estimate for ψ:

$$\|\psi\|_{\mathbb{W}^{-s,r'}(\Omega)}\le c\tau\mathfrak{G}+c\|\mathfrak{d}\|_{\mathbb{W}^{-s,r'}(\Omega)}+c|n|.\tag{2.34}$$

Let us estimate $\mathbf{w}$ and ω. Substituting $F=\varsigma=0$ and $M=\upsilon=0$ into (2.24) we obtain

$$\big\langle\mathbf{w},\mathbf{H}\big\rangle+\big\langle\omega,G\big\rangle=\mathfrak{A}_1(\mathbf{w},\mathbf{h})+\mathfrak{I}_2+\mathfrak{I}_3+R\mathfrak{I}_4+R\mathfrak{I}_5+n\big\langle1,b_{23}g\big\rangle-n+\big\langle\mathscr{D},\mathbf{h}\big\rangle.$$

By virtue of Lemma 2.2 and (2.22), the first term in the right-hand side is bounded:

$$|\mathfrak{A}_1(\mathbf{w},\mathbf{h})|\le c\tau^2(\|\mathbf{H}\|_{\mathcal{W}^{s-1,r}(\Omega)}+\|G\|_{W^{s,r}(\Omega)})\|\mathbf{w}\|_{\mathcal{W}_0^{1-s,r'}(\Omega)}.$$

Next we have

$$|n\big\langle1,b_{23}g\big\rangle|\le c(\|\mathbf{H}\|_{\mathcal{W}^{s-1,r}(\Omega)}+\|G\|_{W^{s,r}(\Omega)})|n|.$$

Obviously

$$|\big\langle\mathscr{D},\mathbf{W}\big\rangle|\le c\|\mathscr{D}\|_{L^1(\Omega)}\|\mathbf{h}\|_{C(\Omega)}\le c(\|\mathbf{H}\|_{\mathcal{W}^{s-1,r}(\Omega)}+\|G\|_{W^{s,r}(\Omega)})\|\mathscr{D}\|_{L^1(\Omega)}.$$

These inequalities together with estimates (2.29)–(2.30) and inequality $R\le\tau^2$ imply

$$\begin{aligned}&\big\langle\mathbf{w},\mathbf{H}\big\rangle+\big\langle\omega,G\big\rangle\le|n|+c\tau\mathfrak{Q}\mathfrak{G}\\&+c\mathfrak{Q}\Big(\|\psi\|_{\mathbb{W}^{-s,r'}(\Omega)}+|n|+\|\mathfrak{d}\|_{\mathbb{W}^{-s,r'}(\Omega)}+\|\mathscr{D}\|_{L^1(\Omega)}\Big).\end{aligned}$$

Combining this result with (2.34), we obtain

$$\big\langle \mathbf{w}, \mathbf{H}\big\rangle + \big\langle \omega, G\big\rangle \le |n| + c\tau\mathfrak{Q}\mathfrak{G} + c\mathfrak{Q}\big(|n| + \|\mathfrak{d}\|_{\mathbb{W}^{-s,r'}(\Omega)} + \|\mathscr{D}\|_{L^1(\Omega)}\big), \tag{2.35}$$

where $\mathfrak{Q} = \mathfrak{Q}(\mathbf{H}, G, 0, 0)$. For $G = 0$ and by the duality principle

$$\|\mathbf{w}\|_{\mathcal{W}_0^{1-s,r'}(\Omega)} = \sup_{\|\mathbf{H}\|_{\mathcal{W}^{s-1,r}(\Omega)}=1} \big\langle \mathbf{H}, \mathbf{w}\big\rangle,$$

we conclude that

$$\|\mathbf{w}\|_{\mathcal{W}_0^{1-s,r'}(\Omega)} \le |n| + c\tau\mathfrak{G} + c\Big(|n| + \|\mathfrak{d}\|_{\mathbb{W}^{-s,r'}(\Omega)} + \|\mathscr{D}\|_{L^1(\Omega)}\Big). \tag{2.36}$$

Next, substituting $\mathbf{H} = \mathbf{h} = 0$, $G = g = F = \varsigma = 0$ into identity (2.24), we arrive at

$$\big\langle \xi, M\big\rangle = \mathfrak{B}(\mathbf{w}, \zeta_2, \upsilon) + \mathfrak{I}_3 + \sigma\big\langle \mathfrak{d}, \upsilon\big\rangle - n.$$

Lemma 1.7 and (2.21) give the estimate for the first term:

$$|\mathfrak{B}(\mathbf{w}, \zeta_2, \upsilon)| \le c\|\mathbf{w}\|_{\mathcal{W}_0^{1-s,r'}(\Omega)}\|\upsilon\|_{W^{s,r}(\Omega)} \le c\|\mathbf{w}\|_{\mathcal{W}_0^{1-s,r'}(\Omega)}\|M\|_{W^{s,r}(\Omega)}.$$

From this and estimates (2.30), (2.21), we obtain

$$\big\langle \xi, M\big\rangle \le c\tau\mathfrak{Q}\mathfrak{G} + c\mathfrak{Q}(\|\mathbf{w}\|_{\mathcal{W}_0^{1-s,r'}(\Omega)} + \|\mathfrak{d}\|_{\mathbb{W}^{-s,r'}(\Omega)}) + |n|.$$

Combining this result with inequality (2.36), we arrive at

$$\big\langle \xi, M\big\rangle \le c\mathfrak{Q}(\tau\mathfrak{G} + |n| + \|\mathfrak{d}\|_{\mathbb{W}^{-s,r'}(\Omega)} + \|\mathscr{D}\|_{L^1(\Omega)}) + c|n|. \tag{2.37}$$

Finally, choosing all test functions in (2.24) equal to 0, we obtain $n = \mathfrak{I}_3$, which together with (2.30) yields

$$|n| \le c\tau\mathfrak{Q}(\mathfrak{G} + \|\mathfrak{d}\|_{\mathbb{W}^{-s,r'}(\Omega)}). \tag{2.38}$$

From (2.33), (2.35), (2.37), combined with (2.38), it follows (2.31), and the proof of Theorem 2.3 is completed. □

Uniqueness of solutions. The important consequence of Theorem 2.3 is the following result on uniqueness of solutions to problem (1.13).

PROPOSITION 2.6. *Under the assumptions of Theorem* 1.6*, there exists a positive* τ_0 *such that, for all* $\tau \in (0, \tau_0]$*, problem* (1.13) *admits a unique solution in the ball* $\mathcal{B}_\tau$*.*

Proof. If for some $\mathbf{N}$ the problem has two distinct solutions $(\vartheta_i, \zeta_i, m_i)$, $i = 1, 2$, with $\vartheta_i \in \mathcal{B}_\tau$, then the corresponding *finite differences* of the solutions $\mathbf{w}$, ψ, ω, and ξ meet all requirements of Theorem 2.3 with $\mathfrak{d} = 0$ and $\mathscr{D} = 0$. Therefore, in view of (2.26) all the elements $\mathbf{w}$, ψ, ω, and ξ are equal to 0, which completes the proof. □

3. Proofs of Theorems 1.9 and 1.11.

Proof of Theorem 1.9. Let us consider a family of matrices $\mathbf{N}(\varepsilon)$ having representation (1.6) and the sequence of corresponding solutions $(\vartheta(\varepsilon), \zeta(\varepsilon), m(\varepsilon))$ to problem (1.13), where $\vartheta(\varepsilon) = (\mathbf{u}(\varepsilon), \pi(\varepsilon), \varphi(\varepsilon))$. By virtue of (1.41), we can assume that, possibly after passing to a subsequence, the sequence $(\vartheta(\varepsilon), \zeta(\varepsilon), m(\varepsilon))$ converges weakly in $(Y^{s,r})^3 \times (X^{s,r})^3 \times \mathbb{R}$ to some element (ϑ, ζ, m), which satisfies equations (1.13) with $\mathbf{N} = \mathbf{I}$ and meets all requirements of Theorem 1.6. Since the solution (ϑ, ζ, m) to problem (1.13) is unique, the limit is independent of the choice of a subsequence,

and the whole sequence converges to the limit (ϑ, ζ, m). It follows from (2.17) that the differences $\vartheta - \vartheta(\varepsilon)$, $\zeta - \zeta(\varepsilon)$, $m - m(\varepsilon)$ satisfy (2.17), with the coefficients $b_{ij}^{(\varepsilon)}$ and the operator $\mathscr{D}_\varepsilon$ given by formula (2.18) with

$$\mathbf{N}_1 = \mathbf{I}, \mathbf{N}_2 = \mathbf{N}(\varepsilon), (\vartheta_1, \zeta_1, m_1) = (\vartheta, \zeta, m), (\vartheta_2, \zeta_2, m_2) = (\vartheta(\varepsilon), \zeta(\varepsilon), m(\varepsilon)).$$

In particular, the operator $\mathscr{D}_\varepsilon$ is defined by the equality

$$\begin{aligned}\mathscr{D}_\varepsilon &= R\varrho(\varepsilon)\Big(\mathbf{u}(\varepsilon)\nabla\mathbf{u}(\varepsilon) - (\mathbf{N}(\varepsilon)^*)^{-1}\big(\mathbf{u}(\varepsilon)\nabla\big(\mathbf{N}(\varepsilon)^{-1}\mathbf{u}(\varepsilon)\big)\big)\Big)\\ &+ (\mathbf{N}(\varepsilon)^*)^{-1}\operatorname{div}\big(\mathfrak{g}^{-1}(\varepsilon)\mathbf{N}(\varepsilon)\mathbf{N}(\varepsilon)^*\nabla(\mathbf{N}(\varepsilon)^{-1}\mathbf{u}(\varepsilon))\big) - \Delta\mathbf{u}(\varepsilon)\end{aligned}$$

and admits the representation $\mathscr{D}_\varepsilon = \varepsilon\mathscr{D}_0(\mathfrak{d}, \mathbf{D}) + \varepsilon^2\mathscr{D}_1(\varepsilon)$, where $\mathscr{D}_0$ is given by (1.45). Moreover, since the norms $\|\varrho(\varepsilon)\|_{C(\Omega)}$ and $\|\mathbf{u}(\varepsilon)\|_{W^{2,2}(\Omega)}$ are uniformly bounded, we have

$$\|\mathscr{D}_1(\varepsilon)\|_{L^2(\Omega)} \le c(\mathbf{U}, \Omega, \sigma). \tag{3.1}$$

Next note that $\mathfrak{g}(\varepsilon)$ admits the decomposition $\mathfrak{g}(\varepsilon) = 1 + \varepsilon\mathfrak{d} + \varepsilon^2\mathfrak{d}_1(\varepsilon)$, where $\mathfrak{d} = \operatorname{Tr}\mathbf{D}$ and the reminder $\mathfrak{d}_1(\varepsilon)$ is uniformly bounded in $C^1(\Omega)$. Proceeding as in the previous section and recalling the equalities $\mathscr{A}_1 = \mathfrak{A}_1 = 0$, we conclude that the finite differences

$$(\mathbf{w}_\varepsilon, \omega_\varepsilon, \psi_\varepsilon) = \varepsilon^{-1}(\vartheta - \vartheta(\varepsilon)), \quad \xi_\varepsilon = \varepsilon^{-1}(\zeta - \zeta(\varepsilon)), \quad n_\varepsilon = \varepsilon^{-1}(m - m(\varepsilon))$$

satisfy the integral identity

$$\begin{aligned}&\int_\Omega \mathbf{w}_\varepsilon\Big(\mathbf{H} - \varrho(\varepsilon)\nabla\mathbf{u}\cdot\mathbf{h} + R\varrho(\varepsilon)\nabla\mathbf{h}^*\mathbf{u}(\varepsilon)\Big)\,dx\\ &- \mathfrak{B}(\mathbf{w}_\varepsilon, \varphi(\varepsilon), \varsigma) - \mathfrak{B}(\mathbf{w}_\varepsilon, \upsilon, \zeta(\varepsilon)) + \big\langle\omega_\varepsilon, G - b_{12}^{(\varepsilon)}\varsigma - b_{22}^{(\varepsilon)}g - \varkappa b_{32}^{(\varepsilon)}\big\rangle\\ &+ \big\langle\psi_\varepsilon, F - b_{11}^{(\varepsilon)}\varsigma - b_{21}^{(\varepsilon)}g - \varkappa b_{31}^{(\varepsilon)} - R\mathbf{u}\cdot\nabla\mathbf{u}\cdot\mathbf{h}\big\rangle + \big\langle\xi_\varepsilon, M - \varkappa b_{34}^{(\varepsilon)}\big\rangle + n_\varepsilon\\ &- n_\varepsilon\big\langle 1, b_{13}^{(\varepsilon)}\varsigma + b_{23}^{(\varepsilon)}g\big\rangle = \big\langle\mathfrak{d} + \varepsilon\mathfrak{d}_1(\varepsilon), b_{10}^{(\varepsilon)}\varsigma + b_{20}^{(\varepsilon)}g + \varkappa b_{30}^{(\varepsilon)} + \sigma\upsilon\big\rangle + \big\langle\mathscr{D}_0 + \varepsilon\mathscr{D}_1(\varepsilon), \mathbf{h}\big\rangle,\end{aligned} \tag{3.2}$$

along with the orthogonality conditions $\big\langle\omega_\varepsilon, 1\big\rangle = 0$. Here $(\mathbf{H}, G, F, M)$ are arbitrary smooth functions such that $G = \Pi G$, and the test functions ς, υ, g, and $\mathbf{h}$ are defined by equations (2.19), (2.20); therefore, the test functions are independent of ε. Recall that the elements $\vartheta_1 = \vartheta$ and $\vartheta_2 = \vartheta(\varepsilon)$ belong to the ball $\mathcal{B}_\tau$ and meet all requirements of Theorem 2.3. Hence, there exists $\tau_1 > 0$, depending only on $\Omega, \mathbf{U}$ and s, r, σ, such that the conditions

$$\lambda^{-1}, \quad R \le \tau^2, \quad \|\mathbf{N}(\varepsilon) - \mathbf{I}\|_{C^1(\Omega)} \le \tau^2, \quad 0 < \tau \le \tau_1,$$

imply

$$\begin{aligned}\|\mathbf{w}_\varepsilon\|_{\mathcal{W}_0^{1-s,r'}(\Omega)} + \|\omega_\varepsilon\|_{\mathbb{W}^{-s,r'}(\Omega)} + \|\psi_\varepsilon\|_{\mathbb{W}^{-s,r'}(\Omega)} + \|\xi_\varepsilon\|_{\mathbb{W}^{-s,r'}(\Omega)} + |n_\varepsilon|\\ \le c(\|\mathscr{D}_0 + \varepsilon\mathscr{D}_1\|_{L^1(\Omega)} + \|\mathfrak{d} + \varepsilon\mathfrak{d}_1\|_{C^2(\Omega)}) \le c.\end{aligned}$$

Therefore, after possibly passing to a subsequence, we can assume that the sequence $\mathbf{w}_\varepsilon$ converges to $\mathbf{w}$ weakly in $\mathcal{W}_0^{1-s,r'}(\Omega)$, and $(\omega_\varepsilon, \psi_\varepsilon, \xi_\varepsilon)$ converge to (ω, ψ, ξ) $(*)$-weakly in $\mathbb{W}^{-s,r'}(\Omega)$ as $\varepsilon \to 0$.

Next, choose $s' > s$ satisfying conditions (1.38). By virtue of Theorem 1.6, there exists $\tau_0' > 0$ (depending only on Ω, $\mathbf{U}$, r, s, σ) such that, for all $\tau \in (0, \tau_0']$, the functions $(\vartheta(\varepsilon), \zeta(\varepsilon))$ are bounded in $(Y^{s',r})^3 \times (X^{s',r})^3$. It follows from this that the family $\mathbf{u}(\varepsilon) = \mathbf{u}_0 + \mathbf{u}(\varepsilon)$ converges to $\mathbf{u}$ strongly in $Y^{s,r}$, and $(\varphi(\varepsilon), \pi(\varepsilon), \zeta(\varepsilon))$ converges to (φ, π, ζ) strongly in $X^{s,r}$. Therefore, by virtue of Lemma 2.1, the sequence $b_{ij}^{(\varepsilon)}$ converges strongly in $X^{s,r}$ to b_{ij}^0. Hence, we can pass to the limit in (3.2). It is easy to see that the limits, the vector field $\mathbf{w}$, and the functionals ψ, ω, ξ are given by a unique weak solution to problem (1.42) and, in addition, meet all requirements of Definition 1.8. It remains to note that, by virtue of Theorem 2.3, the limit is independent of the choice of a subsequence, which completes the proof of Theorem 1.9. ∎

Proof of Theorem 1.11. Assume that r, s, σ, and τ satisfy inequalities (1.31), (1.38), and that $\vartheta = (\mathbf{u}, \pi, \varphi) \in \mathcal{B}_\tau$ is a solution to problem (1.13) given by Theorem 1.6. Denote by $Y_0^{s,r}$ the subspace of the space $Y^{s,r}$ of all functions vanishing on $\partial\Omega$, by $X_{\text{in}}^{s,r}$ and $X_{\text{out}}^{s,r}$ the subspaces of $X^{s,r}$ which consist of all functions vanishing on Σ_{in} and Σ_{out}, respectively, and by $X_\Pi^{s,r}$ the subspace of all function in $X^{s,r}$ having the zero mean value. Introduce the Banach spaces $\mathcal{E} = (Y_0^{s,r})^3 \times X_\Pi^{s,r} \times X_{\text{out}}^{s,r} \times X_{\text{in}}^{s,r} \times \mathbb{R}$ and $\mathcal{F} = (Z^{s,r})^3 \times X_\Pi^{s,r} \times (X^{s,r})^2 \times \mathbb{R}$. Our first task is to show that, for all $\Theta \in \mathcal{F}$, problem (1.52), (1.49) has a unique solution $Y \in \mathcal{E}$. We begin with the observation that, by virtue of Lemma 1.3, the Stokes operator has the bounded inverse

$$\begin{pmatrix} \Delta & -\nabla \\ \operatorname{div} & 0 \end{pmatrix}^{-1} : (Z^{s,r})^3 \times X_\Pi^{s,r} \to (Y_0^{s,r})^3 \times X_\Pi^{s,r}.$$

On the other hand, by virtue of Corollary 1.5, the operators $\mathscr{L}$ and $\mathscr{L}^*$ from (1.48) have the bounded inverses $\mathscr{L}^{-1} : X^{s,r} \to X_{\text{in}}^{s,r}$, $(\mathscr{L}^*)^{-1} : X^{s,r} \to X_{\text{out}}^{s,r}$. Therefore, there exists the bounded operator

$$\begin{pmatrix} \mathscr{L}^* & 0 & 0 \\ 0 & \mathscr{L} & 0 \\ -\mathbb{B}_{1,3} & 0 & 1 \end{pmatrix}^{-1} : (X^{s,r})^2 \times \mathbb{R} \to X_{\text{out}}^{s,r} \times X_{\text{in}}^{s,r} \times \mathbb{R}.$$

It follows from this that, for all $\Theta \in \mathcal{F}$, the equation $\mathfrak{L}\mathbf{Y} = \Theta$ has a unique solution satisfying boundary conditions (1.49) and the inequality $\|\mathbf{Y}\|_{\mathcal{E}} \le c\|\Theta\|_{\mathcal{F}}$, where the constant c is independent of τ. Let us consider the operators $\mathfrak{U}$. By virtue of Lemma 1.7, we have

$$\|\varsigma\nabla\varphi\|_{\mathcal{W}^{s-1,r}(\Omega)} \le c\tau\|\varsigma\|_{X^{s,r}}.$$

It is easy to see that

$$\|\varsigma\nabla\varphi\|_{L^2(\Omega)} \le c\|\varsigma\|_{X^{s,r}}\|\nabla\varphi\|_{L^2(\Omega)} \le c\|\varsigma\|_{X^{s,r}}\|\varphi\|_{X^{s,r}} \le c\tau\|\varsigma\|_{X^{s,r}}.$$

Combining the obtained estimates we get the inequality $\|\varsigma\nabla\varphi\|_{Z^{s,r}} \le c\tau\|\varsigma\|_{X^{s,r}}$. Repetition of these arguments gives the inequality $\|\zeta\nabla v\|_{Z^{s,r}} \le c\|v\|_{X^{s,r}}$. Since the norms $\|b_{ij}^0\|_{X^{s,r}}$ are uniformly bounded, we conclude from this that $\|\mathfrak{U}\mathbf{Y}\|_{\mathcal{F}} \le c\|Y\|_{\mathcal{E}}$. Finally, let us consider the operator $\mathfrak{V}$. Since the space $X^{s,r}$ is the commutative Banach algebra and $\nabla\mathbf{u} \in X^{s,r}$, we have

$$\|R\mathbf{u}\nabla\mathbf{u}\mathbf{h}\|_{X^{s,r}} \le c\tau^2\|\mathbf{h}\|_{X^{s,r}}, \quad \|R\varrho(\nabla\mathbf{u} + \mathbf{u}\nabla)\mathbf{h}\|_{X^{s,r}} \le c\tau^2\|\mathbf{h}\|_{Y^{s,r}}.$$

On the other hand, by virtue of Lemma 2.1 and (1.43), the coefficients b_{ij}^0 in the expression for $\mathfrak{V}$ satisfy the inequalities

$$\|b_{12}^0\|_{X^{s,r}} + \|b_{11}^0\|_{X^{s,r}} + \|b_{31}^0\|_{X^{s,r}} + \|b_{32}^0\|_{X^{s,r}} + \|b_{34}^0\|_{X^{s,r}} \le c\tau,$$

which yield the estimate $\|\mathfrak{V}\mathbf{Y}\|_{\mathcal{F}} \le c\tau\|Y\|_{\mathcal{E}}$. Thus we get that the diagonal matrix operator $\mathfrak{L}$ has the bounded inverse, $\mathfrak{U}$ is the bounded upper triangular (with respect to $\mathfrak{L}$) matrix operator, and $\mathfrak{V}$ is the small bounded operator. Hence, for all sufficiently small τ the operator $\mathfrak{L} - \mathfrak{U} - \mathfrak{V} : \mathcal{E} \to \mathcal{F}$ has the bounded inverse, which implies the existence of an adjoint state satisfying (1.52) and boundary conditions (1.49).

It remains to prove identity (1.53). Fix the adjoint state $\mathbf{Y} = (\mathbf{h}, g, \varsigma, \upsilon, l)$, and set

$$\mathbf{H} = \Delta\mathbf{h} - \nabla g, \quad G = \operatorname{div}\mathbf{h}, \quad F = \mathscr{L}^*\varsigma, \quad M = \mathscr{L}\upsilon.$$

It follows from (1.52) that

$$\begin{gathered}
\mathbf{H} - R\varrho(\nabla\mathbf{u} - \mathbf{u}\nabla)\mathbf{h} + \varsigma\nabla\varphi + \zeta\nabla\upsilon = \Delta\eta\mathbf{U}_\infty + R\varrho\big((\mathbf{u}\nabla\eta)\mathbf{U}_\infty + (\mathbf{u}\mathbf{U}_\infty)\nabla\eta\big), \\
G - \Pi(b^0_{21}\varsigma + b^0_{22}g - \varkappa b^0_{32}l) = \Pi(\nabla\eta\mathbf{U}_\infty), \\
F - R\mathbf{u}\nabla\mathbf{u}\mathbf{h} - b^0_{12}g - b^0_{11}\varsigma - \varkappa b^0_{31}l = (\mathbf{u}\nabla\eta)(\mathbf{u}\mathbf{U}_\infty), \quad M = \varkappa b^0_{34}l.
\end{gathered} \tag{3.3}$$

By virtue of Theorem 1.9, the *material* derivative $(\mathbf{w}, \omega, \psi, \xi, n)$ satisfies the integral identities (1.47). On the other hand, $(F, \mathbf{H}, G, M)$ together with the components of the adjoint state $\mathbf{Y}$ can be regarded as a collection of test functions for this identity. Substituting these test functions into (1.47), using equalities (3.3), and recalling the identity $\langle\omega, 1\rangle = 0$, we obtain

$$\begin{gathered}
L_u(\mathbf{w}, \omega, \psi) + \varkappa(l-1)\Big(\big\langle\psi, b^0_{31}\big\rangle + \big\langle\omega, b^0_{32}\big\rangle + \big\langle\xi, b^0_{34}\big\rangle\Big) \\
+ n - n\big\langle 1, b^0_{13}\varsigma\big\rangle = \big\langle\mathfrak{d}, b^0_{10}\varsigma + b^0_{20}g + \varkappa b^0_{30} + \sigma\upsilon\big\rangle + \big\langle\mathscr{D}_0, \mathbf{W}\big\rangle.
\end{gathered}$$

It follows from (1.52) that $l = \big\langle 1, b^0_{13}\varsigma\big\rangle$, which leads to

$$\begin{aligned}
L_u(\mathbf{w}, \omega, \psi) + \varkappa(l-1)\Big(\varkappa\big(\big\langle\psi, b^0_{31}\big\rangle + \big\langle\omega, b^0_{32}\big\rangle + \big\langle\xi, b^0_{34}\big\rangle\big) - n\Big) \\
= \big\langle\mathfrak{d}, b^0_{10}\varsigma + b^0_{20}g + \varkappa b^0_{30} + \sigma\upsilon\big\rangle + \big\langle\mathscr{D}_0, \mathbf{W}\big\rangle.
\end{aligned} \tag{3.4}$$

Next note that the identities (1.42) imply the following expression for the constant n:

$$n = \varkappa\big(\big\langle\psi, b^0_{31}\big\rangle + \big\langle\omega, b^0_{32}\big\rangle + \big\langle\xi, b^0_{34}\big\rangle + \big\langle\mathfrak{d}, b^0_{30}\big\rangle\big).$$

Substituting this equality into (3.4) and noting that $\mathfrak{d} = \operatorname{Tr}\mathbf{D}$, we obtain (1.53), which completes the proof. □

4. Proof of Theorem 1.4. Our strategy is the following. First we use the classical method of characteristics (see [8] for instance) to show that in the vicinity of each point $P \in \Sigma_{\text{in}} \cup \Gamma$ there exist normal coordinates (y_1, y_2, y_3) such that $\mathbf{u}\nabla_x = \mathbf{e}_1\nabla_y$. Hence the problem of existence of solutions to the transport equation in the neighborhood of $\Sigma_{\text{in}} \cap \Gamma$ is reduced to a boundary value problem for the model equation $\partial_{y_1}\varphi + \sigma\varphi = f$ in a parabolic domain. Next we prove that the boundary value problem for the model equation admits a unique solution in fractional Sobolev space, which leads to the existence and uniqueness of solutions in the neighborhood of the inlet set. Using the existence of a local solution we can reduce problem (1.28) to a boundary value problem for the modified equation, which does not require any boundary data. Application of well-known results on solvability of elliptic-hyperbolic equations in the case $\Gamma = \emptyset$ finally gives the existence and uniqueness of solutions to problems (1.28) and (1.29).

First we introduce some notation which will be used throughout this section. For any $a > 0$ we denote by Q_a the cube $[-a, a]^3$ and by Q_a^+ the slab $[-a, a]^2 \times [0, a]$ in the space of points $y = (y_1, y_2, y_3) \in \mathbb{R}^3$. We will write Y instead of (y_2, y_3) so that $y = (y_1, Y)$.

DEFINITION 4.1. *A standard parabolic neighborhood associated with the constant c_0 is a compact subset of the slab Q_a^+, defined by the inequalities*

$$\mathcal{P}_a = \{y = (y_1, Y) \in Q_a^+ : a^-(Y) \le y_1 \le a^+(Y)\}, \tag{4.1}$$

where $a^\pm : [-a, a] \times [0, a] \mapsto \mathbb{R}$ are continuous, piecewise C^1-functions satisfying the inequalities

$$\begin{gathered} -a \le a^-(Y) \le 0 \le a^+(Y) \le a, \\ -c_0\sqrt{y_3} \le a^-(Y) \le a^+(Y) \le c_0\sqrt{y_3}, \\ |\partial_{y_2} a^\pm(Y)| \le c_0, \quad |\partial_{y_3} a^\pm(Y)| \le c_0/\sqrt{y_3}. \end{gathered} \tag{4.2}$$

Set $Q_{\text{in}} = \{Y : a^-(Y) > -a\}$ and $Q_{\text{out}} = \{Y : a^+(Y) < a\}$. Denote by Σ^y_{in} and Σ^y_{out} the surfaces determined by the relations

$$\begin{aligned} \Sigma^y_{\text{in}} &= \{y : Y \in Q_{\text{in}}, \quad y_1 = a^-(Y)\}, \\ \Sigma^y_{\text{out}} &= \{y : Y \in Q_{\text{out}}, \quad y_1 = a^+(Y)\}. \end{aligned}$$

It is clear that $\partial\mathcal{P}_a = (\partial Q_a \cap \partial\mathcal{P}_a) \cup \Sigma^y_{\text{in}} \cup \Sigma^y_{\text{out}}$.

LEMMA 4.2. *Assume that the C^2-manifold $\Sigma = \partial B$ and the vector field $\mathbf{U} \in C^2(\Sigma)^3$ satisfy conditions (**H**1)–(**H**3). Let $\mathbf{u} \in C^1(\mathbb{R}^3)^3$ be a compactly supported vector field such that*

$$\mathbf{u} = \mathbf{U} \ \text{ on } \ \Sigma, \quad \mathbf{u} = 0 \ \text{ on } \ S.$$

Denote $M = \|\mathbf{u}\|_{C^1(\mathbb{R}^3)}$. Then there are positive constants a, c, C, ρ_c, and R_c, depending only on M, $\partial\Omega$, and $\mathbf{U}$, with the following properties:

(**P1**) *For any point $P \in \Gamma$ there exists a mapping $y \to \mathbf{x}(y)$ which takes diffeomorphically the cube Q_a onto a neighborhood $\mathcal{O}_P$ of P and satisfies the equation*

$$\partial_{y_1}\mathbf{x}(y) = \mathbf{u}(\mathbf{x}(y)) \ \text{ in } \ Q_a \tag{4.3}$$

and the inequalities

$$\|\mathbf{x}\|_{C^1(Q_a)} + \|\mathbf{x}^{-1}\|_{C^1(\mathcal{O}_P)} \le C, \quad |\mathbf{x}(y)| \le C|y|. \tag{4.4}$$

(**P2**) *There is a standard parabolic neighborhood $\mathcal{P}_a$ associated with the constant c such that*

$$\mathbf{x}(\mathcal{P}_a) = \mathcal{O}_P \cup \Omega, \quad \mathbf{x}(\Sigma^y_{\text{in}}) = \Sigma_{\text{in}} \cap \mathcal{O}_P, \quad \mathbf{x}(\Sigma^y_{\text{out}}) = \Sigma_{\text{out}} \cap \mathcal{O}_P. \tag{4.5}$$

(**P3**) *Denote by $G_a \subset \mathcal{P}_a$ the domain*

$$G_a = \{y = (y_1, Y) \in \mathcal{P}_a : Y \in Q_{in}\}, \tag{4.6}$$

and by $B_P(\rho)$ the ball $|x - P| \le \rho$. Then we have the inclusions

$$B_P(\rho_c) \cap \Omega \subset \mathbf{x}(G_a) \subset \mathcal{O}_P \cap \Omega \subset B_P(R_c) \cap \Omega. \tag{4.7}$$

Proof. We start with the proof of (**P1**). Recall condition (**H1**) and fix the standard Cartesian coordinate system (x_1, x_2, x_3) associated with the point $P \in \Gamma$. Let us consider the Cauchy problem.

$$\begin{aligned} \partial_{y_1}\mathbf{x} &= \mathbf{u}(\mathbf{x}(y)) \ \text{ in } \ Q_a, \\ x_1(y) = \Upsilon(y_2), \quad x_2(y) &= y_2 \quad \text{for } \ y_1 = 0, \\ x_3 &= F(\Upsilon(y_2), y_2) + y_3 \quad \text{for } \ y_1 = 0. \end{aligned} \tag{4.8}$$

Without any loss of generality we can assume that $0 < a < k < 1$. For any such a, problem (4.8) has a unique solution of class $C^1(Q_a)$. Denote by $\mathfrak{F}(y) = D_y\mathbf{x}(y)$ the Jacobian matrix function. The calculations show that

$$\mathfrak{F}_0 := \mathfrak{F}(y)\Big|_{y_1=0} = \begin{pmatrix} u_1 & \Upsilon'(y_2) & 0 \\ u_2 & 1 & 0 \\ u_3 & \partial_{y_2}F(\Upsilon(y_2), y_2) & 1 \end{pmatrix}, \quad \mathfrak{F}(0) = \begin{pmatrix} U & \Upsilon'(0) & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

which implies

$$\|\mathfrak{F}(0)^{\pm 1}\| \le C/3, \quad \|\mathfrak{F}_0(y) - \mathfrak{F}(0)\| \le ca, \tag{4.9}$$

where the constants C, c are independent of a.

Differentiation of (4.8) leads to the ordinary differential equation for $\mathfrak{F}$

$$\partial_{y_1}\mathfrak{F} = D_y\mathbf{u}(\mathbf{x})\mathfrak{F}, \quad \mathfrak{F}\Big|_{y_1=0} = \mathfrak{F}_0.$$

From this we get

$$\partial_{y_1}\|\mathfrak{F} - \mathfrak{F}_0\| \le M(\|\mathfrak{F} - \mathfrak{F}_0\| + \|\mathfrak{F}_0\|),$$

and hence $\|\mathfrak{F} - \mathfrak{F}_0\| \le c(M)\|\mathfrak{F}_0\|a$. Combining this result with (4.9) we finally arrive at

$$\|\mathfrak{F}(y) - \mathfrak{F}(0)\| \le ca. \tag{4.10}$$

This inequality along with the implicit function theorem implies the existence of $a > 0$, depending only on M and Ω, such that the mapping $x = \mathbf{x}(y)$ takes diffeomorphically the cube Q_a onto some neighborhood of the point P and satisfies inequalities (4.4).

Let us turn to the proof of (**P2**). We begin with the observation that the manifold $\mathbf{x}^{-1}(\partial\Omega \cap \mathcal{O})$ is defined by the equation

$$\Phi_0(y) := x_3(y) - F(x_1(y), x_2(y)) = 0, \quad y \in Q_a.$$

Let us show that Φ_0 is strictly monotone in y_3 and has the opposite signs on the faces $y_3 = \pm a$. To this end note that the formula for $\mathfrak{F}(0)$ along with (4.10) implies the estimates

$$|\partial_{y_3}x_3(y) - 1| + |\partial_{y_3}x_1(y)| + |\partial_{y_3}x_2(y)| \le ca \ \text{ in } \ Q_a.$$

Thus, we get

$$1 - ca \le \partial_{y_3}\Phi_0(y) = \partial_{y_3}x_3(y) - \partial_{x_i}F(x_1, x_2)\partial_{y_3}x_i(y) \le 1 + ca.$$

It follows from (4.10) that, for $y_3 = 0$, we have $|x_3(y)| \le ca|y|$, which along with (4.4) yields the estimate

$$|\Phi_0(y)| \le |x_3(y)| + |F(x(y))| \le ca|y| + KC|y|^2 \le ca^2 \ \text{ for } \ y_3 = 0.$$

Hence there is a positive a, depending only on M and Ω, such that the inequalities

$$1/2 \le \partial_{y_3}\Phi_0(y) \le 2, \quad \pm\Phi_0(y_1, y_2, \pm a) > 0$$

hold true for all $y \in Q_a$. Therefore, the equation $\Phi_0(y) = 0$ has a unique solution $y_3 = \Phi(y_1, y_2)$ in the cube Q_a. Moreover, the function $\Phi \in C^1([-a, a]^2)$ vanishes for $y_1 = y_3 = 0$. Thus, we get

$$\mathcal{P}_a := \mathbf{x}^{-1}(\mathcal{O} \cap \Omega) = \{\Phi(y_1, y_2) < y_3 < a, \quad |y_1|, |y_2| \le a\}.$$

Note that $|\mathbf{u}(\mathbf{x}(y)) - U\mathbf{e}_1| \le M|\mathbf{x}(y)| \le Ca$. Therefore, we can choose $a = a(M, \Omega)$ such that $2U/3 \le u_1 \le 4U/3$ and $C|u_2| \le U/3$ in Q_a. Recall that $x_1(y) - \Upsilon(x_2(y))$ vanishes at the plane $y_1 = 0$ and

$$\partial_{y_1}\big[x_1(y) - \mathfrak{g}(x_2(y))\big] = u_1(y) - \Upsilon'(x_2(y))u_2(y).$$

We obtain from this that, for a suitable choice of a,

$$|y_1|U/3 \le \big|x_1(y) - \Upsilon(x_2(y))\big| \le |y_1|5U/3 \ \text{ for } \ y \in Q_a. \tag{4.11}$$

Equations (4.8) imply the identity

$$\partial_{y_1}\Phi_0(y) \equiv \nabla F_0(\mathbf{x}(y)) \cdot \mathbf{u}(x(y)) = \nabla F_0(\mathbf{x}(y)) \cdot \mathbf{U}(x(y)) \quad \text{for} \ \ \Phi_0(y) = 0.$$

Combining this result with (1.17) and (4.11), we obtain the estimates

$$|y_1|N^-U/3 \le |\partial_{y_1}\Phi_0(y)| \le |y_1|N^+U5/3,$$

which along with the identity

$$\partial_{y_1}\Phi = -\partial_{y_1}\Phi_0(\partial_{y_3}\Phi_0)^{-1}$$

yield the inequalities

$$\begin{aligned} -c < \partial_{y_1}\Phi(y_1, y_2) \le cy_1 \ \ &\text{for} \ \ -a < y_1 < 0, \\ cy_1 < \partial_{y_1}\Phi(y_1, y_2) \le c \ \ &\text{for} \ \ 0 < y_1 < a, \\ |\partial_{y_2}\Phi(y_1, y_2) \le c, 0 \le \Phi(y_1, y_2) \le cy_1^2&. \end{aligned} \tag{4.12}$$

It is clear that for sufficiently small a, depending only on $\mathbf{U}$ and Ω, the functions $\Phi^{\pm}(y_2) = \Phi(\pm a, y_2)$ admit the estimates $ca^2 \le \Phi^{\pm}(y_2) < a$. Set

$$\begin{aligned} Q_{\text{in}} &= \{Y \in [-a, a] \times [0, a] : \ 0 < y_3 < \Phi^-(y_2)\}, \\ Q_{\text{out}} &= \{Y :\in [-a, a] \times [0, a] : \ 0 < y_3 < \Phi^+(y_2)\}. \end{aligned}$$

It follows from (4.12) that for every $Y \in Q_{in}$ ($Y \in Q_{\text{out}}$) the equation $y_3 = \Phi(y_1, y_2)$ has a unique solution $a^-(Y) < 0$ ($a^+(Y) > 0$). We adopt the convention that $a^{\pm}(Y) = \pm a$ for $y_3 > \Phi^{\pm}(y_2)$. It remains to note that, by virtue of (4.12), the functions $a^{\pm}$ meet all requirements of Lemma 4.2. □

The next lemma shows the existence of the normal coordinates in the vicinity of points of the inlet Σ_{in}.

LEMMA 4.3. *Let vector fields* $\mathbf{u}$ *and* $\mathbf{U}$ *meet all requirements of Lemma* 4.2 *and* $U_n = -\mathbf{U}(P) \cdot \mathbf{n} > N > 0$. *Then there is* $b > 0$, *depending only on* N, Σ, *and* $M = \|\mathbf{u}\|_{C^1(\Omega)}$, *with the following properties. There exists a mapping* $y \to \mathbf{x}(y)$,

which takes diffeomorphically the cube $Q_b = [-b,b]^3$ *onto a neighborhood* $\mathcal{O}_P$ *of* P *and satisfies the equations*

$$\partial_{y_3}\mathbf{x}(y) = \mathbf{u}(\mathbf{x}(y)) \ \text{ in } \ Q_b, \quad \mathbf{x}(y_1,y_2,0) \in \Sigma \cap \mathcal{O}_P \ \ \textit{for} \quad |y_2| \le a, \tag{4.13}$$

and the inequalities

$$\|\mathbf{x}\|_{C^1(Q_b)} + \|\mathbf{x}^{-1}\|_{C^1(\mathcal{O}_P)} \le C_{M,N}, \quad |\mathbf{x}(y)| \le C_M|y|, \tag{4.14}$$

where $C_{M,N} = 3(1+N^{-1})(M^2+2)^{1/2}$. *The inclusions*

$$B_P(\rho_i) \cap \Omega \subset \mathbf{x}(Q_b \cap \{y_3 > 0\}) \subset B_P(R_i) \cap \Omega \tag{4.15}$$

hold true for $\rho_i = C_{M,N}^{-1} b$ *and* $R_i = C_{M,N} b$.

Proof. The proof simulates the proof of Lemma 4.2. Choose the local Cartesian coordinates (x_1,x_2,x_3) centered at P such that in new coordinates $\mathbf{n}(P) = \mathbf{e}_3$. By the smoothness of Σ, there is a neighborhood $\mathcal{O} = [-k,k]^2 \times [-t,t]$ such that the manifold $\Sigma \cap \mathcal{O}$ is defined by the equation

$$x_3 = F(x_1,x_2), \quad F(0,0) = 0, \quad |\nabla F(x_1,x_2)| \le K(|x_1| + |x_2|).$$

The constants k, t, and K depend only on Σ. Let us consider the initial value problem

$$\partial_{y_3}\mathbf{x} = \mathbf{u}(\mathbf{x}(y)) \ \text{ in } \ Q_a, \quad \mathbf{x}\Big|_{y_3=0} = (y_1,y_2,F(y_1,y_2)). \tag{4.16}$$

Without any loss of generality we can assume that $0 < b < k < 1$. It follows from (**H1**) that for any such b problem (4.16) has a unique solution of class $C^1(Q_b)$. Next note that for $y_3 = 0$ we have

$$|\mathbf{x}(y)| \le (K+1)|y|, \quad |\mathbf{u}(\mathbf{x}(y)) - \mathbf{u}(0)| \le M(K+1)|y|. \tag{4.17}$$

Denote by $\mathfrak{F}(y) = D_y\mathbf{x}(y)$. The calculations show that

$$\mathfrak{F}_0 := \mathfrak{F}(y)\Big|_{y_3=0} = \begin{pmatrix} 1 & 0 & u_1 \\ 0 & 1 & u_2 \\ 0 & 0 & u_3 \end{pmatrix}, \quad \mathfrak{F}(0) = \begin{pmatrix} 1 & 0 & u_1(P) \\ 0 & 1 & u_2(P) \\ 0 & 0 & U_n \end{pmatrix},$$

which along with (4.17) implies

$$\|\mathfrak{F}(0)^{\pm 1}\| \le C_{M,N}/3, \quad \|\mathfrak{F}_0(y) - \mathfrak{F}(0)\| \le cb. \tag{4.18}$$

Next, differentiation of (4.16) with respect to y leads to the equation

$$\partial_{y_1}\mathfrak{F} = D_y\mathbf{u}(\mathbf{x})\mathfrak{F}, \quad \mathfrak{F}\Big|_{y_3=0} = \mathfrak{F}_0.$$

Arguing as in the proof of Lemma 4.2 we obtain $\|\mathfrak{F} - \mathfrak{F}_0\| \le c(M)\|\mathfrak{F}_0\|b$. Combining this result with (4.18), we finally arrive at $\|\mathfrak{F}(y) - \mathfrak{F}(0)\| \le cb$. From this and the implicit function theorem we conclude that there is positive b, depending only on M and Σ, such that the mapping $x = \mathbf{x}(y)$ takes diffeomorphically the cube Q_b onto some neighborhood of the point P and satisfies inequalities (4.14). Inclusions (4.15) easily follow from (4.14). ∎

Model equation. Let us consider the following boundary value problem in standard parabolic neighborhood $\mathcal{P}_a$:

$$\partial_{y_1}\varphi(y)+\sigma\varphi(y)=f(y) \ \text{ in } \ \mathcal{P}_a, \quad \varphi(y)=0 \ \text{ for } \ y_1=a^-(y_2,y_3). \tag{4.19}$$

LEMMA 4.4. *Assume that*

$$1/2<s\le 1 \quad \textit{and} \quad 1<r<3/(2s-1). \tag{4.20}$$

Then for any $f\in W^{s,r}(Q_a)\cap L^\infty(Q_a^\phi)$, problem (4.19) *has a unique solution satisfying the inequalities*

$$\begin{aligned}\|\varphi\|_{W^{s,r}(\mathcal{P}_a)}\le c(r,s)\Big(a^{4/r-s}\|f\|_{L^\infty(\mathcal{P}_a)}+a^{1/r}\|f\|_{W^{s,r}(\mathcal{P}_a)}\Big),\\ \|\varphi\|_{L^\infty(Q_a^\phi)}\le\sigma^{-1}\|f\|_{L^\infty(Q_a^\phi)}.\end{aligned} \tag{4.21}$$

Proof. It suffices to prove the lemma for $s<1$. For every $y,z\in\mathbb{R}^3$, we denote by $Y=(y_2,y_3)$, $Z=(z_2,z_3)$, respectively. Obviously, we have

$$\varphi(y)=\int_{a^-(Y)}^{y_1}e^{\sigma(x_1-y_1)}f(x_1,Y)\,dx_1 \ \text{ and } \ \sigma\|\varphi\|_{C(\mathcal{P}_a)}\le\|f\|_{C(\mathcal{P}_a)}. \tag{4.22}$$

Therefore, it suffices to estimate the seminorm $|\varphi|_{s,r,\mathcal{P}_a}$. Choose an arbitrary $y,z\in\mathcal{P}_a$. Without any loss of generality we can assume that $a^-(Z)\le a^-(Y)$. The identity

$$\begin{aligned}\varphi(z)-\varphi(y)=\varphi(z_1,Z)-\varphi(y_1,Z)+\int_{a^-(Z)}^{a^-(Y)}e^{\sigma(x_1-y_1)}f(x_1,Z)\,dx_1\\ +\int_{a^-(Y)}^{y_1}e^{\sigma(x_1-y_1)}(f(x_1,Z)-f(x_1,Y))\,dx_1\end{aligned}$$

implies the estimate

$$|\varphi(z)-\varphi(y)|\le\|f\|_{L^\infty(\mathcal{P}_a)}(2|y_1-z_1|+|a^-(Y)-a^-(Z)|)+\int_{-a}^{a}|(f(x_1,Z)-f(x_1,Y))|\,dx_1,$$

which along with the inequality

$$\left(\int_{-a}^{a}|(f(x_1,Z)-f(x_1,Y))|\,dx_1\right)^r\le a^{r-1}\int_{-a}^{a}|(f(x_1,Z)-f(x_1,Y))|^r\,dx_1$$

leads to the estimate

$$|\varphi|^r_{s,r,\mathcal{P}_a}\le 2\|f\|^r_{C(\mathcal{P}_a)}(I_1+I_2)+a^{(r-1)}I_3. \tag{4.23}$$

Here we denote

$$\begin{aligned}I_1=\int_{Q_a\times Q_a}\frac{|y_1-z_1|^r}{|x-y|^{3+rs}}\,dxdy, \quad I_2=\int_{\mathcal{P}_a\times\mathcal{P}_a}\frac{|a^-(Y)-a^-(Z)|^r}{|x-y|^{3+rs}}\,dxdy,\\ I_3=\int_{[-a,a]^4}\int_{-a}^{a}\frac{|f(x_1,Y)-f(x_1,Z)|^r}{|x-y|^{3+rs}}\,dxdy\,dx_1.\end{aligned}$$

Let us estimate the terms I_j, $i = 1, 2, 3$. We begin with the observation that

$$\int_{[-a,a]^2} \frac{dZ}{|x-y|^{3+rs}} = \frac{1}{|y_1 - z_1|^{3+rs}} \int_{[-a,a]^2} \frac{dZ}{(|Y-Z|^2/|y_1-z_1|^2+1)^{(3+rs)/2}}$$
$$\leq \frac{1}{|y_1-z_1|^{1+rs}} \int_{\mathbb{R}^2} \frac{dZ}{(|Z|^2+1)^{(3+rs)/2}} \leq \frac{c}{|y_1-z_1|^{1+rs}},$$

and hence

$$\int_{[-a,a]^4} \frac{dY\,dZ}{|x-y|^{3+rs}} \leq \frac{ca^2}{|y_1-z_1|^{1+rs}}.$$

From this we obtain

$$I_1 \leq ca^2 \int_{-a}^{a} \left(\int_{-a}^{a} |y_1 - z_1|^{r(1-s)-1}\, dz_1 \right) dy_1 \leq c(r,s) a^{3+r(1-s)}. \tag{4.24}$$

In order to estimate I_2, note that, by Lemma 4.2,

$$|a^-(Y) - a^-(Z)| \leq c|Y-Z|^{1/2}, \quad |a^-(Y) - a^-(Z)| \leq c|Y-Z| \left(\frac{1}{\sqrt{y_3}} + \frac{1}{\sqrt{z_3}} \right).$$

Next, it follows from the assumptions of the lemma that there is $\lambda \in (0,1)$ such that

$$\lambda < 3/r, \quad 0 < (1+\lambda)/2 - s < 1/r. \tag{4.25}$$

Noting that

$$|a^-(Y) - a^-(Z)| \leq c|Y-Z|^{(1+\lambda)/2} ((y_3)^{-\lambda/2} + (z_3)^{-\lambda/2}),$$

we obtain

$$I_2 \leq c \int_{\mathcal{P}_a} (y_3)^{-r\lambda/2} \left(\int_{Q_a} \frac{|Y-Z|^{r(1+\lambda)/2}}{|x-y|^{3+rs}}\, dz \right) dy$$
$$+ c \int_{\mathcal{P}_a} (z_3)^{-r\lambda/2} \left(\int_{Q_a} \frac{|Y-Z|^{r(1+\lambda)/2}}{|x-y|^{3+rs}}\, dy \right) dz$$
$$\leq 2c \int_{\mathcal{P}_a} (y_3)^{-r\lambda/2} \left(\int_{Q_a} \frac{|Y-Z|^{r(1+\lambda)/2}}{|x-y|^{3+rs}}\, dz \right) dy.$$

Next, inequalities (4.25) imply

$$\int_{Q_a} \frac{|Y-Z|^{r(1+\lambda)/2}}{|x-y|^{3+rs}}\, dz$$
$$\leq \int_{-a}^{a} |y_1 - z_1|^{-3-rs} \left(\int_{\mathbb{R}^2} \frac{|Y-Z|^{r(1+\lambda)/2}}{(|Y-Z|^2 |y_1-z_1|^{-2}+1)^{(3+rs)/2}}\, dZ \right)$$
$$\leq \int_{-a}^{a} |y_1 - z_1|^{-rs-1+r(1+\lambda)/2} dz_1 \int_{\mathbb{R}^2} \frac{|Z|^{r(1+\lambda)/2}}{(1+|Z|^2)^{(3+rs)/2}}\, dZ$$
$$\leq c(r,s) \int_{-a}^{a} |y_1 - z_1|^{r(1+\lambda)/2 - rs - 1} dz_1 \leq c(r,s).$$

From this and Lemma 4.2 we conclude that

$$\begin{aligned} I_2 &\le c(r,s)\int_{\mathcal{P}_a}(y_3)^{-r\lambda/2}\,dy \\ &\le c(r,s)\int_{[-a,a]^2}\left(\int_{cy_1^2\le y_3\le a}(y_3)^{-r\lambda/2}\,dy_3\right)dy_1dy_2 \\ &\le ac(r,s)\int_{-a}^{a}|y_1|^{2-r\lambda}\,dy_1 \le ac(r,s). \end{aligned} \tag{4.26}$$

The remaining part of the proof is based on the following proposition.

PROPOSITION 4.5. *Let $f\in W^{s,r}(Q_a)$ and $sr>1$. Then f has an extension $\bar f$ onto $\mathbb{R}^3$, which vanishes outside the set Q_{3a} and satisfies*

$$\|\bar f\|_{W^{s,r}}\le ca^{(3-rs)/r}\|f\|_{L^\infty(Q_a)}+|f|_{s,r,Q_a}. \tag{4.27}$$

Proof. Define an extension of f onto the slab $[-3a,3a]\times[-a,a]^2$ by the formula

$$f(x^\pm)=f(x)\ \text{ for }\ x\in Q_a,\ \text{ where }\ x^\pm=(\pm(2a-x_1),x_2,x_3).$$

It easily follows from the definition of the seminorm $|\cdot|_{r,s,\Omega}$ that

$$\|f\|_{W^{s,r}([-3a,3a]\times[-a,a]^2)}\le 3\|f\|_{L^r(Q_a)}+6|f|_{s,r,Q_a}\le 6\|f\|_{W^{s,r}(Q_a)}.$$

Proceeding in the same way as before, first we can extend f onto the plate $[-3a,3a]^2\times[-a,a]$ and then over the cube Q_{3a}. Obviously, the extended function, still denoted by f, satisfies the inequalities

$$\|f\|_{W^{s,r}(Q_{3a})}\le 216\|f\|_{W^{s,r}(Q_a)},\quad \|f\|_{C(Q_{3a})}\le\|f\|_{C(Q_a)}. \tag{4.28}$$

Next choose $\mu\in C^\infty(\mathbb{R}^3)$ such that $0\le\mu\le 1$, $\mu=1$ in Q_1, and $\mu=0$ outside of Q_{22}. Set $\bar f=f\mu_a$, where $\mu_a(x)=\mu(x/a)$. Next, the interpolation inequality along with the estimate $|\nabla\mu_a|\le ca^{-1}$ implies

$$\|\mu_a\|_{W^{s,r}(\mathbb{R}^3)}\le\|\mu_a\|^{1-s}_{L^r(\mathbb{R}^3)}\|\mu_a\|^{s}_{W^{1,r}(\mathbb{R}^3)}\le ca^{3(1-s)/r}a^{(3-r)s/r}=ca^{(3-rs)/r}.$$

From this and the obvious inequality $\|\mu_af\|_{W^{s,r}(\mathbb{R}^3)}\le\|f\|_{L^\infty(Q_{3a})}\|\mu_a\|_{W^{s,r}(\mathbb{R}^3)}+\|f\|_{W^{s,r}(Q_{3a})}$ we conclude that

$$\|\mu_af\|_{W^{s,r}(\mathbb{R}^3)}\le ca^{(3-rs)/r}\|f\|_{L^\infty(Q_a)}+\|f\|_{W^{s,r}(Q_a)}.$$

Hence $\bar f=\mu_af$ satisfies (4.27), and the proposition follows. ∎

Let us return to the proof of Lemma 4.4. We have

$$\begin{aligned} I_3=&\int_{[-a,a]^4}\int_{-a}^{a}|Z-Y|^{-3-rs}(|z_1-y_1|^2|Z-Y|^{-2}+1)^{-(3+rs)/2} \\ &\times|f(x_1,Z)-f(x_1,Y)|^r\,dxdydx_1\le ca\int_{\mathbb{R}}(|t|^2+1)^{-(3+rs)/2}dt \\ &\times\int_{[-a,a]^5}|Z-Y|^{-2-rs}|f(x_1,Z)-f(x_1,Y)|^r\,dXdYdx_1 \\ =&\,ca\int_{-a}^{a}|f(x_1,\cdot)|^r_{r,s,[-a,a]^2}\,dx_1, \end{aligned}$$

which yields

$$I_3 \le ca\|f\|^r_{L^r(-a,a;W^{s,r}([-a,a]^2))} \le ca\|\bar f\|^r_{L^r(\mathbb{R};W^{s,r}(\mathbb{R}^2))}. \tag{4.29}$$

Recall that, for $s = 0, 1$, the embedding operator $W^{s,r}(\mathbb{R}^3) \hookrightarrow L^r(\mathbb{R}; W^{s,r}(\mathbb{R}^2))$ is bounded. By virtue of Lemma B.1, this results holds true for all $s \in [0, 1]$, which along with Proposition 4.5 and inequality (4.29) implies

$$I_3 \le ca^{4-rs}\|f\|^r_{L^\infty(Q_a)} + a(|f|_{s,r,Q_a})^r.$$

Combining this result with (4.24), (4.26), since $3 + r(1 - s) \ge 4 - rs$, we finally obtain

$$\|f\|^r_{L^\infty(Q_a)}(I_1 + I_2) + I_3 \le ca^{4-sr}\|f\|^r_{L^\infty(Q_a)} + ca|f|^r_{s,r,Q_a}.$$

Substituting this inequality into (4.23) gives (4.21), and the lemma follows. □

Let us consider the following boundary value problem in the slab $Q^+_a = [-a, a]^2 \times [0, a]$:

$$\partial_{y_3}\varphi(y) + \sigma\varphi(y) = f(y) \ \text{ in } \ Q^+_a, \quad \varphi(y) = 0 \ \text{ for } \ y_3 = 0. \tag{4.30}$$

Lemma 4.6. *Problem* (4.30) *has a unique solution satisfying the inequality*

$$\|\varphi\|_{W^{s,r}(Q^+_a)} \le c(r, s)\big(a^{4/r-s}\|f\|_{L^\infty(Q^+_a)} + a^{1/r}\|f\|_{W^{s,r}(Q^+_a)}\big). \tag{4.31}$$

Proof. The proof of Lemma 4.4 can also be used in this case. □

Local existence results. It follows from the assumptions of Theorem 1.4 that the vector field $\mathbf{u}$ and the manifold Σ satisfy all assumptions of Lemma 4.2. Therefore, there exist positive numbers a, ρ_c, and R_c, depending only on Σ and $\|\mathbf{u}\|_{C^1(\Omega)}$, such that for all $P \in \Gamma$, the canonical diffeomorphism $\mathbf{x} : Q_a \mapsto \mathcal{O}_P$ is well defined and meets all requirements of Lemma 4.2. Fix an arbitrary point $P \in \Gamma$ and consider the boundary value problem

$$\mathbf{u} \cdot \nabla\varphi + \sigma\varphi = f \ \text{ in } \ \mathcal{O}_P, \quad \varphi = 0 \ \text{ on } \ \Sigma_{\rm in} \cap \mathcal{O}_P. \tag{4.32}$$

Lemma 4.7. *Suppose that the exponents s, r satisfy condition* (1.31). *Then, for any $f \in C^1(\Omega)$, problem* (4.32) *has a unique solution satisfying the inequalities*

$$|\varphi|_{s,r,B_P(\rho_c)} \le c(\|f\|_{C(B_P(R_c))} + |f|_{s,r,B_P(R_c)}), \quad \|\varphi\|_{C(B_P(\rho_c))} \le \sigma^{-1}\|f\|_{C(B_P(R_c))}, \tag{4.33}$$

where the constant c depends only on Σ, M, σ, s, r and the constant ρ_c is determined by Lemma 4.2.

Proof. We transform (4.33) using the normal coordinates (y_1, y_2, y_3) given by Lemma 4.2. Set $\overline{\varphi}(y) = \varphi(\mathbf{x}(y))$ and $\overline{f}(y) = f(\mathbf{x}(y))$. Next note that (4.3) implies the identity $\mathbf{u}\nabla_x\varphi = \partial_{y_1}\overline{\varphi}(y)$. Therefore the function $\overline{\varphi}(y)$ satisfies the following equation along with the boundary conditions:

$$\partial_{y_1}\overline{\varphi} + \sigma\overline{\varphi} = \overline{f} \text{ in } Q_a \cap \{y_3 > \Phi\}, \quad \overline{\varphi} = 0 \ \text{ for } \ y_3 = \Phi(y_1, y_2),\ y_1 < 0. \tag{4.34}$$

It follows from Lemma 4.4 that problem (4.34) has a unique solution $\overline{\varphi} \in W^{s,r}(G_a)$ satisfying the inequality

$$|\overline{\varphi}|_{s,r,G_a} \le c(\|\overline{f}\|_{C(Q_a)} + |\overline{f}|_{s,r,Q_a}), \quad \|\overline{\varphi}\|_{C(G_a)} \le \sigma^{-1}\|f\|_{C(Q_a)}, \tag{4.35}$$

where the domain G_a is defined by (4.6). It remains to note that, by estimate (4.4), the mappings $\mathbf{x}^{\pm 1}$ are uniformly Lipschitz, which along with inclusions (4.7) implies the estimates

$$|\varphi|_{s,r,B_P(\rho_c)} \le c|\overline{\varphi}|_{s,r,G_a}, \quad |\overline{f}|_{s,r,Q_a} \le c|f|_{s,r,B_P(R_c)}.$$

Combining these results with (4.35) we finally obtain (4.33) and the lemma follows. □

In order to formulate the similar result for interior points of the inlet Σ_{in} we introduce the set

$$\Sigma'_{\text{in}} = \big\{x \in \Sigma_{\text{in}} : \ \operatorname{dist}(x, \Gamma) \ge \rho_c/3\big\}, \tag{4.36}$$

where the constant ρ_c is given by Lemma 4.2. It is clear that

$$\inf_{P \in \Sigma'_{\text{in}}} \mathbf{U}(P) \cdot \mathbf{n}(P) \ge N > 0,$$

where the constant N depends only on M and Σ. It follows from Lemma 4.3 that there are positive numbers b, ρ_i, and R_i such that for each $P \in \Sigma'_{\text{in}}$, the canonical diffeomorphism $\mathbf{x} : Q_b \mapsto \mathcal{O}_P$ is well defined and satisfies the hypotheses of Lemma 4.3. The following lemma gives the local existence and uniqueness of solutions to the boundary value problem

$$\mathbf{u} \cdot \nabla\varphi + \sigma\varphi = f \ \text{ in } \ \mathcal{O}_P, \quad \varphi = 0 \ \text{ on } \ \Sigma_{\text{in}} \cap \mathcal{O}_P. \tag{4.37}$$

LEMMA 4.8. *Suppose that the exponents s, r satisfy condition* (1.38). *Then, for any $f \in C^1(\Omega)$ and $P \in \Sigma'_{\text{in}}$, problem* (4.32) *has a unique solution satisfying the inequalities*

$$\begin{aligned} |\varphi|_{s,r,B_P(\rho_i)} &\le c(\|f\|_{C(B_P(R_i))} + |f|_{s,r,B_P(R_i)}), \\ \|\varphi\|_{C(B_P(R))} &\le \sigma^{-1}\|f\|_{C(B_P(R_i))}, \end{aligned} \tag{4.38}$$

where c depends on Σ, M, σ, and exponents s, r.

Proof. Using the normal coordinates given by Lemma 4.3 we rewrite (4.37) in the form

$$\partial_{y_3}\overline{\varphi} + \sigma\overline{\varphi} = \overline{f} \ \text{ in } \ Q_b, \quad \overline{\varphi} = 0 \ \text{ for } \ y_3 = 0.$$

Applying Lemma 4.4 and arguing as in the proof of Lemma 4.7 we obtain (4.38). □

Existence of solutions near inlet. The next step is based on the well-known geometric lemma (see [21, Chap. 3]).

LEMMA 4.9. *Suppose that a given set $A \subset \mathbb{R}^d$ is covered by balls such that each point $x \in A$ is the center of a certain ball $B_x(r(x))$ of radius $r(x)$. If $\sup r(x) < \infty$, then from the system of the balls $\{B_x(r(x))\}$ it is possible to select a countable system $B_{x_k}(r(x_k))$ covering the entire set A and having multiplicity not greater than a certain number $\mathfrak{n}(d)$ depending only on the dimension d.*

The following lemma gives the dependence of the multiplicity of radii of the covering balls.

LEMMA 4.10. *Assume that a collection of balls $B_{x_k}(r) \subset \mathbb{R}^3$ of constant radius r has the multiplicity $\mathfrak{n}_r$. Then the multiplicity of the collections of the balls $B_{x_k}(R)$, $r < R$, is bounded by the constant $27(R/r)^3\mathfrak{n}_r$.*

Proof. Let $\mathfrak{n}_R$ be a multiplicity of the system $\{B_{x_k}(R)\}$. This means that at least $\mathfrak{n}_R$ balls, say $B_{x_1}(R), \dots, B_{x_{\mathfrak{n}_R}}(R)$, have the common point P. In particular, we

have $B_{x_i}(r) \subset B_P(3R)$ for all $i \le \mathfrak{n}_R$. Introduce the counting function $\iota(x)$ for the collection of balls $B_{x_i}(r)$, defined by

$$\iota(x) = \text{ card}\{i : x \in B_{x_i}(r), 1 \le i \le \mathfrak{n}_r\}.$$

Note that $\iota(x) \le \mathfrak{n}_r$. We have

$$\frac{4\pi}{3}\mathfrak{n}_R r^3 = \sum_{i=1}^{\mathfrak{n}_R} \text{meas } B_{x_i}(r) = \int_{\cup_i B_{x_i}(r)} \iota(x)\, dx \le \mathfrak{n}_r \int_{\cup_i B_{x_i}(r)} dx \le \frac{4\pi}{3}(3R)^3 \mathfrak{n}_r,$$

and the lemma follows. □

We are now in a position to prove the local existence and uniqueness of a solution to the first boundary value problem for the transport equation in the neighborhood of the inlet Σ_{in}. Let Ω_t be the t-neighborhood of the set Σ_{in},

$$\Omega_t = \{x \in \Omega : \text{ dist } (x, \Sigma_{\text{in}}) < t\}.$$

LEMMA 4.11. *Let $t = \min\{\rho_c/2, \rho_i/2\}$ and $T = \max\{R_c, R_i\}$, where the constants ρ_α, R_α are defined by Lemmas* 4.2 *and* 4.3. *Then there exists a constant C, depending only on M, Σ, and σ, such that, for any $f \in C^1(\Omega)$, the boundary value problem*

$$\mathbf{u} \cdot \nabla\varphi + \sigma\varphi = f \ \text{ in } \ \Omega_t, \quad \varphi = 0 \ \text{ on } \ \Sigma_{\text{in}} \tag{4.39}$$

has a unique solution satisfying the inequalities

$$|\varphi|_{s,r,\Omega_t} \le C(\|f\|_{C(\Omega_T)} + |f|_{s,r,\Omega_T}), \quad \|\varphi\|_{C(\Omega_t)} \le \sigma^{-1}\|f\|_{C(\Omega_T)}. \tag{4.40}$$

Proof. It follows from Lemma 4.9 that there is a covering of the characteristic manifold Γ by the finite collection of balls $B_{P_i}(\rho_c/4)$, $1 \le i \le \mathfrak{m}$, $P_i \in \Gamma$, of the multiplicity $\mathfrak{n}$. The cardinality $\mathfrak{m}$ of this collection does not exceed $4\mathfrak{n}(\rho_c)^{-1}L$, where L is the length of Γ. Obviously, the balls $B_{P_i}(\rho_c)$ cover the set

$$\mathcal{V}_\Gamma = \{x \in \Omega : \text{ dist } (x, \Gamma) < \rho_c/2\}.$$

By virtue of Lemma 4.7, in each such ball the solution to problem (4.39) satisfies inequalities (4.33), which leads to the estimate

$$|\varphi|^r_{s,r,\mathcal{V}_\Gamma} \le \sum_i |\varphi|^r_{s,r,B_{P_i}(\rho_c)} \le c\sum_i \|f\|^r_{C(B_{P_i}(R_c))} + c\sum_i |f|^r_{s,r,B_{P_i}(R_c)}, \tag{4.41}$$

where c depends only on M, Σ, and σ. By Lemma 4.10, the multiplicity of the system of balls $B_{P_i}(R_c)$ is bounded from above by $12^3(R_c/\rho_c)^3$, which along with the inclusion $\cup_i B_{P_i}(R_c) \subset \Omega_T$ yields

$$\sum_{i=1}^{\mathfrak{m}} |f|^r_{s,r,B_{P_i}(R_c)} \le 12^3(R_c/\rho_c)^3 |f|^r_{s,r,\Omega_T}.$$

Obviously we have

$$\sum_i \|f\|^r_{C(B_{P_i}(R_c))} \le \mathfrak{m}\|f\|^r_{C(\Omega_T)} \le 4\mathfrak{n}(\rho_c)^{-1}L\|f\|^r_{C(\Omega_T)}.$$

Combining these results with (4.41) we obtain the estimates for the solution to problem (4.39) in the neighborhood of the characteristic manifold Γ,

$$|\varphi|_{s,r,\mathcal{V}_\Gamma} \le c\|f\|_{C(\Omega_T)} + c|f|_{s,r,\Omega_T}. \tag{4.42}$$

Our next task is to obtain the similar estimate in the neighborhood of the compact $\Sigma'_{\rm in} \subset \Sigma_{\rm in}$. To this end, we introduce the set

$$\mathcal{V}_{\rm in} = \{x \in \Omega :\ \text{dist}\ (x, \Sigma'_{\rm in}) < \rho_i/2\},$$

where $\Sigma'_{\rm in}$ is given by (4.36). By virtue of Lemma 4.9, there exists the finite collection of balls $B_{P_k}(\rho_i/4)$, $1 \le k \le \mathfrak{m}$, $P_k \in \Sigma'_{\rm in}$, of the multiplicity $\mathfrak{n}$ which covers $\Sigma'_{\rm in}$. Obviously $\mathfrak{m} \le 16n(\rho_i)^{-2}\text{meas}\ \Sigma_{\rm in}$, and the balls $B_{P_k}(\rho_i)$ cover the set $\mathcal{V}_{\rm in}$. From this and Lemma 4.8 we conclude that

$$|\varphi|^r_{s,r,\mathcal{V}_{\rm in}} \le \sum_k |\varphi|^r_{s,r,B_{P_k}(\rho_i)} \le c\sum_k \|f\|^r_{C(B_{P_k}(R_i))} + c\sum_k |f|^r_{s,r,B_{P_k}(R_i)}.$$

By virtue of Lemma 4.10, the multiplicity of the system of balls $B_{P_i}(R_i)$ is not greater than $12^3(R_i/\rho_i)^3$, which yields

$$\sum_i |f|^r_{s,r,B_{P_i}(R_i)} \le 12^3(R_i/\rho_i)^3|f|^r_{s,r,\Omega_T}.$$

Obviously we have

$$\sum_k \|f\|^r_{C(B_{P_k}(R_i))} \le \mathfrak{m}\|f\|^r_{C(\Omega_T)} \le 16\mathfrak{n}(\rho_k)^{-2}\text{meas}\ \Sigma_{\rm in}\|f\|^r_{C(\Omega_T)}.$$

Thus we get

$$|\varphi|_{s,r,\mathcal{V}_{in}} \le c\|f\|_{C(\Omega_T)} + c|f|_{s,r,\Omega_T}. \tag{4.43}$$

Since $\mathcal{V}_\Gamma$ and $\mathcal{V}_{\rm in}$ cover Ω_t, this inequality along with inequality (4.42) yields (4.40), and the lemma follows. □

Partition of unity. Let us turn to the analysis of the general problem

$$\mathscr{L}\varphi := \mathbf{u}\cdot\nabla\varphi + \sigma\varphi = f \ \text{ in } \ \Omega, \quad \varphi = 0 \ \text{ on } \ \Sigma_{\rm in}. \tag{4.44}$$

The next step is based on the theory of partial differential equations with nonnegative characteristic form. The following lemma is a particular case of general results of Oleinik and Radkevich; we refer to Theorems 1.5.1 and 1.6.2 in [34].

LEMMA 4.12. *Assume that Ω is a bounded domain of the class C^2, the vector field $\mathbf{u}$ belongs to the class $C^1(\Omega)^3$, and $\sigma - \operatorname{div}\mathbf{u}(x) > \delta > 0$. Then, for any $f \in L^\infty(\Omega)$, problem (1.28) has a unique solution such that $\|\varphi\|_{L^\infty(\Omega)} \le \delta^{-1}\|f\|_{L^\infty}$. Moreover, this solution is continuous at the interior points of $\Sigma_{\rm in}$ and vanishes on $\Sigma_{\rm in}$. If, in addition, $\Gamma = \ cl\ (\Sigma_{\rm out} \cap \Sigma_0) \cap cl\ \Sigma_{\rm in}$ is a smooth one-dimensional manifold, then a bounded generalized solution to problem (4.44) is unique.*

The question of smoothness of solutions to boundary value problems for transport equations is more complicated. All known results [19], [34] are related to the case of $\Gamma = \emptyset$. The following lemma is a consequence of Theorem 1.8.1 in the monograph [34].

LEMMA 4.13. *Assume that Ω is a bounded domain of the class C^2 and $\Sigma_{\rm out} = \emptyset$. Furthermore, let the following conditions hold.*

(1) *The vector field* $\mathbf{u}$ *and the function* f *belong to the class* $C^1(\mathbb{R}^3)$.
(2) *There is* $\Omega' \ni \Omega$ *such that the inequality*

$$\sigma - \sup_{\Omega'} \left\{ |\operatorname{div} \mathbf{u}| - \frac{1}{2} \sup_i \sum_{j \neq i} \left| \frac{\partial u_i}{\partial x_j} \right| - \frac{1}{2} \sup_j \sum_{j \neq i} \left| \frac{\partial u_j}{\partial x_i} \right| \right\} > 0$$

is fulfilled. Then a weak solution to problem (1.28) *satisfies the Lipschitz condition in* $\overline{\Omega}$.

Using these results we can construct a strong solution to problem (1.28). Recall that by Lemma 4.11, for any $f \in C^1(\Omega)$, problem (4.44) has a unique strong solution defined in neighborhood Ω_t of the inlet Σ_{in}. On the other hand, Lemma 4.12 guarantees the existence and uniqueness of a bounded weak solution to problem (4.44). The following lemma shows that both solutions coincide in Ω_t.

LEMMA 4.14. *Under the assumptions of Theorem* 1.4 *and Lemma* 4.11*, each bounded generalized solution to problem* (4.44) *coincides in* Ω_t *with the local solution* φ_t.

Proof. Let $\varphi \in L^\infty(\Omega)$ be a weak solution to problem (4.44). Recall that each point $P \in \Gamma$ has a canonical neighborhood $\mathcal{O}_P := \mathbf{x}(Q_a)$, where canonical diffeomorphism $\mathbf{x} : Q_a \mapsto \mathcal{O}_P$ is defined by Lemma 4.2. Choose an arbitrary function $\zeta \in C^1(\Omega)$ vanishing on Σ_{in} and outside of $\mathcal{O}_P$ and set

$$\overline{\varphi}(y) = \varphi(\mathbf{x}(y)), \quad \overline{f}(y) = f(\mathbf{x}(y)), \quad \overline{\zeta}(y) = \zeta(\mathbf{x}(y)), \quad y \in Q_a \cap \{y_3 > \Phi\}.$$

By the definition of a weak solution to the transport equation we have

$$\int_{\mathcal{O}_P \cap \Omega} \big(\sigma \varphi \zeta - \varphi \operatorname{div}(\zeta \mathbf{u}) - f\zeta\big)\, dx = 0.$$

Direct calculations lead to the identity $\operatorname{div}_x(\zeta \mathbf{u}) = \det \mathfrak{F}^{-1} \operatorname{div}_y(\overline{\zeta} \det \mathfrak{F}\, \mathfrak{F}^{-1} \overline{\mathbf{u}})$, in which the notation $\mathfrak{F}$ stands for the Jacobi matrix $\mathfrak{F} = D_y \mathbf{x}(y)$. On the other hand, (4.3) implies the equality $\mathfrak{F}^{-1} \overline{\mathbf{u}} = \mathbf{e_1}$. From this we conclude that

$$\int_{Q_a \cap \{y_3 > \Phi\}} \left(\big(\det \mathfrak{F}\, \overline{\zeta}\big)\big(\sigma \overline{\varphi} - \overline{f}\big) - \overline{\varphi} \frac{\partial}{\partial y_1} (\det \mathfrak{F}\, \overline{\zeta}) \right) dy = 0.$$

Recall that, by Lemma 4.2, $\partial_{y_1} \mathfrak{F}$ is continuous and $\det \mathfrak{F}$ is strictly positive in the cube Q_a. Setting $\xi = \det \mathfrak{F} \zeta$ we conclude that the integral identity

$$\int_{Q_a \cap \{y_3 > \Phi\}} \left(\xi\big(\sigma \overline{\varphi} - \overline{f}\big) - \overline{\varphi} \frac{\partial \xi}{\partial y_1} \right) dy = 0$$

holds true for all functions $\xi \in C_0(Q_a)$ having continuous derivative $\partial_{y_1} \zeta \in C(Q_a)$ and vanishing for $y_3 = \Phi(y_1, y_2)$, $y_1 < 0$. Since $\overline{f}$ is continuously differentiable, $\overline{\varphi}$ belongs to the class $C^1_{loc}(Q_a \cap \{y_3 > \Phi\})$ and satisfies (4.34). On the other hand, $\overline{\varphi_t}$ also satisfies (4.34). Obviously, all solutions to problem (4.34) coincide in the domain G_a, and hence $\overline{\varphi}_t = \overline{\varphi}$ in this domain. Recalling that $B_P(\rho_c) \subset \mathbf{x}(G_a)$ we obtain that $\varphi_t = \varphi$ in the ball $B_P(\rho_c)$. The same arguments show that, for any $P \in \Sigma'_{\text{in}}$, the function φ_t is equal to φ in the ball $B_P(\rho_i)$. It remains to note that the balls $B_P(\rho_c)$ and $B_P(\rho_i)$ cover Ω_t, and the lemma follows. ∎

Furthermore, we split the weak solution $\varphi \in L^\infty(\Omega)$ to problem (4.44) into two parts, namely, the local solution φ_t and the remainder vanishing near the inlet. To this end fix a function $\Lambda \in C^\infty(\mathbb{R})$ such that

$$0 \le \Lambda' \le 3, \quad \Lambda(u) = 0 \quad \text{for} \quad u \le 1 \quad \text{and} \quad \Lambda(u) = 1 \quad \text{for} \quad u \ge 3/2, \tag{4.45}$$

and introduce the one-parametric family of smooth functions

$$\chi_t(x) = \frac{1}{t^3} \int_{\mathbb{R}^3} \Theta\left(\frac{2(x-y)}{t}\right) \Lambda\left(\frac{\text{dist} \ \ (y, \Sigma_{\text{in}})}{t}\right) dy, \tag{4.46}$$

where $\Theta \in C^\infty(\mathbb{R}^3)$ is a standard mollifying kernel supported in the unit ball. It follows that

$$\begin{gathered} \chi_t(x) = 0 \ \ \text{for} \ \ \text{dist} \ (x, \Sigma_{\text{in}}) \le t/2, \quad \chi_t(x) = 1 \ \ \text{for} \ \ \text{dist} \ (x, \Sigma_{\text{in}}) \ge 2t, \\ |\partial^l \chi_t(x)| \le \varpi(l) t^{-l} \ \ \text{for all} \ \ l \ge 0, \end{gathered} \tag{4.47}$$

where $\varpi(l)$ is a constant. Now fix a number $t = t(\Sigma, M)$ satisfying all assumptions of Lemma 4.11 and set

$$\varphi(x) = (1 - \chi_{t/2}(x))\varphi_t(x) + \phi(x). \tag{4.48}$$

By virtue of (4.47) and Lemma 4.14, the function $\phi \in L^\infty(\Omega)$ vanishes in $\Omega_{t/2}$ and in a weak sense satisfies the equations

$$\mathbf{u}\nabla\phi + \sigma\phi = \chi_{t/2} f + \varphi_t \mathbf{u} \nabla \chi_{t/2} =: F \ \ \text{in} \ \ \Omega, \ \ \phi = 0 \ \ \text{on} \ \ \Sigma_{\text{in}}.$$

Next introduce a new vector field $\tilde{\mathbf{u}}(x) = \chi_{t/8}(x)\mathbf{u}(x)$. It easy to see that $\chi_{t/8} = 1$ on the support of ϕ, and hence the function ϕ is also a weak solution to the modified transport equation

$$\tilde{\mathscr{L}}\phi := \tilde{\mathbf{u}}\nabla\phi + \sigma\phi = F \ \ \text{in} \ \ \Omega. \tag{4.49}$$

The advantage of such an approach is that the topology of integral lines of the modified vector field $\tilde{\mathbf{u}}$ drastically differs from the topology of integral lines of $\mathbf{u}$. The corresponding inlet, outgoing set, and characteristic set have the other structure and $\tilde{\Sigma}_{\text{in}} = \emptyset$. In particular, (4.49) does not require boundary conditions. Finally note that the C^1-norm of the modified vector fields has the majorant

$$\|\tilde{\mathbf{u}}\|_{C^1(\Omega)} \le M(1 + 16\varpi(1)t^{-1}), \tag{4.50}$$

where $\varpi(1)$ is a constant from (4.47). The following lemma constitutes the existence and uniqueness of solutions to the modified equation.

LEMMA 4.15. *Suppose that*

$$\sigma > \sigma^*(M, \Sigma) = 4M(1 + 16\varpi(1)t^{-1}) + 1, \quad M = \|\mathbf{u}\|_{C^1(\Omega)}, \tag{4.51}$$

and $0 \le s \le 1$, $r > 1$. *Then, for any* $F \in W^{s,r}(\Omega) \cap L^\infty(\Omega)$, (4.49) *has a unique weak solution* $\phi \in W^{s,r}(\Omega) \cap L^\infty(\Omega)$ *such that*

$$\|\phi\|_{W^{s,r}(\Omega)} \le c\|F\|_{W^{s,r}(\Omega)}, \quad \|\phi\|_{L^\infty(\Omega)} \le \sigma^{-1}\|F\|_{L^\infty(\Omega)}, \tag{4.52}$$

where c *depends only on* r.

Proof. Without any loss of generality we can assume that $F \in C^1(\Omega)$. By virtue of (4.50) and (4.51), the vector field $\tilde{\mathbf{u}}$ and σ meet all requirements of Lemma 4.13. Hence (4.49) has a unique solution $\phi \in W^{1,\infty}(\Omega)$. For $i = 1, 2, 3$ and $\tau > 0$, define the finite difference operator

$$\delta_{i\tau}\phi = \frac{1}{\tau}\big(\phi(x + \tau\mathbf{e}_i) - \phi(x)\big).$$

It is easy to see that

$$\tilde{\mathbf{u}}\nabla\delta_{i\tau}\phi + \sigma\delta_{i\tau}\phi = \delta_{i\tau}F - \delta_{i\tau}\tilde{\mathbf{u}}\nabla\phi(x + \tau\mathbf{e}_i) \ \text{ in } \ \Omega \cap (\Omega - \tau\mathbf{e}_i). \tag{4.53}$$

Next introduce the function $\eta \in C^\infty(\mathbb{R})$ such that $\eta' \geq 0$, $\eta(u) = 0$ for $u \leq 1$, and $\eta(u) = 1$ for $u \geq 1$, and set $\eta_h(x) = \eta\big(\operatorname{dist}(x, \partial\Omega)/h\big)$. Since $\tilde{\Sigma}_{\text{in}} = \emptyset$, the inequality

$$\limsup_{h\to 0} \int_\Omega g\tilde{\mathbf{u}} \cdot \nabla\eta_h(x)\, dx \leq 0 \tag{4.54}$$

holds true for all nonnegative functions $g \in L^\infty(\Omega)$. Choosing $h > \tau$, multiplying both sides of (4.53) by $\eta_h|\delta_{i\tau}\phi|^{r-2}\delta_{i\tau}\phi$, and integrating the result over $\Omega \cap (\Omega - \tau\mathbf{e}_i)$, we obtain

$$\begin{aligned}\int_{\Omega\cap(\Omega-\tau\mathbf{e}_i)} \eta_h|\delta_{i\tau}\phi|^r \left(\sigma - \frac{1}{r}\operatorname{div}\tilde{\mathbf{u}}\right) dx - \int_{\Omega\cap(\Omega-\tau\mathbf{e}_i)} |\delta_{i\tau}\phi|^r \tilde{\mathbf{u}}\nabla\eta_h\, dx \\ = \int_{\Omega\cap(\Omega-\tau\mathbf{e}_i)} \big(\delta_{i\tau}F - \delta_{i\tau}\tilde{\mathbf{u}}\nabla\phi(x + \tau\mathbf{e}_i)\big)\eta_h|\delta_{i\tau}\phi|^{r-2}\delta_{i\tau}\phi\, dx.\end{aligned}$$

Letting $\tau \to 0$ and then $h \to 0$ and using inequality (4.54), we obtain

$$\int_\Omega |\partial_{x_i}\phi|^r \left(\sigma - \frac{1}{r}\operatorname{div}\tilde{\mathbf{u}}\right) dx \leq \int_\Omega \big(\partial_{x_i}F - \partial_{x_i}\tilde{\mathbf{u}}\nabla\phi\big)|\partial_{x_i}\phi|^{r-2}\partial_{x_i}\phi\, dx. \tag{4.55}$$

Next note that

$$\sum_i \partial_{x_i}\tilde{\mathbf{u}}\nabla\phi\big|\partial_{x_i}\phi|^{r-2}\partial_{x_i}\phi \leq 3\|\tilde{\mathbf{u}}\|_{C^1(\Omega)} \sum_i |\partial_{x_i}\phi|^r.$$

On the other hand, since $1/r + 3 \leq 4$, inequalities (4.50) and (4.51) imply

$$\sigma - \left(\frac{1}{r} + 3\right)\|\tilde{\mathbf{u}}\|_{C^1(\Omega)} \geq \sigma - 4M(1 + 16\varpi(1)t^{-1}) \geq 1.$$

From this we conclude that

$$\begin{aligned}\sum_i \int_\Omega |\partial_{x_i}\phi|^r\, dx &\leq \sum_i \int_\Omega |\partial_{x_i}\phi|^{r-1}|\partial_{x_i}F|\, dx \\ &\leq \left(\sum_i \int_\Omega |\partial_{x_i}\phi|^{r/(r-1)}\, dx\right)^{(r-1)/r} \left(\sum_i \int_\Omega |\partial_{x_i}F|^r\, dx\right)^{1/r},\end{aligned}$$

which leads to the estimate

$$\|\nabla\phi\|_{L^r(\Omega)} \leq c(r)\|\nabla F\|_{L^r(\Omega)}. \tag{4.56}$$

Next, multiplying both sides of (4.49) by $|\phi|^{r-2}\eta_h$ and integrating the result over Ω, we get the identity

$$\int_\Omega \left(\sigma - \frac{1}{r}\operatorname{div}\tilde{\mathbf{u}}\right)\eta_h|\phi|^r\,dx - \int_\Omega |\phi|^r\tilde{\mathbf{u}}\nabla\eta_h\,dx = \int_\Omega F\eta_h|\phi|^{r-2}\phi\,dx.$$

The passage $h \to 0$ gives the inequality

$$\int_\Omega \left(\sigma - \frac{1}{r}\operatorname{div}\tilde{\mathbf{u}}\right)|\phi|^r\,dx \le \int_\Omega |F||\phi|^{r-1}\,dx.$$

Recalling that $\sigma - 1/r\operatorname{div}\tilde{\mathbf{u}} \ge 1$, we finally obtain

$$\|\phi\|_{L^r(\Omega)} \le c(r)\|F\|_{L^r(\Omega)}. \tag{4.57}$$

Inequalities (4.56) and (4.57) imply estimate (4.52) for $s = 0, 1$. Hence the linear operator $\tilde{\mathscr{L}}^{-1} : F \mapsto \phi$ is continuous in the Banach spaces $W^{0,r}(\Omega)$ and $W^{1,r}(\Omega)$, and its norm does not exceed $c(r)$. Recall that $W^{s,r}(\Omega)$ is the interpolation space $[L^r(\Omega), W^{1,r}(\Omega)]_{s,r}$. From this and Lemma B.1 we conclude that inequality (4.52) is fulfilled for all $s \in [0, 1]$, which completes the proof. □

Proof of Theorem 1.4. We begin with the proof of statement (i). Fix $\sigma > \sigma^*$, where the constant σ^* depends only on Σ, $\mathbf{U}$, and $\|\mathbf{u}\|_{C^1(\Omega)}$ and is defined by (4.51). Without any loss of generality we can assume that $f \in C^1(\Omega)$. The existence and uniqueness of a weak bounded solution for $\sigma > \sigma^*$ follows from Lemma 4.12. Moreover, by virtue of Lemma 4.12, such a solution satisfies the second inequality in (1.33). Therefore, it suffices to prove estimate (1.33) for $\|\varphi\|_{W^{s,r}(\Omega)}$. Since $W^{s,r}(\Omega) \cap L^\infty(\Omega)$ is the Banach algebra, representation (4.48) together with inequality (4.47) implies

$$\|\varphi\|_{W^{s,r}(\Omega)} \le c(1 + t^{-1})(\|\varphi_t\|_{W^{s,r}(\Omega_t)} + \|\varphi_t\|_{L^\infty(\Omega_t)}) + c\|\phi\|_{W^{s,r}(\Omega)}. \tag{4.58}$$

On the other hand, Lemma 4.15 along with (4.49) yields

$$\|\phi\|_{W^{s,r}(\Omega)} \le c\|F\|_{W^{s,r}(\Omega)} \le c\|\chi_{t/2}f\|_{W^{s,r}(\Omega)} + \|\varphi_t\mathbf{u}\nabla\chi_{t/2}\|_{W^{s,r}(\Omega)}.$$

The first term in the right-hand side is bounded:

$$\|\chi_{t/2}f\|_{W^{s,r}(\Omega)} \le c(1 + t^{-1})(\|f\|_{W^{s,r}(\Omega)} + \|\varphi_t\|_{L^\infty(\Omega)}).$$

In order to estimate the second term we note that, by virtue of (4.47), $\|\mathbf{u}\nabla\chi_{t/2}\|_{C^1(\Omega)} \le cM(1 + t^{-2})$, which gives

$$\|\varphi_t\mathbf{u}\nabla\chi_{t/2}\|_{W^{s,r}(\Omega)} \le cM(1 + t^{-2})(\|\varphi_t\|_{W^{s,r}(\Omega_t)} + \|\varphi_t\|_{L^\infty(\Omega_t)}).$$

Substituting the obtained estimates into (4.58) we arrive at the inequality

$$\|\varphi\|_{W^{s,r}(\Omega)} \le c(M+1)(1+t^{-2})\big(\|\varphi_t\|_{W^{s,r}(\Omega)} + \|\varphi_t\|_{L^\infty(\Omega_t)} + \|f\|_{W^{s,r}(\Omega_t)} + \|f\|_{L^\infty(\Omega)}\big),$$

which along with (4.40) leads to the estimate (1.33). In order to prove statement (ii) of Theorem 1.4, we note that the adjoint equation can be written in the form

$$-\mathbf{u}\nabla\varphi^* + \sigma\varphi^* = f + \varphi^*\operatorname{div}\mathbf{u}.$$

Since

$$\|\operatorname{div}\mathbf{u}\varphi^*\|_{W^{s,r}(\Omega)} \le c\big(\|\operatorname{div}\mathbf{u}\|_{W^{s,r}(\Omega)} + \|\operatorname{div}\mathbf{u}\|_{C(\Omega)}\big)\|\varphi^*\|_{W^{s,r}(\Omega)},$$

we have

$$\|\operatorname{div}\mathbf{u}\varphi^*\|_{W^{s,r}(\Omega)} + \|\operatorname{div}\mathbf{u}\varphi^*\|_{C(\Omega)} \le \delta(\|\varphi^*\|_{W^{s,r}(\Omega)} + \|\varphi^*\|_{C(\Omega)}),$$

and the needed result follows from (i) and the contraction mapping principle. □

Appendix A. Proof of Lemmas 1.7 and 2.2.

Proof of Lemma 1.7. Since $\partial\Omega$ belongs to the class C^1, functions φ, ς have the extensions $\overline{\varphi},\overline{\varsigma} \in W^{s,r}(\Omega)\cap W^{1,2}(\Omega)$ such that $\overline{\varphi},\overline{\varsigma}$ are compactly supported in $\mathbb{R}^d$ and

$$\|\overline{\varphi}\|_{W^{s,r}(\mathbb{R}^d)} \le c\|\varphi\|_{W^{s,r}(\Omega)},\ \ \|\overline{\varsigma}\|_{W^{s,r}(\mathbb{R}^d)} \le c\|\varsigma\|_{W^{s,r}(\Omega)}.$$

By virtue of Definition 1.2 and inequality (1.19), function $\mathbf{w}$ has the extension by 0 outside Ω, denoted by $\overline{\mathbf{w}}$, such that

$$\|\overline{\mathbf{w}}\|_{W^{1-s,r'}(\mathbb{R}^d)} \le c\|\mathbf{w}\|_{\mathcal{W}_0^{1-s,r'}(\Omega)}.$$

Obviously we have

$$\mathfrak{B}(\mathbf{w},\varphi,\varsigma) = -\int_{\mathbb{R}^d} \overline{\mathbf{w}}\cdot\nabla\overline{\varphi}\,\overline{\varsigma}\,dx.$$

The following multiplicative inequality is due to Maz'ya [26]. For all $s_1 > 0$, $r_1 > 1$, and $r_1 s_1 < d$,

$$\|uv\|_{W^{s_1,r_1}(\mathbb{R}^d)} \le c(r_1,s_1,d)(\|v\|_{W^{s_1,d/s_1}(\mathbb{R}^d)} + \|v\|_{L^\infty(\mathbb{R}^d)})\|u\|_{W^{s_1,r_1}(\mathbb{R}^d)}. \tag{5.1}$$

By virtue of (5.1), we have

$$\|\overline{\mathbf{w}}\,\overline{\varsigma}\|_{W^{1-s,r'}(\mathbb{R}^d)} \le c\|\overline{\mathbf{w}}\|_{W^{1-s,r'}(\mathbb{R}^d)}\big(\|\overline{\varsigma}\|_{W^{1-s,d/(1-s)}(\mathbb{R}^d)} + \|\overline{\varsigma}\|_{L^\infty(\mathbb{R}^d)}\big).$$

On the other hand, since $r^{-1} - (s-(1-s))/d \le (1-s)/d$ for $sr > d$, embedding inequality (1.21) yields

$$\|\overline{\varsigma}\|_{W^{1-s,d/(1-s)}(\mathbb{R}^d)} \le c\|\overline{\varsigma}\|_{W^{s,r}(\mathbb{R}^d)},\quad \|\overline{\varsigma}\|_{L^\infty(\mathbb{R}^d)} \le c\|\overline{\varsigma}\|_{W^{s,r}(\mathbb{R}^d)}.$$

Thus we get

$$\|\overline{\mathbf{w}}\,\overline{\varsigma}\|_{W^{1-s,r'}(\mathbb{R}^d)} \le c\|\overline{\mathbf{w}}\|_{W^{1-s,r'}(\mathbb{R}^d)}\|\overline{\varsigma}\|_{W^{s,r}(\mathbb{R}^d)}.$$

Next note that the operator $\nabla : W^{1,r}(\mathbb{R}^d) \mapsto W^{0,r}(\mathbb{R}^d)$, $\nabla : W^{0,r}(\mathbb{R}^d) \mapsto W^{-1,r}(\mathbb{R}^d)$ is continuous. By virtue of the basic property of interpolation spaces, the operator $\nabla : W^{s,r}(\mathbb{R}^d) \mapsto W^{s-1,r}(\mathbb{R}^d)$ is continuous for all $s \in [0,1]$. In particular we have $\|\nabla\overline{\varphi}\|_{W^{s-1,r}(\mathbb{R}^d)} \le c\|\overline{\varphi}\|_{W^{s,r}(\mathbb{R}^d)}$. Since $\overline{\varphi}$ and $\overline{\mathbf{w}}\,\overline{\varsigma}$ are compactly supported in $\mathbb{R}^d$, the duality principle implies

$$\begin{aligned}\int_{\mathbb{R}^d} \overline{\mathbf{w}}\,\overline{\varsigma}\nabla\overline{\varphi}\,dx &\le c\|\overline{\mathbf{w}}\,\overline{\varsigma}\|_{W^{1-s,r'}(\mathbb{R}^d)}\|\nabla\overline{\varphi}\|_{W^{s-1,r}(\mathbb{R}^d)}\\ &\le c\|\overline{\mathbf{w}}\|_{W^{1-s,r'}(\mathbb{R}^d)}\|\overline{\varsigma}\|_{W^{s,r}(\mathbb{R}^d)}\|\overline{\varphi}\|_{W^{s,r}(\mathbb{R}^d)},\end{aligned}$$

which completes the proof. □

Proof of Lemma 2.2. By virtue of (1.19), the extension $\overline{\mathbf{w}}$ satisfies the inequalities

$$\|\overline{\mathbf{w}}\|_{W^{1-s,r'}(\mathbb{R}^3)} \le c\|\mathbf{w}\|_{\mathcal{W}_0^{1-s,r'}(\Omega)}, \quad \|\overline{\mathbf{w}}\|_{W^{1,2}(\mathbb{R}^3)} \le c\|\mathbf{w}\|_{W^{1,2}(\Omega)}.$$

On the other hand, the vector field $\mathbf{h}$ has a compactly supported extension $\overline{\mathbf{h}} : \mathbb{R}^3 \to \mathbb{R}^3$ such that $\|\overline{\mathbf{h}}\|_{W^{1+s,r}(\mathbb{R}^3)} \le c\|\mathbf{h}\|_{W^{1+s,r}(\Omega)}$; however, this extension does not vanish outside Ω. Substituting the expression for $\mathscr{A}$ into the formula for $\mathfrak{A}$ and integrating by parts, we conclude that $\mathfrak{A}(\mathbf{w},\mathbf{h})$ equals

$$\int_{\mathbb{R}^3} \mathfrak{g}^{-1}\Big(\nabla\big(((\mathbf{N}_1^*)^{-1}-\mathbf{I})\overline{\mathbf{w}}\big) : \big(\mathbf{N}_1\mathbf{N}_1^*\nabla(\mathbf{N}_1^{-1^*}\overline{\mathbf{h}})\big) + \nabla\overline{\mathbf{w}} : \big(\mathbf{N}_1\mathbf{N}_1^*\nabla(\mathbf{N}_1^{-1^*}\overline{\mathbf{h}}) - \mathfrak{g}\nabla\overline{\mathbf{h}}\big)\Big)\,dx.$$

Since $\|\mathbf{N}_1^{\pm 1} - \mathbf{I}\|_{C^2(\Omega)} \le c\tau^2$, we have

$$\|((\mathbf{N}_1^*)^{-1}-\mathbf{I})\overline{\mathbf{w}}\|_{W^{1-s,r'}(\mathbb{R}^3)} \le c\tau^2\|\overline{\mathbf{w}}\|_{W^{1-s,r'}(\mathbb{R}^3)},$$
$$\|\mathfrak{g}^{-1}(\mathbf{N}_1\mathbf{N}_1^*\nabla(\mathbf{N}_1^{-1^*}\overline{\mathbf{h}})) - \nabla\overline{\mathbf{h}}\|_{W^{s,r}(\mathbb{R}^3)} \le c\tau^2\|\overline{\mathbf{h}}\|_{W^{1+s,r}(\mathbb{R}^3)}.$$

Application of the duality arguments, similar to that in the proof of Lemma 1.7, completes the proof. □

Appendix B. Interpolation. In this section we recall some results from the interpolation theory; see [6] for the proofs. Let A_0 and A_1 be Banach spaces. For $t > 0$ introduce two nonnegative functions $K : A_0 + A_1 \mapsto \mathbb{R}$ and $J : A_0 \cap A_1 \mapsto \mathbb{R}$ defined by

$$K(t,u,A_0,A_1) = \inf_{\substack{u=u_0+u_1\\ u_i\in A_i}} \|u_0\|_{A_0} + t\|u_1\|_{A_1},$$
$$J(t,u,A_0,A_1) = \max\{\|u\|_{A_0}, t\|u\|_{A_1}\}.$$

For each $s \in (0,1)$, $1 < r < \infty$, the K-interpolation space $[A_0,A_1]_{s,r,K}$ consists of all elements $u \in A_0 + A_1$ having the finite norm

$$\|u\|_{[A_0,A_1]_{s,r,K}} = \left(\int_0^\infty t^{-1-sr}K(t,u,A_0,A_1)^r\,dt\right)^{1/r}. \tag{6.1}$$

On the other hand, J-interpolation space $[A_0,A_1]_{s,r,J}$ consists of all elements $u \in A_0 + A_1$ which admit the representation

$$u = \int_0^\infty \frac{v(t)}{t}\,dt, \quad v(t) \in A_1 \cap A_0 \ \text{ for } \ t \in (0,\infty), \tag{6.2}$$

and have the finite norm

$$\|u\|_{[A_0,A_1]_{s,r,J}} = \inf_{v(t)} \left(\int_0^\infty t^{-1-sr}J(t,v(t),A_0,A_1)^r\,dt\right)^{1/r} < \infty, \tag{6.3}$$

where the infimum is taken over the set of all $v(t)$ satisfying (6.2). The first main result of interpolation theory reads as follows: For all $s \in (0,1)$ and $r \in (1,\infty)$ the spaces $[A_0,A_1]_{s,r,K}$ and $[A_0,A_1]_{s,r,J}$ are isomorphic topologically and algebraically. Hence the introduced norms are equivalent, and we can omit indices J and K. The following simple properties of interpolation spaces directly follow from definitions.

(1) If $A_1 \subset A_0$ is dense in A_0, then $[A_0, A_1]_{s,r} \subset A_0$ is dense in A_0.

(2) If $\tilde{A}_i$, $i = 0, 1$, are closed subspaces of A_i, then $[\tilde{A}_0, \tilde{A}_1]_{s,r} \subset [A_0, A_1]_{s,r}$ and $\|u\|_{[A_0,A_1]_{s,r}} \le \|u\|_{[\tilde{A}_0,\tilde{A}_1]_{s,r}}$. One of the important results of the interpolation theory is the following representation for the interpolation of dual spaces. Let A_i be Banach spaces such that $A_1 \cap A_0$ is dense in $A_0 + A_1$. Then the Banach spaces $[(A_0)', (A_1)']_{s,r'}$ and $([A_0, A_1]_{s,r})'$ are isomorphic topologically and algebraically. Hence the spaces can be identified with equivalent norms.

In particular, if $A_1 \subset A_0$, $A_0' \subset A_1'$ are dense in A_0 and A_1', respectively, then $([A_0, A_1]_{s,r})'$ is the completion of A_0' in the $([A_0, A_1]_{s,r})'$-norm.

The following lemma is the central result of the interpolation theory.

LEMMA B.1. *Let A_i, B_i, $i = 0, 1$, be Banach spaces, and let $T : A_i \mapsto B_i$ be a bounded linear operator. Then, for all $s \in (0, 1)$ and $r \in (1, \infty)$, the operator $T : [A_0, A_1]_{s,r} \mapsto [B_0, B_1]_{s,r}$ is bounded and*

$$\|T\|_{\mathcal{L}([A_0,A_1]_{s,r},[X_0,Y_1]_{s,r})} \le \|T\|^s_{\mathcal{L}(A_0,B_0)} \|T\|^{1-s}_{\mathcal{L}(A_1,B_1)}.$$

Now we show that all basic properties of spaces $\mathcal{W}_0^{s,r}$ determined by Definition 1.2 easily follow from previously mentioned results of the interpolation theory. Let Ω be a bounded domain with a boundary of the class C^1 or $\Omega = \mathbb{R}^d$. It is well known (see [46]) that, for all $s \in (0, 1)$ and $r \in (1, \infty)$, the Sobolev space $W^{s,r}(\Omega) = [L^r(\Omega), W^{1,r}(\Omega)]_{s,r}$. Since $\mathcal{W}^{0,r}(\Omega)$ and $\mathcal{W}_0^{1,r}(\Omega)$ are closed subspaces of $W^{0,r}(\mathbb{R}^d)$ and $W^{1,r}(\mathbb{R}^d)$, respectively, the interpolating space $\mathcal{W}_0^{s,r}$ determined by Definition 1.2 satisfies inequality (1.19).

Next note that, by virtue of pairing (1.22), the space $L^{r'}(\Omega)$ can be identified with $(\mathcal{W}_0^{0,r})'$, which is dense in $\mathcal{W}^{-1,r}(\Omega) = (\mathcal{W}_0^{1,r}(\Omega))'$. Therefore, the space $(\mathcal{W}_0^{s,r})'$ is the completion of $L^{r'}(\Omega)$ in the norm of $(\mathcal{W}_0^{s,r}(\Omega))'$, which is exactly equal to the norm of $\mathcal{W}^{-s,r'}(\Omega)$. Hence $(\mathcal{H}_0^{s,r}(\Omega))' = \mathcal{W}^{-s,r'}(\Omega)$, which leads to the duality principle (1.24).

Proof of Lemma 1.3. Finally we show that Lemma 1.3 is a straightforward consequence of classical results on solvability of the first boundary value problem for the Stokes equations. Note that, by virtue of Theorem 6.1 in [9], for any $\mathbf{F} \in \mathcal{H}^{s-1,r}(\Omega)$ and $G \in H^{s,r}(\Omega)$ with $s = 0, 1$, problem (1.26) has a unique solution $\mathbf{u}, \pi$ satisfying inequality

$$\|\mathbf{u}\|_{H^{s+1,r}(\Omega)} + \|\pi\|_{H^{s,r}(\Omega)} \le c(\Omega, r, s)(\|\mathbf{F}\|_{\mathcal{H}^{s-1,r}(\Omega)} + \|G\|_{H^{s,r}(\Omega)}).$$

Thus the relation $(F, G) \mapsto (\mathbf{u}, \pi)$ determines the linear operator $T : \mathcal{H}^{s-1,r}(\Omega) \times H^{s,r}(\Omega) \mapsto H^{s+1,r}(\Omega) \times H^{s,r}(\Omega)$. Therefore, Lemma 1.3 is a consequence of Lemma B.1. □

Appendix C. Change of variables. In this section we derive the equations (1.7). We will write $\mathbf{u}(y)$ and $\varrho(y)$, $y \in \Omega$, and set

$$y = x + \varepsilon \mathbf{T}(x), \quad \mathbf{M}(x) = \mathbf{I} + \varepsilon \mathbf{T}'(x), \quad \tilde{\mathbf{u}}(x) = \mathbf{u}(y(x)), \quad \varrho_\varepsilon(x) = \varrho(y(x)).$$

Thus we get $\mathbf{u}_\varepsilon = N\tilde{\mathbf{u}}$. The Jacobi matrix $\mathbf{M}$ is connected with the matrix $\mathbf{N}$ by the relations

$$\det \mathbf{M} = (\det N)^{1/2} \equiv \mathfrak{g}, \quad \mathbf{M} = \mathfrak{g}\mathbf{N}^{-1}. \tag{7.1}$$

For any function $\phi \in C^1(\Omega)$ we have $\nabla_y \phi = (\mathbf{M}^*)^{-1} \nabla_x \tilde{\phi}$, where $\tilde{\phi}(x) = \phi(y(x))$. It follows from this that the identities

$$\int_{\tilde{\Omega}} (\mathrm{div}_y \mathbf{u})(y(x))\, \tilde{\phi}(x) \det \mathbf{M}\, dx = \int_{\Omega} (\mathrm{div}_y \mathbf{u})(y) \phi(y) \det\ dy = -\int_{\Omega} \mathbf{u} \cdot \nabla_y \phi\, dy$$
$$= -\int_{\tilde{\Omega}} \tilde{\mathbf{u}} \cdot (\mathbf{M}^*)^{-1} \nabla_x \tilde{\phi}(x) \det \mathbf{M}\, dx = \int_{\tilde{\Omega}} \mathrm{div}_x \big((\det \mathbf{M})\, \mathbf{M}^{-1} \tilde{\mathbf{u}} \big)\, \tilde{\phi}(x)\, dx$$

hold true for all $\phi \in C_0^\infty(\Omega)$. On the other hand, by virtue of (7.1) we have $(\det \mathbf{M})\, \mathbf{M}^{-1} \tilde{\mathbf{u}} = \mathbf{u}_\varepsilon(x)$. This leads to the equalities

$$\begin{aligned} (\mathrm{div}_y \mathbf{u})(y(x)) = \mathfrak{g}^{-1} \mathrm{div}_x \big(\mathbf{N}\, \tilde{\mathbf{u}}(x) \big) \equiv \mathfrak{g}^{-1} \mathrm{div}_x \mathbf{u}_\varepsilon(x), \\ \mathrm{div}_y (\varrho \mathbf{u})(y(x)) = \mathfrak{g}^{-1} \mathrm{div}_x (\varrho_\varepsilon \mathbf{u}_\varepsilon), \end{aligned} \tag{7.2}$$

which imply the modified mass balance equation (1.7b). From (7.2) and the identity $(\mathbf{M}^*)^{-1} = \mathfrak{g}^{-1} \mathbf{N}^*$ we obtain

$$\nabla \left(\lambda \mathrm{div} \mathbf{u} - \frac{R}{\epsilon^2} p(\varrho) \right) = \mathfrak{g}^{-1} \mathbf{N}^* \nabla \left(\lambda \mathfrak{g}^{-1} \,\mathrm{div}\, \mathbf{u}_\varepsilon - \frac{R}{\epsilon^2} p(\varrho_\varepsilon) \right). \tag{7.3}$$

Combining (7.2) with the identity $\Delta = \mathrm{div}\, \nabla$ we obtain

$$\begin{aligned} \Delta \mathbf{u}(y) = \mathfrak{g}^{-1}\, \mathrm{div} \left(\mathbf{N} (\mathbf{M}^*)^{-1} \nabla \tilde{\mathbf{u}} \right) \\ = \mathfrak{g}^{-1}\, \mathrm{div} \left(\mathfrak{g}^{-1} \mathbf{N} \mathbf{N}^* \nabla (\mathbf{N}^{-1} \mathbf{u}_\varepsilon) \right) = \mathfrak{g}^{-1} \mathbf{N}^* \left(\Delta \mathbf{u}_\varepsilon - \mathscr{A}(\mathbf{u}_\varepsilon) \right). \end{aligned} \tag{7.4}$$

Next note that the components $(\mathbf{u} \nabla \mathbf{u})_i$ of the vector $\mathbf{u} \nabla \mathbf{u}$ satisfy the equalities

$$(\mathbf{u} \nabla \mathbf{u})_i = \mathbf{u} \cdot \nabla_y u_i = \tilde{\mathbf{u}} \cdot \left((\mathbf{M}^*)^{-1} \nabla \tilde{u}_i \right) = \mathfrak{g}^{-1} \mathbf{N} \tilde{\mathbf{u}} \cdot \nabla \tilde{u}_i = \mathfrak{g}^{-1} \mathbf{u}_\varepsilon \cdot \nabla \big(\mathbf{N}^{-1} \mathbf{u}_\varepsilon \big)_i.$$

This gives

$$\varrho \mathbf{u} \nabla \mathbf{u} = \mathbf{g}^{-1} \mathbf{N}^* \mathscr{B}(\varrho_\varepsilon, \mathbf{u}_\varepsilon, \mathbf{u}_\varepsilon). \tag{7.5}$$

Substituting (7.3)–(7.5) into mass balance equation (1.2a) and multiplying both sides of the resulting equality by $\mathfrak{g}(\mathbf{N}^*)^{-1}$, we obtain modified equation (1.7a).

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] H. Beirao da Veiga, *Stationary motions and the incompressible limit for compressible viscous limit*, Houston J. Math., 13 (1987), pp. 527–544.

[3] H. Beirao da Veiga, *Existence results in Sobolev spaces for a transport equation*, Ricerche Mat., 36 (1987), pp. 173-184.

[4] J. A. Bello, E. Fernández-Cara, J. Lemoine, and J. Simon, *The differentiability of the drag with respect to the variations of a Lipschitz domain in a Navier–Stokes flow*, SIAM J. Control. Optim., 35 (1997), pp. 626–640.

[5] J. A. Bello, E. Fernandez-Cara, and J. Simon, *Optimal shape design for Navier-Stokes flow*, in System Modelling and Optimization, Lecture Notes in Control and Inform. Sci. 180, D. Kall, ed., Springer-Verlag, Berlin, 1992, pp. 481–489.

[6] J. Bergh and J. Löfström, *Interpolation Spaces. An Introduction*, Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[7] R. J. DiPerna and P. L. Lions, *Ordinary differential equations, transport theory and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.

[8] L. C. Evans, *Partial Differential Equations*, AMS, Providence, RI, 1998.

[9] G. Galdi, *An Introduction to the Mathematical Theory of the Navier–Stokes Equations*, *Vol.* VI, Springer-Verlag, Berlin, Heidelberg, New York, 1998.
[10] G. Fichera, *Sulle equazioni differenziali lineari elliptico-paraboliche del secondo ordine*, Atti Accad. Naz. Lincei, Mem. Cl. Sci. Fis. Mat. Nat., Sez. I, 5 (1956), pp. 1–30.
[11] E. Feireisl, *Dynamics of Viscous Compressible Fluids*, Oxford University Press, Oxford, UK, 2004.
[12] E. Feireisl, A. H. Novotný, and H. Petzeltová, *On the domain dependence of solutions to the compressible Navier–Stokes equations of a barotropic fluid*, Math. Methods Appl. Sci., 25 (2002), pp. 1045–1073.
[13] E. Feireisl, *Shape optimization in viscous compressible fluids*, Appl. Math. Optim., 47 (2003), pp. 59–78.
[14] J. Frehse, S. Goj, and M. Steinhauer, *L^p-estimates for the Navier–Stokes equations for steady compressible flow*, Manuscripta Math., 116 (2005), pp. 265-275.
[15] J. G. Heywood and M. Padula, *On the uniqueness and existence theory for steady compressible viscous flow*, in Fundamental Directions in Mathematical Fluid Mechanics, Adv. Math. Fluid Mech., Birkhäuser, Basel, 2000, pp. 171–189.
[16] L. Hörmander, *Pseudo-differential operators and non-elliptic boundary value problems*, Ann. of Math. (2), 83 (1966), pp. 129–209.
[17] W. Jager and A. Mikelic, *Couette flow over a rough boundary and drag reduction*, Comm. Math. Phys., 232 (2003), pp. 429–455.
[18] B. Kawohl, O. Pironneau, L. Tartar, and J. Zolesio, *Optimal Shape Design*, Lecture Notes in Math. 1740, Springer-Verlag, Berlin, 2000.
[19] J. J. Kohn and L. Nirenberg, *Degenerate elliptic-parabolic equations of second order*, Comm. Pure Appl. Math., 20 (1967), pp. 797–872.
[20] J. R. Kweon and R. B. Kellogg, *Compressible Navier–Stokes equations in a bounded domain with inflow boundary condition*, SIAM J. Math. Anal., 28 (1997), pp. 94–108.
[21] J. R. Kweon and R. B. Kellogg, *Regularity of solutions to the Navier–Stokes equations for compressible barotropic flows on a polygon*, Arch. Ration. Mech. Anal., 163 (2000), pp. 36–64.
[22] L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, *Vol.* 6, *Fluid Mechanics*, Pergamon Press, Oxford, UK, 1987.
[23] N. S. Landkof, *Foundations of Modern Potential Theory*, Springer-Verlag, Berlin, Heidelberg, New York, 1972.
[24] P. L. Lions, *Mathematical Topics in Fluid Dynamics*, *Volume* 1, *Incompressible Models*, Clarendon Press, Oxford, UK, 1996.
[25] P. L. Lions, *Mathematical Topics in Fluid Dynamics*, *Volume* 2, *Compressible Models*, Clarendon Press, Oxford, UK, 1998.
[26] V. G. Maz'ya and T. O. Shaposhnikova, *Multipliers in Spaces of Differential Functions*, Leningrad University, Leningrad, 1986.
[27] B. Mohammadi and O. Pironneau, *Shape optimization in fluid mechanics*, in Annual Review of Fluid Mechanics, Annu. Rev. Fluid Mech. 36, Annual Reviews, Palo Alto, CA, 2004, pp. 255–279.
[28] M. Moubachir and J.-P. Zolesio, *Moving Shape Analysis And Control: Applications to Fluid Structure Interactions*, Chapman & Hall/CRC, Boca Raton, FL, 2006.
[29] A. Noiri, F. Poupaund, and Y. Demay, *An existence theorem for the multi-fluid Stokes problem*, Quart. Appl. Math., 55 (1997), pp. 421–435.
[30] A. Novotny, *About steady transport equation.* I. *L^p-approach in domains with smooth boundaries*, Comment. Math. Univ. Carolin., 37 (1996), pp. 43–89.
[31] A. Novotny, *About steady transport equation.* II. *Schauder estimates in domains with smooth boundaries*, Portugal. Math., 54 (1997), pp. 317–333.
[32] A. Novotný and M. Padula, *Existence and uniqueness of stationary solutions for viscous compressible heat conductive fluid with large potential and small non-potential external forces*, Siberian Math. J., 34 (1993), pp. 120–146.
[33] A. Novotný and I. Straškraba, *Introduction to the Mathematical Theory of Compressible Flow*, Oxford Lecture Ser. Math. Appl. 27, Oxford University Press, Oxford, UK, 2004.
[34] O. A. Oleinik and E. V. Radkevich, *Second order equations with non-negative characteristic form*, AMS, Providence, RI, Plenum Press, New York, London, 1973.
[35] M. Padula, *Existence and uniqueness for viscous steady compressible motions*, Arch. Rational Mech. Anal., 97 (1986), pp. 1–20.
[36] M. Padula, *Steady flows of barotropic viscous fluids*, in Classical Problems in Mechanics, Quad. Mat. 1, Dipartamento di Matematika, Seconda Universita di Napoli, Caserta, Italy, 1997, pp. 253–345.

[37] P. I. Plotnikov and J. Sokolowski, *On compactness, domain dependence, and existence of steady state solutions to compressible isothermal Navier-Stokes equations*, J. Math. Fluid Mech., 7 (2005), pp. 529–573.

[38] P. I. Plotnikov and J. Sokolowski, *Concentrations of solutions to time-discretized compressible Navier-Stokes equations*, Comm. Math. Phys., 258 (2005), pp. 567–608.

[39] P. I. Plotnikov and J. Sokolowski, *Stationary boundary value problems for Navier-Stokes equations with adiabatic index* $\gamma < 3/2$, Doklady Mathematics, 70 (2004), pp. 535–538 (in English); Dok. Akad. Nauk, 397 (2004), pp. 166–169 (in Russian).

[40] P. I. Plotnikov and J. Sokolowski, *Domain dependence of solutions to compressible Navier–Stokes equations*, SIAM J. Control Optim., 45 (2006), pp. 1165–1197.

[41] H. Schlichting, *Boundary-Layer Theory*, McGraw-Hill Ser. Mech. Engrg., McGraw-Hill, New York, 1955.

[42] J. Simon, *Domain variation for drag in Stokes flow*, in Control Theory of Distributed Parameter Systems and Applications, Lecture Notes in Control and Inform. Sci. 159, Springer-Verlag, Berlin, 1991, pp. 28–42.

[43] T. Slawig, *A formula for the derivative with respect to domain variations in Navier–Stokes flow based on an embedding domain method*, SIAM J. Control Optim., 42 (2003), pp. 495–512.

[44] T. Slawig, *An explicit formula for the derivative of a class of cost functionals with respect to domain variations in Stokes flow*, SIAM J. Control Optim., 39 (2000), pp. 141–158.

[45] J. Sokolowski and J.-P. Zolésio, *Introduction to Shape Optimization. Shape Sensitivity Analysis*, Springer Ser. Comput. Math. 16, Springer-Verlag, Berlin, 1992.

[46] L. Tartar, *An introduction to Sobolev spaces and interpolation spaces*, Lecture Notes of the Unione Matematica Italiana 3, Springer-Verlag, Berlin, and UMI, Bologna, Italy, 2007.

SIAM J. MATH. ANAL.
Vol. 40, No. 3, pp. 1201–1245

QUASI-STATIC EVOLUTION FOR A MODEL IN STRAIN GRADIENT PLASTICITY*

ALESSANDRO GIACOMINI† AND LUCA LUSSARDI†

Abstract. We prove the existence of a quasi-static evolution for a model in strain gradient plasticity proposed by Gurtin and Anand concerning isotropic, plastically irrotational materials under small deformations. This is done by means of the energetic approach to rate-independent evolution problems. Finally we study the asymptotic behavior of the evolution as the strain gradient length scales tend to zero recovering in the limit a quasi-static evolution in perfect plasticity.

Key words. variational models, energy minimization, quasi-static evolution, higher-order stress, flow rule

AMS subject classifications. 74D99, 74C05, 74G65, 49J45.

DOI. 10.1137/070708202

1. Introduction.

Since the early attempts of Aifantis [3], strain gradient plasticity models have been proposed in order to capture [illegible] size effects in metals such as [illegible] and [illegible]. These effects, which take place approximately at the scale of 500 nm–50 μm, cannot be modeled by conventional theories of plasticity. This fact led to the development of continuum theories of plasticity that incorporate size dependence by accounting for strain gradients, namely, the gradient of plastic strains. Following the classical papers by Nye [30] and by Ashby [4, 5], strain gradients induce [illegible], and these dislocations together with [illegible] are mainly responsible for size effects.

Several strain gradient theories, different from one another, have been recently proposed by different authors [1, 7, 10, 11, 14, 15, 16, 17, 18, 20, 12, 21]. In this paper we focus on the theory proposed by Gurtin and Anand [19]. In the context of small deformations, and in the absence of plastic rotation, the strain gradient dependence enters the model via a microstress associated to the gradient of the plastic strain and by a free energy dependent of the macroscopic Burgers tensor.

Let $\Omega \subseteq \mathbb{R}^3$ be the reference configuration of the body. The strain $(\mathbf{E}u)_{ij} := (\partial_i u_j + \partial_j u_i)/2$ of the displacement $u : \Omega \to \mathbb{R}^3$ is decomposed as usual in the form

$$\mathbf{E}u = \mathbf{E}^{\mathrm{e}} + \mathbf{E}^{\mathrm{p}}, \tag{1.1}$$

where $\mathbf{E}^{\mathrm{e}} \in M^{3\times 3}_{\mathrm{sym}}$ is the [illegible] while $\mathbf{E}^{\mathrm{p}}$ is referred to as the [illegible]. It is assumed that $\mathbf{E}^{\mathrm{p}}$ has zero trace; i.e., $\mathbf{E}^{\mathrm{p}}$ belongs to the space of deviatoric matrices $M^{3\times 3}_D$. Besides the usual Cauchy stress $\mathbf{T}$, which satisfies the classical macroscopic force balance, the stress configuration of the system is described by [illegible] tensor $\mathbf{T}^{\mathrm{p}}$ and a [illegible] tensor $\mathbb{K}^{\mathrm{p}}$ which satisfy the equilibrium condition

$$\mathbf{T}_D = \mathbf{T}^{\mathrm{p}} - \mathrm{div}\mathbb{K}^{\mathrm{p}}. \tag{1.2}$$

*Received by the editors November 14, 2004; accepted for publication (in revised form) April 22, 2008; published electronically October 29, 2008.
http://www.siam.org/journals/sima/40-3/70820.html

†Dipartimento di Matematica, Facoltà di Ingegneria, Università degli Studi di Brescia, Via Valotti 9, 25133 Brescia, Italy (alessandro.giacomini@ing.unibs.it, luca.lussardi@ing.unibs.it).

Here $\mathbf{T}_D$ denotes the deviatoric part of $\mathbf{T}$, i.e., $\mathbf{T}_D := \mathbf{T} - \frac{1}{3}\mathrm{tr}(\mathbf{T})Id$. The triple $(\mathbf{T}, \mathbf{T}^{\mathrm{p}}, \mathbb{K}^{\mathrm{p}})$ furnishes the internal power expenditure within a subbody $\mathcal{B} \subseteq \Omega$ by means of the relation

$$\mathcal{W}_{\mathrm{int}}(\mathcal{B}) = \int_{\mathcal{B}} (\mathbf{T} : \dot{\mathbf{E}}^{\mathrm{e}} + \mathbf{T}^{\mathrm{p}} : \dot{\mathbf{E}}^{\mathrm{p}} + \mathbb{K}^{\mathrm{p}} : \nabla\dot{\mathbf{E}}^{\mathrm{p}})\, dx,$$

where $\dot{\mathbf{E}}^{\mathrm{e}}$ and $\dot{\mathbf{E}}^{\mathrm{p}}$ derive from a virtual velocity $(\dot{u}, \dot{\mathbf{E}}^{\mathrm{e}}, \dot{\mathbf{E}}^{\mathrm{p}})$ of the system. So $\mathbf{T}^{\mathrm{p}}$ and $\mathbb{K}^{\mathrm{p}}$ are stresses conjugated to the plastic strain and its gradient, $\mathbb{K}^{\mathrm{p}}$ being a [illegible] since it is conjugated to the gradient of the plastic strain. The balance equations for $\mathbf{T}$, $\mathbf{T}^{\mathrm{p}}$, and $\mathbb{K}^{\mathrm{p}}$ follow by equating the internal power expenditure to the power expenditure associated to the external loads. This entails also boundary conditions for the normal components of $\mathbf{T}$ and $\mathbb{K}^{\mathrm{p}}$ which are connected to the imposed traction and [illegible] on parts of the boundary (see section 3 for details).

The [illegible] of the system is a function of the elastic strain $\mathbf{E}^{\mathrm{e}}$ and of the macroscopic Burgers tensor $\mathbf{G} = \mathrm{curl}\mathbf{E}^{\mathrm{p}}$. In the separable quadratic isotropic case, it assumes the form

$$\psi = \mu|\mathbf{E}^{\mathrm{e}}_D|^2 + \frac{1}{2}k|\mathrm{tr}\mathbf{E}^{\mathrm{e}}|^2 + \frac{\mu L^2}{2}|\mathrm{curl}\mathbf{E}^{\mathrm{p}}|^2, \tag{1.3}$$

where μ and k are the elastic shear and bulk moduli and L is an energetic length scale. The presence of $\mathrm{curl}\mathbf{E}^{\mathrm{p}}$ inside the free energy accounts for the incompatibility of the tensor field $\mathbf{E}^{\mathrm{p}}$, and so it is connected to the presence of geometrically necessary dislocations in Ω. By means of ψ, the energetic third order tensor $\mathbb{K}^{\mathrm{p}}_{\mathrm{en}}$ is defined as the symmetric-deviatoric part (in the first two subscripts) of $\frac{\partial\psi}{\partial\mathbf{G}}$. This entails a decomposition of $\mathbb{K}^{\mathrm{p}}$ into dissipative and energetic parts $\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}$ and $\mathbb{K}^{\mathrm{p}}_{\mathrm{en}}$, respectively, with

$$\mathbb{K}^{\mathrm{p}} = \mathbb{K}^{\mathrm{p}}_{\mathrm{diss}} + \mathbb{K}^{\mathrm{p}}_{\mathrm{en}}.$$

Let Ω be subject to body forces $f(t)$ and to traction forces $g(t)$ on a part $\partial_N\Omega$ of its boundary, with $t \in [0, T]$. Let $\partial\Omega$ be [illegible]; i.e., [illegible] occurs (see section 3 for details). Let us assume that a displacement $w(t)$ is imposed on $\partial_D\Omega := \partial\Omega\setminus\overline{\partial_N\Omega}$. The laws governing the evolution $(u(t), \mathbf{E}^{\mathrm{e}}(t), \mathbf{E}^{\mathrm{p}}(t))$ of the system are obtained by the thermodynamical requirement

$$\dot{\psi}(\mathcal{B}) \le \mathcal{W}_{\mathrm{int}}(\mathcal{B}),$$

where $\psi(\mathcal{B})$ is the free energy of the subbody $\mathcal{B}$ obtained by integrating (1.3) over $\mathcal{B}$ and $\dot{\psi}(\mathcal{B})$ denotes its time derivative. In order to match such an inequality, Gurtin and Anand propose a flow rule involving $\dot{\mathbf{E}}^{\mathrm{p}}(t)$, $\nabla\dot{\mathbf{E}}^{\mathrm{p}}(t)$, $\mathbf{T}^{\mathrm{p}}(t)$, $\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t)$, a [illegible] $l > 0$, and a hardening internal variable. This law reduces to the usual flow rules of classical plasticity when the length scales l and L are set to zero. In the [illegible], and neglecting the hardening internal variable, it takes the form

$$\mathbf{T}^{\mathrm{p}}(t,x) = S_Y\frac{\dot{\mathbf{E}}^{\mathrm{p}}(t,x)}{d^p(t,x)}, \qquad \mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t,x) = S_Y\frac{l^2\nabla\dot{\mathbf{E}}^{\mathrm{p}}(t,x)}{d^p(t,x)}. \tag{1.4}$$

Here $\dot{\mathbf{E}}^{\mathrm{p}}(t,x)$ and $\nabla\dot{\mathbf{E}}^{\mathrm{p}}(t,x)$ denote the time derivative of $\mathbf{E}^{\mathrm{p}}(t,x)$ and $\nabla\mathbf{E}^{\mathrm{p}}(t,x)$, respectively, S_Y is the yield strength, and

$$d^p(t,x) := \sqrt{|\dot{\mathbf{E}}^{\mathrm{p}}(t,x)|^2 + l^2|\nabla\dot{\mathbf{E}}^{\mathrm{p}}(t,x)|^2}$$

is an [illegible]. The stresses $\mathbf{T}^{\mathrm{p}}(t)$ and $\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t)$ satisfy the [illegible]

$$\sqrt{|\mathbf{T}^{\mathrm{p}}(x)|^2 + l^{-2}|\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(x)|^2} \le S_Y, \tag{1.5}$$

and (1.4) is valid when relation (1.5) holds with equality; $(\dot{\mathbf{E}}^{\mathrm{p}}(t), \nabla\dot{\mathbf{E}}^{\mathrm{p}}(t)) = (0,0)$ otherwise. Notice that, by setting $l = L = 0$, we have $\mathbb{K}^{\mathrm{p}} = 0$ and $\mathbf{T}^{\mathrm{p}} = \mathbf{T}_D$ and (1.4) reduces to the usual flow rule of the von Mises type.

The aim of the paper is to provide an existence result of an evolution for the Gurtin–Anand model in the rate-independent case without hardening. The case with positive hardening has been considered recently by Reddy, Ebobisse, and McBride [31]. Adopting a primal formulation, they study the problem by means of variational inequalities in abstract Hilbert spaces. In the case without hardening, coercivity estimates fail, and the use of the abstract setting is no longer possible. This fact reflects what happens also at the level of classical plasticity, where perfect plasticity deserves an "ad hoc" treatment (see [32, 8]).

Inspired by the recent paper of Dal Maso, DeSimone, and Mora [8] concerning perfect plasticity, we recast the problem of the evolution for the Gurtin–Anand model in the framework of the energetic approach to rate-independent processes developed in [24, 25, 27, 28, 29].

Let us consider $\Omega \subseteq \mathbb{R}^N$ open, bounded, and with a Lipschitz boundary ($N \ge 2$). By means of variational arguments, we first construct a discretized-in-time evolution $(u_{k,i}, \mathbf{E}^{\mathrm{e}}_{k,i}, \mathbf{E}^{\mathrm{p}}_{k,i})$ relative to the nodes t^i_k of a subdivision $0 = t^0_k < t^1_k < \cdots < t^k_k = T$ of the time interval $[0,T]$ with step T/k.

In order to enforce variationally the stress constraint (1.5), we consider the function

$$\mathbf{E}^{\mathrm{p}} \mapsto S_Y \int_\Omega \sqrt{|\mathbf{E}^{\mathrm{p}}|^2 + l^2|\nabla\mathbf{E}^{\mathrm{p}}|^2}\, dx.$$

Since this map has linear growth in $\nabla\mathbf{E}^{\mathrm{p}}$, in order to perform direct minimization, we are naturally led to consider $\mathbf{E}^{\mathrm{p}}$ as a [illegible] $BV(\Omega; \mathrm{M}^{N\times N}_D)$ and to relax the functional to the form

$$\mathcal{H}(\mathbf{E}^{\mathrm{p}}) := S_Y \int_\Omega \sqrt{|\mathbf{E}^{\mathrm{p}}|^2 + l^2|\nabla\mathbf{E}^{\mathrm{p}}|^2}\, dx + lS_Y|D^s\mathbf{E}^{\mathrm{p}}|(\Omega),$$

where $D^s\mathbf{E}^{\mathrm{p}}$ denotes the singular part of the derivative of $\mathbf{E}^{\mathrm{p}}$.

The minimization problem that we consider in order to construct $(u_{k,i+1}, \mathbf{E}^{\mathrm{e}}_{k,i+1}, \mathbf{E}^{\mathrm{p}}_{k,i+1})$ relative to the boundary displacement $w(t^{i+1}_k)$ once constructed $(u_{k,i}, \mathbf{E}^{\mathrm{e}}_{k,i}, \mathbf{E}^{\mathrm{p}}_{k,i})$ is the following:

$$\min_{(u,\mathbf{E}^{\mathrm{e}},\mathbf{E}^{\mathrm{p}})\in\mathcal{A}\left(w\left(t^{i+1}_k\right)\right)} \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}) - \left\langle \mathcal{L}\left(t^{i+1}_k\right), u\right\rangle + \mathcal{H}\left(\mathbf{E}^{\mathrm{p}} - \mathbf{E}^{\mathrm{p}}_{k,i}\right). \tag{1.6}$$

Here $\mathcal{A}(w(t^{i+1}_k))$ is the class of admissible configurations for $w(t^{i+1}_k)$,

$$\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}) := \int_\Omega \left(\mu|\mathbf{E}^{\mathrm{e}}_D|^2 + \frac{1}{2}k|\mathrm{tr}\mathbf{E}^{\mathrm{e}}|^2\right) dx, \qquad \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}) := \frac{\mu L^2}{2}\int_\Omega |\mathrm{curl}\mathbf{E}^{\mathrm{p}}|^2\, dx,$$
$$\langle \mathcal{L}(t), u\rangle := \int_\Omega f(t)\cdot u\, dx + \int_{\partial_N\Omega} g(t)\cdot u\, d\mathcal{H}^{N-1},$$

where $\mathcal{H}^{N-1}$ denotes the $(N-1)$-dimensional Hausdorff measure.

In order to have a well-defined energy in (1.6), it suffices that the elastic strain $\mathbf{E}^{\mathrm{e}}$ and the Burgers tensor $\mathrm{curl}\mathbf{E}^{\mathrm{p}}$ belong to the space of square integrable functions. As a consequence, the class $\mathcal{A}(t_k^i)$ turns out to be defined as the triples $(u, \mathbf{E}^{\mathrm{e}}, \mathbf{E}^{\mathrm{p}})$, with

$$u \in W^{1,\frac{N}{N-1}}\left(\Omega;\mathbb{R}^N\right), \qquad \mathbf{E}^{\mathrm{e}} \in L^2\left(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}}\right),$$

$$\mathbf{E}^{\mathrm{p}} \in BV\left(\Omega;\mathrm{M}^{N\times N}_{D}\right), \qquad \mathrm{curl}\mathbf{E}^{\mathrm{p}} \in L^2\left(\Omega;\mathrm{M}^{N\times N}\right),$$

which satisfy the boundary condition $u = w(t_k^i)$ on $\partial_D\Omega$, and such that the compatibility condition (1.1) holds. Notice that the requirement $u \in W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N)$ follows by (1.1) and by the assumptions on $\mathbf{E}^{\mathrm{e}}$ and $\mathbf{E}^{\mathrm{p}}$ in view of Korn's inequality. We assume that $f(t) \in L^N(\Omega;\mathbb{R}^N)$ and $g(t) \in L^N(\partial_N\Omega;\mathbb{R}^N)$ so that the work $\mathcal{L}(t)$ of external forces turns out to be well-defined. The displacement on $\partial_D\Omega$ is assumed to be given by the trace of a map in $W^{1,2}(\Omega;\mathbb{R}^N)$.

The minimum problem (1.6) admits solutions in $\mathcal{A}(w(t_k^i))$ provided that the external loads satisfy a suitable safe load condition (see (4.13)–(4.14)) which appears also in the study of evolutions in perfect plasticity. This condition entails some coercivity in BV for $\mathbf{E}^{\mathrm{p}}$ from the interaction between $\mathcal{H}(\mathbf{E}^{\mathrm{p}} - \mathbf{E}^{\mathrm{p}}_{k,i-1})$ and the linear term $\langle \mathcal{L}(t_k^i), u\rangle$. The existence of a solution for (1.6) follows by applying the direct method of the calculus of variations (Lemma 6.1).

The continuous-in-time evolution is obtained by interpolating the discrete evolution $(u_{k,i}, \mathbf{E}^{\mathrm{e}}_{k,i}, \mathbf{E}^{\mathrm{p}}_{k,i})$ and sending $k \to +\infty$ (section 7). If $w \in AC(0,T;W^{1,2}(\Omega;\mathbb{R}^N))$, $f \in AC(0,T;L^N(\Omega;\mathbb{R}^N))$, $g \in AC(0,T;L^\infty(\partial_N\Omega;\mathbb{R}^N))$, and the safe load condition on f, g holds uniformly in time, we prove the convergence towards a quasi-static evolution $t \mapsto (u(t), \mathbf{E}^{\mathrm{e}}(t), \mathbf{E}^{\mathrm{p}}(t)) \in \mathcal{A}(w(t))$ which is absolutely continuous in time and which satisfies the following two conditions:

(a) Global minimality: For every $(v, \mathbf{e}, \mathbf{p}) \in \mathcal{A}(w(t))$

$$\begin{aligned}(1.7)\quad \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(t)) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)) - \langle \mathcal{L}(t), u(t)\rangle \\ \le \mathcal{Q}_1(\mathbf{e}) + \mathcal{Q}_2(\mathrm{curl}\mathbf{p}) - \langle \mathcal{L}(t), v\rangle + \mathcal{H}(\mathbf{p} - \mathbf{E}^{\mathrm{p}}(t));\end{aligned}$$

(b) Energy balance:

$$\begin{aligned}(1.8)\quad \mathcal{E}(t) + \mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}};0,t) = \mathcal{E}(0) + \int_0^t\int_\Omega \mathbf{T}(\tau) : \mathbf{E}\dot{w}(\tau)\,dx\,d\tau \\ - \int_0^t \langle \dot{\mathcal{L}}(\tau), u(\tau)\rangle\,d\tau - \int_0^t \langle \mathcal{L}(\tau), \dot{w}(\tau)\rangle\,d\tau,\end{aligned}$$

where $\mathbf{T}(t)$ is the Cauchy stress tensor, $\mathcal{E}(t) := \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(t)) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)) - \langle \mathcal{L}(t), u(t)\rangle$, $\dot{\mathcal{L}}(t)$ is associated to $\dot{f}(t)$, $\dot{g}(t)$, and $\mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}};0,t)$ defined as

$$\mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}};a,b) := \sup\left\{\sum_{j=1}^k \mathcal{H}\left(\mathbf{E}^{\mathrm{p}}(t_j) - \mathbf{E}^{\mathrm{p}}(t_{j-1})\right) : a = t_0 < t_1 < \cdots < t_k = b\right\}$$

has the role of a [illegible].

We refer to an evolution satisfying (a) and (b) as a [illegible] for the Gurtin–Anand model (Definition 5.1).

The analysis of the global minimality condition (1.7) leads to the existence of stresses $\mathbf{T}^{\mathrm{p}}(t)$, $\mathbb{K}^{\mathrm{p}}(t)$, and $\mathbb{S}^{\mathrm{p}}(t)$ which together with the Cauchy stress $\mathbf{T}(t)$ satisfy the balance of internal and external powers in Ω

$$\int_\Omega \mathbf{T}(t) : \mathbf{e}\, dx + \int_\Omega \mathbf{T}^{\mathrm{p}}(t) : \mathbf{p}\, dx + \int_\Omega \mathbb{K}^{\mathrm{p}}(t) : \nabla \mathbf{p}\, dx + \langle \mathbb{S}^{\mathrm{p}}(t), D^s \mathbf{p} \rangle = \langle \mathcal{L}(t), v \rangle \tag{1.9}$$

for every virtual velocity $(v, \mathbf{e}, \mathbf{p}) \in \mathcal{A}(0)$ (Lemma 8.1). Notice that a new higher-order stress $\mathbb{S}^{\mathrm{p}}(t)$ conjugated to $D^s\mathbf{E}^{\mathrm{p}}$ appears from our approach: This is somehow natural since $D^s\mathbf{E}^{\mathrm{p}}$ is treated at the same level of $\nabla\mathbf{E}^{\mathrm{p}}$. The balance (1.9) entails the usual balance equation for the Cauchy stress (Proposition 8.2), the balance equation (1.2), the stress constraint (1.5), and the confinement $\|\mathbb{S}^{\mathrm{p}}(t)\| \le l S_Y$ for the singular stress $\mathbb{S}^{\mathrm{p}}(t)$ (Proposition 8.3).

The flow rule (1.4) follows from the analysis of the energy balance equality (1.8) (Proposition 8.10). It is also supplemented by a weak flow rule for the singular stress $\mathbb{S}^{\mathrm{p}}(t)$ (Proposition 8.9).

Concerning the uniqueness of the evolution, it turns out that the maps $t \mapsto \mathbf{E}^{\mathrm{e}}(t)$ and $t \mapsto \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)$ are uniquely determined by the initial conditions (Proposition 8.8). As for the maps $t \mapsto u(t)$ and $t \mapsto \mathbf{E}^{\mathrm{p}}(t)$, we suspect that nonuniqueness can occur as in the case of standard perfect plasticity (see [32, section 2.1]), even if we do not have at the moment an explicit counterexample.

In section 9, we study the asymptotic behavior of a quasi-static evolution for the Gurtin–Anand model when the length scales l and L vanish. As noted previously, by setting l and L equal to zero, the model reduces to the classical model of perfect plasticity of von Mises. Under a suitable assumption on the initial configuration, we prove (Theorem 9.2) that the quasi-static evolution for the Gurtin–Anand model converges in a suitable sense to the evolution for elastic-perfectly plastic bodies in the framework proposed by Dal Maso, DeSimone, and Mora [8]. The main difficulty we have to handle is the change in the mathematical setting of the problem, especially concerning the plastic strain. While in the strain gradient context $\mathbf{E}^{\mathrm{p}}$ is a BV function, in [8] it is modeled simply as a Radon measure.

The paper is organized as follows. In section 2 we fix the notation and recall some basic tools we need from the theory of BV functions. In section 3 we give a brief sketch of the Gurtin–Anand model, while in section 4 we settle the mathematical framework that we adopt in the analysis. The main results are stated in section 5. The existence of a quasi-static evolution is obtained in section 7 after exploiting the convergence of the discrete evolution constructed in section 6. Section 8 is devoted to the proof of the balance equations and the flow rule. Finally section 9 contains the asymptotic analysis as the strain gradient effects vanish.

2. Notation and preliminaries. In this section we recall some basic definitions and results employed in the rest of the paper.

Matrices. We will denote by $\mathrm{M}^{N\times N}$ the space of $N \times N$ matrices $\mathbf{A} = (a_{ij})$, with $a_{ij} \in \mathbb{R}$ endowed with the scalar product

$$\mathbf{A} : \mathbf{B} := \sum_{i,j} a_{ij} b_{ij}. \tag{2.1}$$

The norm of $\mathbf{A}$ induced by the scalar product (2.1) is denoted by $|\mathbf{A}|$.

We will denote by $\mathrm{M}^{N\times N}_{\mathrm{sym}}$ the subspace of symmetric matrices and by $\mathrm{M}^{N\times N}_{D}$ the subspace of $\mathrm{M}^{N\times N}_{\mathrm{sym}}$ of matrices $\mathbf{A}$ with zero trace, that is, such that $\mathrm{tr}\mathbf{A} := \sum_i a_{ii} = 0$.

Given $\mathbf{A} \in \mathrm{M}^{N\times N}_{\mathrm{sym}}$, we denote by $\mathbf{A}_D$ its projection on $\mathrm{M}^{N\times N}_D$, i.e.,

$$\mathbf{A}_D := \mathbf{A} - \frac{1}{N}(\mathrm{tr}A)\mathbf{Id}, \tag{2.2}$$

where $\mathbf{Id}$ is the identity matrix.

The symmetrized gradient of an $\mathbb{R}^N$-valued function $u(x)$ is defined as

$$\mathbf{E}u := \frac{\nabla u + \nabla u^T}{2},$$

where $(\nabla u)_{ij} = \frac{\partial u_i}{\partial x_j}$ is the gradient of u and ∇u^T denotes its transpose.

The gradient, the divergence, and the curl of a $\mathrm{M}^{N\times N}$-valued function $\mathbf{A}(x) = (a_{ij}(x))$ are defined as

$$(\nabla \mathbf{A})_{ijk} := \frac{\partial a_{ij}}{\partial x_k}, \qquad (\mathrm{div}\mathbf{A})_i := \sum_j \frac{\partial a_{ij}}{\partial x_j}, \qquad (\mathrm{curl}\mathbf{A})_{ij} := \sum_{p,q} \epsilon_{ipq}\frac{\partial a_{jq}}{\partial x_p},$$

respectively, where ϵ_{ipq} are the standard permutation symbols.

We will indicate by $\mathrm{M}^{N\times N\times N}$ the space of third order tensors $\mathbb{A} = (a_{ijk})$ with scalar product

$$\mathbb{A} : \mathbb{B} := \sum_{i,j,k} a_{ijk}b_{ijk},$$

and $|\mathbb{A}|$ will denote the induced norm of $\mathbb{A}$.

We say that $\mathbb{A} = (a_{ijk}) \in \mathrm{M}^{N\times N\times N}$ is [illegible] if

$$a_{ijk} = a_{jik} \qquad \text{and} \qquad \sum_p a_{ppk} = 0.$$

We write $\mathbb{A} \in \mathrm{M}^{N\times N\times N}_D$.

The divergence of a $\mathrm{M}^{N\times N\times N}$-valued function $\mathbb{A}(x) = (a_{ijk}(x))$ is given by

$$(\mathrm{div}\mathbb{A})_{ij} := \sum_k \frac{\partial a_{ijk}}{\partial x_k}.$$

Functional spaces and measures. Given $A \subseteq \mathbb{R}^N$ open and $1 \le p < +\infty$, we will denote by $L^p(A;\mathbb{R}^M)$ the space of p-summable functions on A with values in $\mathbb{R}^M$ and by $W^{1,p}(A;\mathbb{R}^M)$ the usual Sobolev space of functions in $L^p(A;\mathbb{R}^M)$ whose derivatives in the sense of distributions belong to L^p. Finally, $\mathcal{M}_b(A;\mathbb{R}^M)$ will denote the space of $\mathbb{R}^M$-valued Radon measures on A, and for every $\mu \in \mathcal{M}_b(A;\mathbb{R}^M)$ we will indicate by $|\mu|(A)$ its total mass. We set $\|\mu\|_{\mathcal{M}_b(A;\mathbb{R}^M)} := |\mu|(A)$. We refer the reader to [9] for the main properties concerning Sobolev spaces and Radon measures.

Let us recall some results from the theory of BV functions. We refer the reader to [2] for an exhaustive treatment of the subject.

We say that $u \in BV(A;\mathbb{R}^M)$ if $u \in L^1(A;\mathbb{R}^M)$, and its distributional derivative Du is a vector-valued Radon measure on A. $BV(A;\mathbb{R}^M)$ is a Banach space with respect to the norm

$$\|u\|_{BV(A;\mathbb{R}^M)} := \|u\|_{L^1(A;\mathbb{R}^M)} + |Du|(A).$$

We will denote by $D^s u$ the singular part of Du with respect to the Lebesgue measure $\mathcal{L}^N$ and by ∇u the density of its absolutely continuous part.

We will say that a sequence $(u_n)_{n\in\mathbb{N}}$ in $BV(A;\mathbb{R}^M)$ converges weakly* in $BV(A;\mathbb{R}^M)$ to $u \in BV(A;\mathbb{R}^M)$ if

$$\begin{aligned} &u_n \to u \qquad \text{strongly in } L^1(A;\mathbb{R}^M),\\ &Du_n \overset{*}{\rightharpoonup} Du \qquad \text{weakly* in } \mathcal{M}_b(A;\mathbb{R}^M). \end{aligned} \tag{2.3}$$

The following compactness result holds: If A is open bounded and with a Lipschitz boundary, every bounded sequence in $BV(A;\mathbb{R}^M)$ admits a subsequence converging weakly* in $BV(A;\mathbb{R}^M)$.

Finally we will use throughout the paper the following embedding property of BV: If A is bounded and with a Lipschitz boundary, then $BV(A;\mathbb{R}^M)$ is continuously embedded into $L^q(A;\mathbb{R}^M)$ for every $1 \le q \le \frac{N}{N-1}$, the embedding being compact for every $1 \le q < \frac{N}{N-1}$.

One-dimensional *AC* and *BV* functions with values in Banach spaces. Let X be a reflexive Banach space or the dual of a separable Banach space. We denote by $BV(a,b;X)$ and $AC(a,b;X)$ the space of functions of bounded variation and the space of absolutely continuous functions from $[a,b]$ to X, respectively. We refer the reader to [6] for the main properties of these spaces. We recall that the variation of $f \in BV(a,b;X)$ is defined as

$$\mathcal{V}(f;a,b) := \sup\left\{ \sum_{j=1}^{k} \|f(t_j) - f(t_{j-1})\|_X \,:\, a = t_0 < t_1 < \cdots < t_k = b \right\}. \tag{2.4}$$

If X is reflexive and $f \in AC(a,b;X)$, then the time derivative $\dot f(t)$ exists for a.e. $t \in [a,b]$. If X is the dual of a separable Banach space (and this is interesting when we consider the plastic strains), the time derivative $\dot f(t)$ exists as a weak* limit of difference quotients for a.e. $t \in [a,b]$ (see [8, Theorem 7.1]).

We will often use the following generalization of Helly's theorem [8, Lemma 7.2] (see also [23, Theorem 3.2]): If X is the dual of a separable Banach space and $(f_k)_{k\in\mathbb{N}}$ a sequence in $BV(a,b;X)$ with $\mathcal{V}(f_k;a,b)$ and $\|f_k(a)\|_X$ uniformly bounded, then there exist $f \in BV(a,b;X)$ and a subsequence $(f_{k_j})_{j\in\mathbb{N}}$ such that $f_{k_j}(t) \overset{*}{\rightharpoonup} f(t)$ weakly* in X for every $t \in [a,b]$.

3. The Gurtin–Anand model. In this section we quickly describe the Gurtin–Anand model [19] in strain gradient plasticity which describes the behavior of isotropic, plastically irrotational materials under small deformations. We present the [illegible] case in which the internal hardening variable is neglected.

Let $\Omega \subseteq \mathbb{R}^N$ be the reference configuration of the body. The starting point of the theory is, as usual, the additive decomposition of the displacement strain $\mathbf{E}u = (\nabla u + \nabla u^T)/2$ into elastic and plastic parts

$$\mathbf{E}u = \mathbf{E}^{\mathrm{e}} + \mathbf{E}^{\mathrm{p}}. \tag{3.1}$$

The symmetric matrices $\mathbf{E}^{\mathrm{e}}$ and $\mathbf{E}^{\mathrm{p}}$ are referred to as the [illegible] and the [illegible], respectively. The plastic part $\mathbf{E}^{\mathrm{p}}$ is supposed to be unable to sustain volumetric changes, so that

$$\mathrm{tr}\mathbf{E}^{\mathrm{p}} = 0;$$

that is, $\mathbf{E}^{\mathrm{p}} \in \mathrm{M}^{N\times N}_D$.

Stresses and balance equations. Given a subbody $\mathcal{B} \subseteq \Omega$, besides the usual Cauchy stress $\mathbf{T} \in \mathrm{M}^{N\times N}_{\mathrm{sym}}$ conjugate to $\mathbf{E}^{\mathrm{e}}$, the analysis of its equilibrium involves also stresses $\mathbf{T}^{\mathrm{p}} \in \mathrm{M}^{N\times N}_{D}$ and $\mathbb{K}^{\mathrm{p}} \in \mathrm{M}^{N\times N\times N}_{D}$ conjugate to $\mathbf{E}^{\mathrm{p}}$ and $\nabla\mathbf{E}^{\mathrm{p}}$, respectively. Given the rate-like kinematical descriptors $(\dot{u}, \dot{\mathbf{E}}^{\mathrm{e}}, \dot{\mathbf{E}}^{\mathrm{p}})$, the power expenditure within $\mathcal{B}$ is given by

$$\mathcal{W}_{\mathrm{int}}(\mathcal{B}) = \int_{\mathcal{B}} (\mathbf{T} : \dot{\mathbf{E}}^{\mathrm{e}} + \mathbf{T}^{\mathrm{p}} : \dot{\mathbf{E}}^{\mathrm{p}} + \mathbb{K}^{\mathrm{p}} : \nabla\dot{\mathbf{E}}^{\mathrm{p}})\, dx.$$

$\mathcal{W}_{\mathrm{int}}(\mathcal{B})$ is balanced by the power expenditure of external forces

$$\mathcal{W}_{\mathrm{ext}}(\mathcal{B}) = \int_{\partial\mathcal{B}} (t(\nu)\cdot\dot{u} + \mathbf{K}(\nu) : \dot{\mathbf{E}}^{\mathrm{p}})\, dA + \int_{\mathcal{B}} f\cdot\dot{u}\, dV,$$

where f is the external body force and $t(\nu)$ is the boundary traction (ν is the outward normal to $\mathcal{B}$) which are associated as usual to $\dot{u}$, while $\mathbf{K}(\nu) \in \mathrm{M}^{N\times N}_{D}$ is a [illegible] associated to the plastic strain rate $\dot{\mathbf{E}}^{\mathrm{p}}$. The balance of power expenditures (that is, $\mathcal{W}_{\mathrm{int}}(\mathcal{B}) = \mathcal{W}_{\mathrm{ext}}(\mathcal{B})$ for every subbody $\mathcal{B}$) leads to the equalities $t(\nu) = \mathbf{T}\nu$ and $\mathbf{K}(\nu) = \mathbb{K}^{\mathrm{p}}\nu$ for the traction and microtraction and to the equilibrium equations

$$-\mathrm{div}\mathbf{T} = f \qquad \text{and} \qquad \mathbf{T}^{\mathrm{p}} = \mathbf{T}_D + \mathrm{div}\mathbb{K}^{\mathrm{p}} \quad \text{in } \Omega,$$

where $\mathbf{T}_D$ is the deviatoric part of $\mathbf{T}$ as defined in (2.2). These equations are supplemented by boundary conditions for $\mathbf{T}$ and $\mathbb{K}^{\mathrm{p}}$. If traction forces g are present on a part $\partial_N\Omega$ of the boundary of Ω, we have as usual

$$\mathbf{T}\nu = g \qquad \text{on } \partial_N\Omega.$$

Concerning $\mathbb{K}^{\mathrm{p}}$, assuming [illegible] at the boundary (see [19, section 8]), we are led to the condition

$$\mathbb{K}^{\mathrm{p}}\nu = 0 \qquad \text{on } \partial\Omega.$$

The free energy. The free energy ψ is assumed to depend on $\mathbf{E}^{\mathrm{e}}$ and $\mathrm{curl}\mathbf{E}^{\mathrm{p}}$: In the quadratic separable case ψ has the form

$$\psi = \frac{1}{2}\mathbb{C}\mathbf{E}^{\mathrm{e}} : \mathbf{E}^{\mathrm{e}} + \frac{\mu L^2}{2}|\mathrm{curl}\mathbf{E}^{\mathrm{p}}|^2, \tag{3.2}$$

where $\mathbb{C}$ is the elastic tensor

$$\mathbb{C}\mathbf{E}^{\mathrm{e}} := 2\mu\mathbf{E}^{\mathrm{e}}_D + k(\mathrm{tr}\mathbf{E}^{\mathrm{e}})\mathbf{I}, \tag{3.3}$$

with μ and k the elastic shear and bulk moduli, respectively. The constant $L > 0$ is an [illegible]. The [illegible] $\mathbb{K}^{\mathrm{p}}_{\mathrm{en}}$ is then defined so that the identity

$$\mu L^2 \mathrm{curl}\mathbf{E}^{\mathrm{p}} : \mathrm{curl}\mathbf{A} = \mathbb{K}^{\mathrm{p}}_{\mathrm{en}} : \nabla\mathbf{A} \tag{3.4}$$

holds for every $\mathrm{M}^{N\times N}$-valued function $\mathbf{A}$. In components we have

$$(\mathbb{K}^{\mathrm{p}}_{\mathrm{en}})_{jqp} := \mu L^2 \left[\frac{\partial \mathbf{E}^{\mathrm{p}}_{jq}}{\partial x_p} - \frac{1}{2}\left(\frac{\partial \mathbf{E}^{\mathrm{p}}_{jp}}{\partial x_q} + \frac{\partial \mathbf{E}^{\mathrm{p}}_{qp}}{\partial x_j}\right) + \frac{1}{N}\delta_{jq}\sum_r \frac{\partial \mathbf{E}^{\mathrm{p}}_{rp}}{\partial x_r}\right], \tag{3.5}$$

where δ_{jq} is the usual Krönecker symbol. The stress $\mathbb{K}^{\mathrm{p}}$ is then additively decomposed in the following way:

$$\mathbb{K}^{\mathrm{p}} = \mathbb{K}^{\mathrm{p}}_{\mathrm{diss}} + \mathbb{K}^{\mathrm{p}}_{\mathrm{en}}.$$

Admissibility of the stresses and the flow rule. Neglecting the hardening internal variable, i.e., if we are in the case without hardening nor softening, the admissibility for the stresses involved in the description of the behavior of Ω reads

$$\sqrt{|\mathbf{T}^{\mathrm{p}}(x)|^2 + l^{-2}|\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(x)|^2} \leq S_Y, \tag{3.6}$$

where $l > 0$ is a [illegible] and S_Y is a [illegible].

Assume now that body and traction forces vary with time, i.e., $f = f(t)$ and $g = g(t)$. The flow rule which drives the system requires that if $(\mathbf{T}^{\mathrm{p}}(t,x), \mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t,x))$ is at the yield surface (that is, (3.6) holds with equality), then

$$\begin{cases} \mathbf{T}^{\mathrm{p}}(t,x) = S_Y \dfrac{\dot{\mathbf{E}}^{\mathrm{p}}(t,x)}{\sqrt{|\dot{\mathbf{E}}^{\mathrm{p}}(t,x)|^2 + l^2|\nabla\dot{\mathbf{E}}^{\mathrm{p}}(t,x)|^2}}, \\ \mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t,x) = S_Y \dfrac{l^2\nabla\dot{\mathbf{E}}^{\mathrm{p}}(t,x)}{\sqrt{|\dot{\mathbf{E}}^{\mathrm{p}}(t,x)|^2 + l^2|\nabla\dot{\mathbf{E}}^{\mathrm{p}}(t,x)|^2}}. \end{cases} \tag{3.7}$$

Here $\dot{\mathbf{E}}^{\mathrm{p}}(t,x)$ and $\nabla\dot{\mathbf{E}}^{\mathrm{p}}(t,x)$ denote the time derivative of $\mathbf{E}^{\mathrm{p}}(t,x)$ and $\nabla\mathbf{E}^{\mathrm{p}}(t,x)$, respectively. If $(\mathbf{T}^{\mathrm{p}}(t,x), \mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t,x))$ is well inside the yield surface, then no plastic phenomenon occurs, i.e., $(\dot{\mathbf{E}}^{\mathrm{p}}, \nabla\dot{\mathbf{E}}^{\mathrm{p}}) = (0,0)$. The flow rule (3.7) is a generalization of the von Mises flow rule in perfect plasticity (set $l = L = 0$, and note that $\mathbf{T}^p = \mathbf{T}_D$). It moreover implies that

$$\int_{\mathcal{B}} \dot{\psi}\, dV \leq \mathcal{W}_{\mathrm{ext}}(\mathcal{B}).$$

The previous inequality reflects the thermodynamical requirement that the increase in free energy of $\mathcal{B}$ is less than or equal to the power expended on $\mathcal{B}$.

4. Functional setting. In this section we state the precise mathematical framework that we adopt to study quasi-static evolutions for the Gurtin–Anand model.

The reference configuration. Let the reference configuration be given by $\Omega \subseteq \mathbb{R}^N$, $N \geq 2$, a bounded open set with a Lipschitz boundary. Let $\partial\Omega$ be partitioned into two open (in the relative topology) disjoint sets $\partial_D\Omega$ and $\partial_N\Omega$ with the same boundary Γ such that $\mathcal{H}^{N-2}(\Gamma) < +\infty$.

Admissible configurations. Let the prescribed boundary displacement on $\partial_D\Omega$ be given by (the trace of) a Sobolev function $w \in W^{1,2}(\Omega;\mathbb{R}^N)$. By an admissible configuration relative to the boundary datum w, we will understand a triple $(u, \mathbf{E}^{\mathrm{e}}, \mathbf{E}^{\mathrm{p}})$ such that

$$u \in W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N), \qquad \mathbf{E}^{\mathrm{e}} \in L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}}), \qquad \mathbf{E}^{\mathrm{p}} \in BV(\Omega;\mathrm{M}^{N\times N}_D), \tag{4.1}$$

with

$$u = w \qquad \text{on } \partial_D\Omega, \tag{4.2}$$

$$\mathbf{E}u = \mathbf{E}^{\mathrm{e}} + \mathbf{E}^{\mathrm{p}}, \tag{4.3}$$

and

$$\mathrm{curl}\mathbf{E}^{\mathrm{p}} \in L^2(\Omega;\mathrm{M}^{N\times N}). \tag{4.4}$$

Equality (4.2) is intended in the sense of traces. Notice that, by the embedding properties of BV, (4.3) entails $\mathbf{E}u \in L^{\frac{N}{N-1}}(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}})$; the requirement $u \in W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N)$ is then consistent with the regularity implied by Korn's inequality in view of the boundary condition (4.2). Let us denote by $\mathcal{A}(w)$ the family of admissible configurations for the boundary datum w, i.e.,

$$\mathcal{A}(w) := \{(u, \mathbf{E}^{\mathrm{e}}, \mathbf{E}^{\mathrm{p}}) \text{ such that (4.1)–(4.4) are satisfied}\}. \tag{4.5}$$

Notice that $\mathcal{A}(w) \neq \emptyset$ since $(w, \mathbf{E}w, 0) \in \mathcal{A}(w)$: Moreover $\mathcal{A}(w)$ contains the triples $(v, \mathbf{E}v, 0)$, with $v \in W^{1,2}(\Omega;\mathbb{R}^N)$ and $v = w$ on $\partial_D\Omega$, so that [illegible] deformations are admissible (and this motivates the choice $w \in W^{1,2}(\Omega;\mathbb{R}^N)$).

The free energy. The free energy of the configuration $(u, \mathbf{E}^{\mathrm{e}}, \mathbf{E}^{\mathrm{p}}) \in \mathcal{A}(w)$ is given according to (3.2) by

$$\Psi(\mathbf{E}^{\mathrm{e}}, \mathrm{curl}\mathbf{E}^{\mathrm{p}}) := \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}),$$

where

$$\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}) := \frac{1}{2}\int_\Omega \mathbb{C}\mathbf{E}^{\mathrm{e}} : \mathbf{E}^{\mathrm{e}}\, dx \tag{4.6}$$

and

$$\mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}) := \frac{1}{2}\mu L^2 \int_\Omega |\mathrm{curl}\mathbf{E}^{\mathrm{p}}|^2\, dx. \tag{4.7}$$

Here $\mathbb{C}$ denotes the elasticity tensor (3.3), and $\mu > 0$ is the elastic shear modulus: We assume also that $k > 0$, so that there exist $0 < \alpha_{\mathbb{C}} \leq \beta_{\mathbb{C}} < +\infty$ such that for every $\mathbf{A} \in \mathrm{M}^{N\times N}_{\mathrm{sym}}$ we have

$$\alpha_{\mathbb{C}}|\mathbf{A}|^2 \leq \mathbb{C}\mathbf{A} : \mathbf{A} \leq \beta_{\mathbb{C}}|\mathbf{A}|^2. \tag{4.8}$$

The yield function $\mathcal{H}$. In order to get variationally the constraint for the stresses conjugated to the plastic variables according to (3.6), we are led to consider the [illegible]

$$\mathcal{H}(\mathbf{E}^{\mathrm{p}}) := S_Y \int_\Omega \sqrt{|\mathbf{E}^{\mathrm{p}}|^2 + l^2|\nabla\mathbf{E}^{\mathrm{p}}|^2}\, dx + lS_Y|D^s\mathbf{E}^{\mathrm{p}}|(\Omega) \tag{4.9}$$

defined for every $\mathbf{E}^{\mathrm{p}} \in BV(\Omega;\mathrm{M}^{N\times N}_D)$. Simple arguments on subadditive and positively one-homogeneous functions on measures (see [13]) show that $\mathcal{H}$ is the relaxation under the L^1-norm of the map

$$\mathbf{E}^{\mathrm{p}} \mapsto S_Y \int_\Omega \sqrt{|\mathbf{E}^{\mathrm{p}}|^2 + l^2|\nabla\mathbf{E}^{\mathrm{p}}|^2}\, dx$$

defined for a regular plastic strain $\mathbf{E}^{\mathrm{p}}$, which is connected to the effective flow rate proposed by Gurtin and Anand (see [19, section 6.3]). As a consequence, $\mathcal{H}$ turns out to be naturally involved in an analysis which employs direct methods of the calculus of variations.

We will often use the lower semicontinuity of $\mathcal{H}$ along weakly* converging sequences, which is a direct consequence of the relaxation process through which $\mathcal{H}$ is obtained.

LEMMA 4.1. $(\mathbf{E}^{\mathrm{p}}_n)_{n\in\mathbb{N}}$ $BV(\Omega;\mathrm{M}^{N\times N}_D)$

$$\mathbf{E}^{\mathrm{p}}_n \overset{*}{\rightharpoonup} \mathbf{E}^{\mathrm{p}} \quad * \quad BV\left(\Omega;\mathrm{M}^{N\times N}_D\right)$$

$\mathbf{E}^{\mathrm{p}} \in BV(\Omega;\mathrm{M}^{N\times N}_D)$

$$\mathcal{H}(\mathbf{E}^{\mathrm{p}}) \le \liminf_{n\to+\infty} \mathcal{H}(\mathbf{E}^{\mathrm{p}}_n).$$

Prescribed boundary displacements and body/traction forces. We assume that the prescribed boundary displacement on $\partial_D\Omega$ is given by (the trace of) a function $w(t,x)$ which is absolutely continuous in time with values in the Sobolev space $W^{1,2}(\Omega;\mathbb{R}^N)$, i.e.,

$$w \in AC\left(0,T;W^{1,2}(\Omega;\mathbb{R}^N)\right). \tag{4.10}$$

Moreover we assume that the prescribed body forces in Ω and traction forces on $\partial_N\Omega$ are given by

$$f \in AC(0,T;L^N(\Omega;\mathbb{R}^N)) \qquad \text{and} \qquad g \in AC(0,T;L^N(\partial_N\Omega;\mathbb{R}^N)). \tag{4.11}$$

For every $t\in[0,T]$ let us consider $\mathcal{L}(t) : W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N) \to \mathbb{R}$ given by

$$\langle \mathcal{L}(t), u\rangle := \int_\Omega f(t)\cdot u\,dx + \int_{\partial_N\Omega} g(t)\cdot u\,d\mathcal{H}^{N-1}. \tag{4.12}$$

Here $\mathcal{H}^{N-1}$ denotes the $(N-1)$-dimensional Hausdorff measure, which is a generalization to arbitrary sets of the usual surface measure (see [9]). By means of the Sobolev embedding theorem it is easily seen that $\mathcal{L}(t)$ is a continuous linear functional on $W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N)$.

Throughout the paper we will assume that the prescribed body and traction forces satisfy the following uniform : We assume that for every $t\in[0,T]$ there exists $\rho(t)\in L^N(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}})$ such that

$$\begin{cases} -\mathrm{div}\rho(t) = f(t) & \text{in } \Omega,\\ \rho(t)\nu = g(t) & \text{on } \partial_N\Omega \end{cases} \tag{4.13}$$

and there exists $\alpha>0$ such that for every $\mathbf{A}\in\mathrm{M}^{N\times N}_D$, with $|\mathbf{A}|\le\alpha$, we have

$$|\mathbf{A}+\rho_D(t)| \le S_Y \qquad \text{a.e. in } \Omega. \tag{4.14}$$

Moreover we assume that $t\mapsto\rho(t)$ and $t\mapsto\rho_D(t)$ are absolutely continuous from $[0,T]$ to $L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}})$ and $L^\infty(\Omega;\mathrm{M}^{N\times N}_D)$, respectively. Notice that the trace condition in (4.13) is well-defined in the dual of the traces on $\partial_N\Omega$ of $W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N)$ since ρ is an L^N-field with divergence in L^N. Moreover, for every $(u,\mathbf{E}^{\mathrm{e}},\mathbf{E}^{\mathrm{p}})\in\mathcal{A}(w)$ we have the following representation formula for $\mathcal{L}(t)$ (here we use $\mathcal{H}^{N-2}(\Gamma)<+\infty$):

$$\langle \mathcal{L}(t), u\rangle = -\langle \rho(t)\nu, w\rangle_{\partial_D\Omega} + \int_\Omega \rho(t) : \mathbf{E}^{\mathrm{e}}\,dx + \int_\Omega \rho_D(t) : \mathbf{E}^{\mathrm{p}}\,dx, \tag{4.15}$$

where the first term on the right-hand side should be interpreted as the pairing between $H^{-1/2}(\partial_D\Omega;\mathbb{R}^N)$ and $H^{1/2}(\partial_D\Omega;\mathbb{R}^N)$.

4.2. Notice that for $\mathcal{L}(t)$ to be well-defined in the dual of $W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N)$ it suffices to require $f(t)\in L^{N/2}(\Omega;\mathbb{R}^N)$ (assume that $N\geq 3$, the case $N=2$ being different in view of Sobolev embedding). But in view of the safe load condition (4.13)–(4.14), $\rho(t)$ would be only an element of $L^{N/2}$ with divergence in $L^{N/2}$, so that its normal trace would be defined in the dual of the traces on $\partial\Omega$ of $W^{1,\frac{N}{N-2}}(\Omega;\mathbb{R}^N)$. Then the representation formula (4.15) would no longer be well-defined (since $w\in W^{1,2}(\Omega;\mathbb{R}^N)$).

As a consequence of the safe load condition, we have the following coercivity estimate for $\mathcal{H}$.

LEMMA 4.3. [illegible] $\mathbf{E}^{\mathrm{p}}\in BV(\Omega;\mathrm{M}^{N\times N}_D)$ [illegible]

$$\mathcal{H}(\mathbf{E}^{\mathrm{p}})-\int_\Omega \rho_D(t):\mathbf{E}^{\mathrm{p}}\,dx\geq \frac{\alpha}{2}\|\mathbf{E}^{\mathrm{p}}\|_{L^1(\Omega;\mathrm{M}^{N\times N}_D)} + \min\left\{l\frac{\alpha}{2}, lS_Y\right\}\|D\mathbf{E}^{\mathrm{p}}\|_{\mathcal{M}_b(\Omega;\mathrm{M}^{N\times N\times N}_D)}. \tag{4.16}$$

[illegible] $\alpha_l>0$ [illegible]

$$\mathcal{H}(\mathbf{E}^{\mathrm{p}})-\int_\Omega \rho_D(t):\mathbf{E}^{\mathrm{p}}\,dx\geq \alpha_l\|\mathbf{E}^{\mathrm{p}}\|_{BV(\Omega;\mathrm{M}^{N\times N}_D)}. \tag{4.17}$$

[illegible]. Notice that by Hölder inequality we have

$$S_Y\int_\Omega\sqrt{|\mathbf{E}^{\mathrm{p}}|^2+l^2|\nabla\mathbf{E}^{\mathrm{p}}|^2}\,dx\geq\sup_{(\tau_1,\tau_2)\in\mathcal{K}}\int_\Omega\left[\tau_1:\mathbf{E}^{\mathrm{p}}+\tau_2:\nabla\mathbf{E}^{\mathrm{p}}\right]\,dx,$$

where

$$\mathcal{K}:=\Big\{(\tau_1,\tau_2)\in L^\infty\left(\Omega;\mathrm{M}^{N\times N}_D\right)\times L^\infty\left(\Omega;\mathrm{M}^{N\times N\times N}_D\right)\,:\, \sqrt{|\tau_1|^2+l^{-2}|\tau_2|^2}\leq S_Y \text{ a.e. in } \Omega\Big\}.$$

We deduce that for every $(\tau_1,\tau_2)\in\mathcal{K}$

$$\mathcal{H}(\mathbf{E}^{\mathrm{p}})-\int_\Omega\rho_D(t):\mathbf{E}^{\mathrm{p}}\,dx\geq\int_\Omega\left[(\tau_1-\rho_D(t)):\mathbf{E}^{\mathrm{p}}+\tau_2:\nabla\mathbf{E}^{\mathrm{p}}\right]\,dx+lS_Y|D^s\mathbf{E}^{\mathrm{p}}|(\Omega),$$

so that in view of (4.14) we get

$$\mathcal{H}(\mathbf{E}^{\mathrm{p}})-\int_\Omega\rho_D(t):\mathbf{E}^{\mathrm{p}}\,dx\geq\int_\Omega\left[\tilde\tau_1:\mathbf{E}^{\mathrm{p}}+\tilde\tau_2:\nabla\mathbf{E}^{\mathrm{p}}\right]\,dx+lS_Y|D^s\mathbf{E}^{\mathrm{p}}|(\Omega)$$

for every $\|\tilde\tau_1\|_{L^\infty(\Omega;\mathrm{M}^{N\times N}_D)}\leq\frac{\alpha}{2}$ and $\|\tilde\tau_2\|_{L^\infty(\Omega;\mathrm{M}^{N\times N\times N}_D)}\leq l\frac{\alpha}{2}$. We conclude that

$$\mathcal{H}(\mathbf{E}^{\mathrm{p}})-\int_\Omega\rho_D(t):\mathbf{E}^{\mathrm{p}}\,dx \geq\frac{\alpha}{2}\|\mathbf{E}^{\mathrm{p}}\|_{L^1(\Omega;\mathrm{M}^{N\times N}_D)}+l\frac{\alpha}{2}\|\nabla\mathbf{E}^{\mathrm{p}}\|_{L^1(\Omega;\mathrm{M}^{N\times N\times N}_D)}+lS_Y|D^s\mathbf{E}^{\mathrm{p}}|(\Omega),$$

so that (4.16) holds. Inequality (4.17) follows by choosing $\alpha_l:=\min\{\frac{\alpha}{2},l\frac{\alpha}{2},lS_Y\}$. □

4.4. As mentioned in the introduction, the safe load condition entails in view of Lemma 4.3 some coercivity in BV for $\mathbf{E}^{\mathrm{p}}$ in the step-by-step minimization problems which we use in section 6 in order to construct a quasi-static evolution (see Lemma 6.1). Those problems contain two terms (dissipation and work of external loads) which have linear growth and which can compete, leading, in general, to a loss of compactness: This cannot happen if a safe load condition is assumed.

5. The main results. Let $T > 0$, and let w, f, and g be the prescribed boundary displacements, body forces, and traction forces according to (4.10) and (4.11), respectively. We assume that f and g satisfy the uniform safe load condition (4.13)–(4.14).

We will denote by $\dot{w}(t)$, $\dot{f}(t)$, and $\dot{g}(t)$ the derivative at time $t \in [0,T]$ of w, f, and g, respectively. Notice that these derivatives exist for a.e. $t \in [0,T]$ since the maps are absolutely continuous with values in a reflexive Banach space. We will denote by $\dot{\mathcal{L}}(t)$ the external work associated to $\dot{f}(t)$ and $\dot{g}(t)$.

Given $\mathcal{H}$ as in (4.9), the $\mathcal{H}$-variation on $[a,b] \subseteq [0,T]$ of $t \mapsto \mathbf{E}^{\mathrm{p}}(t)$ is defined as

$$\mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}}; a, b) := \sup\left\{ \sum_{j=1}^{k} \mathcal{H}\left(\mathbf{E}^{\mathrm{p}}(t_j) - \mathbf{E}^{\mathrm{p}}(t_{j-1})\right) \,:\, a = t_0 < t_1 < \cdots < t_k = b \right\}. \tag{5.1}$$

The notion of quasi-static evolution for the Gurtin–Anand model is the following.

DEFINITION 5.1 (quasi-static evolution). [illegible]

$$t \mapsto (u(t), \mathbf{E}^{\mathrm{e}}(t), \mathbf{E}^{\mathrm{p}}(t))$$

[illegible] $[0,T]$ [illegible] $W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N) \times L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}}) \times BV(\Omega;\mathrm{M}^{N\times N}_{D})$ [illegible] $t \in [0,T]$ [illegible] $(u(t), \mathbf{E}^{\mathrm{e}}(t), \mathbf{E}^{\mathrm{p}}(t)) \in \mathcal{A}(w(t))$ [illegible]

(a) Global stability [illegible] $(v, \mathbf{e}, \mathbf{p}) \in \mathcal{A}(w(t))$

$$\begin{aligned} \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(t)) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)) - \langle \mathcal{L}(t), u(t)\rangle \\ \leq \mathcal{Q}_1(\mathbf{e}) + \mathcal{Q}_2(\mathrm{curl}\mathbf{p}) - \langle \mathcal{L}(t), v\rangle + \mathcal{H}(\mathbf{p} - \mathbf{E}^{\mathrm{p}}(t)); \end{aligned} \tag{5.2}$$

(b) Energy balance [illegible] $t \mapsto \mathbf{E}^{\mathrm{p}}(t)$ [illegible] $[0,T]$ [illegible] $BV(\Omega;\mathrm{M}^{N\times N}_{D})$ [illegible]

$$\begin{aligned} \mathcal{E}(t) + \mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}}; 0, t) = \mathcal{E}(0) + \int_0^t \int_\Omega \mathbf{T}(\tau) : \mathbf{E}\dot{w}(\tau)\, dx\, d\tau \\ - \int_0^t \langle \dot{\mathcal{L}}(\tau), u(\tau)\rangle\, d\tau - \int_0^t \langle \mathcal{L}(\tau), \dot{w}(\tau)\rangle\, d\tau, \end{aligned} \tag{5.3}$$

[illegible] $\mathbf{T}(t) := \mathbb{C}\mathbf{E}^{\mathrm{e}}(t)$

$$\mathcal{E}(t) := \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(t)) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)) - \langle \mathcal{L}(t), u(t)\rangle, \tag{5.4}$$

[illegible] $\mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}}; 0, t)$ [illegible] (5.1)

Our first main result is the following existence theorem.

THEOREM 5.2. [illegible] $(u_0, \mathbf{E}^{\mathrm{e}}_0, \mathbf{E}^{\mathrm{p}}_0) \in \mathcal{A}(w(0))$ [illegible]

$$\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}_0) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}_0) - \langle \mathcal{L}(0), u_0\rangle \leq \mathcal{Q}_1(\mathbf{e}) + \mathcal{Q}_2(\mathrm{curl}\mathbf{p}) - \langle \mathcal{L}(0), v\rangle + \mathcal{H}(\mathbf{p} - \mathbf{E}^{\mathrm{p}}_0)$$

[illegible] $(v, \mathbf{e}, \mathbf{p}) \in \mathcal{A}(w(0))$ [illegible] $t \mapsto (u(t), \mathbf{E}^{\mathrm{e}}(t), \mathbf{E}^{\mathrm{p}}(t))$ [illegible] $(u(0), \mathbf{E}^{\mathrm{e}}(0), \mathbf{E}^{\mathrm{p}}(0)) = (u_0, \mathbf{E}^{\mathrm{e}}_0, \mathbf{E}^{\mathrm{p}}_0)$

Theorem 5.2 will be proved in section 7 by exploiting the convergence of a discrete-in-time evolution constructed through variational arguments in section 6.

Our second main result shows that a quasi-static evolution satisfies the required constitutive equations, balance equations, and the flow rule of the Gurtin–Anand model.

THEOREM 5.3. [illegible] $t \mapsto (u(t), \mathbf{E}^{\mathrm{e}}(t), \mathbf{E}^{\mathrm{p}}(t))$ [illegible] $t \mapsto u(t), t \mapsto \mathbf{E}^{\mathrm{e}}(t), t \mapsto \mathbf{E}^{\mathrm{p}}(t), t \mapsto \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)$ [illegible] $[0,T]$, $W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N)$, $L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}})$, $BV(\Omega;\mathrm{M}^{N\times N}_D)$ [illegible] $L^2(\Omega;\mathrm{M}^{N\times N})$ [illegible] $t \to \mathbf{E}^{\mathrm{e}}(t)$ [illegible] $t \mapsto \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)$ [illegible] $t \in [0,T]$

(a) Cauchy stress $\mathbf{T}(t) = \mathbb{C}\mathbf{E}^{\mathrm{e}}(t)$ [illegible] fi [illegible]

$$\begin{cases} -\mathrm{div}\mathbf{T}(t) = f(t) & \text{in } \Omega, \\ \mathbf{T}(t)\nu = g(t) & \text{on } \partial_N\Omega. \end{cases} \tag{5.5}$$

(b) Stresses conjugated to the plastic variables [illegible] $\mathbf{T}^{\mathrm{p}}(t) \in L^\infty(\Omega;\mathrm{M}^{N\times N}_D)$, $\mathbb{K}^{\mathrm{p}}(t) \in L^2(\Omega;\mathrm{M}^{N\times N\times N}_D)$, $\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t) \in L^\infty(\Omega;\mathrm{M}^{N\times N\times N}_D)$, [illegible] $\mathbb{S}^{\mathrm{p}}(t) \in (\mathcal{M}_b(\Omega;\mathrm{M}^{N\times N\times N}_D))^*$ [illegible] fi [illegible] $\mathbb{K}^{\mathrm{p}}_{\mathrm{en}}(t)$ [illegible] (3.5) [illegible] [illegible] $\mathbf{E}^{\mathrm{p}}(t)$ [illegible] [illegible] $\mathbf{T}_D(t) := (\mathbf{T}(t))_D$ [illegible]

$$\mathbb{K}^{\mathrm{p}}(t) = \mathbb{K}^{\mathrm{p}}_{\mathrm{en}}(t) + \mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t) \quad \text{in } \Omega,$$

$$\begin{cases} \mathbf{T}^{\mathrm{p}}(t) = \mathbf{T}_D(t) + \mathrm{div}\mathbb{K}^{\mathrm{p}}(t) & \text{in } \Omega, \\ \mathbb{K}^{\mathrm{p}}(t)\nu = 0 & \text{on } \partial\Omega, \end{cases} \tag{5.6}$$

$$\sqrt{|\mathbf{T}^{\mathrm{p}}(t)|^2 + l^{-2}|\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t)|^2} \le S_Y \quad \text{in } \Omega, \tag{5.7}$$

$$\|\mathbb{S}^{\mathrm{p}}(t)\|_{(\mathcal{M}_b(\mathrm{M}^{N\times N\times N}_D))^*} \le lS_Y,$$

[illegible] $(v, \mathbf{e}, \mathbf{p}) \in \mathcal{A}(0)$

$$\int_\Omega \mathbf{T}(t) : \mathbf{e}\, dx + \int_\Omega \mathbf{T}^{\mathrm{p}}(t) : \mathbf{p}\, dx + \int_\Omega \mathbb{K}^{\mathrm{p}}(t) : \nabla\mathbf{p}\, dx + \langle \mathbb{S}^{\mathrm{p}}(t), D^s\mathbf{p}\rangle = \langle \mathcal{L}(t), v\rangle.$$

(c) The flow rule [illegible] $\dot{\mathbf{E}}^{\mathrm{p}}(t)$ [illegible] $x \in \Omega$ [illegible] $\dot{\mathbf{E}}^{\mathrm{p}}(t)$, $\nabla\dot{\mathbf{E}}^{\mathrm{p}}(t)$, $\mathbf{T}^{\mathrm{p}}(t)$, [illegible] $\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t)$ [illegible] fi [illegible] (3.7) [illegible] fi

Notice that the normal trace which appears in (5.5) is well-defined in $H^{-1/2}(\partial\Omega;\mathbb{R}^N)$ since $\mathbf{T}(t)$ is an L^2-field with divergence in L^2. Similarly, the normal trace in (5.6) is well-defined in $H^{-1/2}(\partial\Omega;\mathbb{R}^{N\times N})$ because $\mathbb{K}^{\mathrm{p}}(t)$ is an L^2-field (by the definition of $\mathbb{K}^{\mathrm{p}}_{\mathrm{en}}(t)$ and by the constraint (5.7) for $\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t)$) with divergence in L^2 (by the balance equation (5.6) and by the constraint (5.7) for $\mathbf{T}^{\mathrm{p}}(t)$).

In section 9 we will analyze the behavior of a quasi-static evolution as the length scales l and L go to zero, i.e., when the strain gradient effects vanish. We will prove (Theorem 9.2) that the quasi-static evolution converges to an evolution for perfect plasticity according to the framework recently proposed by Dal Maso, DeSimone, and Mora in [8].

6. The discrete-in-time evolution. In this section we construct a discretized-in-time evolution for the Gurtin–Anand model employing a step-by-step minimization procedure. The convergence of this approximated evolution to a quasi-static evolution for the Gurtin–Anand model as the time step discretization goes to zero will be proved in the next section.

Let $k \in \mathbb{N}$, $k \ge 1$, and let us set $t^i_k := \frac{i}{k}T$ for every $i = 0, 1, \dots, k$. Let us set

$$u_{k,0} := u_0, \qquad \mathbf{E}^{\mathrm{e}}_{k,0} := \mathbf{E}^{\mathrm{e}}_0, \qquad \mathbf{E}^{\mathrm{p}}_{k,0} := \mathbf{E}^{\mathrm{p}}_0,$$

where $(u_0, \mathbf{E}^{\mathrm{e}}_0, \mathbf{E}^{\mathrm{p}}_0) \in \mathcal{A}(w(0))$ is the initial configuration of the system given by Theorem 5.2.

Supposing to have constructed $(u_{k,i}, \mathbf{E}^{\mathrm{e}}_{k,i}, \mathbf{E}^{\mathrm{p}}_{k,i}) \in \mathcal{A}(w(t^i_k))$, let $(u_{k,i+1}, \mathbf{E}^{\mathrm{e}}_{k,i+1}, \mathbf{E}^{\mathrm{p}}_{k,i+1}) \in \mathcal{A}(w(t^{i+1}_k))$ $(i = 0, \dots, k-1)$ be a solution of the following minimization problem:

$$\min_{(u,\mathbf{E}^{\mathrm{e}},\mathbf{E}^{\mathrm{p}})\in\mathcal{A}\left(w\left(t^{i+1}_k\right)\right)} \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}) - \langle \mathcal{L}(t^{i+1}_k), u\rangle + \mathcal{H}(\mathbf{E}^{\mathrm{p}} - \mathbf{E}^{\mathrm{p}}_{k,i}). \tag{6.1}$$

The existence of a solution for problem (6.1) is established in the following lemma.

LEMMA 6.1. [illegible] (6.1) [illegible]

[illegible]. The result follows by applying the direct method of the calculus of variations. In fact, let

$$(u_n, \mathbf{E}^{\mathrm{e}}_n, \mathbf{E}^{\mathrm{p}}_n) \in \mathcal{A}(w(t^{i+1}_k))$$

be a minimizing sequence for (6.1). By comparison with $(w(t^{i+1}_k), \mathbf{E}w(t^{i+1}_k), 0)$ we get

$$\begin{aligned}\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}_n) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}_n) - \langle \mathcal{L}\left(t^{i+1}_k\right), u_n\rangle + \mathcal{H}(\mathbf{E}^{\mathrm{p}}_n - \mathbf{E}^{\mathrm{p}}_{k,i})\\ \le \mathcal{Q}_1(\mathbf{E}w(t^{i+1}_k)) - \langle \mathcal{L}(t^{i+1}_k), w(t^{i+1}_k)\rangle + \mathcal{H}(\mathbf{E}^{\mathrm{p}}_{k,i}) := C.\end{aligned}$$

By the representation formula (4.15) for $\mathcal{L}(t^{i+1}_k)$ we deduce that

$$\begin{aligned}\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}_n) - \int_\Omega \rho(t^{i+1}_k) : \mathbf{E}^{\mathrm{e}}_n\, dx + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}_n)\\ + \mathcal{H}(\mathbf{E}^{\mathrm{p}}_n - \mathbf{E}^{\mathrm{p}}_{k,i}) - \int_\Omega \rho_D(t^{i+1}_k) : (\mathbf{E}^{\mathrm{p}}_n - \mathbf{E}^{\mathrm{p}}_{k,i})\, dx\\ \le C + \int_\Omega \rho_D(t^{i+1}_k) : \mathbf{E}^{\mathrm{p}}_{k,i}\, dx - \langle \rho(t^{i+1}_k)\nu, w(t^{i+1}_k)\rangle_{\partial_D\Omega}.\end{aligned}$$

By the coercivity of $\mathcal{Q}_1$ and $\mathcal{Q}_2$ in L^2 and by (4.17) we get

$$\|\mathbf{E}^{\mathrm{e}}_n\|^2_{L^2\left(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}}\right)} + \|\mathrm{curl}\mathbf{E}^{\mathrm{p}}_n\|^2_{L^2(\Omega;\mathrm{M}^{N\times N})} + \|\mathbf{E}^{\mathrm{p}}_n - \mathbf{E}^{\mathrm{p}}_{k,i}\|_{BV\left(\Omega;\mathrm{M}^{N\times N}_D\right)} \le C_1$$

for some $C_1 > 0$. Up to a subsequence we may assume that

$$\mathbf{E}^{\mathrm{e}}_n \rightharpoonup \mathbf{E}^{\mathrm{e}} \qquad \text{weakly in } L^2\left(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}}\right)$$

and

$$\mathbf{E}^{\mathrm{p}}_n \overset{*}{\rightharpoonup} \mathbf{E}^{\mathrm{p}} \qquad \text{weakly* in } BV\left(\Omega;\mathrm{M}^{N\times N}_D\right).$$

As a consequence we get $\mathrm{curl}\mathbf{E}^{\mathrm{p}} \in L^2(\Omega;\mathrm{M}^{N\times N})$ and that

$$\mathrm{curl}\mathbf{E}^{\mathrm{p}}_n \rightharpoonup \mathrm{curl}\mathbf{E}^{\mathrm{p}} \qquad \text{weakly in } L^2\left(\Omega;\mathrm{M}^{N\times N}\right).$$

By the compatibility $\mathbf{E}u_n = \mathbf{E}^{\mathrm{e}}_n + \mathbf{E}^{\mathrm{p}}_n$ and by the embedding $BV(\Omega;\mathrm{M}^{N\times N}_D) \hookrightarrow L^{\frac{N}{N-1}}(\Omega;\mathrm{M}^{N\times N}_D)$, we get that $(\mathbf{E}u_n)_{n\in\mathbb{N}}$ is bounded in $L^{\frac{N}{N-1}}(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}})$. In view of the boundary condition $u_n = w_{k,i+1}$ on $\partial_D\Omega$, Korn's inequality implies that $(u_n)_{n\in\mathbb{N}}$ is bounded in $W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N)$. Up to a further subsequence we can thus suppose that

$$u_n \rightharpoonup u \qquad \text{weakly in } W^{1,\frac{N}{N-1}}\left(\Omega;\mathbb{R}^N\right)$$

for some $u \in W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N)$, with $u = w(t_k^{i+1})$ on $\partial_D\Omega$. We clearly have $(u, \mathbf{E}^{\rm e}, \mathbf{E}^{\rm p}) \in \mathcal{A}(w(t_k^{i+1}))$, and by lower semicontinuity ($\mathcal{Q}_1$ and $\mathcal{Q}_2$ are quadratic, $\mathcal{L}(t_k^{i+1})$ is linear, and $\mathcal{H}$ is lower semicontinuous by Lemma 4.1), we deduce that

$$\begin{aligned}\mathcal{Q}_1(\mathbf{E}^{\rm e}) + \mathcal{Q}_2({\rm curl}\mathbf{E}^{\rm p}) - \langle \mathcal{L}(t_k^{i+1}), u\rangle + \mathcal{H}(\mathbf{E}^{\rm p} - \mathbf{E}^{\rm p}_{k,i}) \\ \leq \liminf_{n\to+\infty} \left[\mathcal{Q}_1(\mathbf{E}^{\rm e}_n) + \mathcal{Q}_2({\rm curl}\mathbf{E}^{\rm p}_n) - \langle \mathcal{L}\left(t_k^{i+1}\right), u_n\rangle + \mathcal{H}(\mathbf{E}^{\rm p}_n - \mathbf{E}^{\rm p}_{k,i})\right].\end{aligned}$$

We conclude that $(u, \mathbf{E}^{\rm e}, \mathbf{E}^{\rm p})$ is a minimizer for problem (6.1), so that the proof is concluded. □

The discretized-in-time evolution is obtained by interpolating the data obtained by the minimization procedure described above. Let us set for $t_k^i \leq t < t_k^{i+1}$

$$w_k(t) := w(t_k^i) \qquad \text{and} \qquad \mathcal{L}_k(t) := \mathcal{L}(t_k^i).$$

We collect the main properties of the discretized-in-time evolution (essential for the passage to the limit as the time step discretization goes to zero) in the following proposition.

PROPOSITION 6.2. [illegible] $t \mapsto (u_k(t), \mathbf{E}^{\rm e}_k(t), \mathbf{E}^{\rm p}_k(t))$ [illegible] $t \in [0,T]$ [illegible] $(u_k(0), \mathbf{E}^{\rm e}_k(0), \mathbf{E}^{\rm p}_k(0)) = (u_0, \mathbf{E}^{\rm e}_0, \mathbf{E}^{\rm p}_0)$ [illegible]

(a) $(u_k(t), \mathbf{E}^{\rm e}_k(t), \mathbf{E}^{\rm p}_k(t)) \in \mathcal{A}(w_k(t))$ [illegible] $t \in [0,T]$ [illegible] $(v, \mathbf{e}, \mathbf{p}) \in \mathcal{A}(w_k(t))$ [illegible]

$$\begin{aligned}\mathcal{Q}_1(\mathbf{E}^{\rm e}_k(t)) + \mathcal{Q}_2({\rm curl}\mathbf{E}^{\rm p}_k(t)) - \langle \mathcal{L}_k(t), u_k(t)\rangle \\ \leq \mathcal{Q}_1(\mathbf{e}) + \mathcal{Q}_2({\rm curl}\mathbf{p}) - \langle \mathcal{L}_k(t), v\rangle + \mathcal{H}(\mathbf{E}^{\rm p}_k(t) - \mathbf{p}).\end{aligned} \tag{6.2}$$

(b) [illegible]

$$\mathcal{E}_k(t) := \mathcal{Q}_1(\mathbf{E}^{\rm e}_k(t)) + \mathcal{Q}_2({\rm curl}\mathbf{E}^{\rm p}_k(t)) - \langle \mathcal{L}_k(t), u_k(t)\rangle,$$

[illegible] $t_k^i \leq t < t_k^{i+1}$

$$\begin{aligned}\mathcal{E}_k(t) + \mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\rm p}_k; 0, t) \leq \mathcal{Q}_1(\mathbf{E}^{\rm e}_0) + \mathcal{Q}_2({\rm curl}\mathbf{E}^{\rm p}_0) - \langle \mathcal{L}(0), u_0\rangle \\ + \int_0^{t_k^i}\int_\Omega \mathbb{C}\mathbf{E}^{\rm e}_k(\tau) : \mathbf{E}\dot{w}(\tau)\,dx\,d\tau - \int_0^{t_k^i} \langle \dot{\mathcal{L}}(\tau), u_k(\tau)\rangle\,d\tau \\ - \int_0^{t_k^i} \langle \mathcal{L}_k(\tau), \dot{w}(\tau)\rangle\,d\tau + e_k,\end{aligned} \tag{6.3}$$

[illegible] $e_k \to 0$ [illegible] $k \to +\infty$ [illegible] $\mathcal{D}_{\mathcal{H}}$ [illegible] fi [illegible] (5.1)

(c) [illegible] C [illegible] $k \in \mathbb{N}$ [illegible] $t \in [0,T]$ [illegible]

$$\begin{aligned}\|u_k(t)\|_{W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N)} + \|\mathbf{E}^{\rm e}_k(t)\|_{L^2(\Omega;{\rm M}^{N\times N}_{\rm sym})} \\ + \|{\rm curl}\mathbf{E}^{\rm p}_k(t)\|_{L^2(\Omega;{\rm M}^{N\times N})} + \mathcal{V}(\mathbf{E}^{\rm p}_k; 0, t) \leq C,\end{aligned} \tag{6.4}$$

[illegible] $\mathcal{V}(\mathbf{E}^{\rm p}_k; 0, t)$ [illegible] $\mathbf{E}^{\rm p}_k$ [illegible] $[0,t]$ [illegible] fi [illegible] (2.4) [illegible] $X = BV(\Omega; {\rm M}^{N\times N}_D)$)

[illegible]. For every $t_k^i \leq t < t_k^{i+1}$ let us set

$$u_k(t) := u_{k,i}, \qquad \mathbf{E}^{\rm e}_k(t) := \mathbf{E}^{\rm e}_{k,i}, \qquad \text{and} \qquad \mathbf{E}^{\rm p}_k(t) := \mathbf{E}^{\rm p}_{k,i},$$

where $(u_{k,j}, \mathbf{E}^{\mathrm{e}}_{k,j}, \mathbf{E}^{\mathrm{p}}_{k,j}) \in \mathcal{A}(w(t^j_k))$ is a solution of the minimization problem (6.1). The minimality property (6.2) follows immediately by the subadditivity of $\mathcal{H}$.

Let us prove (6.3). By construction, by comparing $(u_{k,j}, \mathbf{E}^{\mathrm{e}}_{k,j}, \mathbf{E}^{\mathrm{p}}_{k,j})$ with

$$(u_{k,j-1} + w(t^j_k) - w(t^{j-1}_k), \mathbf{E}^{\mathrm{e}}_{k,j-1} + \mathbf{E}w(t^j_k) - \mathbf{E}w(t^{j-1}_k), \mathbf{E}^{\mathrm{p}}_{k,j-1}) \in \mathcal{A}(w(t^j_k)),$$

we get

$$\begin{aligned}(6.5)\quad &\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}_{k,j}) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}_{k,j}) + \mathcal{H}(\mathbf{E}^{\mathrm{e}}_{k,j} - \mathbf{E}^{\mathrm{e}}_{k,j-1}) - \langle \mathcal{L}(t^j_k), u_{k,j}\rangle \\ &\qquad \le \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}_{k,j-1} + \mathbf{E}w(t^j_k) - \mathbf{E}w(t^{j-1}_k)) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}_{k,j-1}) \\ &\qquad\qquad - \langle \mathcal{L}(t^j_k), u_{k,j-1} + w(t^j_k) - w(t^{j-1}_k)\rangle \\ &= \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}_{k,j-1}) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}_{k,j-1}) - \langle \mathcal{L}(t^{j-1}_k), u_{k,j-1}\rangle + \int_{t^{j-1}_k}^{t^j_k}\int_\Omega \mathbb{C}\mathbf{E}^{\mathrm{e}}_k(\tau) : \mathbf{E}\dot w(\tau)\,dx\,d\tau \\ &\qquad\qquad - \int_{t^{j-1}_k}^{t^j_k} \langle \dot{\mathcal{L}}(\tau), u_k(\tau)\rangle\,d\tau - \int_{t^{j-1}_k}^{t^j_k} \langle \mathcal{L}_k(\tau), \dot w(\tau)\rangle\,d\tau + \delta_{k,j},\end{aligned}$$

where

$$\delta_{k,j} := \mathcal{Q}_1(\mathbf{E}w(t^j_k) - \mathbf{E}w(t^{j-1}_k)) - \langle \mathcal{L}(t^j_k) - \mathcal{L}(t^{j-1}_k), w(t^j_k) - w(t^{j-1}_k)\rangle.$$

By summing up from $j = 1$ to $j = i$ we get

$$\begin{aligned}\mathcal{E}_k(t) + \mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}}_k; 0, t) &\le \mathcal{E}_k(0) + \int_0^{t^i_k}\int_\Omega \mathbb{C}\mathbf{E}^{\mathrm{e}}_k(\tau) : \mathbf{E}\dot w(\tau)\,dx\,d\tau \\ &\quad - \int_0^{t^i_k} \langle \dot{\mathcal{L}}(\tau), u_k(\tau)\rangle\,d\tau - \int_0^{t^i_k} \langle \mathcal{L}_k(\tau), \dot w(\tau)\rangle\,d\tau + \sum_{j=1}^{i} \delta_{k,j}.\end{aligned}$$

Since

$$\begin{aligned}\delta_{k,j} &\le \frac{\beta_{\mathbb{C}}}{k}\int_{t^{j-1}_k}^{t^j_k} \|\mathbf{E}\dot w(\tau)\|^2_{L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}})}\,d\tau \\ &\quad + \sup_j \Big\|\mathcal{L}(t^j_k) - \mathcal{L}(t^{j-1}_k)\Big\|_{\left(W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N)\right)^*} \int_{t^{j-1}_k}^{t^j_k} \|\dot w(\tau)\|_{W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N)}\,d\tau,\end{aligned}$$

by setting $e_k := \sum_{j=1}^k \delta_{k,j}$, we get $e_k \to 0$ as $k \to +\infty$. Since

$$\mathcal{E}_k(0) = \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}_0) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}_0) - \langle \mathcal{L}(0), u_0\rangle,$$

inequality (6.3) follows.

Let us prove (6.4). By using the safe load condition on $f$and g, by (4.15) we can rewrite the first inequality of (6.5) in the following form:

$$\begin{aligned}&\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}_{k,j}) - \int_\Omega \rho(t^j_k) : \mathbf{E}^{\mathrm{e}}_{k,j}\,dx + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}_{k,j}) + \mathcal{H}(\mathbf{E}^{\mathrm{e}}_{k,j} - \mathbf{E}^{\mathrm{e}}_{k,j-1}) - \int_\Omega \rho_D(t^j_k) : \mathbf{E}^{\mathrm{p}}_{k,j}\,dx \\ &\qquad \le \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}_{k,j-1} + \mathbf{E}w(t^j_k) - \mathbf{E}w(t^{j-1}_k)) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}_{k,j-1}) \\ &\qquad\qquad - \int_\Omega \rho(t^j_k) : \mathbf{E}^{\mathrm{e}}_{k,j-1}\,dx - \int_\Omega \rho_D(t^j_k) : \mathbf{E}^{\mathrm{p}}_{k,j-1}\,dx,\end{aligned}$$

so that

$$\mathcal{Q}_1(\mathbf{E}^{\rm e}_{k,j}) - \int_\Omega \rho(t^j_k) : \mathbf{E}^{\rm e}_{k,j}\,dx + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\rm p}_{k,j}) + \mathcal{H}(\mathbf{E}^{\rm e}_{k,j} - \mathbf{E}^{\rm e}_{k,j-1})$$
$$- \int_\Omega \rho_D(t^j_k) : (\mathbf{E}^{\rm p}_{k,j} - \mathbf{E}^{\rm p}_{k,j-1})\,dx \le \mathcal{Q}_1(\mathbf{E}^{\rm e}_{k,j-1}) - \int_\Omega \rho(t^{j-1}_k) : \mathbf{E}^{\rm e}_{k,j-1}\,dx$$
$$+ \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\rm p}_{k,j-1}) + \int_{t^{j-1}_k}^{t^j_k}\int_\Omega \mathbb{C}\mathbf{E}^{\rm e}_k(\tau) : \mathbf{E}\dot w(\tau)\,dx\,d\tau$$
$$- \int_{t^{j-1}_k}^{t^j_k}\int_\Omega \dot\rho(\tau) : \mathbf{E}^{\rm e}_k(\tau)\,dx\,d\tau + \tilde\delta_{k,j},$$

where

$$\tilde\delta_{k,j} := \mathcal{Q}_1(\mathbf{E}w(t^j_k) - \mathbf{E}w(t^{j-1}_k)) \le \frac{\beta_{\mathbb{C}}}{k}\int_{t^{j-1}_k}^{t^j_k} \|\mathbf{E}\dot w(\tau)\|^2_{L^2(\Omega;\mathrm{M}^{N\times N}_{\rm sym})}\,d\tau.$$

By summing up from 0 to i we have

$$\mathcal{Q}_1(\mathbf{E}^{\rm e}_{k,i}) - \int_\Omega \rho(t^i_k) : (\mathbf{E}^{\rm e}_{k,i} - \mathbf{E}w(t^i_k))\,dx + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\rm p}_{k,i})$$
$$+ \sum_{j=0}^{i}\left[\mathcal{H}(\mathbf{E}^{\rm e}_{k,j} - \mathbf{E}^{\rm e}_{k,j-1}) - \int_\Omega \rho_D(t^j_k) : (\mathbf{E}^{\rm p}_{k,j} - \mathbf{E}^{\rm p}_{k,j-1})\,dx\right]$$
$$\le \mathcal{Q}_1(\mathbf{E}^{\rm e}_{k,0}) - \int_\Omega \rho(0) : (\mathbf{E}^{\rm e}_{k,0} - \mathbf{E}w_{k,0})\,dx + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\rm p}_{k,0})$$
$$+ \int_0^{t^i_k}\int_\Omega \mathbb{C}\mathbf{E}^{\rm e}_k(\tau) : \mathbf{E}\dot w(\tau)\,dx\,d\tau - \int_0^{t^i_k}\int_\Omega \dot\rho(\tau) : (\mathbf{E}^{\rm e}_k(\tau) - \mathbf{E}w_k(\tau))\,dx\,d\tau + \tilde e_k,$$

where $\tilde e_k := \sum_{j=0}^k \tilde\delta_{k,j} \to 0$ as $k \to +\infty$. Since by (4.17) we have

$$\sum_{j=0}^{i}\left[\mathcal{H}(\mathbf{E}^{\rm e}_{k,j} - \mathbf{E}^{\rm e}_{k,j-1}) - \int_\Omega \rho_D(t^j_k) : (\mathbf{E}^{\rm p}_{k,j} - \mathbf{E}^{\rm p}_{k,j-1})\,dx\right]$$
$$\ge \alpha_l \sum_{j=0}^{i} \|\mathbf{E}^{\rm e}_{k,j} - \mathbf{E}^{\rm e}_{k,j-1}\|_{BV(\Omega;\mathrm{M}^{N\times N}_D)},$$

we deduce that

$$\mathcal{Q}_1(\mathbf{E}^{\rm e}_k(t)) - \int_\Omega \rho(t^i_k) : (\mathbf{E}^{\rm e}_k(t) - \mathbf{E}w_k(t))\,dx + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\rm p}_k(t)) + \alpha_l\mathcal{V}(\mathbf{E}^{\rm e}_k;0,t)$$
$$\le C_1 + \int_0^{t^i_k}\int_\Omega \mathbb{C}\mathbf{E}^{\rm e}_k(\tau) : \mathbf{E}\dot w(\tau)\,dx\,d\tau - \int_0^{t^i_k}\int_\Omega \dot\rho(\tau) : (\mathbf{E}^{\rm e}_k(\tau) - \mathbf{E}w_k(\tau))\;dx\,d\tau \tag{6.6}$$

for some $C_1 > 0$ independent of k and t. Since $\mathcal{Q}_1(\mathbf{E}^{\rm e}_k(t))$ is quadratic, we get that $\|\mathbf{E}^{\rm e}_k(t)\|_{L^2}$ is uniformly bounded in k and t. Hence from (6.6) we deduce also that $\|\mathrm{curl}\mathbf{E}^{\rm p}_k(t)\|_{L^2}$ and $\mathcal{V}(\mathbf{E}^{\rm p}_k;0,t)$ are uniformly bounded with respect to k and t. Since $u_k(t) = w_k(t)$ on $\partial_D\Omega$, by Korn's inequality we have also that $u_k(t)$ is uniformly bounded in $W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N)$ with respect to k and t. The proof of (6.4) is thus concluded. □

7. Existence of a quasi-static evolution and approximation results. In this section we prove that the discrete evolution $t \mapsto (u_k(t), \mathbf{E}^{\mathrm{e}}_k(t), \mathbf{E}^{\mathrm{p}}_k(t))$ given by Proposition 6.2 admits a subsequence converging (in a suitable sense) to a quasi-static evolution for the Gurtin–Anand model. This will be done in Lemmas 7.1, 7.2, and 7.3. Theorem 5.2 will thus follow by combining these lemmas.

LEMMA 7.1. [illegible] $t \mapsto (u_{k_j}(t), \mathbf{E}^{\mathrm{e}}_{k_j}(t), \mathbf{E}^{\mathrm{p}}_{k_j}(t))$ [illegible] $t \mapsto (u_k(t), \mathbf{E}^{\mathrm{e}}_k(t), \mathbf{E}^{\mathrm{p}}_k(t))$ [illegible] $t \mapsto (u(t), \mathbf{E}^{\mathrm{e}}(t), \mathbf{E}^{\mathrm{p}}(t))$ [illegible] $(u(0), \mathbf{E}^{\mathrm{e}}(0), \mathbf{E}^{\mathrm{p}}(0)) = (u_0, \mathbf{E}^{\mathrm{e}}_0, \mathbf{E}^{\mathrm{p}}_0)$ [illegible] $t \in [0,T]$

$$(u(t), \mathbf{E}^{\mathrm{e}}(t), \mathbf{E}^{\mathrm{p}}(t)) \in \mathcal{A}(w(t)),$$

$$u_{k_j}(t) \rightharpoonup u(t) \quad \text{[illegible] } W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N), \tag{7.1}$$

$$\mathbf{E}^{\mathrm{e}}_{k_j}(t) \rightharpoonup \mathbf{E}^{\mathrm{e}}(t) \quad \text{[illegible] } L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}}), \tag{7.2}$$

$$\mathbf{E}^{\mathrm{p}}_{k_j}(t) \overset{*}{\rightharpoonup} \mathbf{E}^{\mathrm{p}}(t) \quad \text{[illegible] } BV(\Omega;\mathrm{M}^{N\times N}_D), \tag{7.3}$$

[illegible]

$$\mathrm{curl}\mathbf{E}^{\mathrm{p}}_{k_j}(t) \rightharpoonup \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t) \quad \text{[illegible] } L^2(\Omega;\mathrm{M}^{N\times N}). \tag{7.4}$$

[illegible] $t \mapsto \mathbf{E}^{\mathrm{p}}(t)$ [illegible] $C > 0$ [illegible] $t \in [0,T]$

$$\|u(t)\|_{W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N)} + \|\mathbf{E}^{\mathrm{e}}(t)\|_{L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}})} + \|\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)\|_{L^2(\Omega;\mathrm{M}^{N\times N})} + \mathcal{V}(\mathbf{E}^{\mathrm{p}};0,t) \le C. \tag{7.5}$$

[illegible] $t \in [0,T]$ [illegible] $(v,\mathbf{e},\mathbf{p}) \in \mathcal{A}(w(t))$ [illegible]

$$\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(t)) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)) - \langle \mathcal{L}(t), u(t)\rangle \le \mathcal{Q}_1(\mathbf{e}) + \mathcal{Q}_2(\mathrm{curl}\mathbf{p}) - \langle \mathcal{L}(t), v\rangle + \mathcal{H}(\mathbf{p} - \mathbf{E}^{\mathrm{p}}(t)). \tag{7.6}$$

[illegible] By Proposition 6.2 we have

$$\|u_k(t)\|_{W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N)} + \|\mathbf{E}^{\mathrm{e}}_k(t)\|_{L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}})} + \|\mathrm{curl}\mathbf{E}^{\mathrm{p}}_k(t)\|_{L^2(\Omega;\mathrm{M}^{N\times N})} + \mathcal{V}(\mathbf{E}^{\mathrm{p}}_k;0,t) \le C \tag{7.7}$$

for some C independent of k and t. Since $\mathbf{E}^{\mathrm{p}}_k(0) = \mathbf{E}^{\mathrm{p}}_0$ and $\mathcal{V}(\mathbf{E}^{\mathrm{p}}_k;0,T) \le C$, the existence of $\mathbf{E}^{\mathrm{p}} \in BV(0,T;BV(\Omega;\mathrm{M}^{N\times N}_D))$ and of a subsequence $t \mapsto (u_{k_j}(t), \mathbf{E}^{\mathrm{e}}_{k_j}(t), \mathbf{E}^{\mathrm{p}}_{k_j}(t))$ such that (7.3) holds follows by applying the generalized version of Helly's theorem proved in [8, Lemma 7.2] (notice that BV can be seen as the dual of a separable Banach space in such a way that the associated convergence with respect to the weak* topology is precisely the weak* convergence defined in (2.3)).

Since weak* convergence in BV implies strong convergence in L^1, by (7.7) we deduce that $\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t) \in L^2(\Omega;\mathrm{M}^{N\times N})$ and that (7.4) holds.

Let us fix $t \in [0,T]$. In view of the coercivity estimate (7.7), we may assume that there exist $\tilde{u} \in W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N)$, $\widetilde{\mathbf{E}^{\mathrm{e}}} \in L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}})$, and a further subsequence k_{j_h} (depending a priori on t) such that

$$u_{k_{j_h}}(t) \rightharpoonup \tilde{u} \qquad \text{weakly in } W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N)$$

and

$$\mathbf{E}^{\mathrm{e}}_{k_{j_h}}(t) \rightharpoonup \widetilde{\mathbf{E}^{\mathrm{e}}} \qquad \text{weakly in } L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}}). \tag{7.8}$$

It follows easily that $(\tilde u, \widetilde{\mathbf{E}^{\mathrm{e}}}, \mathbf{E}^{\mathrm{p}}(t)) \in \mathcal{A}(w(t))$. We claim that for every $(v,\mathbf{e},\mathbf{p}) \in \mathcal{A}(w(t))$ we have

$$\begin{aligned}\mathcal{Q}_1(\widetilde{\mathbf{E}^{\mathrm{e}}}) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)) - \langle \mathcal{L}(t), \tilde u\rangle \\ \le \mathcal{Q}_1(\mathbf{e}) + \mathcal{Q}_2(\mathrm{curl}\mathbf{p}) - \langle \mathcal{L}(t), v\rangle + \mathcal{H}(\mathbf{p}-\mathbf{E}^{\mathrm{p}}(t)).\end{aligned} \tag{7.9}$$

Notice that, in view of (7.9), it turns out that $\tilde u$ and $\widetilde{\mathbf{E}^{\mathrm{e}}}$ are uniquely determined by $\mathbf{E}^{\mathrm{p}}(t)$. In fact, the pair $(\tilde u, \widetilde{\mathbf{E}^{\mathrm{e}}})$ minimizes the convex functional $(v,\mathbf{e}) \mapsto \mathcal{Q}_1(\mathbf{e}) - \langle \mathcal{L}(t), v\rangle$ on the convex set $K := \{(v,\mathbf{e}) : (v,\mathbf{e},\mathbf{E}^{\mathrm{p}}(t)) \in \mathcal{A}(w(t))\}$. Since the functional is strictly convex in $\mathbf{e}$, $\widetilde{\mathbf{E}^{\mathrm{e}}}$ is uniquely determined, and so is $\tilde u$ in view of Korn's inequality. By setting $u(t) := \tilde u$ and $\mathbf{E}^{\mathrm{e}}(t) := \widetilde{\mathbf{E}^{\mathrm{e}}}$, we get that (7.1) and (7.2) hold (without passing to a further subsequence).

In view of (7.7) we deduce that (7.5) holds. Finally, the global stability is given precisely by (7.9).

In order to conclude the proof, we need to prove claim (7.9). Let us set

$$v_h := v - \tilde u + u_{k_{j_h}}(t), \qquad \mathbf{e}_h := \mathbf{e} - \widetilde{\mathbf{E}^{\mathrm{e}}} + \mathbf{E}^{\mathrm{e}}_{k_{j_h}}(t), \qquad \text{and} \qquad \mathbf{p}_h := \mathbf{p} - \mathbf{E}^{\mathrm{p}}(t) + \mathbf{E}^{\mathrm{p}}_{k_{j_h}}(t).$$

We have $(v_h, \mathbf{e}_h, \mathbf{p}_h) \in \mathcal{A}(w_{k_{j_h}}(t))$. By (6.2) we have

$$\begin{aligned}&\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}_{k_{j_h}}(t)) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}_{k_{j_h}}(t)) - \langle \mathcal{L}_{k_{j_h}}(t), u_{k_{j_h}}(t)\rangle \\ &\qquad \le \mathcal{Q}_1(\mathbf{e}_h) + \mathcal{Q}_2(\mathrm{curl}\mathbf{p}_h) - \langle \mathcal{L}_{k_{j_h}}(t), v_h\rangle + \mathcal{H}(\mathbf{p}_h - \mathbf{E}^{\mathrm{p}}_{k_{j_h}}(t)) \\ &\qquad = \mathcal{Q}_1(\mathbf{e} - \widetilde{\mathbf{E}^{\mathrm{e}}} + \mathbf{E}^{\mathrm{e}}_{k_{j_h}}(t)) + \mathcal{Q}_2(\mathrm{curl}\mathbf{p} - \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t) + \mathrm{curl}\mathbf{E}^{\mathrm{p}}_{k_{j_h}}(t)) \\ &\qquad\qquad - \langle \mathcal{L}_{k_{j_h}}(t), v - \tilde u + u_{k_{j_h}}(t)\rangle + \mathcal{H}(\mathbf{p} - \mathbf{E}^{\mathrm{p}}(t)),\end{aligned}$$

so that we get

$$\begin{aligned}0 \le{}& \mathcal{Q}_1(\mathbf{e} - \widetilde{\mathbf{E}^{\mathrm{e}}}) + \int_\Omega \mathbb{C}(\mathbf{e} - \widetilde{\mathbf{E}^{\mathrm{e}}}) : \mathbf{E}^{\mathrm{e}}_{k_{j_h}}(t)\,dx \\ &+ \mathcal{Q}_2(\mathrm{curl}\mathbf{p} - \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)) + \mu L^2 \int_\Omega (\mathrm{curl}\mathbf{p} - \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)) : \mathrm{curl}\mathbf{E}^{\mathrm{p}}_{k_{j_h}}(t)\,dx \\ &\qquad - \langle \mathcal{L}_{k_{j_h}}(t), v - \tilde u\rangle + \mathcal{H}(\mathbf{p} - \mathbf{E}^{\mathrm{p}}(t)).\end{aligned}$$

By letting $h \to +\infty$, in view of (7.8), (7.3), and (7.4) and since $t \mapsto \mathcal{L}(t)$ is absolutely continuous with values in $(W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N))^*$, we obtain

$$\begin{aligned}0 \le{}& \mathcal{Q}_1(\mathbf{e} - \widetilde{\mathbf{E}^{\mathrm{e}}}) + \int_\Omega \mathbb{C}(\mathbf{e} - \widetilde{\mathbf{E}^{\mathrm{e}}}) : \widetilde{\mathbf{E}^{\mathrm{e}}}\,dx \\ &+ \mathcal{Q}_2(\mathrm{curl}\mathbf{p} - \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)) + \mu L^2 \int_\Omega (\mathrm{curl}\mathbf{p} - \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)) : \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)\,dx \\ &\qquad - \langle \mathcal{L}(t), v - \tilde u\rangle + \mathcal{H}(\mathbf{p} - \mathbf{E}^{\mathrm{p}}(t)).\end{aligned}$$

By adding to both sides the term $\mathcal{Q}_1(\widetilde{\mathbf{E}^{\mathrm{e}}}) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)) - \langle \mathcal{L}(t), \tilde u\rangle$, we obtain precisely (7.9), so that the proof is concluded. ∎

We have the following estimate from above for the total energy.

LEMMA 7.2. [illegible] $t \mapsto (u(t), \mathbf{E}^{\mathrm{e}}(t), \mathbf{E}^{\mathrm{p}}(t))$ [illegible] 7.1 [illegible] $t \in [0,T]$ [illegible]

$$\mathcal{E}(t) + \mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}}; 0, t) \le \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}_0) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}_0) - \langle \mathcal{L}(0), u_0\rangle + \int_0^t \int_\Omega \mathbb{C}\mathbf{E}^{\mathrm{e}}(\tau) : \mathbf{E}\dot{w}(\tau)\, dx\, d\tau - \int_0^t \langle \dot{\mathcal{L}}(\tau), u(\tau)\rangle\, d\tau - \int_0^t \langle \mathcal{L}(\tau), \dot{w}(\tau)\rangle\, d\tau, \tag{7.10}$$

[illegible] $\mathcal{E}(t)$ [illegible] $\mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}}; 0, t)$ [illegible] fi[illegible] (5.4) [illegible] (5.1) [illegible]. Let us fix $t \in [0,T]$. By (6.3) we have

$$\mathcal{E}_k(t) + \mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}}_k; 0, t) \le \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}_0) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}_0) - \langle \mathcal{L}(0), u_0\rangle + \int_0^{t^i_k} \int_\Omega \mathbb{C}\mathbf{E}^{\mathrm{e}}_k(\tau) : \mathbf{E}\dot{w}(\tau)\, dx\, d\tau - \int_0^{t^i_k} \langle \dot{\mathcal{L}}(\tau), u_k(\tau)\rangle\, d\tau - \int_0^{t^i_k} \langle \mathcal{L}_k(\tau), \dot{w}(\tau)\rangle\, d\tau + e_k, \tag{7.11}$$

where $e_k \to 0$ as $k \to +\infty$. In view of (7.2), (7.4), and (7.1) and since $\mathcal{L}_k(t) \to \mathcal{L}(t)$ strongly in $(W^{1,\frac{N}{N-1}}(\Omega; \mathbb{R}^N))^*$, we get by lower semicontinuity

$$\mathcal{E}(t) \le \liminf_{j \to +\infty} \mathcal{E}_{k_j}(t).$$

Moreover, by (7.3) and the lower semicontinuity of $\mathcal{H}$ with respect to the weak* convergence in BV, the very definition of $\mathcal{D}_{\mathcal{H}}$ implies that

$$\mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}}; 0, t) \le \liminf_{j \to +\infty} \mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}}_{k_j}; 0, t).$$

By Lebesgue dominated convergence we get as $k \to +\infty$

$$\int_0^{t^i_{k_j}} \int_\Omega \mathbb{C}\mathbf{E}^{\mathrm{e}}_{k_j}(\tau) : \mathbf{E}\dot{w}(\tau)\, dx\, d\tau - \int_0^{t^i_{k_j}} \langle \dot{\mathcal{L}}(\tau), u_{k_j}(\tau)\rangle\, d\tau - \int_0^{t^i_{k_j}} \langle \mathcal{L}_{k_j}(\tau), \dot{w}(\tau)\rangle\, d\tau \to \int_0^t \int_\Omega \mathbb{C}\mathbf{E}^{\mathrm{e}}(\tau) : \mathbf{E}\dot{w}(\tau)\, dx\, d\tau - \int_0^t \langle \dot{\mathcal{L}}(\tau), u(\tau)\rangle\, d\tau - \int_0^t \langle \mathcal{L}(\tau), \dot{w}(\tau)\rangle\, d\tau.$$

Then (7.10) follows passing to the limit in (7.11). □

The following estimate from below for the total energy holds.

LEMMA 7.3. [illegible] $t \mapsto (u(t), \mathbf{E}^{\mathrm{e}}(t), \mathbf{E}^{\mathrm{p}}(t))$ [illegible] 7.1 [illegible] $t \in [0,T]$ [illegible]

$$\mathcal{E}(t) + \mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}}; 0, t) \ge \mathcal{Q}_1\left(\mathbf{E}^{\mathrm{e}}_0\right) + \mathcal{Q}_2\left(\mathrm{curl}\mathbf{E}^{\mathrm{p}}_0\right) - \langle \mathcal{L}(0), u_0\rangle + \int_0^t \int_\Omega \mathbb{C}\mathbf{E}^{\mathrm{e}}(\tau) : \mathbf{E}\dot{w}(\tau)\, dx\, d\tau - \int_0^t \langle \dot{\mathcal{L}}(\tau), u(\tau)\rangle\, d\tau - \int_0^t \langle \mathcal{L}(\tau), \dot{w}(\tau)\rangle\, d\tau, \tag{7.12}$$

[illegible] $\mathcal{E}(t)$ [illegible] $\mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}}; 0, t)$ [illegible] fi[illegible] (5.4) [illegible] (5.1) [illegible]

. Let $t \in [0,T]$, $h \geq 1$, and let us set $s_h^j := \frac{j}{h}t$ for $j = 0, 1, \dots h$. By the global stability condition (7.6), by comparing $(u(s_h^j), \mathbf{E}^{\mathrm{e}}(s_h^j), \mathbf{E}^{\mathrm{p}}(s_h^j))$ with

$$\Big(u(s_h^{j+1}) - w(s_h^{j+1}) + w(s_h^j), \\ \mathbf{E}^{\mathrm{e}}(s_h^{j+1}) - \mathbf{E}w(s_h^{j+1}) + \mathbf{E}w(s_h^j), \mathbf{E}^{\mathrm{p}}(s_h^{j+1})\Big) \in \mathcal{A}(w(s_h^j)),$$

we get

$$\begin{aligned}\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(s_h^{j+1}) - \mathbf{E}w(s_h^{j+1}) + \mathbf{E}w(s_h^j)) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(s_h^{j+1})) \\ - \langle \mathcal{L}(s_h^j), u(s_h^{j+1}) - w(s_h^{j+1}) + w(s_h^j)\rangle + \mathcal{H}(\mathbf{E}^{\mathrm{p}}(s_h^{j+1}) - \mathbf{E}^{\mathrm{p}}(s_h^j)) \\ \geq \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(s_h^j)) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(s_h^j)) - \langle \mathcal{L}(s_h^j), u(s_h^j)\rangle,\end{aligned}$$

which can be rewritten in the following form:

$$\begin{aligned}&\mathcal{Q}_1\left(\mathbf{E}^{\mathrm{e}}\left(s_h^{j+1}\right)\right) + \mathcal{Q}_2\left(\mathrm{curl}\mathbf{E}^{\mathrm{p}}\left(s_h^{j+1}\right)\right) - \langle \mathcal{L}\left(s_h^{j+1}\right), u\left(s_h^{j+1}\right)\rangle \\ &+\mathcal{H}\left(\mathbf{E}^{\mathrm{p}}\left(s_h^{j+1}\right) - \mathbf{E}^{\mathrm{p}}\left(s_h^j\right)\right) \geq \mathcal{Q}_1\left(\mathbf{E}^{\mathrm{e}}\left(s_h^j\right)\right) + \mathcal{Q}_2\left(\mathrm{curl}\mathbf{E}^{\mathrm{p}}\left(s_h^j\right)\right) - \langle \mathcal{L}\left(s_h^j\right), u\left(s_h^j\right)\rangle \\ &+\int_{s_h^j}^{s_h^{j+1}}\int_\Omega \mathbb{C}\overline{\mathbf{E}_h^{\mathrm{e}}}(s) : \mathbf{E}\dot{w}(s)\,dx\,ds - \int_{s_h^j}^{s_h^{j+1}} \langle \dot{\mathcal{L}}(s), \bar{u}_h(s)\rangle\,ds - \int_{s_h^j}^{s_h^{j+1}} \langle \overline{\mathcal{L}_h}(s), \dot{w}(s)\rangle\,ds + \bar{\delta}_{h,j},\end{aligned} \tag{7.13}$$

where for $s_h^j < s \leq s_h^{j+1}$ we set

$$\overline{u_h}(s) := u(s_h^{j+1}), \quad \overline{\mathbf{E}_h^{\mathrm{e}}}(s) := \mathbf{E}^{\mathrm{e}}(s_h^{j+1}), \quad \overline{\mathbf{E}_h^{\mathrm{p}}}(s) := \mathbf{E}^{\mathrm{p}}(s_h^{j+1}), \quad \overline{\mathcal{L}_h}(s) := \mathcal{L}(s_h^{j+1}),$$

and

$$\bar{\delta}_{h,j} := -\mathcal{Q}_1(\mathbf{E}w(s_h^{j+1}) - \mathbf{E}w(s_h^j)) - \langle \mathcal{L}(s_h^{j+1}) - \mathcal{L}(s_h^j), w(s_h^{j+1}) - w(s_h^j)\rangle.$$

By summing up in (7.13) from 0 to $h-1$ we get

$$\begin{aligned}&\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(t)) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)) - \langle \mathcal{L}(t), u(t)\rangle + \sum_{j=0}^{h-1} \mathcal{H}(\mathbf{E}^{\mathrm{p}}(s_h^{j+1}) - \mathbf{E}^{\mathrm{p}}(s_h^j)) \\ &\qquad \geq \mathcal{Q}_1(\mathbf{E}_0^{\mathrm{e}}) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}_0^{\mathrm{p}}) - \langle \mathcal{L}(0), u_0\rangle \\ &\quad + \int_0^t \int_\Omega \mathbb{C}\overline{\mathbf{E}_h^{\mathrm{e}}}(s) : \mathbf{E}\dot{w}(s)\,dx\,ds - \int_0^t \langle \dot{\mathcal{L}}(s), \overline{u_h}(s)\rangle\,ds - \int_0^t \langle \overline{\mathcal{L}_h}(s), \dot{w}(s)\rangle\,ds + \bar{e}_h,\end{aligned}$$

where $\bar{e}_h \to 0$ as $h \to +\infty$. By the very definition of $\mathcal{D}_{\mathcal{H}}$ we get

$$\begin{aligned}\mathcal{E}(t) + \mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}}; 0, t) \geq \mathcal{Q}_1(\mathbf{E}_0^{\mathrm{e}}) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}_0^{\mathrm{p}}) - \langle \mathcal{L}(0), u_0\rangle \\ + \int_0^t \int_\Omega \mathbb{C}\overline{\mathbf{E}_h^{\mathrm{e}}}(s) : \mathbf{E}\dot{w}(s)\,dx\,ds - \int_0^t \langle \dot{\mathcal{L}}(s), \overline{u_h}(s)\rangle\,ds \\ - \int_0^t \langle \overline{\mathcal{L}_h}(s), \dot{w}(s)\rangle\,ds + \bar{e}_h.\end{aligned} \tag{7.14}$$

Since $\mathbf{E}^{\mathrm{p}} \in BV(0,T; BV(\Omega; \mathrm{M}_D^{N\times N}))$, we have that $\mathbf{E}^{\mathrm{p}}(t)$ is continuous in time with respect to the strong topology in $BV(\Omega; \mathrm{M}_D^{N\times N})$ up to a countable set in $[0,T]$. Let $s \in [0,T]$ be a continuity point of $\mathbf{E}^{\mathrm{p}}$, and let $s_n \to s$. Then

$$\mathbf{E}^{\mathrm{e}}(s_n) \rightharpoonup \mathbf{E}^{\mathrm{e}}(s) \qquad \text{weakly in } L^2(\Omega; \mathrm{M}_{\mathrm{sym}}^{N\times N}) \tag{7.15}$$

and

$$u(s_n) \rightharpoonup u(s) \qquad \text{weakly in } W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N). \tag{7.16}$$

In fact up to a subsequence we have that

$$\mathbf{E}^{\mathrm{e}}(s_n) \rightharpoonup \widetilde{\mathbf{E}^{\mathrm{e}}} \qquad \text{weakly in } L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}})$$

and

$$u(s_n) \rightharpoonup \tilde{u} \qquad \text{weakly in } W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N),$$

with $(\tilde{u},\widetilde{\mathbf{E}^{\mathrm{e}}},\mathbf{E}^{\mathrm{p}}(s)) \in \mathcal{A}(w(s))$. Given $(v,\mathbf{e},\mathbf{E}^{\mathrm{p}}(s)) \in \mathcal{A}(w(s))$, by the global stability condition (7.6), comparing $(u(s_n),\mathbf{E}^{\mathrm{e}}(s_n),\mathbf{E}^{\mathrm{p}}(s_n))$ with $(v-w(s)+w(s_n),\mathbf{e}-\mathbf{E}w(s)+\mathbf{E}w(s_n),\mathbf{E}^{\mathrm{p}}(s))$, and taking into account the continuity of $\mathcal{H}$ with respect to the BV-norm, we obtain that $(\tilde{u},\widetilde{\mathbf{E}^{\mathrm{e}}})$ is a minimizer of the convex functional $(v,\mathbf{e}) \mapsto \mathcal{Q}_1(\mathbf{e}) - \langle \mathcal{L}(s), v\rangle$ on the convex set $\mathcal{K} := \{(v,\mathbf{e}) : (v,\mathbf{e},\mathbf{E}^{\mathrm{p}}(s)) \in \mathcal{A}(w(s))\}$. By the uniqueness of the minimizer, we have that $\tilde{u} = u(s)$ and $\widetilde{\mathbf{E}^{\mathrm{e}}} = \mathbf{E}^{\mathrm{e}}(s)$, so that (7.15) and (7.16) follow.

By (7.15) and (7.16) we have that for a.e. every $s \in [0,t]$

$$\overline{\mathbf{E}^{\mathrm{e}}_h}(s) \rightharpoonup \mathbf{E}^{\mathrm{e}}(s) \qquad \text{weakly in } L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}}) \tag{7.17}$$

and

$$\overline{u_h}(s) \rightharpoonup u(s) \qquad \text{weakly in } W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N). \tag{7.18}$$

By taking into account that for every $s \in [0,T]$

$$\overline{\mathcal{L}_h}(s) \to \mathcal{L}(s) \qquad \text{strongly in } \left(W^{1,\frac{N}{N-1}}\left(\Omega;\mathbb{R}^N\right)\right)^*,$$

in view of (7.17) and (7.18), passing to the limit in (7.14) we get by dominated convergence (take into account (7.5)) that (7.12) follows. □

We are in a position to prove Theorem 5.2. Indeed, the evolution $t \mapsto (u(t),\mathbf{E}^{\mathrm{e}}(t),\mathbf{E}^{\mathrm{p}}(t))$ given by Lemma 7.1 is a quasi-static evolution for the Gurtin–Anand model because it satisfies the global stability condition in view of (7.6), and it satisfies the energy balance because of (7.10) and (7.12).

The convergence of the discrete-in-time evolution to the continuous one can be improved in the following way.

PROPOSITION 7.4. [illegible] $t \mapsto (u(t),\mathbf{E}^{\mathrm{e}}(t),\mathbf{E}^{\mathrm{p}}(t))$ [illegible] 7.1 [illegible] $t \in [0,T]$ [illegible] $j \to +\infty$

$$\mathcal{E}_{k_j}(t) \to \mathcal{E}(t) \tag{7.19}$$

[illegible]

$$\mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}}_{k_j};0,t) \to \mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}};0,t). \tag{7.20}$$

[illegible]

$$\mathbf{E}^{\mathrm{e}}_{k_j}(t) \to \mathbf{E}^{\mathrm{e}}(t) \quad \text{[illegible] } L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}}), \tag{7.21}$$

$$\mathrm{curl}\mathbf{E}^{\mathrm{p}}_{k_j}(t) \to \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t) \quad \text{[illegible] } L^2(\Omega;\mathrm{M}^{N\times N}), \tag{7.22}$$

$t \in [0,T]$

$$\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}_{k_j}(t)) \to \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(t)), \qquad \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}_{k_j}(t)) \to \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)) \tag{7.23}$$

$$\langle \mathcal{L}_{k_j}(t), u_{k_j}(t)\rangle \to \langle \mathcal{L}(t), u(t)\rangle. \tag{7.24}$$

. Notice that by lower semicontinuity we have for every $t \in [0,T]$

$$\mathcal{E}(t) \le \liminf_{j\to+\infty} \mathcal{E}_{k_j}(t). \tag{7.25}$$

Moreover, by the lower semicontinuity of $\mathcal{H}$ with respect to the weak* convergence, and by the very definition of $\mathcal{D}_{\mathcal{H}}$, we deduce that for every $t \in [0,T]$

$$\mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}};0,t) \le \liminf_{j\to+\infty} \mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}}_{k_j};0,t). \tag{7.26}$$

By (6.3) and (7.12) we get that

$$\begin{aligned}
\mathcal{E}(t) + \mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}};0,t) &\le \liminf_{j\to+\infty} (\mathcal{E}_{k_j}(t) + \mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}}_{k_j};0,t)) \\
&\le \limsup_{j\to+\infty} \left(\mathcal{E}_{k_j}(t) + \mathcal{D}_{\mathcal{H}}\left(\mathbf{E}^{\mathrm{p}}_{k_j};0,t\right)\right) \\
&\le \limsup_{j\to+\infty} \left[\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}_0) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}_0) - \langle \mathcal{L}(0), u_0\rangle + \int_0^{t^i_{k_j}} \int_\Omega \mathbb{C}\mathbf{E}^{\mathrm{e}}_{k_j}(\tau) : \mathbf{E}\dot{w}(\tau)\, dx\, d\tau \right. \\
&\qquad \left. - \int_0^{t^i_{k_j}} \langle \dot{\mathcal{L}}(\tau), u_{k_j}(\tau)\rangle\, d\tau - \int_0^{t^i_{k_j}} \langle \mathcal{L}_{k_j}(\tau), \dot{w}(\tau)\rangle\, d\tau + e_{k_j}\right] \\
&= \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}_0) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}_0) - \langle \mathcal{L}(0), u_0\rangle + \int_0^t \int_\Omega \mathbb{C}\mathbf{E}^{\mathrm{e}}(\tau) : \mathbf{E}\dot{w}(\tau)\, dx\, d\tau \\
&\qquad - \int_0^t \langle \dot{\mathcal{L}}(\tau), u(\tau)\rangle\, d\tau - \int_0^t \langle \mathcal{L}(\tau), \dot{w}(\tau)\rangle\, d\tau \le \mathcal{E}(t) + \mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}};0,t).
\end{aligned}$$

We conclude that for every $t \in [0,T]$

$$\lim_{j\to+\infty} (\mathcal{E}_{k_j}(t) + \mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}}_{k_j};0,t)) = \mathcal{E}(t) + \mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}};0,t).$$

From (7.25) and (7.26) we deduce that (7.19) and (7.20) hold. Since by lower semicontinuity

$$\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(t)) \le \liminf_{j\to+\infty} \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}_{k_j}(t)) \qquad \text{and} \qquad \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)) \le \liminf_{j\to+\infty} \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}_{k_j}(t)),$$

while

$$\langle \mathcal{L}_{k_j}(t), u_{k_j}(t)\rangle \to \langle \mathcal{L}(t), u(t)\rangle,$$

from (7.19) we deduce that (7.23) and (7.24) hold. In particular (7.21) and (7.22) follow, and the proof is concluded. □

7.5 (convergence for the elastic and the "energetic" plastic strains). Since the maps $t \mapsto \mathbf{E}^{\mathrm{e}}(t)$ and $t \mapsto \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)$ turn out to be uniquely determined by the initial conditions (see Proposition 8.8), we infer that

$$\begin{aligned}
\mathbf{E}^{\mathrm{e}}_k(t) &\to \mathbf{E}^{\mathrm{e}}(t) \qquad \text{strongly in } L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}}), \\
\mathrm{curl}\mathbf{E}^{\mathrm{p}}_k(t) &\to \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t) \qquad \text{strongly in } L^2(\Omega;\mathrm{M}^{N\times N})
\end{aligned}$$

without passing to a subsequence $(k_j)_{j\in\mathbb{N}}$.

8. Balance equations and the flow rule. This section is devoted to the proof of Theorem 5.3; that is, we prove that a quasi-static evolution $t \mapsto (u(t), \mathbf{E}^{\mathrm{e}}(t), \mathbf{E}^{\mathrm{p}}(t))$ for the Gurtin–Anand model satisfies the prescribed regularity and uniqueness properties, the balance equations, and the flow rule.

We need the following lemma.

LEMMA 8.1. [illegible] $t \in [0,T]$ [illegible] $\mathbf{T}^{\mathrm{p}}(t) \in L^\infty(\Omega; \mathrm{M}^{N\times N}_D)$, $\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t) \in L^\infty(\Omega; \mathrm{M}^{N\times N\times N}_D)$, [illegible] $\mathbb{S}^{\mathrm{p}}(t) \in (\mathcal{M}_b(\Omega; \mathrm{M}^{N\times N\times N}_D))^*$ [illegible] $(\mathbf{A}, \mathbb{B}, \mathbb{L}) \in L^1(\Omega; \mathrm{M}^{N\times N}_D) \times L^1(\Omega; \mathrm{M}^{N\times N\times N}_D) \times \mathcal{M}_b(\Omega; \mathrm{M}^{N\times N\times N}_D)$

$$\left| \int_\Omega \mathbf{T}^{\mathrm{p}}(t) : \mathbf{A}\, dx + \int_\Omega \mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t) : \mathbb{B}\, dx + \langle \mathbb{S}^{\mathrm{p}}(t), \mathbb{L} \rangle \right| \le S_Y \int_\Omega \sqrt{|\mathbf{A}|^2 + l^2 |\mathbb{B}|^2}\, dx + l S_Y |\mathbb{L}|(\Omega) \tag{8.1}$$

[illegible] $(v, \mathbf{e}, \mathbf{p}) \in \mathcal{A}(0)$

$$\int_\Omega \mathbf{T}(t) : \mathbf{e}\, dx + \mu L^2 \int_\Omega \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t) : \mathrm{curl}\mathbf{p}\, dx - \langle \mathcal{L}(t), v \rangle = -\int_\Omega \mathbf{T}^{\mathrm{p}}(t) : \mathbf{p}\, dx - \int_\Omega \mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t) : \nabla \mathbf{p}\, dx - \langle \mathbb{S}^{\mathrm{p}}(t), D^s \mathbf{p} \rangle. \tag{8.2}$$

[illegible] $\mathbb{K}^{\mathrm{p}}(t) := \mathbb{K}^{\mathrm{p}}_{\mathrm{en}}(t) + \mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t)$, [illegible] $\mathbb{K}^{\mathrm{p}}_{\mathrm{en}}(t)$ [illegible] (3.4), (3.5) [illegible] $\mathbf{E}^{\mathrm{p}}(t)$, [illegible] $(v, \mathbf{e}, \mathbf{p}) \in \mathcal{A}(0)$ [illegible]

$$\int_\Omega \mathbf{T}(t) : \mathbf{e}\, dx + \int_\Omega \mathbf{T}^{\mathrm{p}}(t) : \mathbf{p}\, dx + \int_\Omega \mathbb{K}^{\mathrm{p}}(t) : \nabla \mathbf{p}\, dx + \langle \mathbb{S}^{\mathrm{p}}(t), D^s \mathbf{p} \rangle = \langle \mathcal{L}(t), v \rangle. \tag{8.3}$$

[illegible]. Let us fix $t \in [0,T]$. From the global stability condition (5.2), for every $(v, \mathbf{e}, \mathbf{p}) \in \mathcal{A}(0)$ and $\varepsilon \in \mathbb{R}$ we get

$$\begin{aligned} &\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(t)) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)) - \langle \mathcal{L}(t), u(t) \rangle \\ &\qquad \le \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(t) + \varepsilon \mathbf{e}) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t) + \varepsilon \mathrm{curl}\mathbf{p}) - \langle \mathcal{L}(t), u(t) + \varepsilon v \rangle + \mathcal{H}(\varepsilon \mathbf{p}) \end{aligned}$$

so that

$$\begin{aligned} &\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(t) + \varepsilon \mathbf{e}) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t) + \varepsilon \mathrm{curl}\mathbf{p}) - \varepsilon \langle \mathcal{L}(t), v \rangle + \mathcal{H}(\varepsilon \mathbf{p}) \\ &\qquad \ge \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(t)) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)). \end{aligned}$$

By taking the left and right derivative for $\varepsilon = 0$ we get

$$\int_\Omega \mathbb{C}\mathbf{E}^{\mathrm{e}}(t) : \mathbf{e}\, dx + \mu L^2 \int_\Omega \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t) : \mathrm{curl}\mathbf{p}\, dx - \langle \mathcal{L}(t), v \rangle + \mathcal{H}(\mathbf{p}) \ge 0$$

and

$$\int_\Omega \mathbb{C}\mathbf{E}^{\mathrm{e}}(t) : \mathbf{e}\, dx + \mu L^2 \int_\Omega \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t) : \mathrm{curl}\mathbf{p}\, dx - \langle \mathcal{L}(t), v \rangle - \mathcal{H}(-\mathbf{p}) \le 0$$

so that, since $\mathbf{T}(t) := \mathbb{C}\mathbf{E}^{\mathrm{e}}(t)$,

$$\left| \int_\Omega \mathbf{T}(t) : \mathbf{e}\, dx + \mu L^2 \int_\Omega \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t) : \mathrm{curl}\mathbf{p}\, dx - \langle \mathcal{L}(t), v \rangle \right| \le \mathcal{H}(\mathbf{p}).$$

The previous inequality shows that the linear functional on $\mathcal{A}(0)$

$$(v,\mathbf{e},\mathbf{p})\mapsto\int_\Omega \mathbb{C}\mathbf{E}^{\mathrm{e}}(t):\mathbf{e}\,dx+\mu L^2\int_\Omega \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t):\mathrm{curl}\mathbf{p}\,dx-\langle\mathcal{L}(t),v\rangle$$

depends indeed only on $\mathbf{p}$.

Let $X\subseteq L^1(\Omega;\mathrm{M}_D^{N\times N})\times L^1(\Omega;\mathrm{M}_D^{N\times N\times N})\times\mathcal{M}_b(\Omega;\mathrm{M}_D^{N\times N\times N})$ be the linear subspace generated by $\{(\mathbf{p},\nabla\mathbf{p},D^s\mathbf{p})\,:\,(v,\mathbf{e},\mathbf{p})\in\mathcal{A}(0)$ for some $v\in W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N),\mathbf{e}\in L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}})\}$. By applying the Hahn–Banach theorem we deduce that the linear functional

$$\varphi(\mathbf{p},\nabla\mathbf{p},D^s\mathbf{p}):=\int_\Omega \mathbf{T}(t):\mathbf{e}\,dx+\mu L^2\int_\Omega \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t):\mathrm{curl}\mathbf{p}\,dx-\langle\mathcal{L}(t),v\rangle \tag{8.4}$$

on the linear space X can be extended in a continuous way to the entire space $L^1(\Omega;\mathrm{M}_D^{N\times N})\times L^1(\Omega;\mathrm{M}_D^{N\times N\times N})\times\mathcal{M}_b(\Omega;\mathrm{M}_D^{N\times N\times N})$ in such a way that

$$|\varphi(\mathbf{A},\mathbb{B},\mathbb{L})|\le S_Y\int_\Omega\sqrt{|\mathbf{A}|^2+l^2|\mathbb{B}|^2}\,dx+lS_Y|\mathbb{L}|(\Omega) \tag{8.5}$$

for every $(\mathbf{A},\mathbb{B},\mathbb{L})\in L^1(\Omega;\mathrm{M}_D^{N\times N})\times L^1(\Omega;\mathrm{M}_D^{N\times N\times N})\times\mathcal{M}_b(\Omega;\mathrm{M}_D^{N\times N\times N})$. By representing φ, in view of (8.5) and (8.4), we obtain that there exist $\mathbf{T}^{\mathrm{p}}(t)\in L^\infty(\Omega;\mathrm{M}_D^{N\times N})$, $\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t)\in L^\infty(\Omega;\mathrm{M}_D^{N\times N\times N})$, and $\mathbb{S}^{\mathrm{p}}(t)\in(\mathcal{M}_b(\Omega;\mathrm{M}_D^{N\times N\times N}))^*$ such that (8.1) and (8.2) hold. Finally, (8.3) follows by (8.2) in view of the very definition of $\mathbb{K}^{\mathrm{p}}_{\mathrm{en}}(t)$. ∎

The following proposition concerns the balance equation for the Cauchy stress.

Proposition 8.2 (balance equations for the Cauchy stress). $t\in[0,T]$

$$\begin{cases}-\mathrm{div}\mathbf{T}(t)=f(t) & \Omega,\\ \mathbf{T}(t)\nu=g(t) & \partial_N\Omega.\end{cases} \tag{8.6}$$

. Let $v\in C^\infty(\overline{\Omega},\mathbb{R}^N)$ such that $v=0$ on $\partial_D\Omega$. By choosing $(v,\mathbf{E}v,0)\in\mathcal{A}(0)$ in (8.3) we deduce that

$$\int_\Omega \mathbf{T}(t):\mathbf{E}v\,dx=\langle\mathcal{L}(t),v\rangle. \tag{8.7}$$

Then clearly $-\mathrm{div}\mathbf{T}(t)=f(t)$ in the sense of distributions in Ω. Since $\mathbf{T}(t)\in L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}})$ and its divergence belongs in particular to $L^2(\Omega;\mathbb{R}^N)$, we have that the normal trace of $\mathbf{T}(t)$ on $\partial\Omega$ is well defined as an element of $H^{-1/2}(\partial\Omega;\mathbb{R}^N)$. By integrating by parts in (8.7) we get immediately the second relation of (8.6). ∎

Concerning the stresses conjugated to the plastic variables, the following result holds.

Proposition 8.3 (stresses conjugated to the plastic variables). $t\in[0,T]$ $\mathbf{T}^{\mathrm{p}}(t)$ $\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t)$ $\mathbb{K}^{\mathrm{p}}(t)$ $\mathbb{S}^{\mathrm{p}}(t)$ 8.1

$$\begin{cases}\mathbf{T}^{\mathrm{p}}(t)=\mathbf{T}_D(t)+\mathrm{div}\mathbb{K}^{\mathrm{p}}(t) & \Omega,\\ \mathbb{K}^{\mathrm{p}}(t)\nu=0 & \partial\Omega,\end{cases} \tag{8.8}$$

$\mathbf{T}_D(t):=(\mathbf{T}(t))_D$ $\mathbf{T}^{\mathrm{p}}(t)$ $\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t)$

$$\sqrt{|\mathbf{T}^{\mathrm{p}}(t,x)|^2+l^{-2}|\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t,x)|^2}\le S_Y \qquad x\in\Omega, \tag{8.9}$$

$\mathbb{S}^{\mathrm{p}}(t)$ fi

$$\|\mathbb{S}^{\mathrm{p}}(t)\|_{(\mathcal{M}_b(\mathrm{M}_D^{N\times N\times N}))^*} \le lS_Y. \tag{8.10}$$

. The stress constraints (8.9) and (8.10) follow by choosing $(\mathbf{A}, \mathbb{B}, 0)$ and $(0, 0, \mathbb{L})$, respectively, in (8.1).

Let us come to (8.8). Let $\mathbf{p} \in C^\infty(\overline{\Omega}, \mathrm{M}_D^{N\times N})$, so that in particular $(0, -\mathbf{p}, \mathbf{p}) \in \mathcal{A}(0)$. Then (8.3) yields

$$-\int_\Omega \mathbf{T}(t) : \mathbf{p}\, dx + \int_\Omega \mathbf{T}^{\mathrm{p}}(t) : \mathbf{p}\, dx + \int_\Omega \mathbb{K}^{\mathrm{p}}(t) : \nabla \mathbf{p}\, dx = 0.$$

Since $\mathbf{p}$ takes values in the space of deviatoric matrices, we can replace $\mathbf{T}(t)$ by $\mathbf{T}_D(t)$ so that we obtain

$$\int_\Omega (\mathbf{T}^{\mathrm{p}}(t) - \mathbf{T}_D(t)) : \mathbf{p}\, dx + \int_\Omega \mathbb{K}^{\mathrm{p}}(t) : \nabla \mathbf{p}\, dx = 0. \tag{8.11}$$

We conclude that the first relation of (8.8) holds. As a consequence, in view of (8.9) and the definition of $\mathbb{K}^{\mathrm{p}}_{\mathrm{en}}(t)$, we have $\mathbb{K}^{\mathrm{p}}(t) \in L^2(\Omega; \mathrm{M}_D^{N\times N\times N})$ with divergence in $L^2(\Omega; \mathrm{M}_D^{N\times N})$, so that its normal trace on $\partial\Omega$ is well defined as an element of $H^{-1/2}(\partial\Omega; \mathbb{R}^{N\times N})$. By integrating by parts in (8.11) we obtain also the second relation of (8.8), and the proof is concluded. □

8.4. Note that relation (8.3) represents the balance of internal and external power expenditures on the whole body Ω (see section 3). Due to our variational approach which requires $\mathbf{E}^{\mathrm{p}}(t) \in BV(\Omega; \mathrm{M}_D^{N\times N})$ so that $D\mathbf{E}^{\mathrm{p}}(t)$ has also a singular part, a stress $\mathbb{S}^{\mathrm{p}}(t)$ associated to $D^s\mathbf{E}^{\mathrm{p}}(t)$ appears in the balance. In order to get a balance equation for a subbody $\mathcal{B} \subset\subset \Omega$ with a sufficiently smooth boundary, we can reason as follows. Let us assume to be in the physical case $N = 3$. As a consequence, admissible displacements v turn out to belong to $L^3(\Omega; \mathbb{R}^3)$. Let us assume for simplicity that $(v, \mathbf{e}, \mathbf{p}) \in \mathcal{A}(0)$ is such that $\mathbf{p}$ belongs also to $L^2(\Omega; M_D^{3\times 3})$, and let $\varphi \in C_c^\infty(\Omega)$. By (8.3) we can write

(8.12)

$$\begin{aligned}
&\int_\Omega \mathbf{T}(t) : (\varphi\mathbf{e})\, dx + \int_\Omega \mathbf{T}^{\mathrm{p}}(t) : (\varphi\mathbf{p})\, dx + \int_\Omega \mathbb{K}^{\mathrm{p}}(t) : (\varphi\nabla\mathbf{p})\, dx + \langle \mathbb{S}^{\mathrm{p}}(t), \varphi D^s\mathbf{p}\rangle \\
&\quad = \int_\Omega \mathbf{T}(t) : (\varphi\mathbf{e} + v \odot \nabla\varphi)\, dx + \int_\Omega \mathbf{T}^{\mathrm{p}}(t) : (\varphi\mathbf{p})\, dx + \int_\Omega \mathbb{K}^{\mathrm{p}}(t) : \nabla(\varphi\mathbf{p})\, dx \\
&\qquad + \langle \mathbb{S}^{\mathrm{p}}(t), D^s(\varphi\mathbf{p})\rangle - \int_\Omega \mathbf{T}(t) : (v \odot \nabla\varphi)\, dx - \int_\Omega \mathbb{K}^{\mathrm{p}}(t) : (\mathbf{p} \otimes \nabla\varphi)\, dx \\
&\qquad\quad = \langle \mathcal{L}(t), \varphi v\rangle - \int_\Omega \mathbf{T}(t) : (v \odot \nabla\varphi)\, dx - \int_\Omega \mathbb{K}^{\mathrm{p}}(t) : (\mathbf{p} \otimes \nabla\varphi)\, dx,
\end{aligned}$$

where the last equality follows since $(\varphi v, \varphi\mathbf{e} + v \odot \nabla\varphi, \varphi\mathbf{p}) \in \mathcal{A}(0)$ (we use $\mathbf{p} \in L^2(\Omega; M_D^{3\times 3})$ to ensure that $\mathrm{curl}(\varphi\mathbf{p}) \in L^2(\Omega; M^{3\times 3})$). Here $(v \odot \nabla\varphi)_{i,j} = (v_i\partial_j\varphi + v_j\partial_i\varphi)/2$. As a consequence we have that the distribution

$$\varphi \mapsto -\int_\Omega \mathbf{T}(t) : (v \odot \nabla\varphi)\, dx - \int_\Omega \mathbb{K}^{\mathrm{p}}(t) : (\mathbf{p} \otimes \nabla\varphi)\, dx$$

turns out to be a measure $\mu \in \mathcal{M}_b(\Omega)$. Moreover, by considering the measure $\eta \in \mathcal{M}_b(\Omega; \mathbb{R}^3)$ given by

$$\int_\Omega \psi\, d\eta := \int_\Omega \mathbf{T}(t) : (v \odot \psi)\, dx + \int_\Omega \mathbb{K}^{\mathrm{p}}(t) : (\mathbf{p} \otimes \psi)\, dx, \qquad \psi \in C_c^\infty\left(\Omega; \mathbb{R}^3\right),$$

we get immediately that $\mathrm{div}\eta = \mu$. According to [22], for every subset $\mathcal{B} \subset\subset \Omega$ with a sufficiently smooth boundary we have that η admits normal trace $\eta \cdot \nu$ on $\partial\mathcal{B}$ defined as an element of the dual of $C^1(\partial\mathcal{B})$, in such a way that the following Gauss–Green formula holds:

$$\int_{\mathcal{B}} d(\mathrm{div}\eta) = \langle \eta \cdot \nu, 1_{\partial\mathcal{B}} \rangle.$$

Let us denote formally $\eta\cdot\nu$ by $[\mathbf{T}(t)\nu\cdot v + \mathbb{K}^{\mathrm{p}}(t)\nu : \mathbf{p}]$, and let $[\mathbb{K}^{\mathrm{p}}(t) : \nabla\mathbf{p} + \mathbb{S}^{\mathrm{p}}(t) : D^s\mathbf{p}]$ be the measure such that

$$\int_\Omega \varphi\, d[\mathbb{K}^{\mathrm{p}}(t) : \nabla\mathbf{p} + \mathbb{S}^{\mathrm{p}}(t) : D^s\mathbf{p}] = \int_\Omega \mathbb{K}^{\mathrm{p}}(t) : (\varphi\nabla\mathbf{p})\, dx + \langle \mathbb{S}^{\mathrm{p}}(t), \varphi D^s\mathbf{p}\rangle.$$

By (8.12) we can write by choosing $\varphi = 1_{\mathcal{B}}$

$$\begin{aligned}(8.13)\qquad \int_{\mathcal{B}} \mathbf{T}(t) : \mathbf{e}\, dx + \int_{\mathcal{B}} \mathbf{T}^{\mathrm{p}}(t) : \mathbf{p}\, dx + [\mathbb{K}^{\mathrm{p}}(t) : \nabla\mathbf{p} + \mathbb{S}^{\mathrm{p}}(t) : D^s\mathbf{p}](\mathcal{B}) \\ = \int_{\mathcal{B}} f(t)\cdot v\, dx + \langle [\mathbf{T}(t)\nu\cdot v + \mathbb{K}^{\mathrm{p}}(t)\nu : \mathbf{p}], 1_{\partial\mathcal{B}}\rangle,\end{aligned}$$

which is a weak form for the balance of power expenditures for the subbody $\mathcal{B}$ relative to the virtual velocity $(v, \mathbf{e}, \mathbf{p}) \in \mathcal{A}(0)$.

The AC regularity in time for the quasi-static evolution is proved in the following proposition. The proof relies heavily on [8, Theorem 5.2]. We exploit the calculations in our context since we aim to understand the precise dependence on the material length scales l and L of the norms involved in the statement (in view of the convergence result of section 9, where we let $L, l \to 0$).

PROPOSITION 8.5. [illegible] $t \mapsto u(t), t \mapsto \mathbf{E}^{\mathrm{e}}(t), t \mapsto \mathbf{E}^{\mathrm{p}}(t)$ [illegible] $t \mapsto \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)$ [illegible] $[0,T]$ [illegible] $W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N)$, $L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}})$, $BV(\Omega;\mathrm{M}^{N\times N}_D)$, $L^2(\Omega;\mathrm{M}^{N\times N})$ [illegible] $t \in [0,T]$ [illegible]

$$(8.14)\qquad \|\dot u(t)\|_{W^{1,\frac{N}{N-1}}} \le C_1(w,\rho,\alpha,\mathbb{C},l)[\|\dot\rho(t)\|_{L^2} + \|\dot\rho_D(t)\|_{L^\infty} + \|\mathbf{E}\dot w(t)\|_{L^2}],$$

$$(8.15)\qquad \|\dot{\mathbf{E}}^{\mathrm{e}}(t)\|_{L^2} \le C_2(w,\rho,\alpha,\mathbb{C})[\|\dot\rho(t)\|_{L^2} + \|\dot\rho_D(t)\|_{L^\infty} + \|\mathbf{E}\dot w(t)\|_{L^2}],$$

$$(8.16)\qquad \|\dot{\mathbf{E}}^{\mathrm{p}}(t)\|_{BV} \le C_3(w,\rho,\alpha,\mathbb{C},l)[\|\dot\rho(t)\|_{L^2} + \|\dot\rho_D(t)\|_{L^\infty} + \|\mathbf{E}\dot w(t)\|_{L^2}],$$

$$(8.17)\qquad \|\mathrm{curl}\dot{\mathbf{E}}^{\mathrm{p}}(t)\|_{L^2} \le C_4(w,\rho,\alpha,\mathbb{C},L)[\|\dot\rho(t)\|_{L^2} + \|\dot\rho_D(t)\|_{L^\infty} + \|\mathbf{E}\dot w(t)\|_{L^2}],$$

[illegible] ρ [illegible] α [illegible] (4.13) (4.14) [illegible] $t \mapsto u(t)$ [illegible] $t \mapsto \mathbf{E}^{\mathrm{p}}(t)$ [illegible] $[0,T]$ [illegible] $BD(\Omega)$ [illegible] $L^1(\Omega;\mathrm{M}^{N\times N}_D)$ [illegible] 9 [illegible] [illegible] $t \in [0,T]$

$$(8.18)\qquad \|\dot u(t)\|_{BD} \le C_5(w,\rho,\alpha,\mathbb{C})[\|\dot\rho(t)\|_{L^2} + \|\dot\rho_D(t)\|_{L^\infty} + \|\mathbf{E}\dot w(t)\|_{L^2}],$$

$$(8.19)\qquad \|\dot{\mathbf{E}}^{\mathrm{p}}(t)\|_{L^1} \le C_6(w,\rho,\alpha,\mathbb{C})[\|\dot\rho(t)\|_{L^2} + \|\dot\rho_D(t)\|_{L^\infty} + \|\mathbf{E}\dot w(t)\|_{L^2}].$$

. Let $t_1, t_2 \in [0,T]$, with $t_1 < t_2$. Since by the very definition of $\mathcal{D}_{\mathcal{H}}$ we have $\mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}}; t_1, t_2) \geq \mathcal{H}(\mathbf{E}^{\mathrm{p}}(t_2) - \mathbf{E}^{\mathrm{p}}(t_1))$, by the energy balance (5.3) we may write

$$\begin{aligned}(8.20)\quad &\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(t_2)) - \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(t_1)) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t_2)) - \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t_1))\\ &\qquad + \mathcal{H}(\mathbf{E}^{\mathrm{p}}(t_2) - \mathbf{E}^{\mathrm{p}}(t_1)) - \langle \mathcal{L}(t_2), u(t_2)\rangle + \langle \mathcal{L}(t_1), u(t_1)\rangle\\ &\qquad \leq \int_{t_1}^{t_2}\int_\Omega \mathbf{T}(\tau) : \mathbf{E}\dot{w}(\tau)\, dx\, d\tau - \int_{t_1}^{t_2} \langle \dot{\mathcal{L}}(\tau), u(\tau)\rangle\, d\tau - \int_{t_1}^{t_2} \langle \mathcal{L}(\tau), \dot{w}(\tau)\rangle\, d\tau.\end{aligned}$$

Let us consider $(v, \mathbf{e}, \mathbf{p}) \in \mathcal{A}(0)$ such that

$$v := u(t_2) - u(t_1) - (w(t_2) - w(t_1)), \quad \mathbf{e} := \mathbf{E}^{\mathrm{e}}(t_2) - \mathbf{E}^{\mathrm{e}}(t_1) - (\mathbf{E}w(t_2) - \mathbf{E}w(t_1)),$$

and $\mathbf{p} := \mathbf{E}^{\mathrm{p}}(t_2) - \mathbf{E}^{\mathrm{p}}(t_1)$. By combining (8.1) and (8.2), we deduce that

$$\begin{aligned}(8.21)\quad &-\int_\Omega \mathbf{T}(t_1) : (\mathbf{E}^{\mathrm{e}}(t_2) - \mathbf{E}^{\mathrm{e}}(t_1) - (\mathbf{E}w(t_2) - \mathbf{E}w(t_1)))\, dx\\ &\qquad - \mu L^2 \int_\Omega \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t_1) : (\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t_2) - \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t_1))\, dx\\ &\qquad + \langle \mathcal{L}(t_1), u(t_2) - u(t_1) - (w(t_2) - w(t_1))\rangle \leq \mathcal{H}(\mathbf{E}^{\mathrm{p}}(t_2) - \mathbf{E}^{\mathrm{p}}(t_1)).\end{aligned}$$

By inserting (8.21) into (8.20) and taking into account (4.15), we obtain

$$\begin{aligned}&\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(t_2) - \mathbf{E}^{\mathrm{e}}(t_1)) + \mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t_2) - \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t_1))\\ &\quad \leq \int_{t_1}^{t_2}\int_\Omega \mathbf{T}(\tau) : \mathbf{E}\dot{w}(\tau)\, dx\, d\tau - \int_{t_1}^{t_2} \langle \dot{\mathcal{L}}(\tau), u(\tau)\rangle\, d\tau - \int_{t_1}^{t_2} \langle \mathcal{L}(\tau), \dot{w}(\tau)\rangle\, d\tau\\ &\qquad + \langle \mathcal{L}(t_2) - \mathcal{L}(t_1), u(t_2)\rangle - \int_\Omega \mathbf{T}(t_1) : (\mathbf{E}w(t_2) - \mathbf{E}w(t_1))\, dx\\ &\qquad\qquad + \langle \mathcal{L}(t_1), w(t_2) - w(t_1)\rangle\\ &\quad = \int_{t_1}^{t_2}\int_\Omega (\mathbf{T}(\tau) - \mathbf{T}(t_1)) : \mathbf{E}\dot{w}(\tau)\, dx\, d\tau - \int_{t_1}^{t_2} \langle \dot{\mathcal{L}}(\tau), u(\tau) - u(t_2)\rangle\, d\tau\\ &\qquad\qquad - \int_{t_1}^{t_2} \langle \mathcal{L}(\tau) - \mathcal{L}(t_1), \dot{w}(\tau)\rangle\, d\tau\\ &= \int_{t_1}^{t_2}\int_\Omega (\mathbf{T}(\tau) - \mathbf{T}(t_1)) : \mathbf{E}\dot{w}(\tau)\, dx\, d\tau - \int_{t_1}^{t_2}\int_\Omega \dot{\rho}(\tau) : (\mathbf{E}^{\mathrm{e}}(\tau) - \mathbf{E}^{\mathrm{e}}(t_2))\, dx\, d\tau\\ &- \int_{t_1}^{t_2}\int_\Omega \dot{\rho}_D(\tau) : (\mathbf{E}^{\mathrm{p}}(\tau) - \mathbf{E}^{\mathrm{p}}(t_2))\, dx\, d\tau - \int_{t_1}^{t_2}\int_\Omega (\rho(\tau) - \rho(t_1)) : \mathbf{E}\dot{w}(\tau)\, dx\, d\tau.\end{aligned}$$

By the coercivity estimate (4.8) for the elasticity tensor $\mathbb{C}$ we deduce that

$$\begin{aligned}(8.22)\quad &\alpha_{\mathbb{C}}\|\mathbf{E}^{\mathrm{e}}(t_2) - \mathbf{E}^{\mathrm{e}}(t_1)\|_{L^2}^2 + \frac{\mu L^2}{2}\|\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t_2) - \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t_1)\|_{L^2}^2\\ &\leq \beta_{\mathbb{C}}\int_{t_1}^{t_2} \|\mathbf{E}^{\mathrm{e}}(\tau) - \mathbf{E}^{\mathrm{e}}(t_1)\|_{L^2}\|\mathbf{E}\dot{w}(\tau)\|_{L^2}\, d\tau + \int_{t_1}^{t_2} \|\dot{\rho}\|_{L^2}\|\mathbf{E}^{\mathrm{e}}(\tau) - \mathbf{E}^{\mathrm{e}}(t_2)\|_{L^2}\, d\tau\\ &+ \int_{t_1}^{t_2} \|\dot{\rho}_D(\tau)\|_{L^\infty}\|\mathbf{E}^{\mathrm{p}}(\tau) - \mathbf{E}^{\mathrm{p}}(t_2)\|_{L^1}\, d\tau + \int_{t_1}^{t_2} \|\rho(\tau) - \rho(t_1)\|_{L^2}\|\mathbf{E}\dot{w}(\tau)\|_{L^2}\, d\tau.\end{aligned}$$

By (4.16) we have for $t_1 \le s \le t_2$

$$\begin{aligned}(8.23)\quad &\frac{\alpha}{2}\|\mathbf{E}^{\mathrm{p}}(t_2)-\mathbf{E}^{\mathrm{p}}(s)\|_{L^1}+\alpha_l\|D\mathbf{E}^{\mathrm{p}}(t_2)-D\mathbf{E}^{\mathrm{p}}(s)\|_{\mathcal{M}_b}\\ &\qquad\le \mathcal{H}\left(\mathbf{E}^{\mathrm{p}}(t_2)-\mathbf{E}^{\mathrm{p}}(s)\right)-\int_\Omega \rho_D(t_2):\left(\mathbf{E}^{\mathrm{p}}(t_2)-\mathbf{E}^{\mathrm{p}}(s)\right)dx,\end{aligned}$$

where $\alpha_l := \min\{l\frac{\alpha}{2}, lS_Y\}$. By combining (8.23) and (8.20) with $t_1 = s$ and using (4.15), we obtain

$$\begin{aligned}(8.24)\quad &\frac{\alpha}{2}\|\mathbf{E}^{\mathrm{p}}(t_2)-\mathbf{E}^{\mathrm{p}}(s)\|_{L^1}+\alpha_l\|D\mathbf{E}^{\mathrm{p}}(t_2)-D\mathbf{E}^{\mathrm{p}}(s)\|_{\mathcal{M}_b}\\ &\quad\le \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(s))-\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(t_2))+\mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(s))-\mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t_2))\\ &\qquad+\langle\mathcal{L}(t_2),u(t_2)\rangle-\langle\mathcal{L}(s),u(s)\rangle+\int_s^{t_2}\!\!\int_\Omega \mathbf{T}(\tau):\mathbf{E}\dot w(\tau)\,dx\,d\tau\\ &\qquad\qquad-\int_s^{t_2}\langle\dot{\mathcal{L}}(\tau),u(\tau)\rangle\,d\tau-\int_s^{t_2}\langle\mathcal{L}(\tau),\dot w(\tau)\rangle\,d\tau\\ &\qquad\qquad\qquad-\int_\Omega \rho_D(t_2)(\mathbf{E}^{\mathrm{p}}(t_2)-\mathbf{E}^{\mathrm{p}}(s))\,dx\\ &\quad\le \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(s))-\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(t_2))+\mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(s))-\mathcal{Q}_2(\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t_2))\\ &\qquad+\int_\Omega \rho(t_2):(\mathbf{E}^{\mathrm{e}}(t_2)-\mathbf{E}^{\mathrm{e}}(s))\,dx+\int_\Omega(\rho(t_2)-\rho(s)):\mathbf{E}^{\mathrm{e}}(s)\,dx\\ &\qquad\qquad+\int_\Omega(\rho_D(t_2)-\rho_D(s)):\mathbf{E}^{\mathrm{p}}(s)\,dx\\ &\qquad-\int_s^{t_2}\!\!\int_\Omega\left\{\dot\rho(\tau):\mathbf{E}^{\mathrm{e}}(\tau)+\dot\rho_D(\tau):\mathbf{E}^{\mathrm{p}}(\tau)-(\mathbf{T}(\tau)-\rho(\tau)):\mathbf{E}\dot w(\tau)\right\}dx\,d\tau.\end{aligned}$$

Notice that

$$\sup_\tau\|\rho(\tau)\|_{L^2},\quad \sup_\tau\|\rho_D(\tau)\|_{L^\infty},$$

and

$$\sup_\tau\|\mathbf{E}^{\mathrm{e}}(\tau)\|_{L^2},\quad \sup_\tau\|\mathbf{E}^{\mathrm{p}}(\tau)\|_{L^1},\quad \sup_\tau\|\mathrm{curl}\mathbf{E}^{\mathrm{p}}(\tau)\|_{L^2}$$

are finite (in fact $t\mapsto\mathbf{E}^{\mathrm{p}}(t)$ has bounded variation, while for $\mathbf{E}^{\mathrm{e}}(t)$ and $\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)$ we can use the energy balance (5.3)). From (8.24) we obtain for every $t_1 \le s \le t_2$

$$\begin{aligned}(8.25)\quad &\frac{\alpha}{2}\|\mathbf{E}^{\mathrm{p}}(t_2)-\mathbf{E}^{\mathrm{p}}(s)\|_{L^1}+\alpha_l\|D\mathbf{E}^{\mathrm{p}}(t_2)-D\mathbf{E}^{\mathrm{p}}(s)\|_{\mathcal{M}_b}\\ &\quad\le C_1\left(\|\mathbf{E}^{\mathrm{e}}(t_2)-\mathbf{E}^{\mathrm{e}}(s)\|_{L^2}+\sqrt{\frac{\mu}{2}}L\|\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t_2)-\mathrm{curl}\mathbf{E}^{\mathrm{p}}(s)\|_{L^2}\right.\\ &\qquad\qquad\left.+\int_s^{t_2}\psi(\tau)\,d\tau\right),\end{aligned}$$

where

$$(8.26)\qquad \psi(\tau) := \|\dot\rho(\tau)\|_{L^2}+\|\dot\rho_D(\tau)\|_{L^\infty}+\|\mathbf{E}\dot w(\tau)\|_{L^2}$$

and C_1 depends on ρ, $\sup_\tau \|\mathbf{E}^{\mathrm{e}}(\tau)\|_{L^2}$, $\sup_\tau \|\mathbf{E}^{\mathrm{p}}(\tau)\|_{L^1}$, $\sup_\tau L\|\mathrm{curl}\mathbf{E}^{\mathrm{p}}(\tau)\|_{L^2}$, and the elasticity tensor $\mathbb{C}$. By (8.22) we conclude that

$$\begin{aligned}
&\alpha_{\mathbb{C}}\|\mathbf{E}^{\mathrm{e}}(t_2)-\mathbf{E}^{\mathrm{e}}(t_1)\|_{L^2}^2+\frac{\mu L^2}{2}\|\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t_2)-\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t_1)\|_{L^2}^2\\
&\quad\le C_2\left(\|\mathbf{E}^{\mathrm{e}}(t_2)-\mathbf{E}^{\mathrm{e}}(t_1)\|_{L^2}+\sqrt{\frac{\mu}{2}}L\|\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t_2)-\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t_1)\|_{L^2}\right)\int_{t_1}^{t_2}\psi(\tau)\,d\tau\\
&\quad+C_2\int_{t_1}^{t_2}\psi(\tau)\left(\|\mathbf{E}^{\mathrm{e}}(\tau)-\mathbf{E}^{\mathrm{e}}(t_1)\|_{L^2}+\sqrt{\frac{\mu}{2}}L\|\mathrm{curl}\mathbf{E}^{\mathrm{p}}(\tau)-\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t_1)\|_{L^2}\right)d\tau\\
&\qquad\qquad+C_2\left(\int_{t_1}^{t_2}\psi(\tau)\,d\tau\right)^2,
\end{aligned}$$

where C_2 depends also on α. By Cauchy's inequality we obtain

$$\begin{aligned}
&\|\mathbf{E}^{\mathrm{e}}(t_2)-\mathbf{E}^{\mathrm{e}}(t_1)\|_{L^2}^2+\frac{\mu L^2}{2}\|\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t_2)-\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t_1)\|_{L^2}^2\\
&\quad\le C_3\int_{t_1}^{t_2}\psi(\tau)\left(\|\mathbf{E}^{\mathrm{e}}(\tau)-\mathbf{E}^{\mathrm{e}}(t_1)\|_{L^2}+\sqrt{\frac{\mu}{2}}L\|\mathrm{curl}\mathbf{E}^{\mathrm{p}}(\tau)-\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t_1)\|_{L^2}\right)d\tau\\
&\qquad\qquad+C_3\left(\int_{t_1}^{t_2}\psi(\tau)\,d\tau\right)^2.
\end{aligned}$$

By means of a Gronwall-type lemma [8, Lemma 5.3] we get in particular that

$$\|\mathbf{E}^{\mathrm{e}}(t_2)-\mathbf{E}^{\mathrm{e}}(t_1)\|_{L^2}+\sqrt{\frac{\mu}{2}}L\|\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t_2)-\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t_1)\|_{L^2}\le C_4\int_{t_1}^{t_2}\psi(\tau)\,d\tau, \tag{8.27}$$

where C_4 depends on ρ, α, $\sup_\tau \|\mathbf{E}^{\mathrm{e}}(\tau)\|_{L^2}$, $\sup_\tau \|\mathbf{E}^{\mathrm{p}}(\tau)\|_{L^1}$, $\sup_\tau L\|\mathrm{curl}\mathbf{E}^{\mathrm{p}}(\tau)\|_{L^2}$, and the elasticity tensor $\mathbb{C}$. As a consequence we get that $t\mapsto \mathbf{E}^{\mathrm{e}}(t)$ and $t\mapsto \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)$ are absolutely continuous from $[0,T]$ to $L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}})$ and $L^2(\Omega;\mathrm{M}^{N\times N})$, respectively. By (8.25), we get that $t\mapsto \mathbf{E}^{\mathrm{p}}(t)$ is absolutely continuous from $[0,T]$ to $BV(\Omega;\mathrm{M}^{N\times N}_D)$.

Let us now come to the proof of (8.14)–(8.17). By the energy balance (5.3), and by the very definition of $\mathcal{H}$, we deduce that

$$\|\mathbf{E}^{\mathrm{p}}(t)\|_{L^1}\le C_5\left(1+\int_0^t(1+\psi(\tau))\|\mathbf{E}^{\mathrm{e}}(\tau)\|_{L^2}\,d\tau+\int_0^t(1+\psi(\tau))\|\mathbf{E}^{\mathrm{p}}(\tau)\|_{L^1}\,d\tau\right),$$

where ψ is as in (8.26) and C_5 depends only on the initial conditions and on w,ρ,α, and $\mathbb{C}$. By means of the classical Gronwall lemma and taking the sup in t, we obtain

$$\sup_{t\in[0,T]}\|\mathbf{E}^{\mathrm{p}}(t)\|_{L^1}\le C_6\left(1+\sup_{t\in[0,T]}\|\mathbf{E}^{\mathrm{e}}(t)\|_{L^2}\right).$$

By the energy balance (5.3) we conclude that $\sup_{t\in[0,T]}\|\mathbf{E}^{\mathrm{e}}(t)\|_{L^2}$ is bounded uniformly independently of l and L, so that the same holds for $\sup_{t\in[0,T]}\|\mathbf{E}^{\mathrm{p}}(t)\|_{L^1}$ and $\sup_{t\in[0,T]}L\|\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)\|_{L^2}$. By (8.27) we conclude that (8.15) and (8.17) hold. Inequalities (8.16) and (8.19) follow by (8.25). Finally the absolute continuity of $t\mapsto u(t)$ from $[0,T]$ to $W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N)$ and inequality (8.14) follow from the compatibility condition $\mathbf{E}u(t):=\mathbf{E}^{\mathrm{e}}(t)+\mathbf{E}^{\mathrm{p}}(t)$ and Korn's inequality. Inequality (8.18) follows in a similar way. □

8.6. Notice that the constants $C_1, \dots, C_6$ of Proposition 8.5 depend also on the initial condition $(u_0, \mathbf{E}^{\mathrm{e}}_0, \mathbf{E}^{\mathrm{p}}_0)$. More precisely, from the previous proof it can be evicted that they depend on $|\mathcal{E}(0)|$ and $\|\mathbf{E}^{\mathrm{p}}_0\|_{L^1(\Omega;\mathrm{M}^{N\times N}_D)}$.

In order to prove the uniqueness result and the flow rule, we need the following lemma.

LEMMA 8.7. $t \in [0,T]$ $\dot u(t), \dot{\mathbf{E}}^{\mathrm{e}}(t), \dot{\mathbf{E}}^{\mathrm{p}}(t)$ $\dot w(t)$

$$\mathcal{H}(\dot{\mathbf{E}}^{\mathrm{p}}(t)) = -\int_\Omega \mathbf{T}(t) : (\dot{\mathbf{E}}^{\mathrm{e}}(t) - \mathbf{E}\dot w(t))\,dx - \mu L^2 \int_\Omega \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t) : \mathrm{curl}\dot{\mathbf{E}}^{\mathrm{p}}(t)\,dx + \langle \mathcal{L}(t), \dot u(t) - \dot w(t)\rangle \tag{8.28}$$

$$\mathcal{H}(\dot{\mathbf{E}}^{\mathrm{p}}(t)) = \int_\Omega \mathbf{T}^{\mathrm{p}}(t) : \dot{\mathbf{E}}^{\mathrm{p}}(t)\,dx + \int_\Omega \mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t) : \nabla\dot{\mathbf{E}}^{\mathrm{p}}(t)\,dx + \langle \mathbb{S}^{\mathrm{p}}(t), D^s\dot{\mathbf{E}}^{\mathrm{p}}(t)\rangle. \tag{8.29}$$

. Since by Proposition 8.5 the map $t \mapsto \mathbf{E}^{\mathrm{p}}(t)$ is absolutely continuous from $[0,T]$ to $BV(\Omega;\mathrm{M}^{N\times N}_D)$, by [8, Theorem 7.1] we obtain

$$\mathcal{D}_{\mathcal{H}}(\mathbf{E}^{\mathrm{p}}; 0, t) = \int_0^t \mathcal{H}(\dot{\mathbf{E}}^{\mathrm{p}}(\tau))\,d\tau.$$

Then by differentiating the energy balance equation (5.3) we obtain for a.e. $t \in [0,T]$

$$\begin{aligned}\int_\Omega \mathbf{T}(t) : \dot{\mathbf{E}}^{\mathrm{e}}(t)\,dx + \mu L^2 \int_\Omega \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t) : \mathrm{curl}\dot{\mathbf{E}}^{\mathrm{p}}(t)\,dx \\ - \langle \dot{\mathcal{L}}(t), u(t)\rangle - \langle \mathcal{L}(t), \dot u(t)\rangle + \mathcal{H}(\dot{\mathbf{E}}^{\mathrm{p}}(t)) \\ = \int_\Omega \mathbf{T}(t) : \mathbf{E}\dot w(t)\,dx - \langle \dot{\mathcal{L}}(t), u(t)\rangle - \langle \mathcal{L}(t), \dot w(t)\rangle\end{aligned}$$

so that (8.28) follows. Since $(\dot u(t) - \dot w(t), \dot{\mathbf{E}}^{\mathrm{e}}(t) - \mathbf{E}\dot w(t), \dot{\mathbf{E}}^{\mathrm{p}}(t)) \in \mathcal{A}(0)$, by (8.2) we get that (8.29) holds. □

Let us now prove the uniqueness result concerning the elastic strain $\mathbf{E}^{\mathrm{e}}(t)$ and the "energetic plastic strain" $\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)$.

PROPOSITION 8.8. $t \mapsto \mathbf{E}^{\mathrm{e}}(t)$ $t \mapsto \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)$ $(u_0, \mathbf{E}^{\mathrm{e}}_0, \mathbf{E}^{\mathrm{p}}_0)$

. Let $t \mapsto (\tilde u(t), \widetilde{\mathbf{E}^{\mathrm{e}}}(t), \widetilde{\mathbf{E}^{\mathrm{p}}}(t))$ be another quasi-static evolution associated to the same initial condition $(u_0, \mathbf{E}^{\mathrm{e}}_0, \mathbf{E}^{\mathrm{p}}_0)$. Let $\widetilde{\mathbf{T}^{\mathrm{p}}}(t)$, $\widetilde{\mathbb{K}^{\mathrm{p}}}(t)$, and $\widetilde{\mathbb{S}^{\mathrm{p}}}(t)$ be the associated stresses according to Lemma 8.1 and Proposition 8.3. By (8.28), (8.1), and (8.2) we have for a.e. $t \in [0,T]$

$$\begin{aligned}-\int_\Omega \mathbf{T}(t) : (\dot{\mathbf{E}}^{\mathrm{e}}(t) - \mathbf{E}\dot w(t))\,dx - \mu L^2 \int_\Omega \mathrm{curl}\mathbf{E}^{\mathrm{p}}(t) : \mathrm{curl}\dot{\mathbf{E}}^{\mathrm{p}}(t)\,dx \\ + \langle \mathcal{L}(t), \dot u(t) - \dot w(t)\rangle = \mathcal{H}(\dot{\mathbf{E}}^{\mathrm{p}}(t)) \\ \geq \int_\Omega \widetilde{\mathbf{T}^{\mathrm{p}}}(t) : \dot{\mathbf{E}}^{\mathrm{p}}(t)\,dx + \int_\Omega \widetilde{\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}}(t) : \nabla\dot{\mathbf{E}}^{\mathrm{p}}(t)\,dx + \langle \widetilde{\mathbb{S}^{\mathrm{p}}}(t), D^s\dot{\mathbf{E}}^{\mathrm{p}}(t)\rangle \\ = -\int_\Omega \widetilde{\mathbf{T}}(t) : (\dot{\mathbf{E}}^{\mathrm{e}}(t) - \mathbf{E}\dot w(t))\,dx - \mu L^2 \int_\Omega \mathrm{curl}\widetilde{\mathbf{E}^{\mathrm{p}}}(t) : \mathrm{curl}\dot{\mathbf{E}}^{\mathrm{p}}(t)\,dx \\ + \langle \mathcal{L}(t), \dot u(t) - \dot w(t)\rangle.\end{aligned}$$

We obtain

$$\int_\Omega \Big[\mathbf{T}(t)-\widetilde{\mathbf{T}}(t)\Big] : \Big[\dot{\mathbf{E}}^{\mathrm{e}}(t)-\mathbf{E}\dot{w}(t)\Big]\,dx+\mu L^2\int_\Omega \Big[\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)-\mathrm{curl}\widetilde{\mathbf{E}^{\mathrm{p}}}(t)\Big] : \mathrm{curl}\dot{\mathbf{E}}^{\mathrm{p}}(t)\,dx \le 0.$$

Similarly we obtain

$$\int_\Omega \Big[\widetilde{\mathbf{T}}(t)-\mathbf{T}(t)\Big] : \Big[\dot{\widetilde{\mathbf{E}^{\mathrm{e}}}}(t)-\mathbf{E}\dot{w}(t)\Big]\,dx+\mu L^2\int_\Omega \Big[\mathrm{curl}\widetilde{\mathbf{E}^{\mathrm{p}}}(t)-\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)\Big] : \mathrm{curl}\dot{\widetilde{\mathbf{E}^{\mathrm{p}}}}(t)\,dx \le 0.$$

By summing the two inequalities we get

$$\begin{aligned}\int_\Omega \Big[\mathbf{T}(t)-\widetilde{\mathbf{T}}(t)\Big] : \Big[\dot{\mathbf{E}}^{\mathrm{e}}(t)-\dot{\widetilde{\mathbf{E}^{\mathrm{e}}}}(t)\Big]\,dx\\ +\mu L^2\int_\Omega \Big[\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)-\mathrm{curl}\widetilde{\mathbf{E}^{\mathrm{p}}}(t)\Big] : \Big[\mathrm{curl}\dot{\mathbf{E}}^{\mathrm{p}}(t)-\mathrm{curl}\dot{\widetilde{\mathbf{E}^{\mathrm{p}}}}(t)\Big]\,dx \le 0,\end{aligned}$$

so that for a.e. $t\in[0,T]$ we have

$$\begin{aligned}\frac{d}{dt}\Bigg(\int_\Omega \mathbb{C}\Big[\mathbf{E}^{\mathrm{e}}(t)-\widetilde{\mathbf{E}^{\mathrm{e}}}(t)\Big] : \Big[\mathbf{E}^{\mathrm{e}}(t)-\widetilde{\mathbf{E}^{\mathrm{e}}}(t)\Big]\,dx\\ +\mu L^2\int_\Omega |\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)-\mathrm{curl}\widetilde{\mathbf{E}^{\mathrm{p}}}(t)|^2\,dx\Bigg) \le 0.\end{aligned}$$

Since we have $\widetilde{\mathbf{E}^{\mathrm{e}}}(0)=\mathbf{E}^{\mathrm{e}}_0=\mathbf{E}^{\mathrm{e}}(0)$ and $\mathrm{curl}\widetilde{\mathbf{E}^{\mathrm{p}}}(0)=\mathrm{curl}\mathbf{E}^{\mathrm{p}}_0=\mathrm{curl}\mathbf{E}^{\mathrm{p}}(0)$, the previous inequality and the absolute continuity of the maps involved yield that $\mathbf{E}^{\mathrm{e}}(t)=\widetilde{\mathbf{E}^{\mathrm{e}}}(t)$ and $\mathrm{curl}\widetilde{\mathbf{E}^{\mathrm{p}}}(t)=\mathrm{curl}\mathbf{E}^{\mathrm{p}}(t)$ for every $t\in[0,T]$, so that the proof is concluded. □

We are now in a position to prove the flow rule for the Gurtin–Anand model. Let us start with the following weak form.

PROPOSITION 8.9 (weak form of the flow rule). [illegible] $t\in[0,T]$ [illegible]

(a) [illegible] $(\mathbf{A},\mathbb{B})\in L^\infty(\Omega;\mathrm{M}^{N\times N}_D)\times L^\infty(\Omega;\mathrm{M}^{N\times N\times N}_D)$ [illegible] $\sqrt{|\mathbf{A}(x)|^2+l^{-2}|\mathbb{B}(x)|^2}\le S_Y$ [illegible] $x\in\Omega$ [illegible]

$$\int_\Omega (\mathbf{T}^{\mathrm{p}}(t)-\mathbf{A}) : \dot{\mathbf{E}}^{\mathrm{p}}(t)\,dx+\int_\Omega (\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t)-\mathbb{B}) : \nabla\dot{\mathbf{E}}^{\mathrm{p}}(t)\,dx \ge 0. \tag{8.30}$$

(b) [illegible] $\mathbb{L}\in(\mathcal{M}_b(\Omega;\mathrm{M}^{N\times N\times N}_D)^*$ [illegible] $\|\mathbb{L}\|_{(\mathcal{M}_b(\Omega;\mathrm{M}^{N\times N\times N}_D))^*}\le lS_Y$ [illegible]

$$\langle \mathbb{S}^{\mathrm{p}}(t)-\mathbb{L}, D^s\dot{\mathbf{E}}^{\mathrm{p}}(t)\rangle \ge 0. \tag{8.31}$$

[illegible]. Recall that for a.e. $t\in[0,T]$

$$\begin{aligned}\mathcal{H}(\dot{\mathbf{E}}^{\mathrm{p}}(t)) &= \mathcal{F}(\dot{\mathbf{E}}^{\mathrm{p}}(t),\nabla\dot{\mathbf{E}}^{\mathrm{p}}(t),D^s\dot{\mathbf{E}}^{\mathrm{p}}(t))\\ &:= S_Y\int_\Omega \sqrt{|\dot{\mathbf{E}}^{\mathrm{p}}|^2+l^2|\nabla\dot{\mathbf{E}}^{\mathrm{p}}|^2}\,dx+lS_Y|D^s\dot{\mathbf{E}}^{\mathrm{p}}|(\Omega).\end{aligned}$$

Since $\mathcal{F}: L^1(\Omega;\mathrm{M}^{N\times N}_D)\times L^1(\Omega;\mathrm{M}^{N\times N\times N}_D)\times\mathcal{M}_b(\Omega;\mathrm{M}^{N\times N\times N}_D)\to[0,+\infty[$ is continuous (with respect to the strong norm), we have $\mathcal{F}(\dot{\mathbf{E}}^{\mathrm{p}}(t),\nabla\dot{\mathbf{E}}^{\mathrm{p}}(t),D^s\dot{\mathbf{E}}^{\mathrm{p}}(t))=$

$\mathcal{F}^{**}(\dot{\mathbf{E}}^{\mathrm{p}}(t), \nabla\dot{\mathbf{E}}^{\mathrm{p}}(t), D^s\dot{\mathbf{E}}^{\mathrm{p}}(t))$, where $*$ denotes the Fenchel transformation. Moreover, we have that $\mathcal{F}^*$ is the indicator function of the set

$$\mathcal{K} := \Big\{(\mathbf{A},\mathbb{B},\mathbb{L}) \in L^\infty\left(\Omega;\mathrm{M}^{N\times N}_D\right) \times L^\infty\left(\Omega;\mathrm{M}^{N\times N\times N}_D\right) \times \left(\mathcal{M}_b\left(\Omega;\mathrm{M}^{N\times N\times N}_D\right)\right)^* : \sqrt{|\mathbf{A}|^2 + l^{-2}|\mathbb{B}|^2} \le S_Y \text{ a.e. in } \Omega \text{ and } \|\mathbb{L}\|_{\mathcal{M}_b^*} \le lS_Y\Big\}.$$

As a consequence, by (8.29) we deduce that for every $(\mathbf{A},\mathbb{B},\mathbb{L}) \in \mathcal{K}$ we have

$$\int_\Omega (\mathbf{T}^{\mathrm{p}}(t) - \mathbf{A}) : \dot{\mathbf{E}}^{\mathrm{p}}(t)\,dx + \int_\Omega (\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t) - \mathbb{B}) : \nabla\dot{\mathbf{E}}^{\mathrm{p}}(t)\,dx + \langle \mathbb{S}^{\mathrm{p}}(t) - \mathbb{L}, D^s\dot{\mathbf{E}}^{\mathrm{p}}(t)\rangle \ge 0.$$

By choosing $\mathbb{L} = \mathbb{S}^{\mathrm{p}}(t)$, which is possible in view of the constraint (8.10), we obtain (8.30). Inequality (8.31) follows by choosing $\mathbf{A} = \mathbf{T}^{\mathrm{p}}(t)$ and $\mathbb{B} = \mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t)$. □

Let us now prove that the weak flow rule (8.30) for the stresses $\mathbf{T}^{\mathrm{p}}(t)$ and $\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t)$ reduces under suitable regularity assumptions to the usual flow rule given by Gurtin and Anand.

PROPOSITION 8.10 (flow rule). [illegible] $t \in [0,T]$ [illegible] $\dot{\mathbf{E}}^{\mathrm{p}}(t)$ [illegible] $x \in \Omega$ [illegible] $\mathbf{T}^{\mathrm{p}}(t)$, $\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t)$, $\dot{\mathbf{E}}^{\mathrm{p}}(t)$, $\nabla\dot{\mathbf{E}}^{\mathrm{p}}(t)$ [illegible]

$$\sqrt{|\mathbf{T}^{\mathrm{p}}(t,x)|^2 + l^{-2}|\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t,x)|^2} < S_Y,$$

[illegible]

$$(\dot{\mathbf{E}}^{\mathrm{p}}(t,x), \nabla\dot{\mathbf{E}}^{\mathrm{p}}(t,x)) = (0,0), \tag{8.32}$$

[illegible]

$$\sqrt{|\mathbf{T}^{\mathrm{p}}(t,x)|^2 + l^{-2}|\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t,x)|^2} = S_Y,$$

[illegible]

$$\begin{cases} \mathbf{T}^{\mathrm{p}}(t,x) = S_Y \dfrac{\dot{\mathbf{E}}^{\mathrm{p}}(t,x)}{\sqrt{|\dot{\mathbf{E}}^{\mathrm{p}}(t,x)|^2 + l^2|\nabla\dot{\mathbf{E}}^{\mathrm{p}}(t,x)|^2}}, \\ \mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t,x) = S_Y \dfrac{l^2\nabla\dot{\mathbf{E}}^{\mathrm{p}}(t,x)}{\sqrt{|\dot{\mathbf{E}}^{\mathrm{p}}(t,x)|^2 + l^2|\nabla\dot{\mathbf{E}}^{\mathrm{p}}(t,x)|^2}}. \end{cases} \tag{8.33}$$

[illegible]. Let K be the convex set defined as

$$K := \left\{(\mathbf{A},\mathbb{B}) \in \mathrm{M}^{N\times N}_D \times \mathrm{M}^{N\times N\times N}_D : \sqrt{|\mathbf{A}|^2 + l^{-2}|\mathbb{B}|^2} \le S_Y\right\}.$$

Let π_K denote the projection onto K, and let π^1_K and π^2_K be its components. Let $(\mathbf{A},\mathbb{B}) \in K$, $\varepsilon > 0$, and let us set

$$\mathcal{C}^\varepsilon_{\mathbf{A},\mathbb{B}} := (\mathbf{T}^{\mathrm{p}}(t) + \varepsilon(\mathbf{A} - \mathbf{T}^{\mathrm{p}}(t,x)), \mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t) + \varepsilon(\mathbb{B} - \mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t,x))) \in L^\infty\left(\Omega;\mathrm{M}^{N\times N}_D\right) \times L^\infty\left(\Omega;\mathrm{M}^{N\times N\times N}_D\right).$$

For every $r > 0$ let us set

$$\mathbf{F} := \begin{cases} \pi^1_K\left(\mathcal{C}^\varepsilon_{\mathbf{A},\mathbb{B}}\right) & \text{in } B(x,r), \\ \mathbf{T}^{\mathrm{p}}(t) & \text{outside } B(x,r) \end{cases}$$

and

$$\mathbb{G} := \begin{cases} \pi_K^2\left(\mathcal{C}^{\varepsilon}_{\mathbf{A},\mathbb{B}}\right) & \text{in } B(x,r), \\ \mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t) & \text{outside } B(x,r). \end{cases}$$

Since $(\mathbf{F}, \mathbb{G})$ are admissible for the weak flow rule (8.30), we obtain

$$\frac{1}{r^N}\left[\int_{B(x,r)} (\mathbf{T}^{\mathrm{p}}(t) - \mathbf{F}) : \dot{\mathbf{E}}^{\mathrm{p}}(t)\, dx + \int_{B(x,r)} (\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t) - \mathbb{G}) : \nabla\dot{\mathbf{E}}^{\mathrm{p}}(t)\, dx\right] \geq 0.$$

Since π_K is a Lipschitz mapping, we have that x is also a Lebesgue point for $\pi_K(\mathcal{C}^{\varepsilon}_{\mathbf{A},\mathbb{B}})$ with Lebesgue value

$$\pi_K(\mathbf{T}^{\mathrm{p}}(t,x) + \varepsilon(\mathbf{A} - \mathbf{T}^{\mathrm{p}}(t,x)), \mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t,x) + \varepsilon(\mathbb{B} - \mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t,x))).$$

By sending $r \to 0$ and considering $0 < \varepsilon < 1$, in view of the convexity of K, we obtain

$$(\mathbf{A} - \mathbf{T}^{\mathrm{p}}(t,x)) : \dot{\mathbf{E}}^{\mathrm{p}}(t,x) + (\mathbb{B} - \mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t,x)) : \nabla\dot{\mathbf{E}}^{\mathrm{p}}(t,x) \leq 0.$$

Since the previous inequality holds for every $(\mathbf{A}, \mathbb{B}) \in K$, we deduce that $(\dot{\mathbf{E}}^{\mathrm{p}}(t,x), \nabla\dot{\mathbf{E}}^{\mathrm{p}}(t,x))$ belongs to the normal cone to K at $(\mathbf{T}^{\mathrm{p}}(t,x), \mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t,x))$. In particular, if $(\mathbf{T}^{\mathrm{p}}(t,x), \mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t,x)) \in \mathrm{int}K$, we get that (8.32) holds, while if $(\mathbf{T}^{\mathrm{p}}(t,x), \mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t,x)) \in \partial K$, (8.33) follows. ∎

8.11. Notice that, in view of the presence of a singular part for $D\mathbf{E}^{\mathrm{p}}(t)$ and of its associated stress, plasticity can develop at a point $x \in \Omega$ also when $\|\mathbb{S}^{\mathrm{p}}\|_{\mathcal{M}_b(\Omega;\mathrm{M}^{N\times N\times N}_D)} = lS_Y$ and $\sqrt{|\mathbf{T}^{\mathrm{p}}(t,x)|^2 + l^{-2}|\mathbb{K}^{\mathrm{p}}_{\mathrm{diss}}(t,x)|^2} < S_Y$.

9. Asymptotic analysis as $l \to 0$ and $L \to 0$. In this section we want to understand the behavior of a quasi-static evolution for the Gurtin–Anand model as the length scales l and L vanish. Our goal is to prove that the quasi-static evolution converges in a suitable sense to an evolution for perfect plasticity. The result is somehow natural, since the strain gradient effects vanish.

More precisely, we prove under suitable assumptions the convergence to a quasi-static evolution for linearly elastic-perfectly plastic bodies recently proposed by Dal Maso, DeSimone, and Mora [8]. The main mathematical problem that we have to face in order to prove such a convergence is that the functional setting of the problem changes, in particular for what concerns the plastic strains. In fact in the strain gradient context, the plastic strain is a BV function (since its gradient enters in the equations), while in [8] it is modeled only as a Radon measure in $\Omega \cup \partial_D\Omega$. Similar problems occur for the displacements, in view of the compatibility condition.

In section 9.1 we briefly recall the model for quasi-static evolution in perfect plasticity recently proposed in [8]. Section 9.2 is devoted to the proof of the convergence result (Theorem 9.2).

9.1. The Dal Maso–DeSimone–Mora model for perfect plasticity. Let us briefly recall the model for quasi-static evolution in perfect plasticity recently proposed in [8]. We formulate the results in the particular form that we need for our asymptotic problem, using the notation of the previous sections.

Let $\Omega \subseteq \mathbb{R}^N$ $(N \geq 2)$ be open bounded, let $\partial_D\Omega$ and $\partial_N\Omega$ have the same boundary Γ (relative to $\partial\Omega$), and let us assume that

$$\partial\Omega \text{ and } \Gamma \text{ are of class } C^2. \tag{9.1}$$

Given $w \in W^{1,2}(\Omega;\mathbb{R}^N)$, the class of admissible configurations for the boundary datum w is given by

$$\mathcal{A}_{\rm pp}(w) := \big\{(u,\mathbf{E}^{\rm e},\mathbf{E}^{\rm p}) \in BD(\Omega)\times L^2(\Omega;\mathrm{M}^{N\times N}_{\rm sym})\times \mathcal{M}_b(\Omega\cup\partial_D\Omega;\mathrm{M}^{N\times N}_D) : \\ \mathbf{E}u = \mathbf{E}^{\rm e}+\mathbf{E}^{\rm p} \text{ in } \Omega, \mathbf{E}^{\rm p} = (w-u)\odot \nu\, d\mathcal{H}^{N-1} \text{ on } \partial_D\Omega\big\}.$$

Here $BD(\Omega)$ denotes the space of functions with bounded deformation on Ω:

$$BD(\Omega) := \big\{u\in L^1(\Omega;\mathbb{R}^N) : \mathbf{E}u \in \mathcal{M}_b(\Omega;\mathrm{M}^{N\times N}_{\rm sym})\big\},$$

which is a Banach space with respect to the norm

$$\|u\|_{BD(\Omega)} := \|u\|_{L^1(\Omega;\mathbb{R}^N)} + \|\mathbf{E}u\|_{\mathcal{M}_b\left(\Omega;\mathrm{M}^{N\times N}_{\rm sym}\right)}.$$

We refer the reader to [33] for the main properties of $BD(\Omega)$. The term $(w-u)$ on $\partial_D\Omega$ is intended in the sense of traces. Finally the subscripts " pp" stand for "perfect plasticity."

Given $\mathbf{E}^{\rm p}\in\mathcal{M}_b(\Omega\cup\partial_D\Omega;\mathrm{M}^{N\times N}_D)$, we set

$$\mathcal{H}_{\rm pp}(\mathbf{E}^{\rm p}) := S_Y|\mathbf{E}^{\rm p}|(\Omega\cup\partial_D\Omega),$$

while for $\mathbf{E}^{\rm e}\in L^2(\Omega;\mathrm{M}^{N\times N}_{\rm sym})$, we consider $\mathcal{Q}_1(\mathbf{E}^{\rm e})$ as defined in (4.6).

Let $t\in[0,T]$, and let the boundary displacement be given by

$$w \in AC(0,T;W^{1,2}(\Omega;\mathbb{R}^N)). \tag{9.2}$$

Let the body and traction forces be given by

$$f\in AC(0,T;L^N(\Omega;\mathbb{R}^N)) \qquad \text{and} \qquad g\in AC(0,T;L^\infty(\partial_N\Omega;\mathbb{R}^N)), \tag{9.3}$$

respectively, and let us denote by $\mathcal{L}(t)$ the associated work as in (4.12). Let us assume that f and g satisfy the uniform safe load condition (4.13)–(4.14). We can simply suppose as in [8] that $t\mapsto\rho(t)$ is absolutely continuous from $[0,T]$ to $L^2(\Omega;\mathrm{M}^{N\times N}_{\rm sym})$, since in view of the regularity of Ω we get $\rho(t)\in L^N(\Omega;\mathrm{M}^{N\times N}_{\rm sym})$ by the embedding result [22, Proposition 2.5].

Given an initial configuration

$$(u_0,\mathbf{E}^{\rm e}_0,\mathbf{E}^{\rm p}_0)\in\mathcal{A}_{\rm pp}(w(0)),$$

a quasi-static evolution $t\mapsto(u(t),\mathbf{E}^{\rm e}(t),\mathbf{E}^{\rm p}(t))$ in the sense of Dal Maso–DeSimone–Mora [8] is a map from $[0,T]$ to $BD(\Omega)\times L^2(\Omega;\mathrm{M}^{N\times N}_{\rm sym})\times\mathcal{M}_b(\Omega\cup\partial_D\Omega;\mathrm{M}^{N\times N}_D)$ with $(u(0),\mathbf{E}^{\rm e}(0),\mathbf{E}^{\rm p}(0)) = (u_0,\mathbf{E}^{\rm e}_0,\mathbf{E}^{\rm p}_0)$ and such that for every $t\in[0,T]$ the following facts hold:

(a) $(u(t),\mathbf{E}^{\rm e}(t),\mathbf{E}^{\rm p}(t))\in\mathcal{A}_{\rm pp}(w(t))$;

(b) Global stability: For every $(v,\mathbf{e},\mathbf{p})\in\mathcal{A}(w(t))$

$$\mathcal{Q}_1(\mathbf{E}^{\rm e}(t)) - \langle\mathcal{L}(t),u(t)\rangle \le \mathcal{Q}_1(\mathbf{e}) - \langle\mathcal{L}(t),v\rangle + \mathcal{H}_{\rm pp}(\mathbf{p}-\mathbf{E}^{\rm p}(t));$$

(b) Energy balance: The function $t\mapsto\mathbf{E}^{\rm p}(t)$ has a bounded variation from $[0,T]$ to $\mathcal{M}_b(\Omega\cup\partial_D\Omega;\mathrm{M}^{N\times N}_D)$ and

$$\mathcal{E}_{\rm pp}(t)+\mathcal{D}_{\rm pp}(\mathbf{E}^{\rm p};0,t) = \mathcal{E}_{\rm pp}(0) + \int_0^t\int_\Omega \mathbf{T}(\tau):\mathbf{E}\dot w(\tau)\,dx\,d\tau \\ - \int_0^t\langle\dot{\mathcal{L}}(\tau),u(\tau)\rangle\,d\tau - \int_0^t\langle\mathcal{L}(\tau),\dot w(\tau)\rangle\,d\tau,$$

where $\mathbf{T}(t) := \mathbb{C}\mathbf{E}^{\mathrm{e}}(t)$,

$$\mathcal{E}_{\mathrm{pp}}(t) := \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(t)) - \langle \mathcal{L}(t), u(t)\rangle$$

and $\mathcal{D}_{\mathrm{pp}}(\mathbf{E}^{\mathrm{p}}; 0, t) := S_Y \mathcal{V}(\mathbf{E}^{\mathrm{p}}; 0, t)$.

In order to prove the convergence result of the next section, we need to recall the pairing between stress and strain which gives a useful representation of the work $\mathcal{L}(t)$ similar to (4.15). Following [8, section 2], for every $t \in [0,T]$ and for every $(v, \mathbf{e}, \mathbf{p}) \in \mathcal{A}(w(t))$ it is possible to define the measure $[\rho_D(t) : \mathbf{p}] \in \mathcal{M}_b(\Omega \cup \partial_D\Omega)$ such that

$$\langle \mathcal{L}(t), v\rangle = -\langle \rho(t)\nu, w(t)\rangle_{\partial_D\Omega} + \int_\Omega \rho(t) : \mathbf{e}\, dx + [\rho_D(t) : \mathbf{p}](\Omega \cup \partial_D\Omega) \tag{9.4}$$

and such that for every $\varphi \in C^1(\overline{\Omega})$

$$\begin{aligned}\int_{\Omega\cup\partial_D\Omega} \varphi\, d[\rho_D(t) : \mathbf{p}] = \langle \mathcal{L}(t), \varphi v\rangle + \langle \rho(t)\nu, \varphi w(t)\rangle_{\partial_D\Omega} \\ - \int_\Omega \rho(t) : \varphi\mathbf{e}\, dx - \int_\Omega \rho(t) : [\nabla\varphi \odot v]\, dx.\end{aligned} \tag{9.5}$$

A similar pairing $[\dot\rho_D(t) : \mathbf{p}] \in \mathcal{M}_b(\Omega \cup \partial_D\Omega)$ can also be defined (for a.e. $t \in [0,T]$), so that (9.4) and (9.5) hold with $\dot\rho_D(t)$, $\dot\rho(t)$, and $\dot{\mathcal{L}}(t)$ in place of $\rho_D(t)$, $\rho(t)$, and $\mathcal{L}(t)$.

9.2. The convergence result as $l, L \to 0$. Let $\Omega \subseteq \mathbb{R}^N$ satisfy (9.1), and let w, f, and g be as in (9.2) and (9.3): Notice that these data are admissible for an evolution for the Gurtin and Anand model. Let us assume that f and g satisfy the uniform safe load condition (4.13)–(4.14).

Let us consider $l_n \to 0$ and $L_n \to 0$, and let us denote by

$$t \mapsto (u_n(t), \mathbf{E}^{\mathrm{e}}_n(t), \mathbf{E}^{\mathrm{p}}_n(t))$$

a quasi-static evolution for the Gurtin–Anand model relative to the data w, f, and g and the material length scales $l = l_n$ and $L = L_n$. Let us denote by $\mathcal{Q}_2^n$, $\mathcal{H}_n$, and $\mathcal{E}_n$ the energies corresponding to $\mathcal{Q}_2$, $\mathcal{H}$, and $\mathcal{E}$, respectively.

Let us assume that the initial configuration $(u_n(0), \mathbf{E}^{\mathrm{e}}_n(0), \mathbf{E}^{\mathrm{p}}_n(0))$ is such that there exist $u_0 \in BD(\Omega)$, $\mathbf{E}^{\mathrm{e}}_0 \in L^2(\Omega; \mathrm{M}^{N\times N}_{\mathrm{sym}})$, and $\mathbf{E}^{\mathrm{p}}_0 \in \mathcal{M}_b(\Omega; \mathrm{M}^{N\times N}_D)$ with

$$u_n(0) \overset{*}{\rightharpoonup} u_0 \qquad \text{weakly* in } BD(\Omega), \tag{9.6}$$

$$\mathbf{E}^{\mathrm{e}}_n(0) \rightharpoonup \mathbf{E}^{\mathrm{e}}_0 \qquad \text{weakly in } L^2(\Omega; \mathrm{M}^{N\times N}_{\mathrm{sym}}), \tag{9.7}$$

and

$$\mathbf{E}^{\mathrm{p}}_n(0) \overset{*}{\rightharpoonup} \mathbf{E}^{\mathrm{p}}_0 \qquad \text{weakly* in } \mathcal{M}_b(\Omega; \mathrm{M}^{N\times N}_D). \tag{9.8}$$

Recall that weak* convergence in $BD(\Omega)$ is given by weak convergence in L^1 for the functions and weak* convergence in the sense of measures for the symmetrized gradients.

Let us assume moreover that convergence for the initial free energies holds, that is,

$$\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}_n(0)) + \mathcal{Q}_2^n(\mathrm{curl}\mathbf{E}^{\mathrm{p}}_n(0)) \to \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}_0). \tag{9.9}$$

We have the following compactness result.

LEMMA 9.1. [illegible] $(u_n(0), \mathbf{E}^{\mathrm{e}}_n(0), \mathbf{E}^{\mathrm{p}}_n(0))$ [illegible] (9.6) (9.9) [illegible]

$$u \in AC(0,T;BD(\Omega)), \qquad \mathbf{E}^{\mathrm{e}} \in AC(0,T;L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}})),$$

[illegible]

$$\mathbf{E}^{\mathrm{p}} \in AC(0,T;\mathcal{M}_b(\Omega\cup\partial_D\Omega;\mathrm{M}^{N\times N}_D))$$

[illegible] $t \in [0,T]$

$$u_n(t) \overset{*}{\rightharpoonup} u(t) \quad \text{[illegible]} \quad BD(\Omega), \tag{9.10}$$

$$\mathbf{E}^{\mathrm{e}}_n(t) \rightharpoonup \mathbf{E}^{\mathrm{e}}(t) \quad \text{[illegible]} \quad L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}}), \tag{9.11}$$

[illegible] $\mathbf{E}^{\mathrm{p}}_n(t) = 0$ [illegible] $\partial_D\Omega$

$$\mathbf{E}^{\mathrm{p}}_n(t) \overset{*}{\rightharpoonup} \mathbf{E}^{\mathrm{p}}(t) \quad \text{[illegible]} \quad \mathcal{M}_b(\Omega\cup\partial_D\Omega;\mathrm{M}^{N\times N}_D). \tag{9.12}$$

[illegible] $t \in [0,T]$

$$(u(t), \mathbf{E}^{\mathrm{e}}(t), \mathbf{E}^{\mathrm{p}}(t)) \in \mathcal{A}_{\mathrm{pp}}(w(t)). \tag{9.13}$$

[illegible]. Let B be an open ball in $\mathbb{R}^N$ such that $\overline{\Omega} \subseteq B$, and let us set $\widetilde{\Omega} := B \setminus \overline{\partial_N\Omega}$. For every $t \in [0,T]$ let us consider $\tilde{u}_n(t) \in W^{1,\frac{N}{N-1}}(\widetilde{\Omega};\mathbb{R}^N)$, $\tilde{\mathbf{E}}^{\mathrm{e}}_n(t) \in L^2(\widetilde{\Omega};\mathrm{M}^{N\times N}_{\mathrm{sym}})$, and $\tilde{\mathbf{E}}^{\mathrm{p}}_n(t) \in BV(\widetilde{\Omega};\mathrm{M}^{N\times N}_D)$ defined as

$$\tilde{u}_n(t) := \begin{cases} u_n(t) & \text{in } \Omega, \\ w(t) & \text{in } \widetilde{\Omega}\setminus\Omega, \end{cases}$$

$$\tilde{\mathbf{E}}^{\mathrm{e}}_n(t) := \begin{cases} \mathbf{E}^{\mathrm{e}}_n(t) & \text{in } \Omega, \\ \mathbf{E}w(t) & \text{in } \widetilde{\Omega}\setminus\Omega, \end{cases}$$

and

$$\tilde{\mathbf{E}}^{\mathrm{p}}_n(t) := \begin{cases} \mathbf{E}^{\mathrm{p}}_n(t) & \text{in } \Omega, \\ 0 & \text{in } \widetilde{\Omega}\setminus\Omega. \end{cases}$$

By (9.6), (9.8), and (9.9) and in view of Remark 8.6 and Proposition 8.5, we deduce that $t \mapsto \tilde{u}_n(t)$, as a map from $[0,T]$ to $BD(\widetilde{\Omega})$, has a variation which is uniformly bounded independently on n. More precisely, the sequence $(u_n)_{n\in\mathbb{N}}$ is equiabsolutely continuous. The same holds for $t \mapsto \tilde{\mathbf{E}}^{\mathrm{e}}_n(t)$ and $t \mapsto \tilde{\mathbf{E}}^{\mathrm{p}}_n(t)$ considered as maps from $[0,T]$ to $L^2(\widetilde{\Omega};\mathrm{M}^{N\times N}_{\mathrm{sym}})$ and $L^1(\widetilde{\Omega};\mathrm{M}^{N\times N}_D)$, respectively.

Recall that $BD(\widetilde{\Omega})$ can be seen as a dual space, with associated weak* convergence given precisely by the weak* convergence in BD previously defined.

Then by considering $BD(\widetilde{\Omega})$ as a dual space and $L^1(\widetilde{\Omega};\mathrm{M}^{N\times N}_D)$ as a subspace of $\mathcal{M}_b(\widetilde{\Omega};\mathrm{M}^{N\times N}_D)$, we may apply the generalized version of Helly's theorem [8, Lemma 7.2] to obtain

$$\tilde{u} \in AC(0,T;BD(\widetilde{\Omega})), \qquad \tilde{\mathbf{E}}^{\mathrm{e}} \in AC(0,T;L^2(\widetilde{\Omega};\mathrm{M}^{N\times N}_{\mathrm{sym}})),$$

and

$$\tilde{\mathbf{E}}^{\mathrm{p}} \in AC(0,T;\mathcal{M}_b(\widetilde{\Omega};\mathrm{M}_D^{N\times N}))$$

such that, up to a subsequence, for every $t\in[0,T]$

$$\tilde{u}_n(t) \overset{*}{\rightharpoonup} \tilde{u}(t) \qquad \text{weakly* in } BD(\widetilde{\Omega}), \tag{9.14}$$

$$\tilde{\mathbf{E}}^{\mathrm{e}}_n(t) \rightharpoonup \tilde{\mathbf{E}}^{\mathrm{e}}(t) \qquad \text{weakly in } L^2(\widetilde{\Omega};\mathrm{M}^{N\times N}_{\mathrm{sym}}), \tag{9.15}$$

and

$$\tilde{\mathbf{E}}^{\mathrm{p}}_n(t) \overset{*}{\rightharpoonup} \tilde{\mathbf{E}}^{\mathrm{p}}(t) \qquad \text{weakly* in } \mathcal{M}_b(\widetilde{\Omega};\mathrm{M}_D^{N\times N}). \tag{9.16}$$

We have clearly that for every $t\in[0,T]$

$$\tilde{u}(t)=w(t), \quad \tilde{\mathbf{E}}^{\mathrm{e}}(t)=\mathbf{E}w(t), \quad \tilde{\mathbf{E}}^{\mathrm{p}}(t)=0 \qquad \text{on } \widetilde{\Omega}\setminus\overline{\Omega}.$$

Let us denote by $u(t)$ and $\mathbf{E}^{\mathrm{e}}(t)$ the restrictions of $\tilde{u}(t)$ and $\tilde{\mathbf{E}}^{\mathrm{e}}(t)$ to Ω, respectively, and let $\mathbf{E}^{\mathrm{p}}(t)$ denote the restriction of $\tilde{\mathbf{E}}^{\mathrm{p}}(t)$ to $\Omega\cup\partial_D\Omega$.

Relations (9.10) and (9.11) follow directly from (9.14) and (9.15). By (9.16), and by taking into account that $\tilde{\mathbf{E}}^{\mathrm{p}}_n(t)=0$ outside $\overline{\Omega}$, we obtain (9.12).

From the compatibility condition

$$\tilde{\mathbf{E}}u_n(t)=\tilde{\mathbf{E}}^{\mathrm{e}}_n(t)+\tilde{\mathbf{E}}^{\mathrm{p}}_n(t),$$

we deduce that in the limit we have

$$\tilde{\mathbf{E}}u(t)=\tilde{\mathbf{E}}^{\mathrm{e}}(t)+\tilde{\mathbf{E}}^{\mathrm{p}}(t)$$

so that

$$\mathbf{E}^{\mathrm{p}}(t)\llcorner\partial_D\Omega=\tilde{\mathbf{E}}^{\mathrm{p}}(t)\llcorner\partial_D\Omega=(w(t)-u(t))\odot\nu\,d\mathcal{H}^{N-1}\llcorner\partial_D\Omega,$$

where $u(t)$ is intended in the sense of traces on $\partial_D\Omega$. We deduce that (9.13) holds, and the proof is concluded. □

The main theorem of the section is the following asymptotic result.

THEOREM 9.2. [illegible] $t\mapsto(u_{l,L}(t),\mathbf{E}^{\mathrm{e}}_{l,L}(t),\mathbf{E}^{\mathrm{p}}_{l,L}(t))$ [illegible] (9.6) (9.9) [illegible] $l,L\to 0$ [illegible] $l_n\to 0$ [illegible] $L_n\to 0$ [illegible] $(l_{n_j},L_{n_j})_{j\in\mathbb{N}}$ [illegible] $t\mapsto(u(t),\mathbf{E}^{\mathrm{e}}(t),\mathbf{E}^{\mathrm{p}}(t))$ [illegible] [8] [illegible]

$$u_j:=u_{l_{n_j},L_{n_j}}, \qquad \mathbf{E}^{\mathrm{e}}_j:=\mathbf{E}^{\mathrm{e}}_{l_{n_j},L_{n_j}}, \qquad \mathbf{E}^{\mathrm{p}}_j:=\mathbf{E}^{\mathrm{p}}_{l_{n_j},L_{n_j}}$$

[illegible] $t\in[0,T]$ [illegible]

$$u_j(t)\overset{*}{\rightharpoonup}u(t) \qquad \text{[illegible] weakly* in } BD(\Omega), \tag{9.17}$$

$$\mathbf{E}^{\mathrm{e}}_j(t)\to\mathbf{E}^{\mathrm{e}}(t) \qquad \text{[illegible] } L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}}), \tag{9.18}$$

[illegible]

$$\mathbf{E}^{\mathrm{p}}_j(t)\overset{*}{\rightharpoonup}\mathbf{E}^{\mathrm{p}}(t) \qquad \text{[illegible] weakly* in } \mathcal{M}_b(\Omega\cup\partial_D\Omega;\mathrm{M}_D^{N\times N}). \tag{9.19}$$

$t \in [0,T]$

$$\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}_j(t)) \to \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(t)) \qquad \mathcal{Q}_2^{n_j}(\mathrm{curl}\mathbf{E}^{\mathrm{p}}_j(t)) \to 0, \tag{9.20}$$

. We divide the proof into three steps.

1: By Lemma 9.1 there exist a subsequence n_j,

$$u \in AC(0,T;BD(\Omega)), \qquad \mathbf{E}^{\mathrm{e}} \in AC(0,T;L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}})),$$

and

$$\mathbf{E}^{\mathrm{p}} \in AC(0,T;\mathcal{M}_b(\Omega\cup\partial_D\Omega;\mathrm{M}^{N\times N}_D))$$

such that by setting $u_j := u_{n_j}$, $\mathbf{E}^{\mathrm{e}}_j := \mathbf{E}^{\mathrm{e}}_{n_j}$, and $\mathbf{E}^{\mathrm{p}}_j := \mathbf{E}^{\mathrm{p}}_{n_j}$, for every $t \in [0,T]$ relations (9.17) and (9.19) hold,

$$\mathbf{E}^{\mathrm{e}}_j(t) \rightharpoonup \mathbf{E}^{\mathrm{e}}(t) \qquad \text{weakly in } L^2(\Omega;\mathrm{M}^{N\times N}_{\mathrm{sym}}), \tag{9.21}$$

and $(u(t),\mathbf{E}^{\mathrm{e}}(t),\mathbf{E}^{\mathrm{p}}(t)) \in \mathcal{A}_{\mathrm{pp}}(w(t))$, so that the triple $(u(t),\mathbf{E}^{\mathrm{e}}(t),\mathbf{E}^{\mathrm{p}}(t))$ is admissible. Finally, from the energy balance (5.3), and by the assumptions for $t=0$, we deduce that for every $t \in [0,T]$

$$\mathcal{Q}_2^{n_j}(\mathrm{curl}\mathbf{E}^{\mathrm{p}}_j(t)) \le C \tag{9.22}$$

for some constant C independent of j and t.

2: Let us fix $t \in [0,T]$. In order to prove that $(u(t),\mathbf{E}^{\mathrm{e}}(t),\mathbf{E}^{\mathrm{p}}(t)) \in \mathcal{A}_{\mathrm{pp}}(w(t))$ satisfies the global stability condition

$$\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(t)) - \langle \mathcal{L}(t), u(t)\rangle \le \mathcal{Q}_1(\mathbf{e}) - \langle \mathcal{L}(t), v\rangle + \mathcal{H}_{\mathrm{pp}}(\mathbf{p}-\mathbf{E}^{\mathrm{p}}(t)) \tag{9.23}$$

for every $(v,\mathbf{e},\mathbf{p}) \in \mathcal{A}_{\mathrm{pp}}(w(t))$, in view of [8, Theorem 3.6] it suffices to prove that the Cauchy stress $\mathbf{T}(t) = \mathbb{C}\mathbf{E}^{\mathrm{e}}(t)$ satisfies the equilibrium conditions

$$\begin{cases} -\mathrm{div}\mathbf{T}(t) = f(t) & \text{in } \Omega,\\ \mathbf{T}(t)\nu = g(t) & \text{on } \partial_N\Omega \end{cases} \tag{9.24}$$

and the constraint

$$|\mathbf{T}_D(t,x)| \le S_Y \qquad \text{for a.e. } x \in \Omega, \tag{9.25}$$

where $\mathbf{T}_D(t) := (\mathbf{T}(t))_D$.

Equation (9.24) follows from the equilibrium equation for the Cauchy stress $\mathbf{T}_j(t) = \mathbb{C}\mathbf{E}^{\mathrm{e}}_j(t)$ given by (8.6) in view of the weak convergence of $\mathbf{T}_j(t)$ to $\mathbf{T}(t)$ which comes from (9.21).

In order to prove (9.25), let us consider the corresponding constraint in the strain gradient context given by (8.9). Let $\mathbb{K}^{\mathrm{p}}_{\mathrm{diss,j}}(t)$, $\mathbb{K}^{\mathrm{p}}_{\mathrm{en},j}(t)$, $\mathbb{K}^{\mathrm{p}}_j(t) = \mathbb{K}^{\mathrm{p}}_{\mathrm{diss},j}(t) + \mathbb{K}^{\mathrm{p}}_{\mathrm{en},j}(t)$ and $\mathbf{T}^{\mathrm{p}}_j(t)$ be the stresses associated to $(u_j(t),\mathbf{E}^{\mathrm{e}}_j(t),\mathbf{E}^{\mathrm{p}}_j(t))$. Notice that, in view of (3.5) and of (9.22), we get

$$\mathbb{K}^{\mathrm{p}}_{\mathrm{en},j}(t) \to 0 \qquad \text{strongly in } L^2(\Omega;\mathrm{M}^{N\times N\times N}_D). \tag{9.26}$$

Moreover by (8.8) and (8.9) we have

$$\mathbf{T}_j^{\mathrm{p}}(t) = (\mathbf{T}_j(t))_D + \mathrm{div}\mathbb{K}_j^{\mathrm{p}}(t), \tag{9.27}$$

and

$$\sqrt{|\mathbf{T}_j^{\mathrm{p}}(t,x)|^2 + l_{n_j}^{-2}|\mathbb{K}_{\mathrm{diss},j}^{\mathrm{p}}(t,x)|^2} \le S_Y \qquad \text{for a.e. } x \in \Omega.$$

In particular we have that $(\mathbf{T}_j^{\mathrm{p}}(t))_{j\in\mathbb{N}}$ is uniformly bounded in $L^\infty(\Omega;\mathrm{M}_D^{N\times N})$ and

$$\mathbb{K}_{\mathrm{diss},j}^{\mathrm{p}}(t) \to 0 \qquad \text{strongly in } L^\infty(\Omega;\mathrm{M}_D^{N\times N\times N}).$$

By (9.26) we conclude that $\mathbb{K}_j^{\mathrm{p}}(t) \to 0$ strongly in $L^2(\Omega;\mathrm{M}_D^{N\times N\times N})$. Notice that from (9.27) we deduce that $\mathrm{div}\mathbb{K}_j^{\mathrm{p}}(t)$ is bounded in $L^2(\Omega;\mathrm{M}_D^{N\times N})$. We obtain

$$\mathrm{div}\mathbb{K}_j^{\mathrm{p}}(t) \rightharpoonup 0 \qquad \text{weakly in } L^2(\Omega;\mathrm{M}_D^{N\times N})$$

so that in view of (9.27) and (9.21)

$$\mathbf{T}_j^{\mathrm{p}}(t) \rightharpoonup \mathbf{T}_D(t) \qquad \text{weakly in } L^2(\Omega;\mathrm{M}_D^{N\times N}). \tag{9.28}$$

Since $\mathbf{T}_j^{\mathrm{p}}(t) \in \mathcal{K} := \{\mathbf{A} \in L^2(\Omega;\mathrm{M}_D^{N\times N}) : |\mathbf{A}| \le S_Y \text{ a.e. in } \Omega\}$, and $\mathcal{K}$ is closed in the weak topology of $L^2(\Omega;\mathrm{M}_D^{N\times N})$, by (9.28) we deduce that (9.25) holds. Hence (9.23) follows, and Step 2 is concluded.

Step 3: [illegible] Since $t \mapsto u_j(t)$ is absolutely continuous from $[0,T]$ to $W^{1,\frac{N}{N-1}}(\Omega;\mathbb{R}^N)$, by integrating by parts we can write the energy balance (5.3) in the following form:

$$\begin{aligned}\mathcal{Q}_1(\mathbf{E}_j^{\mathrm{e}}(t)) + \mathcal{Q}_2^{n_j}(\mathrm{curl}\mathbf{E}_j^{\mathrm{p}}(t)) + \mathcal{D}_{\mathcal{H}_{n_j}}(\mathbf{E}_j^{\mathrm{p}};0,t) - \int_0^t \langle \mathcal{L}(\tau), \dot{u}_j(\tau)\rangle\, d\tau \\ = \mathcal{Q}_1(\mathbf{E}_j^{\mathrm{e}}(0)) + \mathcal{Q}_2^{n_j}(\mathrm{curl}\mathbf{E}_j^{\mathrm{p}}(0)) + \int_0^t \int_\Omega \mathbf{T}_j(\tau) : \mathbf{E}\dot{w}(\tau)\, dx\, d\tau \\ - \int_0^t \langle \mathcal{L}(\tau), \dot{w}(\tau)\rangle\, d\tau.\end{aligned} \tag{9.29}$$

We claim that for every $t \in [0,T]$

$$\begin{aligned}\liminf_{j\to+\infty}\left[\mathcal{D}_{\mathcal{H}_{n_j}}(\mathbf{E}_j^{\mathrm{p}};0,t) - \int_0^t \langle \mathcal{L}(\tau), \dot{u}_j(\tau)\rangle\, d\tau\right] \ge \mathcal{D}_{\mathcal{H}_{\mathrm{pp}}}(\mathbf{E}^{\mathrm{p}};0,t) - \int_0^t \langle \mathcal{L}(\tau), \dot{u}(\tau)\rangle\, d\tau \\ = \mathcal{D}_{\mathcal{H}_{\mathrm{pp}}}(\mathbf{E}^{\mathrm{p}};0,t) - \langle \mathcal{L}(t), u(t)\rangle + \langle \mathcal{L}(0), u(0)\rangle + \int_0^t \langle \dot{\mathcal{L}}(\tau), u(\tau)\rangle\, d\tau.\end{aligned} \tag{9.30}$$

By passing to the limit in (9.29), by (9.21), (9.9), and (9.30) we get for every $t \in [0,T]$

$$\begin{aligned}\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(t)) - \langle \mathcal{L}(t), u(t)\rangle + \mathcal{D}_{\mathcal{H}_{\mathrm{pp}}}(\mathbf{E}^{\mathrm{p}};0,t) \le \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(0)) - \langle \mathcal{L}(0), u(0)\rangle \\ + \int_0^t \int_\Omega \mathbf{T}(\tau) : \mathbf{E}\dot{w}(\tau)\, dx\, d\tau - \int_0^t \langle \dot{\mathcal{L}}(\tau), u(\tau)\rangle\, d\tau - \int_0^t \langle \mathcal{L}(\tau), \dot{w}(\tau)\rangle\, d\tau.\end{aligned}$$

In view of the global stability condition (9.23), by [8, Theorem 4.7] we have that also the opposite inequality holds, so that the energy balance follows. From the previous steps, we conclude that $t \mapsto (u(t), \mathbf{E}^{\mathrm{e}}(t), \mathbf{E}^{\mathrm{p}}(t))$ is a quasi-static evolution according to Dal Maso, DeSimone, and Mora [8].

By (9.29), (9.30), and the energy balance, we deduce that for every $t \in [0,T]$ we have

$$\mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}_j(t)) \to \mathcal{Q}_1(\mathbf{E}^{\mathrm{e}}(t)) \qquad \text{and} \qquad \mathcal{Q}_2^{n_j}(\mathrm{curl}\mathbf{E}^{\mathrm{p}}_j(t)) \to 0$$

so that (9.20) holds. In view of (9.21), we conclude that (9.18) follows.

Let us prove claim (9.30). Recall that by [8, Theorem 7.1] we have the following representation of the dissipation:

$$\mathcal{D}_{\mathcal{H}_{n_j}}(\mathbf{E}^{\mathrm{p}}_j; 0, t) = \int_0^t \mathcal{H}_{n_j}(\dot{\mathbf{E}}^{\mathrm{p}}_j(\tau))\, d\tau.$$

From the representation (4.15) we get

(9.31)
$$\mathcal{D}_{\mathcal{H}_{n_j}}(\mathbf{E}^{\mathrm{p}}_j; 0, t) - \int_0^t \langle \mathcal{L}(\tau), \dot{u}_j(\tau)\rangle\, d\tau = \int_0^t \left[\mathcal{H}_{n_j}(\dot{\mathbf{E}}^{\mathrm{p}}_j(\tau)) - \int_\Omega \rho_D(\tau) : \dot{\mathbf{E}}^{\mathrm{p}}_j(\tau)\, dx\right] d\tau$$
$$- \int_0^t \int_\Omega \rho(\tau) : \dot{\mathbf{E}}^{\mathrm{e}}_j(\tau)\, dx\, d\tau + \int_0^t \langle \rho(\tau)\nu, \dot{w}(\tau)\rangle_{\partial_D\Omega}\, d\tau.$$

Moreover for every $0 \le \tau \le t$

$$\mathcal{H}_{n_j}(\dot{\mathbf{E}}^{\mathrm{p}}_j(\tau)) - \int_\Omega \rho_D(\tau) : \dot{\mathbf{E}}^{\mathrm{p}}_j(\tau)\, dx \ge \int_\Omega \left[S_Y|\dot{\mathbf{E}}^{\mathrm{p}}_j(\tau)| - \rho_D(\tau) : \dot{\mathbf{E}}^{\mathrm{p}}_j(\tau)\right] dx,$$

and the integrand of the right-hand side is positive in view of the safe load condition (4.14). Let $\varphi \in C^1(\overline{\Omega})$, with $0 \le \varphi \le 1$ and $\varphi = 0$ near $\overline{\partial_N\Omega}$. By applying again the representation result [8, Theorem 7.1] for the dissipation $\mathcal{D}_{\mathcal{H}_{\mathrm{pp}}}$ we conclude that

(9.32)
$$\begin{aligned}
&\liminf_{j\to+\infty} \int_0^t \left[\mathcal{H}_{n_j}(\dot{\mathbf{E}}^{\mathrm{p}}_j(\tau)) - \int_\Omega \rho_D(\tau) : \dot{\mathbf{E}}^{\mathrm{p}}_j(\tau)\, dx\right] \\
&\qquad \ge \liminf_{j\to+\infty} \int_0^t \int_\Omega \left[S_Y|\dot{\mathbf{E}}^{\mathrm{p}}_j(\tau)| - \rho_D(\tau) : \dot{\mathbf{E}}^{\mathrm{p}}_j(\tau)\right] dx\, d\tau \\
&\qquad \ge \liminf_{j\to+\infty} \int_0^t \int_\Omega \left[S_Y|\varphi\dot{\mathbf{E}}^{\mathrm{p}}_j(\tau)| - \rho_D(\tau) : \varphi\dot{\mathbf{E}}^{\mathrm{p}}_j(\tau)\right] dx\, d\tau \\
&\qquad\qquad = \liminf_{j\to+\infty} \left[\mathcal{D}_{\mathcal{H}_{\mathrm{pp}}}(\varphi\mathbf{E}^{\mathrm{p}}_j; 0, t) - \int_0^t \int_\Omega \rho_D(\tau) : \varphi\dot{\mathbf{E}}^{\mathrm{p}}_j(\tau)\, dx\, d\tau\right].
\end{aligned}$$

By the very definition of $\mathcal{D}_{\mathcal{H}_{\mathrm{pp}}}$ and by (9.19), it is easy to see that

$$\liminf_{j\to+\infty} \mathcal{D}_{\mathcal{H}_{\mathrm{pp}}}(\varphi\mathbf{E}^{\mathrm{p}}_j; 0, t) \ge \mathcal{D}_{\mathcal{H}_{\mathrm{pp}}}(\varphi\mathbf{E}^{\mathrm{p}}; 0, t). \tag{9.33}$$

On the other hand, the absolute continuity of $t \mapsto \mathbf{E}^{\mathrm{p}}_j(t)$ implies that

(9.34)
$$\int_0^t \int_\Omega \rho_D(\tau) : \varphi\dot{\mathbf{E}}^{\mathrm{p}}_j(\tau)\, dx\, d\tau = \int_\Omega \rho_D(t) : \varphi\mathbf{E}^{\mathrm{p}}_j(t)\, dx - \int_\Omega \rho_D(0) : \varphi\mathbf{E}^{\mathrm{p}}_j(0)\, dx$$
$$- \int_0^t \int_\Omega \dot{\rho}_D(\tau) : \varphi\mathbf{E}^{\mathrm{p}}_j(\tau)\, dx\, d\tau.$$

By integrating by parts, for a.e. $\tau \in [0,t]$ we have

$$\int_\Omega \dot\rho_D(\tau) : \varphi \mathbf{E}^{\mathrm{p}}_j(\tau)\,dx = \langle \dot{\mathcal{L}}(\tau), \varphi u_j(\tau)\rangle + \langle \dot\rho(\tau)\nu, w(\tau)\rangle_{\partial_D\Omega} - \int_\Omega \dot\rho(\tau) : \varphi \mathbf{E}^{\mathrm{e}}_j(\tau)\,dx - \int_\Omega \dot\rho(\tau) : [\nabla\varphi \odot u_j(\tau)]\,dx.$$

In view of the embedding result [22, Proposition 2.5], we get $\dot\rho(\tau) \in L^N(\Omega; \mathrm{M}^{N\times N}_{\mathrm{sym}})$ for a.e. $\tau \in [0,t]$. By (9.21) and (9.17), and since $\varphi = 0$ near $\partial_N\Omega$, we deduce for a.e. $\tau \in [0,t]$ that

$$\lim_{j\to+\infty} \int_\Omega \dot\rho_D(\tau) : \varphi \mathbf{E}^{\mathrm{p}}_j(\tau)\,dx = \langle \dot{\mathcal{L}}(\tau), \varphi u(\tau)\rangle + \langle \dot\rho(\tau)\nu, w(\tau)\rangle_{\partial_D\Omega} - \int_\Omega \dot\rho(\tau) : \varphi \mathbf{E}^{\mathrm{e}}(\tau)\,dx - \int_\Omega \dot\rho(\tau) : [\nabla\varphi \odot u(\tau)]\,dx = \int_{\Omega\cup\partial_D\Omega} \varphi\, d[\dot\rho_D(\tau) : \mathbf{E}^{\mathrm{p}}(\tau)], \tag{9.35}$$

where $[\dot\rho_D(\tau) : \mathbf{E}^{\mathrm{p}}(\tau)]$ is the measure defined in the previous subsection, and the last equality follows by (9.5) (with $\dot\rho_D$ in place of ρ_D). Similarly we obtain

$$\lim_{j\to+\infty} \int_\Omega \rho_D(t) : \varphi \mathbf{E}^{\mathrm{p}}_j(t)\,dx = \int_{\Omega\cup\partial_D\Omega} \varphi\, d[\rho_D(t) : \mathbf{E}^{\mathrm{p}}(t)] \tag{9.36}$$

and

$$\lim_{j\to+\infty} \int_\Omega \rho_D(0) : \varphi \mathbf{E}^{\mathrm{p}}_j(0)\,dx = \int_{\Omega\cup\partial_D\Omega} \varphi\, d[\rho_D(0) : \mathbf{E}^{\mathrm{p}}(0)]. \tag{9.37}$$

By letting $\varphi \to 1_{\Omega\cup\partial_D\Omega}$ we obtain from (9.32), (9.33), (9.34), and (9.35)–(9.37)

$$\liminf_{j\to+\infty} \int_0^t \left[\mathcal{H}_{n_j}(\dot{\mathbf{E}}^{\mathrm{p}}_j(\tau)) - \int_\Omega \rho_D(\tau) : \dot{\mathbf{E}}^{\mathrm{p}}_j(\tau)\,dx\right] \ge \mathcal{D}_{\mathcal{H}_{\mathrm{pp}}}(\mathbf{E}^{\mathrm{p}}; 0, t) - [\rho_D(t) : \mathbf{E}^{\mathrm{p}}(t)](\Omega) + [\rho_D(0) : \mathbf{E}^{\mathrm{p}}(0)](\Omega\cup\partial_D\Omega) + \int_0^t [\dot\rho_D(\tau) : \mathbf{E}^{\mathrm{p}}(\tau)](\Omega\cup\partial_D\Omega)\,d\tau = \mathcal{D}_{\mathcal{H}_{\mathrm{pp}}}(\mathbf{E}^{\mathrm{p}}; 0, t) - \int_0^t [\rho_D(\tau) : \dot{\mathbf{E}}^{\mathrm{p}}(\tau)](\Omega\cup\partial_D\Omega)\,d\tau. \tag{9.38}$$

In conclusion, by passing to the limit in (9.31), by (9.38) and (9.21) we get

$$\liminf_{j\to+\infty} \left[\mathcal{D}_{\mathcal{H}_{n_j}}(\mathbf{E}^{\mathrm{p}}_j; 0, t) - \int_0^t \langle \mathcal{L}(\tau), \dot u_j(\tau)\rangle\,d\tau\right] \ge \mathcal{D}_{\mathcal{H}_{\mathrm{pp}}}(\mathbf{E}^{\mathrm{p}}; 0, t) - \int_0^t [\rho_D(\tau) : \dot{\mathbf{E}}^{\mathrm{p}}(\tau)](\Omega\cup\partial_D\Omega)\,d\tau - \int_0^t \int_\Omega \rho(\tau) : \dot{\mathbf{E}}^{\mathrm{e}}(\tau)\,dx\,d\tau + \int_0^t \langle \rho(\tau)\nu, \dot w(\tau)\rangle_{\partial_D\Omega}\,d\tau = \mathcal{D}_{\mathcal{H}_{\mathrm{pp}}}(\mathbf{E}^{\mathrm{p}}; 0, t) - \int_0^t \langle \mathcal{L}(\tau), \dot u(\tau)\rangle\,d\tau,$$

where the last equality comes from the integration by parts (9.4). We deduce that claim (9.30) holds, and the proof is concluded. ∎

9.3. It is interesting to compare our result with the abstract convergence result recently proposed by Mielke, Roubíček, and Stefanelli [26] concerning the asymptotic behavior of quasi-static evolutions with Γ-converging energies and dissipations. Their framework cannot be easily adapted to our context since the total energy and the dissipation functional do not satisfy trivially a sort of Γ-liminf inequality. This is due to the fact that we skip from a BV to a Radon-measure setting for the plastic strain, so that concentration at the boundary $\partial_D\Omega$ can occur, and this has to be taken into account for the dissipation $\mathcal{H}_{\rm pp}$ and the work of the external loads. The safe load condition is the key point to rearrange the terms in order to work out suitable lower semicontinuity inequalities (see the arguments of Step 3) which are essential for the study of the asymptotics of the problem.

REFERENCES

[1] A. Acharya and J. L. Bassani, *Incompatibility and crystal plasticity*, J. Mech. Phys. Solids, 48 (2000), pp. 1565–1595.

[2] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variations and Free Discontinuity Problems*, Clarendon Press, Oxford, 2000.

[3] E. C. Aifantis, *On the microstructural origin of certain inelastic models*, ASME J. Eng. Mater. Technol., 106 (1984) pp. 326–330.

[4] M. F. Ashby, *The deformation of plastically non-homogeneous alloys*, Philos. Mag., 21 (1970), pp. 399–424.

[5] M. F. Ashby, *The deformation of plastically non-homogeneous alloys*, in Strengthening Methods in Crystals, A. Kelly and R. B. Nicholson, eds., Elsevier, Amsterdam, 1971, pp. 137–192.

[6] H. Brezis, *Operateurs Maximaux Monotones et Semi-Groups de Contractions dans les Espaces de Hilbert*, North-Holland, Amsterdam-London; Elsevier, New York, 1973.

[7] P. Cermelli and M. E. Gurtin, *On the characterization of geometrically necessary dislocations in finite plasticity*, J. Mech. Phys. Solids, 49 (2000), pp. 1539–1568.

[8] G. Dal Maso, A. DeSimone, and M. G. Mora, *Quasistatic evolution problems for linearly elastic-perfectly plastic materials*, Arch. Ration. Mech. Anal., 180 (2006), pp. 237–291.

[9] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, Stud. Adv. Math., CRC Press, Boca Raton, FL, 1992.

[10] N. A. Fleck and J. W. Hutchinson, *Strain gradient plasticity*, Adv. Appl. Mech., 33 (1997), pp. 295–361.

[11] N. A. Fleck and J. W. Hutchinson, *A reformulation of strain gradient plasticity*, J. Mech. Phys. Solids, 49 (2001), pp. 2245–2271.

[12] H. Gao, Y. Huang, W. D. Nix, and J. W. Hutchinson, *Mechanism-based strain gradient plasticity*, I. *Theory*, J. Mech. Phys. Solids, 47 (1999), pp. 1239–1263.

[13] C. Goffman and J. Serrin, *Sublinear functions of measures and variational integrals*, Duke Math. J., 31 (1964), pp. 159–178.

[14] P. Gudmundson, *A unified treatment of strain gradient plasticity*, J. Mech. Phys. Solids, 52 (2004), pp. 1379–1406.

[15] M. E. Gurtin, *On the plasticity of single crystals: Free energy, microforces, plastic strain gradients*, J. Mech. Phys. Solids, 48 (2000), pp. 989–1036.

[16] M. E. Gurtin, *A gradient theory of single-crystal viscoplasticity that accounts for geometrically necessary dislocations*, J. Mech. Phys. Solids, 50 (2002), pp. 5–32.

[17] M. E. Gurtin, *On a framework for small-deformation viscoplasticity: Free energy, microforces, strain gradients*, Int. J. Plasticity, 19 (2003), pp. 47–90.

[18] M. E. Gurtin, *A gradient theory of small-deformation isotropic plasticity that accounts for the Burgers vector and for dissipation due to plastic spin*, J. Mech. Phys. Solids, 52 (2004), pp. 2545–2568.

[19] M. E. Gurtin and L. Anand, *A theory of strain-gradient plasticity for isotropic, plastically irrotational materials*, I. *Small deformations*, J. Mech. Phys. Solids, 53 (2005), pp. 1624–1649.

[20] M. E. Gurtin and A. Needleman, *Boundary conditions in small-deformation, single-crystal plasticity that account for the Burgers vector*, J. Mech. Phys. Solids, 53 (2005), pp. 1–31.

[21] Y. Huang, H. Gao, W. D. Nix, and J. W. Hutchinson, *Mechanism-based strain gradient plasticity*-II, *Analysis*, J. Mech. Phys. Solids, 48 (2000), pp. 99–128.

[22] R. Kohn and R. Temam, *Dual spaces of stresses and strains, with applications to Hencky plasticity*, Appl. Math. Optim., 10 (1983), pp. 1–35.
[23] A. Mainik and A. Mielke, *Existence results for energetic models for rate-independent systems*, Calc. Var. Partial Differential Equations, 22 (2005), pp. 73–99.
[24] A. Mielke, *Evolution of rate-independent systems*, in Evolutionary Equations, Handb. Differ. Equ. II, Elsevier/North-Holland, Amsterdam, 2005, pp. 461–559.
[25] A. Mielke, *Analysis of energetic models for rate-independent materials*, in Proceedings of the International Congress of Mathematicians, Vol. III (Beijing, 2002), Higher Ed. Press, Beijing, 2002, pp. 817–828.
[26] A. Mielke, T. Roubíček and U. Stefanelli, Γ*-limits and relaxations for rate-independent evolutionary problems*, Calc. Var. Partial Differential Equations, 31 (2008), pp. 387–416.
[27] A. Mielke and F. Theil, *A mathematical model for rate independent phase transformations with hysteresis*, in Proceedings of the Workshop on "Models of Continuum Mechanics in Analysis and Engineering," H.-D. Alber, R. Balean, and R. Farwig, eds., Shaker-Verlag, Aachen, 1999, pp. 117–129.
[28] A. Mielke and F. Theil, *On rate-independent hysteresis models*, NoDEA Nonlinear Differential Equations Appl., 11 (2004), pp. 151–189.
[29] A. Mielke, F. Theil, and V. Levitas, *A variational formulation of rate-independent phase transformations using an extremum principle*, Arch. Ration. Mech. Anal., 162 (2002), pp. 137–177.
[30] J. F. Nye, *Some geometrical relations in dislocated crystals*, Acta Metall., 1 (1953), pp. 153–162.
[31] B. D. Reddy, F. Ebobisse, and A. McBride, *Well-posedness of a model of strain gradient plasticity for plastically irrotational materials*, Int. J. Plasticity, 24 (2008), pp. 55–73.
[32] P.-M. Suquet, *Sur les équations de la plasticité: Existence et règularité des solutions*, *[On the equations of plasticity: Existence and regularity of solutions]*, J. Mécanique, 20 (1981), pp. 3–39.
[33] R. Temam, *Problèmes Mathématiques en Plasticité*, Méthodes Mathématiques de l'Informatique 12, Gauthier-Villars, Montrouge, 1983.

SIAM J. MATH. ANAL.
Vol. 40, No. 3, pp. 1246–1271

ON THE NEW MULTISCALE RODLIKE MODEL OF POLYMERIC FLUIDS*

HUI ZHANG† AND PINGWEN ZHANG‡

Abstract. This paper is concerned with the well-posedness for the new rigid rodlike model in a polymeric fluid recently proposed by W.N. E and P.W. Zhang [*Meth. Appl. Anal.*, 13 (2006), pp. 181–198]. The constitutive relations considered in this work are motivated by the kinetic theory. The micro equations involve five independent spatial variables (degrees of freedom): two in the configuration domain and three in the macro flow domain. We obtain the local existence of solutions with large initial data and also global existence of solutions with small Deborah and Reynolds constants in periodic domains.

Key words. polymeric fluid, rodlike model, kinetic theory, global existence

AMS subject classifications. 76B03, 65M12, 35Q35

DOI. 10.1137/050640795

1. Introduction.

The Doi kinetic theory for spatially homogeneous flow of rodlike molecules has successfully described the properties of liquid crystal polymers in a solvent [6]. One of the simplest models of polymeric fluids described by the Doi theory is the rigid rodlike model, which takes into account the macro and micro behavior of the dilute or solute polymers—the effects of flow, Brownian motion, and intermolecular forces on the molecular orientation distribution (see [6, 10, 11, 16, 17, 23]). However, it does not include the so-called distortional elasticity. The Doi theory is valid only in the limit of spatial homogeneity. For small molecule liquid crystals, distortional elasticity has been formulated in the limit of weak distribution as Frank elasticity. This is one ingredient of the Leslie–Ericken theory. Several attempts have been made to find a theory which encompasses both the molecular visco-elasticity and the distortional elasticity. Marrucci and Greco [18] made a molecular theory of distortional elasticity. They proposed a nonlocal mean field nematic potential for LCPs (liquid crystal polymers), which accounts for spatial variations of the molecular orientation distribution. Tsuji and Rey [21, 22] add distortional elasticity to the rodlike model of the Doi theory but did not give a stress tensor. Edwards and Beris [2] give an ad hoc generalization of the Frank elasticity in tensorial form. Ericksen [9] allowed the order parameter to be a variable but still required the orientation distribution to be uniaxial. An extension of Kuzuu and Doi [13] theory to flowing systems of nonhomogeneous liquid crystalline polymers is made by Wang [25], in which the author models the LCP molecules as spheroids of equal shape and size. He derives an intermolecular potential which could be considered as an extension of Marrucci–Greco potential.

*Received by the editors September 21, 2005; accepted for publication (in revised form) July 15, 2008; published electronically October 29, 2008.
http://www.siam.org/journals/sima/40-3/64079.html

†Laboratory of Mathematics and Complex Systems, Ministry of Education, School of Mathematical Sciences, Beijing Normal University, Beijing, 100875, P.R. China (hzhang@bnu.edu.cn). This author is partially supported by an Alexande von Humbolt Fellowship and is also partially supported by the key basic research project of the Ministry of Chinese Education 107016 and the state key basic research project of China 2005CB321704.

‡LMAM and School of Mathematical Sciences, Peking University, Beijing, 100871, P.R. China (pzhang@pku.edu.cn). This author's research is partially supported by the state key basic research project of China 2005CB321704 and the National Science Foundation of China for Distinguished Young Scholars 10225103.

All these approaches are phonomenological in nature. They invariably contain a large number of unknown parameters which in general cannot be determined rationally. Another drawback of the phonomenological theories is the lack of consistency with existing theories and among themselves. Then E and Zhang [8] developed a new model for nonhomogeneous flows of liquid crystalline polymers with a few adjustable parameters that could model a variety of configurations of polymeric liquid crystal molecules. This new model is a combination of macroscopic partial differential equations and microscopic Fokker–Planck equations. In this model, the function $\psi(\mathbf{x}, \mathbf{m}, t)$ describes the distribution of an identical rigid rodlike molecule at $(\mathbf{x}, t)$ with the orientation $\mathbf{m}$. Denoting the velocity and pressure of the fluid by $\mathbf{u}$ and p, the new multiscale rodlike model can be expressed as

$$\begin{aligned}\frac{\partial \psi}{\partial t} + \nabla \cdot (\mathbf{u}\psi) &= \frac{1}{k_B T}\nabla \cdot \big\{[D_{\|}\mathbf{mm} + D_{\perp}(\mathbf{I} - \mathbf{mm})] \cdot (\psi \nabla \mu)\big\} \\ &\quad + \frac{D_r}{k_B T}\mathcal{R} \cdot (\psi \mathcal{R}\mu) - \mathcal{R} \cdot (\mathbf{m} \times \kappa \cdot \mathbf{m}\psi), \quad \mathbf{m} \in \mathbb{S}^2,\end{aligned} \tag{1.1}$$

where k_B is Boltzmann constant and T is the absolute temperature, $D_{\|} \geq 0$ and $D_{\perp} \geq 0$ are translational diffusion coefficients parallel and normal to the orientation of the LCP molecule, $D_r = \frac{\xi_r}{k_B T}$ is the rotary diffusivity and ξ_r is the friction coefficient, ∇ is the gradient operator with respect to the spatial variables $\mathbf{x}$, $\mathcal{R} = \mathbf{m} \times \frac{\partial}{\partial \mathbf{m}}$ is the rotational gradient operator, and $\mathbb{S}^2$ is the unit sphere in $\mathbb{R}^3$. The symbol μ denotes the chemical potential

$$\mu = \ln \psi + \bar{U}, \tag{1.2}$$

and $\bar{U}$ represents the excluded-volume potential [6, 11]

$$\bar{U}(\mathbf{x}, \mathbf{m}, t) = k_B T \alpha \int_{\Omega} \int_{|\mathbf{m}'|=1} B(\mathbf{x}, \mathbf{x}', \mathbf{m}, \mathbf{m}')\psi(\mathbf{x}', \mathbf{m}', t) d\mathbf{m}' d\mathbf{x}'. \tag{1.3}$$

The function B in (1.3) is the interactional factor among rods. Here α denotes the intensity between particles. Now we choose $B(\mathbf{x}, \mathbf{x}', \mathbf{m}, \mathbf{m}') = \frac{1}{\varepsilon^3}\chi(\frac{\mathbf{x}-\mathbf{x}'}{\varepsilon})|\mathbf{m} \times \mathbf{m}'|^2$, where $\chi(\mathbf{x})$ is the smooth kernel; e.g., $\chi(\mathbf{x}) = C \exp(1/(|\mathbf{x}|^2 - 1))$ as $|\mathbf{x}| < 1$, and $\chi(\mathbf{x}) = 0$ as $|\mathbf{x}| \geq 1$, where C is a constant such that $\int_{|\mathbf{x}|\leq 1} \chi(\mathbf{x}) d\mathbf{x} = 1$. $\kappa = (\nabla \mathbf{u})^T$ is the velocity gradient tensor.

Let L_0 be the typical size of the flow region, V_0 be the typical velocity scale, and $T_0 = \frac{L_0}{V_0}$ be a typical convective time scale. Further De is an important parameter called the Deborah number:

$$De = \frac{\frac{\xi_r}{k_B T}}{\frac{L_0}{V_0}} = \frac{\xi_r V_0}{k_B T L_0}. \tag{1.4}$$

It is the ratio of the orientational diffusion time scale of the rods (which is the relevant relaxation time scale) and the convective time scale of the fluid. Set

$$\varepsilon = \frac{L}{L_0}, \tag{1.5}$$

where L is the length of the rods. Thus the nondimensional kinetic equation is

$$\frac{\partial \psi}{\partial t} + \nabla \cdot (\mathbf{u}\psi) = \frac{\varepsilon^2}{De}\nabla \cdot \Big\{ [D_{\|}^{*}\mathbf{mm} + D_{\perp}^{*}(\mathbf{I} - \mathbf{mm})] \cdot (\psi \nabla \mu) \Big\}$$

$$+ \frac{1}{De}\mathcal{R} \cdot (\psi \mathcal{R}\mu) - \mathcal{R} \cdot (\mathbf{m} \times \kappa \cdot \mathbf{m}\psi), \quad \mathbf{m} \in \mathbb{S}^2, \tag{1.6}$$

$$\mu = \ln \psi + U, \tag{1.7}$$

$$U(\mathbf{x}, \mathbf{m}, t) = \alpha \int_{\Omega} \int_{|\mathbf{m}'|=1} B(\mathbf{x}, \mathbf{x}', \mathbf{m}, \mathbf{m}')\psi(\mathbf{x}', \mathbf{m}', t) d\mathbf{m}' d\mathbf{x}'. \tag{1.8}$$

The velocity field satisfies the Navier–Stokes-like equation

$$\rho(\mathbf{u}_t + (\mathbf{u} \cdot \nabla)\mathbf{u}) + \nabla p = \nabla \cdot \tau + F, \tag{1.9}$$

$$\nabla \cdot \mathbf{u} = 0. \tag{1.10}$$

In the LCP system, the extra stress τ is given by two parts, the viscous stress τ_s and the elastic stress τ_e, namely

$$\tau = \tau_s + \tau_e. \tag{1.11}$$

The viscous stress comes from two sources, one from the solvent and the other from the constrain of rods derived in [6],

$$\tau_s = 2\eta_s \mathbf{D} + \frac{1}{2}\xi_r \mathbf{D} : \langle \mathbf{mmmm} \rangle,$$

where $\mathbf{D} = \frac{1}{2}(\kappa + \kappa^T) = \frac{1}{2}(\nabla \mathbf{u} + (\nabla \mathbf{u})^T)$ is the strain rate tensor, η_s is the solvent viscosity, and $\langle (\cdot) \rangle$ denotes averaging with respect to the distribution ψ; i.e., $\langle (g) \rangle = \int_{|\mathbf{m}|=1} g\psi d\mathbf{m}$. The elastic stress is derived through a generalized virtual work principle [6]. The detail can be found in [8]. Now we cite the result from [8],

$$\tau_e = -\langle (\mathbf{m} \times \mathcal{R}\mu)\mathbf{m} \rangle. \tag{1.12}$$

Meanwhile the body force can also be identified as

$$\mathbf{F} = -\langle \nabla \mu \rangle. \tag{1.13}$$

Now let $\eta_p = \xi_r, \eta = \eta_s + \eta_p, \gamma = \frac{\eta_s}{\eta}$, and Re denotes the Reynolds number. Then the nondimensional Navier–Stokes-like equation

$$\mathbf{u}_t + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p = \frac{\gamma}{Re}\triangle \mathbf{u} + \frac{1-\gamma}{2Re}\nabla \cdot (\mathbf{D} : \langle \mathbf{mmmm} \rangle)$$

$$+ \frac{1-\gamma}{ReDe}(\nabla \cdot \tau_e + \mathbf{F}) \text{ for } \mathbf{x} \in \Omega \tag{1.14}$$

$$\nabla \cdot \mathbf{u} = 0, \quad \text{for } \mathbf{x} \in \Omega. \tag{1.15}$$

In this work we mainly investigate the well-posedness of this new multiscale rodlike polymeric model. Moreover, in most cases [6] $D_{\|}^{*}/D_{\perp}^{*} \approx 2$, so we can set $D_{\perp}^{*} = 1$ and $D_{\|}^{*} = 2$ for simple and without the lost of generality. Then (1.6) can be written as

$$\frac{\partial \psi}{\partial t} + \nabla \cdot (\mathbf{u}\psi) = \frac{\varepsilon^2}{De}\nabla \cdot [(\mathbf{I} + \mathbf{mm})(\psi \nabla \mu)]$$

$$+ \frac{1}{De}\mathcal{R} \cdot (\psi \mathcal{R}\mu) - \mathcal{R} \cdot (\mathbf{m} \times \kappa \cdot \mathbf{m}\psi), \quad \mathbf{m} \in \mathbb{S}^2. \tag{1.16}$$

We can verify this system (1.14)–(1.16) satisfies the energy identity in the following way.

Multiplying $\mathbf{u}$ to (1.14) and integrating it over Ω yields

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_\Omega |\mathbf{u}|^2 d\mathbf{x} + \int_\Omega \left[\frac{\gamma}{De}|\nabla \mathbf{u}|^2 + \frac{1-\gamma}{2Re}\langle (\mathbf{mm}:\mathbf{D})^2\rangle\right] d\mathbf{x} \\
&\qquad = \frac{1-\gamma}{ReDe}\int_\Omega [-\tau_e : \nabla \mathbf{u} + \mathbf{F}\cdot\mathbf{u}]\, d\mathbf{x} \\
&\qquad = \frac{1-\gamma}{ReDe}\int_\Omega [\langle (\mathbf{m}\times \mathcal{R}\mu)\mathbf{m}\rangle : \nabla \mathbf{u} - \langle \nabla \mu\rangle \cdot \mathbf{u}]\, d\mathbf{x} \\
&\qquad = \frac{1-\gamma}{ReDe}\int_\Omega \int_{|\mathbf{m}|=1} [(\mathbf{m}\times \mathcal{R}\mu)\mathbf{m}\psi : \nabla \mathbf{u} - \psi \nabla \mu \cdot \mathbf{u}]\, d\mathbf{m}d\mathbf{x}. \qquad (1.17)
\end{aligned}$$

Multiplying μ to (1.16) as well as integrating over in Ω and the unit sphere yields

$$\begin{aligned}
&\int_\Omega \int_{|\mathbf{m}|=1} \frac{\partial \psi}{\partial t}\mu\, d\mathbf{m}d\mathbf{x} + \int_\Omega \left[\frac{\varepsilon^2}{De}\langle \nabla\mu\cdot(\mathbf{I}+\mathbf{mm})\nabla\mu\rangle + \frac{1}{De}\langle \mathcal{R}\mu\cdot\mathcal{R}\mu\rangle\right] d\mathbf{x} \\
&\qquad = \int_\Omega \int_{|\mathbf{m}|=1} \mathbf{m}\times\kappa\cdot\mathbf{m}\psi\cdot\mathcal{R}\mu\, d\mathbf{m}d\mathbf{x} + \int_\Omega \int_{|\mathbf{m}|=1} \mathbf{u}\psi\cdot\nabla\mu\, d\mathbf{m}d\mathbf{x}. \qquad (1.18)
\end{aligned}$$

Additionally, we can calculate

$$\begin{aligned}
&\frac{d}{dt}\int_\Omega \int_{|\mathbf{m}|=1} \left[\psi\ln\psi + \frac{1}{2}\psi U\right] d\mathbf{m}d\mathbf{x} \\
&\qquad = \int_\Omega \int_{|\mathbf{m}|=1} \left[\frac{\partial\psi}{\partial t}\ln\psi + \frac{\partial\psi}{\partial t} + \frac{1}{2}\frac{\partial\psi}{\partial t}U + \frac{1}{2}\psi\frac{\partial U}{\partial t}\right] d\mathbf{m}d\mathbf{x} \\
&\qquad = \int_\Omega \int_{|\mathbf{m}|=1} \left[\frac{\partial\psi}{\partial t}(\ln\psi + U) + \frac{1}{2}\left(\psi\frac{\partial U}{\partial t} - \frac{\partial\psi}{\partial t}U\right)\right] d\mathbf{m}d\mathbf{x} \\
&\qquad = \int_\Omega \int_{|\mathbf{m}|=1} \frac{\partial\psi}{\partial t}\mu\, d\mathbf{m}d\mathbf{x}. \qquad (1.19)
\end{aligned}$$

Combining (1.17)–(1.19), we obtain that the system (1.14)–(1.18) satisfies the energy law:

$$\begin{aligned}
&\frac{d}{dt}\left[\frac{1}{2}\int_\Omega |\mathbf{u}|^2 d\mathbf{x} + \frac{1-\gamma}{ReDe}E(\psi)\right] = -\int_\Omega \left[\frac{\gamma}{De}|\nabla\mathbf{u}|^2 + \frac{1-\gamma}{2Re}\langle(\mathbf{mm}:\mathbf{D})^2\rangle\right] d\mathbf{x} \\
&\qquad - \frac{1-\gamma}{ReDe}\int_\Omega \left[\frac{\varepsilon^2}{De}\langle\nabla\mu\cdot(\mathbf{I}+\mathbf{mm})\nabla\mu\rangle + \frac{1}{De}\langle\mathcal{R}\mu\cdot\mathcal{R}\mu\rangle\right] d\mathbf{x}, \qquad (1.20)
\end{aligned}$$

where $E(\psi)$ is a nonlocal intermolecular potential given by

$$E(\psi) = \int_\Omega \int_{|\mathbf{m}|=1} \psi(\mathbf{x},\mathbf{m},t)\ln\psi(\mathbf{x},\mathbf{m},t) + \frac{1}{2}U(\mathbf{x},\mathbf{m},t)\psi(\mathbf{x},\mathbf{m},t) d\mathbf{m}d\mathbf{x}. \qquad (1.21)$$

By the same way one can see that the original system (1.6)–(1.15) also satisfies the energy law like the form (1.20). Comparing to the Doi model of rodlike polymeric fluid [6], this new model is based on the more rational assumption that the particle distribution function (pdf) ψ is possibly different in the macro variable $\mathbf{x}$. When the pdf is the same at every point $\mathbf{x}$ in the domain Ω, this model is similar to the Doi model. Here the other important difference is the excluded-volume potential (1.3) if we choose $B = |\mathbf{m}\times\mathbf{m}'|^2$ or $B = |\mathbf{m}\times\mathbf{m}'|$ in (1.3), which is the well-known Maier–Saupe or Onsager potential [6]. Now in [8] E and Zhang have proved that the inhomogeneous

property reduces to the Ericksen–Leslie theory in the limit of small Debroah number. Recently numerical simulation results [26, 27] have shown that this model can really describe the anisotropic long-range elasticity of polymeric molecules, and the microstructure and defect dynamics of LCP solution. Moreover they have reported in [26] that there are seven in-plane flow modes in plane Couette flow described by this new model. Four of them have also been reported by Rey and Tsuji [22], and the other three modes are new complicated in-plane modes with inner defects. Furthermore, some significant scaling properties were verified in [26], such as the tumbling period is proportional to the inverse of the shear rate. In plane Poiseuille flow, different local states, such as flow-aligning, tumbling, or wagging, arise in different flow region. There are also some related numerical analysis results for special cases of this model of (1.14)–(1.16), e.g., [3] and references therein. These numerical results require a detailed well-posedness analysis for the system (1.14)–(1.16). This is the main objective of the present work. These related problems for the macroscopic nonlinear elasticity and viscoelasticity cases were recently studied by Sideris and Thomases [23] and Lin, Liu, and Zhang [16]. For the micro-macro model with dumbbell type of potential there are lots of works, e.g., [7, 15, 17] and references therein. In [28], we gave the globally classical existence theory and a numerical analysis for the Dirichlet initial boundary problem of the system (1.14)–(1.16) in a simple case, the 1+1-dimensional case, and the pressure-driven channel flow. More precisely, it is assumed that the rodlike particles rotate in shear plane. That work is a first step towards the better understanding for currently more sophisticated models (1.14)–(1.16).

In this paper we consider the system (1.14)–(1.16) with the initial data

$$\mathbf{u}(\mathbf{x},0)=\mathbf{u}_0,\quad \psi(\mathbf{x},\mathbf{m},0)=\psi_0(\mathbf{x},\mathbf{m}). \tag{1.22}$$

We denote the space of functions by $H^i(\Omega), i\in\mathbb{N}$, which are in $H^i_{loc}(\mathbb{R}^3)$ (i.e., $u|_\Omega$ for every open bounded set Ω) and which are periodic with period Ω: $\mathbf{u}(x_j+Le_j)=\mathbf{u}(x_j)$, $\psi(x_j+Le_j)=\psi(x_j)$, $j=1,2,3$. It is easy to see that ψ is also periodic with respect to the variable $\mathbf{m}$ since $\mathbf{m}\in\mathbb{S}^2$. Denote

$$H^i_d(\Omega)=\text{closure of } \mathcal{V} \text{ in } H^i(\Omega),\ \text{where } \mathcal{V}=C_0^\infty(\Omega)\cap\{\mathbf{u}:\ \text{div } \mathbf{u}=0\}.$$

We can see that $H^i_d(\Omega)$ is a Sobolev space. Moreover, we define the space

$$H^l(\Omega,\mathcal{X}_k)=\left\{\psi:\left|\sum_{j=0}^{l}\sum_{i=0}^{k}\int_\Omega\int_{|\mathbf{m}|=1}|\nabla^j\mathcal{R}^i\psi(\mathbf{x},\mathbf{m},t)|^2d\mathbf{m}d\mathbf{x}<\infty\right.\right\}, \tag{1.23}$$

with the natural topology of a Banach space. Now we state our main results:

THEOREM 1.1 (local existence). [illegible]

*(**A1**)* $\mathbf{u}_0\in H^3_d(\Omega)$ [illegible] $\mathbf{u}_0=0$ $\mathbf{u}_0$ [illegible]

*(**A2**)* $\psi_0\in\cap_{i+j=3}H^i(\Omega,\mathcal{X}_j)$ [illegible] $\mathbf{x}$ [illegible] $\psi_0\ge 0$ [illegible]

$$\int_\Omega\int_{|\mathbf{m}|=1}\psi_0(\mathbf{x},\mathbf{m})d\mathbf{m}d\mathbf{x}=1.$$

[illegible] $\|\mathbf{u}_0\|^2_{H^3_d(\Omega)}+\|\psi_0\|^2_{\cap_{i+j=3}H^i(\Omega,\mathcal{X}_j)}$ [illegible] $T'>0$ [illegible] (1.14)–(1.16) [illegible] (1.22) [illegible] $(\mathbf{u},\psi)$ [illegible]

$$\mathbf{u}\in L^\infty([0,T');H^3_d(\Omega))\cap L^2([0,T);H^4_d(\Omega)); \tag{1.24}$$

$$\psi\in L^\infty([0,T');\cap_{i+j=3}H^i(\Omega,\mathcal{X}_j))\cap L^2([0,T');\cap_{i+j=4}H^i(\Omega,\mathcal{X}_j)), \tag{1.25}$$

[illegible] fi [illegible]

THEOREM 1.2 (global existence). *(**A1**) (**A2**)*

$$\|\mathbf{u}_0\|^2_{H^3_d(\Omega))} + \|\psi_0\|^2_{H^3(\Omega,\mathcal{X}_0)\cap H^2(\Omega,\mathcal{X}_1)} \le B,$$

B (1.14) (1.16) (1.22) $(\mathbf{u},\psi)$

$$\mathbf{u} \in L^\infty([0,\infty); H^3_d(\Omega)), \tag{1.26}$$

$$\psi \in L^\infty([0,\infty); H^3(\Omega,\mathcal{X}_0)\cap H^2(\Omega,\mathcal{X}_1)), \tag{1.27}$$

$$\|\mathbf{u}\|^2_{L^\infty([0,\infty);H^3_d(\Omega))} + \|\psi\|^2_{L^\infty([0,\infty);H^3(\Omega,\mathcal{X}_0)\cap H^2(\Omega,\mathcal{X}_1))} \le B$$

$$Re < \gamma/C_2, \qquad De < \varepsilon^2/C_2;$$

C_2 n Ω

1.1. From the proof of Theorem 1.2 in section 4 we see that C_2 is a large positive constant. Thus Theorem 1.2 requires the Deborah and Reynolds numbers to be small enough.

1.2. From the proof of Theorem 1.1 we can also obtain the "global" result for the two-dimensional system. That is, for given $T > 0$, there exists a solution under the conditions of Theorem 1.1,

$$\mathbf{u} \in L^\infty([0,T]; H^3_d(\Omega)) \cap L^2([0,T); H^4_d(\Omega)); \tag{1.28}$$

$$\psi \in L^\infty([0,T]; \cap_{i+j=3} H^i(\Omega,\mathcal{X}_j)) \cap L^2([0,T'); \cap_{i+j=4} H^i(\Omega,\mathcal{X}_j)). \tag{1.29}$$

But the bound of this solution depends on T.

These results are similar to that of the Navier–Stokes of traditional models of complex fluids in the case of the spatially periodic solutions [5, 24] for $n = 3$. However, in contrast to traditional models of complex fluids [5, 24] which express polymer stress τ using empirical constitutive relations; τ in (1.11) expresses the polymer stress in terms of the microscopic conformations of the polymers. So the model considered in this work is closer to the original system for polymeric fluids in kinetic theory of polymers. But, on the other hand, it causes the difficulty of well-posedness analysis and numerical simulation since we have to study the configuration equation (1.16) which involves five spatial freedom variables, two of them are in the configuration domain and the others are in the macro flow domain. We utilized the properties of the Laplace–Bertrami operator on compact Riemannian manifold to obtain the existence and the preservation of the positivity of the solution to the linearized equation of (1.16). Then the regularity of the solution was strengthened by the energy estimates method. Thus, in virtue of the properties of the distribution function ψ, we can obtain the regularity of the stress τ. Then it is finished by the local well-posedness analysis of the rigid rodlike model by utilizing the Galerkin approximation and energy methods. However, we specially point out that the nonlinear stress (1.11) and the nonlinear body force (1.13) concerned with the solution of the micro-scale model (1.16) let us only obtain the global solution for small Deborah and Reynolds numbers in virtue of the method which we choose in this paper.

The paper is organized as follows. In section 2, we give the iterative scheme of the system to obtain the existence of the local solution and the scheme alternates between

solving an equation of the same type as encountered in incompressible elasticity and solving a linear diffusion equation. Section 3 is devoted to giving the detailed proof of the main lemmas. We will investigate the global existence of the solution in section 4. In this paper C denotes different constant depending only on γ, De, Re, Ω, and ε if there are no special notations. Some times we denote $H_d^p(\Omega), L^p(\Omega)$ by H^p, L^p for brevity.

2. Local solution. In this section we will construct an iterative scheme of the system (1.14)–(1.16) with (1.22) and with which we can obtain the existence of the local solution.

Motivated by the approach [12, 20], we construct an iterative scheme of the system (1.14)–(1.16) with (1.22). Given an iteration ψ^l we determine $\mathbf{u}^{l+1}$ by solving the equations

$$\begin{aligned}&\mathbf{u}_t^{l+1} + (\mathbf{u}^{l+1}\cdot\nabla)\mathbf{u}^{l+1} + \nabla p^{l+1}\\ &\quad = \frac{\gamma}{Re}\Delta\mathbf{u}^{l+1} + \frac{1-\gamma}{2Re}\nabla\cdot(\mathbf{D}:\langle\mathbf{mmmm}\rangle)^{l+1} + \frac{1-\gamma}{ReDe}(\nabla\cdot\tau_e^l + F^l),\end{aligned} \tag{2.1}$$

$$\nabla\cdot\mathbf{u}^{l+1} = 0 \tag{2.2}$$

with the initial condition $\mathbf{u}^{l+1}(\mathbf{x},0) = \mathbf{u}_0(\mathbf{x})$, where

$$(\mathbf{D}:\langle\mathbf{mmmm}\rangle)^{l+1} = \mathbf{D}_{ks}^{l+1}\langle m_i m_j m_k m_s\rangle^l, \tag{2.3}$$

$$\langle m_i m_j m_k m_s\rangle^l = \int_{|\mathbf{m}|=1} m_i m_j m_k m_s \psi^l(\mathbf{x},\mathbf{m},t)d\mathbf{m}, \tag{2.4}$$

$$(\tau_e^l) = -\langle(\mathbf{m}\times\mathcal{R}\mu^l)\mathbf{m}\rangle^l, \tag{2.5}$$

$$\mu^l = \ln\psi^l + U^l, \tag{2.6}$$

$$U^l = \alpha\int_\Omega\int_{|\mathbf{m}'|=1} B(\mathbf{x},\mathbf{x}';\mathbf{m},\mathbf{m}')\psi^l(\mathbf{x}',\mathbf{m}',t)d\mathbf{m}'d\mathbf{x}', \tag{2.7}$$

$$\kappa^{l+1} = (\nabla\mathbf{u}^{l+1})^T,\ F^l = -\langle\nabla\mu^l\rangle. \tag{2.8}$$

Meanwhile, for given $\mathbf{u}^l$, we determine ψ^l from the following initial value problem:

$$\begin{aligned}&\frac{\partial\psi^l}{\partial t} + \nabla\cdot(\mathbf{u}^{\mathbf{l}}\psi^l) = \frac{\varepsilon^2}{De}[\nabla\cdot(\mathbf{I}+\mathbf{mm})\nabla\psi^l + \nabla\cdot(\mathbf{I}+\mathbf{mm})(\psi^l\nabla U^l)]\\ &\qquad + \frac{1}{De}[\mathcal{R}\cdot\mathcal{R}\psi^l + \mathcal{R}\cdot(\psi^l\mathcal{R}U^l)] - \mathcal{R}\cdot(\mathbf{m}\times\kappa^l\cdot\mathbf{m}\psi^l),\end{aligned} \tag{2.9}$$

$$\psi^l(\mathbf{x},\mathbf{m},0) = \psi_0(\mathbf{x},\mathbf{m}). \tag{2.10}$$

Our eventual task is to show that the mapping $\mathcal{M}: \mathbf{u}^l \mapsto \mathbf{u}^{l+1}$ has a fixed point in an appropriate complete space of functions. The fixed point of the mapping is the solution we seek.

We will consider the mapping $\mathcal{M}$ in the function space $S(M,T)$, which is defined as a set of all functions $\mathbf{u}: \Omega\times[0,T]\to\mathbb{R}^n (n=2,3)$ with the following properties:

$$\mathbf{u}\in L^\infty([0,T];H_d^3(\Omega))\cap L^2([0,T];H_d^4(\Omega)), \tag{2.11}$$

$$\|\mathbf{u}\|^2_{L^\infty([0,T];H_d^3(\Omega))} + \|\mathbf{u}\|^2_{L^2([0,T];H_d^4(\Omega))} \le M. \tag{2.12}$$

On $S(M,T)$, we define the metric

$$d(\mathbf{u}_1,\mathbf{u}_2) = \|\mathbf{u}_1-\mathbf{u}_2\|_{L^\infty([0,T];H_d^3(\Omega))} + \|\mathbf{u}_1-\mathbf{u}_2\|_{L^2([0,T];H_d^4(\Omega))}. \tag{2.13}$$

It is easy to verify that $S(M,T)$ is complete with the associated metric, and also it is nonempty for large M, as evidenced in [19]. The properties of the mapping $\mathcal{M}$ is established by proving the next three lemmas.

LEMMA 2.1.

$$\|\psi^l\|_{L^\infty([0,T];H^3(\Omega,\mathcal{X}_0))\cap L^2([0,T];H^4(\Omega,\mathcal{X}_0))} \le K. \tag{2.14}$$

$T' > 0$ (2.1)–(2.8)

$$\mathbf{u}^{l+1} \in L^\infty([0,T'];H^3_d(\Omega)) \cap L^2([0,T'];H^4_d(\Omega)) \tag{2.15}$$

$$\|\mathbf{u}^{l+1}\|^2_{L^\infty([0,T'];H^3_d(\Omega))} + \|\mathbf{u}^{l+1}\|^2_{L^2([0,T'];H^4_d(\Omega))} \le \phi_1(T',K), \tag{2.16}$$

$f = \frac{1-\gamma}{ReDe}(\nabla\cdot\tau^l_e + F^l)$

$$\begin{aligned}\phi_1(T,K) &= [\|\mathbf{u}_0\|^2_{H^3_d(\Omega)} + C\|f\|^2_{L^2([0,T],H^2(\Omega))} + C(K)]e^{CK^2T+CK^4T}\\ &\quad + \|\mathbf{u}_0\|^2_{H^3_d(\Omega)} + C\|f\|^2_{L^2([0,T],H^2(\Omega))} + C(K)T.\end{aligned} \tag{2.17}$$

LEMMA 2.2.

$$\tau^l_e \in L^2([0,T];H^3(\Omega)), \qquad F^l \in L^2([0,T];H^2(\Omega)) \tag{2.18}$$

$$\|\tau^l_e\|_{L^2([0,T];H^3(\Omega))} \le C\|\psi^l\|_{L^2([0,T];H^3(\Omega,\mathcal{X}_0))}, \tag{2.19}$$

$$\|F^l\|_{L^2([0,T];H^2(\Omega))} \le C\|\psi^l\|_{L^2([0,T];H^3(\Omega,\mathcal{X}_0))} \tag{2.20}$$

$\psi^l \in L^2([0,T];H^3(\Omega,\mathcal{X}_0))$

LEMMA 2.3. $\mathbf{u}^l \in S(M,T)$ (2.9)–(2.10)

$$\psi^l \in L^\infty([0,T];\cap_{i+j=3}H^i(\Omega,\mathcal{X}_j)) \cap L^2([0,T];\cap_{i+j=4}H^i(\Omega,\mathcal{X}_j)), \tag{2.21}$$

$$\|\psi^l\|^2_{L^2([0,T],\cap_{i+j=3}H^i(\Omega,\mathcal{X}_j))} \le \|\psi_0\|^2_{\cap_{i+j=3}H^i(\Omega,\mathcal{X}_j)}(CT+CTM)\cdot e^{CT+CMT}. \tag{2.22}$$

By combining Lemmas 2.1–2.3, it follows easily that the map $\mathcal{M}: S(M,T') \longrightarrow S(M,T')$ is a compact operator if M is chosen sufficiently large and T' is chosen sufficiently small. In fact, by using (2.22), we know

$$\|\psi\|^2_{L^2([0,T],H^3(\Omega,\mathcal{X}_0))} \le \|\psi_0\|^2_{H^3(\Omega,\mathcal{X}_0))}(CT+CMT)e^{CT+CMT}.$$

Thus

$$\|f\|_{L^2([0,T],H^2(\Omega))} \le C\|\nabla\cdot\tau^l_e + F^l\|_{L^2([0,T],H^2(\Omega))} \le C\|\psi\|_{L^2([0,T],H^3(\Omega,\mathcal{X}_0))}.$$

Then, from (2.17), we have

$$\begin{aligned}&\|\mathbf{u}^{l+1}\|^2_{L^\infty([0,T'];H^3_d(\Omega))} + \|\mathbf{u}^{l+1}\|^2_{L^2([0,T'];H^4_d(\Omega))}\\ &\le [\|\mathbf{u}_0\|^2_{H^3_d(\Omega)} + \|\psi_0\|^2_{H^3(\Omega,\mathcal{X}_0)}(CT+CMT)e^{CT+CMT} + C(K)]e^{CK^2T+CK^4T}\\ &+ \|\mathbf{u}_0\|^2_{H^3_d(\Omega)} + \|\psi_0\|^2_{H^3(\Omega,\mathcal{X}_0)}(CT+CMT)e^{CT+CMT} + C(K)T.\end{aligned}$$

Now we choose

$$M \geq 6\|\mathbf{u}_0\|^2_{H^3(\Omega)} + 2C(K)(2+T_0) + 12CT_0\|\psi_0\|^2_{H^3(\Omega,\mathcal{X}_0))}, \tag{2.23}$$

$$T' \leq T_0 \triangleq \min\left\{\frac{\ln 2}{C(1+M)}, \frac{\ln 2}{C(K^2+K^4)}, \frac{1}{12C\|\psi_0\|^2_{H^3(\Omega,\mathcal{X}_0))}}\right\}. \tag{2.24}$$

Then $e^{CT'+CMT'} \leq 2, e^{CK^2T'+CK^4T'} \leq 2$, and

$$\begin{aligned}
&C\|\psi_0\|^2_{H^3(\Omega,\mathcal{X}_0)}e^{CT'+CMT'}MT'[1+e^{CK^2T'+CK^4T'}] \leq \frac{M}{2},\\
&\|\mathbf{u}_0\|^2_{H^3(\Omega)}(1+e^{CK^2T'+CK^4T'}) + C(K)[T' + e^{CK^2T'+CK^4T'}]\\
&+ C\|\psi_0\|^2_{H^3(\Omega,\mathcal{X}_0)}e^{CT'+CMT'}MT'[1+e^{CK^2T'+CK^4T'}]T' \leq \frac{M}{2}.
\end{aligned}$$

Thus we have (2.12).

Since $S(M,T')$ is clearly a closed, convex subset of $L^2([0,T],H^4_d(\Omega))$ and is also compact, by the fixed point theorem of Leray and Schauder [14], the conclusion of Theorem 1.1 can be obtained.

3. Proof of lemmas.

3.1. Estimates of u. In this section we will give the proof of Lemma 2.1. Let $\mathbf{u}^{l+1} = w$, $q = p^{l+1}$), and the fourth order tensor

$$A(\mathbf{x},t) = (a_{ijks}), \quad \text{where } a_{ijks}(\mathbf{x},t) = \int_{|\mathbf{m}|=1} m_i m_j m_k m_s \psi^l(\mathbf{x},\mathbf{m},t)d\mathbf{m}. \tag{3.1}$$

(2.1) can be rewritten as

$$\begin{aligned}
w_t + (w\cdot\nabla)w + \nabla q &= \frac{\gamma}{Re}\triangle w + f\\
&\quad + \frac{1-\gamma}{2Re}\nabla\cdot[\mathbf{D} : A(\mathbf{x},t)], \qquad (3.2)\\
\nabla\cdot w &= 0, \qquad (3.3)
\end{aligned}$$

where $f = \frac{1-\gamma}{ReDe}(\nabla\cdot\tau_e^l + F^l)$. In the following we solve this problem to obtain the existence and uniqueness of the solution by using the Galerkin approximation similar to solving the standard Navier–Stokes equation [5, 24]. The difference here is the appearance of the term $\frac{1-\gamma}{2Re}\nabla\cdot[\mathbf{D} : A(\mathbf{x},t)]$. We can see that it is a good term when we give a priori estimates because we will obtain $-\frac{1-\gamma}{2Re}\int_\Omega\langle(\mathbf{D}:\mathbf{mm})^2\rangle d\mathbf{x}$ while multiplying w to (3.2) and integrating it in Ω. The high derivatives estimates are obtained similarly. Thus we will obtain the result in Lemma 2.1 provided that

$$f \in L^2([0,T];H^2(\Omega)), \tag{3.4}$$

$$A \in L^\infty([0,T],H^3(\Omega)) \cap L^2([0,T],H^4(\Omega)). \tag{3.5}$$

The regularity of f and A can be easily obtained from the estimates of τ_e and F in section 3.2 and ones of ψ in section 3.3, respectively. Why we need conditions (3.4) and (3.5), for the aim of self-contained, will be answered in the outline of the proof to Lemma 2.1 when $n = 2$ and $n = 3$ in the appendix. We also refer the readers to [5, 24] for further details.

3.2. Estimates of τ_e and F.

Proof of Lemma 2.2. From the definition of τ_e, it is straightforward to obtain its estimates from the assumption of ψ. In fact,

$$\begin{aligned}\tau_e &= -\int_{|\mathbf{m}|=1}(\mathbf{m}\times\mathcal{R}\mu)\mathbf{m}\psi d\mathbf{m}\\ &= -\int_{|\mathbf{m}|=1}\left[\mathbf{m}\times\left(\frac{1}{\psi}\mathcal{R}\psi+\mathcal{R}U\right)\right]\mathbf{m}\psi d\mathbf{m}\\ &= -\int_{|\mathbf{m}|=1}(\mathbf{m}\times\mathcal{R}\psi)\mathbf{m}d\mathbf{m}-\int_{|\mathbf{m}|=1}(\mathbf{m}\times\mathcal{R}U)\mathbf{m}\psi d\mathbf{m}\\ &= -\mathbf{I}+3\int_{|\mathbf{m}|=1}\mathbf{m}\mathbf{m}\psi d\mathbf{m}-\int_{|\mathbf{m}|=1}(\mathbf{m}\times\mathcal{R}U)\mathbf{m}\psi d\mathbf{m},\end{aligned}$$

where we used the property of operators $\mathcal{R}$ and $\int_{|\mathbf{m}|=1}\cdot\, d\mathbf{m}$ (p. 293 in [6]),

$$\int_{|\mathbf{m}|=1}G(\mathbf{m})\mathcal{R}[F(\mathbf{m})]d\mathbf{m}=-\int_{|\mathbf{m}|=1}F(\mathbf{m})\mathcal{R}[G(\mathbf{m})]d\mathbf{m}. \tag{3.6}$$

From the definition (2.7) of U,

$$\begin{aligned}U &= \alpha\int_\Omega\int_{|\mathbf{m}'|=1}\frac{1}{\varepsilon^3}\chi\left(\frac{\mathbf{x}-\mathbf{x}'}{\varepsilon}\right)|\mathbf{m}\times\mathbf{m}'|^2\psi^l(\mathbf{x}',\mathbf{m}',t)d\mathbf{m}'d\mathbf{x}'\\ &= \alpha\int_\Omega\int_{|\mathbf{m}'|=1}\frac{1}{\varepsilon^3}\chi\left(\frac{\mathbf{x}-\mathbf{x}'}{\varepsilon}\right)\psi^l(\mathbf{x}',\mathbf{m}',t)d\mathbf{m}'d\mathbf{x}'\\ &\quad -\alpha\mathbf{m}\mathbf{m}:\int_\Omega\int_{|\mathbf{m}'|=1}\frac{1}{\varepsilon^3}\chi\left(\frac{\mathbf{x}-\mathbf{x}'}{\varepsilon}\right)\mathbf{m}'\mathbf{m}'\psi^l(\mathbf{x}',\mathbf{m}',t)d\mathbf{m}'d\mathbf{x}';\end{aligned} \tag{3.7}$$

here $\varepsilon>0$ is in (1.5) and $\chi(\mathbf{x})$ is a smooth kernel, and we can see that $U\in C^\infty(\Omega\times\mathbb{S}^2)$. But the bounds of the derivatives of U with respect to $\mathbf{x}$ and $\mathbf{m}$ depend on ε, denoted by $C(\varepsilon)$.

Therefore

$$\|\tau_e\|^2_{L^2(\Omega)}\le C(1+\|\psi\|^2_{L^2(\Omega,\mathcal{X}_0)}) \tag{3.8}$$

since $\psi\in L^2([0,T],H^3(\Omega,\mathcal{X}_0))$. Higher derivatives can be obtained similarly. Thus (2.19) is obtained. By the same way we can obtain

$$F=\langle\nabla\mu\rangle=\int_{|\mathbf{m}|=1}(\nabla\psi+\psi\nabla U)d\mathbf{m}\in L^2([0,T];H^2(\Omega)) \tag{3.9}$$

since $\psi\in L^2([0,T],H^3(\Omega,\mathcal{X}_0))$. Hence (2.20) is obtained by using (3.9).

3.3. Estimates of ψ. In this section we first review some results about the Laplace–Beltrami operator on a compact Riemannian manifold which will be utilized to show the existence of the solution of (2.9)–(2.10). Then we will give the regularity estimates of the solution to (2.9)–(2.10).

3.3.1. Review about the Laplace–Beltrami operator. We recall some known results about the Laplace–Beltrami operator on a compact Riemannian manifold (M_n, g); see (section 4 of Chap. 4) of [1] and [14].

LEMMA 3.1. [illegible] $v(Q,t)$ [illegible] $M_n \times [0,t_0]$ [illegible] $v \geq 0$ [illegible] $M_n \times \{0\}$ [illegible] $\partial M_n \times [0,t_0]$ [illegible] *fi*

$$\partial v/\partial t \geq \triangle_{M_n} v + b^i(Q,t)\partial_i v + c(Q,t)v \tag{3.10}$$

[illegible] b^i, c [illegible] $v \geq 0$ [illegible]. This lemma can be proved similar to the proof of the maximum principle in p. 130 of [1]. Let $C_m = \max_{M_n \times \mathbb{R}^+} |c(Q,t)|$ and $w = e^{-(C_m+1)t}v$. Then w and v have the same sign. Since

$$\partial w/\partial t = e^{-(C_m+1)t}[\partial v/\partial t - (C_m+1)v],$$

we have

$$\partial w/\partial t \geq \triangle_{M_n} w + b^i(Q,t)\partial_i w + [c(Q,t) - C_m - 1]w. \tag{3.11}$$

Assume w is negative somewhere and let (Q,t) be a point where w achieves its minimum. Then $\triangle_{M_n} w \geq 0, \partial_i w = 0$, and $\partial w/\partial t \leq 0$. Thus (3.11) implies $w(Q,t) \geq 0$, which yields a contradiction. □

LEMMA 3.2. [illegible] $g \in L^\infty([0,t_0], L^p(M_n))$ [illegible]

$$v \in W^{1,\infty}([0,t_0], W^{2,p}(M_n))$$

[illegible]

$$\partial v^i/\partial t = \triangle_{M_n} v^i + a^i_j \partial_i v^j + b^i_j v^j + g^i, \tag{3.12}$$

[illegible] $1 \leq \alpha \leq k$ [illegible] $v(P,0) \equiv 0$, $P \in M_n$ [illegible] $v^i (i = 1,2,\cdots,k)$ [illegible] k [illegible] $M_n \times [0,\infty)$ [illegible] $g^i (i = 1,2,\cdots,k)$ [illegible] k [illegible] $M_n \times [0,\infty)$. [illegible] *ffi* [illegible] a^i_j [illegible] b^i_j [illegible]

3.1. In fact, the condition on the coefficients a^i_j and b^i_j in the above lemma can be weakened to be bounded in $L^\infty([0,t_0] \times M_n)$. The proof is analogous.

3.3.2. Estimate of ψ.

Proof of Lemma 2.3. In this section we simply denote

$$\phi(\mathbf{x},\mathbf{m},t) = \psi^l(\mathbf{x},\mathbf{m},t) \qquad W = U^l. \tag{3.13}$$

Then (2.9)–(2.10) can be rewritten as

$$\frac{\partial \phi}{\partial t} + \mathbf{u}^l \cdot \nabla \phi = \frac{\varepsilon^2}{De}[\nabla \cdot (\mathbf{I} + \mathbf{mm})\nabla\phi + \nabla \cdot (\mathbf{I} + \mathbf{mm})(\phi \nabla W)]$$

$$+ \frac{1}{De}[\mathcal{R} \cdot \mathcal{R}\phi + \mathcal{R} \cdot (\phi \mathcal{R} W)] - \mathcal{R} \cdot (\mathbf{m} \times \kappa \cdot \mathbf{m}\phi), \tag{3.14}$$

$$\phi(\mathbf{x},\mathbf{m},0) = \psi_0(\mathbf{x},\mathbf{m}). \tag{3.15}$$

Here we still denote $(\nabla \mathbf{u}^l(\mathbf{x},t))^T$ by κ. Equation (3.14) is a nonlinear differential-integral equation. We first linearize (3.14) replacing W by U^{l-1} and obtaining the solution of this linear equation with the initial data (3.15). If now we transform the Cartesian coordinates $\mathbf{m}$ to the local coordinates (θ,φ) of the unit sphere $\mathbb{S}^2$ in

$\mathbb{R}^3$ by $m_1(\theta,\varphi) = \sin\theta\cos\varphi$, $m_2(\theta,\varphi) = \sin\theta\sin\varphi$, and $m_3 = \cos\varphi$, the operator $\mathcal{R}\cdot\mathcal{R}$ is the Laplace–Beltrami operator on the unit sphere [4]. Thus the operator $\frac{\varepsilon^2}{De}\triangle + \frac{1}{De}\mathcal{R}\cdot\mathcal{R}$ is also the Laplace–Beltrami operator. Moreover, the coefficients of (3.14) are all bounded and smooth when κ is bounded, which is obtained from $\mathbf{u}\in L^\infty([0,T];H^3_d(\Omega))$. Therefore we can utilize the above lemmas to show the existence and nonnegativeness of the solution to the linearized equation when the initial data is nonnegative. In fact, using Lemma 3.2 and Remark 3.1 for $v = \phi - \psi_0$, we obtain that (3.14)–(3.15) possesses a unique global solution $v \in W^{1,\infty}([0,T], H^2(\Omega,\mathcal{X}_0)\cap H^1(\Omega,\mathcal{X}_1)\cap L^2(\Omega,\mathcal{X}_2))$. Therefore, $\psi^l \in W^{1,\infty}([0,T];H^2(\Omega,\mathcal{X}_0)\cap H^1(\Omega,\mathcal{X}_1)\cap L^2(\Omega,\mathcal{X}_2))$ for given $T>0$. Now the coefficient of ϕ is $\frac{\varepsilon^2}{De}\nabla\cdot(\mathbf{I}+\mathbf{mm})\nabla W + \frac{1}{De}\mathcal{R}\cdot\mathcal{R}W - \mathcal{R}\cdot(\mathbf{m}\times\kappa\cdot\mathbf{m})$ and it is bounded since $\psi^{l-1}\in L^\infty([0,T];\cap_{i+j=3}H^i(\Omega,\mathcal{X}_j))$ and $\mathbf{u}^l\in L^\infty([0,T];H^3_d(\Omega))$. Then it follows that the positivity is preserved by Lemma 3.1, and by integrating both sides of (3.14) we find that

$$\int_\Omega\int_{|\mathbf{m}|=1}\phi(\mathbf{x},\mathbf{m},t)d\mathbf{m}d\mathbf{x} = \int_\Omega\int_{|\mathbf{m}|=1}\phi(\mathbf{x},\mathbf{m},0)d\mathbf{m}d\mathbf{x} = 1 \text{ for all } t.$$

We can obtain the solution of (3.14) with (3.15) from the solution of the linearized equation to construct a suitable Sobolev space and use the Schauder fixed point theorem in the standard argument. So here we omit the detail.

Next we will further prove the regularity of the solution $\psi^l(\mathbf{x},\mathbf{m},t)$ if the initial data is more regular.

I. The estimate of ϕ. Multiplying ϕ to (3.14) and integrating on $\mathbb{S}^2$ with respect to $\mathbf{m}$ and in Ω with respect to $\mathbf{x}$ yields

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_\Omega\int_{|\mathbf{m}|=1}|\phi|^2d\mathbf{m}d\mathbf{x} + \frac{\varepsilon^2}{De}\int_\Omega\int_{|\mathbf{m}|=1}(|\nabla\phi|^2+|\mathbf{m}\cdot\nabla\phi|^2)d\mathbf{m}d\mathbf{x}\\
&\quad+\frac{1}{De}\int_\Omega\int_{|\mathbf{m}|=1}|\mathcal{R}\phi|^2d\mathbf{m}d\mathbf{x}\\
&=-\frac{\varepsilon^2}{De}\int_\Omega\int_{|\mathbf{m}|=1}(\mathbf{I}+\mathbf{mm})\phi\nabla W\cdot\nabla\phi d\mathbf{m}d\mathbf{x} - \frac{1}{De}\int_\Omega\int_{|\mathbf{m}|=1}\phi\mathcal{R}W\cdot\mathcal{R}\phi d\mathbf{m}d\mathbf{x}\\
&\quad+\int_\Omega\int_{|\mathbf{m}|=1}\mathbf{m}\times\kappa\cdot\mathbf{m}\phi\cdot\mathcal{R}\phi d\mathbf{m}d\mathbf{x},
\end{aligned}$$

where we used the periodic condition. Thus we can obtain the estimate,

$$\begin{aligned}
&\frac{d}{dt}\int_\Omega\int_{|\mathbf{m}|=1}|\phi|^2d\mathbf{m}d\mathbf{x} + \frac{\varepsilon^2}{De}\int_\Omega\int_{|\mathbf{m}|=1}(|\nabla\phi|^2+|\mathbf{m}\cdot\nabla\phi|^2)d\mathbf{m}d\mathbf{x}\\
&\quad+\frac{1}{De}\int_\Omega\int_{|\mathbf{m}|=1}|\mathcal{R}\phi|^2d\mathbf{m}d\mathbf{x}\\
&\le C\int_\Omega\int_{|\mathbf{m}|=1}|\phi|^2d\mathbf{m}d\mathbf{x} + C|\kappa|\int_\Omega\int_{|\mathbf{m}|=1}|\phi|^2d\mathbf{m}d\mathbf{x},
\end{aligned}\tag{3.16}$$

where we used the fact that $|\kappa|$ is also bounded in $\Omega\times[0,T]$ redefined by a set of measure zero from (3.7) since $\mathbf{u}^l\in L^\infty([0,T];H^3(\Omega))$ and $|\nabla W|, |\mathcal{R}W|$ are bounded by $C(\varepsilon)$. Therefore from (3.16) we obtain

$$\sup_{t\in[0,T]}\int_\Omega\int_{|\mathbf{m}|=1}|\phi|^2d\mathbf{m}d\mathbf{x} \le \|\psi_0\|^2_{L^2(\Omega,\mathcal{X}_0)}e^{CT+CMT} \triangleq N_1,\tag{3.17}$$

and

$$\frac{\varepsilon^2}{De}\int_0^T\int_\Omega\int_{|\mathbf{m}|=1}(|\nabla\phi|^2+|\mathbf{m}\cdot\nabla\phi|^2)d\mathbf{m}d\mathbf{x}dt+\frac{1}{De}\int_0^T\int_\Omega\int_{|\mathbf{m}|=1}|\mathcal{R}\phi|^2d\mathbf{m}d\mathbf{x}dt$$
$$\leq CN_1T+CMN_1T\triangleq N_2. \tag{3.18}$$

This shows that

$$\phi\in L^\infty([0,T];L^2(\Omega,\mathcal{X}_0))\cap L^2([0,T];L^2(\Omega,\mathcal{X}_1))\cap L^2([0,T];H^1(\Omega,\mathcal{X}_0))$$

provided that $\psi_0\in L^2(\Omega,\mathcal{X}_0)$ and $\mathbf{u}^l\in S(M,T)$.

II. The estimate of $\mathcal{R}\phi$, ϕ. Applying the operator $\mathcal{R}$ to (3.14) and multiplying $\mathcal{R}\phi$ to (3.14) and integrating on $\mathbb{S}^2$ with respect to $\mathbf{m}$ and in Ω with respect to $\mathbf{x}$, respectively, yields

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int_\Omega\int_{|\mathbf{m}|=1}|\mathcal{R}\phi|^2d\mathbf{m}d\mathbf{x}+\frac{\varepsilon^2}{De}\int_\Omega\int_{|\mathbf{m}|=1}(|\nabla\mathcal{R}\phi|^2+|\mathbf{m}\cdot\nabla\mathcal{R}\phi|^2)d\mathbf{m}d\mathbf{x}\\&\quad+\frac{1}{De}\int_\Omega\int_{|\mathbf{m}|=1}|\mathcal{R}\cdot\mathcal{R}\phi|^2d\mathbf{m}d\mathbf{x}\\&=-\frac{\varepsilon^2}{De}\int_\Omega\int_{|\mathbf{m}|=1}\{\mathcal{R}\cdot(\mathbf{I}+\mathbf{mm})\cdot\nabla\phi\cdot\nabla\mathcal{R}\phi-\nabla\cdot[(\mathbf{I}+\mathbf{mm})(\phi\nabla W)]\mathcal{R}\cdot\mathcal{R}\phi\}d\mathbf{m}d\mathbf{x}\\&\quad+\frac{1}{De}\int_\Omega\int_{|\mathbf{m}|=1}\mathcal{R}\cdot(\phi\mathcal{R}W)\mathcal{R}\cdot\mathcal{R}\phi d\mathbf{m}d\mathbf{x}+\int_\Omega\int_{|\mathbf{m}|=1}\mathcal{R}\cdot(\mathbf{m}\times\kappa\cdot\mathbf{m}\phi)\mathcal{R}\cdot\mathcal{R}\phi d\mathbf{m}d\mathbf{x}.\end{aligned}$$

Thus, by using $|\nabla W|,|\nabla^2W|$, $|\nabla\mathcal{R}W|$, $|\mathcal{R}W|,|\mathcal{R}^2W|$ are bounded for all $\mathbf{x}\in\Omega$ and $\mathbf{m}\in\mathbb{S}^2$, and $|\kappa|$ is bounded by redefined in a set of measure zero, we have

$$\begin{aligned}&\frac{d}{dt}\int_\Omega\int_{|\mathbf{m}|=1}|\mathcal{R}\phi|^2d\mathbf{m}d\mathbf{x}+\frac{\varepsilon^2}{De}\int_\Omega\int_{|\mathbf{m}|=1}(|\nabla\mathcal{R}\phi|^2+|\mathbf{m}\cdot\nabla\mathcal{R}\phi|^2)d\mathbf{m}d\mathbf{x}\\&\quad+\frac{1}{De}\int_\Omega\int_{|\mathbf{m}|=1}|\mathcal{R}\cdot\mathcal{R}\phi|^2d\mathbf{m}d\mathbf{x}\\&\leq C\int_\Omega\int_{|\mathbf{m}|=1}(|\mathcal{R}\phi|^2+|\nabla_{\mathbf{x}}\phi|^2+|\phi|^2)d\mathbf{m}d\mathbf{x}\\&\quad+C|\kappa|\int_\Omega\int_{|\mathbf{m}|=1}(|\mathcal{R}\phi|^2+|\phi|^2)d\mathbf{m}d\mathbf{x}.\end{aligned}\tag{3.19}$$

Similarly, differentiating (3.14) with respect to $\mathbf{x}$, then multiplying $\nabla\phi$ and integrating it, we obtain

$$\begin{aligned}&\frac{d}{dt}\int_\Omega\int_{|\mathbf{m}|=1}|\nabla\phi|^2d\mathbf{m}d\mathbf{x}+\frac{\varepsilon^2}{De}\int_\Omega\int_{|\mathbf{m}|=1}(|\triangle\phi|^2+|\mathbf{m}\cdot\nabla^2\phi|^2)d\mathbf{m}d\mathbf{x}\\&\quad+\frac{1}{De}\int_\Omega\int_{|\mathbf{m}|=1}|\nabla\mathcal{R}\phi|^2d\mathbf{m}d\mathbf{x}\\&\leq C\int_\Omega\int_{|\mathbf{m}|=1}(|\nabla\phi|^2+|\mathcal{R}\phi|^2+|\phi|^2)d\mathbf{m}d\mathbf{x}\\&\quad+C|\kappa|\int_\Omega\int_{|\mathbf{m}|=1}(|\nabla\phi|^2+|\phi|^2)d\mathbf{m}d\mathbf{x}.\end{aligned}\tag{3.20}$$

Here we used that $|\nabla W|, |\nabla^2 W|$, $|\mathcal{R}W|, |\mathcal{R}^2 W|$ and κ are bounded. Combination of (3.16), (3.19), and (3.20) and application of the Grownwall inequality yields

$$\begin{aligned}
&\sup_{t\in[0,T]} \int_\Omega \int_{|\mathbf{m}|=1} (|\nabla\phi|^2 + |\mathcal{R}\phi|^2 + |\phi|^2) d\mathbf{m}d\mathbf{x} \\
&\quad \le \int_\Omega \int_{|\mathbf{m}|=1} (|\nabla\psi_0|^2 + |\mathcal{R}\psi_0|^2 + |\psi_0|^2) d\mathbf{m}d\mathbf{x}\ e^{CT+CM} \triangleq N_3, \\
&\frac{\varepsilon^2}{De} \int_0^T \int_\Omega \int_{|\mathbf{m}|=1} (|\nabla\phi|^2 + |\triangle\phi|^2 + |\nabla\mathcal{R}\phi|^2) d\mathbf{m}d\mathbf{x}dt \\
&\quad + \frac{1}{De} \int_0^T \int_\Omega \int_{|\mathbf{m}|=1} |\nabla\mathcal{R}\phi|^2 + |\mathcal{R}\cdot\mathcal{R}\phi|^2 d\mathbf{m}d\mathbf{x}dt \le CN_3T + CMN_3T.
\end{aligned}$$

This shows that

$$\begin{aligned}
&\phi \in L^\infty([0,t]; \cap_{i+j=1,i,j\ge 0} H^i(\Omega, \mathcal{X}_j)) \quad \text{and} \\
&\phi \in L^2([0,t]; \cap_{i+j=2,i,j\ge 0} H^i(\Omega, \mathcal{X}_j))
\end{aligned} \tag{3.21}$$

provided that $\psi_0 \in \cap_{i+j=1,i,j\ge 0} H^i(\Omega, \mathcal{X}_j)$.

III. The estimate of $\mathcal{R}^2\phi$, $^2\phi$, and $\mathcal{R}\phi$. Similar to that in section II, we can obtain the following estimates:

$$\begin{aligned}
&\frac{d}{dt} \int_\Omega \int_{|\mathbf{m}|=1} |\nabla\mathcal{R}\phi|^2 d\mathbf{m}d\mathbf{x} + \frac{\varepsilon^2}{De} \int_\Omega \int_{|\mathbf{m}|=1} (|\triangle\mathcal{R}\phi|^2 + |\mathbf{m}\cdot\nabla^2\mathcal{R}\phi|^2) d\mathbf{m}d\mathbf{x} \\
&\quad + \frac{1}{De} \int_\Omega \int_{|\mathbf{m}|=1} |\nabla\mathcal{R}\cdot\mathcal{R}\phi|^2 d\mathbf{m}d\mathbf{x} \\
&\le C \int_\Omega \int_{|\mathbf{m}|=1} (|\nabla\mathcal{R}\phi|^2 + |\nabla^2_{\mathbf{x}}\phi|^2 + |\nabla\phi|^2 + |\mathcal{R}\phi|^2 + |\phi|^2) d\mathbf{m}d\mathbf{x} \\
&\quad + C|\kappa| \int_\Omega \int_{|\mathbf{m}|=1} (|\mathcal{R}^2\phi|^2 + |\mathcal{R}\phi|^2 + |\phi|^2) d\mathbf{m}d\mathbf{x},
\end{aligned} \tag{3.22}$$

$$\begin{aligned}
&\frac{d}{dt} \int_\Omega \int_{|\mathbf{m}|=1} |\mathcal{R}^2\phi|^2 d\mathbf{m}d\mathbf{x} + \frac{\varepsilon^2}{De} \int_\Omega \int_{|\mathbf{m}|=1} (|\nabla\mathcal{R}\cdot\mathcal{R}\phi|^2 + |\mathbf{m}\cdot\nabla\mathcal{R}^2\phi|^2) d\mathbf{m}d\mathbf{x} \\
&\quad + \frac{1}{De} \int_\Omega \int_{|\mathbf{m}|=1} |\mathcal{R}(\mathcal{R}\cdot\mathcal{R}\phi)|^2 d\mathbf{m}d\mathbf{x} \\
&\le C \int_\Omega \int_{|\mathbf{m}|=1} (|\mathcal{R}^2\phi|^2 + |\mathcal{R}\phi|^2 + |\nabla\phi|^2 + |\nabla\mathcal{R}\phi|^2 + |\phi|^2) d\mathbf{m}d\mathbf{x} \\
&\quad + C|\kappa| \int_\Omega \int_{|\mathbf{m}|=1} (|\mathcal{R}^2\phi|^2 + |\mathcal{R}\phi|^2 + |\phi|^2) d\mathbf{m}d\mathbf{x},
\end{aligned} \tag{3.23}$$

$$\begin{aligned}
&\frac{d}{dt}\int_\Omega\int_{|\mathbf{m}|=1}|\nabla^2\phi|^2d\mathbf{m}d\mathbf{x}+\frac{\varepsilon^2}{De}\int_\Omega\int_{|\mathbf{m}|=1}(|\nabla\triangle\phi|^2+|\mathbf{m}\cdot\nabla^3\phi|^2)d\mathbf{m}d\mathbf{x}\\
&\quad+\frac{1}{De}\int_\Omega\int_{|\mathbf{m}|=1}|\mathcal{R}\triangle_{\mathbf{x}}\phi|^2d\mathbf{m}d\mathbf{x}\\
&\le C\int_\Omega\int_{|\mathbf{m}|=1}(|\nabla^2\phi|^2+|\nabla\phi|^2+|\mathcal{R}\phi|^2+|\nabla\mathcal{R}\phi|^2+|\phi|^2)d\mathbf{m}d\mathbf{x}\\
&\quad+C|\nabla\kappa|\int_\Omega\int_{|\mathbf{m}|=1}(|\mathcal{R}\phi|^2+|\nabla\phi|^2+|\phi|^2)d\mathbf{m}d\mathbf{x}\\
&\quad+C|\kappa|\int_\Omega\int_{|\mathbf{m}|=1}(|\nabla\mathcal{R}\phi|^2+|\nabla\phi|^2)d\mathbf{m}d\mathbf{x}. \qquad (3.24)
\end{aligned}$$

Combining the above estimates and applying the Gronwall inequality with $|\nabla\kappa|\in L^2([0,T],H^4(\Omega))$, we can obtain

$$\begin{aligned}
\|\phi\|^2_{L^\infty([0,t];\cap_{i+j=2,i,j\ge0}H^i(\Omega,\mathcal{X}_j))}&\le\|\psi_0\|^2_{\cap_{i+j=2,i,j\ge0}H^i(\Omega,\mathcal{X}_j)}e^{CT+CMT}\triangleq N_4,\\
\|\phi\|^2_{L^2([0,t];\cap_{i+j=3,i,j\ge0}H^i(\Omega,\mathcal{X}_j))}&\le CN_4T+CMN_4T.
\end{aligned}$$

This implies that

$$\begin{aligned}
&\phi\in L^\infty([0,t];\cap_{i+j=2,i,j\ge0}H^i(\Omega,\mathcal{X}_j))\quad\text{and}\\
&\phi\in L^2([0,t];\cap_{i+j=3,i,j\ge0}H^i(\Omega,\mathcal{X}_j)) \qquad (3.25)
\end{aligned}$$

provided that $\psi_0\in\cap_{i+j=2,i,j\ge0}H^i(\Omega,\mathcal{X}_j)$.

IV. The estimate of $\mathcal{R}^i\ \ ^j\phi(i+j=3)$. Analogously to that in section III, we can obtain the following estimates:

$$\begin{aligned}
&\frac{d}{dt}\int_\Omega\int_{|\mathbf{m}|=1}|\mathcal{R}^3\phi|^2d\mathbf{m}d\mathbf{x}+\frac{\varepsilon^2}{De}\int_\Omega\int_{|\mathbf{m}|=1}(|\nabla\mathcal{R}^3\phi|^2+|\mathbf{m}\cdot\nabla\mathcal{R}^3\phi|^2)d\mathbf{m}d\mathbf{x}\\
&\quad+\frac{1}{De}\int_\Omega\int_{|\mathbf{m}|=1}|\mathcal{R}^4\phi|^2d\mathbf{m}d\mathbf{x}\\
&\le C\int_\Omega\int_{|\mathbf{m}|=1}(|\mathcal{R}^3\phi|^2+|\mathcal{R}^2\phi|^2+|\mathcal{R}\phi|^2+|\phi|^2\\
&\quad+|\nabla\mathcal{R}^2\phi|^2+|\nabla\mathcal{R}\phi|^2+|\nabla\phi|^2)d\mathbf{m}d\mathbf{x}\\
&\quad+C|\kappa|\int_\Omega\int_{|\mathbf{m}|=1}(|\mathcal{R}^3\phi|^2+|\mathcal{R}^2\phi|^2+|\mathcal{R}\phi|^2+|\phi|^2)d\mathbf{m}d\mathbf{x}, \qquad (3.26)\\
&\frac{d}{dt}\int_\Omega\int_{|\mathbf{m}|=1}|\nabla\mathcal{R}^2\phi|^2d\mathbf{m}d\mathbf{x}+\frac{\varepsilon^2}{De}\int_\Omega\int_{|\mathbf{m}|=1}(|\nabla^2\mathcal{R}^2\phi|^2+|\mathbf{m}\cdot\nabla^2\mathcal{R}^2\phi|^2)d\mathbf{m}d\mathbf{x}\\
&\quad+\frac{1}{De}\int_\Omega\int_{|\mathbf{m}|=1}|\mathcal{R}^3\nabla\phi|^2d\mathbf{m}d\mathbf{x}\\
&\le C\int_\Omega\int_{|\mathbf{m}|=1}(|\nabla^2\mathcal{R}\phi|^2+|\nabla^2\phi|^2+|\nabla\mathcal{R}^2\phi|^2+|\nabla\mathcal{R}\phi|^2\\
&\quad+|\nabla\phi|^2+|\mathcal{R}^2\phi|^2+|\phi|^2)d\mathbf{m}d\mathbf{x}\\
&\quad+C|\kappa|\int_\Omega\int_{|\mathbf{m}|=1}|\nabla\phi|^2d\mathbf{m}d\mathbf{x}+C|\nabla\kappa|\int_\Omega\int_{|\mathbf{m}|=1}|\phi|^2d\mathbf{m}d\mathbf{x}, \qquad (3.27)
\end{aligned}$$

$$
\begin{aligned}
&\frac{d}{dt}\int_\Omega\int_{|\mathbf{m}|=1}|\nabla^2\mathcal{R}\phi|^2d\mathbf{m}d\mathbf{x}+\frac{\varepsilon^2}{De}\int_\Omega\int_{|\mathbf{m}|=1}(|\nabla^3\mathcal{R}\phi|^2+|\mathbf{m}\cdot\nabla^3\mathcal{R}\phi|^2)d\mathbf{m}d\mathbf{x}\\
&\quad+\frac{1}{De}\int_\Omega\int_{|\mathbf{m}|=1}|\nabla^2\mathcal{R}^2\phi|^2d\mathbf{m}d\mathbf{x}\\
&\le C\int_\Omega\int_{|\mathbf{m}|=1}(|\nabla^3\phi|^2+|\nabla^2\mathcal{R}\phi|^2+|\nabla\mathcal{R}\phi|^2+|\mathcal{R}\phi|^2\\
&\quad+|\nabla^2\phi|^2+|\nabla\phi|^2+|\phi|^2)d\mathbf{m}d\mathbf{x}\\
&\quad+C|\kappa|\int_\Omega\int_{|\mathbf{m}|=1}(|\nabla\phi|^2+|\nabla\mathcal{R}\phi|^2+|\nabla^2\mathcal{R}\phi|^2+|\nabla\mathcal{R}^2\phi|^2)d\mathbf{m}d\mathbf{x}\\
&\quad+C|\nabla\kappa|\int_\Omega\int_{|\mathbf{m}|=1}(|\mathcal{R}^2\phi|^2+|\nabla\mathcal{R}\phi|^2+|\mathcal{R}\phi|^2+|\phi|^2)d\mathbf{m}d\mathbf{x},
\end{aligned}
\tag{3.28}
$$

$$
\begin{aligned}
&\frac{d}{dt}\int_\Omega\int_{|\mathbf{m}|=1}|\nabla^3\phi|^2d\mathbf{m}d\mathbf{x}+\frac{\varepsilon^2}{De}\int_\Omega\int_{|\mathbf{m}|=1}(|\nabla^4\phi|^2+|\mathbf{m}\cdot\nabla^4\phi|^2)d\mathbf{m}d\mathbf{x}\\
&\quad+\frac{1}{De}\int_\Omega\int_{|\mathbf{m}|=1}|\nabla^3\mathcal{R}\phi|^2d\mathbf{m}d\mathbf{x}\\
&\le C\int_\Omega\int_{|\mathbf{m}|=1}(|\nabla^3\phi|^2+|\nabla^2\phi|^2+|\nabla\phi|^2+|\phi|^2)d\mathbf{m}d\mathbf{x}\\
&\quad+C|\kappa|\int_\Omega\int_{|\mathbf{m}|=1}|\nabla^3\phi|^2d\mathbf{m}d\mathbf{x}+C|\nabla\kappa|\int_\Omega\int_{|\mathbf{m}|=1}|\nabla^2\phi|^2d\mathbf{m}d\mathbf{x}\\
&\quad+C\|\phi\|^2_{\mathcal{X}_0}\int_\Omega|\nabla^3\kappa|^2d\mathbf{x}+C\|\nabla^3\kappa\|^2_{L^2}\int_\Omega\int_{|\mathbf{m}|=1}(|\nabla^2\phi|^2+|\nabla\phi|^2)d\mathbf{m}d\mathbf{x}\\
&\quad+C\|\nabla^2\kappa\|^2_{L^2}\int_\Omega\int_{|\mathbf{m}|=1}(|\nabla\phi|^2+|\phi|^2)d\mathbf{m}d\mathbf{x}.
\end{aligned}
\tag{3.29}
$$

In the last estimate we used the following estimates:

$$
\begin{aligned}
&\int_\Omega\left[|\nabla^2\kappa|\int_{|\mathbf{m}|=1}|\nabla\phi||\nabla^3\mathcal{R}\phi|d\mathbf{m}\right]d\mathbf{x}\\
&\le\int_\Omega|\nabla^2\kappa|\|\nabla\phi\|_{L^2(\mathbb{S}^2)}\|\nabla^3\mathcal{R}\phi\|_{L^2(\mathbb{S}^2)}d\mathbf{x}\\
&\le\|\nabla^2\kappa\|_{L^3(\Omega)}\|\|\nabla\phi\|_{L^2(\mathbb{S}^2)}\|_{L^6(\Omega)}\|\|\nabla^3\mathcal{R}\phi\|_{L^2(\mathbb{S}^2)}\|_{L^2(\Omega)}\\
&\le\|\nabla^2\kappa\|_{H^{\frac12}(\Omega)}\|\|\nabla\phi\|_{L^2(\mathbb{S}^2)}\|_{H^1(\Omega)}\|\|\nabla^3\mathcal{R}\phi\|_{L^2(\mathbb{S}^2)}\|_{L^2(\Omega)}\\
&\le\|\nabla^2\kappa\|_{H^4(\Omega)}\|\phi\|_{H^2(\Omega,\mathcal{X}_0)}\|\|\nabla^3\mathcal{R}\phi\|_{L^2(\mathbb{S}^2)}\|_{L^2(\Omega)}\\
&\le\frac{1}{De}\|\nabla^3\mathcal{R}\phi\|^2_{L^2(\Omega,\mathcal{X}_0)}+C(De)\|\mathbf{u}\|^2_{H^4(\Omega)}\|\phi\|^2_{H^2(\Omega,\mathcal{X}_0)}.
\end{aligned}
\tag{3.30}
$$

Combining the above estimates and applying the Gronwall inequality by using $\mathbf{u}\in L^2([0,T],H^4(\Omega))$, we can obtain

$$
\begin{aligned}
\|\phi\|_{L^\infty([0,t];\cap_{i+j=3,i,j\ge0}H^i(\Omega,\mathcal{X}_j))}&\le\|\psi_0\|_{\cap_{i+j=3,i,j\ge0}H^i(\Omega,\mathcal{X}_j)}e^{CT+CMT}\triangleq N_6,\\
\|\phi\|_{L^2([0,t];\cap_{i+j=4,i,j\ge0}H^i(\Omega,\mathcal{X}_j))}&\le CN_6T+CMN_6T.
\end{aligned}
$$

This implies that

$$
\begin{aligned}
&\phi\in L^\infty([0,t];\cap_{i+j=3,i,j\ge0}H^i(\Omega,\mathcal{X}_j))\quad\text{and}\\
&\phi\in L^2([0,t];\cap_{i+j=4,i,j\ge0}H^i(\Omega,\mathcal{X}_j))
\end{aligned}
\tag{3.31}
$$

provided that $\psi_0\in\cap_{i+j=3,i,j\ge0}H^i(\Omega,\mathcal{X}_j)$.

Up to now, we have completed the proof of Lemma 2.3.

4. Global solution. In this section, we will show that the local solution obtained in Theorem 1.1 is actually defined for $t \in \mathbb{R}^+$ if the Deborah and Reynolds numbers are small enough. To this end we derive some a priori bounds, satisfied by that solution. Notably, C_1 in this section denotes different constants depending only on n and Ω.

4.1. Some a priori estimates. Recall that the local solution $(\mathbf{u}, \psi)$ obtained in Theorem 1.1 satisfies (1.24) and (1.25) together with (1.14) and (1.16). Now we give the detail of a priori estimates for the case $n = 3$. For $n = 2$, it can be similarly obtained.

For (1.14), the inequality (A.36) in the appendix implies that there exists constant C such that

(4.1)
$$\frac{1}{2}\frac{d}{dt}\|\mathbf{u}\|^2_{H^3_d} + \left(\frac{\gamma}{Re} - 5\epsilon - \frac{6\epsilon}{11}\right)\|\mathbf{u}\|^2_{H^4} \le \frac{5\epsilon}{11}\|\psi\|^2_{H^4(\Omega,\mathcal{X}_0)} + \frac{C_1}{\epsilon}[\|\mathbf{u}\|^{14}_{H^1} + \|\mathbf{u}\|^4_{H^1} + \|\mathbf{u}\|^{24}_{H^2} + \|\psi\|^{20}_{H^1(\Omega,\mathcal{X}_0)} + \|\psi\|^{24}_{H^2(\Omega,\mathcal{X}_0)} + \|\psi\|^8_{H^3(\Omega,\mathcal{X}_0)} + \|\psi\|^2_{H^3(\Omega,\mathcal{X}_0)}], \quad (n = 3).$$

Here in the left side of the above inequality we omit the term $\frac{1-\gamma}{2Re}\int_\Omega \langle |\mathbf{mm} : \nabla^3 D|^2\rangle d\mathbf{x}$ since it is positive. The main difficulty to obtain (4.1) is to estimate

$$\int_\Omega \left|\nabla^3\left(\mathbf{D} : \int_{|\mathbf{m}|=1} \mathbf{mmmm}\nabla\psi d\mathbf{m}\right)\nabla^4\mathbf{u}\right| d\mathbf{x}.$$

Now we will use the following inequalities:

$$\begin{aligned}
&\int_\Omega |\nabla^2\mathbf{D}\int_{|\mathbf{m}|=1}\mathbf{mmmm}\nabla\psi d\mathbf{m}\nabla^4\mathbf{u}|d\mathbf{x}\\
&\quad\le C_1\|\nabla^4\mathbf{u}\|_{L^2}\|\nabla^2\mathbf{D}\|_{L^3}\|\nabla\psi\|_{L^6(\Omega,\mathcal{X}_0)}\\
&\quad\le C_1\|\mathbf{u}\|_{H^4}\|\nabla^2\mathbf{D}\|_{H^{\frac{1}{2}}}\|\nabla\psi\|_{H^1(\Omega,\mathcal{X}_0)}\\
&\quad\le C_1\|\mathbf{u}\|_{H^4}\|\mathbf{u}\|_{H^{3+\frac{1}{2}}}\|\psi\|_{H^2(\Omega,\mathcal{X}_0)}\\
&\quad\le C_1\|\mathbf{u}\|_{H^4}\|\mathbf{u}\|^{\frac{5}{6}}_{H^4}\|\mathbf{u}\|^{\frac{1}{6}}_{H^1}\|\psi\|_{H^2(\Omega,\mathcal{X}_0)}\\
&\quad\le \epsilon\|\mathbf{u}\|^2_{H^4} + \frac{C_1}{\epsilon}\|\mathbf{u}\|^2_{H^1}\|\psi\|^{12}_{H^2(\Omega,\mathcal{X}_0)},
\end{aligned} \tag{4.2}$$

where ϵ is a positive constant and will be chosen later. Here we used the Hölder inequality ($\int_\Omega |abc|d\mathbf{x} \le \|a\|_{L^2}\|a\|_{L^3}\|c\|_{L^6}$), Sobolev embedding theorems ($H^1(\Omega) \subset L^6(\Omega), H^{1/2}(\Omega) \subset L^3(\Omega)$ for $n = 3$), the interpolation inequality ($\|\mathbf{u}\|_{H^{3+\frac{1}{2}}} \le C_1\|\mathbf{u}\|^{\frac{5}{6}}_{H^4}\|\mathbf{u}\|^{\frac{1}{6}}_{H^1}$), and Young's inequality ($ab \le \epsilon a^{12/11} + \frac{1}{\epsilon}b^{12}$), respectively. Similarly, we have

$$\begin{aligned}
&\int_\Omega |\nabla\mathbf{D}\int_{|\mathbf{m}|=1}\mathbf{mmmm}\nabla^2\psi d\mathbf{m}\nabla^4\mathbf{u}|d\mathbf{x}\\
&\quad\le \epsilon\|\mathbf{u}\|^2_{H^4} + \frac{C_1}{\epsilon}\|\mathbf{u}\|^2_{H^1}\|\psi\|^4_{H^3(\Omega,\mathcal{X}_0)}.
\end{aligned}$$

Further, a different estimate from the above two is

$$\begin{aligned}
&\int_\Omega |\mathbf{D}\int_{|\mathbf{m}|=1}\mathbf{mmmm}\nabla^3\psi d\mathbf{m}\nabla^4\mathbf{u}|d\mathbf{x}\\
&\quad\le C_1\|\nabla^4\mathbf{u}\|_{L^2}\|\mathbf{D}\|_{L^6}\|\nabla^3\psi\|_{L^3(\Omega,\mathcal{X}_0)}\\
&\quad\le C_1\|\mathbf{u}\|_{H^4}\|\mathbf{D}\|_{H^1}\|\nabla^3\psi\|_{H^{\frac{1}{2}}(\Omega,\mathcal{X}_0)}\\
&\quad\le C_1\|\mathbf{u}\|_{H^4}\|\mathbf{u}\|_{H^2}\|\psi\|^{\frac{1}{6}}_{H^1(\Omega,\mathcal{X}_0)}\|\psi\|^{\frac{5}{6}}_{H^4(\Omega,\mathcal{X}_0)}\\
&\quad\le \epsilon\left[\|\mathbf{u}\|_{H^4}\|\psi\|^{\frac{5}{6}}_{H^4(\Omega,\mathcal{X}_0)}\right]^{\frac{12}{11}}+\frac{C_1}{\epsilon}\|\mathbf{u}\|^{12}_{H^2}\|\psi\|^{10}_{H^1(\Omega,\mathcal{X}_0)}\\
&\quad\le \epsilon\left[\|\mathbf{u}\|^{\frac{12}{11}}_{H^4}\|\psi\|^{\frac{10}{11}}_{H^4(\Omega,\mathcal{X}_0)}\right]+\frac{C_1}{\epsilon}\|\mathbf{u}\|^{12}_{H^2}\|\psi\|^{10}_{H^1(\Omega,\mathcal{X}_0)}\\
&\quad\le \frac{6\epsilon}{11}\|\mathbf{u}\|^2_{H^4}+\frac{5\epsilon}{11}\|\psi\|^2_{H^4(\Omega,\mathcal{X}_0)}+\frac{C_1}{\epsilon}\|\mathbf{u}\|^{12}_{H^2}\|\psi\|^{10}_{H^1(\Omega,\mathcal{X}_0)}.
\end{aligned}\tag{4.3}$$

For the convect term we used the same estimate in [24]

$$\int|(\mathbf{u}\cdot\nabla)\mathbf{u}\triangle^3\mathbf{u}|dx\le\epsilon\|\mathbf{u}\|^2_{H^4}+\frac{C_1}{\epsilon}\|\mathbf{u}\|^{14}_{H^1},$$

and by using the results of Lemma 2.2 we have

$$\begin{aligned}
&\int_\Omega|\nabla^3\tau_e\nabla^4\mathbf{u}|d\mathbf{x}\le\epsilon\|\mathbf{u}\|^2_{H^4}+\frac{C_1}{\epsilon}\|\tau_e\|^2_{H^3}\le\epsilon\|\mathbf{u}\|^2_{H^4}+\frac{C_1}{\epsilon}\|\psi\|^2_{H^3(\Omega,\mathcal{X}_0)},\\
&\int_\Omega|\nabla^2F\nabla^4\mathbf{u}|d\mathbf{x}\le\epsilon\|\mathbf{u}\|^2_{H^4}+\frac{C_1}{\epsilon}\|F\|^2_{H^2}\le\epsilon\|\mathbf{u}\|^2_{H^4}+\frac{C_1}{\epsilon}\|\psi\|^2_{H^3(\Omega,\mathcal{X}_0)}.
\end{aligned}$$

Combination of all yields (4.1). For the equation of ψ, we can obtain the following estimates:

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\psi\|^2_{H^3(\Omega,\mathcal{X}_0)}+\left(\frac{\varepsilon^2}{De}-5\epsilon-\frac{\epsilon}{11}\right)\|\psi\|^2_{H^4(\Omega,\mathcal{X}_0)}+\left(\frac{1}{De}-\frac{5\epsilon}{11}\right)\|\psi\|^2_{H^3(\Omega,\mathcal{X}_1)}\\
&\le\frac{5\epsilon}{11}\|\mathbf{u}\|^2_{H^4}+\frac{C_1}{\epsilon}[\|\mathbf{u}\|^4_{H^3}+\|\mathbf{u}\|^{24}_{H^2}+\|\mathbf{u}\|^4_{H^2}+\|\psi\|^4_{H^3(\Omega,\mathcal{X}_0)}+\|\psi\|^4_{H^1(\Omega,\mathcal{X}_0)}\\
&+\|\psi\|^{24}_{H^1(\Omega,\mathcal{X}_1)}+\|\psi\|^4_{H^2(\Omega,\mathcal{X}_1)}+\|\psi\|^2_{H^3(\Omega,\mathcal{X}_0)}].
\end{aligned}\tag{4.4}$$

Here the difficulties are the estimates to the nonlinear terms $\mathbf{u}\cdot\nabla\psi$ and $\mathcal{R}\cdot(\mathbf{m}\times\kappa\cdot\mathbf{m}\psi)$ in the equation. To overcome them we use the following inequalities similar to (4.2):

$$\begin{aligned}
&\int_\Omega\int_{|\mathbf{m}|=1}|\nabla^2\mathbf{u}\nabla\psi\nabla^4\psi|d\mathbf{m}d\mathbf{x}\le\epsilon\|\psi\|^2_{H^4(\Omega,\mathcal{X}_0)}+\frac{C_1}{\epsilon}\|\mathbf{u}\|^2_{H^3}\|\psi\|^2_{H^2(\Omega,\mathcal{X}_0)},\\
&\int_\Omega\int_{|\mathbf{m}|=1}|\nabla\mathbf{u}\nabla^2\psi\nabla^4\psi d\mathbf{m}d\mathbf{x}\le\epsilon\|\psi\|^2_{H^4(\Omega,\mathcal{X}_0)}+\frac{C_1}{\epsilon}\|\mathbf{u}\|^2_{H^2}\|\psi\|^2_{H^3(\Omega,\mathcal{X}_0)},\\
&\int_\Omega\int_{|\mathbf{m}|=1}|\mathbf{u}\nabla^3\psi\nabla^4\psi d\mathbf{m}d\mathbf{x}\le\epsilon\|\psi\|^2_{H^4(\Omega,\mathcal{X}_0)}+\frac{C_1}{\epsilon}\|\mathbf{u}\|^{12}_{H^1}\|\psi\|^2_{H^1(\Omega,\mathcal{X}_0)}.
\end{aligned}$$

For the other term $\int_\Omega\int_{|\mathbf{m}|=1}|\nabla^2\mathcal{R}\cdot(\mathbf{m}\times\kappa\cdot\mathbf{m}\psi)\nabla^4\psi|d\mathbf{m}d\mathbf{x}$, we used the estimates

obtained similar to (4.3):

$$\begin{aligned}
&\int_\Omega \int_{|\mathbf{m}|=1} |\nabla^2 \kappa \mathcal{R}\psi \nabla^4 \psi| d\mathbf{m} d\mathbf{x} \\
&\qquad \le \frac{6\epsilon}{11}\|\psi\|^2_{H^4(\Omega,\mathcal{X}_0)} + \frac{5\epsilon}{11}\|\mathbf{u}\|^2_{H^4} + \frac{C_1}{\epsilon}\|\mathbf{u}\|^2_{H^1}\|\psi\|^{12}_{H^1(\Omega,\mathcal{X}_1)}, \\
&\int_\Omega \int_{|\mathbf{m}|=1} |\kappa \nabla^2 \mathcal{R}\psi \nabla^4 \psi| d\mathbf{m} d\mathbf{x} \\
&\qquad \le \frac{6\epsilon}{11}\|\psi\|^2_{H^4(\Omega,\mathcal{X}_0)} + \frac{5\epsilon}{11}\|\psi\|^2_{H^3(\Omega,\mathcal{X}_1)} + \frac{C_1}{\epsilon}\|\mathbf{u}\|^{12}_{H^2}\|\psi\|^2_{H^1(\Omega,\mathcal{X}_1)}, \\
&\int_\Omega \int_{|\mathbf{m}|=1} |\nabla \kappa \nabla \mathcal{R}\psi \nabla^4 \psi| d\mathbf{m} d\mathbf{x} \le \epsilon\|\psi\|^2_{H^4(\Omega,\mathcal{X}_0)} + \frac{C_1}{\epsilon}\|\mathbf{u}\|^2_{H^3}\|\psi\|^2_{H^2(\Omega,\mathcal{X}_1)}.
\end{aligned}$$

Now from (4.4), we see that it is necessary to estimate $\|\psi\|^2_{H^2(\Omega,\mathcal{X}_1)}$. But it is not difficult to obtain the following inequality in the similar way with the estimate (4.4).

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\psi\|^2_{H^2(\Omega,\mathcal{X}_1)} + \left(\frac{\varepsilon^2}{De} - 4\epsilon\right)\|\psi\|^2_{H^3(\Omega,\mathcal{X}_1)} + \frac{1}{De}\|\psi\|^2_{H^2(\Omega,\mathcal{X}_2)} \\
&\le \frac{C_1}{\epsilon}\left[\|\mathbf{u}\|^4_{H^3} + \|\mathbf{u}\|^{\frac{16}{3}}_{H^2} + \|\mathbf{u}\|^{16}_{H^1} + \|\psi\|^4_{H^1(\Omega,\mathcal{X}_1)} + \|\psi\|^8_{H^1(\Omega,\mathcal{X}_1)} + \|\psi\|^2_{H^2(\Omega,\mathcal{X}_1)}\right].
\end{aligned} \tag{4.5}$$

For the case $n = 2$, we can obtain a priori estimates by using similar approaches, and we omit the details and give the results directly as follows:

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\mathbf{u}\|^2_{H^3_d} + \left(\frac{\gamma}{Re} - 5\epsilon - \frac{9\epsilon}{16}\right)\|\mathbf{u}\|^2_{H^4} \le \frac{7\epsilon}{16}\|\psi\|^2_{H^4(\Omega,\mathcal{X}_0)} + \frac{C_1}{\epsilon}[\|\mathbf{u}\|^4_{H^1} + \|\mathbf{u}\|^{18}_{H^2} \\
&+ \|\psi\|^{18}_{H^2(\Omega,\mathcal{X}_0)} + \|\psi\|^{\frac{9}{2}}_{H^3(\Omega,\mathcal{X}_0)}], \quad (n = 2).
\end{aligned} \tag{4.6}$$

The a priori estimates for ψ are

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\psi\|^2_{H^3(\Omega,\mathcal{X}_0)} + \left(\frac{\varepsilon^2}{De} - 5\epsilon - \frac{9\epsilon}{16}\right)\|\psi\|^2_{H^4(\Omega,\mathcal{X}_0)} + \left(\frac{1}{De} - \frac{4\epsilon}{13}\right)\|\psi\|^2_{H^3(\Omega,\mathcal{X}_1)} \\
&\le \frac{7\epsilon}{16}\|\mathbf{u}\|^2_{H^4} + \frac{C_1}{\epsilon}[\|\mathbf{u}\|^4_{H^3} + \|\mathbf{u}\|^4_{H^1} + \|\mathbf{u}\|^8_{H^1} + \|\mathbf{u}\|^{18}_{H^1} \\
&+ \|\psi\|^4_{H^1(\Omega,\mathcal{X}_1)} + \|\psi\|^{18}_{H^1(\Omega,\mathcal{X}_1)} + \|\psi\|^4_{H^3(\Omega,\mathcal{X}_0)}].
\end{aligned} \tag{4.7}$$

4.2. Global existence. In this subsection we will give the proof of Theorem 1.2. Let

$$Y(t) = \|\mathbf{u}\|^2_{H^3(\Omega)} + \|\psi\|^2_{H^3(\Omega,\mathcal{X}_0)} + \|\psi\|^2_{H^2(\Omega,\mathcal{X}_1)}. \tag{4.8}$$

Combining the above estimates (4.1), (4.4), and (4.5) we have

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}Y(t) + \left(\frac{\gamma}{Re} - 6\epsilon\right)\|\mathbf{u}\|^2_{H^4} \\
&+ \left(\frac{\varepsilon^2}{De} - 5\epsilon - \frac{6\epsilon}{11}\right)\|\psi\|^2_{H^4(\Omega,\mathcal{X}_0)} + \left(\frac{1}{De} - \frac{5\epsilon}{11}\right)\|\psi\|^2_{H^3(\Omega,\mathcal{X}_1)} \\
&+ \left(\frac{\varepsilon^2}{De} - 4\epsilon\right)\|\psi\|^2_{H^3(\Omega,\mathcal{X}_1)} + \frac{1}{De}\|\psi\|^2_{H^2(\Omega,\mathcal{X}_2)} \\
&\le \frac{C_1}{\epsilon}(Y^{12} + Y^{10} + Y^7 + Y^4 + Y^2 + Y).
\end{aligned} \tag{4.9}$$

Now we choose

$$\epsilon = \min\left\{\frac{\gamma}{7Re}, \frac{\varepsilon^2}{7De}\right\}, \quad \text{and} \quad \beta = \frac{C_1}{\epsilon}. \tag{4.10}$$

Since $Y(t) \le \|\mathbf{u}\|^2_{H^4} + \|\psi\|^2_{H^4(\Omega,\mathcal{X}_0)} + \|\psi\|^2_{H^3(\Omega,\mathcal{X}_1)}$, (4.9) implies

$$Y'(t) + 2\epsilon Y(t) \le 2\beta(Y^{12} + Y^{10} + Y^7 + Y^4 + Y^2 + Y). \tag{4.11}$$

LEMMA 4.1. [illegible] (4.11) [illegible] $0 < B < B_0$ [illegible] B_0 [illegible]

$$B^{11} + B^9 + B^6 + B^3 + B + 1 - \frac{\epsilon}{2\beta} = 0. \tag{4.12}$$

[illegible] $Y(0) \le B$ [illegible] $Y(t)$ [illegible] B [illegible] $t \ge 0$

[illegible]. We will prove it by contradiction. Suppose that there exists a t such that $Y(t) > B$, and define $t^* = \inf\{t \in \mathbb{R}^+, Y(t) > B\}$, then $Y(t^*) = B$ and $Y'(t^*) \ge 0$. However from (4.11) and the hypothesis made on B we deduce

$$\begin{aligned} Y'(t^*) &\le -2\epsilon Y(t^*) + 2\beta[Y^{11}(t^*) + Y^9(t^*) + Y^7(t^*) + Y^4(t^*) + Y^2(t^*) + Y(t^*)] \\ &\le -2\epsilon B + 2B\beta\frac{\epsilon}{2\beta} = -\epsilon B < 0, \end{aligned}$$

which contradicts the above statement. Therefore $Y(t) \le B$ for all $t \in \mathbb{R}^+$. □

[illegible] 1.2. We have seen in subsection 4.1 that a specific norm of the local solution obtained in Theorem 1.1 satisfies an inequality with the form (4.11), where C_1 depends on the domain Ω and n while ϵ depends on $\gamma, Re, \varepsilon^2, De$ from (4.10). Lemma 4.1 shows that there exists a constant B_0, depending on initial data such that

$$Y(t) \le B, \text{ for } t \in \mathbb{R}^+, \quad \text{if} \quad Y(0) \le B < B_0.$$

But B_0 is the unique positive solution of (4.12). From (4.12), we can easily see that (4.12) possesses a unique positive solution if and only if

$$1 - \frac{\epsilon}{2\beta} < 0, \tag{4.13}$$

which implies from (4.10) that

$$Re < \frac{\gamma}{7\sqrt{2C_1}} \quad \text{and} \quad De < \frac{\varepsilon^2}{7\sqrt{2C_1}}. \tag{4.14}$$

This completes the proof of Theorem 1.2 for $n = 3$. □

Denoting

$$Z(t) = \|\mathbf{u}\|^2_{H^3(\Omega)} + \|\psi\|^2_{H^3(\Omega,\mathcal{X}_0)},$$

the addition of (4.6) and (4.7) yields

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}Z(t) + \left(\frac{\gamma}{Re} - 6\epsilon\right)\|\mathbf{u}\|^2_{H^4} \\ &+ \left(\frac{\varepsilon^2}{De} - 6\epsilon\right)\|\psi\|^2_{H^4(\Omega,\mathcal{X}_0)} + \left(\frac{1}{De} - \frac{4\epsilon}{13}\right)\|\psi\|^2_{H^3(\Omega,\mathcal{X}_1)} \\ &\le \frac{C_1}{\epsilon}(Z^9 + Y^4 + Y^{9/4} + Y^2) \quad \text{for } n = 2. \end{aligned} \tag{4.15}$$

By the same way the result of Theorem 1.2 for $n = 2$ can be obtained.

Appendix A. Proof of Lemma 2.1. Now we solve (3.2)–(3.3) with the initial value $\mathbf{u}(\mathbf{x},0)=\mathbf{u}_0$ by using the Galerkin approximation. Let V be the space of all divergence-free vector in $H^1(\Omega)$, and let $\{\omega^i | i\in\mathbb{N}\}$ be a basis for V. We seek an approximation to w of the form

$$w^N(\mathbf{x},t)=\Sigma_{n=1}^N\rho^n(t)\omega^n(\mathbf{x}). \tag{A.1}$$

The function w^N satisfies, instead of (3.2) and (3.3),

$$\begin{aligned} w_t^N+(w^N\cdot\nabla)w^N&=\frac{\gamma}{Re}\triangle w^N+f\\ &\quad+\frac{1-\gamma}{2Re}\nabla\cdot\left[\mathbf{D}^N:A(\mathbf{x},t)\right], \end{aligned} \tag{A.2}$$

$$w^N(\mathbf{x},0)=P_N\mathbf{u}_0(\mathbf{x}), \tag{A.3}$$

where P_N is the orthogonal projector in H_d^1 onto $W_N=Span\{\omega^i, i=1,\cdots,N\}$. The existence and uniqueness of a solution w_N to (A.2)–(A.3) with periodic boundary conditions defined on some interval $(0,T_N), T_N>0$, is clear; in fact, the following estimate shows that $T_N=T$ for $n=2$ and T_N suitably small for $n=3$.

Multiplying w^N to (A.2) and integrating it, we get

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\|w^N\|_{L^2}^2+\frac{\gamma}{Re}\|\nabla w^N\|_{L^2}^2+\frac{1-\gamma}{2Re}\int_\Omega\langle|\mathbf{D}^N:\mathbf{mm}|^2\rangle d\mathbf{x}\\ &\le\frac{\gamma}{2Re}\|w^N\|_{L^2}^2+\frac{Re}{2\gamma}\|f\|_{H^{-1}}^2. \end{aligned} \tag{A.4}$$

This implies that

$$\frac{d}{dt}\|w^N\|_{L^2}^2+\frac{\gamma}{Re}\|\nabla w^N\|_{L^2}^2\le\frac{Re}{\gamma}\|f\|_{H^{-1}}^2. \tag{A.5}$$

It shows that

$$\int_0^T\|\nabla w^N(t)\|_{L^2}^2dt\le K_1, \tag{A.6}$$

where

$$K_1=K_1\left(\mathbf{u}_0,f,\frac{\gamma}{Re},T\right)=\frac{Re}{\gamma}\left(\|\mathbf{u}_0\|_{L^2}^2+\frac{Re}{\gamma}\int_0^T\|f(t)\|_{H^{-1}}^2dt\right). \tag{A.7}$$

For $0<s<T$, by (A.5),

$$\|w^N(s)\|_{L^2}^2\le K_2, \tag{A.8}$$

where

$$K_2=K_2\left(\mathbf{u}_0,f,\frac{\gamma}{Re},T\right)=\frac{\gamma}{Re}K_1. \tag{A.9}$$

Multiplying $\triangle w^N$ by (A.2) and integrating it, we get

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\|\nabla w^N\|_{L^2}^2+\frac{\gamma}{Re}\|\nabla^2 w^N\|_{L^2}^2+\frac{1-\gamma}{2Re}\int_\Omega\langle|\nabla\mathbf{D}^N:\mathbf{mm}|^2\rangle d\mathbf{x}\\ &\le\frac{1}{4}\frac{\gamma}{Re}\|\nabla w^N\|_{L^2}^2+\frac{Re}{\gamma}\|f\|_{L^2}^2\\ &\quad+\int|(w^N\cdot\nabla)w^N\triangle w^N|d\mathbf{x}+\frac{1-\gamma}{2Re}\int|\mathbf{D}^N\nabla A\triangle w^N|d\mathbf{x}. \end{aligned} \tag{A.10}$$

Now the following a priori of $\int |(w^N \cdot \nabla) w^N \triangle w^N| dx$ are different, depending on the dimension. We use the relation [24]

$$\int |(w \cdot \nabla) w \triangle w| dx \le C \|w\|_{L^2}^{1/2} \|\triangle w\|_{L^2}^{3/2} \|\nabla w\|_{L^2}, \qquad (n = 2), \tag{A.11}$$

$$\int |(w \cdot \nabla) w \triangle w| dx \le C \|w\|_{L^2}^{1/4} \|\triangle w\|_{L^2}^{7/4} \|\nabla w\|_{L^2}, \qquad (n = 3), \tag{A.12}$$

and the estimates

$$\int |\mathbf{D}^N \nabla A \triangle w^N| d\mathbf{x} \le C \|\triangle w^N\|_{L^2}^{4/3} \|\nabla^2 A\|_{L^2} \|\nabla w^N\|_{L^2}^{2/3}, \quad (n = 2), \tag{A.13}$$

$$\int |\mathbf{D}^N \nabla A \triangle w^N| d\mathbf{x} \le C \|\triangle w^N\|_{L^2}^{3/2} \|\nabla^2 A\|_{L^2} \|\nabla w^N\|_{L^2}^{1/2}, \quad (n = 3). \tag{A.14}$$

By Young's inequality (A.10) implies

$$\begin{aligned} \frac{d}{dt} \|\nabla w^N\|_{L^2}^2 + \frac{1}{2} \frac{\gamma}{Re} \|\Delta w^N\|_{L^2}^2 &\le \frac{2Re}{\gamma} \|f\|_{L^2}^2 \\ &+ C \|w^N\|_{L^2}^2 \|\nabla w^N\|_{L^2}^4 + C \|\nabla^2 A\|_{L^2}^4 \|\nabla w^N\|_{L^2}^2, \quad (n = 2), \end{aligned} \tag{A.15}$$

$$\begin{aligned} \frac{d}{dt} \|\nabla w^N\|_{L^2}^2 + \frac{3}{2} \frac{\gamma}{Re} \|\Delta w^N\|_{L^2}^2 &\le \frac{2Re}{\gamma} \|f\|_{L^2}^2 \\ &+ C \|w^N\|_{L^2}^2 \|\nabla w^N\|_{L^2}^8 + C \|\nabla^2 A\|_{L^2}^4 \|\nabla w^N\|_{L^2}^2, \quad (n = 3). \end{aligned} \tag{A.16}$$

Since $\psi^l \in L^\infty([0,T], H^3(\Omega, \mathcal{X}_0))$, we can obtain

$$|\nabla A(x,t)| \in L^\infty([0,T]; H^2(\Omega)). \tag{A.17}$$

Then by the same way as the a priori estimate of Theorem 3.2 in [24] we can get for $n = 2$ in virtue of (2.14) and (3.1)

$$\sup_{t \in [0,T]} \|\nabla w^N\|_{L^2}^2 \le K_3, \tag{A.18}$$

where

$$K_3 = K_3 \left(\mathbf{u}_0, f, \frac{\gamma}{Re}, T \right) = \left(\|\mathbf{u}_0\|_{H^1}^2 + C \int_0^T \|f(t)\|_{L^2}^2 dt \right) \exp(CK_1 K_2 + CK^3 T).$$

And

$$\int_0^T \|\triangle w^N\|_{L^2}^2 dt \le K_4, \tag{A.19}$$

where

$$K_4 = K_4(\mathbf{u}_0, f, \frac{\gamma}{Re}, T) = C\gamma \left(\|\mathbf{u}_0\|_{H^1}^2 + C \int_0^T \|f(t)\|_{L^2}^2 dt + CK_2 K_3^2 T + CK^3 K_3 T \right).$$

For $n = 3$, we will estimate in the following from (A.16). Let $y(t) = \|\nabla w^N\|_{L^2}^2 + 1, a(t) = C\|w^N\|_{L^2}^2, b(t) = C\|\nabla^2 A\|_{L^2}^4$, and $c(t) = \frac{2Re}{\gamma} \|f\|_{L^2}^2$. Then from (A.16), we have the inequality

$$\frac{dy}{dt} \le [a(t) + b(t) + c(t)] y^4, \tag{A.20}$$

and

$$\int_0^T [a(t)+b(t)+c(t)]dt \le T(CK_2+CK^6+CK^2), \tag{A.21}$$

where we used the condition of (2.14) and the estimate (3.8) and above estimates. Then they imply that for $t\in[0,T]$ and $T<1/[(CK_2+CK^4+\frac{Re}{\gamma}K^2)y^3(0)]\triangleq T_0$

$$y(t)\le \frac{y(0)}{\sqrt[3]{1-(CK_2+CK^4+\frac{Re}{\gamma}K^2)y^3(0)T}}\triangleq K_5. \tag{A.22}$$

Therefore, when $T<T_0$, we have

$$\sup_{s\in[0,T]}\|\nabla w^N\|_{L^2}^2\le K_5 \tag{A.23}$$

and

$$\int_0^T \|\triangle w^N\|_{L^2}^2 dt\le K_6, \tag{A.24}$$

where

$$K_6=K_6(\mathbf{u}_0,f,\frac{\gamma}{Re},T)=C\left(\|\mathbf{u}_0\|_{H^1}^2+C\int_0^T\|f(t)\|_{L^2}^2dt+CK_1K_5^4T+CK^4K_5T\right).$$

Multiplying $\triangle^2 w^N$ by (A.2) and integrating it, we get

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\nabla^2 w^N\|_{L^2}^2+\frac{\gamma}{Re}\|\nabla^3 w^N\|_{L^2}^2+\frac{1-\gamma}{4Re}\int_\Omega\langle|\nabla^2\mathbf{D}^N:\mathbf{mm}|^2\rangle d\mathbf{x}\\
&\quad\le\frac{1}{4}\frac{\gamma}{Re}\|\nabla\triangle w^N\|_{L^2}^2+C\|\nabla f\|_{L^2}^2\\
&\qquad+\frac{1-\gamma}{2Re}\int[|\mathbf{D}^N\nabla^2A\nabla\triangle w^N|+|\nabla\mathbf{D}^N\nabla A\nabla\triangle w^N|]dx\\
&\qquad+\int|(w^N\cdot\nabla)w^N\triangle^2w^N|dx.
\end{aligned}\tag{A.25}$$

In the following we will estimate (A.25). In virtue of the inequality (p. 31 [24])

$$\int|(w^N\cdot\nabla)w^N\triangle^r w^N|dx\le\frac{1}{4}\frac{\gamma}{Re}\|w^N\|_{H^{r+1}}^2+C\|w^N\|_{H^1}^{2r}\quad(n=2), \tag{A.26}$$

$$\int|(w^N\cdot\nabla)w^N\triangle^r w^N|dx\le\frac{1}{4}\frac{\gamma}{Re}\|w^N\|_{H^{r+1}}^2+C\|w^N\|_{H^1}^{4r+2}\quad(n=3). \tag{A.27}$$

And the estimates

$$\int|\mathbf{D}^N\nabla^2A\nabla^3w^N|d\mathbf{x}\le C\|\nabla^3w^N\|_{L^2}\|\nabla A\|_{L^2}^{1/3}\|\nabla^3A\|_{L^2}^{2/3}\|\nabla^2w^N\|_{L^2},\quad(n=2),$$

$$\int|\mathbf{D}^N\nabla^2A\nabla^3w^N|d\mathbf{x}\le C\|\nabla^3w^N\|_{L^2}\|\nabla A\|_{L^2}^{1/4}\|\nabla^3A\|_{L^2}^{3/4}\|\nabla^2w^N\|_{L^2},\quad(n=3),$$

$$\int|\nabla\mathbf{D}^N\nabla A\nabla^3w^N|d\mathbf{x}\le C\|\nabla^3w^N\|_{L^2}^{5/3}\|\nabla^2A\|_{L^2}\|\nabla w^N\|_{L^2}^{1/3},\quad(n=2),$$

$$\int|\nabla\mathbf{D}^N\nabla A\nabla^3w^N|d\mathbf{x}\le C\|\nabla^3w^N\|_{L^2}^{7/4}\|\nabla^2A\|_{L^2}\|\nabla w^N\|_{L^2}^{1/4}.\quad(n=3).$$

Thus, we have

$$\frac{d}{dt}\|w^N\|_{H^2}^2+\frac{1}{2}\frac{\gamma}{Re}\|w^N\|_{H^3}^2\le C\|w^N\|_{H^1}^4+\frac{Re}{\gamma}\|f\|_{H^1}^2$$
$$+C\|A\|_{H^1}^{2/3}\|A\|_{H^3}^{4/3}\|w^N\|_{H^2}^2+C\|A\|_{H^2}^6\|w^N\|_{H^2}^2,\quad (n=2), \tag{A.28}$$
$$\frac{d}{dt}\|w^N\|_{H^2}^2+\frac{1}{2}\frac{\gamma}{Re}\|w^N\|_{H^3}^2\le C\|w^N\|_{H^1}^{10}+\frac{Re}{\gamma}\|f\|_{H^1}^2$$
$$+C\|A\|_{H^1}^{1/2}\|A\|_{H^3}^{3/2}\|w^N\|_{H^2}^{3/2}+C\|A\|_{H^2}^8\|w^N\|_{H^1}^2,\quad (n=3). \tag{A.29}$$

Similarly we can obtain the following estimates: for $n=2$,

$$\sup_{t\in[0,T]}\|w^N(t)\|_{H^2}^2\le [CK_2^2T+CKK_2T+C\|f\|_{L^2(0,T;H^1)}^2+\|\mathbf{u}_0\|_{H^2}^2]e^{CK^2T+CK^6T}\triangleq K_7, \tag{A.30}$$

and

$$\int_0^T\|w^N(t)\|_{H^3}^2dt\le K_8, \tag{A.31}$$

where

$$K_8=C\left(CK_3^2T+C\|f\|_{L^2(0,T;H^1)}^2+\|\mathbf{u}_0\|_{H^2}^2+CK^2K_7K_1T+CK^6K_7T\right).$$

While for $n=3$,

$$\sup_{t\in[0,T]}\|w^N(t)\|_{H^2}^2\le K_9, \tag{A.32}$$

where

$$K_9=\left[CK_5^5T+CK^8K_1+\frac{Re}{\gamma}\|f\|_{L^2(0,T;H^1)}^2+\|\mathbf{u}_0\|_{H^2}^2\right]e^{CK^2T}.$$

Further, we have

$$\int_0^T\|w^N(t)\|_{H^3}^2dt\le K_{10}, \tag{A.33}$$

where

$$K_{10}=C\left(CK_5^5T+C\|f\|_{L^2(0,T;H^1)}^2+\|\mathbf{u}_0\|_{H^2}^2+CK^8K_1T+CK^2K_9T\right).$$

This shows for T given,

$$w^N\in L^\infty([0,T];H^2(\Omega))\cap L^2([0,T];H^3(\Omega)) \tag{A.34}$$

provided that $f\in L^2([0,T];H^1(\Omega))$ and $A\in L^\infty([0,T];H^3(\Omega))\cap L^2([0,T];H^4(\Omega))$.

Similarly we can obtain the estimates

$$\frac{d}{dt}\|w^N\|_{H^3}^2+\frac{1}{2}\frac{\gamma}{Re}\|w^N\|_{H^4}^2\le C\|w^N\|_{H^1}^6+\frac{Re}{\gamma}\|f\|_{H^2}^2$$
$$+C\|A\|_{H^2}^9\|w^N\|_{H^1}^2+C\|A\|_{H^3}^4\|w^N\|_{H^1}^2$$
$$+C\|A\|_{H^1}^{4/9}\|A\|_{H^4}^{14/9}\|w^N\|_{H^2}^2,\quad (n=2), \tag{A.35}$$
$$\frac{d}{dt}\|w^N\|_{H^3}^2+\frac{1}{2}\frac{\gamma}{Re}\|w^N\|_{H^4}^2\le C\|w^N\|_{H^1}^{14}+\frac{Re}{\gamma}\|f\|_{H^2}^2$$
$$+C\|A\|_{H^2}^{12}\|w^N\|_{H^1}^2+C\|A\|_{H^3}^4\|w^N\|_{H^1}^2$$
$$+C\|A\|_{H^1}^{1/3}\|A\|_{H^4}^{5/3}\|w^N\|_{H^2}^2,\quad (n=3). \tag{A.36}$$

From (A.35), for $n = 2$ we have

$$\sup_{t\in[0,T]} \|w^N(t)\|^2_{H^3} \le K_{11}, \tag{A.37}$$

where

$$K_{11} = \left[\|u_0\|^2_{H^3} + \frac{Re}{\gamma}\|f\|^2_{L^2(0,T;H^2)} + CK_3^3T + CK^4K_1 + CK^9K_1\right] e^{CK^4T}.$$

Furthermore, we have

$$\int_0^T \|w^N(t)\|^2_{H^4} dt \le K_{12}, \tag{A.38}$$

where

$$K_{12} = C\left(\|u_0\|^2_{H^3} + C\|f\|^2_{L^2(0,T;H^2)} + CK_3^3T + CK^4K_1T + CK^2K_{11}T + CK^2K_7T\right).$$

Thus we know that $\phi_1(K,T) = K_{12} + K_{11}$ for $n = 2$.

From (A.36), for $n = 3$ we have

$$\sup_{t\in[0,T]} \|w^N(t)\|^2_{H^3} \le K_{13}, \tag{A.39}$$

where

$$K_{13} = [\|u_0\|^2_{H^3} + C\|f\|^2_{L^2(0,T;H^2)} + CK_3^7T + CK^{12}K_3T + CK^4K_3T]e^{CK^2T}.$$

Moreover,

$$\int_0^T \|w^N(t)\|^2_{H^4} dt \le K_{14}, \tag{A.40}$$

where

$$K_{14} = C\Big(\|u_0\|^2_{H^3} + C\|f\|^2_{L^2(0,T;H^2)} + CK_3^7T + CK^{12}K_3T + CK^4K_3T + CK^2K_{13}T\Big).$$

Thus we know that $\phi_1(K,T) = K_{14} + K_{13}$ for $n = 3$.

Therefore, this implies that for T given,

$$w^N \in L^\infty([0,T];H^3(\Omega)) \cap L^2([0,T];H^4(\Omega)) \tag{A.41}$$

provided that $f \in L^2([0,T];H^2(\Omega))$ and $A \in L^\infty([0,T],H^3(\Omega)) \cap L^2([0,T],H^4(\Omega))$.

Passing $N \to \infty$, we can obtain the estimate (2.16).

Acknowledgments. The authors are very grateful to the referee for his many valuable suggestions. We are very grateful to Professors Weinan E and Chun Liu for their helpful discussions.

REFERENCES

[1] T. Aubin, *Some Nonlinear Problems in Riemannian Geometry*, Springer, New York, 1998.
[2] A.N. Beris and B.J. Edwards, *Thermodynamics of Flowing Systems with Internal Microstructure*, Oxford Science, New York, 1994.

[3] C.V. Chaubal, A. Srinivasan, Ö. Eğecioğlu, and L.G. Leal, *Smoothed particle hydrodynamics techniques for the solution of kinitic theory problems*, Part 1: Method, J. Non-Newtonian Fluid Mesh., 70 (1997), pp. 125–154.

[4] P. Constantin, I.G. Kevrekids, and E.S. Titi, *Asmptotic states of a smoluchowski equation*, Discrete Continuous Dynamical Systems, 11 (2004), pp. 101–112.

[5] P. Constantin and C. Foias, *Navier-Stokes Equations*, The University of Chicago Press, Chicago and London, 1988.

[6] M. Doi and S.F. Edwards, *The Theory of Polymer Dynamics*, Oxford University Press, Oxford, 1986.

[7] Q. Du, C. Liu, and P. Yu, *FENE dumbbell model and its several linear and nonlinear closure approximations*, Multiscale Model. Simul., 4 (2005), pp. 709–731.

[8] W.N. E and P.W. Zhang, *A molecular kinetic theory of inhomogeneous liquid crystal flow and the small Deborah number limit*, Meth. Appl. Anal., 13 (2006), pp. 181–198.

[9] J. Ericksen, *Liquid crystals with variable degree of orientation*, Arch. Rat. Mech. Anal., 113 (1991), pp. 97–120.

[10] V. Faraoni, M. Grosso, S. Crescitelli, and P.L. Maffettone, *The rigid-rod model for nematic polymers: An analysis of the shear flow problem*, J. Rheol., 43 (1999), pp. 829–843.

[11] P.G. de Gennes and J. Prost, *The Physics of Liquid Crystals*, 2nd ed., Oxford Science Publications, New York, 1993.

[12] C. Guillope and J.C. Saut, *Existence results for the flow of viscoelastic fluids with a differential constitutive law*, Nonl. Anal. TMA, 15 (1990), pp. 849–869.

[13] N. Kuzuu and M. Doi, *Constitutive equation for nematic liquid crystals under weak velocity gradient derived from a molecular kinetic equation*, J. Phys. Soc. Japan, 52 (1983), pp. 3486–3494.

[14] O. Ladyzenskaja, V. Solonnikov, and N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type, Translations of Mathematical Monographs*, Am. Math. Soc., Providence, RI, 1968.

[15] T.J. Li, H. Zhang and P.W. Zhang, *Local existence for the dumbbell model of polymeric fluids*, Comm. Partial Differential Equations, 29 (2004), pp. 903–923.

[16] F.H. Lin, C. Liu, and P. Zhang, *On hydrodynamics of viscoelastic fluids*, Comm. Pure Appl. Math., 58 (2005), pp. 1–35.

[17] F.H. Lin, C. Liu, and P. Zhang, *On a micro-macro model for polymeric fluids near equilibrium*, Comm. Pure Appl. Math., 60 (2006), pp. 838–866.

[18] G. Marrucci and F. Greco, *The elastic constants of Maier-Saupe rodlike molecule nematics*, Mol. Cryst. Liq. Cryst., 206 (1991), pp. 17–30.

[19] M. Renardy, *Local existence of solutions of the Dirichlet initial boundary problem for incompressible hypoelastic materials*, SIAM J. Math. Anal., 21 (1990), pp. 1369–1385.

[20] M. Renardy, *An existence theorem for model equations resulting from kinetic theories*, SIAM J. Math. Anal., 22 (1991), pp. 313–327.

[21] A.D. Rey and T. Tsuji, *Orientation mode selection mechanisms for sheared nematic liquid crystalline materials*, Phys. Rev. E, 57 (1998), pp. 5610–5625.

[22] A.D. Rey and T. Tsuji, *Recent advances in theoretical liquid crystal rheology*, Macromol. Theory Simul., 7 (1998), pp. 623–639.

[23] T.C. Sideris and B. Thomases, *Global existence for 3D incompressible isotropic elastodynamics via the incompressible limit*, Comm. Pure Appl. Math., 58 (2005), pp. 750–788.

[24] R. Teman, *Navier-Stokes Equations and Nonlinear Functional Analysis*, SIAM, Philadelphia, 1995.

[25] Q. Wang, *A hydrodynamic theory for solutions of nonhomogeneous nematic liquid crystalline polymers of different configurations*, J. Chem. Phys., 116 (2002), pp. 9120–9136.

[26] H.J. Yu and P.W. Zhang, *A kinetic-hydrodynamic simulation of microstructure of liquid crystal polymers in plane shear flow*, J. Non-Newtonian Fluid Mech., 141 (2007), pp. 116–127.

[27] H.J. Yu, G.H. Ji, and P.W. Zhang, *A nonhomogeneous kinetic model of liquid crystal polymers and its thermodynamic closure approximation*, preprint.

[28] H. Zhang and P.W. Zhang, *A theoretical and numerical study for the rod-like model of a polymeric fluid*, J. Comp. Math., 22 (2004), pp. 319–330.

SIAM J. MATH. ANAL.
Vol. 40, No. 3, pp. 1272–1289

A DESCRIPTION OF SEISMIC ACOUSTIC WAVE PROPAGATION IN POROUS MEDIA VIA HOMOGENIZATION*

ANVARBEK MEIRMANOV†

Abstract. We consider a linear system of differential equations describing the joint motion of an elastic porous body and a fluid occupying the porous space. A rigorous justification is performed for the homogenization procedures under various conditions imposed on the physical parameters as the dimensionless size of the pores tends to zero, while the porous body is geometrically periodic and the process's characteristic time is sufficiently small. Such models describe the propagation of seismic acoustic waves. In the present paper, we derive the homogenized equations, which are different types of nonstandard wave equations depending on the relations between the physical parameters. The proofs are based on Nguetseng's two-scale convergence method of homogenization in periodic structures.

Key words. Stokes equations, Lamé's equations, wave equation, two-scale convergence, homogenization of periodic structures

AMS subject classifications. 35M20, 74F10, 76S05

DOI. 10.1137/070697483

1. Introduction. In the present paper, we deal with a problem of joint motion of a deformable solid (the) perforated by a system of channels or pores (the) and a fluid occupying the pore space. In a domain $\Omega \subset \mathbf{R}^3$, the dimensionless displacement vector $\boldsymbol{w}$ of the continuum medium in the dimensionless variables

$$\boldsymbol{x}' = L\boldsymbol{x}, \quad t' = \tau t, \quad \boldsymbol{w}' = \frac{L^2}{g\tau^2}\boldsymbol{w}, \quad \rho_s' = \rho_0\rho_s, \quad \rho_f' = \rho_0\rho_f, \quad \boldsymbol{F}' = g\boldsymbol{F}$$

satisfies the differential equation

$$\bar{\rho}\frac{\partial^2 \boldsymbol{w}}{\partial t^2} = \mathrm{div}P + \bar{\rho}\boldsymbol{F}, \tag{1.1}$$

where

$$P = \bar{\chi}P^f + (1-\bar{\chi})P^s, \tag{1.2}$$

$$P^f = \alpha_\mu D\left(x, \frac{\partial \boldsymbol{w}}{\partial t}\right) - \left(p_f - \alpha_\nu \mathrm{div}\frac{\partial \boldsymbol{w}}{\partial t}\right) I, \tag{1.3}$$

$$P^s = \alpha_\lambda D(x, \boldsymbol{w}) + \alpha_\eta(\mathrm{div}\boldsymbol{w})I, \tag{1.4}$$

$$p_f + \bar{\chi}\alpha_p \mathrm{div}\boldsymbol{w} = 0. \tag{1.5}$$

Hereafter, we use the notation

$$D(x, \boldsymbol{u}) = (1/2)\left(\nabla_x \boldsymbol{u} + (\nabla_x \boldsymbol{u})^T\right), \quad \bar{\rho} = \bar{\chi}\rho_f + (1-\bar{\chi})\rho_s,$$

*Received by the editors July 17, 2007; accepted for publication (in revised form) June 24, 2008; published electronically October 31, 2008. This work was partially supported by RFBR grant 08-05-00265.

http://www.siam.org/journals/sima/40-3/69748.html

†Mathematics Department, Belgorod State University, ul.Pobedi 85, 308015 Belgorod, Russia (meirmanov@bsu.edu.ru).

where I is the unit tensor, the given function $\bar{\chi}(\boldsymbol{x})$ is the characteristic function of the pore space, the given function $\boldsymbol{F}(\boldsymbol{x},t)$ is the dimensionless vector of distributed mass forces, P^f is the liquid stress tensor, P^s is the stress tensor in the solid skeleton, and p_f is the liquid pressure.

Equations (1.1)–(1.5) mean that the displacement vector $\boldsymbol{w}$ satisfies the Stokes equations in the pore space Ω_f and the Lamé equations in the solid skeleton Ω_s.

On the "solid skeleton–pore space" common boundary Γ, the displacement vector $\boldsymbol{w}$ and the liquid pressure p_f satisfy the usual continuity condition

$$[\boldsymbol{w}](\boldsymbol{x}_0,t)=0,\quad \boldsymbol{x}_0\in\Gamma,\ t\geq 0, \tag{1.6}$$

and the momentum conservation law in the form

$$[P\cdot\mathbf{n}](\boldsymbol{x}_0,t)=0,\quad \boldsymbol{x}_0\in\Gamma,\ t\geq 0, \tag{1.7}$$

where $\mathbf{n}(\boldsymbol{x}_0)$ is the unit normal to the boundary at the point $\boldsymbol{x}_0\in\Gamma$ and

$$[\varphi](\boldsymbol{x}_0,t)=\varphi_{(s)}(\boldsymbol{x}_0,t)-\varphi_{(f)}(\boldsymbol{x}_0,t),$$
$$\varphi_{(s)}(\boldsymbol{x}_0,t)=\lim_{\substack{\boldsymbol{x}\to\boldsymbol{x}_0\\ \boldsymbol{x}\in\Omega_s}}\varphi(\boldsymbol{x},t),\quad \varphi_{(f)}(\boldsymbol{x}_0,t)=\lim_{\substack{\boldsymbol{x}\to\boldsymbol{x}_0\\ \boldsymbol{x}\in\Omega_f}}\varphi(\boldsymbol{x},t).$$

The problem is endowed with the homogeneous initial and boundary conditions

$$\boldsymbol{w}(\boldsymbol{x},0)=0,\quad \frac{\partial\boldsymbol{w}}{\partial t}(\boldsymbol{x},0)=0,\quad \boldsymbol{x}\in\Omega, \tag{1.8}$$

$$\boldsymbol{w}(\boldsymbol{x},t)=0,\quad \boldsymbol{x}\in S=\partial\Omega,\quad t\geq 0. \tag{1.9}$$

The dimensionless constants α_i $(i=\tau,\nu,\ldots)$ are defined by the formulas

$$\alpha_\mu=\frac{2\mu\tau}{L^2\rho_0},\quad \alpha_\lambda=\frac{2\lambda\tau^2}{L^2\rho_0},\quad \alpha_\nu=\frac{\nu\tau}{L^2\rho_0},$$

$$\alpha_p=\rho_f c_f^2\frac{\tau^2}{L^2},\quad \alpha_\eta=\frac{\eta\tau^2}{L^2\rho_0}=\rho_s c_s^2\frac{\tau^2}{L^2},$$

where μ is the fluid viscosity, ν is the bulk fluid viscosity, λ and η are elastic Lamé's constants, c_f is the speed of sound in fluids, c_s is the speed of sound in solids, L is the characteristic size of the domain under study, τ is the characteristic time of the process, ρ_f and ρ_s are the respective mean dimensionless densities of the liquid and solid phases correlated with the mean density of water ρ_0, and g is the value of acceleration due to gravity.

The corresponding mathematical model described by system (1.1)–(1.9) is commonly used (see [2], [9]) and contains a natural small parameter ε, which is the pore characteristic size l divided by the characteristic size L of the entire porous body:

$$\varepsilon=\frac{l}{L}.$$

Our aim is to derive all possible limiting regimes (the homogenized equations) as $\varepsilon\searrow 0$. Such an approximation significantly simplifies the original problem and at the same time preserves all of its main features. But even this approach is too difficult

to be realized, and some additional simplifying assumptions are necessary. In terms of geometrical properties of the medium, it is most expedient to simplify the problem by postulating that the porous structure is periodic.

We impose the following constraints.

1.

(1) The domain $\Omega = (0,1)^3$ is a periodic repetition of an elementary cell $Y^\varepsilon = \varepsilon Y$, where $Y = (0,1)^3$ and the quantity $1/\varepsilon$ is an integer so that Ω always contains an integer number of elementary cells Y^ε.

(2) Let Y_s be the "solid part" of Y, and let the "liquid part" Y_f of Y be its open complement. We write $\gamma = \partial Y_f \cap \partial Y_s$ and assume that γ is a Lipschitz continuous surface.

(3) The pore space Ω_f^ε is a periodic repetition of the elementary cell εY_f, and the solid skeleton Ω_s^ε is a periodic repetition of the elementary cell εY_s. The Lipschitz continuous boundary $\Gamma^\varepsilon = \partial\Omega_s^\varepsilon \cap \partial\Omega_f^\varepsilon$ is a periodic repetition in Ω of the boundary $\varepsilon\gamma$.

(4) Ω_s^ε Ω_f^ε

Here the essential assumptions are those last three on the geometry of the elementary cells Y_s and Y_f and the domains Ω_s^ε and Ω_f^ε. As for the first assumption, we take the simplest structure of Ω (namely, the cube) just to simplify the procedure. In principle, for the domain Ω we can choose any bounded domain with a Lipschitz continuous boundary $S = \partial\Omega$.

Under these assumptions, we have

$$\bar{\chi}(\boldsymbol{x}) = \chi^\varepsilon(\boldsymbol{x}) = \chi\left(\frac{\boldsymbol{x}}{\varepsilon}\right),$$

$$\bar{\rho} = \rho^\varepsilon(\boldsymbol{x}) = \chi^\varepsilon(\boldsymbol{x})\rho_f + (1 - \chi^\varepsilon(\boldsymbol{x}))\rho_s,$$

where $\chi(\boldsymbol{y})$ is the characteristic function of Y_f in Y.

We assume that all dimensionless parameters depend on the small parameter ε and the (finite or infinite) limits exist:

$$\lim_{\varepsilon\searrow 0} \alpha_\mu(\varepsilon) = \mu_0, \quad \lim_{\varepsilon\searrow 0} \alpha_\lambda(\varepsilon) = \lambda_0, \quad \lim_{\varepsilon\searrow 0} \alpha_\nu(\varepsilon) = \nu_0,$$

$$\lim_{\varepsilon\searrow 0} \alpha_\eta(\varepsilon) = \eta_0, \quad \lim_{\varepsilon\searrow 0} \alpha_p(\varepsilon) = p_*,$$

$$\lim_{\varepsilon\searrow 0} \frac{\alpha_\mu}{\varepsilon^2} = \mu_1, \quad \lim_{\varepsilon\searrow 0} \frac{\alpha_\lambda}{\varepsilon^2} = \lambda_1.$$

The first research aiming to find the limiting regimes in the case where the skeleton was an absolutely rigid body was carried out by Sanchez-Palencia and Tartar. Sanchez-Palencia [9, sect. 7.2] formally obtained Darcy's law of filtration using the method of two-scale asymptotic expansions, and Tartar [9, Appendix] rigorously justified the homogenization procedure. Using the same method of two-scale expansions, Burridge and Keller [2] formally derived a system of Biot's equations from problem (1.1)–(1.9) in the case where the parameter α_μ was of order ε^2 and the rest of the coefficients were fixed independent of ε. Under the same assumptions as in [2], a rigorous justification of Biot's model was given by Nguetseng [8] and later by Clopeaut et al. [3]. The most general case of problem (1.1)–(1.9) where

$$\mu_0,\ \lambda_0^{-1},\ \nu_0,\ p_*^{-1},\ \eta_0^{-1} < \infty$$

was studied in [6].

All these authors used Nguetseng's two-scale convergence method [7, 5].

In the present paper, we use the same method to investigate all possible limiting regimes in problem (1.1)–(1.9) in the cases where

$$\nu_0,\, p_*,\, \eta_0 < \infty; \quad \mu_0 = \lambda_0 = 0, \quad 0 < p_*,\, \eta_0.$$

These cases correspond to the seismic acoustic wave propagation, where all the processes on distances of tens of thousands of meters ($L \nearrow \infty$) come to an end in several seconds ($\tau \searrow 0$).

We show that the homogenized equations are different types of nonstandard wave equations for a two- or one-velocity continuum (Theorem 2.2).

This is a very interesting fact: initially a one-velocity continuum becomes a two-velocity continuum after the homogenization procedure, which appears to be the result of different smoothness of the solution in the solid and liquid components:

$$\int_\Omega \alpha_\mu(\varepsilon)\chi^\varepsilon |\nabla \boldsymbol{w}^\varepsilon|^2 dx \le C_0, \quad \int_\Omega \alpha_\lambda(\varepsilon)(1-\chi^\varepsilon)|\nabla \boldsymbol{w}^\varepsilon|^2 dx \le C_0,$$

where C_0 is a constant independent of the small parameter ε. To preserve the best properties of the solution, we must use the well-known extension lemma [1, 4] and extend the solution from the solid part to the liquid part and conversely. At this stage, the criteria μ_1 and λ_1 become crucial. Namely, let $\boldsymbol{w}_f^\varepsilon$ ($\boldsymbol{w}_s^\varepsilon$) be an extension of the liquid (solid) displacements to the solid (liquid) part, and let $\mu_1 = \lambda_1 = \infty$. Then the limiting (homogenized) system describes the one-velocity continuum. This is because of the fact that each of the sequences $\{\boldsymbol{w}^\varepsilon\}$, $\{\boldsymbol{w}_f^\varepsilon\}$, and $\{\boldsymbol{w}_s^\varepsilon\}$ two-scale converges to a function independent of the fast variable. This statement easily follows from Nguetseng's theorem.

If $\mu_1 < \infty$ and $\lambda_1 = \infty$ or $\mu_1 = \infty$ and $\lambda_1 < \infty$, then the homogenized systems describe the two-velocity continuum.

Finally, we note that, in practice, to solve a real physical problem in, for example, acoustics, one does not want to carry out the limiting procedure but, instead, wants to find a simple and reliable mathematical model describing the process. But there is only one exact (sufficiently reliable) mathematical model (1.1)–(1.9) with given physical constants (densities, viscosities, etc.), the characteristic size L of the physical domain under study, and the characteristic time τ of the physical process. The small parameter ε and the dimensionless quantities α_μ, α_λ, α_p, ... are functions of them. Changing the values of L and τ within some reasonable limits, one may find some rules for the behavior of the dimensionless quantities as the small parameter tends to zero. All possible limits of these quantities are described by conditions on μ_0, λ_0, μ_1, ... and, as was mentioned above, each homogenized system corresponds to a given combination of them. Thus, for a given physical situation, there exists a combination of dimensionless criteria, which would suggest the choice of the form of the homogenized system for obtaining the exact mathematical model. Therefore, to find all possible homogenized systems is very important from both mathematical and practical standpoints.

2. Main results. There are various equivalent (in the sense of distributions) forms of representation of (1.1) and boundary conditions (1.6)–(1.7). In what follows, it is convenient to write them in the form of the integral identities.

We say that four functions $\left(\boldsymbol{w}^\varepsilon, p_f^\varepsilon, p_s^\varepsilon, q^\varepsilon\right)$ are a generalized solution of problem (1.1)–(1.9) if they satisfy the regularity conditions

$$\boldsymbol{w}^\varepsilon,\ \nabla\boldsymbol{w}^\varepsilon,\ p_f^\varepsilon,\ p_s^\varepsilon,\ q^\varepsilon \in L^2(\Omega_T) \tag{2.1}$$

in the domain $\Omega_T = \Omega \times (0,T)$, boundary condition (1.9) in the trace sense, (1.5) and the equations

$$p_s^\varepsilon + (1-\chi^\varepsilon)\alpha_\eta \mathrm{div}\boldsymbol{w}^\varepsilon = 0, \tag{2.2}$$

$$q^\varepsilon = p_f^\varepsilon + \frac{\alpha_\nu}{\alpha_p}\frac{\partial p_f^\varepsilon}{\partial t} \tag{2.3}$$

a.e. in Ω_T, and, finally, the integral identity

$$\left.\begin{aligned}\int_{\Omega_T}\Bigg(\rho^\varepsilon \boldsymbol{w}^\varepsilon\cdot\frac{\partial^2\boldsymbol{\varphi}}{\partial t^2} - \chi^\varepsilon\alpha_\mu D(\boldsymbol{x},\boldsymbol{w}^\varepsilon) : D\left(x,\frac{\partial\boldsymbol{\varphi}}{\partial t}\right) - \rho^\varepsilon \boldsymbol{F}\cdot\boldsymbol{\varphi}\\ + \{(1-\chi^\varepsilon)\alpha_\lambda D(x,\boldsymbol{w}^\varepsilon) - (q^\varepsilon + p_s^\varepsilon)I\} : D(x,\boldsymbol{\varphi})\Bigg)\,dxdt = 0\end{aligned}\right\} \tag{2.4}$$

for all smooth vector-functions $\boldsymbol{\varphi} = \boldsymbol{\varphi}(\boldsymbol{x},t)$ such that $\boldsymbol{\varphi}|_{\partial\Omega} = \boldsymbol{\varphi}|_{t=T} = \partial\boldsymbol{\varphi}/\partial t|_{t=T} = 0$.

In this definition, we changed the form of representation of the stress tensor P in the integral identity (2.4) by introducing two new unknown functions, q^ε and p_s^ε, which in a certain sense have the meaning of pressure. In what follows, we call functions q^ε and p_s^ε the liquid and the solid pressure, respectively, and regard (2.3) as the state equation and equations (1.5) and (2.2) as the continuity equations. This special choice of the continuity and state equations simplifies the use of the homogenization procedure.

In (2.4), by $A : B$ we denote the convolution (or, equivalently, the inner tensor product) of two second-rank tensors along the both indices, i.e., $A : B = \mathrm{tr}\,(B^* \circ A) = \sum_{i,j=1}^3 A_{ij}B_{ji}$.

Theorems 2.1–2.2 are the main results of the paper.

THEOREM 2.1. [illegible] $\boldsymbol{F}$ [illegible] $L^2(\Omega)$ [illegible] $\varepsilon > 0$ [illegible] $[0,T]$ [illegible] (1.1)–(1.9)

$$\max_{0\le t\le T}\left\|\left|\frac{\partial\boldsymbol{w}^\varepsilon}{\partial t}\right| + \sqrt{\alpha_\mu}\chi^\varepsilon|\nabla\boldsymbol{w}^\varepsilon| + (1-\chi^\varepsilon)\sqrt{\alpha_\lambda}|\nabla\boldsymbol{w}^\varepsilon|\right\|_{2,\Omega}(t) \le C_0, \tag{2.5}$$

$$\|q^\varepsilon\|_{2,\Omega_T} + \|p_f^\varepsilon\|_{2,\Omega_T} + \|p_s^\varepsilon\|_{2,\Omega_T} \le C_0, \tag{2.6}$$

[illegible] C_0 [illegible] ε

THEOREM 2.2. [illegible] 2.1 [illegible] $\{\varepsilon > 0\}$ [illegible] $\boldsymbol{w}_f^\varepsilon, \boldsymbol{w}_s^\varepsilon \in L^\infty(0,T; W_2^1(\Omega))$ [illegible]

$$\boldsymbol{w}_f^\varepsilon = \boldsymbol{w}^\varepsilon \text{ [illegible] } \Omega_f^\varepsilon \times (0,T), \qquad \boldsymbol{w}_s^\varepsilon = \boldsymbol{w}^\varepsilon \text{ [illegible] } \Omega_s^\varepsilon \times (0,T),$$

[illegible] $\{p_f^\varepsilon\}$, $\{q^\varepsilon\}$, $\{p_s^\varepsilon\}$, $\{\boldsymbol{w}^\varepsilon\}$, $\{\chi^\varepsilon\boldsymbol{w}^\varepsilon\}$, $\{(1-\chi^\varepsilon)\boldsymbol{w}^\varepsilon\}$, $\{\boldsymbol{w}_f^\varepsilon\}$ [illegible] $\{\boldsymbol{w}_s^\varepsilon\}$ [illegible] $\varepsilon \searrow 0$ [illegible] $L^2(\Omega_T)$ [illegible] p_f, q, p_s, $\boldsymbol{w}$, $\boldsymbol{w}^f$, $\boldsymbol{w}^s$, $\boldsymbol{w}_f$, [illegible] $\boldsymbol{w}_s$ [illegible]

(I) $\mu_1 = \lambda_1 = \infty$, $\boldsymbol{w}_f = \boldsymbol{w}_s = \boldsymbol{w}$ Ω_T $\boldsymbol{w}$, p_f, q p_s

$$\hat{\rho}\frac{\partial^2 \boldsymbol{w}}{\partial t^2} = -\frac{1}{m}\nabla q + \hat{\rho}\boldsymbol{F}, \tag{2.7}$$

$$\frac{1}{p_*}p_f + \frac{1}{\eta_0}p_s + \mathrm{div}\boldsymbol{w} = 0, \tag{2.8}$$

$$q = p_f + \frac{\nu_0}{p_*}\frac{\partial p_f}{\partial t}, \quad \frac{1}{m}q = \frac{1}{1-m}p_s, \tag{2.9}$$

$$\boldsymbol{w}(\boldsymbol{x}, 0) = \frac{\partial \boldsymbol{w}}{\partial t}(\boldsymbol{x}, 0) = 0, \quad \boldsymbol{x} \in \Omega, \tag{2.10}$$

$$\boldsymbol{w}(\boldsymbol{x}, t) \cdot \boldsymbol{n}(\boldsymbol{x}) = 0, \quad \boldsymbol{x} \in S,\ t > 0, \tag{2.11}$$

$\hat{\rho} = m\rho_f + (1-m)\rho_s$ $m = \int_Y \chi dy$

(II) $\mu_1 = \infty$ $\lambda_1 < \infty$ Ω_T $\boldsymbol{w}^f = m\boldsymbol{w}_f$, $\boldsymbol{w}^s$, p_f, q p_s (2.9)

$$\rho_f m\frac{\partial^2 \boldsymbol{w}_f}{\partial t^2} + \rho_s\frac{\partial^2 \boldsymbol{w}^s}{\partial t^2} = -\frac{1}{m}\nabla q + \hat{\rho}\boldsymbol{F} \tag{2.12}$$

$$\frac{1}{p_*}p_f + \frac{1}{\eta_0}p_s + m\mathrm{div}\boldsymbol{w}_f + \mathrm{div}\boldsymbol{w}^s = 0, \tag{2.13}$$

$$\frac{\partial \boldsymbol{w}^s}{\partial t} = (1-m)\frac{\partial \boldsymbol{w}_f}{\partial t} + \int_0^t B_1^s(t-\tau)\cdot \boldsymbol{z}^s(\boldsymbol{x}, \tau)d\tau, \tag{2.14}$$

$$\boldsymbol{z}^s(\boldsymbol{x}, t) = -\frac{1}{m}\nabla q(\boldsymbol{x}, t) + \rho_s \boldsymbol{F}(\boldsymbol{x}, t) - \rho_s\frac{\partial^2 \boldsymbol{w}_f}{\partial t^2}(\boldsymbol{x}, t),$$

$\lambda_1 > 0$

$$\rho_s\frac{\partial^2 \boldsymbol{w}^s}{\partial t^2} = \rho_s B_2^s \cdot \frac{\partial^2 \boldsymbol{w}_f}{\partial t^2} + ((1-m)I - B_2^s)\cdot\left(-\frac{1}{m}\nabla q + \rho_s \boldsymbol{F}\right) \tag{2.15}$$

$\lambda_1 = 0$ (2.9), (2.12)–(2.15) (2.10) (2.11) $\boldsymbol{w} = m\boldsymbol{w}_f + \boldsymbol{w}^s$

(2.14)–(2.15) $B_1^s(t)$ B_2^s (5.37) (5.39) $((1-m)I - B_2^s)$

(III) $\mu_1 < \infty$ $\lambda_1 = \infty$ Ω_T $\boldsymbol{w}^f$, $\boldsymbol{w}^s = (1-m)\boldsymbol{w}_s$, p_f, q p_s (2.9)

$$\rho_f \frac{\partial^2 \boldsymbol{w}^f}{\partial t^2} + \rho_s(1-m)\frac{\partial^2 \boldsymbol{w}_s}{\partial t^2} = -\frac{1}{m}\nabla q + \hat{\rho}\boldsymbol{F} \tag{2.16}$$

$$\frac{1}{p_*}p_f + \frac{1}{\eta_0}p_s + \mathrm{div}\boldsymbol{w}^f + (1-m)\mathrm{div}\boldsymbol{w}_s = 0, \tag{2.17}$$

$$\frac{\partial \boldsymbol{w}^f}{\partial t} = m\frac{\partial \boldsymbol{w}_s}{\partial t} + \int_0^t B_1^f(t-\tau)\cdot \boldsymbol{z}^f(\boldsymbol{x},\tau)d\tau, \tag{2.18}$$

$$\boldsymbol{z}^f(\boldsymbol{x},t) = -\frac{1}{m}\nabla q(\boldsymbol{x},t) + \rho_f \boldsymbol{F}(\boldsymbol{x},t) - \rho_f \frac{\partial^2 \boldsymbol{w}_s}{\partial t^2}(\boldsymbol{x},t),$$

$\mu_1 > 0$

$$\rho_f \frac{\partial^2 \boldsymbol{w}^f}{\partial t^2} = \rho_f B_2^f \cdot \frac{\partial^2 \boldsymbol{w}_s}{\partial t^2} + (mI - B_2^f)\cdot\left(-\frac{1}{m}\nabla q + \rho_f \boldsymbol{F}\right) \tag{2.19}$$

$\mu_1 = 0$ (2.9), (2.16)–(2.19) (2.10) (2.11) $\boldsymbol{w} = \boldsymbol{w}^f + (1-m)\boldsymbol{w}_s$

(2.18)–(2.19) $B_1^f(t)$ B_2^f (5.44)–(5.45) $(mI - B_2^f)$

(IV) $\mu_1 < \infty$ $\lambda_1 < \infty$ Ω_T $\boldsymbol{w}$, p_f, q p_s (2.8) (2.9)

$$\frac{\partial \boldsymbol{w}}{\partial t} = \int_0^t B(t-\tau)\cdot\nabla q(\boldsymbol{x},\tau)d\tau + \boldsymbol{f}(\boldsymbol{x},t), \tag{2.20}$$

$B(t)$ $\boldsymbol{f}(\boldsymbol{x},t)$ (5.58) (5.59) (2.8), (2.9), (2.20) (2.10) (2.11)

As was mentioned above, even in the most simple case (I) with $\nu_0 = 0$, Theorem 2.2 gives the standard wave equation for the solid pressure p_s but with a completely new speed of sound in the mixture, which includes the porosity, densities, and speeds of sound in the solid and liquid components.

In the next simple case (IV) with $\nu_0 = 0$, Theorem 2.2 gives a new wave equation for the solid pressure in the form

$$\frac{\partial p_s}{\partial t} = \int_0^t \mathrm{div}\big(\widetilde{B}(t-\tau)\cdot\nabla p_s(\boldsymbol{x},\tau)\big)d\tau. \tag{2.21}$$

Here $\widetilde{B}(0) = c^2 I$, where the time derivative of the matrix $\widetilde{B}(t)$ is generally unbounded at $t = 0$. This equation has no simple solutions like traveling waves and requires a special analysis even for the smooth matrix $\widetilde{B}(t)$.

The rest of the homogenized models described by Theorem 2.2 are much more complicated than the model (2.21). This is natural, because one cannot expect that a simple model gives an "accurate" approximation of the very complicated original model (1.1)–(1.9).

3. Preliminaries.

3.1. Two-scale convergence. The justification of Theorem 2.2 is based on a systematic use of the two-scale convergence method, which was proposed by Nguetseng [7] and has been recently used in a wide range of homogenization problems (see, for example, the survey [5]).

DEFINITION 3.1. $\{w^\varepsilon\} \subset L^2(\Omega_T)$ $W \in L^2(\Omega_T \times Y)$

$$\lim_{\varepsilon \searrow 0} \int_{\Omega_T} w^\varepsilon(\boldsymbol{x},t)\, \sigma\left(\boldsymbol{x},t,\frac{\boldsymbol{x}}{\varepsilon}\right) d\boldsymbol{x}dt = \int_{\Omega_T}\int_Y W(\boldsymbol{x},t,\boldsymbol{y})\, \sigma(\boldsymbol{x},t,\boldsymbol{y}) d\boldsymbol{y}d\boldsymbol{x}dt \tag{3.1}$$

$\sigma = \sigma(\boldsymbol{x},t,\boldsymbol{y}) \in C^\infty(\Omega_T \times Y)$ $\boldsymbol{y}$ Ω_T

The existence and the main properties of weakly convergent sequences are established by the following fundamental theorem [7, 5].

THEOREM 3.2 (Nguetseng's theorem). 1. $L^2(\Omega_T)$ $W \in L^2(\Omega_T \times Y)$

2. $\{w^\varepsilon\}$ $\{\varepsilon \nabla_x w^\varepsilon\}$ $L^2(\Omega_T)$ $W = W(\boldsymbol{x},t,\boldsymbol{y})$ $\boldsymbol{y}$ $\{w^\varepsilon\}$ W, $\nabla_y W \in L^2(\Omega_T \times Y)$ $\{w^\varepsilon\}$ $\{\varepsilon \nabla_x w^\varepsilon\}$ W $\nabla_y W$

COROLLARY 3.3. $\sigma \in L^2(Y)$ $\sigma^\varepsilon(\boldsymbol{x}) = \sigma(\boldsymbol{x}/\varepsilon)$ $\{w^\varepsilon\} \subset L^2(\Omega_T)$ $W \in L^2(\Omega_T \times Y)$ $\{\sigma^\varepsilon w^\varepsilon\}$ σW

3.2. An extension lemma. A typical difficulty in homogenization problems like problem (1.1)–(1.7) arises in passing to the limit as $\varepsilon \searrow 0$ because of the fact that the bounds on the displacement gradient $\nabla \boldsymbol{w}^\varepsilon$ may be different in the liquid and solid components. The classical approach to overcoming this difficulty consists in constructing an extension of the displacement field defined merely on Ω_s or Ω_f to the whole Ω. The following lemma is valid due to the well-known results from [1, 4, 8]. We formulate it in the form convenient for us.

LEMMA 3.4. 1 $\boldsymbol{w}^\varepsilon \in \overset{\circ}{W}{}_2^1(\Omega)$ $\boldsymbol{w}_f^\varepsilon, \boldsymbol{w}_s^\varepsilon \in W_2^1(\Omega)$ Ω_f^ε Ω_s^ε w^ε

$$\chi^\varepsilon(\boldsymbol{x})(\boldsymbol{w}_f^\varepsilon(\boldsymbol{x}) - \boldsymbol{w}^\varepsilon(\boldsymbol{x})) = 0, \quad (1-\chi^\varepsilon(\boldsymbol{x}))(\boldsymbol{w}_s^\varepsilon(\boldsymbol{x}) - \boldsymbol{w}^\varepsilon(\boldsymbol{x})) = 0, \quad \boldsymbol{x} \in \Omega, \tag{3.2}$$

$$\|\boldsymbol{w}_i^\varepsilon\|_{2,\Omega} \le C\|\boldsymbol{w}^\varepsilon\|_{2,\Omega_i^\varepsilon}, \quad \|D(x,\boldsymbol{w}_i^\varepsilon)\|_{2,\Omega} \le C\|D(x,\boldsymbol{w}^\varepsilon)\|_{2,\Omega_i^\varepsilon},\ i = f,s, \tag{3.3}$$

C Y ε

3.3. Some notation. Further we denote the following:
(1)

$$\langle\Phi\rangle_Y = \int_Y \Phi dy, \quad \langle\Phi\rangle_{Y_f} = \int_Y \chi\Phi dy, \quad \langle\Phi\rangle_{Y_s} = \int_Y (1-\chi)\Phi dy.$$

(2) If $\mathbf{a}$ and $\mathbf{b}$ are two vectors, then the matrix $\mathbf{a}\otimes\mathbf{b}$ is defined by the formula

$$(\mathbf{a}\otimes\mathbf{b})\cdot\mathbf{c} = \mathbf{a}(\mathbf{b}\cdot\mathbf{c})$$

for any vector $\mathbf{c}$.

4. Proof of Theorem 2.1. Estimates (2.5)–(2.6) follow from the energy equality in the form

$$\frac{d}{dt}\left\{\int_\Omega \rho^\varepsilon\left(\frac{\partial \boldsymbol{w}^\varepsilon}{\partial t}\right)^2 + \alpha_\lambda\int_\Omega (1-\chi^\varepsilon)D(x,\boldsymbol{w}^\varepsilon) : D(x,\boldsymbol{w}^\varepsilon)dx\right.$$
$$\left. + \alpha_p\int_\Omega \chi^\varepsilon(\mathrm{div}\boldsymbol{w}^\varepsilon)^2 dx + \alpha_\eta\int_\Omega (1-\chi^\varepsilon)(\mathrm{div}\boldsymbol{w}^\varepsilon)^2 dx\right\} + \alpha_\nu\int_\Omega \chi^\varepsilon\left(\mathrm{div}\frac{\partial \boldsymbol{w}^\varepsilon}{\partial t}\right)^2 dx$$
$$+ \alpha_\mu\int_\Omega \chi^\varepsilon D\left(x,\frac{\partial \boldsymbol{w}^\varepsilon}{\partial t}\right) : D\left(x,\frac{\partial \boldsymbol{w}^\varepsilon}{\partial t}\right)dx = \int_\Omega \rho^\varepsilon\frac{\partial \boldsymbol{F}}{\partial t}\cdot\frac{\partial \boldsymbol{w}^\varepsilon}{\partial t}dx \tag{4.1}$$

if we use Hölder, Gronwall, and Korn inequalities and extension Lemma 3.4. In turn, the energy equality (4.1) follows from (1.1) if we express the stress tensor P and the liquid pressure p_f using state equations (1.2)–(1.4) and continuity equation (1.5), multiply the result by $\partial\boldsymbol{w}^\varepsilon/\partial t$, and integrate by parts. Note that all terms on the "solid skeleton–pore space" interface Γ^ε disappear due to boundary conditions (1.6)–(1.7).

The same estimates (2.5)–(2.6) guarantee the existence and uniqueness of the generalized solution for problem (1.1)–(1.9).

5. Proof of Theorem 2.2.

5.1. Weak and two-scale limits of sequences of displacements and pressures. First, we use Lemma 3.4 and conclude that there are functions $\boldsymbol{w}_f^\varepsilon$, $\boldsymbol{w}_s^\varepsilon \in L^\infty(0,T;W_2^1(\Omega))$ such that

$$\boldsymbol{w}_f^\varepsilon = \boldsymbol{w}^\varepsilon \text{ in } \Omega_f^\varepsilon\times(0,T), \quad \boldsymbol{w}_s^\varepsilon = \boldsymbol{w}^\varepsilon \text{ in } \Omega_s^\varepsilon\times(0,T).$$

By Theorem 2.1, the sequences $\{p_f^\varepsilon\}$, $\{q^\varepsilon\}$, $\{p_s^\varepsilon\}$, $\{\boldsymbol{w}^\varepsilon\}$, $\{\boldsymbol{w}_f^\varepsilon\}$, $\{\sqrt{\alpha_\mu}\nabla\boldsymbol{w}_f^\varepsilon\}$, $\{\boldsymbol{w}_s^\varepsilon\}$, and $\{\sqrt{\alpha_\lambda}\nabla\boldsymbol{w}_s^\varepsilon\}$ are bounded in $L^2(\Omega_T)$. Hence there exists a subsequence of small parameters $\{\varepsilon>0\}$ and functions p_f, q, p_s, $\boldsymbol{w}$, $\boldsymbol{w}_f$, and $\boldsymbol{w}_s$ such that

$$p_f^\varepsilon \rightharpoonup p_f, \quad q^\varepsilon \rightharpoonup q, \quad p_s^\varepsilon \rightharpoonup p_s, \quad \boldsymbol{w}^\varepsilon \rightharpoonup \boldsymbol{w}, \quad \boldsymbol{w}_f^\varepsilon \rightharpoonup \boldsymbol{w}_f, \quad \boldsymbol{w}_s^\varepsilon \rightharpoonup \boldsymbol{w}_s \tag{5.1}$$

weakly in $L^2(\Omega_T)$ as $\varepsilon\searrow 0$.

Note also that

$$(1-\chi^\varepsilon)\alpha_\lambda D(x,\boldsymbol{w}_s^\varepsilon)\to 0, \quad \chi^\varepsilon\alpha_\mu D(x,\boldsymbol{w}_f^\varepsilon)\to 0 \tag{5.2}$$

strongly in $L^2(\Omega_T)$ as $\varepsilon\searrow 0$.

Relabeling if necessary, we assume that the sequences themselves converge.

By Nguetseng's theorem, there exist functions $P_f(\boldsymbol{x},t,\boldsymbol{y})$, $P_s(\boldsymbol{x},t,\boldsymbol{y})$, $Q(\boldsymbol{x},t,\boldsymbol{y})$, $\boldsymbol{W}(\boldsymbol{x},t,\boldsymbol{y})$, $\boldsymbol{W}_f(\boldsymbol{x},t,\boldsymbol{y})$, and $\boldsymbol{W}_s(\boldsymbol{x},t,\boldsymbol{y})$ that are one-periodic in $\boldsymbol{y}$ and satisfy the condition that the sequences $\{p_f^\varepsilon\}$, $\{p_s^\varepsilon\}$, $\{q^\varepsilon\}$, $\{\boldsymbol{w}^\varepsilon\}$, $\{\boldsymbol{w}_f^\varepsilon\}$, and $\{\boldsymbol{w}_s^\varepsilon\}$ two-scale converge to $P_f(\boldsymbol{x},t,\boldsymbol{y})$, $P_s(\boldsymbol{x},t,\boldsymbol{y})$, $Q(\boldsymbol{x},t,\boldsymbol{y})$, $\boldsymbol{W}(\boldsymbol{x},t,\boldsymbol{y})$, $\boldsymbol{W}_f(\boldsymbol{x},t,\boldsymbol{y})$, and $\boldsymbol{W}_s(\boldsymbol{x},t,\boldsymbol{y})$, respectively.

LEMMA 5.1. If $\mu_1 = \infty$ ($\lambda_1 = \infty$), then $\boldsymbol{W}_f(\boldsymbol{x},t,\boldsymbol{y}) = \boldsymbol{w}_f(\boldsymbol{x},t)$, $\chi(\boldsymbol{y})\boldsymbol{W}(\boldsymbol{x},t,\boldsymbol{y}) = \chi(\boldsymbol{y})\boldsymbol{w}_f(\boldsymbol{x},t)$, and $\boldsymbol{w}^f = \langle \boldsymbol{W}\rangle_{Y_f} = m\boldsymbol{w}_f$ $\big(\boldsymbol{W}_s(\boldsymbol{x},t,\boldsymbol{y}) = \boldsymbol{w}_s(\boldsymbol{x},t)$, $(1-\chi(\boldsymbol{y}))\boldsymbol{W}(\boldsymbol{x},t,\boldsymbol{y}) = (1-\chi(\boldsymbol{y}))\boldsymbol{w}_s(\boldsymbol{x},t)$ and $\boldsymbol{w}^s = \langle \boldsymbol{W}\rangle_{Y_s} = (1-m)\boldsymbol{w}_s\big)$

Proof. Suppose that $\mu_1 = \infty$, and let $\Psi(\boldsymbol{x},t,\boldsymbol{y})$ be an arbitrary smooth scalar function periodic in $\boldsymbol{y}$. The sequence $\{\sigma_{ij}^\varepsilon\}$, where

$$\sigma_{ij}^\varepsilon = \int_\Omega \sqrt{\alpha_\lambda}\frac{\partial w_{f,i}^\varepsilon}{\partial x_j}(\boldsymbol{x},t)\Psi(\boldsymbol{x},t,\boldsymbol{x}/\varepsilon)dx, \quad \boldsymbol{w}_f^\varepsilon = (w_{f,1}^\varepsilon, w_{f,2}^\varepsilon, w_{f,3}^\varepsilon),$$

is uniformly bounded in ε. Therefore,

$$\int_\Omega \varepsilon\frac{\partial w_{f,i}^\varepsilon}{\partial x_j}(\boldsymbol{x},t)\Psi(\boldsymbol{x},t,\boldsymbol{x}/\varepsilon)dx = \frac{\varepsilon}{\sqrt{\alpha_\lambda}}\sigma_{ij}^\varepsilon \to 0$$

as $\varepsilon \searrow 0$, which is equivalent to

$$\int_\Omega\int_Y W_{f,i}(\boldsymbol{x},t,\boldsymbol{y})\frac{\partial\Psi}{\partial y_j}(\boldsymbol{x},t,\boldsymbol{y})dxdy = 0, \quad \boldsymbol{W}_f = (W_{f,1}, W_{f,2}, W_{f,3}),$$

or $\boldsymbol{W}_f(\boldsymbol{x},t,\boldsymbol{y}) = \boldsymbol{w}_f(\boldsymbol{x},t)$. Therefore taking the two-scale limit as $\varepsilon \searrow 0$ in the relation $\chi^\varepsilon(\boldsymbol{w}^\varepsilon - \boldsymbol{w}_f^\varepsilon) = 0$, we arrive at

$$\chi(\boldsymbol{y})\boldsymbol{W}(\boldsymbol{x},t,\boldsymbol{y}) = \chi(\boldsymbol{y})\boldsymbol{w}_f. \qquad \square$$

5.2. Micro- and macroscopic equations. We start the proof of the theorem from the macro- and microscopic equations related to the continuity equations.

LEMMA 5.2. For almost all $\boldsymbol{x} \in \Omega$ and $\boldsymbol{y} \in Y$, the sequences $\{p_f^\varepsilon\}$, $\{p_s^\varepsilon\}$, $\{q^\varepsilon\}$, $\{\boldsymbol{w}^\varepsilon\}$, $\{\boldsymbol{w}_f^\varepsilon\}$, and $\{\boldsymbol{w}_s^\varepsilon\}$ two-scale converge to functions satisfying the relations

$$Q = q\chi/m,\ P_f = p_f\chi/m,\ P_s = p_s(1-\chi)/(1-m),\ Q = P + \nu_0 p_*^{-1}\partial P/\partial t; \tag{5.3}$$

$$q/m = p_s/(1-m), \quad q = p_f + \nu_0 p_*^{-1}\partial p_f/\partial t; \tag{5.4}$$

$$p_f/p_* + p_s/\eta_0 + \mathrm{div}\boldsymbol{w} = 0; \tag{5.5}$$

$$\boldsymbol{w}(\boldsymbol{x},t)\cdot\boldsymbol{n}(\boldsymbol{x}) = 0, \quad \boldsymbol{x}\in S,\ t>0; \tag{5.6}$$

$$\mathrm{div}_y\boldsymbol{W} = 0; \tag{5.7}$$

$$\boldsymbol{W} = \chi\boldsymbol{W}_f + (1-\chi)\boldsymbol{W}_s. \tag{5.8}$$

Proof. In order to prove (5.3), into (2.4) we substitute the test function $\boldsymbol{\psi}^\varepsilon = \varepsilon\boldsymbol{\psi}(\boldsymbol{x},t,\boldsymbol{x}/\varepsilon)$, where $\boldsymbol{\psi}(\boldsymbol{x},t,\boldsymbol{y})$ is an arbitrary one-periodic function of $\boldsymbol{y}$ that is finite on Y_f (or finite on Y_s or finite on Y). Passing to the limit as $\varepsilon \searrow 0$, we obtain

$$\nabla_y Q = 0,\ \boldsymbol{y}\in Y_f;\ \nabla_y P_s = 0,\ \boldsymbol{y}\in Y_s;\ \nabla_y(Q+P_s) = 0,\ \boldsymbol{y}\in Y. \tag{5.9}$$

Next, fulfilling the two-scale passage to the limit in the state equation (2.3) and in the relations

$$(1-\chi^\varepsilon)q^\varepsilon = 0, \quad \chi^\varepsilon p_s^\varepsilon = 0$$

we arrive at the last equation in (5.3) and the relations

$$(1-\chi)Q=0,\quad \chi P_s=0,$$

which together with the first two equations in (5.9) prove the first three equations in (5.3).

The second equation in (5.4) is the result of integration of the last equation in (5.3) over the domain Y_f.

The first relation in (5.4) follows from (5.3) and the last equation in (5.9): the sequence $\{(q^\varepsilon+p_s^\varepsilon)\}$ two-scale converges to $(Q+P_s)=(q+p_s)$.

Equations (5.5)–(5.7) appear as a result of the two-scale passage to the limit in (1.5) and (2.2) with the proper test functions being involved. Thus, for example, (5.5) and (5.6) arise if we consider the linear combination of (1.5) and (2.2)

$$\frac{1}{\alpha_p}p_f^\varepsilon+\frac{1}{\alpha_\eta}p_s^\varepsilon+\mathrm{div}\boldsymbol{w}^\varepsilon=0, \tag{5.10}$$

multiply it by an arbitrary function independent of the "fast" variable $\boldsymbol{x}/\varepsilon$, and then pass to the limit as $\varepsilon\searrow 0$. To prove (5.7), it suffices to consider the two-scale limiting relations in (5.10) as $\varepsilon\searrow 0$ with the test functions $\varepsilon\psi(\boldsymbol{x}/\varepsilon)\,h(\boldsymbol{x},t)$, where ψ and h are arbitrary smooth functions.

To prove (5.8), it suffices to consider the two-scale limiting relations in

$$\boldsymbol{w}^\varepsilon=\chi^\varepsilon\boldsymbol{w}_f^\varepsilon+(1-\chi^\varepsilon)\boldsymbol{w}_s^\varepsilon. \qquad \square$$

LEMMA 5.3. [illegible] $(\boldsymbol{x},t)\in\Omega_T$ [illegible]

$$\rho_f\frac{\partial^2\boldsymbol{w}^f}{\partial t^2}+\rho_s\frac{\partial^2\boldsymbol{w}^s}{\partial t^2}=-\frac{1}{m}\nabla q+\hat\rho\boldsymbol{F} \tag{5.11}$$

[illegible]. Substituting a test function of the form $\boldsymbol{\psi}=\boldsymbol{\psi}(\boldsymbol{x},t)$ into integral identity (2.4) and passing to the limit as $\varepsilon\searrow 0$, we arrive at (5.11). $\square$

LEMMA 5.4. [illegible] $\mu_1=\infty$ [illegible] $\lambda_1<\infty$ [illegible] Y_s [illegible] $\boldsymbol{W}^s=(1-\chi)\boldsymbol{W}$, $\boldsymbol{w}_f$ [illegible] q [illegible]

$$\rho_s\frac{\partial^2\boldsymbol{W}^s}{\partial t^2}=\lambda_1\triangle_y\boldsymbol{W}^s-\nabla_y R^s-\frac{1}{m}\nabla q+\rho_s\boldsymbol{F},\quad \boldsymbol{y}\in Y_s, \tag{5.12}$$

$$\boldsymbol{W}^s=\boldsymbol{w}_f,\quad \boldsymbol{y}\in\gamma, \tag{5.13}$$

[illegible] $\lambda_1>0$ [illegible]

$$\rho_s\frac{\partial^2\boldsymbol{W}^s}{\partial t^2}=-\nabla_y R^s-\frac{1}{m}\nabla q+\rho_s\boldsymbol{F},\quad \boldsymbol{y}\in Y_s, \tag{5.14}$$

$$(\boldsymbol{W}^s-\boldsymbol{w}_f)\cdot\boldsymbol{n}=0,\quad \boldsymbol{y}\in\gamma, \tag{5.15}$$

[illegible] $\lambda_1=0$

$$\boldsymbol{W}^s(\boldsymbol{y},0)=\frac{\partial\boldsymbol{W}^s}{\partial t}(\boldsymbol{y},0)=0,\quad \boldsymbol{y}\in Y_s. \tag{5.16}$$

(5.15) $\boldsymbol{n}$ [illegible] γ [illegible]. The differential equations (5.12) and (5.14) follow as $\varepsilon \searrow 0$ from integral equality (2.4) with the test function $\boldsymbol{\psi} = \boldsymbol{\varphi}(x\varepsilon^{-1}) \cdot h(\boldsymbol{x}, t)$, where $\boldsymbol{\varphi}$ is solenoidal and finite in Y_s.

The boundary condition (5.13) is a consequence of the two-scale convergence of the sequence $\{\sqrt{\alpha_\lambda}\nabla_x \boldsymbol{w}^\varepsilon\}$ to the function $\sqrt{\lambda_1}\nabla_y \boldsymbol{W}(\boldsymbol{x}, t, \boldsymbol{y})$. By this convergence, the function $\nabla_y \boldsymbol{W}(\boldsymbol{x}, t, \boldsymbol{y})$ is L^2-integrable in Y. The boundary condition (5.15) follows from (5.7)–(5.8) and the relation $\boldsymbol{W}_f = \boldsymbol{w}_f$. □

In the same way, one can prove the following lemma.

LEMMA 5.5. [illegible] $\mu_1 < \infty$ [illegible] $\lambda_1 = \infty$ [illegible] Y_f [illegible] $\boldsymbol{W}^f = \chi \boldsymbol{W}$, $\boldsymbol{w}_s$ [illegible] q [illegible]

$$\rho_f \frac{\partial^2 \boldsymbol{W}^f}{\partial t^2} = \mu_1 \triangle_y \frac{\partial \boldsymbol{W}^f}{\partial t} - \nabla_y R^f - \frac{1}{m}\nabla q + \rho_f \boldsymbol{F}, \quad \boldsymbol{y} \in Y_f, \tag{5.17}$$

$$\boldsymbol{W}^f = \boldsymbol{w}_s, \quad \boldsymbol{y} \in \gamma, \tag{5.18}$$

[illegible] $\mu_1 > 0$ [illegible]

$$\rho_f \frac{\partial^2 \boldsymbol{W}^f}{\partial t^2} = -\nabla_y R^f - \frac{1}{m}\nabla q + \rho_f \boldsymbol{F}, \quad \boldsymbol{y} \in Y_f, \tag{5.19}$$

$$(\boldsymbol{W}^f - \boldsymbol{w}_s) \cdot \boldsymbol{n} = 0, \quad \boldsymbol{y} \in \gamma, \tag{5.20}$$

[illegible] $\mu_1 = 0$ [illegible]

$$\boldsymbol{W}^f(\boldsymbol{y}, 0) = \frac{\partial \boldsymbol{W}^f}{\partial t}(\boldsymbol{y}, 0) = 0, \quad \boldsymbol{y} \in Y_f. \tag{5.21}$$

LEMMA 5.6. [illegible] $\mu_1 < \infty$, $\lambda_1 < \infty$, [illegible] $\tilde{\rho} = \rho_f \chi + \rho_s(1 - \chi)$ [illegible] Y, [illegible] $\boldsymbol{W}$ [illegible] q [illegible]

$$\left.\begin{aligned} &\tilde{\rho}\partial^2 \boldsymbol{W}/\partial t^2 + 1/m \nabla q - \tilde{\rho}\boldsymbol{F} \\ &= \operatorname{div}_y\{\mu_1 \chi D(y, \partial \boldsymbol{W}/\partial t) + \lambda_1(1 - \chi)D(y, \boldsymbol{W}) - RI\} \end{aligned}\right\} \tag{5.22}$$

[illegible]

$$\boldsymbol{W}(\boldsymbol{y}, 0) = \frac{\partial \boldsymbol{W}}{\partial t}(\boldsymbol{y}, 0) = 0, \quad \boldsymbol{y} \in Y. \tag{5.23}$$

In the proof of the last lemma, we additionally use Nguetseng's theorem, which states that the sequence $\{\varepsilon D(x, \boldsymbol{w}^\varepsilon)\}$ two-scale converges to the function $D(y, \boldsymbol{W})$.

5.3. Homogenized equations. Lemmas 5.2 and 5.3 imply the following lemma.

LEMMA 5.7. [illegible] $\mu_1 = \lambda_1 = \infty$ [illegible] $\boldsymbol{w}_f = \boldsymbol{w}_s = \boldsymbol{w}$ [illegible] Ω_T [illegible] $\boldsymbol{w}$, p_f, q [illegible] p_s [illegible]

$$\hat{\rho}\frac{\partial^2 \boldsymbol{w}}{\partial t^2} = -\frac{1}{m}\nabla q + \hat{\rho}\boldsymbol{F}, \tag{5.24}$$

$$\frac{1}{p_*}p_f + \frac{1}{\eta_0}p_s + \operatorname{div}\boldsymbol{w} = 0, \tag{5.25}$$

$$q = p_f + \frac{\nu_0}{p_*}\frac{\partial p_f}{\partial t}, \quad \frac{1}{m}q = \frac{1}{1-m}p_s, \tag{5.26}$$

$$\boldsymbol{w}(\boldsymbol{x},0) = \frac{\partial \boldsymbol{w}}{\partial t}(\boldsymbol{x},0) = 0, \quad \boldsymbol{x} \in \Omega, \tag{5.27}$$

$$\boldsymbol{w}(\boldsymbol{x},t)\cdot\boldsymbol{n}(\boldsymbol{x}) = 0, \quad \boldsymbol{x} \in S,\ t > 0. \tag{5.28}$$

LEMMA 5.8. $\mu_1 = \infty$ $\lambda_1 < \infty$ Ω_T $\boldsymbol{w}_f$, $\boldsymbol{w}^s$, p_f, q p_s (5.26)

$$\rho_f m\frac{\partial^2 \boldsymbol{w}_f}{\partial t^2} + \rho_s\frac{\partial^2 \boldsymbol{w}^s}{\partial t^2} = -\frac{1}{m}\nabla q + \hat{\rho}\boldsymbol{F}, \tag{5.29}$$

$$\frac{1}{p_*}p_f + \frac{1}{\eta_0}p_s + m\mathrm{div}\boldsymbol{w}_f + \mathrm{div}\boldsymbol{w}^s = 0, \tag{5.30}$$

$$\frac{\partial \boldsymbol{w}^s}{\partial t} = (1-m)\frac{\partial \boldsymbol{w}_f}{\partial t} + \int_0^t B_1^s(t-\tau)\cdot\boldsymbol{z}^s(\boldsymbol{x},\tau)d\tau, \tag{5.31}$$

$$\boldsymbol{z}^s(\boldsymbol{x},t) = -\frac{1}{m}\nabla q(\boldsymbol{x},t) + \rho_s\boldsymbol{F}(\boldsymbol{x},t) - \rho_s\frac{\partial^2 \boldsymbol{w}_f}{\partial t^2}(\boldsymbol{x},t),$$

$\lambda_1 > 0$

$$\rho_s\frac{\partial^2 \boldsymbol{w}^s}{\partial t^2} = \rho_s B_2^s\cdot\frac{\partial^2 \boldsymbol{w}_f}{\partial t^2} + ((1-m)I - B_2^s)\cdot\left(-\frac{1}{m}\nabla q + \rho_s\boldsymbol{F}\right) \tag{5.32}$$

$\lambda_1 = 0$ (5.26), (5.29)–(5.32) (5.27) (5.28) $\boldsymbol{w} = m\boldsymbol{w}_f + \boldsymbol{w}^s$ (5.31)–(5.32) $B_1^s(t)$ B_2^s (5.37) (5.39) $((1-m)I - B_2^s)$. Equation (5.29) follows directly from (5.11). The continuity equation (5.30) follows from (5.5) if we take into account that

$$\boldsymbol{w} = m\boldsymbol{w}_f + \boldsymbol{w}^s.$$

To find the last two equations (5.31) and (5.32), we just have to solve the system of microscopic equations (5.7), (5.12)–(5.16) and use the formula

$$\boldsymbol{w}^s = \langle \boldsymbol{W}\rangle_{Y_s}.$$

There are two different cases.

(a) If $\lambda_1 > 0$, then the solution of the system of microscopic equations (5.7), (5.12), and (5.13) supplemented with the homogeneous initial data (5.16) is given by the formulas

$$\boldsymbol{W}^s = \int_0^t \left(\boldsymbol{v}(\boldsymbol{x},\tau) + \sum_{i=1}^{3} \boldsymbol{W}^{s,i}(\boldsymbol{y}, t-\tau) z_i^s(\boldsymbol{x},\tau) \right) d\tau,$$

$$R^s = \int_0^t \sum_{i=1}^{3} R^{s,i}(\boldsymbol{y}, t-\tau) z_i^s(\boldsymbol{x},\tau) d\tau, \quad \boldsymbol{z}^s = \left(z_1^s, z_2^s, z_3^s\right),$$

and the functions $\boldsymbol{W}^{s,i}(\boldsymbol{y},t)$ and $R^{s,i}(\boldsymbol{y},t)$ are defined by virtue of the periodic initial boundary value problem

$$\rho_s \frac{\partial^2 \boldsymbol{W}^{s,i}}{\partial t^2} - \lambda_1 \triangle \boldsymbol{W}^{s,i} + \nabla R^{s,i} = 0, \quad \boldsymbol{y} \in Y_s,\ t > 0, \tag{5.33}$$

$$\operatorname{div}_y \boldsymbol{W}^{s,i} = 0, \quad \boldsymbol{y} \in Y_s,\ t > 0, \tag{5.34}$$

$$\boldsymbol{W}^{s,i} = 0, \quad \boldsymbol{y} \in \gamma,\ t > 0, \tag{5.35}$$

$$\boldsymbol{W}^{s,i}(y,0) = 0, \quad \rho_s \frac{\partial \boldsymbol{W}^{s,i}}{\partial t}(\boldsymbol{y},0) = \boldsymbol{e}_i, \quad \boldsymbol{y} \in Y_s. \tag{5.36}$$

In (5.36), $\boldsymbol{e}_i$ is the standard Cartesian basis vector.

Therefore,

$$B_1^s(t) = \left\langle \sum_{i=1}^{3} \frac{\partial \boldsymbol{W}^{s,i}}{\partial t} \right\rangle_{Y_s} (t) \otimes \boldsymbol{e}_i. \tag{5.37}$$

Note that (5.33) is understood in the sense of distributions and the function $B_1^s(t)$ has no time derivative at $t = 0$.

(b) If $\lambda_1 = 0$, then in solving the system (5.7), (5.14), (5.15), and (5.16), we first find the pressure $R^s(\boldsymbol{x}, t, \boldsymbol{y})$ by solving the Neumann problem for the Laplace equation in Y_s in the form

$$R^s(\boldsymbol{x}, t, \boldsymbol{y}) = \sum_{i=1}^{3} R_{s,i}(\boldsymbol{y}) z_i^s(\boldsymbol{x}, t),$$

where $R_{s,i}(\boldsymbol{y})$ is the solution of the problem

$$\triangle_y R_{s,i} = 0, \quad \boldsymbol{y} \in Y_s; \quad \nabla_y R_{s,i} \cdot \boldsymbol{n} = \boldsymbol{n} \cdot \boldsymbol{e}_i, \quad \boldsymbol{y} \in \gamma; \quad \langle R_{s,i} \rangle_{Y_s} = 0. \tag{5.38}$$

Formula (5.32) is the result of integration of (5.14) over the domain Y_s and

$$B_2^s = \sum_{i=1}^{3} \langle \nabla R_{s,i} \rangle_{Y_s} \otimes \boldsymbol{e}_i, \tag{5.39}$$

where the matrix $B = ((1-m)I - B_2^s)$ is symmetric and strictly positive definite. In fact, let $\tilde{R} = \sum_{i=1}^3 R_{s,i}\xi_i$ for any unit vector $\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_3)$. Then

$$(B \cdot \boldsymbol{\xi}) \cdot \boldsymbol{\xi} = \langle (\boldsymbol{\xi} - \nabla \tilde{R})^2 \rangle_{Y_f}$$

and $(B \cdot \boldsymbol{\xi}) \cdot \boldsymbol{\xi} = 0$ if and only if $\tilde{R}$ is a linear function in $\boldsymbol{y}$. On the other hand, it follows from the assumption about the geometry of the domain Y_s that all linear periodic functions on Y_s are constant. Finally, the normalization condition $\langle R_{s,i} \rangle_{Y_s} = 0$ yields that $\tilde{R} = 0$. However, this is impossible, because the functions $R_{s,i}$ are linearly independent. □

LEMMA 5.9. [illegible] $\mu_1 < \infty$ [illegible] $\lambda_1 = \infty$ [illegible] Ω_T [illegible] $\boldsymbol{w}^f$, $\boldsymbol{w}_s$, p_f, q [illegible] p_s [illegible] (5.26) [illegible]

$$\rho_f \frac{\partial^2 \boldsymbol{w}^f}{\partial t^2} + \rho_s (1-m) \frac{\partial^2 \boldsymbol{w}_s}{\partial t^2} = -\frac{1}{m} \nabla q + \hat{\rho} \boldsymbol{F}, \tag{5.40}$$

[illegible]

$$\frac{1}{p_*} p_f + \frac{1}{\eta_0} p_s + \mathrm{div} \boldsymbol{w}^f + (1-m) \mathrm{div} \boldsymbol{w}_s = 0, \tag{5.41}$$

[illegible]

$$\frac{\partial \boldsymbol{w}^f}{\partial t} = m \frac{\partial \boldsymbol{w}_s}{\partial t} + \int_0^t B_1^f (t-\tau) \cdot \boldsymbol{z}^f(\boldsymbol{x}, \tau) d\tau, \tag{5.42}$$

[illegible]

$$\boldsymbol{z}^f(\boldsymbol{x}, t) = -\frac{1}{m} \nabla q(\boldsymbol{x}, t) + \rho_f \boldsymbol{F}(\boldsymbol{x}, t) - \rho_f \frac{\partial^2 \boldsymbol{w}_s}{\partial t^2}(\boldsymbol{x}, t),$$

[illegible] $\mu_1 > 0$ [illegible]

$$\rho_f \frac{\partial^2 \boldsymbol{w}^f}{\partial t^2} = \rho_f B_2^f \cdot \frac{\partial^2 \boldsymbol{w}_s}{\partial t^2} + (mI - B_2^f) \cdot \left(-\frac{1}{m} \nabla q + \rho_f \boldsymbol{F} \right) \tag{5.43}$$

[illegible] $\mu_1 = 0$ [illegible] (5.26), (5.40)–(5.43) [illegible] (5.27) [illegible] (5.28) [illegible] $\boldsymbol{w} = \boldsymbol{w}^f + (1-m)\boldsymbol{w}_s$ [illegible] (5.42)–(5.43) [illegible] $B_1^f(t)$ [illegible] B_2^f [illegible] (5.44)–(5.45) [illegible] $(mI - B_2^f)$ [illegible] f_i [illegible]. The proof of this lemma repeats that of the previous lemma. Here we have to solve the system of microscopic equations (5.7), (5.17)–(5.21) and use the formula

$$\boldsymbol{w}^f = \langle \boldsymbol{W} \rangle_{Y_f}.$$

Thus,

$$B_1^f(t) = \left\langle \sum_{i=1}^3 \frac{\partial \boldsymbol{W}^{f,i}}{\partial t} \right\rangle_{Y_f} (t) \otimes \boldsymbol{e}_i, \tag{5.44}$$

$$B_2^f = \sum_{i=1}^{3} \langle \nabla R_{f,i} \rangle_{Y_f} \otimes \boldsymbol{e}_i, \tag{5.45}$$

where the functions $\boldsymbol{W}^{f,i}(\boldsymbol{y},t)$ solve the periodic initial boundary value problem

$$\rho_f \frac{\partial^2 \boldsymbol{W}^{f,i}}{\partial t^2} - \mu_1 \triangle \frac{\partial \boldsymbol{W}^{f,i}}{\partial t} + \nabla R^{f,i} = 0, \quad \boldsymbol{y} \in Y_f,\ t > 0, \tag{5.46}$$

$$\operatorname{div}_y \boldsymbol{W}^{f,i} = 0, \quad \boldsymbol{y} \in Y_f,\ t > 0, \tag{5.47}$$

$$\boldsymbol{W}^{f,i} = 0, \quad \boldsymbol{y} \in \gamma,\ t > 0, \tag{5.48}$$

$$\boldsymbol{W}^{f,i}(y,0) = 0, \quad \rho_f \frac{\partial \boldsymbol{W}^{f,i}}{\partial t}(y,0) = \boldsymbol{e}_i, \quad \boldsymbol{y} \in Y_f, \tag{5.49}$$

and the functions $R_{f,i}(\boldsymbol{y})$ solve the periodic boundary value problem

$$\triangle_y R_{f,i} = 0, \quad \boldsymbol{y} \in Y_f; \quad \nabla_y R_{f,i} \cdot \boldsymbol{n} = \boldsymbol{n} \cdot \boldsymbol{e}_i, \quad \boldsymbol{y} \in \gamma; \quad \langle R_{f,i} \rangle_{Y_f} = 0. \tag{5.50}$$

Note that, as before, the matrix $(mI - B_2^f)$ is symmetric and strictly positive definite. □

The proof of Theorem 2.2 is completed by the following lemma.

LEMMA 5.10. [illegible] $\mu_1 < \infty$ [illegible] $\lambda_1 < \infty$ [illegible] Ω_T [illegible] $\boldsymbol{w}$, p_f, q, p_s [illegible] (5.25) [illegible] (5.26) [illegible]

$$\frac{\partial \boldsymbol{w}}{\partial t} = \int_0^t B(t-\tau) \cdot \nabla q(\boldsymbol{x},\tau) d\tau + \boldsymbol{f}(\boldsymbol{x},t), \tag{5.51}$$

[illegible] $B(t)$ [illegible] $\boldsymbol{f}(\boldsymbol{x},t)$ [illegible] (5.58) [illegible] (5.59) [illegible] (5.25), (5.26), (5.51) [illegible] (5.27) [illegible] (5.28) [illegible]. To derive the momentum conservation law (5.51), we must solve the system of microscopic equations (5.7), (5.22) with the initial conditions (5.23) and use the formula

$$\boldsymbol{w} = \langle \boldsymbol{W} \rangle_Y.$$

Let

$$\boldsymbol{W} = \int_0^t \sum_{i=1}^{3} \left\{ \boldsymbol{W}^{q,i}(\boldsymbol{y}, t-\tau) \frac{\partial q}{\partial x_i}(\boldsymbol{x},\tau) + \boldsymbol{W}^{F,i}(\boldsymbol{y}, t-\tau) F_i(\boldsymbol{x},\tau) \right\} d\tau,$$

$$R = \int_0^t \sum_{i=1}^{3} \left\{ R^{q,i}(\boldsymbol{y}, t-\tau) \frac{\partial q}{\partial x_i}(\boldsymbol{x},\tau) + R^{F,i}(\boldsymbol{y}, t-\tau) F_i(\boldsymbol{x},\tau) \right\} d\tau,$$

where $\boldsymbol{F} = \sum_{i=1}^{3} F_i \boldsymbol{e}_i$.

Then the pair $\{\boldsymbol{W}, R\}$ is a solution of system (5.7), (5.22) and (5.23) if and only if the functions $\{\boldsymbol{W}^{q,i}(\boldsymbol{y},t), R^{q,i}(\boldsymbol{y},t)\}$ and $\{\boldsymbol{W}^{F,i}(\boldsymbol{y},t), R^{F,i}(\boldsymbol{y},t)\}$ are periodic in $\boldsymbol{y}$ solutions of the equations

$$\mathrm{div}_y\left\{\mu_1\chi D\left(y,\frac{\partial \boldsymbol{W}^{q,i}}{\partial t}\right)+\lambda_1(1-\chi)D(y,\boldsymbol{W}^{q,i})-R^{q,i}I\right\}=\tilde{\rho}\frac{\partial^2\boldsymbol{W}^{q,i}}{\partial t^2}, \tag{5.52}$$

$$\mathrm{div}_y\boldsymbol{W}^{q,i}=0, \tag{5.53}$$

$$\mathrm{div}_y\left\{\mu_1\chi D\left(y,\frac{\partial \boldsymbol{W}^{F,i}}{\partial t}\right)+\lambda_1(1-\chi)D(y,\boldsymbol{W}^{F,i})-R^{F,i}I\right\}=\tilde{\rho}\frac{\partial^2\boldsymbol{W}^{F,i}}{\partial t^2}, \tag{5.54}$$

$$\mathrm{div}_y\boldsymbol{W}^{F,i}=0 \tag{5.55}$$

in the domain Y for $t>0$ and satisfy the initial conditions

$$\boldsymbol{W}^{q,i}(\boldsymbol{y},0)=0,\quad \tilde{\rho}\frac{\partial \boldsymbol{W}^{q,i}}{\partial t}(\boldsymbol{y},0)=-\frac{1}{m}\boldsymbol{e}_i,\quad \boldsymbol{x}\in Y, \tag{5.56}$$

$$\boldsymbol{W}^{F,i}(\boldsymbol{y},0)=0,\quad \frac{\partial \boldsymbol{W}^{F,i}}{\partial t}(\boldsymbol{y},0)=\boldsymbol{e}_i,\quad \boldsymbol{x}\in Y. \tag{5.57}$$

Here $\boldsymbol{e}_i$ is the standard Cartesian basis vector.

Therefore,

$$B(t)=\sum_{i=1}^{3}\left\langle\frac{\partial \boldsymbol{W}^{q,i}}{\partial t}(\boldsymbol{y},t)\right\rangle_Y\otimes\boldsymbol{e}_i, \tag{5.58}$$

$$\mathbf{f}(\boldsymbol{x},t)=\int_0^t\sum_{i=1}^{3}\left\langle\frac{\partial \boldsymbol{W}^{F,i}}{\partial t}(\boldsymbol{y},t-\tau)\right\rangle_Y F_i(\boldsymbol{x},\tau)d\tau. \tag{5.59}$$

The solvability and uniqueness of problems (5.52), (5.53), (5.56) and (5.54), (5.55), (5.57) follow directly from the energy identity

$$\frac{1}{2}\int_Y\left(\tilde{\rho}\left(\frac{\partial \boldsymbol{W}^{j,i}}{\partial t}(\boldsymbol{y},t)\right)^2+\lambda_1 D\big(y,\boldsymbol{W}^{j,i}(\boldsymbol{y},t)\big):D\big(y,\boldsymbol{W}^{j,i}(\boldsymbol{y},t)\big)\right)dy$$

$$+\int_0^t\int_Y\mu_1 D\left(y,\frac{\partial \boldsymbol{W}^{j,i}}{\partial\tau}(\boldsymbol{y},\tau)\right):D\left(y,\frac{\partial \boldsymbol{W}^{j,i}}{\partial\tau}(\boldsymbol{y},\tau)\right)dyd\tau=\frac{1}{2}\beta^j$$

for $i=1,2,3$ and $j=q,F$.

Here

$$\beta^q=\left\langle\frac{1}{\tilde{\rho}}\right\rangle_Y,\quad \beta^F=\langle\tilde{\rho}\rangle_Y.$$

As before, equations (5.51) are understood in the sense of distributions and the function $B(t)$ has no time derivative at $t=0$. □

Acknowledgment. The author acknowledges the anonymous referees for kind advice and for helpful remarks.

REFERENCES

[1] E. Acerbi, V. Chiado Piat, G. Dal Maso, and D. Percivale, *An extension theorem from connected sets and homogenization in general periodic domains*, Nonlinear Anal., 18 (1992), pp. 481–496.

[2] R. Burridge and J. B. Keller, *Poroelasticity equations derived from microstructure*, J. Acoust. Soc. Amer., 70 (1981), pp. 1140–1146.

[3] T. Clopeau, J. L. Ferrin, R. P. Gilbert, and A. Mikelić, *Homogenizing the acoustic properties of the seabed: Part* II, Math. Comput. Modelling, 33 (2001), pp. 821–841.

[4] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, New York, 1994.

[5] D. Lukkassen, G. Nguetseng, and P. Wall, *Two-scale convergence*, Int. J. Pure Appl. Math., 2 (2002), pp. 35–86.

[6] A. Meirmanov, *Nguetseng's two-scale convergence method for filtration and seismic acoustic problems in elastic porous media*, Siberian Math. J., 48 (2007), pp. 519–538.

[7] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–623.

[8] G. Nguetseng, *Asymptotic analysis for a stiff variational problem arising in mechanics*, SIAM J. Math. Anal., 21 (1990), pp. 1394–1414.

[9] E. Sanchez-Palencia, *Nonhomogeneous Media and Vibration Theory*, Lecture Notes in Phys. 129, Springer-Verlag, Berlin, 1980.

SIAM J. MATH. ANAL.
Vol. 40, No. 4, pp. 1291–1336

BOUNDARY BEHAVIOR OF SOLUTIONS OF A CLASS OF GENUINELY NONLINEAR HYPERBOLIC SYSTEMS*

JULIAN GEVIRTZ†

Abstract. We study the set of boundary singularities of arbitrary classical solutions of genuinely nonlinear 2×2 planar hyperbolic systems of the form $D_k R_k = 0$, where D_k denotes differentiation in the direction $e^{i\theta_k(R_1,R_2)}$, $k = 1, 2$, and where the defining functions θ_k satisfy (i) $\theta_2 = \theta_1 + \frac{\pi}{2}$, and (ii) $A \leq |\frac{\partial\theta_k(R)}{\partial R_k}| \leq B$, for all $R \in \mathbb{R}^2$, $k = 1, 2$, for some positive constants A, B. We show that for any system of this kind there is a $\tau < 1$ such that for any locally Lipschitz solution R in a smoothly bounded domain G, the set of points of ∂G at which R fails to have a nontangential limit has Hausdorff dimension at most τ, and, on the other hand, for any such system for which the $\theta_k \in C^\infty(\mathbb{R}^2)$, we construct a C^∞ solution R on a half-plane $\mathbb{H}$ for which the set of points of $\partial\mathbb{H}$ at which R fails to have a nontangential limit has positive Hausdorff dimension. These results are immediately applicable to constant principal strain mappings, which are defined in terms of a system of this kind for which θ_1 is a linear function of R_1 and R_2.

Key words. boundary behavior, constant principal strain mapping, genuinely nonlinear hyperbolic system, Hausdorff dimension, nontangential limit

AMS subject classifications. 35L40, 28A78

DOI. 10.1137/070705507

1. Introduction.

For hyperbolic systems in two independent variables x and t, most often associated with space and time, one usually studies the Cauchy problem, in which one seeks a solution $u(x,t)$, $t \geq 0$, for which $u(x,0)$ coincides with a given $u_0(x)$, the questions considered including well-posedness, global existence, blow-up, and behavior of solutions as $t \to \infty$. In the nonlinear case, discussion is often limited to initial data with a small range, and, even for such data, generalized solutions must be considered.

In this paper we concern ourselves with the following inverse question for a certain family of genuinely nonlinear 2×2 hyperbolic systems: What can be said about the boundary values of an [illegible] classical solution in a plane domain G? Here "classical" can be taken to mean C^∞, although the treatment we give will be valid for locally Lipschitz solutions. In the first place, we are interested in systems whose formulation imposes no a priori limit on the range of characteristic directions, that is, systems such that for a characteristic given parametrically by $z(s)$, $\arg\{z'(s)\}$ can potentially cover all of $\mathbb{R}$, in contrast to what is implicitly the case in the standard space-time context. Second, we are interested in statements valid for [illegible] classical solutions rather than ones known to arise from some form of initial value problem. Because of this generality, even in geometrically simple domains such as disks or half-planes characteristics can be quite contorted curves. Although the specific focus of this paper is the size of the set of boundary points at which classical solutions can fail to have nontangential limits, it would be reasonable to investigate other aspects of their behavior and that of the associated characteristics. In any event, given the nonstandard nature of the boundary value question and of several of the issues that arise in dealing with it, we shall begin with a somewhat detailed discussion of a system

*Received by the editors October 16, 2007; accepted for publication March 31, 2008; published electronically November 5, 2008.
http://www.siam.org/journals/sima/40-4/70550.html
†2005 North Winthrop Road, Muncie, IN 47304-2536 (jgevirtz@gmail.com).

for which it is physically meaningful, namely the system which describes smooth planar mappings with constant principal stretches (cps-mappings), about which we have previously written [ChG], [G1], [G2], [G3], [G4], [G5]. It is in fact the study of the boundary behavior of such mappings that is the main goal of this paper, and we have chosen to work in a wider context only because it is possible to do so with little additional effort, and because this broader approach suggests some interesting questions.

A [illegible] with principal stretch factors $m_1 \neq m_2$ is a mapping $f : G \to \mathbb{C}$ with locally Lipschitz continuous Jacobian

$$J_f = T(-\phi)\sigma(m_1, m_2)T(\theta),$$

where

$$T(\theta) = \begin{bmatrix} \cos\ \theta & \sin\ \theta \\ -\sin\ \theta & \cos\ \theta \end{bmatrix} \quad \text{and} \quad \sigma(m_1, m_2) = \begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix}.$$

As is explained in the cited references, apart from regularity considerations, functions θ and ϕ will correspond to such a mapping on a simply connected domain G if and only if they satisfy the autonomous quasi-linear hyperbolic system

$$D_1(m_1\theta - m_2\phi) = 0, \quad D_2(m_2\theta - m_1\phi) = 0, \tag{1.1}$$

where

$$D_1 u = (\cos\theta)u_x + (\sin\theta)u_y \quad \text{and} \quad D_2 u = (-\sin\theta)u_x + (\cos\theta)u_y.$$

The characteristics of a solution are the integral curves of the fields $e^{i\theta}$ and $ie^{i\theta}$. It turns out that a net $\mathcal{N}$ made up of two mutually orthogonal families of curves covering a simply connected G is the net of characteristics of a cps-mapping if and only if for any two curves C_1, C_2 belonging to one of the families of $\mathcal{N}$ the change in the inclination of the tangent is the same along all subarcs of curves of the [illegible] family which join C_1 to C_2. Nets with this property are known as [illegible]. (See [CS], [G3], [Hem], [Hen], [Hi], [Pr].) The theory of cps-mappings we have developed is based on direct application of (1.1) together with this Hencky–Prandtl (HP) property and the equations

$$D_2D_1\theta = [D_1\theta]^2 \quad \text{and} \quad D_1D_2\theta = -[D_2\theta]^2, \tag{1.2}$$

which are also effectively equivalent to (1.1) and which are very special cases of equations derived by Lax [L] in the context of considerably more general genuinely nonlinear 2×2 hyperbolic systems in the plane and used by him in connection with the inevitability of singularity formation. The [illegible] (1.2) imply that if a characteristic C has curvature κ_0 at p, then the orthogonal characteristic arc emanating from p towards the concave side of C can have length at most $\frac{1}{\kappa_0}$, that is, that the boundary of G must be encountered after moving at most a distance of $\frac{1}{\kappa_0}$ along this orthogonal characteristic. A [illegible] of this kind is a [illegible] for the theory we are developing and plays a fundamental role in all that is to follow.

When regarded as deformations with constant principal strains, cps-mappings are of concrete interest as models in a number of physically interesting contexts (see [Y]).

Consider, for example, a thin liquid film on a plane surface which upon solidification takes on a rectangular cryptocrystalline structure; that is, at each point a suitably oriented minute square of the original liquid becomes a rectangular crystal whose side lengths are constant multiples of the side length of the square. In this light global geometric results for cps-mappings tell one about the extent to which the shape of the original film can change as a result of such solidification and about how matter is moved around in the process, and statements about the existence of boundary limits of θ (and, in light of (1.1), of ϕ, and consequently of the Jacobian of the mapping) tell one to what extent the cryptocrystalline structure is present at the very edge of the solidified lamina. Applied to the system (1.1) the main result of this paper says that [illegible] $\tau < 1$ [illegible] G [illegible] θ [illegible] ϕ [illegible] $S \subset \partial G$ [illegible] τ. On the other hand, the construction of section 5 shows that [illegible] θ [illegible].

Beyond their immediate physical significance, cps-mappings constitute a particularly important and tractable class of planar quasi-isometric mappings, for which we believe they will ultimately be shown to display extremal behavior for many of the as yet unsolved distortion questions (see [J1], [J2]). In this direction a very significant inroad was made a few years ago by Gutlyanskii and Martio [GM], who showed that the [illegible] given by

$$g(re^{i\psi}) = re^{i(\psi+\psi_0 \frac{\log r}{\log \rho})}$$

are extremal for the problem of determining for given $\rho > 1$ and $\psi_0 > 0$ the smallest ratio $\frac{m_1}{m_2} > 1$, such that there is a quasi-isometric mapping f of the annulus $1 < |z| < \rho$ onto itself with stretching bounds m_1, m_2 which satisfies the boundary conditions

$$f(z) = z \quad \text{and} \quad f(\rho z) = \rho e^{i\psi_0} z \quad \text{for} \quad |z| = 1.$$

It is in fact not hard to see that g is indeed a cps-mapping of the entire punctured plane $\mathbb{C}\backslash\{0\}$ with

$$m_1 = \frac{\sqrt{a^2+4}+a}{2} \quad \text{and} \quad m_2 = \frac{\sqrt{a^2+4}-a}{2},$$

where $a = \frac{\psi_0}{\log \rho}$. We call these spiral mappings because the corresponding inclination functions are of the form

$$\theta(re^{i\psi}) = \psi + \alpha, \tag{1.3}$$

where $\tan \alpha = m_2$, which is to say that the characteristics form two mutually orthogonal families of logarithmic spirals, all members of each of which are rotations of each other.

We now describe the class of 2×2 systems for which we treat the boundary limit question. Let $\theta_k = \theta_k(R_1, R_2)$, $k = 1, 2$. For given $R_1(x, y)$, $R_2(x, y)$ and any $u = u(x, y)$ we write

$$D_k u = \cos \theta_k(R_1(x,y), R_2(x,y)) \frac{\partial u}{\partial x} + \sin \theta_k(R_1(x,y), R_2(x,y)) \frac{\partial u}{\partial y}. \tag{1.4}$$

It is well known that in general an autonomous 2×2 quasi-linear homogeneous hyperbolic system for unknown functions f and g is formally equivalent to a system of

the form

$$D_k R_k = 0, \quad k = 1, 2,$$

with appropriate inclination functions $\theta_k(R_1, R_2)$. The relationship between the [illegible] R_1, R_2 and f, g is of the form $R_k(x, y) = F_k(f(x, y), g(x, y))$. Henceforth we write $\Theta = (\theta_1, \theta_2)$ and use the term [illegible] to mean a C^∞ mapping $\Theta : \mathbb{R}^2 \to \mathbb{R}^2$, although the smoothness requirement could be weakened substantially. Obviously, a solution of the system Θ in a domain G of the plane is then a pair of functions $R_1(x, y)$, $R_2(x, y)$ for which R_k is constant on each integral curve (henceforth referred to as a k [illegible]) of the field $e^{i\theta_k(R_1(x,y),R_2(x,y))}$, $k = 1, 2$. A system is said to be [illegible] if the derivatives $\frac{\partial \theta_k}{\partial R_k}$, $k = 1, 2$, never vanish. It is clear that the system (1.1) for the θ and ϕ associated with cps-mappings is already in Riemann invariant form with $R_i = m_i\theta - m_j\phi$, so that in this case the two inclination functions are given by

$$\theta_1 = \theta = \frac{m_1 R_1 - m_2 R_2}{m_1^2 - m_2^2} \quad \text{and} \quad \theta_2 = \theta + \frac{\pi}{2}. \tag{1.5}$$

(Here we have used the convention, in force throughout this paper, to the effect that $\{i, j\} = \{1, 2\}$.) This system is obviously genuinely nonlinear and in fact is the simplest possible such system in that the two families of characteristics are mutually orthogonal and Θ is a linear function of $R = (R_1, R_2)$. We now define the family of systems with which we work.

DEFINITION 1.1. [illegible] normal system [illegible] Θ [illegible]

(i) $\theta_2 = \theta_1 + \frac{\pi}{2}$ [illegible]

(ii) [illegible] A, $B > 0$ [illegible] $A \le |\frac{\partial \theta_k}{\partial R_k}| \le B$ [illegible] $R \in \mathbb{R}^2$

The only hyperbolic systems with which we deal will be normal systems, for which we use the symbol θ to denote θ_1. We shall show that [illegible] Θ [illegible] $\tau = \tau(\Theta) < 1$ [illegible] $G \subset \mathbb{C}$ [illegible] R [illegible] Θ [illegible] G [illegible] ∂G [illegible] R [illegible] ff [illegible] τ. This will follow as an immediate consequence (Corollary 4.3) of our principal result, Main Theorem 2.2, which deals with boundary singularities of a class of functions effectively more general than the class of solutions of normal systems. An outline of the proof, which is actually made simpler by this somewhat greater generality, is given early in section 2, just after the statement of the main theorem. Furthermore, in section 5 we shall show that [illegible] Θ [illegible] C^∞ [illegible] R [illegible] $\mathbb{H}$ [illegible] $\partial\mathbb{H}$ [illegible] R [illegible] ff [illegible]

2. Quasi-HP functions, the characteristic length bound, and related matters. Let $G \subset \mathbb{C}$ be a domain. If θ is a locally Lipschitz function on G, the integral curves of the fields $e^{i\theta(z)}$ and $ie^{i\theta(z)}$ will be called the 1- and 2-characteristics of θ, respectively. The term [illegible] will refer to the complete integral curve, and we will use the term [illegible] to refer to either of the two arcs into which a full characteristic is divided by one of its points. (Characteristics which are closed curves will not arise in this paper since we are dealing with single-valued functions θ.) As indicated in the introduction we will use the convention $\{i, j\} = \{1, 2\}$ throughout. Arcs of k-characteristics will be called k-arcs, or, less specifically, characteristic arcs. With reference to a given θ, a characteristic arc joining points $a, b \in D$ will be denoted by ab and we shall use the abbreviation

$$\Delta\theta(ab) = \theta(b) - \theta(a).$$

A domain $Q \subset G$ will be said to be a [illegible] of θ, and we write $Q = abcd$, if ∂Q is a Jordan curve lying in G containing four points a, b, d, c which occur in that order when ∂Q is traversed in the positive (negative) sense and such that ab and cd are i-arcs and ac and bd are j-arcs. We say that ab and cd are [illegible] of each other with respect to or along any j-characteristic passing through ab. For a curve parameterized by $z = z(s)$, $\alpha < s < \beta$, we use the terms "to the right of C" and "to the left of C" in the obvious sense, so that, for example, if C is a characteristic arc and $p \in C$, we describe an orthogonal characteristic arc or a half-characteristic as emanating from p to the right or left of C.

We shall denote the 2-dimensional measure of $X \subset \mathbb{C}$ by $\mu(X)$ and the 1-dimensional measure of a set A by $\lambda(A)$. The parameter s will always refer to arc length. We use the notation $N(a, r) = \{z : |z - a| < r\}$ and denote the line segment joining a to b by $\overline{ab}$. The overline will also be used to denote closure, but this should cause no confusion. For $X, Y \subset \mathbb{C}$, $\operatorname{dist}(X, Y) = \inf\{|y - x| : x \in X, y \in Y\}$, and for $a \in \mathbb{C}$, $\operatorname{dist}(a, X) = \operatorname{dist}(\{a\}, X)$.

DEFINITION 2.1. [illegible] $K \geq 1$ [illegible] θ [illegible] G [illegible] K-quasi-HP function [illegible] K-quasi-HP property [illegible] $abcd \subset G$ [illegible]

$$\frac{1}{K}|\Delta\theta(ac)| \leq |\Delta\theta(bd)| \leq K|\Delta\theta(ac)|. \tag{2.1}$$

[illegible] 1 [illegible] 2 [illegible] K [illegible] K-quasi-HP net

A simple continuity argument shows that this definition implies that the $\Delta\theta(ac)$ and $\Delta\theta(bd)$ in (2.1) must in fact have the same sign (unless both vanish). It is also obvious that a 1-quasi-HP net is an HP net. Note that while the HP property is a local condition that implies its global counterpart, this is not the case for the K-quasi-HP property when $K > 1$. [illegible] Θ [illegible] K [illegible], where $K = K(\Theta)$. Indeed, if A and B are as in Definition 1.1, then it is clear that we can take $K(\Theta) = B/A$.

We can now state our main theorem.

MAIN THEOREM 2.2. [illegible] $\tau = \tau(K) < 1$ [illegible] θ [illegible] K [illegible] G [illegible] C^2 [illegible] ∂G [illegible] θ [illegible] [illegible] τ

Because the proof of this theorem, to be given in section 4, is quite involved and depends on the prior development of a considerable amount of machinery in this and the following section, we shall briefly explain here how it proceeds. As we shall show (see Proposition 3.29) for any point $p \in \partial G$ at which θ does not have a nontangential limit either there is a nontrivial fan of characteristics emanating from p or, for $k = 1$ or 2, every neighborhood of p completely contains a full k-characteristic of θ; in the latter case we say that p is a k-singularity. Since, as will be apparent, any quasi-HP function can have at most a countable number of fans (see Proposition 3.24), we need only show that the Hausdorff dimension, $\dim(S)$, of the set S of k-singularities satisfies $\dim(S) \leq \tau < 1$. For this to be the case it is enough that there be some $\delta = \delta(K) > 0$ such that any almost straight arc $A \subset \partial G$ has a subarc of length at least $\delta\lambda(A)$ which has at most a countable set of k-singularities of θ. If for some θ there were no such δ, then for any N there would have to be an almost straight arc $A \subset \partial G$ such that there is an N-element set S_N of k-singularities of θ which are essentially uniformly distributed along A. For sufficiently large N, starting with a

set of N small k-characteristics, one very close to each of the points of S_N, we show that there must be a k-characteristic C (where the term "k-characteristic" is used here in an appropriate sense—see the discussion of extended characteristics between Propositions 3.13 and 3.14) whose endpoints lie on A and are at least $\delta'\lambda(A)$ apart (where $\delta' > 0$ depends solely on K) and which is tangent to A at a point $m \in A$, where m is appropriately bounded away from the endpoints of C. We then use this to obtain a subarc A' of A which contains m, whose length is bounded below by $\delta''\lambda(A)$, where δ'', like δ', depends only on K, and on which θ can have at most a countable number of k-singularities (see Proposition 3.33), thereby arriving at the desired contradiction. A good measure of the complexity of the proof lies in establishing the existence of the extended characteristic C, which is carried out in section 4, but which depends on properties of the net of extended characteristics of quasi-HP functions developed in section 3. We begin with the following proposition, which is immediate.

PROPOSITION 2.3 (invariance of K). [illegible] $a, b \in \mathbb{C}$ [illegible] $a \neq 0$ [illegible] $\theta(az+b) - \arg a$ [illegible] $\frac{1}{a}(G-b)$

The next proposition is a special case of [G1, Lemma 2]; a somewhat simpler proof than the one given there is included for the sake of completeness. A function θ is said to be a locally L-Lipschitz function on G if each point of G has a neighborhood in which θ satisfies a Lipschitz condition with constant L.

PROPOSITION 2.4. [illegible] L [illegible] $Q = abcd \subset G$ [illegible] $|b-a| = l$, $|c-a| = \epsilon \leq l$ [illegible] $\operatorname{dist}(Q, \partial G) = \eta > 0$ [illegible] $\overline{l} = \overline{l}(L, \eta) > 0$ [illegible]

$$|d-c| = l - \epsilon\Delta\theta(ab) + O(\epsilon^2 + l^3) \qquad l \leq \overline{l}, \tag{2.2}$$

[illegible] O [illegible] L [illegible] η [illegible]. In what follows, when we say that some quantity is big-O of some expression, we mean that this is so for all l less than some positive number which depends only on L and η, and that the constant corresponding to the big-O depends only on L and η. Let ab and cd be i-arcs. Without loss of generality we can assume that $a = 0$ and $b = l$. Let $\frac{\pi}{2} + \alpha$ and $\frac{\pi}{2} + \beta$ be the inclinations of the tangents to the j-characteristics at a and b, respectively, and let $\frac{\pi}{2} + \alpha' = \arg\{c-a\}$, $\frac{\pi}{2} + \beta' = \arg\{d-b\}$. Clearly, α, β, α', and β' are all $O(l)$ and $\beta - \alpha = \Delta\theta(ab)$. It easily follows from the Lipschitz condition that $\alpha' = \alpha + O(\epsilon)$. We have $d = \epsilon i e^{i\alpha'} + te^{i\gamma}$, where $t = |d-c| = O(l)$ and $\gamma = O(l)$. We also have $d = l + sie^{i\beta'}$, where $s = O(l)$. Just as $\alpha' = \alpha + O(\epsilon)$, one sees that $\beta' = \beta + O(s)$. From the two expressions for d we have that

$$\epsilon i e^{i\alpha'} + te^{i\gamma} = l + sie^{i\beta'},$$

so that considering real and imaginary parts we have

$$-\epsilon \sin\alpha' + t\cos\gamma = l - s\sin\beta' \tag{2.3}$$

and

$$\epsilon\cos\alpha' + t\sin\gamma = s\cos\beta'.$$

The latter equation implies that

$$s(1 + O(l^2)) = \epsilon(1 + O(l^2)) + O(l^2),$$

so that $s = \epsilon + O(l^2)$. Thus, since $\beta' = \beta + O(s)$, we have

$$\beta' = \beta + O(\epsilon + l^2). \tag{2.4}$$

From (2.3) it now follows that

$$\begin{aligned} t(1 + O(l^2)) &= l + \epsilon \sin \alpha' - s \sin \beta' \\ &= l + \epsilon(\alpha' + O(l^3)) - (\epsilon + O(l^2))(\beta' + O(l^3)), \end{aligned}$$

so that from the fact that $\alpha' = \alpha + O(\epsilon)$ and (2.4) it follows that

$$\begin{aligned} |d - c| = t &= l + \epsilon(\alpha' - \beta') + O(l^3) \\ &= l + \epsilon(\alpha + O(\epsilon)) - \epsilon(\beta + O(\epsilon + l^2)) + O(l^3) \\ &= l - \epsilon(\beta - \alpha) + O(\epsilon^2 + l^3). \end{aligned}$$

Since $\beta - \alpha = \Delta\theta(ab)$, we are done. □

Henceforth we use the notation $Df(s)$ to denote $f'(s)$.

PROPOSITION 2.5 (length change estimate). [illegible] θ [illegible] K [illegible] G [illegible] $Q = abcd$ [illegible] $z(s)$ $0 \le s \le \alpha$ [illegible] $z(0) = a$ [illegible] ab [illegible] $s \in [0, \alpha]$ [illegible] $E(s) = z(s)w(s)$ [illegible] ca [illegible] $z(s)$ [illegible] $D\theta(z(s_0)) = \kappa_0$ [illegible] $s_0 \in (0, \alpha)$ [illegible] $\lambda(E(s_0)) = \lambda_0$ [illegible] $\delta > 0$ [illegible] $\tau > 0$ [illegible] $s \in [0, \alpha]$ [illegible] $|z(s) - z(s_0)| < \tau$,

$$|w(s) - w(s_0)| = |z(s) - z(s_0)|(1 - P\lambda_0\kappa_0 + R),$$

[illegible] $\frac{1}{K+\delta} \le P \le K + \delta$, [illegible] $|R| \le \delta$

[illegible]. By replacing G with an appropriate subdomain which contains Q, we can assume that θ is L-Lipschitz on G for some L. Let

$$\psi_0 = \psi_0(s) = \theta(z(s)) - \theta(z(s_0)).$$

For notational convenience we assume that $s > s_0$, the opposite case being effectively the same. Let $u_0(\sigma)$, $0 \le \sigma \le \lambda_0$, be the arc length parameterization of $E(s_0)$ with $u_0(0) = z(s_0)$. For each $\sigma \in [0, \lambda_0]$, let $u(\sigma)$ be the point of $E(s)$ joined to $u_0(\sigma)$ by an i-arc. Let $M(s)$ and $m(s)$ denote the maximum and minimum of $|u(\sigma) - u_0(\sigma)|$ for $0 \le \sigma \le \lambda_0$. Obviously, $M(s) \to 0$ as $s \to 0$. Let $\overline{\tau}$ be so small that

$$M(\overline{\tau}) < \overline{l}(L, \operatorname{dist}(Q, \partial G)), \tag{2.5}$$

where $\overline{l}$ is as in Proposition 2.4. For $\mu \in (0, 1]$ let $\tau_0 \le \overline{\tau}$ be such that

$$|\psi_0(s) - \kappa_0|z(s) - z(s_0)|| < \mu(|\kappa_0| + 1)|z(s) - z(s_0)| \quad \text{for} \quad |s - s_0| < \tau_0. \tag{2.6}$$

We can define an increasing sequence of numbers in $[0, \lambda_0]$ as follows. Let $\sigma_0 = 0$, and let σ_{k+1} be the first number in $[\sigma_k, \lambda_0]$ (as long as there is one) for which $\epsilon_k = |u_0(\sigma_{k+1}) - u(\sigma_k)| = \mu l_k$, where $l_k = |u_0(\sigma_k) - u(\sigma_k)|$. By (2.6) we have

$$|\psi_0 - \kappa_0 l_0| < \mu(|\kappa_0| + 1)l_0. \tag{2.7}$$

Bearing in mind that by the K-quasi-HP property $\Delta\theta(u_0(\sigma_k)u(\sigma_k)) = P_k\psi_0$, where P_k is between $\frac{1}{K}$ and K, we may apply Proposition 2.4 (in light of (2.5)) to conclude that

$$l_{k+1} = l_k - \mu l_k P_k \psi_0 + O((\mu^2 + l_k)l_k^2). \tag{2.8}$$

We continue defining σ_{k+1} until we come to $k = N$, for which $|u_0(\sigma_N) - w(s_0)| < \mu l_N$. Such an N exists, and indeed $(N+1)\mu m(s) \le \lambda_0$. We denote by $T \ge 1$ the constant of the big-O in (2.8). Note that, by the preceding proposition, T depends only on L and $\operatorname{dist}(Q, \partial G))$, and so is independent of the value of μ we ultimately decide to work with.

We now restrict μ to satisfy

$$0 < \mu \le \mu_1 = \frac{1}{T(1+\Lambda_0)\Lambda_0} \le 1, \tag{2.9}$$

where

$$\Lambda_0 = 1 + K(|\kappa_0| + 2)\lambda_0.$$

Note that μ_1 depends only on K, κ_0, λ_0, and T. We show that for $l_0 \le \tau_1 = \min\{\tau_0, \mu^2\}$ we have $l_k \le \Lambda_0 l_0$ for all k. Assume inductively that $l_j \le \Lambda_0 l_0$, $0 \le j \le k$. We now use the facts that $P_k \le K$, and $\Sigma\mu l_j = \Sigma\epsilon_j \le \lambda_0$, and that $|\psi_0| \le (|\kappa_0|+1)l_0$ in light of (2.9) and (2.7). It follows from (2.8) that

$$l_{k+1} = l_0 + \sum_{j=0}^{k} l_{j+1} - l_j = l_0 + \sum_{j=0}^{k}(-\mu l_j P_j \psi_0 + O((\mu^2 + l_j)l_j^2))$$

$$\le l_0 + K|\psi_0|\lambda_0 + T\sum_{j=0}^{k}(\mu^2 + l_j)l_j^2 \le l_0(1 + K(|\kappa_0| + 1)\lambda_0) + \sum_{j=0}^{k} T(\mu^2 + l_j)l_j^2.$$

However,

$$\sum_{j=0}^{k} T(\mu^2 + l_j)l^2 \le T\sum_{j=0}^{k}(\mu^2 + \Lambda_0 l_0)\frac{\epsilon_j}{\mu} l_j \le T\frac{(\mu^2 + \Lambda_0 l_0)\Lambda_0 l_0 \lambda_0}{\mu}$$

$$\le T\frac{(\mu^2 + \Lambda_0\mu^2)\Lambda_0 l_0 \lambda_0}{\mu} \le T\mu(1+\Lambda_0)\Lambda_0 l_0 \lambda_0. \tag{2.10}$$

But by (2.9) this last expression is at most $l_0\lambda_0 \le K l_0 \lambda_0$, so that

$$l_{k+1} \le l_0(1 + K(|\kappa_0| + 1)\lambda_0) + K l_0 \lambda_0 = \Lambda_0 l_0,$$

so that indeed $l_k \le \Lambda_0 l_0$, for all k and all $l_0 \le \tau$, provided only that μ satisfies (2.9) and $l_0 \le \tau_1$. Since $M(s) \to 0$ as $s \to 0$, there is a $\tau_2 = \tau_2(\mu) \le \tau_1$ such that

$$(1-\mu)\lambda_0 \le \sum_{k=0}^{N-1} \epsilon_k \le \lambda_0 \quad \text{for} \quad l_0 \le \tau_2. \tag{2.11}$$

Assume the sequence of σ_k stops at $k = N$. Then $|w(s_0) - u(\sigma_N)| < \mu l_N \le \mu\Lambda_0 l_0$, and, since we are assuming $l_0 = |z(s) - z(s_0)| < \tau_2 \le \mu^2$, Proposition 2.4 implies that

$$||w(s) - w(s_0)| - l_N| \le \mu\Lambda_0 l_0 K\psi_0 + T(\mu^2\Lambda_0^2 l_0^2 + \Lambda_0^3 l_0^3).$$

From this it is easy to see that there is some $\mu \le \mu_1$, which depends only on K, κ_0, λ_0, and T, such that

$$||w(s) - w(s_0)| - l_N| \le \mu l_0. \tag{2.12}$$

As previously,

$$l_N - l_0 = \sum_{k=0}^{N-1} l_{k+1} - l_k = -\sum_{k=0}^{N-1} (\mu l_k P_k \psi_0 + O((\mu^2 + l_k) l_k^2)).$$

By (2.10), the sum of the big-O terms of the immediately preceding displayed line is bounded above by $\Lambda_1 \mu l_0$, where $\Lambda_1 = T(1 + \Lambda_0)\Lambda_0$. For $\psi_0 \ge 0$ it follows from this together with (2.11) and (2.12) that

$$\begin{aligned} |w(s) - w(s_0)| - l_0 &= l_N - l_0 + (|w(s) - w(s_0)| - l_N) \\ &\ge -\sum_{k=0}^{N-1} K\mu l_k \psi_0 - (\Lambda_1 + 1)\mu l_0 \ge -K\psi_0\lambda_0 - (\Lambda_1 + 1)\mu l_0. \end{aligned}$$

But since by (2.7), $\psi_0 \le (\kappa_0 + (|\kappa_0| + 1)\mu) l_0$, it follows that $|w(s) - w(s_0)| - l_0 \ge A l_0$, where

$$A = -K\lambda_0\kappa_0 - \mu(K\lambda_0(|\kappa_0| + 1) + \Lambda_1 + 1).$$

Since (2.7) also implies that $\psi_0 \ge (\kappa_0 - (|\kappa_0| + 1)\mu) l_0$, in an analogous manner one sees that for $\psi_0 \ge 0$

$$\begin{aligned} |w(s) - w(s_0)| - l_0 &\le -\sum_{k=0}^{N-1} \frac{\mu l_k \psi_0}{K} + (\Lambda_1 + 1)\mu l_0 \\ &\le \left(-\frac{1}{K}(1 - \mu)\lambda_0(\kappa_0 - \mu(|\kappa_0| + 1) + (\Lambda_1 + 1)\mu \right) l_0, \end{aligned}$$

so that $|w(s) - w(s_0)| - l_0 \le B l_0$, where

$$B = -\frac{1}{K}\lambda_0\kappa_0 + \mu\left(\frac{1}{K}\lambda_0\kappa_0 + \frac{(1 - \mu)}{K}(|\kappa_0| + 1) + \Lambda_1 + 1 \right).$$

By considering separately the cases $\kappa_0 = 0$ and $\kappa_0 \ne 0$ we obtain the desired conclusion when $\psi_0 \ge 0$ with an appropriate $\mu = \mu(K, \kappa_0, \lambda_0, T, \delta)$. The case $\psi_0 \le 0$ is the same apart from straightforward reversal of inequalities. □

Let ϕ be a real-valued function defined on (α, β), and let $s_0 \in (\alpha, \beta)$. We denote by $D^-\phi(s_0)$ and $D^+\phi(s_0)$ the lower and upper limits of $\frac{\phi(s) - \phi(s_0)}{s - s_0}$ as $s \to s_0$. The remaining three propositions of this section all follow easily from Proposition 2.5 together with elementary measure theory.

PROPOSITION 2.6 (curvature-characteristic length bound). [illegible] $z = z(s)$, $\alpha < s < \beta$, [illegible] K [illegible] θ [illegible] G [illegible] J [illegible] j [illegible] $z(s_0)$ [illegible] C [illegible] ∂G [illegible] $D^+\theta(z(s_0)) \le \frac{K}{\lambda(J)}$.

In particular this says that if C is an i-arc whose curvature $\kappa(p)$ at p exists, then the j-half-characteristic emanating from p towards the [illegible] side of C (that is, to the left or right of C according as $\kappa(p) > 0$ or $\kappa(p) < 0$) has length at most $\frac{K}{\kappa(p)}$.

PROPOSITION 2.7 (length monotonicity). *abcd* θ G $z = z(s)$, $\alpha \le s \le \beta$, $z(\alpha) =$ a ab E λ ab $D\theta(z(s)) \le 0$ E (ab *abcd* E) E' cd E ac $\lambda(E') \ge \lambda(E)$

In section 4 we shall make use of the following two lower bounds for the area of certain regions made up of families of characteristic arcs of a K-quasi-HP function θ; the constants η and η' which appear in them depend solely on K.

PROPOSITION 2.8 (area bounds). C i K θ $w \in C$, $C'(w)$ j C w λ_j $U = \cup\{C'(w) : w \in C\}$

(i) $\mu(U) \ge \eta\lambda(C)\lambda_j$

(ii) $\mu(U) \ge \eta'\lambda_j^2|\Delta\theta(C)|$

3. Extended characteristics and regular and singular boundary behavior. Our approach to boundary behavior requires the examination of curves which are in effect characteristics whose interiors (i.e., sets of nonendpoints) contain boundary points; the subtleties that arise in this connection require careful discussion. Hereafter the symbol $\mathcal{G}$ will denote the family of all Jordan domains $G \subset \mathbb{C}$ for which ∂G is a C^2 curve, and $\mathcal{G}(\rho) \subset \mathcal{G}$ will denote the family of all $G \in \mathcal{G}$ such that for each $p \in \partial G$ the interior of one of the circles of radius ρ tangent to ∂G at p is contained in G and that of the other such circle is contained in $\mathbb{C}\backslash\overline{G}$. Obviously, for $G \in \mathcal{G}(\rho)$ the unsigned curvature of ∂G is everywhere bounded by $\frac{1}{\rho}$, and $\mathcal{G} = \cup\{\mathcal{G}(\rho) : \rho > 0\}$. Furthermore, $\mathrm{HP}(G, K)$ will denote the family of K-quasi-HP functions on G. Although sometimes the hypotheses of the propositions of this section do not state so explicitly, they always deal with $G \in \mathcal{G}$ and $\theta \in \mathrm{HP}(G, K)$. (Although the results of this paper apply to unbounded and multiply connected domains as well, the proof of the main theorem itself will entail only consideration of Jordan domains, and in fact we will be able to work largely with Jordan domains of the kind we call "characteristic subdomains," as defined below.) By an C we shall henceforth mean a continuous one-to-one mapping $z = z(t)$ of a closed interval $[\alpha, \beta]$ into the closure $\overline{G}$ of G, and $z((\alpha, \beta))$ will be referred to as the of C. As in section 2, when an arc is considered to be oriented, we use the term "to (towards) the right (left) of C" to refer to the part of C (immediately) to the right (left) of C, and the term will refer to a complete integral curve of θ or $\theta + \frac{\pi}{2}$ in G. The following proposition was proved in [G5, Proposition 2.8] for HP nets. Although the proof is virtually the same in the present more general context we include it here for the sake of completeness.

PROPOSITION 3.1. *C* G $z = z(s)$, $s \in (\alpha, \beta)$ $\lim_{s\to\beta} z(s)$ ∂G

. Clearly, the conclusion holds if $\beta \ne \infty$, so that we assume $\beta = \infty$. First of all, we show that $\mathrm{dist}(z(s), \partial G) \to 0$ as $s \to \infty$. If this were not true, then there would be a $z_0 \in G$ and an $\epsilon > 0$ such that for some sequence $\{s_i\}$ tending to ∞, $z(s_i) \to z_0$, but $z([s_i, s_{i+1}]) \cap \partial N(z_0, \epsilon) \ne \emptyset$. But from this it would follow that some orthogonal characteristic crosses C twice, an impossible occurrence in light of the simple connectivity of G. We can now show that, in fact, $z(s) \to b \in \partial G$ as $s \to \infty$. If this is not so, the foregoing then implies that there is an arc E of ∂G, $\lambda(E) > 0$, each point of which is an accumulation point of $C_\gamma = \{z(s) : s > \gamma\}$ for each $\gamma \in (\alpha, \infty)$. Since G is bounded and $\beta = \infty$, C cannot be a straight line, so that from the characteristic length bound it follows that there is an orthogonal

half-characteristic C' of finite length which joins some $z(\sigma)$ to a point $e \in \partial G$. Since C cannot cross C' twice in G, $C_\sigma \subset G \backslash C'$. Let z_1, z_2 be distinct points of $E \backslash \{e\}$. For each $\delta > 0$, C_σ has a subarc $pp' \subset N(\partial G, \delta) \backslash C'$, with $p, p' \in N(z_1, \delta)$ and a point $p'' \in pp' \cap N(z_2, \delta)$. For obvious topological reasons, for each sufficiently small δ, there must be a point q on pp' which is joined to a point in $N(z_1, \delta)$ by an orthogonal characteristic arc B of length at least $|z_1 - z_2| - 2\delta$ such that the curvature of C at q exists and tends to infinity as $\delta \to 0$ and C is concave towards the side from which B emanates. But this clearly violates the characteristic length bound (Proposition 2.6), as indicated in the sentence immediately following it. □

DEFINITION 3.2. [illegible] C [illegible] $z(\alpha)$, $z(\beta) \in \partial G$ [illegible] $z((\alpha, \beta)) \subset G$ [illegible] elementary i-characteristic

In other words, an elementary characteristic is a full characteristic together with its endpoints, which are well defined by the preceding proposition. Note that for each $p \in \partial G$, $\{p\}$ is a [illegible] elementary i-characteristic. Subarcs of elementary i-characteristics will be called i [illegible] (or simply i-arcs, for short).

LEMMA 3.3. [illegible] $G \in \mathcal{G}$ [illegible] $B' = B'(G) \in (0, 1]$ [illegible] [illegible] $a, b \in \partial G$ [illegible] C_1 [illegible] C_2 [illegible] [illegible] ∂G [illegible] a [illegible] b ($a = b$, $C_1 = \{a\}$ [illegible] $C_2 = \partial G$) [illegible]

$$\operatorname{dist}(z, C_1) + \operatorname{dist}(z, C_2) \geq B' \operatorname{dist}(z, \{a, b\}) \quad \text{[illegible]} \quad z \in G.$$

[illegible]. This is self-evident. □

PROPOSITION 3.4 (bounded length of characteristics). [illegible] $G \in \mathcal{G}$ [illegible] $M = M(G, K)$ [illegible] $\lambda(C) \leq M$ [illegible] C [illegible] $\theta \in \mathrm{HP}(G, K)$

[illegible]. Let θ be a K-quasi-HP function on G, and let C be an elementary i-characteristic of θ. We regard C as being oriented and let $A = C \cap \partial G$ (A, being the set of endpoints of C, has at most two points). Let $d_0 = \operatorname{diam}(G)$, and let $d = \sup\{\operatorname{dist}(z, A) : z \in C\} \leq d_0$. Clearly,

$$\mu(\{z \in G : \operatorname{dist}(z, A) \leq 8r\}) \leq Br^2, \tag{3.1}$$

where $B = 128\pi$. For $k \geq 0$, let

$$G_k = \left\{ z \in G : \frac{d}{2^k} \leq \operatorname{dist}(z, A) \leq \frac{d}{2^{k-1}} \right\}$$

and

$$C_k = C \cap G_k,$$

so that $C \cap G = \cup\{C_k : k \geq 1\}$. For each nonendpoint p of C, let $J(p)$ denote the elementary j-characteristic containing p. Obviously, $J(p) \cap J(p') \cap G = \emptyset$ for $p \neq p'$. For $k \geq 1$ and $p \in C_k$ let $l(p)$ and $r(p)$ be the first points encountered when moving along $J(p)$ from p to the left and right of C, respectively, which are not in the interior of $G_{k-1} \cup G_k \cup G_{k+1}$. Each $q \in \{l(p), r(p)\}$ is either on ∂G or is in G and satisfies one of $\operatorname{dist}(q, A) = \frac{d}{2^{k+1}}$ or $\operatorname{dist}(q, A) = \frac{d}{2^{k-2}}$. Say $q \notin \partial G$ and $\operatorname{dist}(q, A) = \frac{d}{2^{k+1}}$. Then $|p - q| \geq \frac{d}{2^{k+1}}$, since otherwise $\operatorname{dist}(p, A) \leq |p - q| + \operatorname{dist}(q, A) < \frac{d}{2^k}$, which contradicts the fact that $p \in C_k$. If $q \notin \partial G$ and $\operatorname{dist}(q, A) = \frac{d}{2^{k-2}}$, then $|p - q| \geq \frac{d}{2^{k+1}}$ since otherwise

$$\operatorname{dist}(p, A) \geq \operatorname{dist}(q, A) - |p - q| > \frac{d}{2^{k-2}} - \frac{d}{2^{k+1}} > \frac{d}{2^{k-1}},$$

which is inconsistent with $p \in C_k$. Thus $|p-q| \geq \frac{d}{2^{k+1}}$ if $q \notin \partial G$. Hence, if at least one of $l(p)$, $r(p)$ is not in ∂G, we have for the open subarc $J_1(p)$ of $J(p)$ with endpoints $l(p)$, $r(p)$ that $\lambda(J_1(p)) \geq \frac{d}{2^{k+1}}$. If, on the contrary, $\{l(p), r(p)\} \subset \partial G$, then it follows from the preceding lemma that

$$\lambda(J_1(p)) \geq |l(p) - p| + |p - r(p)| \geq \frac{B'd}{2^k} > \frac{B'd}{2^{k+1}},$$

so that this bound holds in all cases for $p \in C_k$ since $B' \leq 1$. Since $J_1(p) \backslash \{l(p), r(p)\} \subset G_{k-1} \cup G_k \cup G_{k+1}$ for $p \in C_k$, it follows from (3.1) that

$$\begin{aligned} B\left(\frac{d}{2^{k+1}}\right)^2 &\geq \mu\left(\left\{z \in G : \frac{d}{2^{k+1}} \leq \operatorname{dist}(z, \partial G) \leq \frac{d}{2^{k-2}}\right\}\right) \\ &= \mu(G_{k-1} \cup G_k \cup G_{k+1}) \\ &\geq \mu(\cup\{J_1(p) : p \in C_k\}) \geq \eta\lambda(C_k)\frac{B'd}{2^{k+1}}, \end{aligned}$$

by Proposition 2.8(i). From this we have that $\lambda(C_k) \leq \frac{Bd}{\eta B' 2^{k+1}}$. But since $C = \cup\{C_k : k \geq 1\}$, we conclude that $\lambda(C) \leq \frac{Bd}{2B'\eta} \leq \frac{Bd_0}{2B'\eta}$. Since $B = 128\pi$ and B' and d_0 depend only on G, and η depends only on K, we are done. □

DEFINITION 3.5. [illegible] C [illegible] $p \in \partial G$ [illegible] exit G [illegible] p [illegible] C [illegible] $z = z(s)$, $0 \leq s \leq L$ [illegible] $\lim_{s\to 0} \theta(z(s))$ [illegible] C [illegible] regularly [illegible] $z(0)$ [illegible] singularly

PROPOSITION 3.6. [illegible] $G \in \mathcal{G}$ [illegible] $\theta \in \mathrm{HP}(G, K)$ [illegible] $C = ab \subset G$ [illegible] i [illegible] θ [illegible] E [illegible] a [illegible] b [illegible] G [illegible] $E \cap C = \{a, b\}$ [illegible] $\operatorname{diam}(C) \leq 5\lambda(E)$

[illegible]. Let D be the interior of the simple closed curve $C \cup E$. Obviously, $D \subset G$. Let $p \in C$ and $q \in E$ satisfy

$$|p - q| = \sup\{|z - w| : z \in C \text{ and } w \in E\}.$$

Let C' be the elementary j-characteristic containing p. Since a j-characteristic can have at most one point in common with an i-characteristic, there is a point $q' \in E$ such that C' contains a subarc $J = pq'$ whose interior lies in D. Let $\operatorname{diam}(C) = |z_1 - z_2|$, where z_1, $z_2 \in C$. Then

$$\operatorname{diam}(C) = |z_1 - z_2| \leq |z_1 - a| + |a - b| + |b - z_2| \leq 2|p - q| + \lambda(E),$$

since $|z_1 - a|$, $|b - z_2| \leq |p - q|$. But

$$|p - q| \leq |p - q'| + |q' - q| \leq \lambda(J) + \lambda(E),$$

so that

$$\operatorname{diam}(C) \leq 2\lambda(J) + 3\lambda(E).$$

Let J be parameterized by $w(s)$, $0 \leq s \leq l$, with $w(0) = p$. Let I^+ and I^- be the set $s \in (0, l)$ at which $D\theta(w(s)) \geq 0$ and $D\theta(w(s)) \leq 0$, respectively. For $s \in I^+$ ($s \in I^-$) let E^+ (E^-) be the set of points of E joined to $w(s)$ by an i-arc emanating to the

right (left) of J. It follows by a simple argument based on Proposition 2.7 (length monotonicity) that $\lambda(I^+) \le \lambda(E^+)$ and $\lambda(I^-) \le \lambda(E^-)$. But then

$$\lambda(J) \le \lambda(I^+) + \lambda(I^-) \le \lambda(E^+) + \lambda(E^-) \le \lambda(E),$$

from which the desired bound follows immediately. □

PROPOSITION 3.7. [illegible] $p \in \partial G$ [illegible] $q_1, q_2 \in G$ [illegible] $C_1 = pq_1$ [illegible] $C_2 = pq_2$ [illegible] i [illegible] K [illegible] θ [illegible] G [illegible] p [illegible] $J \subset G$ [illegible] j [illegible] c_1 [illegible] $C_1 \cap G$ [illegible] c_2 [illegible] $C_2 \cap G$ [illegible] C_1 C_2 [illegible] J [illegible] T [illegible] $P \subset J$ [illegible] J [illegible] T [illegible] $\lambda(P) = 0$

[illegible]. Suppose not. Then, since the elementary i-characteristic passing through any point of J must exit at p, after replacing the original J by an appropriate subarc and changing C_1 and C_2 accordingly, we can assume that $\lambda(P \cap J) = \epsilon > 0$, $\lambda(N) < \frac{\epsilon}{8}$, and $\lambda(\theta(J) < \frac{1}{100K}$, where $N = J \backslash P$. Let C_k be parameterized by $z_k(s)$, $0 \le s \le \lambda_k$, with $z(0) = p$. Let $\{k, l\} = \{1, 2\}$. Let $J_k(s)$ denote the j-arc joining $z_k(s)$ to a point $w_k(s) \in C_l \cap G$. Note that we know only that $J_k(s)$ is defined for $s \le \lambda_k$ sufficiently near λ_k. It follows from length monotonicity (Proposition 2.7) that $\lambda(J_k(s)) \ge \lambda(P_k(s)) \ge \epsilon$, where $P_k(s)$ is the set of points of $J_k(s)$ joined to points of P by an i-arc. By the quasi-HP property $\lambda(\theta(J_k(s))) < \frac{1}{100}$, so that $J_k(s)$ is almost straight and in particular the distance between its endpoints is at least $\frac{1}{2}\lambda(J_k(s)) \ge \frac{1}{2}\lambda(P_k(s)) \ge \frac{\epsilon}{2}$. Let ξ_k be the infimum of all s for which $J_k(s)$ is defined. Since the distance between the endpoints of $J_k(s)$ is at least $\frac{\epsilon}{2}$, it is clear that at least one of ξ_1, ξ_2 must be positive, and for definiteness we assume that $\xi_1 > 0$. For $\sigma \in (\xi_1, \lambda_1]$ let $J_1(\sigma)$ be parameterized by $\zeta(s, \sigma)$, $0 \le s \le \lambda(J_1(\sigma))$, with $\zeta(0, \sigma) \in C_1$. It is clear that there are δ, $T > 0$ such that

$$\operatorname{dist}(\zeta(s, \sigma), C_1 \cup C_2) \ge Ts \quad \text{for} \quad s \in (0, \delta), \quad \sigma \in (\xi_1, \lambda_1]. \tag{3.2}$$

From the fact that a j-arc can intersect an i-arc at most once in G it easily follows that for each point $z \in C_1 \cap G$ there is a $\delta_1 = \delta_1(z)$ such that

$$|z - \zeta(s, \sigma)| \ge \delta_1 \quad \text{for} \quad s \in [\delta, \lambda(J_1(\sigma))], \quad \sigma \in (\xi_1, \lambda_1]. \tag{3.3}$$

From (3.2) and (3.3) together with the fact that for $\sigma \in (\xi_1, \lambda_1]$, $J_1(\sigma) = J_2(\sigma')$ for some $\sigma' \in (\xi_2, \lambda_2]$ it follows that as $\sigma \to \xi_1$, $J_1(\sigma)$ tends to an arc J_0 which contains p, which joins $z_1(\xi_1)$ to the point $z_2(\xi_2)$ of C_2, and the distance between whose endpoints is at least $\frac{\epsilon}{2}$. Furthermore, either J_0 consists of a j-arc joining $z_1(\xi_1)$ to p or (in the case that $\xi_2 > 0$) it consists of such an arc together with another j-arc joining p to $z_2(\xi_2)$. One of the endpoints of J_0, which we henceforth call q, is at a distance of at least $\frac{\epsilon}{4}$ from p. By renaming, if necessary, we can assume that $q = z_1(\xi_1) \in C_1$. Let J_0' be the subarc of J_0 which joins q to p. Let A denote the arc pq of C_1. Then J_0' and A form the two sides of a "characteristic bilateral" B. Let P_0' be the subset of points of J_0' which correspond to (i.e., are joined by i-arcs to) points of P (that is, P_0' is, apart from a set of linear measure 0, the set of points at which J_0 is nonconcave towards the inside of B), and let $N_0' \subset J_0'$ be the subset corresponding to N. Since, by length monotonicity, $\lambda(N_0') \le \lambda(N) \le \frac{\epsilon}{8}$, it follows that

$$\lambda(P_0') \ge \lambda(J_0') - \lambda(N_0') \ge \frac{\epsilon}{4} - \frac{\epsilon}{8} = \frac{\epsilon}{8}. \tag{3.4}$$

Let J_0' be parameterized by $z_0(s)$, $0 \le s \le \lambda(J_0')$, with $z_0(0) = p$, and for $s \in (0, \lambda(J_0'))$ let $A(s)$ be the part of the elementary i-characteristic through $z_0(s)$ in $\overline{B}$, so that $A(s)$ joins $z_0(s)$ to p (since its interior can cross neither A nor J_0'). It follows from Proposition 3.6 that $\operatorname{diam}(A(s)) \to 0$ as $s \to 0$. Let qp' be an arc of A for which $\lambda(\theta(qp')) < \frac{1}{100K}$. Since $\lambda(\theta(W)) < \frac{1}{100}$ for any translate W of either J_0' or qp' by the quasi-HP property, it follows that one can translate qp' in B all the way down $J_0' \backslash \{p\}$ from q without meeting the boundary of B, since for simple geometric reasons all these translates, being essentially perpendicular to the virtually straight arc J_0', must stay away from p. If $A'(s)$ denotes the translate of qp' with initial point $z_0(s)$, then $A'(s) \subset A(s)$, and therefore $\operatorname{diam}(A'(s)) \to 0$ as $s \to \lambda(J_0')$. This means that each point $z \in qp'$ is joined to p in B by a j-arc $J''(z)$ such that if $P''(z)$ is the set of points of $J''(z)$ joined to points of P by i arcs in T, then by (3.4) and length monotonicity $\lambda(P''(z)) \ge \frac{\epsilon}{8}$ and by the quasi-HP property $\lambda(\theta(J''(z))) < \frac{1}{100}$, so that

$$|z - p| \ge \frac{\epsilon}{16}. \tag{3.5}$$

We can now repeat this process starting with $J''(p')$ instead of $J_0' = J''(q)$ and continue doing so to obtain in the end $J''(z)$ for all $z \in A \backslash \{p\}$. However, by the argument we just gave we now have (3.5) for all $z \in A$, which is absurd since p is an endpoint of A. Therefore $\lambda(P) = 0$. □

PROPOSITION 3.8. [illegible] G [illegible]. This follows easily from Proposition 3.6. □

PROPOSITION 3.9. [illegible]. Assume to the contrary that points $p_1 \ne p_2$ of ∂G are joined by distinct elementary i-characteristics C_1 and C_2. It follows from Proposition 3.8 that the elementary i-characteristic through any point of the simply connected domain D bounded by $C_1 \cup C_2$ must also have endpoints p_1 and p_2. But then it follows easily from Proposition 3.7 that all j-characteristic arcs in D are straight line segments. But this contradicts Proposition 3.7. □

PROPOSITION 3.10. [illegible] G [illegible] G [illegible] $p \in G$ [illegible]. The proof is an easy consequence of Proposition 3.7 and the quasi-HP property. □

DEFINITION 3.11. [illegible] θ [illegible] G [illegible] C_0 [illegible] i [illegible] θ [illegible] $a, b \in \partial G$ [illegible] B [illegible] ∂G [illegible] a, b [illegible] D [illegible] G [illegible] $\partial D = C_0 \cup B$ [illegible] i-characteristic subdomain

When we wish to indicate the elementary i-characteristic involved, we will denote the i-characteristic subdomain by (D, C_0). The arc $B = \overline{\partial D - C_0} \subset \partial G$, called the [illegible] of (D, C_0) and denoted by $\operatorname{bot}(D)$, will be considered to have the order "$<$" corresponding to the positive orientation of ∂D. We shall freely use interval notation as well as the terms "to the right of," "to the left of," "between," etc., when dealing with $\operatorname{bot}(D)$. Furthermore, if ab is an elementary i-characteristic joining points a, b of $\operatorname{bot}(D)$, it will be understood that $a \le b$, unless otherwise indicated. When dealing with a characteristic subdomain (D, C_0), we shall work with the class $\mathcal{I}(D)$ of nontrivial elementary i-characteristics C which join points p, q of $\operatorname{bot}(D)$. Note that $C_0 \in \mathcal{I}(D)$. If $C = pq \in \mathcal{I}(D)$, then "above" C refers to the part of D not in the closed region bounded by $C \cup [p, q]$. If $F \subset \overline{D}$ is a compact set for which $F \cap \operatorname{bot}(D) \ne \emptyset$, we

say that an elementary i-characteristic $E = pq \in \mathcal{I}(D)$ [illegible] F, and write $F \preceq E$, if F is contained in the closed set bounded by the simple closed curve $[p, q] \cup pq$. We deal only with sets F for which each component of F has points in $\mathrm{bot}(D)$. For two such sets F_1, $F_2 \subset \overline{D}$ we say that F_1 [illegible] F_2 and write $F_1 \leq F_2$ if for all f_1, f_2 with $f_k \in F_k \cap \mathrm{bot}(D)$, $k = 1, 2$, there holds $f_1 \leq f_2$. It is clear that if C_1, $C_2 \in \mathcal{I}(D)$, then one of the following is true: $C_1 \preceq C_2$, $C_2 \preceq C_1$, $C_1 \leq C_2$, or $C_2 \leq C_1$. The first and second of these possibilities include the case $C_1 = C_2$.

In what follows, $L(G, p, d)$ will denote the segment $\overline{pq}$ of length d which is orthogonal to ∂G at p, which emanates from p into G and which is oriented from p to q.

PROPOSITION 3.12. [illegible] $\omega_0 = \omega_0 < 1$ [illegible] $\omega_1 = \omega_1(K, \rho)$ [illegible] $G \in \mathcal{G}(\rho)$, $p \in \partial G$, $\theta \in \mathrm{HP}(G, K)$ [illegible] C [illegible] $z(s)$, $0 \leq s \leq \lambda(C)$ [illegible] i [illegible] ρ [illegible]

$$C \cap L(G, p, \omega_0 \rho) = \{z(0)\}, \tag{3.6}$$

[illegible] $L(G, p, \omega_0\rho)$ [illegible] j [illegible] [illegible] C [illegible] ρ [illegible] ∂G [illegible] $w(s)$, $0 \leq s < \lambda(\partial G)$, [illegible] $w(0) = p$ [illegible] $|\arg\{z'(0)\} - \arg\{w'(0)\}| \leq \frac{\pi}{2}$ [illegible]

$$\arg\{z'(0)\} \geq \arg\{w'(0)\} - \omega_1 \sqrt{|z(0) - p|}$$

[illegible] C [illegible] $L(G, p, \xi_0)$ [illegible]

$$\arg\{z'(0)\} \geq \arg\{w'(0)\} + \omega_1 \sqrt{|z(0) - p|}$$

[illegible] C [illegible] $L(G, p, \omega_0)$ [illegible]. Clearly, it is enough to handle the case in which C emanates to the right of $L(G, p, \rho)$. Without loss of generality we can assume that $p = 0 = \arg\{w'(0)\}$, so that in particular $D = N(-\rho i, \rho) \subset \mathbb{C} \backslash G$ and $\partial N(-\rho i, \rho)$ is tangent to ∂G at 0. Let $d = |z(0) - p|$ and let $R_d = \{di + te^{-i\alpha} : t \geq 0\}$, $\alpha = \alpha(d) \in (0, \frac{\pi}{2})$ be the ray emanating to the right of $L(G, 0, \rho)$ from the point $di \in L(G, 0, \rho)$ which is tangent to ∂D, and let the point of tangency be z_d. Then $\alpha = \cos^{-1}(\frac{\rho}{\rho + d})$, so that α is bounded below and above by $\sqrt{1 - \frac{\rho}{\rho+d}}$ and $2\sqrt{1 - \frac{\rho}{\rho+d}}$, respectively. Thus

$$2\sqrt{d/\rho} \geq \alpha \geq \frac{\sqrt{d/\rho}}{2} \quad \text{for} \quad d \leq \rho. \tag{3.7}$$

Let T_d be the curvilinear triangle bounded by $L(G, 0, d)$, the line segment $[di, z_d]$ and the (shorter) arc of ∂D with endpoints 0, z_d. Then in light of (3.7) it is easy to see that

$$\mathrm{diam}(T_d) \leq \rho\alpha + d \leq 2\rho\sqrt{d/\rho} + d = 2\sqrt{\rho d} + d \quad \text{for} \quad d \leq \rho, \tag{3.8}$$

so that

$$\mathrm{diam}(T_d) < \rho \quad \text{for} \quad d \leq \frac{\rho}{16}. \tag{3.9}$$

Now assume that C is as in the hypothesis with $z(0) = di$, where $d \leq \frac{\rho}{16}$, and that

$$-\beta = \arg\{z'(0)\} \leq -N\alpha.$$

Here $N > 1$ is a number yet to be determined. Since, by (3.9), $\operatorname{diam}(C) \geq \rho > \operatorname{diam}(T_d)$, and since obvious angle considerations imply that C enters T_d at di, in light of assumption (3.6), C must exit T_d at a point of the segment $[di, z_d]$. Say s_0 is the smallest value of s for which $z(s) \in [di, z_d]$. Then obviously $\theta([0, s_0]) \supset [-\beta, -\alpha]$, so that $[0, s_0]$ has a subinterval $[s_1, s_2]$ for which $\theta([s_1, s_2]) = [-\beta, -\alpha]$. Since $\beta - \alpha \leq \frac{\pi}{2}$, and since the length of a curve given by $y = f(x)$, $x_1 \leq x \leq x_2$ for which $|f'(x)| \leq 1$ is at most $\sqrt{2}|x_2 - x_1|$, it follows that

$$s_2 - s_1 = \lambda(z([s_1, s_2]) \leq \sqrt{2}|z(s_2) - z(s_1)|$$

$$< 2\operatorname{diam}(T_d) \leq 2(2\sqrt{\rho d} + d),$$

by (3.8). Thus by the mean value theorem there must be some $s_0 \in [s_1, s_2]$ for which

$$D^+\theta(z(s_0)) \geq \frac{(N-1)\alpha}{2(2\sqrt{\rho d} + d)}. \tag{3.10}$$

But the curvature bound (Proposition 2.6) together with the hypothesis regarding the j-characteristics implies that $D^+\theta(z(s_0)) \leq \frac{K}{\rho}$, so that in light of (3.10) and the lower bound in (3.7) we have

$$\frac{\rho(N-1)\sqrt{d/\rho}}{8\sqrt{\rho d} + 4d} \leq K.$$

But the left-hand side of this inequality is $(N-1)/(8 + 4\sqrt{d/\rho})$, which is bounded below by $\frac{N-1}{9}$ for $d \leq \frac{\rho}{16}$, so that we have a contradiction for $N = 9K + 2$. Thus if $d \leq \frac{\rho}{16}$, we must have

$$\arg\{z'(0)\} \geq -2\frac{9K+2}{\sqrt{\rho}}\sqrt{d} \geq -\frac{22K\sqrt{d}}{\sqrt{\rho}},$$

in light of the upper bound in (3.7) and the fact that $K \geq 1$. Thus we have proved the proposition with $\omega_0 = \frac{1}{16}$ and $\omega_1 = \frac{22K}{\sqrt{\rho}}$. □

PROPOSITION 3.13. [illegible] $\xi_0 = \xi_0(K, \rho)$ [illegible] $\xi_1 = \xi_1(K, \rho)$ [illegible] $G \in \mathcal{G}(\rho)$, $\theta \in \mathrm{HP}(G, K)$ [illegible] $C = ab$ [illegible] i [illegible] (D, C) [illegible] C [illegible] $A = \operatorname{bot}(D)$ [illegible] $z(s)$, $0 \leq s \leq \lambda(C)$, [illegible] $w(s)$, $0 \leq s \leq \lambda(A)$, [illegible] $z(0) = w(0) = a$, a [illegible] $\operatorname{bot}(D)$ [illegible] $p = w(\sigma_0) \in A$, [illegible] $\operatorname{dist}(p, \partial G \backslash A) \geq \rho$, [illegible] $L(G, p, \xi_0)$ [illegible] C, [illegible] $z(s)$ [illegible]

$$|\arg\{z'(s)\} - \arg\{w'(\sigma_0)\}| \leq \xi_1\sqrt{|z(s) - p|}. \tag{3.11}$$

[illegible]. Without loss of generality we may assume that $p = 0 = \arg\{w'(\sigma_0)\}$. Let ω_0 and ω_1 be as in the preceding proposition. Let $L(\epsilon) = L(G, p, \epsilon)$. We assume for the moment that $z(s) \in L(\omega_0\frac{\rho}{2})$ is the only point of C on this segment. Since immediately to the left of C there are points in the complement of D and the interior of the segment $L(|z(s)|)$ lies in D, it is clear that if we write $\arg\{z'(s_0)\} = e^{i\tau}$, with $-\pi < \tau \leq \pi$, then $|\tau| \leq \frac{\pi}{2}$. Let E be either of the two subarcs of C, one of whose endpoints is $z(s)$ and the other of which is a point $q \in G$ for which $|z(s) - q| = \frac{\rho}{2}$. It follows from the preceding proposition that if $\omega_1(K, \frac{\rho}{2})\sqrt{|z(s)|} \leq \frac{\pi}{4}$, then C cannot

be tangent to $L(\omega_0\frac{\rho}{2})$. Thus, of the two arcs E, one moves to the right of $L(\omega_0\frac{\rho}{2})$ as we move along it away from $z(s)$ and the other moves to the left. But then by the preceding proposition we have (3.11) with

$$\xi_1 = \omega_1\left(K, \frac{\rho}{2}\right) \tag{3.12}$$

for any $z(s)$ for which

$$|z(s)| \leq \min\left\{\omega_0\frac{\rho}{2}, \left(\frac{\pi}{4\omega_1(K,\rho/2)}\right)^2\right\}.$$

Let

$$\xi_0 = \min\left\{\omega_0\frac{\rho}{10}, \left(\frac{\pi}{4\omega_1(K,\rho/2)}\right)^2, \frac{\pi\rho}{21K}\right\}. \tag{3.13}$$

Assume $\epsilon \leq \xi_0$ and that $L(\epsilon)$ contains at least two points of C. Then there are s', $s'' \in (0, \lambda(C))$, $s' < s''$, such that the interior of $L(\min\{|z(s')|, |z(s'')|\})$ contains no point of C (and is therefore contained in D) and $z(s')z(s'') \cap L(\epsilon) = \{z(s'), z(s'')\}$. By Proposition 3.6, $\operatorname{diam}(z(s')z(s'')) \leq 5\epsilon < \omega_0\frac{\rho}{2}$. There are the following two cases.

(i) $|z(s')| < |z(s'')|$. In this case, simple topological arguments show that $z(s')$ $z(s'')$ must lie to the right of $L(\epsilon)$. Also, by the foregoing and our definition of ξ_0, one easily has

$$|\arg\{z'(s')\} - \arg\{w'(\sigma_0)\}| \leq \frac{\pi}{4}.$$

But then since C crosses $L(\epsilon)$ again at $z(s'')$, we must have that for some $t' < t''$ in (s', s''), $\arg\{z'(t'')\} - \arg\{z'(t')\} \geq \frac{\pi}{2} - \frac{\pi}{4} = \frac{\pi}{4}$, so that by mean value considerations as in the proof of the preceding proposition together with the curvature bound we see that there must be a point $\sigma' \in (t', t'')$ at such that

$$\frac{K}{\rho/2} \geq \frac{d\arg\{z'(s)\}}{ds}|_{s=\sigma'} \geq \frac{\pi/4}{2\operatorname{diam}(z(s')z(s''))} \geq \frac{\pi}{10\epsilon} \geq \frac{\pi}{10\xi_0},$$

which implies that $\xi_0 \geq \frac{\pi\rho}{20K}$, a contradiction, since $\xi_0 \leq \frac{\pi\rho}{21K}$.

(ii) $|z(s')| > |z(s'')|$. In this case, simple topological arguments show that $z(s')$ $z(s'')$ must lie to the left of $L(\epsilon)$, and we proceed analogously to the way we did in case (i).

Thus, for $\epsilon \leq \xi_0$, $L(\epsilon)$ can contain at most one point of C. This completes the proof of the proposition with ξ_1 and ξ_0 as defined in (3.12) and (3.13), respectively. □

We now extend the notion of characteristic to include certain arcs whose interiors contain points of ∂G. Although what follows does not give the most exhaustive extension possible, it is sufficient for our present needs. Let (D_0, C_0) be an i-characteristic subdomain of G, and consider a monotone decreasing sequence $\{C_k = a_kb_k : k \geq 0\}$ in $\mathcal{I}(D_0)$. In other words, $a_k \leq a_{k+1} < b_{k+1} \leq b_k$, $k \geq 0$, so that the arcs $A_k = [a_k, b_k]$ of $\operatorname{bot}(D_0) = A_0$ are nested. Let $D_k \subset D_0$ be the i-characteristic subdomain bounded by $C_k \cup A_k$, so that $D_k \supset D_{k+1}$, $k \geq 0$. We regard C_k as being oriented from a_k to b_k. Let $a_k \to a$ and $b_k \to b$. We define C to be the set of limits of sequences $\{z_l\}$, where $z_l \in C_{k_l}$, $k_l \to \infty$. Any such C will be called an [illegible] a [illegible] b. An extended characteristic consisting of a single point $p \in \partial G$ will be called

[illegible]. The following proposition, in the statement and proof of which the notation is the same as that of the immediately preceding sentences, contains the basic properties of extended characteristics. Note that $M(G,K)$ is, as in Proposition 3.4, an upper bound on the length of elementary characteristics of K-quasi-HP functions on G.

PROPOSITION 3.14. [illegible] C [illegible] a [illegible] b [illegible] $a = b$, [illegible] $C = \{a\}$ [illegible] C [illegible] a [illegible] b [illegible] $\lambda(C) \leq M(G,K)$ [illegible] C [illegible] $z(s)$, $0 \leq s \leq \lambda(C)$, [illegible] $z(0) = a$, [illegible] $C \cap \partial G \subset [a,b]$ [illegible] $C \cap [a,b]$ [illegible] $A_0 = \mathrm{bot}(D_0)$ [illegible] s [illegible] z [illegible] $(0, \lambda(C))$, [illegible] $t \in (0, \lambda(C))$, $z(t)$ [illegible] $\partial G \backslash (a,b)$ [illegible] j [illegible] $J(t)$ [illegible] C [illegible]

$$D^+ \arg\{z'(t)\} \leq \frac{K}{\lambda(J(t))}. \tag{3.14}$$

[illegible]. We begin by observing that

$$A_0 \cap C \subset [a,b]. \tag{3.15}$$

To see this, note that for each $w \in A_0 \backslash [a,b]$ there is an n for which $w \notin A_n$. Since $C_n \cap A_0 = \{a_n, b_n\}$, w is not in C_n either, and therefore $w \notin \overline{D_n}$. But then, by the monotonicity of $\{\overline{D_k}\}$,

$$\mathrm{dist}(w, C_k) \geq \mathrm{dist}(w, \overline{D_k}) \geq \mathrm{dist}(w, \overline{D_n}) > 0 \text{ for } k \geq n,$$

from which the desired conclusion follows at once. From (3.15) it follows immediately that $C \cap \partial G \subset [a,b]$.

Next we note that if $a = b$, then Proposition 3.6 implies that $\mathrm{diam}(C_k) \to 0$, so that $C = \{a\}$. For the remainder of the proof we therefore assume that $a < b$.

By this assumption and Proposition 3.4 on the boundedness of the lengths of characteristics there exist l_1 and l_2 such that $0 < l_1 \leq \lambda(C_k) \leq l_2 < \infty$. Let C_k be parameterized by $z = z_k(s)$, $0 \leq s \leq \lambda(C_k)$. Let $0 < \epsilon < |b-a|/3$, and let n_0 be such that

$$|a_k - a|,\ |b_k - b| < \epsilon \quad \text{for} \quad k \geq n_0.$$

For $k \geq n_0$ let

$$\alpha_k = \sup\{s : z_k(s) \in \partial N(a,\epsilon)\}$$

and

$$\beta_k = \inf\{s : z_k(s) \in \partial N(b,\epsilon)\},$$

and let $E_k = E_k(\epsilon) = z_k([\alpha_k, \beta_k])$. Then

$$\delta = \inf\{\mathrm{dist}(E_k, \partial G \backslash A_k) : k \geq n_0\} > 0, \tag{3.16}$$

since otherwise there would be a point of C in $A_0 \backslash [a,b]$, in contradiction of (3.15). For $k \geq n_0$ and $s \in [\alpha_k, \beta_k]$, let $J_k(s)$ denote the j-half-characteristic emanating to the left of C_k and joining $z_k(s)$ to a point $w_k(s)$ of $\partial G \backslash A_k$. In light of (3.16),

$$\lambda(J_k(s)) \geq \delta > 0 \quad \text{for} \quad k \geq n_0, \quad s \in [\alpha_k, \beta_k]. \tag{3.17}$$

This means that for each $\epsilon > 0$ we have an upper bound on the curvature to the left of E_k. More precisely, the curvature bound implies that

$$D^+\theta(z_k(s)) \le \frac{K}{\lambda(J_k(s))} \le \frac{K}{\delta} \quad \text{for} \quad k \ge n_0, \quad s \in [\alpha_k, \beta_k]. \tag{3.18}$$

In addition, the curvature bound trivially implies that

$$D^-\theta(z_k(s)) \ge -\frac{K}{\operatorname{dist}(z_k(s), \partial G)}. \tag{3.19}$$

There is a neighborhood U of ∂G such that for each $z \in U$ there is a unique $p(z) \in \partial G$ for which $|z - p(z)| = \min\{\operatorname{dist}(z, \zeta) : \zeta \in \partial G\}$ and for which p is continuous. For q in ∂G, let $e^{i\phi(q)}$ be the positively oriented unit tangent to ∂G at q, so that, in U, $e^{i\phi(p(z))}$ is continuous. It follows from (3.18) and Proposition 3.13 that there is a $\Lambda = \Lambda(\epsilon)$ such that

$$|z_k'(s) - e^{i\phi(p(z_k(s)))}| \le \Lambda\sqrt{|z_k(s) - p(z_k(s))|} \quad \text{for} \quad k \ge n_0, \quad s \in [\alpha_k, \beta_k]. \tag{3.20}$$

Bounds (3.18), (3.19), and (3.20) imply that the family $\{z_k'(s)\}$ is uniformly bounded and equicontinuous (that is, there is a single modulus of continuity valid for all $z_k'(s)$, $k \ge n_0$, on the respective $[\alpha_k, \beta_k]$). From this it follows that there are α, β, and a sequence $\{k_l\}$ such that $\alpha_{k_l} \to \alpha$ and $\beta_{k_l} \to \beta$ and $z_{k_l}(s) \to z_\epsilon(s)$ uniformly (in the obvious sense), where z_ϵ is continuously differentiable and parameterizes an arc $C(\epsilon)$ which joins a point of $\partial N(a, \epsilon)$ to a point of $\partial N(b, \epsilon)$ in the part of $\overline{D_0}$ lying outside of both these circles. The arc $C(\epsilon)$ is simple, since otherwise there would be some $q \in (a, b)\backslash N(\{a, b\}, \epsilon)$ such that for arbitrarily small $\tau > 0$, $L(G, q, \tau)$ intersects some E_k more than once, which would contradict Proposition 3.13. Clearly, $\lambda(C(\epsilon)) \le M(G, K)$. The nested nature of the D_k then implies that (at least for sufficiently small ϵ) the entire sequence converges to $C(\epsilon)$. Also, it is clear that $C(\epsilon')$ is an extension of $C(\epsilon)$ for $\epsilon' < \epsilon$. We have that $C = \cup\{C(\epsilon) : \epsilon > 0\}$, since $\operatorname{diam}(z_k([0, \alpha_k]))$ and $\operatorname{diam}(z_k([\beta_k, \lambda(C_k)]))$ tend to 0 uniformly in k as $\epsilon \to 0$ by Proposition 3.6. It follows from $C = \cup\{C(\epsilon) : \epsilon > 0\}$ that $\lambda(C) \le M(G, K)$ and that C is a simple arc with endpoints a, b, and furthermore that C is parameterized by a function $z(s)$ which is continuously differentiable on $(0, \lambda(C))$. That the two possible orderings of the points of $C \cap [a, b]$ coincide follows from the fact that C is a simple arc and (3.15).

The existence and uniqueness of $J(t)$ is trivial when $z(t) \in G$. When $z(t) \in (a, b)$ the existence of $J(t)$ follows from a straightforward compactness argument. In light of (3.20), for such t, any corresponding $J(t)$ must be orthogonal to ∂G at $z(t)$, so that the uniqueness of $J(t)$ follows from Proposition 3.10. Bound (3.14) follows from (3.18). □

We will refer to an extended characteristic C joining a, $b \in \partial G$ as ab; points of $ab \cap \partial G$ will be called [illegible], and contact points other than a and b will be called [illegible]. It is clear that θ can be continuously extended to $G \cup (ab\backslash\{a, b\})$. For what is to follow it is important to understand that if (D_0, C_0) is an i-characteristic subdomain, then, in addition to the extended characteristics constructed above, D_0 might contain extended characteristics C' joining points $a' < b'$ arising from a sequence of i-characteristic subdomains $\{(D_k', C_k')\}$, where $C_k' = a_k'b_k'$ with $a_{k+1}' \le a_k' < b_k' \le b_{k+1}'$ (where the order is with respect to $A_0 = \operatorname{bot}(D_0)$, as above). Here the C_k' are contained in D_0, but the other part of the boundary of D_k' is $\partial G\backslash(a_k', b_k')$ (where here again the interval notation refers to the order on

$A_0 = \text{bot}(D_0)$). Note that if for such an extended characteristic C', $C' \cap \partial G$ has points other than a' and b' (that is, if C' is not simply an elementary characteristic), then the contact points will not occur monotonically with respect to the order on A_0 when C' is traversed from a' to b'. [illegible] D_0 the extended characteristics C constructed originally will be referred to as [illegible], and this other kind of extended characteristic C' with proper contact points will be said to be [illegible]. We consider that an extended characteristic ab exits G at all of its contact points, and we use the terms "exits regularly" and "exits singularly" at a or b in the obvious fashion. Clearly, ab exits regularly at its proper contact points.

PROPOSITION 3.15. [illegible] C [illegible] i [illegible] p [illegible] $z(s)$, $0 \le s \le L$, [illegible] $z(0) = p$ [illegible] $\phi(s) = \arg\{z'(s)\}$ [illegible] $\lim_{s\to 0} \phi(s)$ [illegible] $s \in (0, L)$, [illegible] $E(s)$ [illegible] j [illegible] $z(s)$ [illegible] $\epsilon, T > 0$, [illegible] $P(\epsilon, T) = \{s < \epsilon : \phi'(s) \ge T\}$ [illegible] $N(\epsilon, T) = \{s < \epsilon : \phi'(s) \le -T\}$ [illegible] $\epsilon, T > 0$, $P(\epsilon, T)$ [illegible] $N(\epsilon, T)$ [illegible] $\text{diam}(E(s)) \to 0$ [illegible] $s \to 0$ [illegible] $E(s)$ [illegible] p

[illegible]. Assume to the contrary that $\lambda(P(\epsilon, T)) > 0$ for some ϵ_0, $T_0 > 0$, $\lambda(P(\epsilon_0, T_0)) = 0$, so that $\phi'(s) \le T_0$ a.e. on $(0, \epsilon_0]$. Then $\phi = \phi^+ + \phi^-$ on $(0, \epsilon_0]$, where ϕ^+ is Lipschitz continuous and nondecreasing and ϕ^- is continuous and nonincreasing. From this in turn it follows that either ϕ^- has a finite limit as $s \to 0$ or it tends to ∞ as $s \to 0$, so that in fact the latter is the case since C exits singularly. But this means that C spirals around p, which is clearly impossible. Thus $P(\epsilon, T)$ must have positive measure for all $\epsilon, T > 0$. One sees similarly that $N(\epsilon, T)$ has positive measure. That $\text{diam}(E(s)) \to 0$ follows immediately from the characteristic length bound. To see that the last sentence of the statement is true, assume to the contrary that $E(\sigma)$ joins $z(\sigma)$ to p. Then, since distinct j-characteristics have no common point in G and can only cross C at one point in G, $E(s)$ exits at p for $0 < s \le \sigma$. But this is impossible in light of Proposition 3.7 and the fact that both $P(\epsilon, 1)$ and $N(\epsilon, 1)$ have positive measure. □

Let (D, C_0) be a characteristic subdomain of G, and let $F \subset D \cup \text{bot}(D)$ be a compact set for which $F \cap \text{bot}(D) \ne \emptyset$. The family $\mathcal{C}(F) = \{C \in \mathcal{I}(D) : F \preceq C\}$ is clearly linearly ordered with respect to the relation $\preceq$. From this it easily follows that there is a unique monotone extended i-characteristic $E_0 = ab$ for which $F \cap \text{bot}(D) \subset [a, b]$ and $E_0 \preceq C$ for all $C \in \mathcal{C}(F)$.

NOTATION 3.16. [illegible] i [illegible] E_0 [illegible] $\min_D(F)$

In the statement and proof of the following proposition all order relations are with respect to (D, C_0).

PROPOSITION 3.17 (structure of $\min_D(A \cup B)$). [illegible] (D, C_0) [illegible] i [illegible] G [illegible] $a_1 < a_2 < b_1 < b_2$ [illegible] $a_1a_2 = A$ [illegible] $b_1b_2 = B$ [illegible] $\mathcal{I}(D)$ [illegible] $U = \min_D(A \cup B) = ef$, $e < f$ [illegible]

(i) U [illegible] A [illegible] B [illegible] $ef \in \mathcal{C}(A \cup B)$ [illegible] i [illegible] ab [illegible] $e \le a \le b \le f$ [illegible]

(ii) ea [illegible] $L \in \mathcal{I}(D)$ [illegible] $A \preceq L \preceq U$, [illegible] $\preceq L$ [illegible]

(iii) bf [illegible] $L \in \mathcal{I}(D)$ [illegible] $B \preceq L \preceq U$, [illegible] $A \preceq L$ [illegible]

[illegible]. It follows from the hypotheses that $e < f$ (i.e., that $e \ne f$). We show that

if (i) does not hold, then at least one of (ii) or (iii) does. Let

$$e' = \sup\{z : z \in \text{bot}(D) \cap \min_D(A \cup B),\ z \le a_1\}$$

and

$$f' = \inf\{z : z \in \text{bot}(D) \cap \min_D(A \cup B),\ z \ge b_2\}.$$

Since (i) does not hold it is easy to see that the subarc $e'f'$ of ef is a member of $\mathcal{C}(A \cup B)$, so that in fact $e'f' = ef$. Obviously, $ef \cap (A \cup B) \subset \{e, f\}$, since otherwise ef would have to be A or B. Let w_0 be any point of ef other than e or f, and consider a small j-arc E whose initial point is w_0, which extends to the right of ef (that is, into the simply connected domain bounded by $ef \cup [e, f]$) and which is disjoint from $A \cup B$. Let E be parameterized by $z = w(s)$, $0 \le s \le \lambda_0$, with $w(\lambda_0) = w_0$, and let $C(s)$ denote the elementary i-characteristic through $w(s)$. Note that

$$C(s_1) \preceq C(s_2) \quad \text{for} \quad s_1 \le s_2. \tag{3.21}$$

Let the left and right endpoints of $C(s)$ be $l(s)$ and $r(s)$. By the minimality of ef it must be that for no $s \in (0, \lambda_0)$ can we have both $l(s) \le a_1$ and $r(s) \ge b_2$. From (3.21) it therefore follows that either for all $s \in (0, \lambda_0)$, $l(s) > a_1$, or for all $s \in (0, \lambda_0)$, $r(s) < b_2$. Assume the latter occurs. Then in fact $r(s) \le b_1$ for all $s \in (0, \lambda_0)$, since otherwise $C(s)$ would cross B in D, which is impossible since they are distinct elementary i-characteristics. Consider the i-characteristic subdomain $(D', C(0))$ bounded by $C(0)$ and $\partial G \backslash (l(0), r(0))$. If we take any sequence $\{s_k\}$ in $(0, \lambda_0)$ which tends monotonically to λ_0, and consider the corresponding sequence of elementary i-characteristics $C_k = C(s_k)$ in D', it is clear that they will give rise to an extended i-characteristic $C' = ab$ (with, as always, $a < b$ with respect to the order for $\text{bot}(D)$), which is nonmonotone with respect to the characteristic subdomain D, which contains ef as a subarc and for which $b \le b_1$. Clearly, $bf \cap ef = \{f\}$, so that $bf \preceq ef$. If we let

$$c' = \sup\{z : z \in \text{bot}(D) \cap bf,\ z \le b_1\}$$

and

$$c'' = \inf\{z : z \in \text{bot}(D) \cap bf,\ z \ge b_2\},$$

then $L = c'c'' \in \mathcal{I}(D)$, so that from the minimality of ef it follows that $c' \ge a_2$ and we have conclusion (iii). In exactly the same manner one obtains conclusion (ii) in the case that $s \in (0, \lambda_0)$, $l(s) > a_1$ for all $s \in (0, \lambda_0)$. □

COMMENT 3.18. (ii) ab $L' \in \mathcal{I}(D)$ $B \preceq L' \preceq U$ (ii) (iii) ab A (iii) (i) ef A B A B

DEFINITION 3.19. $p \in \partial G$ regular boundary point θ $\theta(z)$ $z \to p$ singular boundary point

PROPOSITION 3.20. [illegible] $p \in \partial G$ [illegible] p [illegible] $\phi \in (0, \pi)$ [illegible] ∂G [illegible] p

[illegible]. This follows immediately from the Peano existence theorem for (local) solutions of the initial value problem $y' = F(x, y)$, $y(x_0) = y_0$ when F is continuous in a neighborhood of (x_0, y_0). □

DEFINITION 3.21. [illegible] $p \in \partial G$ [illegible] θ [illegible] (θ, π) [illegible] p [illegible] θ [illegible] singularity of type 0 (singularity of type 1) [illegible] p [illegible] θ [illegible] [illegible] p [illegible] θ [illegible] singularity of type 2 [illegible] p

The following proposition is needed to make use of fans of characteristics in what is to follow.

PROPOSITION 3.22. [illegible] $C_1 = pq_1$ [illegible] $C_2 = pq_2$ [illegible] [illegible] i [illegible] [illegible] A [illegible] [illegible] ∂G [illegible] q_1 [illegible] q_2 [illegible] [illegible] p [illegible] T [illegible] [illegible] [illegible] $C_1 \cup C_2 \cup A$ [illegible] [illegible]

(i) [illegible] $d > 0$ [illegible] $z \in N(p, d) \cap T$ [illegible] [illegible] i [illegible] [illegible] z [illegible] p

(ii) p [illegible] [illegible] C

[illegible]. Let W be the set of all $z \in T$ for which the elementary i-characteristic $C(z)$ through z does not exit at p. For $z \in W$, $C(z)$ has no points in common with either $C_1 \backslash \{q_1\}$ or $C_2 \backslash \{q_2\}$, so that both endpoints of $C(z)$ lie on A. For each $z \in W$, let $D(z)$ denote the interior of the domain bounded by $C(z)$ and the arc of A whose endpoints are those of $C(z)$. For $w_1, w_2 \in T$, $C(w_1) \cap T$ and $C(w_2) \cap T$ are either identical or disjoint, so that for w_1, $w_2 \in W$, $D(w_1)$ and $D(w_2)$ are either nested or disjoint. Let $\xi = \frac{1}{2}\text{dist}(p, A) > 0$. If $z \in N(p, \frac{\xi}{2}) \cap W$, then it follows from Proposition 2.8(i) that $\mu(D(z)) \geq \eta\xi^2$ since $\lambda(C(z) \cap N(p, \xi)) \geq \xi$ and each of the j-arcs joining a point of $C(z) \cap N(p, \xi)$ to A in $D(z)$ has length at least ξ. Thus if $z_1, \ldots, z_n \in N(p, \frac{\xi}{2}) \cap W$ are such that the corresponding $D(z_k)$ are disjoint, then $n \leq \frac{\mu(G)}{\eta\xi^2}$. If (i) is not true, a pigeonhole (area exhaustion) argument then shows that there is a sequence $\{z_k : k \geq 1\}$ of points of W tending to p such that $\{D(z_k)\}$ is an increasing sequence. But our extended characteristic construction then gives us a monotone extended characteristic C (with respect to the characteristic subdomain bounded by $C(z_1)$ and the complement in ∂G of the subarc of A joining the endpoints of $C(z_1)$) for which $p \in C$. But the endpoints of C are in A since those of all the $D(z_k)$ are in A, so that p is a proper contact point of C. Thus C satisfies (ii). □

DEFINITION 3.23 (i-fan). [illegible] θ [illegible] i-fan [illegible] $p \in \partial G$ [illegible] p [illegible] [illegible] i [illegible] [illegible] θ [illegible] p [illegible] [illegible] [illegible] i [illegible] [illegible] p [illegible] $\mathcal{F}_i(p)$ [illegible] p [illegible] vertex [illegible]

Proposition 3.22 implies that if C_1 and C_2 are distinct elementary i-characteristics in $\mathcal{F}_i(p)$, then there is a $d > 0$ such that all points of $N(p, d)$ between them belong to members of $\mathcal{F}_i(p)$, so that the interior of $\cup\{C : C \in \mathcal{F}_i(p)\}$ is nonempty. Furthermore, it follows from Proposition 3.7 that the j-characteristic through any point in the interior of $\cup\{C : C \in \mathcal{F}_i(p)\}$ is strictly concave towards the side facing p. These two facts in turn imply that if there are i-fans at $p_1 \neq p_2$, then the interiors of $\cup\{C : C \in \mathcal{F}_i(p_k)\}$, $k = 1, 2$, must be disjoint. Since $\mu(\cup\{C : C \in \mathcal{F}_i(p)\}) > 0$, we have the following.

PROPOSITION 3.24.

PROPOSITION 3.25. $R = (R_1, R_2)$

. Let F be an extended i-characteristic which exits regularly at p. Without loss of generality we can assume that the positive direction along ∂G at p is that of the positive real axis. Let C be a nontrivial arc of F parameterized by $z = z(s)$, $0 \le s \le L$, with $z(0) = p$. Again, without loss of generality we can assume that $\gamma = \arg\{z'(s)\} = \theta(z(s))$. We consider the following three possibilities separately.

(i) $\gamma \in (0, \frac{\pi}{2}) \cup (\frac{\pi}{2}, \pi)$. In this case we can assume that C is an arc of an elementary i-characteristic exiting at p. To be specific, we assume that $\gamma \in (0, \frac{\pi}{2})$. Let q be an interior point of C. For any $\epsilon > 0$ there exist nontrivial j-characteristic arcs $E^+ = E^+(\epsilon)$ and $E^- = E^-(\epsilon)$ emanating from q to the right and left of C, respectively, such that $\lambda(\theta(E^+ \cup E^-)) < \epsilon$. Since there is no fan at p, none of the j-characteristics through any point of $E^+ \cup E^-$ other than q exits at p. Let $C(\epsilon)$ be an initial segment of C such that $\lambda(\theta(C \backslash \{p\})) < \epsilon$. Let $\alpha < \gamma$; we show that $\theta(z) \to \gamma$ as $z \to p$ between C and the ray $\arg\{z\} = \alpha$. It is easy to see that some initial segment of this ray (that is, the portion of the ray contained in $N(p, \xi)$ for some $\xi > 0$) is covered by a collection $\mathcal{Q}(\epsilon)$ of characteristic quadrilaterals Q, one of whose i-sides is a subarc of $C(\epsilon)$ and one of whose j-sides is a translate of an initial subarc of $E^+(\epsilon)$. The quasi-HP property then implies that there is a $\delta = \delta(\epsilon)$ such that if $|z - p| < \delta$ and z is between C and the ray $\arg\{z\} = \alpha$, then $|\theta(z) - \gamma| < 2K\epsilon$. Since ϵ is arbitrary, $\theta(z) \to \gamma$ as $z \to p$ between C and the ray. It is also clear that for sufficiently small ϵ the translate of E^- down to p is a nontrivial initial arc of a j-characteristic C' which exits regularly at p and which forms with ∂G an acute angle of size $\frac{\pi}{2} - \gamma$. What we have shown in regard to C now implies that if $\beta < \frac{\pi}{2} - \gamma$, then $\theta(z) + \frac{\pi}{2} \to \gamma + \frac{\pi}{2}$ as $z \to p$ between C' and the ray $\arg\{z\} = \pi - \beta$, so that $\theta(z) \to \gamma$ as $z \to p$ in that curvilinear sector. Finally, it is easy to see that $\theta(z) \to \gamma$ as $z \to p$ between C and C'. Since $\alpha < \gamma$ and $\beta < \frac{\pi}{2} - \gamma$ are arbitrary, we have the desired regularity at p. The case $\gamma \in (\frac{\pi}{2}, \pi)$ is handled in the same manner apart from minor changes of a notational nature. In the case that θ arises from the solution of a normal system it is clear from (ii) of Definition 1.1 that both $R_i(z)$ and $R_j(z)$ have limits as $z \to p$ in the sector $\alpha < \arg\{z\} < \beta$, so that the second conclusion is valid in this case.

(ii) $\gamma = \frac{\pi}{2}$. Here again we can take C to be an arc of an elementary i-characteristic exiting at p and proceed as in case (i). For any $\alpha < \frac{\pi}{2}$ it follows as in case (i) that $\theta(z) \to \gamma$ as $z \to p$ between C and the ray $\arg\{z\} = \alpha$. Again it is immediate that the same holds in the curvilinear sector between C and the ray $\arg\{z\} = \pi - \alpha$. The second conclusion is likewise immediate.

(iii) $\gamma \in \{0, \pi\}$. We deal with the case $\gamma = 0$, the case $\gamma = \pi$ being essentially the same. Here we can define $E^-(\epsilon)$ as in case (i), but because there is no fan at p the translate of $E^-(\epsilon)$ down to p is a nontrivial arc of a j-characteristic which is orthogonal to ∂G at p. This puts us in case (ii), so we are done. □

DEFINITION 3.26. $p \in \partial G$ i-singularity

The following is an immediate consequence of Proposition 3.15.

PROPOSITION 3.27. j C p

PROPOSITION 3.28. [illegible] 2-[illegible] i-[illegible] $i = 1$ [illegible] 2 [illegible]. Let C be any elementary i-characteristic, and let D be the characteristic subdomain bounded by C and the arc of ∂G containing p. Let $E = ef = \min_D(\{p\})$. If $E = p$, we are done, so that $E = ef$ is a nontrivial extended characteristic. If $p \in E$, Proposition 3.25 implies that (i) $p \in \{e, f\}$, since extended characteristics exit regularly at their proper contact points. If $p \notin E$, the minimality of E implies that (ii) E is an elementary i-characteristic. We deal with possibility (i) first. Let E be parameterized by $z(s)$, $0 \le s \le l$, with $z(0) = p$. Since p is a singularity of type 2, for no $\sigma > 0$ is it true that $z((0, \sigma)) \subset G$. On the other hand, it follows from the regularity of ∂G and Proposition 3.25 that for no $\sigma > 0$ is it true that $z((0, \sigma)) \subset \partial G$. It is then clear that E contains a sequence $\{C_k\}$ of subarcs which are elementary i-characteristics for which $\operatorname{diam}(C_k)$ and $\operatorname{dist}(C_k, p)$ tend to 0 as $k \to \infty$, so that p is indeed an i-singularity. We finish by showing that possibility (ii) cannot occur; to do so we proceed as in the proof of Proposition 3.17. Let w_0 be any point of E other than e or f, and consider a small j-arc $J \subset G$, which extends from w_0 to the right of E (i.e., into D). Let J be parameterized by $w(s)$, $0 \le s \le \lambda_0$, with $w(\lambda_0) = w_0$, and let $C(s) = l(s)r(s)$ be the elementary i-characteristic through $w(s)$. As in the the proof of Proposition 3.17, it follows from the minimality of E that either $l(s) > p$ for all $s \in (0, \lambda_0)$ or $r(s) < p$ for all $s \in (0, \lambda_0)$. Assume, for definiteness, that the latter occurs. By considering what happens as $s \to 0$ we obtain, as in that proof, a nonmonotone extended i-characteristic ab (with $a < b$ with respect to the order on $\operatorname{bot}(D)$) which contains ef as a proper subarc, for which $b < p$ and for which $bf \preceq ef$. But $p \notin bf$ by hypothesis, so that bf must have a subarc gh which is an elementary i-characteristic for which $g < p < h$, which contradicts the minimality of E. □

The following is an immediate consequence of Propositions 3.25, 3.27, and 3.28 and Definitions 3.21 and 3.26.

PROPOSITION 3.29. [illegible] $p \in \partial G$ [illegible]

(i) p [illegible] 0 [illegible]

(ii) p [illegible] 1 [illegible] k [illegible] $k = 1$ [illegible] 2

(iii) p [illegible] 2 [illegible] k [illegible] $k = 1$ [illegible] 2

COMMENT 3.30. [illegible] 2.2 [illegible] i [illegible] α [illegible] ff. [illegible] 0 [illegible] $\alpha < 1$

PROPOSITION 3.31 (type 1 singularities with fans). [illegible] C [illegible] i [illegible] $p \in \partial G$ [illegible] $q \in G$ [illegible] $z(s)$, $0 \le s \le l$, [illegible] $z(0) = p$ [illegible] $\eta < \frac{\pi}{2}$

$$\phi + \eta \le \arg\{z'(s)\} \le \phi + \pi - \eta \quad \text{[illegible]} \quad 0 < s \le l, \tag{3.22}$$

[illegible] ϕ [illegible] ∂G [illegible] p, $\lim_{s \to 0} \arg\{z'(s)\}$ [illegible] C [illegible] i [illegible] p [illegible]. For $0 < \xi < \pi$, let $R(p, \xi)$ denote the (open) ray

$$R(p, \xi) = \{z : \arg\{z - p\} = \xi\}.$$

Without loss of generality we can assume that $\phi = 0$. It follows immediately from hypothesis (3.22) that C lies in the sector $\{z : \eta \le \arg\{z - p\} \le \pi - \eta\}$. The hypothesis that $\lim_{s \to 0} \arg\{z'(s)\}$ does not exist implies that there is some δ, $0 < \delta \le \frac{\eta}{K+1}$, such

that for arbitrarily small $\sigma_1 > \sigma_2 > 0$

$$|\arg\{z'(\sigma_1)\} - \arg\{z'(\sigma_2)\}| \geq 2\delta. \tag{3.23}$$

For $0 < s \leq l$, let $F(s) = a(s)b(s)$ be the elementary j-characteristic through $z(s)$. It follows from the last sentence of Proposition 3.15 that $a(l) \neq p \neq b(l)$. Let D be the j-characteristic subdomain bounded by $F(l)$ and the arc of ∂G joining $a(l)$ to $b(l)$ and containing p. Clearly, $F(s) \cap C \cap G = \{z(s)\}$, and $a(s)$ is nonincreasing and $b(s)$ is nondecreasing with respect to the order on $\mathrm{bot}(D)$. Since, by Proposition 3.15, $\mathrm{diam}(F(s)) \to 0$ as $s \to 0$, by replacing C by a sufficiently short initial subarc we can assume that

$$R(p, \xi) \cap \partial D \subset F(l), \quad \frac{\delta}{2} \leq \xi \leq \pi - \frac{\delta}{2}. \tag{3.24}$$

Now consider any pair of numbers $\sigma_1 \neq \sigma_2$ in $(0, l)$ for which (3.23) holds. Obviously, for at least one of them, call it σ, there must hold $|\arg\{z'(\sigma)\} - \frac{\pi}{2}| \geq \delta$. Then $F(\sigma)$ must have a subarc $W = uv$ such that $\lambda(\theta(W)) = \delta$ and $z(\sigma) \in \{u, v\}$. To see this, say, for example, that $\arg\{z'(\sigma)\} \leq \frac{\pi}{2} - \delta$. If, along the arc of $F(\sigma)$ emanating to the left of C, the argument of the tangent were always within δ of its argument at $z(\sigma)$, then $F(\sigma)$ would never cross $R(p, \pi - \delta)$, and it would therefore not be able to exit D. An analogous argument may be used in the case that $\arg\{z'(\sigma)\} \geq \frac{\pi}{2} + \delta$.

From this it follows that there is a sequence $s_k \to 0$ of such numbers σ for all of which the corresponding j-arc $W_k = u_k v_k$ lies on one side of C or the other. To be specific, say $u_k = z(s_k)$ and W_k lies to the right of C. It follows from the quasi-HP property that if $W' = z(s)v$ is any translate of W_k along C in the direction of increasing s, then

$$\delta/K \leq \lambda(\theta(W')) \leq K\delta.$$

From this and the fact that $\delta \leq \frac{\eta}{K+1}$ it follows in turn that for $s > s_k$, along any i-characteristic arc whose initial point is in W_k and which is parallel to the subarc $z(s_k)z(s)$ of C, the inclination of the tangent is in the interval

$$[\eta - K\delta, \pi - \eta + K\delta] \subset [(K+1)\delta - K\delta, \pi - (K+1)\delta + K\delta] = [\delta, \pi - \delta].$$

For each k let

$$S_k = \{R(z, \xi) : z \in W_k \text{ and } \xi \in [\delta, \pi - \delta]\}.$$

Now $\mathrm{diam}(W_k) \leq \mathrm{diam}(F(s_k))$ and the latter tends to 0 as $k \to \infty$, so that in light of (3.24) by eliminating a finite number of elements of $\{s_k\}$ and renaming we can assume that all the $R(z, \xi)$ in each of the S_k meet ∂D a point of F before exiting G. From this it follows that the translates of $z(s_k)z(l)$ along W_k all lie in G. This means that for any $s \in [s_k, l]$ the complete translate of W_k along C from $z(s_k)$ up to $z(s)$ belongs to G. For the translate T_k of W_k up to $z(l)$ we have that $z(l) \in T_k \subset F(l)$, and by the quasi-HP property $\lambda(\theta(T_k)) \geq \delta/K$. This bound implies that there is a positive lower bound on the length of the T_k. Thus $\cap\{T_k : k \geq 1\} = T$ is a nontrivial arc of $F(l)$, one of whose endpoints is $z(l)$. However, the translate of T down to $z(s_k)$ is contained in $W_k \subset F(s_k)$. Since $\mathrm{dist}(F(s), \{p\}) \to 0$ as $s \to 0$, it follows that all i-characteristics through T exit at p, and this establishes the existence of the desired fan. □

We need the following elementary lemma.

LEMMA 3.32. [illegible] f [illegible] $[0,T]$ [illegible] $\int_0^T f(x)dx = A > 0$ [illegible] $\rho \in [0, \frac{1}{2})$

$$\lambda(\{\xi : f(\xi) \geq \rho\frac{A}{T} \quad \textit{and} \quad \int_0^\xi f(x)\,dx \geq \rho A\}) > 0.$$

[illegible]. Let $f(x) \leq M$ a.e. on $[0,T]$, and let $Q = \{\xi : f(\xi) \geq \rho\frac{A}{T}\}$. Obviously, $\lambda(Q) > 0$. Let Q^* be the set of density points of Q, so that $\lambda(Q^*) = \lambda(Q) > 0$. Let $\xi_0 = \sup\{\xi : \xi \in Q^*\}$. If the set $Q^* \cap \{\xi : \int_0^\xi f(x)\,dx \geq \rho A\}$ had measure 0, then for each $\delta > 0$ there would be a point $\xi \in (\xi_0 - \delta, \xi_0)$ for which $\int_0^\xi f(x)\,dx < \rho A$, so that

$$\begin{aligned} A &= \int_0^\xi f(x)dx + \int_\xi^{\xi_0} f(x)dx + \int_{\xi_0}^T f(x)dx < \rho A + \delta M + (T - \xi_0)\frac{\rho A}{T} \\ &< 2\rho A + \delta M < A, \end{aligned}$$

for δ sufficiently small, so that $\lambda(Q^* \cap \{\xi : \int_0^\xi f(x)\,dx \geq \rho A\})$ must be positive. ∎

The following proposition plays a fundamental role in the proof of Main Theorem 2.2.

PROPOSITION 3.33 (essentially singularity-free boundary arcs). [illegible] $\overline{\eta} = \overline{\eta}(K,\rho) < \rho$ [illegible] $G \in \mathcal{G}(\rho)$ [illegible] $\theta \in \mathrm{HP}(G,K)$ [illegible] C [illegible] i [illegible] θ [illegible] $p \in \partial G$ [illegible] $E = pp'$ [illegible] j [illegible] ∂G [illegible] p [illegible] D [illegible] j [illegible] G [illegible] E [illegible] A [illegible] ∂G [illegible] $b \in A \cap C$ [illegible] C [illegible] A [illegible] F [illegible] $[b,p']$ [illegible] A [illegible] $\mathrm{dist}(p,F) \geq \rho$ [illegible] $[p,p'']$ [illegible] A [illegible] $|p'' - p| = \overline{\eta}$ [illegible] θ

[illegible]. Without loss of generality we may assume that $p = 0$, that the positively oriented tangent to ∂G at p has the direction of the positive real axis, and that A extends to the right of p. Let C be parameterized by $z(s) = x(s)+iy(s)$, with $z(0) = p$. We shall introduce positive constants, which will depend only on K and ρ. In most cases these constants will be denoted by the same symbol B, but the use of a single symbol to denote different constants will not cause confusion or be misleading since any statement in which B plays a role will be valid if B is taken to be any suitably large number and since this convention is used only finitely many times. The symbol B_1, on the other hand, will refer to a specific constant, which again depends only on K and ρ and which will have the same value every time it is used.

If we give A in nonparametric form by $y = g(x)$, then

$$|g(x)| \leq \frac{x^2}{\rho} \quad \text{and} \quad |g'(x)| \leq \frac{2x}{\rho} \quad \text{for} \quad 0 \leq x \leq \frac{1}{10\rho}. \tag{3.25}$$

Clearly, Proposition 3.12 holds for extended characteristics with the obvious wording changes. Any j-arc emanating to the left of C from any point z of the subarc pb of C will exit G at a point of F, so that the length of any such j-arc is a least $\rho - |p - z|$. Since $\mathrm{dist}(z(s), \partial G) \leq |z(s)| \leq |s|$, it follows from Proposition 3.12 that $|\arg\{z'(s)\}| \leq \frac{1}{2}$ for $0 \leq s \leq \frac{1}{B}$, so that for these values of s we have

$$\frac{1}{B} \leq x'(s) \leq 1 \quad \text{and} \quad |y'(s)| \leq 1,$$

and we can represent C nonparametrically by $y = f(x)$, where $f(x(s)) = y(s)$. For $0 \le s \le \frac{1}{B}$ we have

$$|f'(x(s))| = |\tan(\arg\{z'(s)\})| \le 2|\arg\{z'(s)\}| \le B\sqrt{\operatorname{dist}(z(s), A)}$$
$$\le B\sqrt{|z(s) - x(s)| + \operatorname{dist}(x(s), A)} \le B\sqrt{f(x(s)) + x(s)^2},$$

where the second inequality follows from Proposition 3.12 and the fourth inequality follows from the first bound of (3.25) in the form $\operatorname{dist}(x(s), A)) \le \frac{(x(s))^2}{\rho}$. Thus there is a $B_1 \ge 1$ such that

$$|f'(x)| \le B_1\sqrt{|f(x)| + x^2} \quad \text{and} \quad |f(x)| \le B_1 x \quad \text{for } 0 \le x \le \frac{1}{B_1}. \tag{3.26}$$

For $0 \le x \le \frac{1}{B_1}$ we then have the following. In the first place,

$$|f'(x)| \le B_1\sqrt{B_1 x + x^2} \le 2^{1/2} B_1^{3/2} x^{1/2},$$

so that in fact $|f(x)| \le \frac{2}{3} 2^{1/2} B_1^{3/2} x^{\frac{3}{2}} < B_1^{3/2} x^{\frac{3}{2}}$. Repeating this argument we have that $|f'(x)| \le 2^{1/2} B_1^{7/4} x^{\frac{3}{4}}$, so that in fact $|f(x)| \le \frac{4}{7} 2^{1/2} B_1^{\frac{7}{4}} x^{\frac{7}{4}} < B_1^{\frac{7}{4}} x^{\frac{7}{4}}$, so that $|f'(x)| \le 2^{1/2} B_1^{\frac{15}{8}} x^{\frac{7}{8}}$, and so on. Thus, in fact,

$$|f(x)| \le B_1^2 x^2 \quad \text{for} \quad 0 \le x \le \frac{1}{B_1} \tag{3.27}$$

and

$$|f'(x)| \le \sqrt{2} B_1^2 x \quad \text{for} \quad 0 \le x \le \frac{1}{B_1}. \tag{3.28}$$

Before proceeding we remind the reader that each time the symbol B appears, it may have a value larger than it did at its previous appearance. For any $t \in [0, \frac{1}{2B}]$, let $J(t)$ be the j-half-characteristic emanating to the right of C (that is, downwards) from $t + if(t) = \alpha(t)$ and joining it to a point of A. Note that $J(t)$ will reduce to a single point if $\alpha(t) \in A$. Let $J(t)$ be given parametrically by $\zeta_t(s)$ with $\zeta_t(0) = \alpha(t)$. Let $t_0 = x_0^2$. Then, by (3.25) and (3.27), $|f(t_0) - g(t_0)| \le B x_0^4$ for $0 < x_0 \le \frac{1}{B}$. For notational convenience we can assume without loss of generality that $\arg\{z'(s)\} = \theta(z(s))$. We show that, with an appropriately large value of B, $0 < x_0 \le \frac{1}{B}$ implies that

$$|\theta(\zeta_{t_0}(s_1)) - \theta(\zeta_{t_0}(0))| \le x_0 = \sqrt{t_0} \tag{3.29}$$

for every $\zeta_{t_0}(s_1) \in J(t_0)$. If this were not true, there would be an $s_1 \in (0, \lambda(J(t_0)))$ such that $|\theta(\zeta_{t_0}(s_1)) - \theta(\zeta_{t_0}(0))| = x_0$. Let s_2 be the smallest number in $[0, s_1]$ for which $|\theta(\zeta_{t_0}(s_2)) - \theta(\zeta_{t_0}(0))| = x_0$, so that

$$\lambda(\theta(\zeta_{t_0}([0, s_2]))) \le 2x_0. \tag{3.30}$$

By increasing B if necessary, $t_0 = x_0^2 < \frac{1}{B^2}$ will be so small that

$$s_2 \le 2|f(t_0) - g(t_0)| \le B x_0^4. \tag{3.31}$$

By the preceding lemma (with $\rho = \frac{1}{3}$, $T = s_2$, $A = x_0$) it then follows that there is an $s_3 \in (0, s_2)$ for which

$$|D\theta(\zeta_{t_0}(s_3))| \ge \frac{x_0}{3s_2} \ge \frac{1}{Bx_0^3}$$

and

$$|\theta(\zeta_{t_0}(s_3)) - \theta(\zeta_{t_0}(0))| \geq \frac{x_0}{3}, \tag{3.32}$$

where the expressions inside the absolute values on the left-hand side of these last two bounds have the same sign. Let R be the i-half-characteristic through $\zeta_{t_0}(s_3)$ emanating from $J(t)$ towards its concave side (that is, to the left of $J(t_0)$ if $d\arg\{\zeta'_{t_0}(s)\}/ds$ is positive at s_3 and to the right if it is negative). It follows from the characteristic length bound (Proposition 2.5) that

$$\lambda(R) \leq Bx_0^3. \tag{3.33}$$

We concentrate on the case that R emanates to the left of $J(t_0)$; the opposite case is much easier to handle, as we indicate below. By (3.27), (3.31), and (3.33), R exits G at a point $g(\beta)$ with

$$\begin{aligned} \beta &\leq t_0 + |f(t_0)| + |\zeta_{t_0}(0) - \zeta_{t_0}(s_3)| + \lambda(R) \\ &\leq x_0^2 + O(x_0^3) \leq 2x_0^2 \quad \text{for} \quad x_0 \leq \frac{1}{B}. \end{aligned} \tag{3.34}$$

For $t \geq t_0$ let $\overline{J}(t)$ be the j-arc joining $\alpha(t) \in C$ to a point $r(t) \in R$, so that $\overline{J}(t_0) = \zeta_{t_0}([0, s_3])$. Obviously, $\overline{J}(t)$ is not defined for all $t \in [t_0, \frac{1}{2B_1}]$, but, for any t for which it is defined, $\overline{J}(t_0)$ and $\overline{J}(t)$ are the j-sides of a characteristic quadrilateral $Q(t)$, whose i-sides are the arc $\alpha(t_0)\alpha(t)$ of C and the arc $\zeta_{t_0}(s_3)r(t)$ of R. From the K-quasi-HP property and (3.30) it follows that

$$\lambda(\theta(\overline{J}(t))) \leq K\lambda(\theta(\overline{J}(t_0))) \leq K\lambda(\theta(\zeta_{t_0}([0, s_2]))) \leq 2Kx_0, \tag{3.35}$$

and by (3.28)

$$\lambda(\theta(\zeta_{t_0}(s_3)r(t))) \leq K\lambda(\theta(\alpha(t_0)\alpha(t))) \leq 2K\tan^{-1}(Bt) \leq Bt. \tag{3.36}$$

But for any $t \leq 10x_0^2$ for which $J(t)$ is defined, we have from (3.25) and (3.28) that

$$\begin{aligned} |\theta(\zeta_t(s))| &\leq \theta(\alpha(t)) + 2Kx_0 \leq \tan^{-1}(Bt) + 2Kx_0 \\ &\leq 20Bx_0^2 + 2Kx_0 \leq \frac{1}{100} \end{aligned}$$

for $x_0 \leq \frac{1}{B}$. From this together with (3.34) it is clear that $g(\beta)$ is the endpoint of $\overline{J}(t_1)$ for some $t_1 \in [x_0^2, 10x_0^2]$. It follows from the quasi-HP property that R, when oriented from $\zeta_{t_0}(s_3)$ to $r(t_1)$, has a well-defined (one-sided) tangent, whose argument we denote by ϕ_0. But then we have from (3.28), (3.32), (3.36), and the quasi-HP property that

$$\phi_0 \geq \frac{x_0}{3} - |\theta(\alpha(t_0))| - Bt_1 \geq \frac{x_0}{2} - \theta(\alpha(x_0^2)) - 10Bx_0^2 \geq \frac{x_0}{3} - Bx_0^2,$$

so that

$$\phi_0 \geq \frac{x_0}{4}, \tag{3.37}$$

for $x_0 \leq \frac{1}{B}$. By (3.25) and (3.24)

$$|\tan^{-1}(g'(\beta))| \leq |g'(\beta)| \leq \frac{2\beta}{\rho} \leq \frac{4}{\rho}x_0^2 \quad \text{for} \quad x_0 \leq \frac{1}{B}.$$

Taking into account the direction from which R crosses A at $g(\beta)$, one sees that $\phi_0 \le |\tan^{-1}(g'(\beta))|$, which is a contradiction for $x_0 < \min\{\frac{1}{B}, \frac{\rho}{17}\}$. If R emanates to the right of $J(t_0)$, we get the same contradiction more easily because in that case we move back along C towards p and there are smaller bounds throughout. This establishes (3.29) for $x_0 \le \frac{1}{B}$ (with, of course, an appropriately large B).

Thus, for $\sqrt{t} \le \frac{1}{B}$, $\lambda(\theta(J(t))) \le \sqrt{t}$. From this together with (3.27) it follows that

$$|\theta(z)| \le \sqrt{t} + Bt \le \frac{1}{100} \quad \text{for} \quad z \in J_t, \quad 0 < t < \frac{1}{B}. \tag{3.38}$$

Finally, after increasing B if necessary, we have $|\tan^{-1}(g'(t))| \le \frac{1}{100}$ for $|t| \le \frac{1}{B}$. Let $l_t = \lambda(J_t)$, and let $q(t) = \zeta_t(l_t) \in A$. We now show that $\{q(t) : 0 \le t \le \frac{1}{B}\}$ is an arc of A, for which we have to only show that $q(t)$ is continuous on $(0, \frac{1}{B})$. Clearly, we have from (3.38)

$$\left| \arg\left\{ \frac{\partial \zeta_t(s)}{\partial s} \right\} + \frac{\pi}{2} \right| \le \frac{1}{100} \quad \text{for} \quad t \le \frac{1}{B} \quad \text{and} \quad 0 \le s \le \lambda(t). \tag{3.39}$$

If $l_{t_0} = 0$, then what we have already shown implies that $l_t \to 0$ as $t \to t_0$, from which it immediately follows that $q(t) \to q(t_0)$ as $t \to t_0$. Thus consider a fixed t_0 for which $l_{t_0} > 0$. Let $\delta > 0$ be small, and let $s \in (l_{t_0} - \delta, l_{t_0})$. Let the elementary i-characteristic through $\zeta_{t_0}(s)$ be parameterized by $z = w_s(\sigma)$ with $w_s(0) = \zeta_{t_0}(s)$. It is clear that there is some $\delta' \in (0, \delta]$ such that for all $\sigma \in [-\delta', \delta']$, $w_s(\sigma) \in J(t(\sigma))$ (where $t(\sigma)$ is some continuous function of σ) and $|\theta(w_s(\sigma)) - \theta(w_s(0))| \le \frac{1}{100}$. Then it is clear from (3.39) that for $|\sigma| \le \delta'$, $J(t(\sigma))$ intersects A at a point within 2δ of $q(t_0)$. The desired continuity follows immediately since there is a positive δ'' such that for $|t - t_0| < \delta''$, $J(t)$ intersects $w_s((-\delta', \delta'))$.

From this and (3.38) it follows that, in fact, for $0 < t < \frac{1}{B}$, $|\theta(z)| \le 2\sqrt{t}$ for z in the domain bounded by J_t and the arcs of C and A with endpoints p and $q(t)$. The desired conclusion now follows immediately from Propositions 3.24, 3.25, and 3.31. □

We end this section with a technical proposition to be used in the proof of the main theorem given in section 4. The numbers $\eta = \eta(K)$ and $\eta' = \eta'(K)$ arising here are those of Proposition 2.8.

PROPOSITION 3.34 (trapped area bound). [illegible] $\eta_1 = \eta_1(K) \le \min\{1, \eta(K)\}$ [illegible] (D, V) [illegible] $U = u_1 u_2 \in \mathcal{I}(D)$ [illegible] $|p - q| \ge \xi$ [illegible] $p, q \in \mathrm{bot}(D)$ [illegible] $[u_1, u_2] \subset [p, q]$ [illegible] $\overline{D}$ [illegible] $u \in \{u_1, u_2\}$ [illegible] $U' \subset \overline{D}$ [illegible] U V [illegible] u u' [illegible] $u' < u_1$ $u' > u_2$ [illegible] $u = u_1$ $u = u_2$ [illegible] u' [illegible] U) [illegible] $|\Delta\theta(U')| \ge \frac{\pi}{4}$ [illegible] D' D [illegible] U V [illegible] $\eta_1 \xi^2$ [illegible]. For definiteness we assume that $u = u_1$, so that $u' < u_1$. Let $V = v_1 v_2$. Let U' be parameterized by $w = w(s)$, $0 \le s \le L$, with $w(0) = u_1$, and for $s \in [0, L)$ denote by $U(s) \preceq V$ the elementary i-characteristic containing $w(s)$, so that in particular $U(0) = U$. One of the endpoints $a(s)$ of $U(s)$ lies below U', and the other endpoint $b(s)$ is in $[v_1, v_2] \backslash (u_1, u_2)$. There are the following two possibilities.

A. The point $b(s) \ge u_2$ (i.e., lies to the right of u_2) for $0 \le s < L$. In this case $U \preceq U(s)$, and consequently $\lambda(U(s)) \ge \xi$, for all $s \in [0, L)$, so that by second

area bound (Proposition 2.8(ii)) the area of $\cup\{U(s) : 0 < s < L\} \subset D'$ is at least $\eta' \frac{\pi}{4}\xi^2$.

B. There is some $s \in (0, L)$ for which $b(s) \leq u_1$. Since $b(0) = u_2$, there is a decreasing sequence $\{s_k\}$ tending to some $\sigma \in [0, L)$ such that $b(\sigma) \geq u_2$, but $b(s_k) \leq u_1$ for $k \geq 1$. But then there is a nonmonotone (with respect to D) extended i-characteristic C joining $a_0 = \lim_{k\to\infty} a(s_k)$ to a point $b_0 = \lim_{k\to\infty} b(s_k) \leq a_0 \leq u_1$ for which $b(\sigma)$ is a proper contact point. Then C has subarcs $U(\sigma) = a(\sigma)b(\sigma)$ and $W = de$, which are elementary i-characteristics and for which $U(\sigma) \preceq W$ and $U(\sigma) \cap W \subset \{b(\sigma)\}$, and where $d \in [v_1, u']$, $e \in [b(\sigma), v_2]$. Since we therefore have $[u_1, u_2] \subset [d, e]$, we conclude that $\lambda(W) \geq |e - d| \geq \xi$. Regard W as oriented from e to d and consider the subarc W' of W of length $\frac{\xi}{2}$ which starts at e, and for each $p \in W'$ consider the j-arc $J(p)$ emanating to the left of W and which joins p to a point $q(p) \in \mathrm{bot}(D)$. Because $J(p)$ cannot intersect C at any interior point of C other than p, we have $d \leq q(p) \leq u_1$. Since the length of any curve joining e to a point to the left of u_1 must be at least ξ, it follows that $\lambda(J(p)) \geq \frac{\xi}{2}$. The first area bound (Proposition 2.8(i)) then implies that

$$\mu\Big(\bigcup\{J(p) : p \in W'\}\Big) \geq \eta\frac{\xi^2}{4}.$$

The conclusions of Cases A and B together imply that the area between U and V is at least $\min\{\eta\frac{\xi^2}{4}, \eta'\frac{\pi}{4}\xi^2\} = \eta_1\xi^2$, where $\eta_1 = \min\{\frac{\eta}{4}, \frac{\pi}{4}\eta'\}$. □

4. Proof of Main Theorem 2.2. The main ingredients in the proof are Proposition 3.33 about essentially singularity-free boundary arcs, Proposition 3.29, Proposition 3.17 regarding the structure of $\min_D(A \cup B)$, and the trapped area bound (Proposition 3.34). We also use the preservation of quasi-HP conditions under linear changes of variable, which allows one to normalize the arc one is working with to have any convenient length. In reference to a quasi-HP function θ on $G \in \mathcal{G}$, we will say that $E \subset \partial G$ is [illegible] (abbreviated ESF-i) if it has at most countably many i-singularities of θ. As is explained in detail in the final paragraph of the proof, it is enough to prove that there is a number $\tau = \tau(K) < 1$ such that on any suitably small boundary arc the set of all i-singularities has Hausdorff dimension at most τ. In the treatment of such arcs there is considerable freedom in the choice of the various explicitly given numerical constants that we use, and for the most part they are far bigger (or smaller) than necessary and have most often been chosen either for the sake of convenience or to avoid the necessity of going into careful geometric arguments. In the same vein, many of the bounds we give are far from being sharp, so that, if we state that the value of some expression is bounded below by 1, it may be clear that it is in fact bounded below by a considerably larger number. Furthermore, in some instances we use symbols to denote numerical constants which could without much effort be determined explicitly.

Let $G \in \mathcal{G}(\rho)$, so that the unsigned curvature of ∂G is everywhere at most $\frac{1}{\rho}$. For $p \in \partial G$ let $w(p) = e^{i\phi(p)}$ be the unit tangent to ∂G (with positive orientation) at p. Let δ be any positive for which

(1) $\delta \leq \frac{\rho}{10000}$.

This implies in addition that

(2) on any arc B of ∂G of length 100δ, $\lambda(w(B)) < \frac{1}{100}$, and that

(3) for any point $p \in \partial G$ and any $r \leq 20\delta$, $\partial N(p, r)$ meets ∂G in two points, so that for such r, $N(p, r) \cap G$ is essentially a semidisk.

The corners of $\partial N(p,9\delta)\cap G$ can be replaced by circular arcs of radius $\frac{\delta}{90}$ and then smoothed at the joins to form a C^2 curve which is the boundary of a subdomain

$$\Sigma = \Sigma(p,\delta) \subset N(p,10\delta)\cap G$$

of G with the following properties:

(1a) $\Sigma \in \mathcal{G}(\frac{\delta}{100})$.

(2a) The arc $N(p,4\delta)\cap\partial\Sigma$, to be referred to as the bottom of Σ, and which is an arc of ∂G, has unsigned curvature bounded above by $\frac{1}{10000\delta}$.

(3a) $\mu(\Sigma) \le 200\delta^2$.

Now consider the $X = X(p) = \frac{1}{\delta}(\Sigma(p,\delta) - p)$ which has the following properties:

(1b) $X \in \mathcal{G}(\frac{1}{100})$.

(2b) $B = N(p,4)\cap\partial X$, to be referred to as the bottom of X, is a C^2 arc on which the unsigned curvature is bounded above by $\frac{1}{10000}$.

(3b) $\mu(X) \le 200$.

Note that what we call the bottom B of X is very close to a straight line segment of length 8 centered at the "midpoint" p of what is close to being the straight portion of the boundary of a semidisk of radius 9. Note also that by Proposition 2.3, $\theta(p+\delta z)$ is a K-quasi-HP function on this $X(p)$. A domain X having these three properties will be called a [illegible].

Until the end of the proof of the main theorem we work exclusively with quasi-HP functions θ defined on a normalized domain X. It is clear from property (1b) that there is a universal constant $\gamma_0 \in (0,1)$ such that for any such X

$$|z_1 - z_2| \ge \gamma_0 \mathrm{dist}_{\mathrm{arc}}(z_1,z_2) \quad \text{for} \quad z_1, z_2 \in \partial X, \tag{4.1}$$

where $\mathrm{dist}_{\mathrm{arc}}(z_1,z_2)$ denotes the length of the shorter arc of ∂X with endpoints z_1, z_2. We shall prove the following proposition from which, as we will subsequently show, the main theorem follows almost immediately. We stress that for an arc F of the bottom of X, $\mathrm{diam}(F)$ is the distance between its endpoints.

PROPOSITION 4.1. [illegible] $\Delta_0 = \Delta_0(K) > 0$ [illegible] θ [illegible] K [illegible] X [illegible] X [illegible] i [illegible] Δ_0 [illegible]. It is clear from Proposition 2.3 that we can assume that $p = 0$ and that $w(p) = 1$. We establish the desired conclusion as follows (once again, the constants chosen are unnecessarily big or small). Let H_0 be an integer for which

$$H_0 \ge \max\left\{\frac{10^6}{\gamma_0^2\eta_1}, 10000\right\},$$

and let

$$\epsilon_0 = \overline{\eta}\left(K, \frac{\gamma_0}{10H_0}\right),$$

where $\eta_1 = \eta_1(K)$ is the constant of the trapped area bound (Proposition 3.34) and $\overline{\eta}$ is the constant of Proposition 3.33 about essentially singularity-free boundary arcs; H_0, η_1, and ϵ_0 depend solely on K. In particular we have

$$\epsilon_0 \le \frac{\gamma_0}{10H_0} < \frac{1}{10H_0} \le \frac{1}{10^5}.$$

Our strategy will be to show that there is some constant $\gamma_1 = \gamma_1(K)$ such that the assumption that

(4.2) the bottom B of X has no ESF-i arc of diameter ϵ_0

leads to the conclusion that the bottom of X has an ESF-i arc of diameter $\frac{\epsilon_0}{10^6}$ or one of diameter γ_1, so that B has an ESF-i arc of diameter $\Delta_0 = \min\{\frac{\epsilon_0}{10^6}, \gamma_1\}$. This underlying assumption (4.2) will be in force throughout and will be used many times.

The proof the existence of such a Δ_0 is divided into three steps.

1. There is an elementary i-characteristic $C_0 = \overline{a}\overline{b}$ which has the following properties:

(i) $|\overline{b} - \overline{a}| \geq 1$.

(ii) At least one of $\overline{a}$, $\overline{b}$ lies in B.

COMMENT 4.2.

We begin the proof of Step 1 by noting that by (4.2) there is an i-singularity on B within ϵ_0 of the left endpoint of B, so that there is an elementary i-characteristic C_0' which joins two points of B, both of which are within $2\epsilon_0$ of the left endpoint of B. We work with the right half B_R of B; the endpoints of B_R are 0 and a point d to the right of 0, with $|d| = 4$. In light of our assumption (4.2), it follows that there are two elementary i-arcs $F_1 = a_1a_2$ and $F_2 = b_1b_2$ joining points of B_R such that

$$|a_1 - a_2|, \quad |b_1 - b_2| < \frac{1}{H_0}$$

and

$$\frac{1}{H_0} < \text{dist}(F_1, 0), \quad \text{dist}(F_2, d) < \frac{2}{H_0}.$$

Consider the characteristic subdomain (D, C_0') for which $\text{bot}(D)$ contains the $\partial X \backslash B$, so that the only part of X not in D is the union of C_0' and the interior of the (small) subdomain bounded by C_0' and the arc of B which joins its endpoints. $\text{bot}(D)$. We apply Proposition 3.17 to $\min_D(F_1 \cup F_2)$ with $A = F_1$ and $B = F_2$ and examine in turn each of the three cases of its conclusion.

(i). If there were a contact point p of $\min_D(F_1 \cup F_2)$ between F_1 and F_2 for which

$$\text{dist}(F_1, p), \quad \text{dist}(F_2, p) > \frac{1}{H_0},$$

then Proposition 3.33 would imply that there is an ESF-i subarc of B_R of diameter ϵ_0, contrary to (4.2). Thus, for every contact point p of $\min_D(F_1 \cup F_2)$ between F_1 and F_2, one of

$$\text{dist}(F_1, p) \leq \frac{1}{H_0} \quad \text{or} \quad \text{dist}(F_2, p) \leq \frac{1}{H_0}$$

must hold. Say for definiteness that the former holds for some contact point p between F_1 and F_2, and let $\overline{a}$ be the rightmost such point. Then it is clear that $\min_D(F_1 \cup F_2)$ has an elementary subcharacteristic C_0 joining $\overline{a} \in B$ to a point $\overline{b} \in \text{bot}(D)$, where $\overline{b}$ is either between F_1 and F_2 but within $\frac{1}{H_0}$ of F_2 or to the right of F_2 (with respect to the order on $\text{bot}(D)$). In either event, $|\overline{b} - \overline{a}| \geq 1$. The case $\text{dist}(F_2, p) \leq \frac{1}{H_0}$ is handled similarly.

For the other two cases we let a, b, e, f, U, and L be as in the conclusion of Proposition 3.17. Furthermore, let the left and right endpoints of C_0' be l and r (with respect to the order on $\text{bot}(D)$, so that since $\text{bot}(D)$ is very close to a segment of $\mathbb{R}$, l lies to the right of r in the usual sense).

(ii). Let α and β be the left and right endpoints of L. If $|\beta - \alpha| \geq 1$, then we can take L for C_0 since α is to the left of F_1 and therefore is in B. Thus we may assume that $|\beta - \alpha| < 1$. Starting at a and traversing ∂X in the negative direction the indicated points occur in the following order: a, β, a_2, a_1, α, e, l, r, f, b (with some equalities possible). If $|e - f| < 1$, then the arc $e\alpha$ joins e, $\alpha \in B$ and $|e - \alpha| \geq 2$. It is then easy to see that either $e\alpha$ has a proper contact point g for which $|g - e|$, $|g - \alpha| > \frac{1}{H_0}$, which (in light of Proposition 3.33) would violate (4.2), or $e\alpha$ has a subarc C_0 which is an elementary i-characteristic with the desired properties. Thus we need only consider what happens if $|e - f| \geq 1$. But in this case $C_0 = ef$ has the desired properties since it is an elementary i-characteristic and $e \in B$.

(iii). Here we assume that we are not simultaneously in Case (ii), since were that so we would be done. The point b is obviously between F_1 and F_2, but, since we are not in Case (ii), a is between l and F_1. Let $L = pq$, with $p < q$. If $|b - p| \geq \frac{1}{H_0}$, then, again by Proposition 3.33, B would have an ESF-i arc of diameter ϵ_0 contradicting (4.2). Thus we can assume that $|b - p| < \frac{1}{H_0}$. If $|p - q| < 1$, then we will have $|b - a| > 1$ and we can take for C_0 any elementary i-characteristic closely approximating the extended characteristic ab (recall the construction of extended characteristics in section 3), since its endpoints can be made arbitrarily close to a and b, and consequently can be chosen to lie in B. On the other hand, if the distance between p and q is at least 1, then we can take C_0 to be the elementary i-characteristic pq, since $p \in B$. Thus we are done with Step 1.

Step 1 implies that there is a subarc P of ∂X and an elementary i-characteristic $C_0 = \overline{a}\overline{b}$ with $\overline{b} - \overline{a} \geq 1$ joining the two endpoints of P such that P contains a subarc I of B, the distance between whose endpoints is at least 1 and one of whose endpoints is an endpoint of C_0. To be specific we assume that the left endpoint of I is the one I has in common with C_0. We shall henceforth work with the characteristic subdomain (D_0, C_0) bounded by $P \cup C_0$, so that in particular all order related statements will refer to the order on $\text{bot}(D_0) = P$, unless otherwise indicated. We give P the usual arc length parameterization $z = z(s)$, $0 \leq s \leq L$, with increasing s corresponding to the positive orientation on ∂X and such that $z(0)$ is the left endpoint of I. Obviously, $z([0,1]) \subset I$. By the bound of $\frac{1}{10000}$ on the unsigned curvature of B it is clear that

$$|z(s_2) - z(s_1)| \geq \frac{9}{10}|s_2 - s_1| \quad \text{for} \quad z(s_1),\, z(s_2) \in I. \tag{4.3}$$

(Obviously, a constant considerably closer to 1 than $\frac{9}{10}$ would also work here.) We also observe that for α, $\beta \in \text{bot}(D_0)$

$$|\beta - \alpha| \geq \gamma_0 \min\{1, \lambda([\alpha', \beta'])\} \quad \text{for all } [\alpha', \beta'] \subset [\alpha, \beta]. \tag{4.4}$$

This follows from (4.1) and the fact that one of the arcs of ∂X joining α to β contains

$[\alpha', \beta']$ and the other contains the endpoints of C_0, the distance between which is at least 1. It is also easy to see that

$$\operatorname{dist}(z(0), P\backslash B) \geq 1. \tag{4.5}$$

2. There is a subarc $J = pq$ of I with $|q - p| \geq \frac{1}{40}$ such that p and q are joined by an elementary i-characteristic in D_0.

To establish this we again use an argument based on the case that results when Proposition 3.17 is applied. However, for this step this proposition must be applied, in conjunction with the trapped area bound (Propositions 3.34), in a sequential manner. By our underlying assumption (4.2) there are elementary i-characteristics $A_1, \ldots, A_{H_0} \in \mathcal{I}(D_0)$ for which the arc $I_k \subset I \subset \operatorname{bot}(D_0)$ joining the endpoints of A_k lies in the middle third of $z([\frac{k-1}{H_0}, \frac{k}{H_0}])$ (i.e., lies in $z([\frac{k-\frac{2}{3}}{H_0}, \frac{k-\frac{1}{3}}{H_0}])$), $1 \leq k \leq H_0$. Note that by (4.3)

$$\operatorname{dist}(I_{k_1}, I_{k_2}) \geq \frac{9}{10H_0}\left(|k_2 - k_1| - \frac{1}{3}\right). \tag{4.6}$$

Assume inductively that we have elementary i-characteristics C_k, $0 \leq k \leq t$, such that C_k envelopes only $A_{l(k)}, A_{l(k)+1}, \ldots, A_{r(k)}$, but none of the other A_n, such that

$$C_0 \succeq C_1 \succeq \cdots \succeq C_t,$$

$$1 = l(0) \leq l(1) \leq \cdots \leq l(t) < r(t) \leq r(t-1) \leq \cdots \leq r(0) = H_0$$

and such that $r(k) - l(k)$ is strictly decreasing and satisfies

$$r(k) - l(k) \geq \frac{H_0}{3}, \quad 0 \leq k \leq t. \tag{4.7}$$

We apply Proposition 3.17 with $A = A_{l(t)}$ and $B = A_{r(t)}$ to analyze $C = \min_{D_0}(A_{l(t)} \cup A_{r(t)})$ and consider separately each of the three cases in the conclusion of that proposition. More specifically, we will show three things: first, that if Case (iii) does not occur, either we will have obtained the desired J (and will therefore stop) or we will be able to produce C_{t+1} with

$$r(t+1) - l(t+1) = r(t) - l(t) - 1; \tag{4.8}$$

second, that when Case (iii) occurs we have the desired J; and third, that Case (iii) will have to occur long before (4.7) can be violated.

(i). There can be no contact points of C between $A_{l(t)+1}$ and $A_{r(t)-1}$, since, were there to be such a point p, the distance from p to each of the endpoints of C would, by (4.6), have to be at least $\frac{9}{10H_0}(\frac{2}{3}) > \frac{1}{10H_0}$, so that we can apply Proposition 3.33 to conclude that there is an ESF-i arc of diameter ϵ_0 on B, which contradicts (4.2). If there is a contact point of C between $A_{r(t)-1}$ and $A_{r(t)}$, we will have obtained the desired conclusion because in that case C would have to have a subarc E which is an elementary i-characteristic with left endpoint to the left of $A_{l(t)+1}$ and right endpoint between $A_{r(t)-1}$ and $A_{r(t)}$, so we stop the process here, Step 2 having been established with $J = E$ in light of (4.7). If there is no contact point of C between $A_{r(t)-1}$ and $A_{r(t)}$, then there is a contact point of C between $A_{l(t)}$ and $A_{l(t)+1}$ and C has a subarc C_{t+1} which is an elementary characteristic whose left endpoint is between $A_{l(t)}$ and $A_{l(t)+1}$ and whose right endpoint lies to the right of $A_{r(t)}$. We have $l(t+1) = l(t)+1$ and $r(t+1) = r(t)$, so that (4.8) holds.

For the other two cases we use the notation of Proposition 3.17, namely L, U, ef, and ab, where $ef = U = C = \min_{D_0}(A_{l(t)} \cup A_{r(t)})$.

(ii) (iii). Because we are not in Case (iii), b lies to the right of $A_{r(t)}$. We have that a is between $A_{l(t)}$ and $A_{l(t)+1}$, since otherwise $|e-a| \geq \frac{9}{10H_0}(\frac{2}{3}) > \frac{1}{10H_0}$, by (4.6), and then, since $[e,a] \subset [e,b]$,

$$|e-b| \geq \gamma_0 \min\{1, \lambda([e,a])\} \geq \gamma_0 \min\{1, |e-a|\} > \frac{\gamma_0}{10H_0}$$

by (4.4), so that by Proposition 3.33 applied to the contact point e there is an ESF-i arc of diameter ϵ_0 on B, which contradicts (4.2). From the definition of extended characteristics it follows that there are elementary i-characteristics $U' = a'b' \preceq U$ with a' and b' arbitrarily close to a and b, respectively. The point a' can therefore be taken to lie between $A_{l(t)}$ and $A_{l(t)+1}$, and b' can be taken to lie to the right of $A_{r(t)}$. In this case we let C_{t+1} be any such U' and set $l(t+1) = l(t)+1$ and $r(t+1) = r(t)$, so that we again have (4.8).

(iii). We will show that the first time this case occurs we will have obtained the desired conclusion of Step 2. Thus we assume that when we apply Proposition 3.17 to produce $C = \min_{D_0}(A_{l(t)} \cup A_{r(t)})$ we are in Case (iii). Now we can only say that a lies somewhere to the right of the left endpoint of C_t and that f lies somewhere to the right of $A_{r(t)}$, but, since we cannot use Proposition 3.33 to any effect (because f is not necessarily in B), we cannot conclude, in analogy with the preceding case, that b lies between $A_{r(t)-1}$ and $A_{r(t)}$. If $f \in B$, then by (4.3) and (4.7) it follows that $|f-e| \geq \frac{1}{6}$. On the other hand, if $f \notin B$, then since $f \in P$, (4.5) implies that $|f - z(0)| \geq 1$. Since by (4.7), $|e - z(0)| < \frac{7}{10}$, we have that $|f-e| \geq \frac{3}{10}$. Therefore, no matter what, $|f-e| \geq \frac{1}{6}$. Let D' be the subdomain of D_0 bounded by $ab \cup [a,b]$. It follows immediately from (4.3) that $\lambda([a,b]) \leq \frac{10}{9}|b-a|$. From the definition of extended characteristics it follows that for any $\epsilon > 0$ there are elementary i-characteristics E inside D', joining points of a', $b' \in [a,b]$ and containing points e' and f' which are within ϵ of a, b, e, f, respectively. But it follows from Proposition 3.6 that

$$\frac{1}{6} - 2\epsilon \leq |f'-e'| \leq \operatorname{diam}(e'f') \leq \operatorname{diam}(a'b') \leq 5\lambda([a',b']) \leq \frac{50}{9}|b'-a'|,$$

so that with ϵ appropriately small we see that $|b'-a'| > \frac{1}{40}$. But taking ϵ sufficiently small, the corresponding E will serve as the desired J.

Thus either we stop at t with the desired J or go on to $t+1$, the latter occurring only if we are in Case (i) or if we are in Case (ii) but not Case (iii). We now show that we must actually arrive at Case (iii) long before (4.7) can be violated. Say that we have arrived at $t = T < \frac{H_0}{3}$. Let $T' = [\frac{T}{2}]$. At least one of the following two things must have occurred:

(A) At least T' of the times that we passed from t to $t+1$ we will have done so because we are in Case (i) and there is a proper contact point p_t of $\min_{D_0}(A_{l(t)} \cup A_{r(t)})$ between $A_{l(t)}$ and $A_{l(t)+1}$ and p_t is the left endpoint of C_{t+1}.

(B) At least T' of the times that we passed from t to $t+1$ we will have done so because we are in Case (ii) but not Case (iii).

We deal with possibility (A) first. Say $k_1 < k_2$ are two values of t for which we are in Case (i). Let C' be the elementary j-characteristic, one of whose endpoints is p_{k_2}. Let its second endpoint be p'. If C' were to cross C_{k_1+1}, then, since a j-characteristic can have at most one point in common with an i-characteristic, the other endpoint of

C' would lie to the left of $A_{l(k_1)}$ or to the right of $A_{r(k_1)}$, and the distance between its endpoints would, by (4.6), have to be at least $\frac{9}{10H_0}(\frac{2}{3}) > \frac{1}{2H_0}$. However (as we have seen happen before), by (4.7) together with Proposition 3.33, B would then have an ESF-i arc of diameter ϵ_0, contradicting (4.2). Thus C' does not cross C_{k_1+1}, so that $p' \in \text{bot}(D_0)$. If α, β are points of $\text{bot}(D_0)$ with α lying to the left of C_{k_2+1} and β to its right, then $\lambda(\alpha\beta) \geq \frac{1}{4}$ since $T < \frac{H_0}{3}$, so that by (4.4)

$$|\beta - \alpha| \geq \frac{\gamma_0}{4}. \tag{4.9}$$

If p' lies to the right of C_{k_2+1}, then each point of $C_{k_2+1} \cap X$ is joined to a point in $\text{bot}(D_0)$ to the right of C_{k_2+1} by a j-half-characteristic lying between C_{k_2+1} and $C' \preceq C_{k_1+1}$. Then a simple argument (which was used in the proof of Proposition 3.34) based on Proposition 2.8(i) and the fact that for each point of the first half of C_{k_2+1}, which has length at least $\frac{\gamma_0}{8}$, the corresponding j-half-characteristic has length at least $\frac{\gamma_0}{8}$ shows that the area between C_{k_2+1} and C_{k_1+1} must be at least $\eta(\frac{\gamma_0}{8})^2 = \frac{\eta\gamma_0^2}{64} \geq \frac{\eta_1\gamma_0^2}{64}$. If, on the other hand, $p' < p_{k_2}$, then $|\Delta\theta(C')| \geq \frac{\pi}{4}$. Also, as we just saw, (4.9) holds for all α, $\beta \in \text{bot}(D_0)$ with α lying to the left of C_{k_2+1} and β to its right. Thus by the trapped area bound (Proposition 3.34) it follows that the area between C_{k_2+1} and C_{k_1+1} is at least $\frac{\eta_1\gamma_0^2}{16} > \frac{\eta_1\gamma_0^2}{64}$, so that this lower bound holds no matter which side of C_{k_2+1} the point p' lies on. But then $\frac{(T'-1)\eta_1\gamma_0^2}{64} \leq \mu(X) \leq 200$, so that $T' \leq \frac{12800}{\gamma_0^2\eta_1} + 1$. This in turn means that $T \leq \frac{30000}{\gamma_0^2\eta_1} < \frac{H_0}{10}$ (since γ_0 and η_1 are both in $(0,1)$).

Possibility (B) is handled in a similar manner. Say $k_1 < k_2$ are two values of t for which we are in Case (ii) but not in Case (iii) when passing from C_t to C_{t+1}. From the construction of C_{k_1+1} we have that $U = \min_{D_0}(A_{l(k_1)} \cup A_{r(k_1)}) = u_1u_2$ is a subarc of a nonmonotone extended i-characteristic W which has another subarc S which is also an elementary i-characteristic for which $S \preceq U$ and $A_{l(k_1)} \preceq S$. It follows in addition from that construction that S precedes $C_{k_1+1} \preceq U$, so that in particular S precedes A_k for all $k \geq l(k_1)+1$. There is an elementary j-characteristic S' which joins u_1 to a point $u' \in [u_1, u_2]$. It follows immediately from (4.7) and (4.4) that $|u_2 - u_1| \geq \frac{\gamma_0}{4}$. We claim that $S' \cap D_0 \cap C_{k_2+1} = \emptyset$. To see this, assume to the contrary that $z \in S' \cap D_0 \cap C_{k_2+1}$. It then follows from (4.4) and (4.6) that $|u' - u_1| \geq \frac{9\gamma_0}{10H_0}(\frac{2}{3}) > \frac{\gamma_0}{2H_0}$, so that by Proposition 3.33, B would have an ESF-i arc of diameter a least ϵ_0 in contradiction to our underlying assumption (4.2). Now we apply the trapped area bound to the i-characteristic subdomain (D_1, C_{k_2+1}), where D_1 is bounded by the curve made up of C_{k_2+1} together with $\partial X \backslash (\alpha_1, \alpha_2)$, where $\alpha_1 < \alpha_2$ are the endpoints of C_{k_2+1} (that is, D_1 is the part of X that remains when the part of D_0 on and below C_{k_2+1} is removed). In particular $S' \backslash \{u_1, u'\} \subset D_1$ and $U \in \mathcal{I}(D_1)$. It is also clear from (4.7) and (4.4) that (4.9) holds for α and β in $\text{bot}(D_1)$ on opposite sides of C_{k_1+1}. For the same reason that $S' \cap D_0 \cap C_{k_2+1} = \emptyset$, we have that u' must lie between u and α_1 with respect to the order on $\text{bot}(D_0)$. This means that both endpoints of S' are in B, so that we clearly have $\Delta\theta(S') \geq \frac{\pi}{4}$. It also means that with respect to the order on $\text{bot}(D_1)$, u_1 lies between u' and u_2. We can now apply the trapped area proposition (with (D_1, C_{k_2+1}) playing the role of (D, V)) to deduce that the area between U and C_{k_2+1} is bounded below by $\frac{\eta_1\gamma_0^2}{16}$, so that this same lower bound holds for the area between C_{k_1} and C_{k_2+1}. If the values of t in question are $t_1 < t_2 < \cdots < t_{T'}$, and if R_k is the region between C_{t_k} and $C_{t_{k+1}+1}$, then $R_1, R_3, \ldots$

are disjoint, so that

$$\left(\frac{T'-2}{2}\right)\frac{\eta_1\gamma_0^2}{16} \le \mu(X) \le 200,$$

and therefore $T' \le \frac{6400}{\gamma_0^2\eta_1} + 2$. Thus, as in the case that (A) holds, we again have $T < \frac{H_0}{10}$. This concludes the proof of Step 2.

Step 3. Now we work with the $J = pq \subset I$, with $|q-p| \ge \frac{1}{40}$, whose endpoints are joined by a $C = pq \in \mathcal{I}(D_0)$. Clearly, there are two points $a_1, a_2 \in J$ such that if we consider $X_k = \Sigma(a_k, \frac{1}{10^6})$, $k = 1, 2$, where the Σ (as defined in the third paragraph of this section) are with respect to the normalized domain X with which we are now working, then

$10^6(X_k - a_k)$ is a normalized domain, $k = 1, 2$,
$\operatorname{dist}(X_k, \{p, q\}\} \ge \frac{1}{500}$, $k = 1, 2$,
$\operatorname{dist}(X_1, X_2) \ge \frac{1}{500}$, and
the bottoms of X_1 and X_2 are in J.

By what we have shown it then follows that either the bottom one of X_1 or X_2 has an ESF-i of length $\frac{\epsilon_0}{10^6}$, in which case we have reached the desired conclusion, or for both $k = 1$ and $k = 2$ the bottom of X_k has a subarc p_kq_k, with

$$|q_k - p_k| \ge \frac{1}{4 \cdot 10^7} \tag{4.10}$$

whose endpoints are joined by an elementary i-characteristic C_k. In this latter case, since C joins the endpoints of J, $C_k \preceq C$, $k = 1, 2$, we can apply Proposition 3.17 to $\max_{D'}(C_1 \cup C_2)$, where D' is the characteristic subdomain bounded by $C \cup [p, q]$. But in light of the above conditions satisfied by the X_k and (4.10), and since all contact points of $\max_{D'}(C_1 \cup C_2)$ are in J, Proposition 3.33 immediately implies that there is a $\gamma_1 = \gamma_1(K)$ for which J has an ESF-i arc of J of diameter γ_1. Thus we have proved that the bottom of a normalized domain has an ESF-i arc of length $\Delta_0 = \min\{\frac{\epsilon_0}{10^6}, \gamma_1\}$. □

[illegible]. It follows from the opening discussion and the proposition we have just proved that there are $\delta_0 = \delta_0(G)$, $\alpha_0 = \alpha_0(K) \in (0, 1)$ such that any arc B of ∂G of diameter at most δ_0 has an ESF-i subarc B', for which $\operatorname{diam}(B') > \alpha_0 \operatorname{diam}(B)$. It is then clear that there are numbers $\xi_0 = \xi_0(K)$, $\tau_0 = \tau_0(K)$, both in $(0, 1)$, such that every arc B with $\operatorname{diam}(B) \le \delta_0$ has two disjoint subarcs B_1 and B_2 for which

$$(\operatorname{diam}(B_1))^{\tau_0} + (\operatorname{diam}(B_2))^{\tau_0} < \xi_0(\operatorname{diam}(B))^{\tau_0},$$

$$\operatorname{diam}(B_1), \operatorname{diam}(B_2) < (1 - \alpha_0)\operatorname{diam}(B),$$

$$B \backslash (B_1 \cup B_2) \text{ is ESF-}i.$$

Let $\operatorname{diam}(B) \le \delta_0$. We start with two such arcs B_1 and B_2 corresponding to B, then we apply the same fact to each of these to get four arcs of diameter $l_1, \ldots, l_4 < (1-\alpha_0)^2$ for which $\sum l_k^{\tau_0} < \xi_0^2\delta_0^{\tau_0}$, and such that the complement of their union is ESF-i, then we do so again to get eight arcs of length at most $(1-\alpha_0)^3$ for which the corresponding $\sum l_k^{\tau_0}$ is less than $\xi_0^3\delta_0^{\tau_0}$ and for which the complement of their union is ESF-i, and so on. At the nth stage we have 2^n arcs $B_k^{(n)}$ for which $\sum_k (\operatorname{diam}(B_k^{(n)}))^{\tau_0} < \xi_0^n\delta_0^{\tau_0}$, which tends to 0, and for which $B \backslash (\cup_k B_k^{(n)})$ is ESF-i. At this stage the diameter of

the largest of the 2^n arcs is at most $(1-\alpha_0)^n$. But then B is the union of a set of τ_0-dimensional Hausdorff measure 0 and a set that has at most countably many i-singularities. In light of Proposition 3.29 and the fact that there are at most countably many vertices of fans on ∂G (Proposition 3.24) it follows that the set of singularities on B, and therefore the set of singularities on all of ∂G, has Hausdorff dimension at most τ_0. This finishes the proof of the main theorem.

COROLLARY 4.3. [illegible] $\tau = \tau(\Theta) < 1$ [illegible] $G \subset \mathbb{C}$ [illegible] R, Θ [illegible] G [illegible] ∂G [illegible] R [illegible] ff [illegible] τ [illegible]. This is an immediate consequence of the main theorem and Propositions 3.20 and 3.25. □

5. Construction of solutions with singularity sets of positive Hausdorff dimension. Although we have chosen to carry out our construction in $\mathbb{H}$ to avoid cumbersome arguments, suitable, largely straightforward changes will allow an analogous procedure to be carried out in any smoothly bounded domain. Throughout this section Θ will denote any fixed normal system. We begin with a brief discussion of characteristic initial value problems for Θ since our construction largely proceeds by joining together solutions of appropriate instances of such problems in domains sharing a boundary characteristic.

Let C_k be C^∞ curves parameterized by $z_k(s)$, $0 \le s \le l_k$, $k = 1, 2$, for which $z_i(0) = z_j(0)$ and for which $z_i'(0)$ and $z_j'(0)$ are mutually orthogonal. (In fact l_k can be infinite, as is the case, for example, when C_k is a ray.) We want to construct a solution of the normal system Θ for which C_k is an k-characteristic arc, $k = 1, 2$. From the definition of normal system, for any ρ_1 there is a discrete set of values ρ_2 such that $e^{i\theta(\rho_1,\rho_2)}$ and $ie^{i\theta(\rho_1,\rho_2)}$ are tangent at $z_1(0)$ to C_1 and C_2, respectively. Given any such [illegible] $\rho = (\rho_1, \rho_2)$ there are unique continuous functions $R_k(s)$ such that $e^{i\theta(\rho_1,R_2(s))}$ and $ie^{i\theta(R_1(s),\rho_2)}$ are tangent to C_1 and C_2 at the points $z_1(s)$ and $z_2(s)$, respectively. For $t = (t_1, t_2)$ we define

$$\overline{R}(t) = (R_1(t_2), R_2(t_1)), \quad 0 \le t_k \le l_k, \quad k = 1, 2,$$

and

$$\overline{\theta}(t) = \theta(R_1(t_2), R_2(t_1)), \quad 0 \le t_k \le l_k, \quad k = 1, 2.$$

We seek $\zeta : [0, l_1] \times [0, l_2] \to \mathbb{C}$, where $\zeta(t) = \xi(t) + i\eta(t)$, for which $e^{i\overline{\theta}(t)}$ and $ie^{i\overline{\theta}(t)}$ are tangent to the curves $\zeta([0, l_1], t_2)$ and $\zeta(t_1, [0, l_2])$, respectively, at the point $\zeta(t)$ and which satisfies the initial conditions

$$\zeta(t_1, 0) = z_1(t_1) \quad \text{and} \quad \zeta(t_2, 0) = z_2(t_2). \tag{5.1}$$

The tangency condition can be written as the linear hyperbolic system

$$\cos\overline{\theta}\frac{\partial\eta}{\partial t_1} - \sin\overline{\theta}\frac{\partial\xi}{\partial t_1} = 0, \quad \sin\overline{\theta}\frac{\partial\eta}{\partial t_2} + \cos\overline{\theta}\frac{\partial\xi}{\partial t_2} = 0. \tag{5.2}$$

The problem (5.2) with initial conditions (5.1) is well posed and has a C^∞ solution $\overline{\theta}$ for any C^∞ initial curves C_1 and C_2. First consider the case in which the $\zeta(t) = \zeta(C_1, C_2, \rho, t)$ determined in this manner has an everywhere nonzero Jacobian

determinant and is globally one-to-one. We then have a solution to the system Θ in $\zeta([0,l_1]\times[0,l_2])$, which is a characteristic quadrilateral, given by

$$R(\zeta(t)) = \overline{R}(t). \tag{5.3}$$

We refer to this solution as $R(C_1, C_2, \rho, z)$ and to the characteristic quadrilateral in which it is defined as $Q(C_1, C_2, \rho)$. Even if ζ is not one-to-one on $[0,l_1]\times[0,l_2]$, then (5.3) gives a multivalued solution of Θ on any domain $\zeta(E)$, provided that ζ is a local diffeomorphism on E. Let i and j be such that $\arg\{z_j'(0)\} = \arg\{z_i'(0)\} + \frac{\pi}{2}$. It is a well-known property of the characteristic initial value problem for genuinely nonlinear systems that if $\frac{d\arg\{z_i'(t)\}}{dt} \le 0$ on $[0,l_i]$ and $\frac{d\arg\{z_j'(s)\}}{ds} \ge 0$ on $[0,l_j]$, then ζ will be locally diffeomorphic on $[0,l_1]\times[0,l_2]$. This is simply a reflection of the fact that, in light of the quasi-HP nature of the the associated inclination function and the length monotonicity property (Proposition 2.7), under these hypotheses the curvature of j-characteristics will not blow up as one moves along i-characteristics away from the convex side of a j-characteristic C, and analogously when the roles of i and j are interchanged. We note that this property holds in the particular case in which one of the initial characteristics is a line segment or ray. We also note that it holds even if one or both of the initial curves is not a simple arc, so that ζ will be a local homeomorphism if, for example, one of these curves is a circle covered several times.

We next need to discuss briefly the smooth adjunction of line segments and circles to C^∞ curves in a specific, constructive fashion; this is another essential element of the construction process. Let C be a C^∞ curve parameterized by $Z(s)$, $0 \le s \le l+\delta$. We want to perform the adjunction with no change in Z on $[0,l]$. Let κ_0 be any number. It is clear that there is an operator $F(Z,l,\delta,\kappa_0,\tau)$ such that $w = F(Z,l,\delta,\kappa_0,\tau) \in C^\infty([0,\infty))$ has the following properties:

(i) $w(s) = Z(s)$, $0 \le s \le l$.

(ii) If $w'(s) = e^{i\phi(s)}$, then $\phi(s) = (s-l-\delta)\kappa_0 + \tau$, $s \ge l+\delta$.

It is clear how this can be done. Indeed, if $Z'(s) = e^{i\alpha(s)}$ on $[0,l+\delta]$, we let $\beta(s)$ be the continuous function which coincides with $\alpha(s)$ on $[0,l+\delta/3]$, which is given by $(s-l-\delta)\kappa_0 + \tau$ for $s \ge l + 2\delta/3$ and is linear on $[l+\delta/3, l+2\delta/3]$. We let σ be a specific nonnegative C^∞ function on $\mathbb{R}$ with support in $(-1,1)$ and $\int_{-\infty}^{\infty}\sigma(s)ds = 1$ and then convolve $\beta(s)$ with $\frac{10}{\delta}\sigma(\frac{10s}{\delta})$. The desired $w = F(Z,l,\delta,\kappa_0,\tau)$ is defined by with $w'(s) = e^{i\beta(s)}$, $w(0) = Z(0)$. We note that if C is convex to the right (left) and κ_0, $\tau - \kappa_0\delta/3 - \alpha(l+\delta/3) \ge 0$ (≤ 0), then the curve given by $F(Z,l,\delta,\kappa_0,\tau)$ is concave towards the same side as C.

We now introduce notation and terminology to be used in our construction. We denote by $\mathcal{CL}$ the family of C^∞ arcs with initial and terminal straight subarcs. Let $\mathcal{S}(\epsilon)$ denote the class of C^∞ curves parameterized by $z = z(s)$, $0 \le s \le L$, with the following properties:

(i) $\Im\{z(0)\} = \Im\{z(L)\} = -\epsilon$.

(ii) $\arg\{z'(0)\} = \frac{\pi}{2} + \epsilon$ and $\arg\{z'(L)\} = -(\frac{\pi}{2} + \epsilon)$.

(iii) $\frac{d\arg\{z'(s)\}}{ds} \le 0$, $0 \le s \le L$.

(iv) $C \subset N(\partial N(0,1)\cap\mathbb{H}, 2\epsilon)$.

(v) $C \in \mathcal{CL}$.

For $t > 0$ and $\alpha \in \mathbb{R}$ we denote $\{tS + \alpha : S \in \mathcal{S}(\epsilon)\}$ by $\mathcal{S}(\epsilon, t, \alpha)$. If C is any arc joining $a < b$ in $\overline{\mathbb{H}}$, $B(C)$ will denote the set whose boundary is $C \cup [a,b]$. For two such arcs C_1, C_2 with $C_1 \subset B(C_2)$ we denote the closure of $B(C_2)\backslash B(C_1)$ by

$E(C_1, C_2)$. In what follows we will often deal with a C^∞ solution R defined only in a neighborhood in $B(C)$ of C, where C is an i-arc of R (by which we mean that R can be extended to a C^∞ solution in a two-sided neighborhood of C and that C is an i-arc of this extension). In this case we shall call C a [illegible] of R. An i-characteristic C of a solution R will be said to satisfy the [illegible] (OSC) if the j-characteristics passing through each point $c \in C$ all contain a straight line subarc containing c in its interior; obviously, it is enough for a single point $c \in C$ to have this property for the OSC to hold. We extend the use of the term OSC in the obvious manner to apply to one-sided characteristics. If D is a domain separated into two subdomains D_1 and D_2 by an arc C, and $R^{(1)}$ and $R^{(2)}$ are C^∞ solutions in these domains, respectively, for which C is a one-sided i-characteristic satisfying the OSC, then together $R^{(1)}$ and $R^{(2)}$ give a C^∞ solution in all of D. Our construction uses this simple fact, which allows for the smooth pasting together of solutions of characteristic initial value problems in contiguous domains.

Next we discuss a specific class of characteristic initial value problems in which one of the initial characteristics is a ray. Let C be parameterized by $z(s)$ satisfying

$$D \arg\{z'(s)\} \leq 0, \quad 0 \leq s \leq L, \quad \text{with } \Im\{z(0)\}, \quad \Im\{z(L)\} < 0, \tag{5.4}$$

and

$$\arg\{z'(0)\} \geq \frac{\pi}{2} + \epsilon \quad \text{and} \quad \arg\{z'(L)\} \leq -\left(\frac{\pi}{2} + \epsilon\right), \tag{5.5}$$

and consider the solution of the initial characteristic value problem where $C_i = C$ and C_j is a ray orthogonal to C at $z(0)$ and emanating to the left of C. Then the j-characteristics of any corresponding solution are all rays orthogonal to C_i and the i-characteristics are the orthogonal trajectories of this family of rays, and in fact are the curves $C(r)$, $r > 0$, parameterized by $z(t)+rz'(t)i$, $0 \leq t \leq L$. As r tends to ∞ the $C(r)$ tend to arcs of a circle of radius r and radian measure $\arg\{z'(0)\} - \arg\{z'(L)\}$.

We next construct two special C^∞ arcs U_i, where it is understood that the ϵ referred to below is some sufficiently small positive number. The arc U_1, parameterized by $u_1(s)$, $0 \leq s \leq \lambda_1$ will have the following properties:

(i) $D \arg\{u_1'(s)\} \leq 0$, $0 \leq s \leq \lambda_1$.
(ii) $\Im\{u_1(0)\} = -\epsilon$ and $u_1(\lambda_1) \in \mathbb{R}$.
(iii) $U_1 \in \mathcal{CL}$.
(iv) $u_1'(0) = e^{i(\frac{\pi}{2}+\epsilon)}$.
(v) $u_1'(s) = e^{-i\pi/4}$, $\lambda_1 - d \leq s \leq \lambda_1$ for some $d > 0$.
(vi) $U_1 \cap N(0, 2) = \emptyset$.

(vii) Let any C' be any curve in $\mathcal{S}(\epsilon)$, and let $\alpha = \alpha(C)$ be such that the initial point of $C = C' + \alpha \in \mathcal{S}(\epsilon, 1, \alpha)$ is $p = -1 - \epsilon i$. Let $\rho \in \mathbb{R}^2$ be such that $ie^{i\theta(\rho)}$ is tangent to C at its left endpoint p. Then there is a solution to the system Θ in $E(C) = E(C \cap \mathbb{H}, U_1 \cap \mathbb{H})$ for which $R(p) = \rho$, for which C is a 2-characteristic and U_1 is an 1-characteristic, and for which both C and U_1 satisfy the OSC.

We stress that the idea is that for the appropriate translate C of [illegible] $C' \in \mathcal{S}(\epsilon)$ and any admissible value ρ of R at the initial point (i.e., left endpoint) c_l of C we have a solution in the domain $E(C \cap \mathbb{H}, U_1 \cap \mathbb{H})$. If we can construct such a curve, it is obvious that it will have a right-hand counterpart U_2, having properties corresponding to (i)–(vii). In regard to ( ), U_2 will terminate in a line segment of slope $\frac{\pi}{4}$ at its left end. In regard to (vii), U_2 and C will be 2- and 1-characteristics, respectively, and the initial value ρ will be the value of R at the [illegible] endpoint of C, which in any

case determines its value at the right endpoint of C since the total curvature of C is exactly $\pi + 2\epsilon$. For lack of a better term we shall refer to U_1 and U_2 as [illegible] for the system Θ. It is clear that translates $U_i + \alpha$, $\alpha \in \mathbb{R}$ of U_i have an analogous universal property.

We show that such a U_1 exists, the existence of U_2 being identical apart from trivial details. Let $A \leq |\frac{\partial\theta}{\partial R_k}| \leq B$ on $\mathbb{R}^2$ (as in the definition of normal system), so that the net of characteristics of any solution of the system Θ in a domain G is a K-quasi-HP net on G with $K = \frac{B}{A}$. Let W be any C^∞ arc parameterized by $w(s)$, $0 \leq s \leq L$, for which $D\arg\{w'(s)\} \leq 0$ and which has total curvature $-\Delta < 0$. Let $V \in \mathcal{CL}$ be parameterized by $v(s)$, $0 \leq s \leq 1$, with $v(0) = w(0)$, $v'(0) = iw'(0)$, for which $0 \geq D\arg\{v'(s)\} \geq -\delta$. One easily show that for any admissible corner value ρ the characteristic coordinate mapping $\zeta(V, W, \rho, t)$ is a local diffeomorphism on all of the corresponding characteristic coordinate rectangles $S = [0,1] \times [0, L]$ for all such W, provided that $\delta K(L + K\Delta) < 1$. To do so, one simply subdivides S into small subrectangles and solves the corresponding initial value problem piece by piece, the key point being that the length change estimate (Proposition 2.5) implies that no blow-up will occur (that is, the Jacobian determinant will never vanish). Indeed, in light of the convexity of W the lengths of its translates increase, while in light of the concavity of V the lengths of its translates decrease. The characteristic length bound as applied to the translates of W shows that they all have length at most $L + K\Delta$, so that by the characteristic length bound the lengths of the translates of V will remain bounded away from 0 if $\delta K(L + K\Delta) < 1$. We start with $W = W_0$, $V = V_0$ and inductively define ρ_{k+1} to be the value of $R(V_k, W_k, \rho_k, \zeta(V_k, W_k, \rho_k, (1, 0)))$, that is, the value of $R(V_k, W_k, \rho_k, p)$, where p is the the final endpoint of V_k, and we define W_{k+1} to be the final translate of W_k along V_k to the end of V_k, that is, $W_{k+1} = \zeta(V_k, W_k, \rho_k, (1, [0, L_k]))$, where $L_k = \lambda(W_k)$. By the foregoing $L_{k+1} \leq L_k + K\Delta$, and the total curvature of V_k can be any $-\delta_k$, for which $0 \leq \delta_k \leq \frac{1}{2K(L_k + K\Delta)}$. We obviously have $L_k \leq L_0 + kK\Delta$, so that the only restriction on δ_k is $0 \leq \delta_k \leq \frac{1}{2K(L_0 + (k+1)K\Delta)}$, this upper bound obviously being the general term of a divergent series for $\Delta > 0$. Let $\epsilon = \frac{1}{100}$, so that if $W_0 = C$, where C is the appropriate translate of any element of $\mathcal{S}(\epsilon)$ as indicated in (vii), then $\pi - \frac{1}{10} \leq \lambda(W_0) \leq \pi + \frac{1}{10}$ and $\Delta = \pi + \epsilon$. Clearly, there is some positive integer M such that for some sufficiently small $\eta \in (0, \frac{\pi}{4})$ there is a curve V, the union of $V_0, \ldots, V_M$, parameterized by $v(s)$, $0 \leq s \leq M$, such that V terminates in a straight line segment whose initial point is in $\mathbb{H}$ and whose slope is $-\tan\eta$. We note that by reducing the size of ϵ if necessary, we can assume that the first subarc V_0 of $V \subset V^1$ intersects $\mathbb{R}$ at a point $v_0(s)$ for which $\arg\{v_1'(s)\} \in (-\pi, -\pi + \delta_0) \subset (-\pi, -\frac{3\pi}{4})$. In addition, we can allow the length of the final straight segment of V to be as large as we want, so that we can take the imaginary part of the terminal point of V to be $-\epsilon$. We emphasize that V is independent of W_0. We next extend V at its right end by smoothly adjoining a ray with slope exactly $-\frac{\pi}{4}$ (by means of $F(v, \lambda(V) - \frac{\epsilon}{2}, \frac{\epsilon}{2}, 0, -\frac{\pi}{4})$, as described above, where $v(s)$ parameterizes V); this extended curve will be called V^1. We now consider the characteristic initial value problem with V^1 as the 1-arc and, as the 2-arc, a segment A of length r orthogonal to it at its initial point and emanating to the left of V^1. Because V^1 satisfies the OSC, the solution so generated will be C^∞ in the domain made up of $E(C \cap \mathbb{H}, V^1 \cap \mathbb{H})$ together with translates of A around V^1. Note that the corresponding solution is constant in the quarter-plane made up of all the translates of A emanating from points on the ray in which V^1 ends. From this it is clear that for $r = \lambda(A)$ sufficiently large, an appropriate subarc of the translate $V^1(r)$

of V^1 to the outer endpoint of A has properties (i), (v), (vi), and (vii). Properties (ii), (iii), and (iv) can be achieved in the following manner. First, we take r so large that at the left intersection point of $V^1(r)$ with $\mathbb{R}$ the angle formed by $V^1(r)$ is within $\frac{\epsilon}{4}$ of $\frac{\pi}{2}$. Then we remove the arc of $V^1(r)$ joining its initial point to a point with imaginary part between $-\frac{\epsilon}{2}$ and 0 and at which the direction of the tangent vector is within $\frac{\epsilon}{2}$ of $\frac{\pi}{2}$. Finally, we can use the operator F to smoothly adjoin an arc to the remaining part of $V^1(r)$ in such a way that (ii), (iii), and (iv) hold. The resulting curve is our U_1. Note that in (vii) we do not require the solution to bear any particular relation to the part of U_1 that lies in the lower half-plane.

The left and right endpoints of U_1 will be called e_1 and m_1; the left and right endpoints of U_2 will be called m_2 and e_2 (m for middle and e for end, for reasons to be made clear momentarily). It is clear that U_1 depends solely on Θ and the parameter ϵ and that once these have been fixed, the solution R generated in $E(C)$ depends on C but that its values on U_1 depend only on $\rho = R(c_l)$, c_l being the left endpoint of C; analogous statements hold for U_2. For $\rho \in \mathbb{R}^2$ we denote by $f_k(\rho)$ the value of this solution R at m_k. We also observe that in the case of U_1 we always have

$$\theta(f_1(\rho)) - \theta(\rho) = \theta(m_1) - \theta(c_l) = -\left(\frac{5\pi}{4} + \epsilon\right). \tag{5.6}$$

As indicated above, in the case of U_2 we still take as ρ the value of R at the left endpoint c_l of the initial curve C used in the above construction which, as we have pointed out, uniquely determines the value of R at the right endpoint of C since the total curvature of all arcs in the family $\mathcal{S}(\epsilon)$ is $\pi + 2\epsilon$. Here we have that

$$\theta(f_2(\rho)) - \theta(\rho) = \theta(m_2) - \theta(c_l) = -(\pi + 2\epsilon) + \left(\frac{5\pi}{4} + \epsilon\right) = \frac{\pi}{4} - \epsilon. \tag{5.7}$$

The f_k are continuous functions of ρ and there is a number B_0 (that depends only on the bounds on the $\frac{\partial\theta}{\partial R_k}$ associated with Θ) such that

$$\|f_k(\rho) - \rho\| \le B_0, \quad \rho \in \mathbb{R}^2, \quad k = 1, 2, \tag{5.8}$$

where the norm is just the Euclidean norm in $\mathbb{R}^2$ (actually, any norm will do). Both the continuity and the bound come from the simple observation that on any i-characteristic, R_i is constant and R_j is a Lipschitz continuous function of the tangent inclination (with the Lipschitz constant depending solely on the system Θ).

For our next step we consider translates $U_1' = U_1 - m_1$ and $U_2' = U_2 - m_2$ which have the common point 0. Note that U_1' and U_2' are orthogonal to each other at 0 and so can be used as initial curves for a characteristic initial value problem, U_k' being a k-arc. For ϵ sufficiently small (so that U_1 and U_2 are virtually orthogonal to the horizontal line $\Im\{z\} = -\epsilon$ at e_1 and e_2, respectively) convexity considerations easily show that for any admissible corner value $\rho = R(0)$ the mapping $\zeta(U_1', U_2', \rho, t)$ is one-to-one on the corresponding characteristic coordinate rectangle. To see that this is indeed the case, note that since the change in θ from the left end of U_1' to the right end of U_2' is $-(\frac{3\pi}{2} + 2\epsilon)$ (that is, the change in θ along U_1' + the change in θ along U_2', both from left to right, is $-(\frac{3\pi}{2} + 2\epsilon)$), the total change along the top two sides of the characteristic quadrilateral is also $-(\frac{3\pi}{2} + 2\epsilon)$. The resulting solution $R_\rho(z) = R(U_1', U_2', \rho, z)$ and the quadrilateral $Q_\rho = Q(U_1', U_2', \rho)$ itself depend solely on $\rho = R(0)$. Note also that because U_1', $U_2' \in \mathcal{CL}$, the other two sides of Q_ρ also belong to $\mathcal{CL}$.

Next we next show how we can produce curves $X_k \in \mathcal{S}(\epsilon, t_k, \alpha_k)$, such that $B(X_k \cap \overline{\mathbb{H}}) \supset Q_\rho \cap \mathbb{H}$, and a solution in $B(X_k \cap \overline{\mathbb{H}}) \backslash (Q_\rho \cap \mathbb{H})$ which gives a C^∞ extension of R_ρ for which X_k is a one-sided k-arc which satisfies the OSC. We show how to get X_2, the case of X_1 being identical apart from obvious details.

First, we extend U_2' on the right by continuously adjoining a circle of radius $\frac{\epsilon}{10}$ lying in the open lower half-plane by using $F(\overline{u}_2 - m_2, \lambda(U_2) - \delta, \delta, \frac{10}{\epsilon}, -(\frac{\pi}{2} + \epsilon))$, where $\overline{u}_2$ parameterizes U_2 from left to right (i.e., with $\overline{u}_2(0) = m_2$) and where δ is so small that $\overline{u}_2([\lambda(U_2) - \delta, \lambda(U_2)])$ is a line segment. This in turn allows us to extend the other 2-side of Q_ρ rightwards and eventually downwards (in light of the quasi-HP property) until it gets to a point p for which $\Im\{p\} = -\frac{\epsilon}{2}$; call this extended 2-side T. The corresponding solution is defined in the simply covered domain $E = E(U_1' \cup U_2', T')$, where T' is the subarc of T which joins a point of $\mathbb{R}$ to the left of U_1' to a point of $\mathbb{R}$ to the right of U_2'. This means that if we change T below $\mathbb{R}$ to form a new curve T_2, which we subsequently use as the 2-arc for a characteristic initial value problem with a straight initial 1-arc outside of E, then the solution is compatible with the restriction to E of the solution we have so far. Using an appropriate instance of the operator F, we obtain an extension T_2 of T by smoothly adjoining to T at its right end a segment in such a way that the argument of the tangent at the right endpoint of T_2 is less than or equal to $-(\pi/2 + \epsilon)$. Obviously, $T_2 \in \mathcal{CL}$. It is also clear from the construction that T_2 depends solely on $\rho = R(0)$ and that all derivatives of its arc length parameterization depend continuously on ρ. This outside curve T_2 clearly satisfies the hypotheses (5.4) and (5.5) of the straight line characteristic initial value construction. Using a straight 1-arc of length r_0 we therefore obtain an "almost circular" 2-arc S parallel T_2 which is contained in $N(0, (1 + \epsilon/2)r_0) \backslash (0, (1 - \epsilon/2)r_0)$. Let $q_1 < q_2$ be the points of $S \cap \mathbb{R}$. By choosing r_0 sufficiently large we can be certain that the interior angles of this S at q_1 and q_2 are within $\frac{\epsilon}{2}$ of $\frac{\pi}{2}$. Simple curvature considerations show that there is a single, sufficiently large value of r_0 that works for all admissible corner values $\rho = R(0)$. Let p_1, $p_2 \in T_2$ lie on the straight 1-characteristics which terminate at q_1 and q_2, respectively. It is clear that for the solution our process has generated, $R(p_1)$ and $R(p_2)$ are continuous functions of ρ which, moreover, satisfy bounds of the form (5.8) (with $f_k(\rho)$ replaced by $R(p_k)$ and with an appropriate B_0). The solution is defined and C^∞ in the part of $\mathbb{H}$ between $U_1' \cup U_2'$ and S, and S is a 2-characteristic arc satisfying the OSC. Finally, we smoothly alter S below $\mathbb{R}$ to obtain an S' such that for appropriate $\xi = \xi(\rho)$, the arc $X_2 = \xi S'$ belongs to $\mathcal{S}(\epsilon)$. If x_l is the left endpoint of X_2, then $g_2(\rho) = R(x_l)$ is a continuous function of ρ, and similarly for the value $g_1(\rho)$ of R at the left endpoint of the analogously constructed X_1. Here again one easily sees that in the case of X_2

$$\theta(g_2(\rho)) - \theta(\rho) = \frac{3\pi}{4} + \epsilon - \frac{\pi}{2} = \frac{\pi}{4} + \epsilon \tag{5.9}$$

and in the case of X_1 that

$$\theta(g_1(\rho)) - \theta(\rho) = \frac{3\pi}{4} + \epsilon. \tag{5.10}$$

Let $U = U_1' \cup U_2'$. By the universal property of U_1', for appropriate $t^{(2)}$ and $\alpha^{(2)}$ we can use $t^{(2)}X_2 + \alpha^{(2)}$ as the 2-arc under U_1' (see property (vii) in the discussion of the universal arcs given above), and similarly we can use some $t^{(1)}X_1 + \alpha^{(1)}$ as the 1-arc under U_2. These $t^{(k)}X_k + \alpha^{(k)}$ in turn come from $t^{(k)}U + \alpha^{(k)}$. If we perform the construction of X_k using the value $\rho^{(k)}$ of R at the center $\alpha^{(k)}$ of $t^{(k)}U + \alpha^{(k)}$,

then the value of R at the left endpoint of X_k is $g_k(\rho^{(k)})$, so that the value of the solution R (corresponding to property (vii) of universal arcs) in the part of the $B(U_j')$ between $t^{(i)}U + \alpha^{(i)}$ and U_j' at 0 is $f_j(g_i(\rho^{(i)}))$. We claim that there is a $\rho^{(i)}$ for which $f_j(g_i(\rho^{(i)}))$ is any given admissible corner value ρ_0 for the characteristic initial value problem with 1- and 2-arcs U_1 and U_2. To see that such a $\rho^{(i)}$ exists in the case that $i = 2$, $j = 1$, for example, note that from (5.6) and (5.9) we have that

$$\theta(f_1(g_2(\rho^{(2)}))) - \theta(\rho^{(2)}) = \theta(f_1(g_2(\rho^{(2)}))) - \theta(g_2(\rho^{(2)})) + \theta(g_2(\rho^{(2)})) - \theta(\rho^{(2)})$$

$$= -\left(\frac{5\pi}{4} + \epsilon\right) + \frac{\pi}{4} + \epsilon = -\pi.$$

Thus, given ρ_0, the point $\rho^{(2)}$ must lie on the curve

$$\theta(R) = \theta(\rho_0) + \pi \tag{5.11}$$

in the R-plane. The level curves $\theta(R) = \theta_0$ in the R-plane are the graphs $R_i = h_{i,\theta_0}(R_j)$ of monotone functions, for which h_{i,θ_0}' are uniformly bounded and uniformly bounded away from 0. Thus as $\rho^{(2)}$ moves along the curve (5.11) it follows from the continuity of f_1 and g_2 and the bound (5.8) and the corresponding bound for the g_k that $f_1(g_2(\rho^{(2)}))$ traces out the entire curve $\theta(R) = \theta(\rho_0)$, so that there is indeed a (unique) $\rho^{(1,2)}$ for which $f_1(g_2(\rho^{(2)})) = \rho_0$. The case of $f_2(g_1(\rho^{(1)})) = \rho_0$ is handled in the same manner, using (5.7) and (5.10) instead of (5.6) and (5.9).

It is now clear that we can iterate this construction to produce a solution $R^* = (R_1^*, R_2^*)$ in $B(U_1' \cap \overline{\mathbb{H}}) \cup B(U_2' \cap \overline{\mathbb{H}})$, and, in fact, a solution in all of $\mathbb{H}$, by appropriately extending the solution we have in X_1 (or in X_2, for that matter). More specifically, the solution in $B(U_1' \cap \overline{\mathbb{H}}) \cup B(U_2' \cap \overline{\mathbb{H}})$ is obtained by placing a suitable similar copy of the form $tU + \alpha$ under each of U_k' in the way described and then under each of the corresponding $tU_k' + \alpha$ smaller similar copies of the form $tU + \alpha$ (with a smaller value of t, of course). At the nth stage (where we regard the initial U as corresponding to the 0th stage) we have 2^n disjoint similar copies $U_l^{(n)}$, $1 \le l \le 2^n$, of U. Let $I_l^{(n)}$ be the interval of $\mathbb{R}$ joining the left and right endpoints of $U_l^{(n)} \cap \overline{\mathbb{H}}$. It is an immediate consequence of the shape of the U_k' that (once ϵ has been fixed) there is a constant $\delta_1 = \delta_1(\Theta) > 0$ such that for all points $p \in I_l^{(n)}$

$$\lambda(\theta(R^*(\{z : \delta_1 < \arg\{z - p\} < \pi - \delta_1,\ z \in U_l^{(n)} \cap \mathbb{H}\}))) \ge \delta_1.$$

Since, on U_i', θ and R_j are bi-Lipschitz functions of each other, it follows that there are $r_1 = r_1(\Theta)$, $\delta_2 = \delta_2(\Theta) > 0$ such that for all $p \in I_l^{(n)}$ the range of R_k^* in

$$N(r_1^n, p) \cap \{z : \delta_1 < \arg\{z - p\} < \pi - \delta_1\}$$

is an interval of length at least δ_2, $k = 1, 2$. Let $M^{(n)} = \cup\{I_l^{(n)} : 1 \le l \le 2^n\}$. Then $M = \cap\{M^{(n)} : n \ge 1\}$ consists entirely of boundary singularities of the solution R^* that we have constructed. To those who have read about Cantor sets it is probably obvious that M has positive Hausdorff dimension, but for completeness we include an appropriately modified version of the argument given by Falconer [F] for the classical excluded middle third case.

First of all, it is clear from the self-similar nature of the construction that there is a number $\gamma \in (0, 1)$ such that the minimum distance between the 2^n intervals

making up the set $M^{(n)}$ is bounded below by γ^n. We show that the τ-dimensional Hausdorff measure of M is positive, where τ is defined by

$$\gamma^\tau = \frac{1}{2}.$$

Let $\{G_k\}$ be an open cover of M, which we can assume to be finite since M is compact. Let $\max\{\text{diam}(G_k)\} < 1$. For each k there is a nonnegative integer $l = l(k)$ such that

$$\gamma^{l+1} \le \text{diam}(G_k) < \gamma^l.$$

From the definition of γ it follows that such a G_k can have a nonempty intersection with at most one of the intervals that make up $M^{(l)}$. Consequently, for $p \ge l(k)$ at most $2^{p-l(k)}$ of the intervals making up $M^{(p)}$ can have a nonempty intersection with such a G_k. For each $p \ge l = l(k)$ we therefore have

$$2^{p-l} = \frac{2^p}{2^l} = \frac{2^p(\gamma^{l+1})^\tau}{2^l\gamma^\tau l\gamma^\tau} \le \frac{2^p(\text{diam}(G_k))^\tau}{(2\gamma^\tau)^l\gamma^\tau}.$$

Let p be so large that $\min\{\text{diam}(G_k)\} \ge \gamma^{p+1}$, so that $p \ge l(k)$ for all k. Since there are 2^p intervals in $M^{(p)}$, if we denote by N_k the number of intervals touched by G_k, then

$$2^p \le \sum_k N_k \le \sum_k \frac{2^p(\text{diam}(G_k))^\tau}{(2\gamma^\tau)^{l(k)}\gamma^\tau} = \sum_k \frac{2^p(\text{diam}(G_k))^\tau}{\gamma^\tau},$$

so that

$$\sum_k (\text{diam}(G_k))^\tau \ge \gamma^\tau = \frac{1}{2}.$$

This means that the τ-dimensional Hausdorff measure of M is at least $\frac{1}{2}$, so that M has Hausdorff dimension at least τ.

6. A few concluding remarks. We briefly discuss some of the issues and problems suggested by the foregoing. In the first place, it would be interesting to determine whether Corollary 4.3 is true for genuinely nonlinear systems (see the definitions of the terms "system" and "genuinely nonlinear" between relations (1.4) and (1.5)) which are not necessarily normal. Probably, though, some hypothesis in the spirit of (ii) of Definition 1.1 as well as some bound like

$$\epsilon < |\theta_1(R) - \theta_2(R)| < \pi - \epsilon \quad \text{for all } R \in \mathbb{R}^2$$

is necessary, so that, using the approach of section 5, or otherwise, one might try to construct a solution of a 2×2 genuinely nonlinear hyperbolic system not satisfying one or the other or both of these conditions and which has a set of boundary singularities of Hausdorff dimension 1. In a wider context one can ask if Corollary 4.3 has any counterparts for an appropriate class of sufficiently nonlinear $n \times n$ planar hyperbolic systems. In reference to normal systems, our analysis leaves open the question of whether in a half-plane $\mathbb{H}$, for example, there can be a solution for which the set of boundary singularities of type 1 has positive Hausdorff dimension. Corresponding to such a solution there would have to be set $A \subset \partial G$ of positive Hausdorff dimension such that for each $a \in A$ there is a characteristic C_a exiting at a but for which

hypothesis (3.22) of Proposition 3.31 does not hold. Also open is the question of whether the word "nontangential" is necessary in the conclusion of Main Theorem 2.2.

As we shall show elsewhere, the ideas of section 5 can be used to construct in an arbitrary Jordan domain cps-mappings with arbitrary (distinct) principal stretch factors which have infinitely many isolated singularities, and in fact these singularities can be of spiral type (see [G3] for the classification of isolated singularities of cps-mappings). This raises the question of how such singularities can be distributed, and in this regard we conjecture that there is some absolute constant $\gamma > 1$ such that if $\{a_n\}$ is the sequence of isolated singularities of any cps-mapping in any smoothly bounded Jordan domain for which $\{\text{dist}(a_n, \partial G)\}$ is nonincreasing, then

$$\sum_n \text{dist}(a_n, \partial D)\gamma^n < \infty. \tag{6.1}$$

Note that this was shown with $\gamma = 1$ in [G3, Corollary 4.1]. More generally, there are other 2×2 genuinely nonlinear systems for which there exist corresponding unambiguously defined nets of characteristics which have isolated singularities, and for any such system one could attempt to obtain a classification of such singularities along the lines of [G3] and seek a bound of type (6.1) on their density.

REFERENCES

[CS] C. Carathéodory and E. Schmidt, *Über die Hencky-Prandtlschen Kurven*, Z. Angew. Math. Mech., 3 (1923), pp. 468–475.

[ChG] M. Chuaqui and J. Gevirtz, *Constant principal strain mappings on 2-manifolds*, SIAM J. Math. Anal., 32 (2000), pp. 734–759.

[F] K. Falconer, *Fractal Geometry. Mathematical Foundations and Applications*, John Wiley and Sons, Chichester, UK, 1990.

[G1] J. Gevirtz, *On planar mappings with prescribed principal strains*, Arch. Rational Mech. Anal., 117 (1992), pp. 295–320.

[G2] J. Gevirtz, *A diagonal hyperbolic system for mappings with prescribed principal strains*, J. Math. Anal. Appl., 176 (1993), pp. 390–403.

[G3] J. Gevirtz, *Hencky-Prandtl nets and constant principal strain mappings with isolated singularities*, Ann. Acad. Sci. Fenn. Math., 25 (2000), pp. 187–238.

[G4] J. Gevirtz, *Singularity sets of constant principal strain deformations*, J. Math. Anal. Appl., 263 (2001), pp. 600–625.

[G5] J. Gevirtz, *Boundary values and the transformation problem for constant principal strain mappings*, Int. J. Math. Math. Sci., 2003 (2003), pp. 739–776.

[GM] V. Gutlyanskii and O. Martio, *Rotation estimates and spirals*, Conform. Geom. Dyn., 5 (2001), pp. 6–20.

[Hem] W. S. Hemp, *Optimum Structures*, Clarendon Press, Oxford, UK, 1973.

[Hen] H. Hencky, *Über einige statisch bestimmte Fälle des Gleichgewichts in plastischen Körpern*, Z. Angew. Math. Mech., 3 (1923), pp. 241–251.

[Hi] R. Hill, *The Mathematical Theory of Plasticity*, Clarendon Press, Oxford, UK, 1964.

[J1] F. John, *On quasi-isometric mappings*, I, Comm. Pure Appl. Math., 21 (1968), pp. 77–110.

[J2] F. John, *On quasi-isometric mappings*, II, Comm. Pure Appl. Math., 22 (1969), pp. 265–278.

[L] P. Lax, *Development of singularities of solutions of nonlinear hyperbolic partial differential equations*, J. Mathematical Phys., 5 (1964), pp. 611–613.

[Pr] L. Prandtl, *Über die Eindringungsfestigkeit (Härte) plastischer Baustoffe und die Festigkeit von Schneiden*, Z. Angew. Math. Mech., 1 (1921), pp. 15–20.

[Y] W.-L. Yin, *Two families of finite deformations with constant strain invariants*, Mech. Res. Comm., 10 (1983), pp. 127–132.

SIAM J. MATH. ANAL.
Vol. 40, No. 4, pp. 1337–1391

ASYMPTOTIC STABILITY OF ASCENDING SOLITARY MAGMA WAVES*

GIDEON SIMPSON† AND MICHAEL I. WEINSTEIN†

Abstract. Coherent structures, such as solitary waves, appear in many physical problems, including fluid mechanics, optics, quantum physics, and plasma physics. A less studied setting is found in geophysics, where highly viscous fluids couple to evolving material parameters, modeling partially molten rock, magma, in the Earth's interior. Solitary waves are also found here, but the equations lack useful mathematical structures such as an inverse scattering transform or even a variational formulation. A common question in all of these applications is whether or not these structures are stable to perturbation. We prove that the solitary waves in this earth science setting are asymptotically stable and accomplish this without any preexisting Lyapunov stability. This holds true for a family of equations, extending beyond the physical parameter space. Furthermore, this extends existing results on well-posedness to data in a neighborhood of the solitary waves.

Key words. solitary waves, stability, viscously deformable porous media, magma

AMS subject classifications. 74J35, 35B40, 74L05

DOI. 10.1137/080712271

1. Introduction.

Coherent structures, such as solitary waves, appear in many physical problems, including fluid mechanics, optics, quantum physics, and plasma physics. A less studied setting is found in geophysics, where highly viscous fluids couple to evolving material parameters, modeling partially molten rock, magma, in the Earth's interior. Solitary waves are also found here, but the equations lack the useful structures such as an inverse scattering transform or even a variational formulation.

A important question in all of these applications is whether or not the coherent structures are stable to perturbation. We prove that the solitary waves in this earth science setting are asymptotically stable and accomplish this without any preexisting Lyapunov stability.

1.1. Magma: Porous flow in a viscously deformable media.

Models of magma in the Earth's interior couple Stokes flow of the viscous melt to the slow, creeping deformation of the porous rock. These force-balance equations couple to transport equations for the volume fraction of melt, the *porosity*. Formulations may be found in [5, 20, 35, 36, 41]. Nonlinearity appears in fluxes and through the material properties, the *permeability* and *viscosity* of the porous, deformable rock, which depend nonlinearly on the porosity. Consequently, such models are known, from computations, to feature localization and generate coherent structures; see [1, 2, 3, 15, 35, 36, 41, 42, 43, 44, 51]. The physical assumptions and their implications will be discussed in a forthcoming review article [39].

Under certain assumptions (small fluid fraction, absence of large-scale shear, no melting, etc.), such a system reduces to a single scalar equation for the porosity's

*Received by the editors January 2, 2008; accepted for publication (in revised form) June 20, 2008; published electronically November 5, 2008. This work was funded in part by the U.S. National Science Foundation (NSF) Collaboration in Mathematical Geosciences (CMG), Division of Mathematical Sciences (DMS) grant DMS-05-30853, NSF Integrative Graduate Education and Research Traineeship (IGERT) grant DGE-02-21041, and NSF grants DMS-04-12305 and DMS-07-07850.
http://www.siam.org/journals/sima/40-4/71227.html
†Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027 (grs2103@columbia.edu, miw2103@columbia.edu).

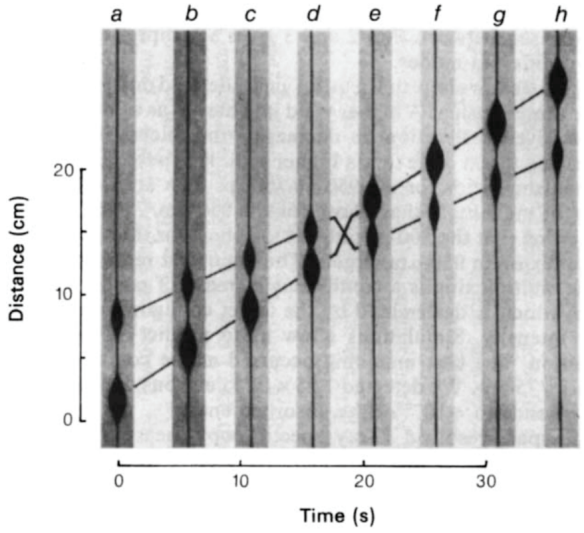


FIG. 1. *Solitary waves colliding and propagating in an experiment with honey. From Figure* 1 *of* [37]. *Reprinted by permission from Macmillan Publishers Ltd.*

evolution [2, 3, 41, 42]. The d-dimensional equation is

$$\partial_t \phi + \partial_z \left(\phi^n\right) - \nabla \cdot \left[\phi^n \nabla \left(\phi^{-m} \partial_t \phi\right)\right] = 0, \quad \mathbf{x} \in \mathbb{R}^d, t > 0, \tag{1.1}$$

with the boundary conditions that $\phi(\mathbf{x}, t) \to 1$ as $|\mathbf{x}| \to \infty$. $\nabla = (\partial_x, \partial_y, \partial_z)$ for $d = 3$ and $\nabla = (\partial_x, \partial_z)$ for $d = 2$. The nonlinearity n comes from the relationship between the permeability, K, and the porosity of the rock, $K \propto \phi^n$. m relates to the bulk viscosity, ζ, to the porosity of the rock, $\zeta \propto \phi^{-m}$. In the physical regime, these exponents have values $2 \leq n \leq 3$ and $0 \leq m \leq 1$ [12, 13, 20, 31, 35, 36, 45, 46, 52].

Equation (1.1) appears elsewhere in earth science as a model for convective mantle plumes. Manifesting themselves as *hot spots* at the surface, these plumes transport warm buoyant material. Examples include the Hawaiian Island chain and Iceland. Modeled as the flow of a viscous fluid up a conduit embedded in a higher viscosity medium, an equation of the form (1.1) was derived in [24]. There, the equation is one-dimensional ($d = 1$), the exponents are $(n, m) = (2, 1)$, and the dependent variable ϕ is the cross-sectional area of the pipe.

Numerical simulations of (1.1) in one, two, and three dimensions were performed in [2, 3, 35, 36, 51], where stable, radially symmetric, *solitary traveling waves* were observed. In [23], it was shown that in one dimension, solitary waves, $U_c(x - ct)$, in excess of the reference state, $\phi \equiv 1$, exist for $n > 1$. In the context of conduit flow, discussed in the preceding paragraph, analogue experiments using viscous syrups appear in [24, 37, 49]; robust solitary waves appeared as predicted; see Figure 1.

1.2. Stability of solitary waves. We consider (1.1) in one dimension,

$$\partial_t \phi + \partial_x \left(\phi^n\right) - \partial_x \left[\phi^n \partial_x \left(\phi^{-m} \partial_t \phi\right)\right] = 0, \quad \lim_{|x| \to \infty} \phi(x, t) = 1, \tag{1.2}$$

where the z coordinate has been relabeled x. A cursory explanation for the solitary waves' stability may be found in [50]. Under a small amplitude scaling, (1.2) is, to leading order, governed by the Korteweg–de Vries (KdV) equation. Since KdV solitons

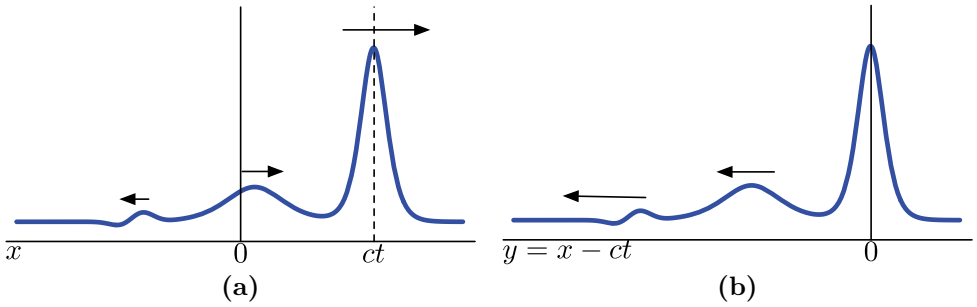


FIG. 2. *The largest solitary wave in* (a) *travels faster than smaller solitary waves and dispersive waves behind it. In the frame of this wave, the rest of the solution appears to move leftward, as in* (b).

are stable, on a timescale for which KdV approximates (1.2) its solitary waves should also be stable.

Based on observations of numerical experiments, we expect a slightly perturbed solitary wave to evolve into another wave with similar amplitude and phase. It will be accompanied by some small amplitude dispersive waves and, perhaps, another solitary wave of smaller amplitude. The leading wave will *outrun* these other disturbances, cease interacting with them, and stabilize.

Some intuition for this stability may be found in two properties. First, taller solitary waves travel with greater speed, c, than smaller ones. In the frame of the largest solitary wave, $y = x - ct$, the other waves travel leftward. Second, in the frame of the leading solitary wave, small perturbations of the reference state, $\phi(x,t) = 1 + \psi(x - ct, t)$ and $|\psi| \ll 1$, evolve under the linear flow

$$\partial_t \psi - c\partial_y \psi + n\partial_y \psi - \partial_y^2 \partial_t \psi + c\partial_y^3 \psi = 0.$$

The dispersion relation and group velocity are

$$\omega(k) = \frac{nk}{1+k^2} - ck, \quad \omega'(k) = n\frac{1-k^2}{(1+k^2)^2} - c.$$

Since the solitary waves of (1.2) that we study travel with speed $c > n > 1$, both phase and group velocities are negative for all k. Therefore, small dispersive waves also travel leftward. These two mechanisms are diagrammed in Figure 2, and we will exploit them to prove the main theorems.

As the system is conservative, perturbations, such as a small solitary wave, will not vanish in a translation invariant norm. A suitable norm will register leftward motion as decay. We will use exponentially weighted norms, in the frame of the leading solitary wave. These norms are defined in section 1.4.

Several paths to proving stability are available. One method is to seek constants of motion that can be combined into a metric centered at the solitary wave. Since the metric is time independent, if the perturbation is initially small, it will remain so. This elegant method relies on the calculus of variations, and for equations such as KdV and nonlinear Schrödinger (NLS) it may be used to prove *orbital stability* [4, 8, 47, 48]. A solitary wave, U_c, is said to be orbitally stable if for data sufficiently close to it

$$\inf_{y\in\mathbb{R}} \|u(t) - U_c(\cdot + y)\|_X < \delta \quad \text{for all } t$$

for some $\delta > 0$ and an appropriate norm $\|\cdot\|_X$. Typically, the norm is equivalent to $L^2(\mathbb{R})$ or $H^1(\mathbb{R})$.

However, in general, (1.2) lacks a sufficient number of conservation laws for this approach. Indeed, in [3], the authors searched for an additional conservation law, in hopes of proving orbital stability [7]. There are many such equations, including some of the Boussinesq systems [29] and many of the "compacton" equations [32], which also lack such structure, yet appear, in numerical experiments, to possess stable solitary waves. Note that when $n + m = 0$, (1.2) is Hamiltonian, and we have investigated this, proving orbital stability in [40]. We wish to consider the general case, which includes the physically interesting cases $(n, m) = (2, 1)$ and $(3, 0)$.

Another approach to stability is to linearize the problem $u_t = N(u, u_x, u_{xx}, \dots)$ about a solitary wave and establish linear stability. Then one seeks a way to perturbatively "boost" this to prove stability for the nonlinear flow. This may rely on direct spectral analysis of the linearized evolution operator. We employ this method, following the work of Pego and Weinstein [28, 29] and Miller and Weinstein [21]. Through this, we prove that the solitary waves are *asymptotically stable*, our main result. By this we mean that in an appropriate norm, $\|\cdot\|_Y$,

$$\|u(t) - U_c\|_Y \to 0 \quad \text{as } t \to +\infty$$

for data sufficiently close to U_c.

We note that another method recently appeared in the work of Martel and Merle [19]. Without linearizing, the authors employ a virial inequality to directly prove asymptotic stability of generalized KdV solitary waves in $H^1_{\text{loc}}(\mathbb{R})$.

Our problem is an example of an equation for which one can prove asymptotic stability in the absence of orbital stability. Upon reflection, it is clear that the asymptotic stability of generalized KdV and Benjamin–Bona–Mahony (BBM) solitary waves could have been proven without using the orbital stability results.

1.3. Main results and outline. The main results are as follows.

THEOREM 1.1. *There exists $c_\star > n$ such that for all $c_0 \in (n, c_\star]$, if $\phi_{c_0}(x - \theta_0)$ is a solitary wave solution of (1.2), then there exist constants $K_* > 0$, $\epsilon_* > 0$, and $a > 0$, such that for $\epsilon \le \epsilon_*$, if*

$$\|v_0\|_{H^1} + \|e^{ax} v_0\|_{H^1} \le \epsilon,$$

then

(a) (1.2) *has a solution with data $\phi_0(x) = \phi_{c_0}(x + \theta_0) + v_0(x)$ for all time;*
(b) *there exist c_∞, θ_∞, K_*, and $\kappa > 0$ such that*

$$\|\phi(\cdot, t) - \phi_{c_\infty}(\cdot - c_\infty t + \theta_\infty)\|_{H^1} \le K_* \epsilon, \tag{1.3}$$

$$\|e^{ax}\left[\phi(\cdot + c_\infty t - \theta_\infty, t) - \phi_{c_\infty}(\cdot)\right]\|_{H^1} \le K_* \epsilon e^{-\kappa t}, \tag{1.4}$$

$$|c_\infty - c_0| + |\theta_\infty - \theta_0| \le K_* \epsilon. \tag{1.5}$$

COROLLARY 1.2. *Let $n + m = 0$. If $\partial_c \mathcal{N}[\phi_c] > 0$, $\mathcal{N}$ defined in (2.8), then Theorem 1.1 holds for all $c > c_\star$, except for a discrete set with no accumulation point.*

Remark 1.3. For general $n > 1$, Theorem 1.1 is limited to $c \le c_\star$ because we are only able to rigorously treat the spectrum of a linear operator perturbatively. This can be extended by a numerical computation of the spectrum. See sections 3.3–3.4.

The feature of (1.2) that allows us to prove nonlinear stability from the linear stability is a nonnegative invariant, denoted $\mathcal{N}[\phi]$, and defined in (2.8). The Taylor expansion of $\mathcal{N}$ about a solitary wave is

$$\mathcal{N}[\phi_c + v] = \mathcal{N}[\phi_c] + \langle \delta\mathcal{N}[\phi_c], v\rangle + \left\langle \delta^2\mathcal{N}[\phi_c]v, v\right\rangle + O\left(\|v\|_{H^1}^3\right).$$

The first variation does not vanish and the second variation is not a positive definite quadratic form. However, in the frame of the solitary wave, the perturbation v is migrating to $-\infty$. Therefore

$$\langle \delta\mathcal{N}[\phi_c], v\rangle \to 0 \quad \text{as } t \to +\infty.$$

The second variation may be decomposed as $\delta^2\mathcal{N}[\phi_c] = P + Q$, P a positive quadratic form and $Q = Q(x)$ a localized function. Then since the perturbation moves leftward

$$\begin{aligned} &\langle Pv, v\rangle \geq \kappa^2 \langle v, v\rangle, \\ &\langle Qv, v\rangle \to 0 \quad \text{as } t \to +\infty. \end{aligned}$$

Asymptotically,

$$\|v\|_{H^1} \leq K\Delta\mathcal{N}, \quad \Delta\mathcal{N} = |\mathcal{N}[\phi_c + v] - \mathcal{N}[\phi_c]|,$$

giving a Lyapunov-type bound on the perturbation. Control of $\Delta\mathcal{N}$ is essential in using this estimate to prove nonlinear stability. Though we rely on the invariance of $\mathcal{N}$, if $\Delta\mathcal{N}$ were merely bounded in terms of the data, it would be sufficient.

However, more is needed to formalize this into a proof, notably a sense in which the perturbation recedes from the solitary wave. This is accomplished by analyzing the spectrum of the linearized evolution operator in a weighted space, in which the perturbation will decay.

The plan of the proof is as follows:

(I) In section 2 we review properties of (1.2) and establish regularity properties of the solitary waves.

(II) In section 3, we prove that the linearized operator, A_a, has the property that there exists $\varepsilon > 0$ such that

$$\sigma(A_a) \cap \{\Re\lambda \geq -\varepsilon\} = \{0\}$$

and zero is an eigenvalue of algebraic multiplicity two.

(III) In section 4, we prove

$$\|w(t)\|_{H^1} = \|e^{A_a t}w_0\|_{H^1} \leq Ke^{-bt}\|w_0\|_{H^1}$$

for appropriate w_0, K, and b positive constants.

(IV) In section 5, we make several estimates, including a formalization of the Lyapunov bound. We also formulate equations for the speed and phase parameters of the solitary wave $(c(t), \theta(t))$, coupling them to the infinite-dimensional system for the perturbation.

(V) In section 6, we prove the main results, asymptotic stability and global existence of data near a solitary wave solution.

Some remarks are made in section 7, and additional details are located in the appendices.

1.4. Notation. Generic constants will typically be denoted by the capital letters K, M, and N, sometimes with tildes, overlines, or primes. Subscripts, such as M_γ, may appear in order to indicate that M depends on γ. We avoid using C as a generic constant, as c appears throughout the paper as the speed parameter, and an operator $C(\lambda)$ appears in section 3.

Functions will typically live in spaces $H^k(\mathbb{R}) = W^{1,k}(\mathbb{R})$, k a nonnegative integer, the spaces of square integrable functions with square integrable (weak) derivatives up to order k. We will frequently omit writing $\mathbb{R}$. The $L^p(\mathbb{R})$ spaces will also appear, in particular L^2 and L^∞. While we write $\|f\|_{H^k}$ for the norm of a function in H^k, we only write $\|f\|_p$ for the norm of a function in L^p.

We will be interested in functions in the exponentially weighted space,

$$H^k_a = \left\{u : e^{ax}u(x) \in H^k\right\}$$

for $k = 0, 1, 2, \ldots$, and $a > 0$ with associated norm

$$\|u\|_{H^k_a} = \|e^{ax}u\|_{H^k}.$$

We also define the norm

$$\|f\|_{H^1 \cap H^1_a} \equiv \|f\|_{H^1} + \|f\|_{H^1_a}.$$

The exponential weight will *always* be a positive number; we will often omit the assumption $a > 0$ in statements.

Frequently, we will have an operator T defined on a weighted space, H^k_a, but wish to make computations in the unweighted space. To T we associate $T_a = e^{ax}Te^{-ax}$, an operator on H^k. For the differentiation operator, $\partial_x \mapsto D_a = \partial_x - a$.

2. Preliminaries.

2.1. Properties of the equation in a weighted space. Much of the analysis involves studying (1.2) in an exponentially weighted space. We therefore state the following extension of the well-posedness results obtained in [38] for H^k spaces.

THEOREM 2.1 (local existence in time and continuous dependence upon data). *Given $0 < a < 1$, let $\phi_0(x)$ satisfy*

$$\|\phi_0 - 1\|_{H^1 \cap H^1_a} \leq R < \infty, \tag{2.1}$$

$$\inf_x \phi_0(x) \geq \alpha_0 > 0, \tag{2.2}$$

$$\inf_x \phi_0(x)^m - a^2\phi_0(x)^n \geq \beta_0 > 0. \tag{2.3}$$

Then there exist $T_{\text{local}} > 0$ and $\phi(x,t) - 1 \in C^1([0, T_{\text{local}}), H^1 \cap H^1_a)$, a solution of (1.2) with data ϕ_0, satisfying

$$\|\phi(\cdot,t) - 1\|_{H^1 \cap H^1_a} \leq 2R, \tag{2.4}$$

$$\inf_x \phi(x,t) \geq \frac{1}{2}\alpha_0, \tag{2.5}$$

$$\inf_x \phi(x,t)^m - a^2\phi(x,t)^n \geq \frac{1}{2}\beta_0 \tag{2.6}$$

for $t < T_{\text{local}}$.

Moreover, there is a maximal time of existence $T_{\max}$*, such that if* $T_{\max} < \infty$*, then*

$$\lim_{t \to T_{\max}} \|\phi(\cdot,t) - 1\|_{H^1 \cap H^1_a} + \left\| \frac{1}{\phi(\cdot,t)} \right\|_\infty + \left\| \frac{1}{\phi(\cdot,t)^m - a^2 \phi(\cdot,t)^n} \right\|_\infty = \infty. \tag{2.7}$$

Remark 2.2. When $a = 0$, (2.3) is unnecessary; this case was treated in [38]. The importance of this condition for $a > 0$ will be discussed in section 2.3.

THEOREM 2.3. *Given* $0 < a < 1$*, let* $\phi^{(j)} - 1 \in C^1([0,T]; H^1 \cap H^1_a)$*,* $j = 1, 2$*, be two solutions of* (1.2) *such that*

$$\|\phi^{(j)}(\cdot,t) - 1\|_{H^1 \cap H^1_a} \le R < \infty,$$
$$\inf_x \phi^{(j)}(x,t) \ge \alpha_0 > 0,$$
$$\inf_x (\phi(x,t)^{(j)})^m - a^2 (\phi(x,t)^{(j)})^n \ge \beta_0 > 0.$$

There exists a constant $K = K(R, \alpha_0, \beta_0, a)$*, such that*

$$\|\phi^{(1)}(\cdot,t) - \phi^{(2)}(\cdot,t)\|_{H^1 \cap H^1_a} \le e^{Kt} \|\phi_0^{(1)} - \phi_0^{(2)}\|_{H^1 \cap H^1_a} \quad \text{for } t \le T.$$

Additionally, (1.2) possesses the conservation law

$$\mathcal{N}[\phi] = \begin{cases} \int \left(\frac{1}{2} \phi^{-2m} \phi_x^2 + \phi \log(\phi) - \phi + 1 \right) dx & \text{if } n + m = 1, \\ \int \left(\frac{1}{2} \phi^{-2m} \phi_x^2 + \phi - 1 - \log(\phi) \right) dx & \text{if } n + m = 2, \\ \int \left(\frac{1}{2} \phi^{-2m} \phi_x^2 + \frac{\phi^{2-n-m} - 1 + (n+m-2)(\phi - 1)}{(n+m-1)(n+m-2)} \right) dx & \text{for all other } n \text{ and } m. \end{cases} \tag{2.8}$$

$\mathcal{N}$ is well defined for ϕ bounded from below away from zero and $\|\phi - 1\|_{H^1} < \infty$. It is also locally convex about $\phi \equiv 1$. See section 3 of [38] for details.

Remark 2.4. The physical significance of $\mathcal{N}$ remains elusive. In the Hamiltonian case, $n + m = 0$, $\mathcal{N}$ revealed itself to be the generalized momentum of the equation [40].

2.2. Solitary waves and their analytic properties. Let us review the properties of the solitary waves associated with (1.2). In particular, we identify their decay and regularity properties and introduce the KdV scaling for later use.

Substituting the traveling wave ansatz, $\phi_c(x,t) = \phi_c(x - ct)$, into (1.2) with boundary conditions

$$\lim_{y \to \pm\infty} \phi_c(y) = 1, \quad \lim_{y \to \pm\infty} \partial_y^j \phi_c(y) = 0 \quad \text{for } j = 1, 2, \ldots \tag{2.9}$$

we have, after one integration,

$$-c(\phi_c - 1) + \phi_c^n - 1 + c\phi_c^n \partial_y \left(\phi_c^{-m} \partial_y \phi_c \right) = 0. \tag{2.10}$$

Letting $u_c = 1 - \phi_c$, u_c satisfies

$$-cu_c + (u_c + 1)^n - 1 + c(u_c + 1)^n \partial_y \left((u_c + 1)^{-m} \partial_y u_c \right) = 0. \tag{2.11}$$

Equation (2.10) may also be integrated up to a first order equation,

$$\frac{1}{2} \phi_c^{-2m} (\partial_x \phi_c)^2 - F_1(\phi_c; c) = 0, \tag{2.12}$$

after applying the boundary condition $\phi_c \to 1$ at $\pm\infty$. F_1 depends on the particular exponents:

$$F_1(x;c) = \begin{cases} \frac{x^{1-n}-1}{1-n} + \left(1-\frac{1}{c}\right)\frac{x^{-n}-1}{n} - \frac{1}{c}\log(x) & \text{if } m=1, \\ x-1-\left(1-\frac{1}{c}\right)\log(x) - \frac{1}{c}\frac{x^n-1}{c} & \text{if } n+m=1, \\ \log(x) + \left(1-\frac{1}{c}\right)\left(\frac{1}{x}-1\right) - \frac{1}{c}\frac{x^{n-1}-1}{n-1} & \text{if } n+m=2, \\ \frac{x^{2-n-m}-1}{2-n-m} - \left(1-\frac{1}{c}\right)\frac{x^{1-n-m}-1}{1-n-m} - \frac{1}{c}\frac{x^{1-m}-1}{1-m} & \text{otherwise.} \end{cases} \tag{2.13}$$

Using (2.12), an equivalent second order, self-adjoint equation for the solitary waves is

$$F_2(\phi_c;c) = -\partial_x^2\phi_c, \tag{2.14}$$

$$F_2(x;c) = x^{m-n}\left[-(x-1) + c^{-1}(x^n-1) - 2mx^{n-m-1}F_1(x;c)\right]. \tag{2.15}$$

Let us introduce the KdV scaling. Define

$$\boxed{\gamma = \sqrt{\frac{c-n}{c}}.} \tag{2.16}$$

Applying the scalings,

$$\xi = \gamma(x-ct), \quad u_c(y) = \frac{\gamma^2}{n-1}U(\xi(y);\gamma), \tag{2.17}$$

(2.11) becomes

$$-U + \frac{1}{2}U^2 + \partial_\xi^2 U = O(\gamma^2). \tag{2.18}$$

Remark 2.5. The parameter γ, (2.16), will be used throughout the paper. Because it uniquely maps $c \in (n,\infty)$ onto $(0,1)$, we will use c and γ interchangeably.

We summarize what is known about (2.11) and (2.18) in the following two results.

THEOREM 2.6. *For any $c > n > 1$, (2.11) has a unique positive, even solution u_c, going to zero at $\pm\infty$. In the KdV scaling, (2.18), U is real analytic in the arguments $(\xi,\gamma) \in \mathbb{R}\times[0,1)$. When $\gamma = 0$*

$$U(\xi;0) = U_\star(\xi) = 3\mathrm{sech}^2\left(\frac{1}{2}\xi\right).$$

Furthermore, for γ in any compact subset of $[0,1)$

$$\partial_\xi^j U(\xi;\gamma)e^{\pm\xi}(\mathrm{sign}(\xi))^j \to K_j(\gamma) \quad \text{as } \xi \to \pm\infty \text{ for } j = 0,1,2.$$

COROLLARY 2.7. *Given a compact interval $[0,\gamma_0] \subset [0,1)$, there exists a constant K such that for all $\gamma \in [0,\gamma_0]$,*

$$|\partial_\xi^j U(\xi;\gamma)| \le Ke^{-|\xi|} \quad \text{for } j = 0,1,2,\ -\infty < \xi < \infty, \tag{2.19}$$

$$|\partial_y^j u_c(y)| \le K\frac{\gamma^{2+j}}{n-1}e^{-\gamma|y|} \quad \text{for } j = 0,1,2,\ -\infty < y < \infty. \tag{2.20}$$

Proof. From [23], solitary waves exist, provided $c > n > 1$ and $m \in \mathbb{R}$. Writing the problem as a two-dimensional dynamical system, we may apply the stable manifold theorem about the hyperbolic point $(0,0)$ to deduce the exponential decay, as in [29, Theorem 2.1, Corollary 2.2]. ∎

Remark 2.8. When the parameter γ is small, ϕ_c is in the regime of small amplitude, long waves, where KdV appears as the leading order equation in a perturbation expansion of (1.2), as in [50].

COROLLARY 2.9. *For each $c > n > 1$, the solitary wave solution, $\phi_c - 1$, lies in $H^\infty(\mathbb{R})$. Furthermore, there exists $\sigma_0 > 0$ such that the solitary wave ϕ_c may be analytically continued off the real axis into the strip $\{z : |\Im z| < \sigma_0\}$.*

Proof. This is a consequence of Corollary 2.7 and [9, Corollary 4.1.6]; see Appendix A.1. ∎

COROLLARY 2.10. *Given a solitary wave ϕ_c, $c > n$, assume $0 < a < \gamma$. Then $\phi_c - 1 \in H^\infty_a$.*

COROLLARY 2.11. *Let $n > 1$.*

(a) *The mapping $c \mapsto \phi_c - 1$ is $C^1\left((n,\infty); H^2\right)$. In fact the mapping is analytic.*
(b) *The mapping $c \mapsto \phi_c - 1$ is analytic, and, for fixed x, $c \mapsto \phi_c(x)$ is an analytic function of c.*
(c) *Given $a < \frac{1}{2}$, the mapping is also $C^1\left(\left(n/(1-4a^2),\infty\right); H^2 \cap H^2_a\right)$.*

Proof. All parts are proved using the implicit function theorem, applied to the functional

$$\mathcal{F}[c,u] = \partial_x^2 u + F_2(1+u;c).$$

See Appendix A.2 for details and [6] for a statement and proof of the implicit function theorem for analytic mappings. ∎

2.3. Remarks and assumptions on the exponential weight. We see in Theorem 2.1 and Corollary 2.11 that the particular exponential weight restricts what data and which solitary waves will be permissible. For the solitary wave result, this restriction comes from the decay rate associated with the speed; see Corollary 2.7.

In the case of the existence theorem, (1.2) may be written as

$$\begin{aligned}\partial_t \phi &= -\left\{I - \partial_x\left[\phi^n \partial_x\left(\phi^{-m}\cdot\right)\right]\right\}^{-1} \partial_x\left(\phi^n\right)\\ &= -\phi^m\left\{\phi^{-m}\left[I - \partial_x\left(\phi^n\partial_x\left(\phi^{-m}\cdot\right)\right)\right]^{-1}\right\}\partial_x\left(\phi^n\right)\\ &= -\phi^m H_\phi^{-1}\partial_x(\phi^n).\end{aligned} \tag{2.21}$$

The operator H_ϕ is

$$H_\phi = \phi(x)^m - \partial_x\left(\phi(x)^n\partial_x\cdot\right). \tag{2.22}$$

This is a bounded operator on $L^2 \to H^1$, provided ϕ is continuous and bounded from below away from zero. However, the exponential weight introduces a second constraint. Consider solving $H_\phi u = f$, $f \in L^2_a$, for $u \in H^1_a$. Letting $g = e^{ax}f$ and $v = e^{ax}u$, this is equivalent to solving

$$\left[\phi^m - D_a\left(\phi^n D_a\right)\right] v = g, \quad v \in H^1.$$

Multiplying by v and integrating by parts,

$$\int \left(\phi^m - a^2\phi^n\right) v^2 + \phi^n\left(\partial_x v\right)^2 dx = \int gv dx.$$

A unique solution exists, provided a and ϕ satisfy $\inf_x \phi(x)^m - a^2\phi(x)^n > 0$; this is condition (2.3).

We invert these restrictions; given a solitary wave of speed c, we will assume that a is sufficiently small so that these, and other, properties hold. There are three restrictions in what follows.

Let

$$a_1 = \frac{1}{3}\gamma(c). \tag{2.23}$$

$\phi_c - 1$ will then be in $H^2_{a_1}$ and $H^2_{2a_1}$, as will all solitary waves of nearby speed. Let

$$a_2 = \frac{1}{2}\inf_x \phi_c(x)^{(m-n)/2}. \tag{2.24}$$

Then for an $a \leq \min\{a_1, a_2\}$, ϕ_c will satisfy (2.3), with

$$\inf_x \phi_c(x)^m - a^2\phi_c(x)^n \geq \frac{3}{4}\inf_x \phi_c(x)^m > 0.$$

Hence, the solitary waves will live in a set on which the existence theorem applies. Moreover, for all data ϕ_0 sufficiently close to ϕ_c in the H^1 norm, an analogous lower bound will exist.

Remark 2.12. $\inf_x \phi_c(x)^{(m-n)/2}$ is related to a physical length scale known as the *compaction length* [20]. This length, $\delta_{\text{comp.}}$, is given by

$$\delta_{\text{comp.}}(x) = \sqrt{\phi(x)^{n-m}}.$$

It measures the distances over which there will be geometrical rearrangement of the material, appearing macroscopically as changes in ϕ, in response to viscous stresses.

The stipulation $a < \inf_x \phi(x)^{(m-n)/2}$ may be interpreted as requiring the length scale associated with the exponential weight, a^{-1}, to never be smaller than this intrinsic, spatially varying length. However, we draw no physical conclusions from this observation.

Finally, there is a constraints related to the *essential spectrum* of a linear operator, discussed in section 3. Let

$$a_3 = \sqrt{\frac{2c}{n + 2c + \sqrt{n\,(n+8c)}}}\sqrt{\frac{c-n}{c}}. \tag{2.25}$$

This will ensure that for any $a \leq a_3$, the essential spectrum is located in a specific part of the complex plane. Let

$$a_\star(c) = \min\left\{a_1(c), a_2(c), \frac{1}{2}a_3(c)\right\}. \tag{2.26}$$

Then for any $a \leq a_\star(c)$, all of these properties will be satisfied for ϕ sufficiently close in H^1 to ϕ_c.

2.4. Ansatz and linearization. Given a perturbed solitary wave solution, ϕ, of (1.2), assume that there exists decomposition of ϕ into a (time-dependent) solitary wave of some speed $c(t)$ and phase $\theta(t)$ and a perturbation, v; this decomposition's

existence will be proved in section 5.1. Transforming our coordinate system into the frame of this modulating solitary wave,

$$y(x,t) = x - \int_0^t c(s)ds + \theta(t), \tag{2.27}$$

$$\phi(x,t) = \phi_{c(t)}(y(x,t)) + v(y(x,t),t) = \phi_c(y,t) + v(y,t). \tag{2.28}$$

The perturbation, v, is governed by

$$\partial_t v = A_c v - \dot\theta \partial_y v - \dot c \partial_c \phi_c - \dot\theta \partial_y \phi_c + \mathcal{F}_1[v;\phi_c], \tag{2.29}$$

where

$$A_c v = \phi_c^m H_{\phi_c}^{-1} \partial_y L_c v, \tag{2.30}$$

$$L_c v = -c\phi_c^n \partial_y^2 \left(\phi_c^{-m} v\right) + \left[c - n\phi_c^{-1} + cn(\phi_c^{-1} - 1)\right] v, \tag{2.31}$$

and (B.6) gives the definition of $\mathcal{F}_1[v;\phi_c]$, composed of terms nonlinear in v. We make two remarks about (2.29). First, the linear operator A_c is time dependent; $c = c(t)$, and we would prefer to work with a time-*independent* linear operator. Second, the appearance of the term $\dot\theta \partial_y v$ will prove problematic for studying the equation in H^1.

The first problem is addressed by adding and subtracting A_{c_0}, and considering the difference $A_c - A_{c_0}$ as another perturbation of the linear flow. To remove the $\partial_y v$ term, we introduce a renormalized time,

$$\tau = {c_0}^{-1} \left[\int_0^t c(s)ds - \theta(t)\right]. \tag{2.32}$$

The asymptotic stability proof of BBM required a similar transformation [21]. In addition, the problem will be considered in the weighted space H^1_a, with $w(y,t) = e^{ay}v(y,t)$. All together, we have the following.

PROPOSITION 2.13 (perturbation equation). *The perturbation to a solitary wave of speed c_0 associated with the ansatz in* (2.28)*, v, and its weighted perturbation $w = e^{ay}v$ evolves according to the equations in t- and τ-time, respectively.*

$$\partial_t v = A_c v - \dot\theta \partial_y v - \dot c \partial_c \phi_c - \dot\theta \partial_y \phi_c + \mathcal{F}_1[v;\phi_c], \tag{2.33}$$

$$\partial_t w = A_{c,a} w - \dot\theta \left(\partial_y - a\right) w - e^{ay}\left(\dot c \partial_c \phi_c + \dot\theta \partial_y \phi_c\right) + \mathcal{G}_1[w,v;\phi_c], \tag{2.34}$$

$$\partial_\tau v = A_{c_0} - \frac{c_0}{c - \dot\theta}\left(c\partial_c\phi_c + \dot\theta\partial_y\phi_c\right) + S[c_0, c, \dot\theta]v + \frac{c_0}{c-\dot\theta}\mathcal{F}_1[v;\phi_c], \tag{2.35}$$

$$\begin{aligned} \partial_\tau w &= A_{c_0,a} w - \frac{c_0}{c-\dot\theta} e^{ay}\left(\dot c\partial_c\phi_c + \dot\theta\partial_y\phi_c\right) + S_a[c_0,c,\dot\theta]w + \frac{c_0}{c-\dot\theta}\mathcal{G}_1[w,v;\phi_c] \\ &= A_{c_0,a} w + \mathcal{G}\left[w, v; c_0, c, \dot\theta\right]. \end{aligned} \tag{2.36}$$

The operator S and the terms $\mathcal{F}_1$ and $\mathcal{G}_1$ are given explicitly in Appendix B.

Proof. $\mathcal{G}_1$ is obtained from $\mathcal{F}_1$ by substituting $e^{-ay}w$ for one of the v's; $e^{ay}\mathcal{F}_1[v;\phi_c] = \mathcal{G}_1[e^{ay}v, v;\phi_c]$. The details appear in Appendix B. ∎

Note: From here on, we assume c_0 to be fixed and will suppress its appearance in the linear operators A_{c_0} and $A_{c_0,a}$.

3. Spectral properties of the linearized operator. In this section we analyze the spectrum of A and A_a. We will use c in place of c_0 and x in place of y as the independent variable.

$$AY = \lambda Y, \tag{3.1}$$

$$A = \left\{ I - \partial_x \left[\phi_c^n \partial_x \left(\phi_c^{-m} \cdot \right) \right] \right\}^{-1} \partial_x L_c, \tag{3.2}$$

$$L_c = -c\phi_c^n \partial_x^2 \left(\phi_c^{-m} \cdot \right) + \left[c - n\phi_c^{-1} + cn \left(\phi_c^{-1} - 1 \right) \right], \tag{3.3}$$

$$A_a = e^{ax} A e^{-ax}. \tag{3.4}$$

Our goal is to prove that the linearized problem is asymptotically stable: $\| e^{A_a t} w_0 \|_{H^1} \to 0$ as $t \to +\infty$. We will actually show something much stronger, that this convergence to zero happens exponentially fast. Our strategy is that of Pego and Weinstein [28, 29] and Miller and Weinstein [21]. We will

1. identify the essential spectrum of A_a by showing it to be a relatively compact perturbation of a constant coefficient operator;
2. rule out the point spectrum (eigenvalues of finite multiplicity) of A_a for $|\lambda|$ sufficiently large via an operator estimate;
3. use the *Evans function*, an infinite-dimensional analogue of the characteristic polynomial, to rule out nonzero point spectra of A_a in the set of "small" λ, which will be compact;
4. show decay in time of the C_0-semigroup $e^{A_a t}$ associated with A_a.

The spectral analysis is handled in this section, and the semigroup theory in the following section.

The principle result of this section is as follows.

THEOREM 3.1 (spectrum of linearized operator).

(a) *Let $a \in (0, a_\star(\gamma)]$. The essential spectrum of A_a denoted by $\sigma_{\text{ess}}(A_a)$ is a curve lying in the open left half-plane, with rightmost point $-\omega$,*

$$-\omega = \max\{ \Re z | z \in \sigma_{\text{ess}}(A_a) \} < 0. \tag{3.5}$$

(b) *There exist $\gamma_\star \in (0,1)$ and $\hat{\vartheta} > 0$ such that for each $\gamma \in (0, \gamma_\star]$ and $a \in (\gamma\hat{\vartheta}, a_\star(\gamma)]$, there exists $\varepsilon(\gamma, a) > 0$ such that the only eigenvalue of A_a with $\Re\lambda \geq -\varepsilon$ is $\lambda = 0$, and this is an eigenvalue of algebraic multiplicity two.*

(c) *In the Hamiltonian case, $n + m = 0$, part* (b) *may be extended for $\gamma \in (\gamma_\star, 1)$ to all but a discrete set with no accumulation point.*

The spectrum is pictured in Figure 3.

3.1. Essential spectrum. We make use of the definition of the essential spectrum of an operator from [33, 34], which states that for a closed, densely defined operator A on a Banach space X,

$$\sigma_{\text{ess}}(A) = \bigcap_{C \in \mathcal{K}(X)} \sigma(A + C), \tag{3.6}$$

where $\mathcal{K}(X)$ denotes the set of compact operators on X. Other definitions are possible and well known; see Chapter IX of [10] for a discussion of how these definitions relate to one another.

$\sigma(A) \setminus \sigma_{\text{ess}}(A)$ then consists of point spectra. This is so because, by Theorem 7.27 of [34], if λ is not in the essential spectrum, then $\lambda I - A$ is Fredholm with index zero.

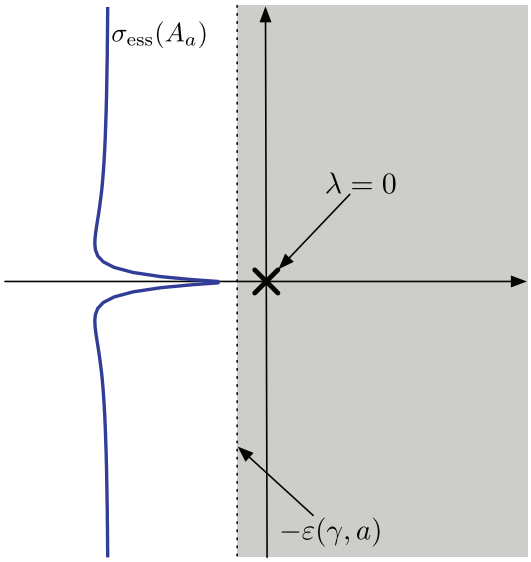


FIG. 3. *The spectrum of the operator A_a. The only eigenvalue with $\Re\lambda \geq -\varepsilon$ is $\lambda = 0$.*

Hence, it has a closed range and a finite kernel. Therefore, it must be an eigenvalue of finite multiplicity.

To prove Theorem 3.1 (a), we express A_a as a perturbation of a constant coefficient operator, A_a^∞, obtained by setting $\phi_c(x)$ equal to its asymptotic state, 1:

$$A_a^\infty = (I - D_a^2)^{-1} D_a(-cD_a^2 + c - n). \tag{3.7}$$

The difference between A_a and A_a^∞, given explicitly in (B.16), may be shown to be an A_a^∞-compact operator. Hence, A_a is a relatively compact perturbation of A_a^∞ and

$$\sigma_{\text{ess}}(A_a) = \sigma_{\text{ess}}(A_a^\infty).$$

Upon examination of the Fourier symbol of A_a^∞, the essential spectrum of A_a is

$$\boxed{\sigma_{\text{ess}}(A_a) = \left\{ \frac{(i\ell - a)(-c(i\ell - a)^2 + c - n)}{1 - (i\ell - a)^2}, \quad \ell \in \mathbb{R} \right\}} \tag{3.8}$$

and

$$-\omega = \max\{\Re z | z \in \sigma_{\text{ess}}(A_a)\} = -ac + \frac{an}{1 - a^2} < 0.$$

$\sigma_{\text{ess}}(A_a)$ lies in the open left half-plane if $0 < a < \gamma$. This set is pictured in Figure 3.

In addition, for $0 < a \leq a_\star < a_3$, a_3 defined by (2.25), the spectrum moves rightward as $a \to 0$. This is because a_3 is the value for which $-\omega$ is leftmost in $\mathbb{C}$, maximizing the rate of decay of e^{ay} as $y \to -\infty$.

3.2. Large eigenvalues. As in [21], we will study the eigenvalues of A_a by considering separately a large $|\lambda|$ regime and a small $|\lambda|$ regime.

Rewriting the linear operator A_a as

$$A_a = c\phi_c^m D_a\left(\phi_c^{-m}\cdot\right) - n\phi_c^m H_{\phi_c,a}^{-1} D_a\left(\phi_c^{-1}\cdot\right) + cm\phi_c^m H_{\phi_c,a}^{-1}\left(\phi_c^{-1}\partial_y\phi_c\cdot\right) + cn\phi_c^m H_{\phi_c,a}^{-1} D_a\left[\left(\phi_c^{-1}-1\right)\cdot\right], \tag{3.9}$$

we note that λ is an L^2 eigenvalue of A_a if and only if it is also an L^2 eigenvalue of $\tilde{A}_a = \phi_c^{-m} A_a \phi_c^m$, given by

$$\tilde{A}_a = cD_a - nH_{\phi_c,a}^{-1} D_a\left(\phi_c^{m-1}\cdot\right) + cmH_{\phi_c,a}^{-1}\left(\phi_c^{m-1}\partial_y\phi_c\cdot\right) + cnH_{\phi_c,a}^{-1} D_a\left[\phi_c^m\left(\phi_c^{-1}-1\right)\cdot\right]. \tag{3.10}$$

We will rule out eigenvalues of $\tilde{A}_a$, thus ruling them out for A_a. This is equivalent to studying A_a in a space weighted by $\phi_c^{-m}(x)$, a strictly positive, smooth, and bounded function. $\tilde{A}_a$ is also a relatively compact perturbation of $\tilde{A}_a^\infty = A_a^\infty$; they share the same essential spectrum.

Let the operator $C(\lambda)$ satisfy

$$C(\lambda) = \left(\lambda I - \tilde{A}_a^\infty\right)^{-1}\left(\tilde{A}_a - \tilde{A}_a^\infty\right).$$

PROPOSITION 3.2.

(a) *The operator $C(\lambda)$ is compact for λ not in σ_{ess}. In particular, $C(\lambda)$ is compact for all λ with $\Re\lambda > -\omega$.*

(b) *For any $\lambda \in \mathbb{C}\setminus\sigma_{\text{ess}}(A_a)$, we have that λ is an eigenvalue of A_a if and only if $1 \in \sigma(C(\lambda))$.*

(c) *Let $\lambda \in \mathbb{C}\setminus\sigma_{\text{ess}}(A_a)$. A sufficient condition for λ not to be an eigenvalue of A_a is that $\|C(\lambda)\| < 1$, with norm either L^2 or H^1, depending on which space is under consideration.*

Proof. Using the equivalence of eigenvalues of A_a and $\tilde{A}_a$, parts (b) and (c) will follow once (a) is established; see [21]. The Fourier symbol of $(\lambda I - \tilde{A}_a^\infty)^{-1}$ is

$$\frac{1-(\imath\ell-a)^2}{\lambda(1-(\imath\ell-a)^2) - (\imath\ell-a)(-c(\imath\ell-a)^2 + c - n)}.$$

This operator is bounded for λ not in the essential spectrum. The difference, given explicitly in (B.17), is a sum of Hilbert–Schmidt compact operators composed with bounded operators on $L^2 \to L^2$; hence $C(\lambda)$ is compact on this space.

For $C(\lambda)$ to be a compact operator on $H^1 \to H^1$, it will be sufficient to prove that

$$(I-\partial_x^2)^{1/2} C(\lambda)(I-\partial_x^2)^{-1/2} = (\lambda I - \tilde{A}_a^\infty)^{-1}(I-\partial_x^2)^{1/2}(\tilde{A}_a - \tilde{A}_a^\infty)(I-\partial_x^2)^{-1/2}$$

is compact on $L^2 \to L^2$. $(\lambda I - A_a^\infty)^{-1}$ is still bounded, and by commuting operators, it may be proven that $(I-\partial_x^2)^{1/2}(\tilde{A}_a - \tilde{A}_a^\infty)(I-\partial_x^2)^{-1/2}$ is compact. ∎

PROPOSITION 3.3. *Let $\delta \in (0,1)$ be fixed.*

(a) *Let $c_\star > n$ and*

$$\hat{\vartheta} \in \left(0, \sup_{c\in(n,c_\star]} a_\star(c)/\gamma(c)\right). \tag{3.11}$$

Then there exists $M = M(c_\star, \hat{\vartheta}) > 0$ *such that for* $c \in (n, c_\star]$ *and* $a \in (\gamma(c)\hat{\vartheta}, a_\star(c)]$, *if*

$$\Re\lambda \geq -\frac{1}{2}ac\gamma^2$$

and at least one of

$$|\Im\lambda| > cM\gamma^3,$$
$$\Re\lambda > cM\gamma^3$$

holds, then $\|C(\lambda)\|_{L^2\to L^2} < 1-\delta$.

(b) *This result also holds for* $\|C(\lambda)\|_{H^1\to H^1}$.

Proof. If $M \geq 1$, then by the assumptions on $\Re\lambda$ and $\Im\lambda$, λ is not in $\sigma_{\mathrm{ess}}(A_a)$; hence we may apply Proposition 3.2 (a) to conclude that $C(\lambda)$ is a compact operator. If the norm of $C(\lambda)$ is less than one, part (c) of that proposition will imply it is not an eigenvalue; we seek an $M \geq 1$ for which $\|C(\lambda)\|$ can be made sufficiently small.

Let $\gamma_0 = \gamma(c_\star)$. For all (γ, a) in $\{(\gamma, a) \mid \gamma \in [0, \gamma_0], a \leq a_\star(\gamma)\}$, there exist K_1 and K_2 such that

$$\begin{aligned}\|C(\lambda)\|_{L^2\to L^2} \leq{}& \|(\lambda I - A_a^\infty)^{-1} H_{1,a}^{-1} D_a\|_{L^2\to L^2}\left(cK_1\gamma^2\right)\\ &+ \|(\lambda I - A_a^\infty)^{-1} H_{1,a}^{-1}\|_{L^2\to L^2}\left(cK_2\gamma^3\right).\end{aligned}$$

This comes from expanding the difference $\tilde{A}_a - \tilde{A}_a^\infty$ and commutating operators; see (B.17). $(\lambda I - A_a^\infty)^{-1} H_{1,a}^{-1} D_a$ and $(\lambda I - A_a^\infty)^{-1} H_{1,a}^{-1}$ are constant coefficient operators and we will treat them in Fourier space.

Thus, if we can prove that, for λ satisfying the hypotheses,

$$\sup_{\ell\in\mathbb{R}}\left|\frac{(\imath\ell - a)\, cK_1\gamma^2}{\lambda\left[1-(\imath\ell-a)^2\right] - (\imath\ell - a)\left[-c(\imath\ell-a)^2 + c\gamma^2\right]}\right| < \frac{1-\delta}{2},$$

$$\sup_{\ell\in\mathbb{R}}\left|\frac{cK_2\gamma^3}{\lambda\left[1-(\imath\ell-a)^2\right] - (\imath\ell - a)\left[-c(\imath\ell-a)^2 + c\gamma^2\right]}\right| < \frac{1-\delta}{2},$$

we will be done.

Introducing the scalings $\lambda = c\Lambda\gamma^3$, $\ell = \gamma\xi$, and $a = \gamma\vartheta$, this is equivalent to identifying $M \geq 1$ such that when $\Re\Lambda \geq -\frac{1}{2}\vartheta$ and $\Re\Lambda > M$ or $|\Im\Lambda| > M$ then both

$$\sup_{\xi\in\mathbb{R}}\left|\frac{(\imath\xi - \vartheta)\, K_1}{\Lambda\left[1-\gamma^2(\imath\xi-\vartheta)^2\right] - (\imath\xi - \vartheta)\left[-(\imath\xi-\vartheta)^2 + 1\right]}\right| < \frac{1-\delta}{2}, \tag{3.12}$$

$$\sup_{\xi\in\mathbb{R}}\left|\frac{K_2}{\Lambda\left[1-\gamma^2(\imath\xi-\vartheta)^2\right] - (\imath\xi - \vartheta)\left[-(\imath\xi-\vartheta)^2 + 1\right]}\right| < \frac{1-\delta}{2} \tag{3.13}$$

are satisfied. By our assumption on a, $\vartheta \in (\hat{\vartheta}, 1)$.

Squaring both sides, (3.12) and (3.13) may be rewritten as two polynomial inequalities, $P_1(\Im\Lambda, \Re\Lambda, \xi) > 0$ and $P_2(\Im\Lambda, \Re\Lambda, \xi) > 0$, respectively. We will show that

for appropriately chosen Λ, the inequalities hold for all ξ. P_1 and P_2 are treated similarly. We study P_2:

$$\begin{aligned}
P_2(\Im\Lambda, \Re\Lambda, \xi) &= \alpha(\Im\Lambda)^2 + \beta\Im\Lambda + \eta_1(\Re\Lambda)^2 + \eta_2\Re\Lambda + \eta_3,\\
\alpha &= \gamma^2\xi^4 + 2\xi^2\gamma^2(1+\gamma^2\vartheta^4) + (1-\gamma^2\vartheta^2)^2,\\
\beta &= -2\xi^5\gamma^2 - 2\xi^3(1+\gamma^2(1+2\vartheta^2)) - 2\xi(1+(-3+\gamma^2)\vartheta^2+\gamma^2\vartheta^4),\\
\eta_1 &= \alpha,\\
\eta_2 &= 2\vartheta\xi^4\gamma^2 + 2\vartheta\xi^2(3+\gamma^2(-1+2\vartheta^2)) + 2\vartheta(1-\vartheta^2)(1-\gamma^2\vartheta^2),\\
\eta_3 &= \xi^6 + \vartheta^2(1-\vartheta^2)^2 + \xi^4(2+3\vartheta^2) + \xi^2(1+3\vartheta^4) - \varpi^2,\\
\varpi &= 2K_2/(1-\delta).
\end{aligned}$$

Using analysis similar to that for P_1 in [21], we first consider P_2 as quadratic in $\Im\Lambda$.[1] Examining its discriminant,

$$\begin{aligned}
\text{discriminant} = \gamma^4 &\left[-4\xi^8\left(\vartheta+\gamma^2\Re\Lambda\right)^2 + O(\xi^6)\right]\\
&+ \gamma^2\left[-24\vartheta^2\xi^6 + O(\xi^4)\right]\\
&- 36\xi^4\vartheta^3 + O(\xi^2).
\end{aligned}$$

If $0 \geq \Re\Lambda \geq -\vartheta/2$, then there exists $M_0 > 0$ such that for all $\Im\Lambda$, $\gamma \in [0,\gamma_0]$, $\vartheta \in (\hat{\vartheta}, 1)$, and $|\xi| > M_0$ the discriminant is negative and $P_2 > 0$. Furthermore, since the coefficient α is always positive, there exists $R_1 > 0$ such that $P_2 > 0$ when $\gamma \in [0,\gamma_0]$, $|\Im\Lambda| > R_1$, $0 \geq \Re\Lambda \geq -\vartheta/2$, and $|\xi| \leq M_0$.

For $\gamma \in [0,\gamma_0]$ and $\vartheta \in (\hat{\vartheta}, 1)$, both η_1 and η_2 are positive; if $P_2(\Im\Lambda, \Re\Lambda, \xi) > 0$, then $P_2(\Im\Lambda, \Re\Lambda + K, \xi) > 0$ for any $K > 0$. Therefore, $P_2 > 0$ for all ξ if $\gamma \in [0,\gamma_0]$, $|\Im\Lambda| > R_1$, and $\Re\Lambda \geq -\vartheta/2$. Also, $P_2 > 0$ for $|\xi| > M_0$ and all $\Im\Lambda$ if $\Re\Lambda \geq -\vartheta/2$.

We must still treat the case of $|\Im\Lambda| \leq R_1$ and $|\xi| \leq M_0$ simultaneously. Consider P_2 as quadratic in $\Re\Lambda$. η_1 and η_2 are positive for all ξ, and η_3 is bounded from below. Therefore there is some $R_2 > 0$ such that $P_2 > 0$ for all ξ if $\gamma \in [0,\gamma_0]$, $|\Im\Lambda| \leq R_1$, and $\Re\Lambda > R_2$. Thus, for any $\vartheta \in (\hat{\vartheta}, 1)$, $\gamma_0 \in [0,\gamma_0]$, there exist $R_1 > 0$ and $R_2 > 0$ such that $P_2 > 0$ for all ξ if

$$\begin{gathered}
\gamma \in [0,\gamma_0],\\
\Re\Lambda \geq -\vartheta/2,\\
|\Im\Lambda| \leq R_1 \quad \text{or} \quad \Re\Lambda > R_2.
\end{gathered}$$

For $C(\lambda) : H^1 \to H^1$, the proof is similar, with constants $\tilde{K}_1$ and $\tilde{K}_2$ in place of K_1 and K_2. □

[1] Though this proof largely follows that for the analogous proposition in [21], we have made two modifications. The first is to introduce $\hat{\vartheta} > 0$ in (3.11). Though this restriction may not be needed, by placing a lower bound on ϑ, estimates are more obviously uniform in γ and ϑ. A second change is to take $\Re\lambda \geq -\vartheta/2$ instead of $\Re\Lambda \geq -\vartheta/(4\gamma^2)$. Our condition is less sharp but, again, makes the intermediary bounds clearly uniform in γ and ϑ.

3.3. Small eigenvalues: The Evans function. In this section we rule out eigenvalues of A_a in the set $|\lambda| \leq M\gamma^3$. This is done using the *Evans function*, an analytic function that vanishes at eigenvalues of A_a. The Evans function, $D = D(\lambda;\gamma)$, is constructed for the eigenvalue problem using particular solutions of an associated dynamical system,

$$\dot{\mathbf{y}} = B(x,\lambda,\gamma)\mathbf{y}, \tag{3.14}$$

$$\mathbf{y} = O(e^{\mu_1 x}) \quad \text{as } x \to +\infty, \tag{3.15}$$

and the adjoint system,

$$\dot{\mathbf{z}} = -\mathbf{z}B(x,\lambda,\gamma), \tag{3.16}$$

$$\mathbf{z} = O(e^{\mu_1 x}) \quad \text{as } x \to -\infty. \tag{3.17}$$

μ_1 will be the eigenvalue of smallest real part of B^∞, the limit as $x \to \pm\infty$ of B. When certain conditions, described in Theorem 3.4, are met, the Evans function exists and may be explicitly defined as

$$D(\lambda;\gamma) = \mathbf{z}(x;\lambda,\gamma)\cdot\mathbf{y}(x;\lambda,\gamma). \tag{3.18}$$

The idea is to measure the angle between the subspace of solutions decaying at $+\infty$ with the subspace decaying at $-\infty$; hence the appearance of the dot product. The Evans function has an equivalent formulation in terms of the determinant of the fundamental solution of (3.14). For a more complete discussion of the Evans function, see [27].

THEOREM 3.4 (see [27, 29]). *Let Ω be a simply connected subset of $\mathbb{C}^2$. Suppose that the system* (3.14) *satisfies the following hypotheses:*

(i) *$B : \mathbb{R}\times\Omega \to \mathbb{C}^{n\times n}$ is continuous in (x,λ,γ) and analytic in (λ,γ) for fixed x.*

(ii) *$B^\infty(\lambda,\gamma) = \lim_{x\to\pm\infty} B(x,\lambda,\gamma)$ exists for all $(\lambda,\gamma)\in\Omega$. The limit is attained uniformly on compact subsets of Ω.*

(iii) *The integral*

$$\int_{-\infty}^{\infty} \|B(x,\lambda,\gamma) - B^\infty(\lambda,\gamma)\|dx$$

converges for all $(\lambda,\gamma)\in\Omega$ and the convergence is uniform on compact subsets of Ω.

(iv) *For every $(\lambda,\gamma)\in\Omega$, the matrix $B^\infty(\lambda,\gamma)$ has a unique eigenvalue of smallest real part, which is simple, denoted μ_1.*

Then $D(\lambda;\gamma)$ is well defined and analytic on Ω, such that $D(\lambda;\gamma) = 0$ if and only if 3.14 *has a solution $\mathbf{y}(x)$ satisfying* (3.15) *and*

$$\mathbf{y}(x) = o(e^{\mu_1 x}) \quad \text{as } x \to -\infty. \tag{3.19}$$

3.3.1. The KdV Evans function. In the case of the KdV equation, the eigenvalue problem may be scaled to

$$\partial_x L_{\mathrm{KdV}} Y = \partial_x\left(-\partial_x^2 Y + Y - 3\mathrm{sech}^2\left(\frac{1}{2}x\right)Y\right) = \Lambda Y. \tag{3.20}$$

Because the speed parameter has been scaled out, there is only one eigenvalue parameter, Λ.

Making the identification

$$\mathbf{y} = \left(Y, \quad \partial_x Y, \quad L_{\mathrm{KdV}} Y\right)^T, \tag{3.21}$$

we see that $\mathbf{y}$ satisfies the dynamical system

$$\dot{\mathbf{y}} = B_{\mathrm{KdV}}(x,\Lambda)\mathbf{y}, \tag{3.22}$$

$$B_{\mathrm{KdV}}(x,\Lambda) = \begin{pmatrix} 0 & 1 & 0 \\ 1 - 3\mathrm{sech}\left(\frac{1}{2}x\right)^2 & 0 & -1 \\ \Lambda & 0 & 0 \end{pmatrix}. \tag{3.23}$$

A complete description of the associated Evans may be found in [28]. We summarize as follows.

THEOREM 3.5 (the KdV Evans function).

(a) *The Evans function* $D_{\mathrm{KdV}}(\Lambda)$ *associated with* (3.20) *is given by*

$$D_{\mathrm{KdV}}(\Lambda) = \left(\frac{\mu_1(\Lambda)+1}{\mu_1(\Lambda)-1}\right)^2,$$

where $\mu_1(\Lambda)$ *denotes the root of* $\mu^3 - \mu + \Lambda = 0$ *of minimal real part.*

(b) *The domain of* $D_{\mathrm{KdV}}(\Lambda)$ *is the slit complex plane*

$$\Delta_{\mathrm{KdV}} = \mathbb{C} \setminus \left(-\infty, -\sqrt{\frac{4}{27}}\right].$$

(c) *The essential spectrum of* $A_{\mathrm{KdV}} : L^2_a \to L^2_a$ *is a curve contained entirely in the domain* $\{\lambda : \Re\lambda < -\epsilon\}$ *for some* $\epsilon > 0$. *Furthermore, if* $\Delta^+_{\mathrm{KdV}}(a)$ *denotes the component of* $\mathbb{C}\setminus\sigma_{\mathrm{ess}}(A_{\mathrm{KdV}})$ *that contains the right half-plane, then* $D_{\mathrm{KdV}}(\Lambda)$ *has no zeros in* $\Delta^+_{\mathrm{KdV}}(a)$ *except for a zero of multiplicity two at* $\Lambda = 0$.

3.3.2. The Evans function applied. The eigenvalue problem $AY = \lambda Y$,

$$\{I - \partial_x[\phi_c^n \partial_x(\phi_c^{-m}\cdot)]\}^{-1}\partial_x\{-c\phi_c^n\partial_x^2(\phi_c^{-m}\cdot) + [c - n\phi_c^{-1} + cn(\phi_c^{-1}-1)]\}Y = \lambda Y,$$

is equivalent to

$$\boxed{\partial_x L_c Y - \lambda\left[I - \partial_x\left(\phi_c^n\partial_x\left(\phi_c^{-m}\cdot\right)\right)\right]Y = 0.} \tag{3.24}$$

Defining

$$\mathbf{y} = \left(\phi_c^{-m}Y, \quad \partial_x\left(\phi_c^{-m}Y\right), \quad L_cY + \lambda\phi_c^n\partial_x\left(\phi_c^{-m}Y\right)\right)^T, \tag{3.25}$$

$\mathbf{y}$ solves the dynamical system

$$\dot{\mathbf{y}} = B(x,\lambda,c)\mathbf{y}, \tag{3.26}$$

$$B(x,\lambda,c) = \begin{pmatrix} 0 & 1 & 0 \\ c^{-1}\phi_c^{m-n}\left[c - n\phi_c^{-1} + cn\left(\phi_c^{-1}-1\right)\right] & \lambda/c & -c^{-1}\phi_c^{-n} \\ \lambda\phi_c^m & 0 & 0 \end{pmatrix}. \tag{3.27}$$

The matrix B may be decomposed as $B = B^\infty(\lambda, c) + R(x, \lambda, c)$.

(3.28)
$$B^\infty(\lambda, c) = \begin{pmatrix} 0 & 1 & 0 \\ \gamma^2 & \lambda/c & -c^{-1} \\ \lambda & 0 & 0 \end{pmatrix},$$

(3.29)
$$R(x, \lambda, \gamma) = \begin{pmatrix} 0 & 0 & 0 \\ c^{-1}\phi_c^{m-n}\left[c - n\phi_c^{-1} + cn\left(\phi_c^{-1} - 1\right)\right] - \gamma^2 & 0 & -c^{-1}\left(\phi_c^{-n} - 1\right) \\ \lambda\left(\phi_c^m - 1\right) & 0 & 0 \end{pmatrix}.$$

Also note that for the corresponding adjoint eigenvalue problem

(3.30)
$$\boxed{-L_c^\star \partial_x W - \lambda\left[I - \phi_c^{-m}\partial_x\left(\phi_c^n \partial_x \cdot\right)\right] W = 0}$$

under the identifications

(3.31)
$$\mathbf{z} = \left(-c\partial_x\left(\phi_c^n \partial_x W\right) - \lambda\phi_c^n \partial_x W, \quad c\phi_c^n \partial_x W, \quad W\right)$$

$\mathbf{z}$ solves

(3.32)
$$\dot{\mathbf{z}} = -\mathbf{z}B(x, \lambda, \gamma).$$

THEOREM 3.6 (properties of the Evans function).

(a) *The Evans function is defined and analytic on the set* $\Omega \subset \mathbb{C}^2$,

(3.33)
$$\boxed{\Omega = \{(\lambda, \gamma) \mid \gamma \in (0,1) \text{ and } \lambda \in \Omega_\gamma\}}$$

with

(3.34)
$$\boxed{\Omega_\gamma = \{\lambda \mid \Re\lambda > -\lambda_0\} \setminus \left(-\lambda_0, -\lambda_{\text{cut}}(\gamma)\right],}$$

where

$$\lambda_{\text{cut}} = \sqrt{\frac{1}{8}}\sqrt{8c^2 + 20cn - n^2 - 8c\sqrt{n^2 + 8cn} - n\sqrt{n^2 + 8cn}}$$

(3.35)
$$= \frac{2}{3\sqrt{3}} n\gamma^3 + \frac{8}{9\sqrt{3}} n\gamma^5 + O(\gamma^7)$$

and $\lambda_0 = n\sqrt{27/16}$.

(b) *Given* $(\lambda, \gamma) \in \Omega$ *and* $a \le a_\star(\gamma)$, *if* λ *is to the right of* $\sigma_{\text{ess}}(A_a)$, *then the following are equivalent:*
- $D(\lambda; \gamma) = 0$.
- λ *is an* L^2 *eigenvalue of* A_a.

(c) *For such zeros of* D, *the algebraic multiplicity of* λ *as an eigenvalue of* A_a *is equal to the order of* λ *as a zero of* $D(\lambda; \gamma)$.

(d) $D(0; \gamma) = \partial_\lambda D(0; \gamma) = 0$, *and hence it is an eigenvalue of algebraic multiplicity* at least *two.*

Remark 3.7. With this construction, for $\Re\lambda < 0$, $\lambda \in \Omega_\gamma$, the characteristic polynomial, (3.36), not only has a unique root of minimal real part, but all roots have distinct real part as well. This is stronger than is needed.

Remark 3.8. λ_{cut} is labeled as such because there is a branch cut in the Evans function there.

Remark 3.9. The eigenvalue at the origin is related to the solitary waves being a two-parameter (speed and phase) family of solutions of (1.2). The presence of the corresponding bound states, explicitly given in Proposition 3.23, could lead to a weak instability of algebraic growth in t. However, we will project out these modes by allowing our leading order solitary wave to modulate its speed and phase.

Proof of Theorem 3.6. Part (a) requires the verification of the hypotheses of Theorem 3.4 for this system. Applying the properties of the ϕ_c and Corollaries 2.9 and 2.11, and through examination of (3.27), B is clearly continuous in its three arguments for $\lambda \in \mathbb{C}$ and $c \geq n$. In addition, for fixed x, it will be analytic in (λ, c), or, equivalently, (λ, γ). Thus property (i) of Theorem 3.4 holds.

By Corollary 2.7, the limiting matrix B^∞ exists and $B - B^\infty$ is in L^1. This will be uniform on compact subsets of $\mathbb{C} \times [0, 1)$, establishing properties (ii) and (iii) of Theorem 3.4.

Lastly, we must verify the existence of μ_1, the unique eigenvalue of minimal real part. We divide this into two parts, $\Re\lambda \geq 0$ and $\Re\lambda < 0$. The characteristic polynomial, $\mathcal{P}(\mu)$, of B^∞ is

$$\boxed{c\mathcal{P}(\mu) = (\lambda - c\mu)(1 - \mu^2) + n\mu.} \tag{3.36}$$

Following the analysis in section 2(c) of [27] for a similar polynomial in the case of generalized BBM, one confirms that property (iv) holds for $\Re\lambda \geq 0$ and all γ, and hence the Evans function exists in $\{\Re\lambda \geq 0\} \times [0, 1)$.

Using the analysis in [21] for Theorem 2.7 of the polynomial, one can conclude the existence of some $\lambda_1 > 0$ and identify $\widetilde{\Omega}(\gamma)$, such that for all $\gamma \in (0, 1)$, a unique root of minimal real part exists for λ in the set

$$\{\lambda : \quad \Re\lambda < -\lambda_1\} \setminus (-\lambda_1, -\widetilde{\Omega}(\gamma)].$$

Alternatively, we give a more precise analysis of (3.36) in Appendix C that yields values of λ_0 and $\widetilde{\Omega}$ given in the proposition. This concludes the proof of part (a).

To prove part (b), we need a lemma regarding the location of $\sigma_{\text{ess}}(A_a)$.

LEMMA 3.10. *Let $\gamma \in (0, 1)$ and $a \leq a_\star(\gamma)$. Let $\Omega_+ = \Omega_+(\gamma, a)$ denote the component of $\mathbb{C} \setminus \sigma_{\text{ess}}(A_a)$ containing the origin. Then for $\lambda \in \Omega_+ \cap \Omega_\gamma$, the roots of the characteristic polynomial satisfy the relation*

$$\Re\mu_1 < -a < \Re\mu_{j\neq 1}. \tag{3.37}$$

Proof. By inspection, if $\lambda \in \sigma_{\text{ess}}(A_a)$, there is a root μ_j of (3.36) with $\Re\mu_j = -a$. Conversely, if there is a root with real part $-a$, then λ is in the essential spectrum. Hence the characteristic polynomial has a root with real part $-a$ if and only if $\lambda \in \sigma_{\text{ess}}(A_a)$

As noted in [27, section 2 (c)], for large $|\lambda|$, the roots of the characteristic polynomial (3.36) are

$$-1 + O\left(|\lambda|^{-1}\right), \quad 1 + O\left(|\lambda|^{-1}\right), \quad \lambda/c + O\left(|\lambda|^{-1}\right).$$

So for large λ in the right half-plane, (3.37) holds because $a < 1$. Now suppose for some $\lambda \in \Omega_+ \cap \Omega_\gamma$ the inequality were false. Because the $\Re\mu_j$ depend continuously on λ, equality would have to hold for some λ, but then it must be that $\lambda \in \sigma_{\rm ess}(A_a)$, which we have assumed is not the case. □

We now prove part (b) of Theorem 3.6. If $D(\lambda;\gamma) = 0$, then there is a solution to the ODE, $\dot{\mathbf{y}} = B\mathbf{y}$, such that

$$\mathbf{y}(x) = O(e^{\mu_1 x}) \quad \text{as } x \to +\infty \quad \text{and} \quad \mathbf{y}(x) = o(e^{\mu_1 x}) \quad \text{as } x \to -\infty.$$

Hence, by Part 1(d) of Proposition 1.6 of [27], for sufficiently small ϵ,

$$\mathbf{y}(x) = O(e^{\mu_\star x + \epsilon|x|}) \quad \text{as } x \to -\infty.$$

Letting $W(x) = e^{ax}\phi_c(x)^m y_1(x)$,

$$W(x) = O(e^{(\mu_\star + a)x + \epsilon|x|}) \quad \text{as } x \to -\infty \quad \text{and} \quad W(x) = O(e^{(\mu_1+a)x}) \quad \text{as } x \to +\infty.$$

From (3.37), $\mu_1 + a < 0 < \mu_\star + a$, so W will decay exponentially fast at $\pm\infty$. Hence it is an L^2 solution to the eigenvalue problem $A_a W = \lambda W$.

Conversely, if we have an L^2 eigenfunction, then it must satisfy

$$W(x) = O(e^{(\mu_1+a)x}) \quad \text{as } x \to +\infty \quad \text{and} \quad W(x) = o(e^{(\mu_1+a)x}) \quad \text{as } x \to -\infty.$$

$Y(x) = e^{-ax}W(x)$ will then satisfy (3.24) in a classical sense, although it may not be in L^2. However it will satisfy the necessary decay estimates on $\mathbf{y}$, constructed from Y as in (3.25), such that $D(\lambda;\gamma) = 0$. This concludes the proof of part (b).

The proof of part (c) follows that of Lemma 2.9 from [28]. First, it is proved that if, for a given γ, λ is a zero of order k of $D(\lambda;\gamma)$, then λ is an L^2 eigenvalue of A_a of algebraic multiplicity at least k. It is then shown that it cannot have algebraic multiplicity greater than k. We omit repeating these details. Part (d) is then a consequence of (c) and the calculations in Appendix D that $D(0;\gamma) = \partial_\lambda D(0;\gamma) = 0$ for all γ. □

Remark 3.11. For $\Re\lambda \le 0$ with $D(\lambda;\gamma) = 0$, it is likely, but not proved, that λ is not an L^2 eigenvalue of A.

3.3.3. The Evans function in the KdV scaling. We now introduce $D_\star(\Lambda;\gamma)$, the Evans function for (3.24) under the KdV scalings introduced in section 2.2:

$$\xi = \gamma x, \quad \lambda = c\Lambda\gamma^3, \quad \phi_c(x) = 1 + \frac{\gamma^2}{n-1}U(\xi(x);\gamma). \tag{3.38}$$

The eigenvalue problem is now

$$\begin{aligned}\partial_\xi L_\gamma Y = {}& \partial_\xi\left[-\left(1+\frac{\gamma^2}{n-1}U\right)^n \partial_\xi^2\left(\left(1+\frac{\gamma^2}{n-1}U\right)^{-m}\cdot\right)\right]Y \\ & + \partial_\xi\left[\left(1+\frac{\gamma^2}{n-1}U\right)^{-1}(1-U)\right]Y \\ & = \Lambda\left[I - \gamma^2\partial_\xi\left(\left(1+\frac{\gamma^2}{n-1}U\right)^n \partial_\xi\left(\left(1+\frac{\gamma^2}{n-1}U\right)^{-m}\cdot\right)\right)\right]Y.\end{aligned} \tag{3.39}$$

Recall that $U = U(\xi;\gamma)$ is the solution of (2.18), and for $\gamma = 0$, $U = U_\star$, the KdV soliton.

We can construct a dynamical system formulation of (3.39), defining the vector $\mathbf{Y}$ as

$$\mathbf{Y} = \begin{pmatrix} \left(1+\frac{\gamma^2}{n-1}U\right)^{-m} Y \\ \partial_\xi\left[\left(1+\frac{\gamma^2}{n-1}U\right)^{-m} Y\right] \\ L_\gamma Y + \gamma^2\Lambda\left(1+\frac{\gamma^2}{n-1}U\right)^{n}\partial_\xi\left[\left(1+\frac{\gamma^2}{n-1}U\right)^{-m} Y\right] \end{pmatrix}, \tag{3.40}$$

which satisfies

$$\dot{\mathbf{Y}} = B_\star(\xi,\Lambda,\gamma)\mathbf{Y}, \tag{3.41}$$

(3.42)

$$B_\star(\xi,\Lambda,\gamma) = \begin{pmatrix} 0 & 1 & 0 \\ \left(1+\frac{\gamma^2}{n-1}U(\xi;\gamma)\right)^{m-n-1}(1-U(\xi;\gamma)) & \gamma^2\Lambda & -\left(1+\frac{\gamma^2}{n-1}U(\xi;\gamma)\right)^{-n} \\ \Lambda\left(1+\frac{\gamma^2}{n-1}U(\xi;\gamma)\right)^{m} & 0 & 0 \end{pmatrix}.$$

As $\xi\to\infty$, the matrix is

$$B_\star^\infty(\Lambda,\gamma) = \begin{pmatrix} 0 & 1 & 0 \\ 1 & \gamma^2\Lambda & -1 \\ \Lambda & 0 & 0 \end{pmatrix}, \tag{3.43}$$

which has the characteristic polynomial

$$P_\star(\nu;\Lambda,\gamma) = \nu^3 - \gamma^2\Lambda\nu^2 - \nu + \Lambda. \tag{3.44}$$

A few remarks about the scaled problem. The assumptions stated in Theorem 3.4 remain the same, except now the matrix under inspection is $B_\star$, with eigenvalue parameters (Λ,γ). $\mathbf{Y}$ and $\mathbf{y}$ are related:

$$\mathbf{y}(x) = \begin{pmatrix} Y_1(\xi(x)) \\ \gamma Y_2(\xi(x)) \\ c\gamma^2 Y_3(\xi(x)) \end{pmatrix}.$$

At $\gamma = 0$, (3.39) is

$$\partial_\xi L_0 Y = \partial_\xi\left[-\partial_\xi^2 Y + (1-U_\star(\xi;0)Y\right] = \Lambda Y,$$

the KdV eigenvalue problem, (3.20), and

$$B_\star(\xi,\Lambda,0) = \begin{pmatrix} 0 & 1 & 0 \\ 1-U_\star(\xi) & 0 & -1 \\ \Lambda & 0 & 0 \end{pmatrix}$$

is the matrix for the KdV dynamical system.

PROPOSITION 3.12 (scaled Evans function).
(a) *$D_\star(\Lambda;\gamma)$ is defined and analytic on the set $\Delta \subset \mathbb{C}^2$,*

$$\Delta = \{(\Lambda,\gamma) \mid \gamma \in [0,1) \text{ and } \lambda \in \Delta_\gamma\}, \tag{3.45}$$

where, for $\gamma > 0$,

$$\begin{aligned}\Delta_\gamma &= \left\{\Lambda \mid \Re\Lambda > -\frac{\lambda_0}{c\gamma^3}\right\} \setminus \left(-\frac{\lambda_0}{c\gamma^3}, -\frac{\lambda_{\text{cut}}(\gamma)}{c\gamma^3}\right] \\ &= \left\{\Lambda \mid \Re\Lambda > -\gamma^{-3}\sqrt{\frac{27}{16}} + O(\gamma^{-1})\right\} \\ &\qquad \setminus \left(-\gamma^{-3}\sqrt{\frac{27}{16}} + O(\gamma^{-1}), -\sqrt{\frac{4}{27}} + O(\gamma^2)\right].\end{aligned}$$

λ_0 and $\widetilde{\Omega}$ are as defined by Theorem 3.6. When $\gamma = 0$,

$$\Delta_0 = \Delta_{\text{KdV}} = \mathbb{C} \setminus \left(-\infty, -\sqrt{4/27}\right].$$

(b) *For fixed $\gamma \in (0,1)$ and $\Lambda \in \Delta_\gamma$,*

$$D_\star(\Lambda;\gamma) = D(c\Lambda\gamma^3;\gamma).$$

(c) *For $\Lambda \in \Delta_0$,*

$$D_\star(\Lambda;0) = D_{\text{KdV}}(\Lambda).$$

Proof. The proof of these statements follows that of Proposition 2.8 in [21] and Theorems 4.9–4.11 of [29].

Proof of part (a). For $\gamma > 0$, as in Theorem 3.6, we must identify a set in $\mathbb{C}^2$ in which the hypotheses of Theorem 3.4 are valid. Parts (i)–(iii) are obvious as the solitary wave $U(\xi;\gamma)$ decays exponentially in ξ and, for fixed ξ, will be analytic in γ. We are left to verify part (iv). The characteristic polynomial of $B_\star^\infty$ is (3.44). As noted in [21], the roots of $P_\star$ are related to those of P, (3.36), by

$$\mu(\lambda,\gamma) = \gamma\nu(\Lambda,\gamma).$$

So $P_\star$ will have a unique root of minimal real part for a given Λ and γ when P has such a unique root for $\lambda = c\Lambda\gamma^3$. Therefore, for $\gamma \in (0,1)$, if Λ is in the set

$$\Delta_\gamma = \frac{1}{c\gamma^3}\Omega_\gamma,$$

(iv) will be satisfied. If $\gamma = 0$, (3.41) and (3.42) coincide with the KdV system, for which $\Delta_0 = \mathbb{C} \setminus (-\infty, -\sqrt{4/27}]$. Clearly, as $\gamma \to 0$, Δ_γ limits to Δ_0.

Proof of part (b). From part (a), $\lambda = c\Lambda\gamma^3 \in \Omega_\gamma$ and, by construction,

$$\begin{aligned} y_1(x,\lambda,\gamma) &\sim e^{\mu_1 x} \quad \text{as } x \to +\infty, \\ Y_1(\xi,\Lambda,\gamma) &\sim e^{\nu_1 \xi} \quad \text{as } \xi \to +\infty. \end{aligned}$$

At $\xi = \gamma x$, $\mu_1 = \gamma\nu_1$, $\lambda = c\Lambda\gamma^3$, $y_1(x,\lambda,\gamma) = Y_1(\xi,\Lambda,\gamma)$. Using the *transmission coefficient* interpretation of the Evans function, we then have

$$\mathbf{y}(x,\lambda,\gamma) \sim D(\lambda;\gamma)e^{\mu_1 x}\,(1, \quad \mu_1, \quad \lambda/\mu_1)^T \quad \text{as } x \to -\infty,$$
$$\mathbf{Y}(\xi,\Lambda,\gamma) \sim D_\star(\Lambda;\gamma)e^{\nu_1 \xi}\,(1, \quad \nu_1, \quad \Lambda/\nu_1)^T \quad \text{as } \xi \to -\infty,$$

implying

$$D(c\Lambda\gamma^3;\gamma) = D_\star(\Lambda;\gamma).$$

Proof of part (c). Trivially, when $\gamma = 0$, this is the KdV Evans function problem exactly. ∎

PROPOSITION 3.13. *Let* $\varepsilon_2 \in (0, \sqrt{4/27})$ *and* $M > 0$. *Set*

$$\mathcal{O} = \{\Re\Lambda \geq -\varepsilon_2, \quad |\Lambda| < M\}.$$

Then for all γ *sufficiently small,* $\mathcal{O} \subset \Delta_\gamma$ *and*

$$\lim_{\gamma\to 0} D_\star(\Lambda;\gamma) = D_\star(\Lambda;0) = D_{\mathrm{KdV}}(\Lambda)$$

with uniform convergence for $\Lambda \in \mathcal{O}$.

Proof. Clearly, for γ sufficiently close to zero, $\mathcal{O} \subset \Delta_\gamma$. Since $\mathcal{O}$ is compact and $D_\star$ is analytic in both arguments,

$$\sup_{\Lambda\in\mathcal{O}} |D_\star(\Lambda;\gamma) - D_\star(\Lambda;0)|$$

may be made arbitrary small by taking γ sufficiently close to zero. ∎

THEOREM 3.14. *There exist* $\gamma_\star \in (0,1)$ *and* $\hat{\vartheta} > 0$ *such that for all* $\gamma \in (0,\gamma_\star]$ *and* $a \in (\gamma\hat{\vartheta}, a_\star(\gamma)]$, *there exists* $\varepsilon = \varepsilon(\gamma,a) > 0$, *such that*

- *the only eigenvalue of* A_a *with* $\Re\lambda \geq -\varepsilon$ *is* $\lambda = 0$, *with algebraic multiplicity two;*
- *the only zero of* $D(\lambda;\gamma)$ *with* $\Re\lambda \geq -\varepsilon$ *is* $\lambda = 0$, *a root of order two.*

Proof. Adapting the proof of Theorem 2.1 in [21] we break a half-plane into two parts: a bounded set about the origin and its unbounded complement. The operator estimates from Proposition 3.3 will rule out eigenvalues in the unbounded part, and the Evans function in the KdV scaling will control eigenvalues in the bounded part.

Let us apply Proposition 3.3 (a) with $\delta = 1/4$, $c_\star = 4n$, and

$$\hat{\vartheta} = \frac{1}{4}\sup_{c\in(n,c_\star]} \frac{a_\star(c)}{\gamma(c)}.$$

Then there exists $M > 0$, such that for any $c \in (n,4n]$, $a \in (\gamma\hat{\vartheta}, a_\star(\gamma)]$, and $\lambda \in \Omega_{\mathrm{U}}$,

$$\Omega_{\mathrm{U}} = \left\{\lambda : \Re\lambda \geq -\frac{1}{2}ac\gamma^2, \quad |\lambda| > cM\gamma^3\right\},$$

$\|C(\lambda)\|_{L^2\to L^2} < 1$. By Proposition 3.2 (c), such λ cannot be L^2 eigenvalues of A_a.

Using this M, let $\varepsilon_2 = \frac{1}{4}$ and set

$$\mathcal{O} = \{\Lambda : \Re\Lambda \geq -\varepsilon_2, \quad |\Lambda| \leq M\}.$$

Define

$$m = \min\{|D_{\mathrm{KdV}}(\Lambda)| \min \Lambda \in \partial\mathcal{O}\}.$$

By Proposition 3.13, there exists a $\gamma_\star \leq 1/2$ such that for all $\gamma \in [0, \gamma_\star]$,

$$|D_\star(\Lambda;\gamma) - D_\star(\Lambda;0)| = |D_\star(\Lambda;\gamma) - D_{\mathrm{KdV}}(\Lambda)| < m \quad \text{for all } \Lambda \in \partial\mathcal{O}.$$

By Theorem 3.5, the only root of D_{KdV} in $\mathcal{O}$ is at $\Lambda = 0$, with multiplicity two. Applying Rouché's theorem (see, for example, [16, 17]), for all $\gamma \in [0, \gamma_\star]$, $D_\star(\cdot,\gamma)$ and D_{KdV} have the same number of roots in $\mathcal{O}$. By Proposition 3.12 (b), if $\gamma \in (0, \gamma_\star]$ and $\Lambda \in \mathcal{O}$, $D_\star(\Lambda;\gamma) = D(c\Lambda\gamma^3;\gamma)$; therefore on the set Ω_{B},

$$\Omega_{\mathrm{B}} = c\gamma^3\mathcal{O} = \left\{\lambda : \Re\Lambda \geq -\varepsilon_2 c\gamma^3, \quad |\lambda| \leq cM\gamma^3\right\}.$$

$D(\lambda;\gamma)$ also has only two zeros. Theorem 3.6 (d) asserts that these two roots are at the origin, corresponding to an eigenvalue of algebraic multiplicity two. By Theorem 3.6 (b), A_a then has no *nonzero* L^2 eigenvalues in Ω_{B}.

Combining Ω_{U} and Ω_{B}, if we set

$$\varepsilon(\gamma, a) = \min\left\{\frac{1}{2}ac\gamma^2, \quad \varepsilon_2 c\gamma^3, \quad \frac{1}{2}\omega, \quad \frac{1}{2}\lambda_{\mathrm{cut}}(\gamma), \quad \frac{1}{2}\lambda_0\right\}, \tag{3.46}$$

then by Theorem 3.6 (b), $\lambda = 0$ is both the only eigenvalue of A_a and the only root of $D(\lambda;\gamma)$ with $\Re\lambda \geq -\varepsilon$. □

COROLLARY 3.15. *For $\gamma \in (0, \gamma_\star]$, $\partial_c\mathcal{N}[\phi_c] \neq 0$.*

Proof. This is a consequence the preceding theorem and the relation $\partial_\lambda^2 D(0;\gamma) \propto \partial_c\mathcal{N}[\phi_c]$; see Appendix D. □

3.4. The Evans function beyond the KdV scaling. For $\gamma > \gamma_\star$, one may compute the Evans function numerically to assert that a given λ is not an eigenvalue. Moreover, the winding number of the image of the Evans function evaluated on the line $\Re\lambda = -\lambda_1 < 0$ equals the number of zeros in the set $\Re\lambda > -\lambda_1$. Therefore, one might evaluate the Evans function on such a line and exam the plot, as we do in Figures 4 and 5.

These plots indicate that $\lambda = 0$ is the only zero in the closed right half-plane for the values of c, n, and m under consideration. Up to the acceptance of these numerics, this extends Theorem 3.14. Note that we do not evaluate out to $-\lambda_1 + \imath\infty$, but merely compute at sufficiently large values of λ such that we are near the asymptotic value of the Evans function. It may be proven that there exists $D_\infty(\gamma)$ such that

$$\lim_{|\lambda|\to\infty, \lambda\in\Omega} D(\lambda;\gamma) = D_\infty(\gamma).$$

A further numerical computation will reveal that this value is nonzero.

In the Hamiltonian case, $n + m = 0$, an analytical result is possible for $\gamma > \gamma_\star$. The linearized operator, A, may be written as $A = J_c L_c$,

$$J_c = \left[I - \partial_x\left(\phi_c^n \partial_x\left(\phi_c^n \cdot\right)\right)\right]^{-1}\partial_x, \qquad J_c^\star = -J_c, \tag{3.47}$$

$$L_c = -c\phi_c^n\partial_x^2\left(\phi_c^n\right) + \left(c - n\phi_c^{-1} + cn\left(\phi_c^{-1} - 1\right)\right), \qquad L_c^\star = L_c. \tag{3.48}$$

This structure permits an extension of Theorem 3.14 beyond the KdV regime, given below in Theorem 3.21. However, this is absent for general n and m.

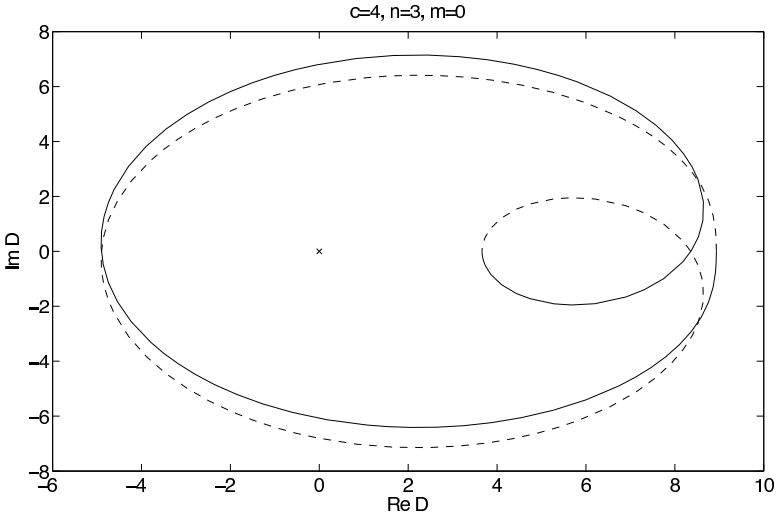


FIG. 4. $D(\cdot; c = 4)$ *evaluated on a portion of the strip* $\Re\lambda = -1/5$. *Since* $D(\overline{\lambda}) = \overline{D(\lambda)}$ *we only compute* $\Im\lambda > 0$, *and then reflect; this is the dashed curve. The curve wraps around the origin, marked by x, twice. In this case,* $n = 3$ *and* $m = 0$.

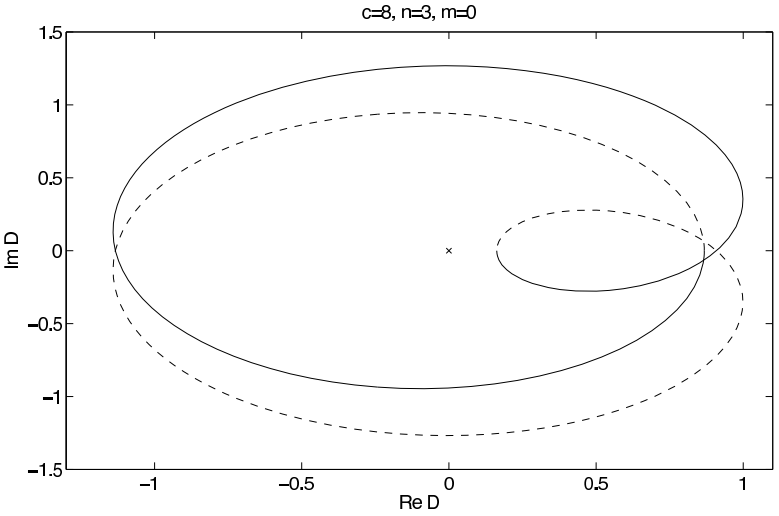


FIG. 5. $D(\cdot; c = 8)$ *evaluated on a portion of the strip* $\Re\lambda = -1$. *Since* $D(\overline{\lambda}) = \overline{D(\lambda)}$ *we only compute* $\Im\lambda > 0$, *and then reflect; this is the dashed curve. The curve wraps around the origin, marked by x, twice. In this case,* $n = 3$ *and* $m = 0$.

Remark 3.16. This section is the only place where the analyticity of the Evans function in the γ argument is used. In turn, this is the only place requiring analyticity of ϕ_c in c from Corollary 2.11 (b). For the results in the preceding section, *joint continuity* of $D(\lambda; \gamma)$ in its two arguments is sufficient.

LEMMA 3.17. *In the Hamiltonian case, the following are equivalent for* $\Re\lambda \geq 0$, $a \leq a_\star(\gamma)$*:*

- λ *is an* L^2 *eigenvalue of* A.
- $D(\lambda; \gamma) = 0$.

Proof. For $\Re\lambda > 0$, we know that $\mu_\star = \min\{\Re\mu_2, \Re\mu_3\} > 0$, so an eigenfunction, having submaximal growth at $-\infty$, must decay exponentially fast. This is very similar to the relation in the case of the weighted operator from Theorem 3.6 (b).

For $\Re\lambda = 0$, the proof relies on the JL structure of the operator A. See the proof of [27, Theorem 3.6]. ∎

LEMMA 3.18. *If λ is a nonzero purely imaginary eigenvalue, then $\partial_\lambda D(\lambda;\gamma) \neq 0$.*

Proof. The proof is by contradiction. Let $Y+$ be the corresponding eigenfunction, $AY^+ = \lambda Y^+$. It may be proven that the subspace $\mathcal{Y} = \operatorname{span}\{Y^+, \overline{Y^+}\}$ satisfies

$$\langle L_c u, v\rangle = 0 \quad \text{for all } u, v \in \mathcal{Y}.$$

As $\mathcal{Y} \cap \ker(L_c) = \{0\}$, we may apply Lemma 3.3 of [27] to conclude $\dim \mathcal{Y} \leq 1$, a contradiction. See Lemma 3.3 from [28] for details. ∎

LEMMA 3.19 (monotonicity of functional analytically confirmed for $n = 2$). *In the Hamiltonian case, assuming $\partial_c \mathcal{N}[\phi_c] \neq 0$ for all c, then for all c, there are no eigenvalues with $\Re\lambda > 0$.*

Proof. By Theorem 3.14, the result holds for $c \leq c_\star$. We argue by contradiction to extend it beyond $c_\star$. Assume for some $\gamma_0 > \gamma_\star$ that there exists λ_0, $\Re\lambda_0 > 0$, such that $D(\lambda_0;\gamma_0) = 0$. If $\Im\lambda_0 \neq 0$, then $D(\overline{\lambda_0};\gamma_0) = 0$ and there would be two eigenvalues in the right half-plane. But by Theorem 3.1 of [27], $A = J_c L_c$ has no more eigenvalues (counting multiplicity) with $\Re\lambda > 0$ than L_c does with $\Re\lambda < 0$. As is discussed in [40], L_c has exactly one negative eigenvalue; this is a contradiction, so λ_0 is real.

Because the number of zeros in the right half-plane is at most one, counting multiplicity, we know $\partial_\lambda D(\lambda_0;\gamma_0) \neq 0$. Applying the implicit function theorem, we get an analytic function, $\lambda(\gamma)$, defined in a neighborhood of γ_0, such that $\lambda(\gamma_0) = \lambda_0$ and $D(\lambda(\gamma);\gamma) = 0$.

Let $\mathcal{O}$ be the maximal domain of analyticity of $\lambda(\gamma)$. In a sufficiently small neighborhood of γ_0, $\Re\lambda(\gamma) > 0$. For real-valued γ in this neighborhood, we must have, by the above argument about complex-conjugates, that $\Im\lambda(\gamma) = 0$. Considering the power series expansion of $\lambda(\gamma)$, about γ_0, $\lambda(\gamma)$ will be real-valued for real-valued $\gamma \in \mathcal{O}$.

Let

$$\gamma_1 = \inf\{\gamma \in (\gamma_\star/2, \gamma_0) \cap \mathcal{O} \mid D(\lambda(\gamma);\gamma) = 0\}.$$

For all $\gamma \in [\gamma_1, \gamma_0]$, we must have $\lambda(\gamma) > 0$. Suppose this is not the case. Then, by continuity, for some γ, we must have $\lambda(\gamma) = 0$. But this would imply that $\lambda = 0$ was a root of multiplicity three, contradicting the assumption on $\mathcal{N}$, which ensures it is a root of multiplicity two.

$\lambda(\gamma)$ may be analytically continued down till at least $\gamma_\star/2$. If not, then $\gamma_1 > \gamma_\star/2$ and $\partial_\lambda D(\lambda(\gamma_1);\gamma_1) \neq 0$ since this root must be simple. Therefore we could apply the implicit function theorem again and extend $\lambda(\gamma)$ below γ_1, contradicting its minimality.

But then $D(\lambda(\gamma_\star);\gamma_\star) = 0$, and $\lambda(\gamma_\star) > 0$, contradicting Theorem 3.14. ∎

Remark 3.20. An analogous result may be found in Theorem 3.4 [27] for generalized KdV and generalized BBM. However, the argument there is very different because it may be proven that $D(\lambda) \to 1$ as $|\lambda| \to \infty$. This does not hold for the Evans function associated with (1.2), due to the appearance of a nonlinearity in the dispersive term.

THEOREM 3.21 (monotonicity of functional analytically confirmed for $n = 2$). *In the Hamiltonian case, assuming $\partial_c \mathcal{N}[\phi_c] \neq 0$ for all c, Theorem* 3.14 *may be extended to all $\gamma \in (\gamma_\star, 1)$, except for a discrete set whose only possible accumulation point is $\gamma = 1$.*

Proof. By Lemma 3.19, if Theorem 3.4 were false for some $\gamma > \gamma_\star$, it would be due to a zero appearing on the imaginary axis. We will prove by contradiction that the set

$$E = \{\gamma \in [0,1) \mid \text{there exists } \beta > 0 \text{ such that } D_\star(\imath\beta;\gamma) = 0\} \tag{3.49}$$

has no accumulation points. We consider only positive β's because if $\imath\beta$ is a root, then so is $-\imath\beta$. This follows the proofs of Theorem 3.6 of [28] and Theorem 2.1 of [21].

Assuming E has a limit point, there exists a sequence, $\gamma_j \in E$, $\gamma_j \to \gamma_0 \in E$ as $j \to \infty$. Taking a subsequence if necessary, γ_j and γ_0 are bounded away from $\gamma = 1$. We will now rule out large eigenvalues, and then argue by contradiction to rule out small eigenvalues.

Applying Proposition 3.3 to this range of γ values, there will exist $M > 0$, such that the corresponding $\beta_j > 0$ must satisfy $\beta_j \leq M$. Taking a subsequence if necessary, $\beta_j \to \beta_0$, $\beta_0 \leq M$, $D_\star(\imath\beta_j;\gamma_j) = 0$, and, by continuity, $D_\star(\imath\beta_0;\gamma_0) = 0$.

Note that $\beta_0 \neq 0$. $\Lambda = 0$ is always a root of order at least two. The assumption $\partial_c \mathcal{N}[\phi_c] \propto \partial_\lambda^2 D(0;\gamma) = (c^2\gamma^6)^{-1}\partial_\Lambda^2 D_\star(0;\gamma) \neq 0$ forces it to be a root of order exactly two. But $\beta_0 = 0$ would imply that it was a zero of at least four, a contradiction.

Applying Lemma 3.18, $\partial_\Lambda D_\star(\imath\beta_j;\gamma_j) \neq 0$ and $\partial_\Lambda D_\star(\imath\beta_0;\gamma_0) \neq 0$. By the implicit function theorem, there is an analytic function $\Lambda_0(\gamma)$ defined in a neighborhood of γ_0, such that $\Lambda_0(\gamma_0) = \imath\beta_0$ and $D_\star(\Lambda_0(\gamma);\gamma) = 0$. By considering the power series expansion of $\Lambda_0(\gamma)$ about γ_0, we see, by taking γ_j sufficiently close to γ_0, that $\Lambda_0(\gamma)$ is purely imaginary for real γ in its maximal domain of analyticity, $\mathcal{O}$.

Let

$$\gamma_1 = \inf\{\gamma \in [0,\gamma_0) \cap \mathcal{O} \mid D_\star(\Lambda_0(\gamma);\gamma) = 0\}.$$

For all $\gamma \in [\gamma_1, \gamma_0]$, we must have $\Im\Lambda_0(\gamma) > 0$. If not, then by continuity there would exist $\gamma \in (\gamma_1, \gamma_0)$, for which $\Lambda_0(\gamma) = 0$, yielding a contradiction as before.

Suppose $\gamma_1 > 0$. $\Im\Lambda_0(\gamma_1) > 0$ because, if not, then by continuity there would exist $\gamma \in (\gamma_1, \gamma_0)$, for which $\Lambda_0(\gamma) = 0$, leading to a contradiction again. Therefore, we may be sure that $\partial_\Lambda D_\star(\Lambda_0(\gamma_1);\gamma_1) \neq 0$. We may then apply the implicit function theorem, allowing us to continue Λ_0 below γ_1, another contradiction. Therefore $\gamma_1 = 0$. But then $\Im\Lambda_0(0) > 0$ and $0 = D_\star(\Lambda_0(0);0) = D_{\mathrm{KdV}}(\Lambda_0(0))$, a contradiction. ∎

Remark 3.22. This result is limited by our inability to analytically evaluate the functional $\mathcal{N}[\phi_c]$. The authors were similarly stymied in [40], where the orbital stability of the solitary waves relies on proving $\partial_c \mathcal{N}[\phi_c] > 0$. Here, as there, one may numerically evaluate the functional and observe its monotonicity in the speed argument. See [40] for the case $n = 2$.

This result, along with (3.8) and Theorem 3.14, completes the proof of Theorem 3.1.

3.5. The generalized kernel.

PROPOSITION 3.23. *Let $c > n$, $a \le a_\star(c)$ and assume $\partial_c \mathcal{N}[\phi_c] \neq 0$. Define*

$$\tilde{\xi}_1 = \partial_x \phi_c,$$

$$\tilde{\xi}_2 = \partial_c \phi_c,$$

$$\tilde{\eta}_1 = \Theta \left[I - \phi_c^{-m} \partial_x \left(\phi_c^n \partial_x \cdot \right) \right] \int_{-\infty}^{x} \left(L_c^\star \right)^{-1} \left[I - \phi_c^{-m} \partial_x \left(\phi_c^n \partial_x \cdot \right) \right] \int_{-\infty}^{x} \frac{\partial_x \phi_c}{\phi_c^{n+m}} dx,$$

$$\tilde{\eta}_2 = \Theta \left[I - \phi_c^{-m} \partial_x \left(\phi_c^n \partial_x \cdot \right) \right] \int_{-\infty}^{x} \frac{\partial_x \phi_c}{\phi_c^{n+m}} dx,$$

$$\Theta = \left(\partial_c \mathcal{N}[\phi_c] \right)^{-1}.$$

For $j = 1, 2$, set

$$\xi_j = e^{ay} \tilde{\xi}_j \quad \textit{and} \quad \eta_j = e^{-ay} \tilde{\eta}_j.$$

Then $\{\xi_1, \xi_2\}$ and $\{\eta_1, \eta_2\}$ are biorthogonal bases for $\ker_g(A_a)$ and $\ker_g(A_a^\star)$, $\langle \xi_i, \eta_j \rangle = \delta_{ij}$. They satisfy the relations

$$A_a \xi_1 = 0, \qquad A_a^\star \eta_2 = 0,$$

$$A_a \xi_2 = -\xi_1, \qquad A_a^\star \eta_1 = -\eta_2.$$

Proof. It is easy to verify $A\tilde{\xi}_1 = 0$, $A\tilde{\xi}_2 = -\xi_1$, and $A^\star \tilde{\eta}_2 = 0$.

$$\ker(L_c^\star) = \operatorname{span} \left\{ \phi_c^{-n-m} \partial_x \phi_c \right\}.$$

The kernel is orthogonal to $\tilde{\eta}_2$ because $\tilde{\eta}_2$ is an even function, while the kernel is odd. Therefore, $\tilde{\eta}_1$ is well defined and $A^\star \tilde{\eta}_1 = -\tilde{\eta}_2$. □

4. Semigroup decay. The following result proves the convective stability of solitary waves under the *linearized* flow. As we will do in section 6, this may be employed to prove full nonlinear stability.

THEOREM 4.1 (linearized stability). *Assume $\gamma \in (0, 1)$, $a \le a_\star(\gamma)$, and there exists $\varepsilon > 0$ such that $\lambda = 0$ is the only eigenvalue of A_a with $\Re\lambda \ge -\varepsilon$. Then the initial-value problem*

$$w_t = A_a w,$$

$$w(0) = w_0 \in H^1 \cap \ker{}_g(A_a^\star)^\perp$$

has a unique solution $w(t) = e^{A_a t} w_0 \in C_0([0, \infty); H^1)$ with

$$\|w(t)\|_{H^1} \le C e^{-bt} \|w_0\|_{H^1} \tag{4.1}$$

for some $C > 0$ and $b > 0$.

Remark 4.2. There exists a $b_{\max} > 0$ such that (4.1) will hold for all $b \in (0, b_{\max})$. In particular, for $\gamma \in (0, \gamma_\star]$ and $a \in (\gamma \hat{\vartheta}, a_\star(\gamma)]$, $b_{\max} \ge \varepsilon$; see Theorem 3.1.

Proof of Theorem 4.1. This is based on a result due to Prüss [30].

THEOREM 4.3. *Let B be the infinitesimal generator of a C_0-semigroup on a Hilbert space Z. Let $b > 0$. If there exists $M > 0$ such that*

$$\|(\lambda I - B)^{-1}\|_{Z \to Z} \le M \quad \textit{for all } \Re\lambda > -b,$$

then $\|e^{Bt}\|_{Z \to Z} \le e^{-bt}$.

Following the approach in [21] for Proposition 3.1, we first show that A_a is the infinitesimal generator of a C_0-semigroup on H^1. Examining the Fourier symbol of A_a^∞, A_a^∞ is such a semigroup. As (B.16) shows, $A_a - A_a^\infty$ is a bounded operator, so we may apply Theorem 3.1.1 of [25], which states that bounded perturbations of infinitesimal generators are also infinitesimal generators.

Consider the Hilbert space

$$Z = H^1 \cap \ker_{\mathrm{g}}(A_a^\star)^\perp$$

equipped with the H^1 norm and the operator

$$B = A_a|_Z$$

the restriction of A_a to Z. B inherits from A_a that it is the infinitesimal generator of a C_0-semigroup on Z. Also note that $\sigma(B) = \sigma(A_a) \setminus \{0\}$ by Theorem III-6.17 of [14].

Recall from Theorem 3.1 that $\sigma_{\mathrm{ess}}(A_a)$ is contained in left half-plane, and all points in $\sigma(A_a) \setminus \sigma_{\mathrm{ess}}(A_a)$ are eigenvalues of finite multiplicity. By the assumption on the spectrum of A_a, the spectrum of B is contained in the open left half-plane.

We will now prove there exists a uniform bound on resolvent of B for $\Re\lambda > -b$ for some $b > 0$. For $\lambda \in \rho(B)$,

$$(\lambda I - B)^{-1} = (\lambda I - A_a)^{-1}|_Z,$$

so

$$\|(\lambda I - B)^{-1}\|_{Z\to Z} \le \|(\lambda I - A_a)^{-1}\|_{H^1\to H^1}.$$

The resolvent of A_a may be written as

$$(\lambda I - A_a)^{-1} = (I - C(\lambda))^{-1}(\lambda I - A_a^\infty)^{-1},$$

with $C(\lambda)$ as defined in Proposition 3.2. Since $\Re\sigma_{\mathrm{ess}}(A_a) \le -\omega < 0$, we must have $b < \omega$. Then, for $\Re\lambda \ge -b$, the Fourier symbol of $(\lambda I - A_a^\infty)^{-1}$ is uniformly bounded, so

$$\|(\lambda I - A_a^\infty)^{-1}\|_{H^1\to H^1} \le M' \quad \text{for some } M' > 0.$$

By Proposition 3.3, for $\Re\lambda \ge -b \ge -\frac{1}{2}ac\gamma^2$, there exists $R' > 0$ such that

$$\|(I - C(\lambda))^{-1}\|_{H^1\to H^1} \le 2 \quad \text{for } |\lambda| > R'.$$

Therefore,

$$\|(\lambda I - B)^{-1}\|_{Z\to Z} \le \|(\lambda - A_a)^{-1}\|_{H^1\to H^1} \le 2M' \quad \text{for } \Re\lambda \ge -b \text{ and } |\lambda| > R'.$$

For $|\lambda| \le R'$ and $\Re\lambda \ge -\varepsilon$, B has no eigenvalues. B is a closed operator, and therefore $(\lambda I - B)^{-1}$ is holomorphic on this compact set (see Theorem III-6.7 of [14]) giving the bound

$$\|(\lambda I - B)^{-1}\|_{Z\to Z} \le M'' \quad \text{for some } M'' > 0.$$

Hence for all $\Re\lambda \ge -\min\{b, \varepsilon\}$,

$$\|(\lambda - B)^{-1}\|_{Z\to Z} \le \max\{2M', M''\},$$

and we may apply Theorem 4.3. ∎

5. Prelude to nonlinear stability. We need three results about our system before we can prove Theorem 1.1. First, we establish a criterion for when a decomposition of ϕ into a modulating solitary wave and a perturbation is possible. Then we derive equations for the evolution of the parameters associated with this modulating solitary wave. Finally, we relate the H^1 norm to the H^1_a norm of the perturbation.

5.1. Local existence of decomposition and continuation principles. In analyzing the weighted perturbation w, we wish to treat the nonlinear terms perturbatively, with the leading order behavior governed by the linear operator A_a. As noted in Proposition 3.23, the operator has a two-dimensional kernel. To prevent the appearance of secular terms, the perturbation must be orthogonal to $\ker_g(A_a^\star)$; this reveals how the decomposition of ϕ into a perturbation and a modulated solitary wave, (2.28), occurs. This follows the strategy of [26] and [21].

PROPOSITION 5.1. *Let $c_0 > n$, $a \le a_\star(c_0)$, and $t_1 > 0$. Given $\delta_1 > 0$, there exists $\delta_0 > 0$ such that for any $\phi - 1 \in C^1([0,t_1]; H^1 \cap H^1_a)$ satisfying*

$$\sup_{t\le t_1} \|e^{a(\cdot+\theta_0)}(\phi(\cdot,t) - \phi_{c_0}(\cdot - c_0 t + \theta_0))\|_{H^1} \le \delta_0 \quad \text{for some } \theta_0 \in \mathbb{R} \tag{5.1}$$

there exists a unique mapping $t \mapsto (\theta(t), c(t))$ in $C^1([0,t_1];\mathbb{R}^2)$ such that

$$\sup|\theta(t) - \theta_0| + \sup|\dot\theta(t)| + \sup|c(t) - c_0| + \sup|\dot c(t)| \le \delta_1, \quad t \le t_1, \tag{5.2}$$

$$\mathcal{T}_k[\phi - 1, \theta, c] = \left\langle \phi(x,t) - \phi_{c(t)}(y), \tilde\eta_j(y)\right\rangle = 0 \quad \text{for } j = 1,2 \text{ and } t \in [0,t_1], \tag{5.3}$$

where $y = x - \int_0^t c(s)ds + \theta(t)$.

The number δ_0 may be chosen as a decreasing function of t_1.

Proof. The proof is via the implicit function theorem. In this context, the Banach spaces are $C^1([0,t_1]; H^1_a)$ and $C^1([0,t_1];\mathbb{R}^2)$, the latter space equipped with the norm

$$\sup_{t\le t_1}|\theta(t)| + \sup_{t\le t_1}|\dot\theta(t)| + \sup_{t\le t_1}|c(t)| + \sup_{t\le t_1}|\dot c(t)|.$$

The functional is

$$\mathcal{T} = (\mathcal{T}_1, \mathcal{T}_2)^T : C^1([0,t_1]; H^1_a) \times C^1([0,t_1];\mathbb{R}^2) \to C^1([0,t_1];\mathbb{R}^2),$$

and it is C^1 in its arguments, permitting the use of the implicit function theorem.

Setting

$$\mathbf{U}_0 = (\phi_{c_0}(x - c_0 t) - 1, \quad 0, \quad c_0)$$

we see $\mathcal{T}[\mathbf{U}_0] = 0$. The Fréchet derivative at $\mathbf{U}_0$ is

$$\begin{aligned} &D\mathcal{T}[\mathbf{U}_0]\,(\delta\phi, \quad \delta\theta, \quad \delta c) \\ &\quad = \begin{pmatrix} \left\langle e^{a(x-c_0t)}\delta\phi(x,t), \eta_1(x-c_0t)\right\rangle - \int_0^1 \delta c(s)ds + \delta\theta(t) \\ \left\langle e^{a(x-c_0t)}\delta\phi(x,t), \eta_2(x-c_0t)\right\rangle + \delta c(t) \end{pmatrix}. \end{aligned} \tag{5.4}$$

The derivative acting on the (θ, c) is $\left(\begin{smallmatrix} I & -B \\ 0 & I \end{smallmatrix}\right)$, $(Bf)(t) = \int_0^t f(s)ds$. This block operator has a bounded inverse on $C^1([0,t_1];\mathbb{R}^2) \to C^1([0,t_1];\mathbb{R}^2)$. By the implicit function theorem, there exists $\delta_0 > 0$, such that if

$$\sup_{t\le t_1}\|\phi(\cdot - \theta_0, t) - \phi_{c_0}(\cdot - c_0 t)\|_{H^1_a} < \delta_0 \quad \text{for some } \theta_0 \in \mathbb{R},$$

then there exists a mapping in $C^1([0,t_1];\mathbb{R}^2)$, $t \mapsto (\tilde{\theta}(t), c(t))$, satisfying

$$\sup|\tilde{\theta}(t)| + \sup|\dot{\tilde{\theta}}(t)| + \sup|c(t)-c_0| + \sup|\dot{c}(t)| < \delta_1 \quad \text{for } t \le t_1,$$
$$\mathcal{T}[\phi(\cdot - \theta_0, t)) - 1, \tilde{\theta}(t), c(t)] = 0 \quad \text{for } t \le t_1.$$

Defining $\theta(t) = \tilde{\theta}(t) + \theta_0$, and applying the change of variables $x \mapsto \tilde{x} + \theta_0$, we obtain the form given in (5.1), (5.2), and (5.3).

Finally, because our norms are taken as the supremum over $t \le t_1$, if δ_0 works for t_1, then it will also work for any $t_2 \le t_1$. This allows δ_0 to be constructed as a decreasing function of t_1. ∎

PROPOSITION 5.2. *Let $c_0 > n$, $a \le a_\star(c_0)$ and assume $\phi - 1 \in C^1([0,T_{\max}]; H^1 \cap H^1_a)$, $T_{\max} > 0$ solves* (1.2).

Given $\delta_1 > 0$, there exist $\delta_0 > 0$ and $\delta_0' > 0$, such that for any $t_0 \in [0, T_{\max})$, if the decomposition of ϕ, $v(y,t) = \phi(x,t) - \phi_{c(t)}(y)$, $y = x - \int_0^t c(s)ds + \theta(t)$, and $(\theta, c) \in C^1([0,t_0];\mathbb{R}^2)$ satisfies

$$\sup_{t \le t_0} \|v(t)\|_{H^1_a} \le \delta_0/3, \tag{5.5}$$

$$\sup_{t \le t_0} |c(t) - c_0| \le \delta_0', \tag{5.6}$$

$$\mathcal{T}[\phi - 1, \theta, c](t) = 0 \quad \textit{for } t \in [0, t_0], \tag{5.7}$$

then there is a unique extension of (θ, c) in $C^1([0, t_0 + t_\star];\mathbb{R}^2)$ for some $t_\star > 0$ such that

$$\mathcal{T}[\phi - 1, \theta, c](t) = 0 \quad \textit{for } t \le t_0 + t_\star \le T_{\max}, \tag{5.8}$$

$$\begin{aligned} &\sup|\theta(t) - \theta(t_0)| + \sup|\dot{\theta}(t)| \\ &\quad + \sup|c(t) - c(t_0)| + \sup|\dot{c}(t)| \le \delta_1, \quad t \in [t_0, t_0 + t_\star]. \end{aligned} \tag{5.9}$$

Proof. This follows the proof of Proposition 5.2 in [28], although we are forced to modify it as we do not know a priori that a solution exists for all time.

Given δ_1, let $\delta_0 > 0$ be the value from Proposition 5.1 with $t_1 = \frac{1}{2}(T_{\max} - t_0)$. Set $\tilde{\phi}(x,t) = \phi(x, t + t_0)$, $\theta_1 = -\int_0^{t_0} c(s)ds + \theta(t_0)$. Let δ_0' be sufficiently small such that

$$\|\phi_{c'} - \phi_{c_0}\|_{H^1 \cap H^1_a} \le \delta_0/3 \quad \text{for all } c' \text{ such that} \quad |c' - c_0| \le \delta_0'. \tag{5.10}$$

Then, since

$$\begin{aligned} \left\| e^{a(\cdot+\theta_1)} \left(\tilde{\phi}(0) - \phi_{c_0}(\cdot + \theta_1) \right) \right\|_{H^1} &\le \left\| e^{a(\cdot+\theta_1)} \left(\tilde{\phi}(0) - \phi_{c_1}(\cdot + \theta_1) \right) \right\|_{H^1} \\ &\quad + \| e^{a(\cdot+\theta_1)} \left(\phi_{c_1}(\cdot + \theta_1) - \phi_{c_0}(\cdot + \theta_1) \right) \|_{H^1} \\ &= \|v(t_0)\|_{H^1_a} + \|\phi_{c_1} - \phi_{c_0}\|_{H^1_a} \le \frac{2}{3}\delta_0, \end{aligned}$$

we have

$$\begin{aligned} \left\| e^{a(\cdot+\theta_1)} \left(\tilde{\phi}(t) - \phi_{c_0}(\cdot - c_0 t + \theta_1) \right) \right\|_{H^1} &\le \left\| e^{a(\cdot+\theta_1)} \left(\tilde{\phi}(t) - \tilde{\phi}(0) \right) \right\|_{H^1} + \frac{2}{3}\delta_0 \\ &= e^{a\theta_1} \|\phi(t + t_0) - \phi(t_0)\|_{H^1 \cap H^1_a} + \frac{2}{3}\delta_0. \end{aligned}$$

As ϕ is continuous in time, there exists a $t_\star \le t_1$, such that

$$\sup_{t\le t_\star} \|e^{a(\cdot+\theta_1)}\left(\tilde\phi(t) - \phi_{c_0}(\cdot - c_0 t + \theta_1)\right)\|_{H^1} \le \delta_0.$$

Therefore, $\tilde\phi$ satisfies the hypotheses of Proposition 5.1. We have a unique $(\tilde\theta(t), \tilde c(t))$ with $(\tilde\theta(0), \tilde c(0)) = (\theta(t_0), c(t_0))$. This gives us the extension $(\theta(t), c(t)) = (\tilde\theta(t - t_0) - \theta_0 + \theta(t_0), \tilde c(t - t_0))$ for $t \in [t_0, t_0 + t_\star]$ and (5.9) will hold. ∎

5.2. Modulation equations. Given that the perturbation must be orthogonal to $\ker_{\mathrm{g}}(A_a^\star)$, the associated constraints give a pair of ODEs, coupled to the perturbation, giving a complete system of three equations for the three dependent variables.

Let P denote the projection onto $\ker_{\mathrm{g}}(A_a^\star)$. Assuming this space is two-dimensional, we use the biorthogonal bases given in Proposition 3.23 to define projection onto this space and its complement,

$$P = \langle \eta_1, \cdot\rangle\, \xi_1 + \langle \eta_2, \cdot\rangle\, \xi_2, \tag{5.11}$$

$$Q = I - P. \tag{5.12}$$

The secular terms are excised from the perturbation equation, (2.36), by requiring that

$$w_\tau = A_a w + Q\mathcal{G}, \tag{5.13}$$

$$P\mathcal{G} = 0, \tag{5.14}$$

$$Pw(\tau = 0) = 0. \tag{5.15}$$

Constraint (5.14) corresponds to

$$\langle \eta_j, \mathcal{G}\rangle = 0 \quad \text{for } j = 1, 2. \tag{5.16}$$

These two equations govern $c(t)$ and $\theta(t)$, completing our system.

Defining $p_1(y,t) = \partial_y \phi_{c(t)}(y) - \partial_y \phi_{c_0}(y)$ and $p_2(y,t) = \partial_c \phi_{c(t)}(y) - \partial_c \phi_{c_0}(y)$, the derived system is

(5.17)

$$\begin{pmatrix} 1 + \langle \tilde\eta_1, p_1\rangle & \langle \tilde\eta_1, p_2\rangle \\ \langle \tilde\eta_2, p_1\rangle & 1 + \langle \tilde\eta_2, p_2\rangle \end{pmatrix} \begin{pmatrix} \dot\theta \\ \dot c \end{pmatrix} = \frac{c - \dot\theta}{c_0} \begin{pmatrix} \left\langle \eta_1, S_a[c_0, c, \dot\theta] w \right\rangle \\ \left\langle \eta_2, S_a[c_0, c, \dot\theta] w \right\rangle \end{pmatrix} + \begin{pmatrix} \langle \eta_1, \mathcal{G}_1\rangle \\ \langle \eta_2, \mathcal{G}_1\rangle \end{pmatrix}.$$

However, the right-hand side still has $\dot\theta$ dependence. Observe that

(5.18)

$$\begin{aligned} \frac{c - \dot\theta}{c_0} S_a[c_0, c, \dot\theta] w = {} & cm \left[\partial_y \log\left(\frac{\phi_{c_0}}{\phi_c}\right)\right] w - \dot\theta m \partial_y \log(\phi_{c_0})\, w \\ & + \Big[\frac{c}{c_0} n \phi_{c_0}^m H_{\phi_{c_0},a}^{-1} D_a \left[\left(\phi_{c_0}^{-1} - c_0(\phi_{c_0}^{-1} - 1)\right)\cdot\right] \\ & \quad - n\phi_c^m H_{\phi_c,a}^{-1} D_a \left[\left(\phi_c^{-1} - c(\phi_c^{-1} - 1)\right)\cdot\right]\Big] w \\ & - \frac{\dot\theta}{c_0} \phi_{c_0}^m H_{\phi_{c_0},a}^{-1} D_a \left[\left(\phi_{c_0}^{-1} - c_0(\phi_{c_0}^{-1} - 1)\right)\cdot\right] w \\ & - cm \left[\phi_{c_0}^m H_{\phi_{c_0},a}^{-1} \left[\partial_y (\log \phi_{c_0})\,\cdot\right] - \phi_c^m \mathcal{H}_{\phi_c,a}^{-1} \left[\partial_y (\log \phi_c)\,\cdot\right]\right] w \\ & + \dot\theta m \phi_{c_0}^m H_{\phi_{c_0},a}^{-1} \left[\partial_y (\log \phi_{c_0})\,\cdot\right] w. \end{aligned}$$

Defining

$$\widetilde{S}_a = \widetilde{S}_a^1 + \widetilde{S}_a^2 + \widetilde{S}_a^3, \tag{5.19}$$

$$\widetilde{S}_a^1 = cm\left(\phi_{c_0}^{-1}\partial_y\phi_{c_0} - \phi_c^{-1}\partial_y\phi_c\right), \tag{5.20}$$

$$\begin{aligned}\widetilde{S}_a^2 = n\Big\{ &\frac{c}{c_0}\phi_{c_0}^m H_{\phi_{c_0},a}^{-1} D_a\left[\left(\phi_{c_0}^{-1} - c_0(\phi_{c_0}^{-1}-1)\right)\cdot\right] \\ &- \phi_c^m H_{\phi_c,a}^{-1} D_a\left[\left(\phi_c^{-1} - c(\phi_c^{-1}-1)\right)\cdot\right]\Big\},\end{aligned} \tag{5.21}$$

$$\widetilde{S}_a^3 = -cm\left[\phi_{c_0}^m H_{\phi_{c_0},a}^{-1}\left(\phi_{c_0}^{-1}\partial_y\phi_{c_0}\cdot\right) - \phi_c^m H_{\phi_c,a}^{-1}\left(\phi_c^{-1}\partial_y\phi_c\cdot\right)\right], \tag{5.22}$$

and

$$\begin{aligned}T_a = &-m\phi_{c_0}^{-1}\partial_y\phi_{c_0} - {c_0}^{-1}\phi_{c_0}^m H_{\phi_{c_0},a}^{-1} D_a\left[\left(\phi_{c_0}^{-1} - c_0(\phi_{c_0}^{-1}-1)\right)\cdot\right] \\ &+ m\phi_{c_0}^m H_{\phi_{c_0},a}^{-1}\left(\phi_{c_0}^{-1}\partial_y\phi_{c_0}\cdot\right),\end{aligned} \tag{5.23}$$

the right-hand side of (5.17) may be written as

$$\begin{pmatrix}\left\langle \eta_1, \widetilde{S}_a w\right\rangle \\ \left\langle \eta_2, \widetilde{S}_a w\right\rangle\end{pmatrix} + \dot\theta\begin{pmatrix}\langle \eta_1, T_a w\rangle \\ \langle \eta_2, T_a w\rangle\end{pmatrix} + \begin{pmatrix}\langle \eta_1, \mathcal{G}_1\rangle \\ \langle \eta_2, \mathcal{G}_1\rangle\end{pmatrix}.$$

Equation(5.17) may be solved algebraically so that $\dot\theta$ appears only on the left-hand side,

$$\begin{aligned}\mathcal{B}(t)\begin{pmatrix}\dot\theta \\ \dot c\end{pmatrix} &= \begin{pmatrix}1 + \langle\tilde\eta_1, p_1\rangle - \langle \eta_1, T_a w\rangle & \langle\tilde\eta_1, p_2\rangle \\ \langle\tilde\eta_2, p_1\rangle - \langle \eta_2, T_a w\rangle & 1 + \langle\tilde\eta_2, p_2\rangle\end{pmatrix}\begin{pmatrix}\dot\theta \\ \dot c\end{pmatrix} \\ &= \begin{pmatrix}\left\langle \eta_1, \widetilde{S}_a w\right\rangle \\ \left\langle \eta_2, \widetilde{S}_a w\right\rangle\end{pmatrix} + \begin{pmatrix}\langle \eta_1, \mathcal{G}_1\rangle \\ \langle \eta_2, \mathcal{G}_1\rangle\end{pmatrix},\end{aligned} \tag{5.24}$$

where $\mathcal{B}(t) = I + O(|c(t)-c_0|) + O(\|w\|_{L^2})$; for sufficiently small $|c(t)-c_0| + \|w\|_{L^2}$, $\mathcal{B}$ is invertible. Thus we have equations for $\dot\theta$ and $\dot c$, closing the system for (v, c, θ).

Remark 5.3. In (5.24), we see that when $\mathcal{B}(t)$ is inverted, the right-hand side of the system is continuous in t. Therefore, provided $c(t)$, $\theta(t)$, and $w(t)$ are continuous in t, $c(t)$ and $\theta(t)$ will actually be C^1.

5.3. Lyapunov bound. Using the functional $\mathcal{N}[\phi]$, defined in (2.8), we have the following.

PROPOSITION 5.4. *Let $c_0 > n$, $a \le a_\star(c_0)$, and let $\phi(x,t)$ be a solution to* (1.2) *in $C^1([0,T]; H^1 \cap H_a^1)$ with data*

$$\phi_0 = \phi_{c_0}(x+\theta_0) + v_0(x), \quad \theta_0 \in \mathbb{R}.$$

Assume the decomposition $\phi(x,t) \to (v(y,t), \theta(t), c(t))$ exists for $t \le T$ and

$$|c(t)-c_0| + \|v(\cdot,t)\|_{H^1} \le \delta_1 < 1 \quad \text{for } t \le T,$$

$$\partial_c \mathcal{N}[\phi_c]\Big|_{c=c_0} \neq 0.$$

Then there exist constants K and K' such that

$$\begin{aligned}\|v\|_{H^1}^2\left(1-K'\|v\|_{H^1}\right) \le K\Big(&|\Delta\mathcal{N}| + \|w\|_{L^2} + \|w\|_{L^2}^2 \\ &+ |c(t)-c_0| + |c(t)-c_0|^2 + |c(t)-c_0|^3\Big),\end{aligned} \tag{5.25}$$

$$\Delta\mathcal{N} = \mathcal{N}[\phi_{c(t)}+v] - \mathcal{N}[\phi_{c_0}].$$

Remark 5.5. This proposition, which relates the unweighted norm of the perturbation to the weighted norm, is where the invariance of $\mathcal{N}$ is particularly useful. In [29], the Boussinesq equations lacked such an invariant; thus, only linear stability was proven. Strictly speaking, it would be sufficient for $\Delta\mathcal{N}$ to be bounded in terms of the data; the conservation of $\mathcal{N}$ is not essential, though it is helpful.

Proof. Taylor expanding $\mathcal{N}$ about a solitary wave with perturbation z, we obtain

$$\mathcal{N}[\phi_{c_0}+z] = \mathcal{N}[\phi_{c_0}] + \langle \delta\mathcal{N}[\phi_{c_0}], z\rangle + \frac{1}{2}\left\langle \delta^2\mathcal{N}[\phi_{c_0}]z, z\right\rangle + O(\|z\|_{H^1}^3).$$

From [38], the first and second variations are

$$\begin{aligned}\langle \delta\mathcal{N}[\phi_{c_0}], z\rangle &= \int\left(\frac{1-\phi_{c_0}^{1-n-m}}{n+m-1} + m\phi_{c_0}^{-2m-1}\left(\partial_x\phi_{c_0}\right)^2 - \partial_x^2\phi_{c_0}^{-2m}\phi_{c_0}\right) z\,dx \\ &= \Theta^{-1}\left\langle\tilde{\eta}_2, z\right\rangle,\end{aligned} \tag{5.26}$$

$$\begin{aligned}\left\langle \delta^2\mathcal{N}[\phi_{c_0}]z, z\right\rangle = &\int\left(\phi_{c_0}^{-n-m} - m(1+2m)\phi_{c_0}^{-2m-2}\left(\partial_x\phi_{c_0}\right)^2\right) z^2\,dx \\ &+ \int\left(2m\phi_{c_0}^{-2m-1}\partial_x^2\phi_{c_0}\right) z^2\,dx + \int\phi_{c_0}^{-2m}\left(\partial_x z\right)^2 dx,\end{aligned} \tag{5.27}$$

where $\tilde{\eta}_2$ and Θ are as defined in Proposition 3.23.

Take $z(y,t) = \phi_{c(t)}(y) - \phi_{c_0}(y) + v(y,t) = \phi(x,t) - \phi_{c_0}(y)$. Then

$$\begin{aligned}\langle \delta\mathcal{N}[\phi_{c_0}], z\rangle &= \left\langle \Theta^{-1}\tilde{\eta}_2, z(y,t)\right\rangle \\ &= \Theta^{-1}\left\langle\tilde{\eta}_2, \phi_{c(t)}(y) - \phi_{c_0}(y)\right\rangle + \Theta^{-1}\left\langle\tilde{\eta}_2, v(y,t)\right\rangle.\end{aligned} \tag{5.28}$$

Using the continuity of $c \mapsto \phi_c - 1$,

$$\Theta^{-1}\left\langle\tilde{\eta}_2, \phi_{c(t)}(y) - \phi_{c_0}(y)\right\rangle \le K|c(t)-c_0|. \tag{5.29}$$

The term with the perturbation, v, may be bounded by

$$\Theta^{-1}\left\langle\tilde{\eta}_2, v(y,t)\right\rangle = \Theta^{-1}\left\langle\eta_2, w(y,t)\right\rangle \le \Theta^{-1}\|\eta_2\|_{L^2}\|w\|_{L^2} \le K\|w\|_{L^2}. \tag{5.30}$$

Now we bound the second variation. For brevity, let

$$\Phi_{c_0} = -m(1+2m)\phi_{c_0}^{-2m-2}\left(\partial_x\phi_{c_0}\right)^2 + 2m\phi_{c_0}^{-2m-1}\partial_x^2\phi_{c_0}.$$

Then

$$\begin{aligned}\left\langle \delta^2\mathcal{N}[\phi_{c_0}]z, z\right\rangle &= \int\phi_{c_0}^{-2m}\left(\partial_x z\right)^2 + \phi_{c_0}^{-n-m}z^2 + \int\Phi_{c_0}z^2 \\ &\ge K\|z\|_{H^1}^2 + \left\langle\Phi_{c_0}, \left(\phi_{c(t)}-\phi_{c_0}\right)^2\right\rangle \\ &\quad + 2\left\langle\Phi_{c_0}, \left(\phi_{c(t)}-\phi_{c_0}\right)v\right\rangle + \left\langle\Phi_{c_0}, v^2\right\rangle \\ &\ge K_1\|v\|_{H^1}^2 - K_2|c(t)-c_0|^2 \\ &\quad - K_3|c(t)-c_0| - K_4\left\langle\Phi_{c_0}e^{-ay}, w^2\right\rangle.\end{aligned} \tag{5.31}$$

We would like $\Phi_{c_0} e^{-2ay} \in L^\infty$, so that the last term may be estimated by $\|w\|_{L^2}^2$. Since $a \leq a_\star(\gamma) < \frac{1}{2}\gamma$, we may do this.

Lastly, we have the remainder term $\mathcal{R}[\phi_{c_0}, z]$. Because of the a priori bound on $\|v\|_{H^1}$, this may be estimated as

$$|\mathcal{R}[\phi_{c_0}, z]| \leq K\|z\|_{H^1}^3 \leq K\left(|c(t)-c_0|^3 + \|v\|_{H^1}^3\right). \tag{5.32}$$

Combining these estimates, (5.29), (5.30), (5.31), and (5.32), gives us

$$\begin{aligned}\|v\|_{H^1}^2(1 - D\|v\|_{H^1}) \leq K&\left(|\Delta\mathcal{N}| + |c(t)-c_0| + |c(t)-c_0|^2 + |c(t)-c_0|^3\right)\\ &+ K\left(\|w\|_{L^2} + \|w\|_{L^2}^2\right). \qquad \square\end{aligned}$$

6. Proof of main results. Before proving Theorem 1.1, we make an a priori estimate.

6.1. A priori estimates.

PROPOSITION 6.1. *Let $c_0 > n$, $a \leq a_\star(c_0)$, and assume there exists $\varepsilon > 0$ such that*

$$\sigma(A_a) \cap \{\Re\lambda \geq -\varepsilon\} = \{0\}, \quad \lambda = 0 \text{ is an eigenvalue of algebraic multiplicity two.}$$

Let $T > 0$. There exist $\delta_\star \in (0,1)$ and $K_\star \geq 1$ such that if the $\phi(x,t) - 1 \in C^1\left([0,T]; H^1 \cap H^1_a\right)$ solves (1.2) and satisfies, for $t \leq T$,

$$\inf_x \phi(x,t) \geq \alpha_0 > 0, \tag{6.1}$$

$$\inf_x \phi(x,t)^m - a^2\phi(x,t)^n \geq \beta_0 > 0, \tag{6.2}$$

and furthermore

(i) *the decomposition $\phi(x,t) \mapsto (v(y,t), c(t), \theta(t))$ exists for $t \leq T$;*

(ii) *for $t \leq T$*

$$\begin{aligned}&\sqrt{|\Delta\mathcal{N}|} + \sqrt{\|w(t)\|_{H^1}} + \sqrt{|c(t)-c_0|}\\ &\quad + |\theta(t)-\theta_0| + \left|1 - \frac{c_0}{c(t)-\dot\theta(t)}\right| + \|v(t)\|_{H^1} \leq \delta_\star;\end{aligned} \tag{6.3}$$

(iii) *the data satisfies*

$$\sqrt{|c(0)-c_0|} + |\theta(0)-\theta_0| + \sqrt{|\Delta\mathcal{N}|} + \sqrt{\|w(0)\|_{H^1}} \leq \epsilon < \delta_\star; \tag{6.4}$$

then for $t \in [0,T]$,

$$\sqrt{e^{\kappa t}\|w(t)\|_{H^1}} + \sqrt{|c(t)-c_0|} + |\theta(t)-\theta_0| + \left|1 - \frac{c_0}{c(t)-\dot\theta(t)}\right| + \|v(t)\|_{H^1} \leq K_\star\epsilon, \tag{6.5}$$

with $\kappa = \kappa(\alpha_0, \beta_0, \delta_\star) \in (0, b_{\max})$.

Proof. The strategy for proving this proposition is to show that the left-hand side of (6.5) may be estimated in terms of their data, (6.4), using (6.3). This largely follows the proof in [21], with a few changes. The need to estimate $|1 - c_0/(c-\dot\theta)|$ in terms of the data will require use of the modulation equations, (5.24), to control $\dot\theta$, and to control $\|w\|_{H^1}$, we will need to work in τ-time. Thus we make the following estimates.

Temporal change of variables. First, let us assume that $\delta_\star \leq \frac{1}{2}$. Then, since $c_0 > n > 1$, $\frac{1}{2}c_0 \leq c(t) \leq \frac{3}{2}c_0$. Furthermore, this initial choice of $\delta_\star$ ensures

$$\frac{1}{2} \leq \frac{c_0}{c(t) - \dot{\theta}(t)} = \frac{d\tau}{dt} \leq \frac{3}{2},$$

so the change of variables $\tau = \tau(t)$ is well defined.

Time derivatives of modulation parameters. Examining (5.24), $\mathcal{B}(t) = I + O(|c(t) - c_0|) + O(\|w\|_{L^2})$; so there exists $\delta_{\mathcal{B}} > 0$ such that for $\delta_\star \leq \delta_{\mathcal{B}}$, $\mathcal{B}(t)$ will be invertible. Therefore

(6.6)
$$|\dot{\theta}(t)| + |\dot{c}(t)| \leq |\mathcal{B}(t)^{-1}| \left(\|\widetilde{S}_a w\|_{L^2} + \|\mathcal{G}_1\|_{L^2} \right)_{L^2} \leq K_1 \left(|c - c_0| + \|v\|_{H^1} \right) \|w\|_{L^2},$$

permitting the estimate

$$|\dot{\theta}(t)| + |\dot{c}(t)| \leq K_1 \delta_\star^3. \tag{6.7}$$

As terms of the form $1/(c - \dot{\theta})$ will appear, we will assume that

$$\delta_\star \leq \left(\frac{1}{2K_1} \right)^{1/3} = \delta_\theta,$$

so $|\dot{\theta}| \leq \frac{1}{2}$, and this quotient will be well defined and bounded.

Weighted perturbation. In τ-time, $\|w(\tau)\|_{H^1}$ is

$$w(\tau) = e^{A_a \tau} w(\tau(t = 0)) + \int_{\tau(0)}^{\tau} e^{A_a(\tau - s)} Q\mathcal{G}(s) ds.$$

By the semigroup decay estimate of Theorem 4.1, there exist $K_2 > 0$ and $b_{\max} > 0$ such that

$$\|w(\tau)\|_{H^1} \leq K_2 e^{-b\tau} \|w(\tau(0))\|_{H^1} + K_2 \int_{\tau(0)}^{\tau} e^{-b(\tau - s)} \|Q\mathcal{G}(s)\|_{H^1} ds$$

for any $b \in (0, b_{\max})$. Estimating $\mathcal{G}$, we obtain

$$\begin{aligned} \|\mathcal{G}\|_{H^1} &\leq K \left[\left| \frac{c_0}{c - \dot{\theta}} \right| \left(|\dot{\theta}| + |\dot{c}| \right) + \left(|c - c_0| + \left| 1 - \frac{c_0}{c - \dot{\theta}} \right| \right) \|w\|_{H^1} \right] \\ &\quad + K \left[\left| \frac{c_0}{c - \dot{\theta}} \right| \|v\|_{H^1} \|w\|_{H^1} \right] \\ &\leq K_3 \delta_\star \|w\|_{H^1}. \end{aligned} \tag{6.8}$$

We have made use of (6.6) to control $\dot{\theta}$ and $\dot{c}$ in terms of $\|w\|_{H^1}$.

Therefore,

$$\|w(\tau)\|_{H^1} \leq K_2 e^{-b\tau} \|w(\tau(0))\|_{H^1} + K_2 K_3 \delta_\star \int_{\tau(0)}^{\tau} e^{-b(\tau - s)} \|w(s)\|_{H^1} ds. \tag{6.9}$$

Defining $\psi(s) = e^{bs} \|w(s)\|_{H^1}$, (6.9) becomes

$$\psi(\tau) \leq K_2 \|w(0)\|_{H^1} + K_2 K_3 \delta_\star \int_{\tau(0)}^{\tau} \psi(s) ds,$$

for which we may apply Gronwall's inequality to get

$$\|w(\tau)\|_{H^1} \leq K_2 \|w(\tau(0))\|_{H^1} e^{-(b-K_2K_3\delta_\star)(\tau-\tau(0))}. \tag{6.10}$$

So for $\delta_\star$ small enough, $b - K_2K_3\delta_\star > 0$ and we induce decay in the H^1 norm of w. In particular, suppose that $\delta_\star \leq \delta_b = \frac{1}{2}b/(K_2K_3)$ and let

$$b' = b - K_2K_3\delta_\star.$$

We then return to t-time,

$$\tau - \tau(0) = \frac{1}{c_0}\int_0^t c(s)ds + \frac{1}{c_0}\left(\theta(0) - \theta(t)\right) \geq \frac{1}{c_0}\left(c_0 - \delta_\star\right)t - 2\frac{\delta_\star}{c_0}.$$

Therefore,

$$\|w(t)\|_{H^1} \leq \tilde{K}_2 \|w(t=0)\|_{H^1} e^{-\kappa t} \tag{6.11}$$

with $\kappa = \frac{1}{2}b'$. We now have $e^{\kappa t}\|w\|_{H^1}$ estimated in terms of the data.

Unweighted perturbation. Applying this to Proposition 5.4, we have the estimate

$$\begin{aligned}
\|v(t)\|_{H^1} &\leq K\left(\sqrt{|\Delta\mathcal{N}|} + \sqrt{|c(t)-c_0|} + |c(t)-c_0| + |c(t)-c_0|^{3/2}\right) \\
&\quad + K\left(\sqrt{\|w(t)\|_{L^2}} + \|w(t)\|_{H^1}\right) \\
&\leq K\left(\sqrt{|\Delta\mathcal{N}|} + \sqrt{|c(t)-c_0|}(1+\delta_\star+\delta_\star^2) + \sqrt{\|w(t)\|_{L^2}}(1+\delta_\star)\right) \\
&\leq K\left(\sqrt{|\Delta\mathcal{N}|} + \sqrt{|c(t)-c_0|} + \sqrt{\|w(0)\|_{L^2}}\right).
\end{aligned} \tag{6.12}$$

If we had control of $\sqrt{|c(t)-c_0|}$, then we would also control $\|v(t)\|_{H^1}$ in terms of the data.

Deviation in c from c_0. Estimating $|c(t)-c_0|$ using (6.6) and (6.11), we get

$$\begin{aligned}
|c(t)-c_0| &\leq |c(0)-c_0| + \int_0^t |\dot{c}(s)|ds \leq |c(0)-c_0| \\
&\quad + \int_0^t K_1\left(|c(s)-c_0| + \|v(s)\|_{H^1}\right)\|w(s)\|_{L^2}ds \\
&\leq |c(0)-c_0| + K_1\delta_\star\int_0^t \|w(s)\|_{H^1}ds \leq |c(0)-c_0| \\
&\quad + K_1\delta_\star\int_0^t K_2\|w(t0)\|_{H^1}e^{-\kappa s}ds \\
&\leq |c(0)-c_0| + K_1K_2\delta_\star\|w(0)\|_{H^1}/\kappa.
\end{aligned}$$

So we now have $|c(t)-c_0|$ in terms of data, which we rewrite as

$$\sqrt{|c(t)-c_0|} \leq K_4\left(\sqrt{|c(0)-c_0|} + \sqrt{\|w(0\|_{H^1}}\right), \tag{6.13}$$

which in turn gives

$$\|v(t)\|_{H^1} \leq K_5\left(\sqrt{|\Delta\mathcal{N}|} + \sqrt{|c(0)-c_0|} + \sqrt{\|w(0)\|_{H^1}}\right). \tag{6.14}$$

Deviation in θ from θ_0. As with the speed parameter,

$$\begin{aligned}|\theta(t)-\theta_0| &\le |\theta(0)-\theta_0| + \int_0^t |\dot\theta(s)|ds \le |\theta(0)-\theta_0| \\ &\quad + \int_0^t K_1\left(|c(s)-c_0| + \|v(s)\|_{H^1}\right)\|w(s)\|_{L^2}ds \\ &\le |\theta(0)-\theta_0| + K_1\delta_\star\int_0^t \|w(s)\|_{H^1}ds \le |c(0)-c_0| \\ &\quad + K_1\delta_\star\int_0^t K_2\|w(t0)\|_{H^1}e^{-\kappa s}ds \\ &\le |\theta(0)-\theta_0| + K_1K_2\delta_\star\|w(0)\|_{H^1}/\kappa.\end{aligned}$$

This is rewritten as

$$|\theta(t)-\theta_0| \le K_7\left(|\theta(0)-\theta_0| + \sqrt{\|w(0\|_{H^1}}\right). \tag{6.15}$$

Another estimate on the temporal change of variables.

$$\left|1-\frac{c_0}{c-\dot\theta}\right| \le \frac{|c-c_0|+|\dot\theta|}{|c-\dot\theta|} \le 2\left(|c-c_0|+|\dot\theta|\right).$$

Then using (6.6) and (6.13)

$$\begin{aligned}\left|1-\frac{c_0}{c-\dot\theta}\right| &\le K\left(\sqrt{|c(0)-c_0|}+\|w(t)\|_{H^1}\right)K \le \left(|c(0)-c_0| + K_2\|w(0)\|_{H^1}\right) \\ &\le K_6\left(\sqrt{|c(0)-c_0|}+\sqrt{\|w(0)\|_{H^1}}\right).\end{aligned} \tag{6.16}$$

Combining (6.11), (6.13), (6.14), (6.15), and (6.16), we have (6.5) with $\delta_\star = \min\{\frac{1}{2}, \delta_{\mathcal{A}}, \delta_b, \delta_\tau, \delta_\theta\}$, $K_\star = \max\{\tilde K_2, K_4, K_5, K_6, K_7\}$. ∎

6.2. Main result. We now prove Theorem 1.1. Let $c_\star$ and $\hat\vartheta$ be as in Theorem 3.1 (b). Take $c_0 \in (n, c_\star]$ and $a \in (\gamma\hat\vartheta, a_\star(c_0)]$. By Theorem 3.1 (b), there exists $\varepsilon > 0$ such that the only eigenvalue of $A_{c_0,a}$ with $\Re\lambda \ge -\varepsilon$ is at the origin. Thus we satisfy the first assumption of Proposition 6.1.

Define $\mathcal{T}$ to be the set of nonnegative numbers, T, such that, given c_0, a, and $v_0 \in H^1 \cap H^1_a$,

- a solution exists, $\phi - 1 \in C([0,T), H^1 \cap H^1_a)$, satisfying

$$\inf_x \phi(x,t) \ge \frac{1}{4} = \alpha_0 > 0, \tag{6.17}$$

$$\inf_x\left(\phi(x,t)^m - a^2\phi(x,t)^n\right) \ge \frac{1}{4}\inf_x\left(\phi_{c_0}(x)^m - a^2\phi_{c_0}(x)^n\right) = \beta_0 > 0; \tag{6.18}$$

- a decomposition of ϕ into $(v(y(x,t),t), \theta(t), c(t))$ exists for $t \in [0,T)$;
- (6.3) holds for $t \in [0,T)$.

Set $T_* = \sup\mathcal{T}$. We will first show that there exists $\epsilon_* > 0$, such that for $\epsilon \le \epsilon_*$, if

$$\|v_0\|_{H^1} + \|v_0\|_{H^1_a} \le \epsilon,$$

then $T_* > 0$. This will be proved using the continuous dependence upon the data. Using Proposition 6.1, we will then prove $T_* = \infty$.

Let $\delta_\star$ and $K_\star$ be as in Proposition 6.1.

The most difficult part of the proof will be ensuring the persistence of (6.3). Consider that, at $t = 0$, the left-hand side of that equation may be estimated with

$$\begin{aligned}\mathrm{LHS}(t=0) \leq\ & |\Delta\mathcal{N}| + \|v_0\|_{H^1} + \|\partial_y \phi_{c_0}\|_{H^1}|\theta_0 - \theta(0)| \\ & + \sqrt{|c(0) - c_0|} + |\theta(0) - \theta_0| + \|\phi_{c_0} - \phi_{c(0)}\|_{H^1} \\ & + \left|1 - \frac{c_0}{c(0) - \dot\theta(0)}\right| + \sqrt{e^{a\theta(0)}\|v_0\|_{H^1_a}} \\ & + \sqrt{\|\partial_y \phi_{c_0}\|_{H^1_a}|\theta_0 - \theta(0)|} + \sqrt{\|\phi_{c_0} - \phi_{c(0)}\|_{H^1_a}}.\end{aligned} \tag{6.19}$$

There exists a choice of δ' and ϵ' such that if

$$|c(0) - c_0| + |\dot c(0)| + |\theta(0) - \theta_0| + |\dot\theta(0)| \leq \delta' \tag{6.20}$$

$$\|v_0\|_{H^1 \cap H^1_a} \leq \epsilon', \tag{6.21}$$

then the right-hand side of (6.19) will be bounded by $\frac{1}{2}\delta_\star$. Set $\delta_1 = \min\left\{\frac{1}{4}, \delta'\right\}$. From Propositions 5.1 and 5.2, let δ_0, δ_0' be the corresponding values for δ_1.

There exists $\epsilon_{\text{exist}} \in (0,1)$ such that if $\|v_0\|_{H^1 \cap H^1_a} \leq \epsilon_{\text{exist}}$, then

$$\begin{aligned}&\inf_x \left[\phi_{c_0}(x + \theta_0) + v_0(x)\right] \geq 2\alpha_0, \\ &\inf_x \left[\left(\phi_{c_0}(x + \theta_0) + v_0(x)\right)^m - a^2\left(\phi_{c_0}(x + \theta_0) + v_0(x)\right)^n\right] \geq 2\beta_0.\end{aligned}$$

By Theorem 2.1, there exist $t_1 > 0$ and a solution in $C^1([0, t_1], H^1 \cap H^1_a)$ satisfying (6.17) and (6.18). Furthermore, we will have the a priori $H^1 \cap H^1_a$ bound that

$$\sup_{t \leq t_1} \|\phi(t) - 1\|_{H^1 \cap H^1_a} \leq 2\|\phi_{c_0}(\cdot + \theta_0) + v_0 - 1\|_{H^1 \cap H^1_a} \leq 2(\|\phi_{c_0} - 1\| + 1).$$

Set

$$\epsilon_1 = \min\left\{\epsilon_{\text{exist}}, \epsilon', \frac{1}{2}e^{-a\theta_0}\delta_0\right\} \tag{6.22}$$

and let $\|v_0\| \leq \epsilon_1$. As noted, the solution exists, satisfying (6.17) and (6.18), until at least $t_1 > 0$. At $t = 0$,

$$\|e^{a(\cdot + \theta_0)}\left(\phi_0 - \phi_{c_0}(\cdot + \theta_0)\right\|_{H^1} \leq e^{a\theta_0}\|v_0\|_{H^1} \leq \frac{1}{2}\delta_0,$$

so by the continuity of ϕ in time, we have

$$\|e^{a(\cdot + \theta_0)}\left(\phi(t) - \phi_{c_0}(\cdot - c_0 t + \theta_0)\right\|_{H^1} \leq \delta_0 \quad \text{for some } t_2 \in (0, t_1).$$

Therefore the decomposition exists, with the δ_1 bound on the modulation parameters, up till $t_2 > 0$.

Also at $t = 0$, using the δ_1 bound on the parameters,

$$\sqrt{|\Delta\mathcal{N}|} + \sqrt{\|w(0)\|_{H^1}} + \sqrt{|c(0) - c_0|} + \left|1 - \frac{c_0}{c(0) - \dot\theta(0)}\right| + \|v(0)\|_{H^1} \leq \frac{1}{2}\delta_\star. \tag{6.23}$$

All of these terms are continuous in time, and there exists some $t_3 \in (0, t_2)2$, for which this remains smaller than $\delta_\star$. Therefore, for $\epsilon_* \le \epsilon_1$, $t_3 \in \mathcal{T}$ and $T_* > 0$.

Continuing to infinity. A few more constraints on ϵ_* are needed to continue out to $t = \infty$. There exists $\epsilon_{\alpha\beta}$ such that for $\epsilon \le \epsilon_{\alpha\beta}$, if

$$\sqrt{|c - c_0|} + \|v\|_{H^1} \le \epsilon,$$

then

$$\inf_y \left[\phi_c(y) + v(y)\right] \ge \alpha_0,$$
$$\inf_y \left[(\phi_c(y) + v_0(y))^m - a^2 (\phi_c(y) + v_0(y))^n\right] \ge \beta_0.$$

Let $\epsilon_2 > 0$ be so small that

$$K_\star \epsilon_2 = \min\left\{\frac{\delta_\star}{3}, \sqrt{\delta_0/3}, \sqrt{\delta_0'}, \frac{1}{2}\epsilon_{\alpha\beta}\right\} \tag{6.24}$$

and set

$$\epsilon_* = \min\left\{\epsilon_1, \epsilon_2\right\}. \tag{6.25}$$

Now, assume $\|v_0\|_{H^1 \cap H^1_a} \le \epsilon \le \epsilon_*$. As above, for this data we will have $T_* > 0$. Assume $T_* < \infty$. For any $T < T_*$, on the interval $[0, T]$, the solution exists with (6.17) and (6.18), as does the decomposition, and (6.3) holds.

Then

$$\begin{aligned}\|\phi(t) - 1\|_{H^1 \cap H^1_a} &\le \max\left\{1, e^{a\left(\int_0^t c(s)ds - \theta(t)\right)}\right\}\left(\|\phi_{c(t)} - 1\|_{H^1 \cap H^1_a} + \|v(t)\|_{H^1 \cap H^1_a}\right) \\ &\le e^{a((c_0+\delta_\star)T_* + |\theta_0| + \delta_\star)}\left(\sup_{|c-c_0|\le\delta_\star} \|\phi_c - 1\|_{H^1 \cap H^1_a} + \delta_\star + \delta_\star^2\right) < \infty,\end{aligned}$$

and this bound is uniform in $T < T_*$. By assumption, (6.17) and (6.18) hold for $t \in [0, T]$, uniformly in $T < T_*$, which may written as

$$\left\|\frac{1}{\phi(\cdot, t)}\right\|_{L^\infty} \le \frac{1}{\alpha_0} < \infty \quad \text{and} \quad \left\|\frac{1}{\phi(\cdot, t)^m - a^2\phi(\cdot, t)^n}\right\|_{L^\infty} \le \frac{1}{\beta_0} < \infty.$$

Therefore, according to (2.7), $\phi(x, t)$ may be extended beyond T_* by some amount $t_2 > 0$. Hence, if $T_* \ne \infty$, it must be a failure for either the decomposition to continue to exist or for (6.3) to hold.

Again using the Proposition 6.1 and our choice of ϵ_*,

$$\|w(\cdot, t)\|_{H^1_a} \le \delta_0/3 \quad \text{and} \quad |c(t) - c_0| \le \delta_0' \quad \text{for all } t \le T, \text{ uniformly in } T < T_*.$$

Since ϕ exists until at least $T_* + t_2$, we may apply Proposition 5.2 to extend the decomposition for some amount $t_\star \le t_2$ also beyond T_*.

By assumption,

$$\sqrt{|c(t) - c_0|} + \|v(t)\|_{H^1} \le K_\star \epsilon_* \le \frac{1}{2}\epsilon_{\alpha\beta} \quad \text{for } t < T_*.$$

Again, by continuity, these remain bounded by $\epsilon_{\alpha\beta}$ until some time $t_3 \in (0, t_\star)$ beyond T_*, so (6.17) and (6.18) also persist beyond T_*.

We may now apply Proposition 6.1 past T_*. This gives

$$\sqrt{\|w(\cdot,t)\|_{H^1}}+\sqrt{|c(t)-c_0|}+|\theta(t)-\theta_0|+\left|1-\frac{c_0}{c(t)-\dot\theta(t)}\right|+\|v(\cdot,t)\|_{H^1}\leq K_\star\epsilon_*\leq\frac{1}{3}\delta_\star$$

for $t\leq T<T_*$. As $\sqrt{|\Delta\mathcal{N}|}$ is time invariant and smaller than $\frac{1}{2}\delta_\star$,

$$\begin{aligned}\sqrt{|\Delta\mathcal{N}|}+\sqrt{\|w(\cdot,t)\|_{H^1}}+\sqrt{|c(t)-c_0|}+|\theta(t)-\theta_0|\\ +\left|1-\frac{c_0}{c(t)-\dot\theta(t)}\right|+\|v(\cdot,t)\|_{H^1}\leq K_\star\epsilon_*\leq\frac{5}{6}\delta_\star\end{aligned}$$

for $t\leq T$, uniformly in $T<T_*$. But all of these functions are continuous for $t\in[0,T_*+t_3]$; so for some $t_4>0$, this expression remains bounded by $\delta_\star$. This contradicts $T_*<\infty$. So a solution exists for all time satisfying (6.17), (6.18), along with a decomposition and (6.3).

Since we may then apply Proposition 6.1 for all time, we will always have (6.5). By virtue of $K_\star\epsilon_*<\delta_\star\leq\delta_{\mathcal{B}}$, we will be able to invert the matrix $\mathcal{B}(t)$ for the modulation equations, (5.24). Therefore $|\dot c(t)|+|\dot\theta(t)|\leq K\epsilon e^{-\kappa t}$ and

$$\lim_{t\to\infty}c(t)=c_\infty$$

exists, and if we define

$$\lim_{t\to\infty}\left(\theta(t)+\int_0^t(c(s)-c_\infty)\,ds\right)=\theta_\infty,$$

then

$$\begin{aligned}&\|\phi(\cdot,t)-\phi_{c_\infty}(\cdot-c_\infty t+\theta_\infty)\|_{H^1}\\ &\quad\leq K_\star\epsilon+\|\phi_{c_\infty}(\cdot-c_\infty t+\theta_\infty)-\phi_{c(t)}\left(\cdot-\textstyle\int_0^t c(s)ds+\theta(t)\right)\|_{H^1}\\ &\quad\leq K_\star\epsilon+K|c(t)-c_\infty|+\|\partial_y\phi_{c_\infty}\|_{H^1}\left|\theta(t)+\int_0^t(c(s)-c_\infty)\,ds-\theta_\infty\right|\\ &\quad\leq K_*\epsilon.\end{aligned}$$

Similarly

$$\|\phi(\cdot+c_\infty t-\theta_\infty,t)-\phi_{c_\infty}\|_{H^1_a}\leq K_*\epsilon e^{-\kappa t}.\qquad\square$$

6.3. Remarks. This proof is equally applicable in the Hamiltonian case, $n+m=0$, for values of c not in the discrete set, E, of points for which A_a has an imaginary eigenvalue.

7. Summary and discussion. We have thus shown that in the space $H^1\cap H^1_a$, the solitary waves are asymptotically stable. This dovetails with an extension of global existence to data in a neighborhood of the solitary waves. In the Hamiltonian case, we can extend it beyond $c_\star$ via analytic continuation, as was done in [21], and this is analytically verified for $n=2$, with computations in [40] suggesting it is true for all $n>1$. Furthermore, to the extent that we will accept a computation of the Evans function as proof, our result generalizes to large amplitude solitary waves with $c>c_\star$.

To our knowledge, this is the first result for which asymptotic stability is established for a conservative PDE in the absence of a variational principle.

Open problems include a weakening of the assumption of exponential decay on the perturbation. This might be accomplished through the use of an algebraic spatial weight, which would require the perturbation to decay algebraically rapidly. Yet less restrictive would be to use the approach of Merle and his colleagues [11, 18, 19, 22]. However, there is a tradeoff in both of these approaches: weakening the assumption on the spatial decay rate of the perturbation also weakens what can be proved about the rate at which the perturbation decays in time.

Finally we remark that the multidimensional case of (1.1) is wide open. While there is no existence theory for the two- and three-dimensional problems, perhaps a similar approach, of working in a neighborhood of a solitary wave, could be applied proving both existence and stability.

Appendix A. Properties of solitary waves.

A.1. Analyticity. Here we provide a proof of Corollary 2.9. Let us restate the crucial theorem.

THEOREM A.1 (Corollary 4.1.6 of [9]). *Suppose that f is a solution of the convolution equation $f = K * G(f)$ such that $f \in L^2 \cap L^\infty$ and $\lim_{|x|\to\infty} f(x) = 0$. If the Fourier transform $\hat{K}$ of the integral kernel K satisfies the decay condition $|\hat{K}(\xi)| \le A_1(1 + A_2|\xi|^m)$ for some constants $A_1, A_2 > 0$ and $m \ge 1$, and G is an infinitely differentiable function whose domain contains the range $R(f)$ of f, having all its derivatives bounded on $R(f)$ and satisfying the condition $G(0) = 0$, then $f, G(f) \in H^\infty$. In addition, if G is an analytic function on an open set U containing $R(f)$, G is continuous up to the boundary of ∂U of U, and*

$$\mathrm{d}\left(\partial U, R(f)\right) > \|K\|_{L^2}, \tag{A.1}$$

then there exists a constant $\sigma_0 > 0$ such that f and $G(f)$ both have analytic extensions to the strip $\{z \in \mathbb{C} : |\Im z| < \sigma_0\}$.

Let $u_c = \phi_c - 1$. By Theorem 2.6 and Corollary 2.7, u_c is positive, in $L^\infty \cap L^2$, and decays exponentially fast at $\pm\infty$. Using (2.14) and (2.15), the equation may be written as

$$-\gamma^2 u_c + \partial_x^2 u_c + \int_0^{u_c} \partial_\tau^3 F_2(1+\tau; c)\frac{(u_c + 1 - \tau)^2}{2} d\tau = 0. \tag{A.2}$$

Define

$$G(z) = \int_0^{z} \partial_\tau^3 F_2(1+\tau; c)\frac{(z + 1 - \tau)^2}{2} d\tau. \tag{A.3}$$

Taking a Fourier transform of (A.2), the equation is

$$-\gamma^2 \widehat{u_c}(\xi) - \xi^2 \widehat{u_c}(\xi) + \widehat{G(u_c)}(\xi) = 0.$$

This becomes the nonlinear convolution equation

$$u_c(x) = K * G(u_c)(x), \tag{A.4}$$

$$\hat{K}(\xi) = \frac{1}{\gamma^2 + \xi^2}. \tag{A.5}$$

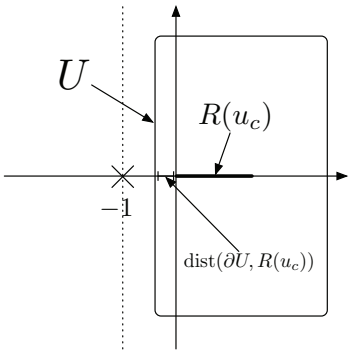


FIG. 6. *A plot of a possible domain U and the range of u_c, $R(u_c)$. Note that the distance between these sets as drawn is the distance into the left-hand side of the complex plane that U extends, and that any such ovoid will be acceptable as long as it stays to the right of $\Re z = -1$.*

For purposes of satisfying (A.1), let us take $\tilde{K}(x) = \alpha K(x)$ and $\tilde{G}(z) = \alpha^{-1} G(z)$ for $\alpha > 0$, where α is to be determined. Under this trivial scaling, $u_c = \tilde{K} * \tilde{G}(u_c)$.

$\widehat{\tilde{K}}$ satisfies the decay estimate for Theorem A.1. $\tilde{G}(z)$ will have a singularity at $z = -1$ but is otherwise analytic. The range of u_c is the finite segment

$$R(u_c) = [0, u_{\max}],$$

and $\tilde{G}$ is infinitely differential there, with all derivatives bounded. $G(0) = 0$. Hence, the first part of Theorem A.1 applies: u_c and $\tilde{G}(u_c)$ are in H^∞.

Now, consider the set U in Figure 6. In this figure, $\mathrm{d}(\partial U, R(u_c))$ is the distance U stretches into the left half-plane:

$$\|\tilde{K}\|_2 = \alpha \sqrt{\frac{\pi}{2}} \gamma^{-3/2}.$$

Picking α so small that the norm is less than 1, we can find a U such that the distance between ∂U and $R(u_c)$ exceeds $\|\tilde{K}\|_2$, satisfying (A.1) and proving analyticity in a strip. □

A.2. Continuity as a function of speed. Consider the functional

$$\mathcal{F}[c, u] = \partial_x^2 u + F_2(1 + u; c) \tag{A.6}$$

as a mapping from $H^2 \times \mathbb{R} \to L^2$. The solitary wave $u_c = \phi_c - 1$ satisfies $\mathcal{F}[c, u_c] = 0$. Using this functional, we prove Corollary 2.11 via the implicit function theorem. Given a particular $\hat{c} > n$, set $\hat{u} = u_{\hat{c}}$.

Let H^2_{even} and L^2_{even} be the subspaces of H^2 and L^2, respectively, of only even functions. Define the sets

$$M_0 = \left(\hat{c} - \frac{\hat{c} - n}{2}, \hat{c} + \frac{\hat{c} - n}{2} \right) \subset \mathbb{R},$$

$$N_0 = \left\{ u \in H^2_{\mathrm{even}} : \quad \|u - \hat{u}\|_{H^2} \leq \frac{1}{2} \right\} \subset H^2_{\mathrm{even}},$$

$$Z = L^2_{\mathrm{even}}.$$

Note that set M_0 is bounded away from zero and all of the functions in N_0 are uniformly bounded from below by $\frac{1}{2}$. Therefore $\mathcal{F}$ is well defined on $M_0 \times N_0$ and will be a C^1 mapping on this set into Z.

Set

$$T = \frac{\delta \mathcal{F}}{\delta c}[\hat{c}, \hat{u}] = \partial_c F_2(1+\hat{u}; \hat{c}),$$
$$S = \frac{\delta \mathcal{F}}{\delta u}[\hat{c}, \hat{u}] = \partial_x^2 + \partial_u F_2(1+\hat{u}; \hat{c}).$$

T and S are bounded operators on $\mathbb{R} \to L^2_{\text{even}}$ and $H^2_{\text{even}} \to L^2_{\text{even}}$, respectively.

Let $f \in L^2_{\text{even}}$ and consider the problem $Su = f$. As an elliptic problem, this has a solution, provided $f \perp \ker\left(S^\dagger\right)$. Note that $S\partial_x \hat{u} = 0$. S is self-adjoint, has smooth coefficients, and is in one spatial dimension; this is the unique element of the kernel. But $\partial_x \hat{u}$ is an odd function, hence f is orthogonal to it and the equation has a solution u satisfying a bound

$$\|u\|_{H^2} \leq K \|u\|_{L^2}.$$

Because the coefficients in S and the right-hand side, f, are all even functions, $\tilde{u}(x) = u(-x)$ also solves $Su = f$. By the uniqueness of the solution, $u = \tilde{u}$, so u is an even function. Therefore $u \in H^2_{\text{even}}$ and the map $S : H^2_{\text{even}} \to L^2_{\text{even}}$ is onto with bounded inverse.

The kernel of S, restricted to $u \in H^2_{\text{even}}$, is trivial, so the implicit function theorem may be applied to conclude the existence of a function $\mathcal{G} : M_1 \to H^2_{\text{even}}$, $\hat{c} \in M_1 \subset M_0$, such that

$$\mathcal{F}[c, \mathcal{G}(c)] = 0$$

for all $c \in M_1$. The mapping $\mathcal{G}$ is C^1. For any such c,

$$\partial_x^2 \mathcal{G}(c) + F_2(\mathcal{G}(c) + 1; c) = 0.$$

This is just the solitary wave equation. Therefore $\mathcal{G}(c) = u_c = \phi_c - 1$, and the mapping $c \mapsto \phi_c - 1$ is $C^1(\mathbb{R}; H^2)$. The analyticity of the mapping may be proven by checking the analyticity of $\mathcal{F}$ in a neighborhood of $(\hat{c}, \hat{u})$.

To prove continuity in $H^2 \cap H^2_a$, the proof is similar. Fixing $a < \frac{1}{2}$, and taking $\hat{c} \in (n/(1-4a^2), \infty)$, we let $M_a = M_0 \cap (n/(1-4a^2), \infty)$, $N_a = N_0 \cap H^2_a$, and $Z_a = Z \cap L^2_a$. We must check that $S : H^2_{\text{even}} \cap H^2_a \to L^2_{\text{even}} \cap L^2_a$ is onto with bounded inverse. This is accomplished using the previous result and studying $S_a = D_a^2 + \partial_u F_2(1+\hat{u}; \hat{c})$ on $H^2_{\text{even}} \to L^2_{\text{even}}$.

Appendix B. Perturbation expansions. Here we provide some explicit calculations, including those for Proposition 2.13.

Note the expansions

$$\phi^n = \phi_c^n + n\phi_c^{n-1} v + f_n[\phi_c, v]v,$$
$$\phi^m = \phi_c^m + m\phi_c^{m-1} v + f_m[\phi_c, v]v,$$
$$f_p[a,b] = \int_0^1 \left[p\,(a+\tau b)^{p-1} - pa^{p-1}\right] d\tau,$$

and

$$\begin{aligned} H_\phi^{-1} = H_{\phi_c}^{-1} - H_{\phi_c}^{-1} B\left[n\phi_c^{n-1}v + f_n[\phi_c,v]v, m\phi_c^{m-1}v + f_m[\phi_c,v]v\right] H_{\phi_c}^{-1} \\ + H_{\phi_c}^{-1} B[\phi^n - \phi_c^n, \phi^m - \phi_c^m]\left(H_\phi^{-1} - H_{\phi_c}^{-1}\right), \end{aligned} \tag{B.1}$$

$$B[a,b]u = -\partial_x\left(a\partial_x u\right) + bu. \tag{B.2}$$

Recall (1.2),

$$\partial_t\phi = \dot{c}\partial_c\phi_c + \left(\dot{\theta} - c\right)\partial_y\phi_c + \left(\dot{\theta} - c\right)\partial_y v + v_t = -\left(\phi_c + v\right)^n H_{\phi_c+v}^{-1}\partial_y\left(\phi_c + v\right)^n.$$

Using the above expansions and the solitary wave equation, $c\partial_y\phi_c = \phi_c^m H_{\phi_c}^{-1}\partial_y\left(\phi_c^n\right)$, this may be expanded into

$$v_t = \phi_c^m H_{\phi_c}^{-1}\partial_y\left[-c\phi_c^n\partial_y^2\left(\phi_c^{-m}v\right) + cv - n\phi_c^{n-1}v - cn\phi_c^{n-1}\partial_y\left(\phi_c^{-m}\partial_y\phi_c\right)v\right] \tag{B.3}$$

$$\begin{aligned} &- \dot{\theta}\partial_y v - \dot{c}\partial_c\phi_c - \dot{\theta}\partial_y\phi_c \\ &- f_m[\phi_c,v]vH_{\phi_c}^{-1}\partial_y(\phi^n) + m\phi_c^{m-1}vH_{\phi_c}^{-1}\left(H_\phi - H_{\phi_c}\right)H_\phi^{-1}\partial_y(\phi^n) \\ &- m\phi_c^{m-1}vH_{\phi_c}^{-1}\partial_y\left(\phi^n - \phi_c^n\right) - \phi_c^m H_{\phi_c}^{-1}\partial_y\left(f_n[\phi_c,v]v\right) \end{aligned} \tag{B.4}$$

$$\begin{aligned} &+ \phi_c^m H_{\phi_c}^{-1} B[f_n[\phi_c,v]v, f_m[\phi_c,v]v]H_\phi^{-1}\partial_y\left(\phi^n\right) \\ &- \phi_c^m H_{\phi_c}^{-1} B[n\phi_c^{n-1}v, m\phi_c^{m-1}v]H_{\phi_c}^{-1}\left(H_\phi - H_{\phi_c}\right)H_\phi^{-1}\partial_y\left(\phi^n\right) \\ &+ \phi_c^m H_{\phi_c}^{-1} B[n\phi_c^{n-1}v, m\phi_c^{m-1}v]H_{\phi_c}^{-1}\partial_y\left(\phi^n - \phi_c^n\right). \end{aligned} \tag{B.5}$$

(B.3) is a linear term. (B.4) will decay to zero as $\theta(t)$ and $c(t)$, the modulating parameters, approach their asymptotic limits. (B.5) is purely nonlinear in v.

We define $\mathcal{F}_1[v;\phi_c]$, the term nonlinear in v, as

$$\begin{aligned} \mathcal{F}_1[v;\phi_c] = &- f_m[\phi_c,v]vH_{\phi_c}^{-1}\partial_y(\phi^n) + m\phi_c^{m-1}vH_{\phi_c}^{-1}\left(H_\phi - H_{\phi_c}\right)H_\phi^{-1}\partial_y(\phi^n) \\ &- m\phi_c^{m-1}vH_{\phi_c}^{-1}\partial_y\left(\phi^n - \phi_c^n\right) - \phi_c^m H_{\phi_c}^{-1}\partial_y\left(f_n[\phi_c,v]v\right) \\ &+ \phi_c^m H_{\phi_c}^{-1} B[f_n[\phi_c,v]v, f_m[\phi_c,v]v]H_\phi^{-1}\partial_y\left(\phi^n\right) \\ &- \phi_c^m H_{\phi_c}^{-1} B[n\phi_c^{n-1}v, m\phi_c^{m-1}v]H_{\phi_c}^{-1}\left(H_\phi - H_{\phi_c}\right)H_\phi^{-1}\partial_y\left(\phi^n\right) \\ &+ \phi_c^m H_{\phi_c}^{-1} B[n\phi_c^{n-1}v, m\phi_c^{m-1}v]H_{\phi_c}^{-1}\partial_y\left(\phi^n - \phi_c^n\right). \end{aligned} \tag{B.6}$$

The operator S is given by

$$\begin{aligned} S[c_0,c,\dot{\theta}] = &\, mc_0\left(\phi_{c_0}^{-1}\partial_y\phi_{c_0} - \frac{c}{c-\dot{\theta}}\phi_c^{-1}\partial_y\phi_c\right) \\ &+ n\phi_{c_0}^m H_{c_0}^{-1}\partial_y\left[\left(\phi_{c_0}^{-1} - c_0(\phi_{c_0}^{-1} - 1)\right)\cdot\right] \\ &- \frac{nc_0}{c-\dot{\theta}}\phi_c^m H_{\phi_c}^{-1}\partial_y\left[\left(\phi_c^{-1} - c(\phi_c^{-1} - 1)\right)\cdot\right] \\ &- c_0 m\left\{\phi_{c_0}^m H_{\phi_{c_0}}^{-1}\left[\phi_{c_0}^{-1}\partial_y\phi_{c_0}\cdot\right] - \frac{c}{c-\dot{\theta}}\phi_c^m H_{\phi_c}^{-1}\left[\phi_c^{-1}\partial_y\phi_c\cdot\right]\right\}. \end{aligned} \tag{B.7}$$

Finally, the terms making up $\mathcal{G}_1$ from (2.36):

$$\mathcal{G}_1 = \tilde{\mathcal{G}}_1 + \tilde{\mathcal{G}}_2 + \tilde{\mathcal{G}}_3 + \tilde{\mathcal{G}}_4 + \tilde{\mathcal{G}}_5 + \tilde{\mathcal{G}}_6 + \tilde{\mathcal{G}}_7, \tag{B.8}$$

$$\tilde{\mathcal{G}}_1 = -f_m[\phi_c, v] w H_{\phi_c}^{-1} \partial_y(\phi^n), \tag{B.9}$$

$$\tilde{\mathcal{G}}_2 = m\phi_c^{m-1} w H_{\phi_c}^{-1} \left(H_\phi - H_{\phi_c}\right) H_\phi^{-1} \partial_y \left(\phi^n\right), \tag{B.10}$$

$$\tilde{\mathcal{G}}_3 = -m\phi_c^{m-1} w H_{\phi_c}^{-1} \partial_y \left(n\phi_c^{n-1} v + f_n[\phi_c, v] v\right), \tag{B.11}$$

$$\tilde{\mathcal{G}}_4 = -\phi_c^m H_{\phi_c,a}^{-1} D_a \left(f_n[\phi_c, v] w\right), \tag{B.12}$$

$$\tilde{\mathcal{G}}_5 = \phi_c^m H_{\phi_c,a}^{-1} B_a \left[f_n[\phi_c, v] w, f_m[\phi_c, v] w\right] H_\phi^{-1} \partial_y \left(\phi^n\right), \tag{B.13}$$

$$\tilde{\mathcal{G}}_6 = -\phi_c^m H_{\phi_c,a}^{-1} B_a \left[n\phi_c^{n-1} w, m\phi_c^{m-1} w\right] H_{\phi_c}^{-1} \left(H_\phi - H_{\phi_c}\right) H_\phi^{-1} \partial_y \left(\phi^n\right), \tag{B.14}$$

$$\tilde{\mathcal{G}}_7 = \phi_c^m H_{\phi_c,a}^{-1} B_a \left[n\phi_c^{n-1} w, m\phi_c^{m-1} w\right] H_{\phi_c}^{-1} \partial_y \left(\phi^n - \phi_c^n\right). \tag{B.15}$$

The difference between A_a and A_a^∞ may be written as

$$\begin{aligned} A_a - A_a^\infty = & -cm\phi_c^{-1}\partial_y\phi_c - (\phi_c^m - 1) H_{\phi_c,a}^{-1} D_a \left(n\phi_c^{n-1}\cdot\right) \\ & + cm\phi_c^m H_{\phi_c,a}^{-1} \left(\phi_c^{-1}\partial_y\phi_c\cdot\right) - cn\phi_c^m H_{\phi_c,a}^{-1} D_a \left[(1-\phi_c^{-1})\cdot\right] \\ & + H_{1,a}^{-1} D_a \left[n(1-\phi_c^{n-1})\cdot\right] + H_{1,a}^{-1} \left(\phi_c^m - 1\right) H_{\phi_c,a}^{-1} D_a \left(n\phi_c^{n-1}\cdot\right) \\ & + H_{1,a}^{-1} D_a \left(1-\phi_c^n\right) D_a H_{\phi_c,a}^{-1} D_a \left(n\phi_c^{n-1}\cdot\right). \end{aligned} \tag{B.16}$$

In the space weighted by $\phi_c(x)^{-m}$, this difference is

$$\begin{aligned} \tilde{A}_a - \tilde{A}_a^\infty = & nH_{1,a}^{-1} D_a \left[\left(1-\phi_c^{m-1}\right)\cdot\right] + nH_{1,a}^{-1} D_a \left(\phi_c^m - 1\right) H_{1,a}^{-1} \left(\phi_c^{m-1}\cdot\right) \\ & - nH_{1,a}^{-1} \left[D_a \left(\phi_c^m - 1\right)\right] H_{1,a}^{-1} \left(\phi_c^{m-1}\cdot\right) \\ & - nH_{1,a}^{-1} \left(\phi_c^m - 1\right) H_{1,a}^{-1} \left(\phi_c^m - 1\right) H_{\phi_c,a}^{-1} D_a \left(\phi_c^{m-1}\cdot\right) \\ & + nH_{1,a}^{-1} \left(\phi_c^m - 1\right) H_{1,a}^{-1} D_a \left(\phi_c^n - 1\right) D_a H_{\phi_c,a}^{-1} D_a \left(\phi_c^{m-1}\cdot\right) \\ & - nH_{1,a}^{-1} D_a \left(\phi_c^n - 1\right) D_a H_{\phi_c,a}^{-1} D_a \left(\phi_c^{m-1}\cdot\right) + cmH_{1,a}^{-1} \left(\phi_c^{m-1}\partial_y\phi_c\cdot\right) \\ & + cmH_{1,a}^{-1} \left(\phi_c^m - 1\right) H_{\phi_c,a}^{-1} \left(\phi_c^{m-1}\partial_y\phi_c\cdot\right) \\ & - cmH_{1,a}^{-1} D_a \left(\phi_c^n - 1\right) D_a H_{\phi_c,a}^{-1} \left(\phi_c^{m-1}\partial_y\phi_c\cdot\right) \\ & + cnH_{1,a}^{-1} D_a \left[\phi_c^m \left(\phi_c^{-1} - 1\right)\cdot\right] \\ & + cnH_{1,a}^{-1} \left(\phi_c^m - 1\right) H_{\phi_c,a}^{-1} D_a \left[\phi_c^m \left(\phi_c^{-1} - 1\right)\cdot\right] \\ & - cnH_{1,a}^{-1} D_a \left(\phi_c^n - 1\right) D_a H_{\phi_c,a}^{-1} D_a \left[\phi_c^m \left(\phi_c^{-1} - 1\right)\cdot\right]. \end{aligned} \tag{B.17}$$

Appendix C. Analysis of the characteristic polynomial. In this section we prove that (3.36), $(\lambda - c\mu)(1-\mu^2) + n\mu = 0$, has a unique root of minimal real part on a slit half-plane

$$\{\lambda : \Re\lambda > -\lambda_0\} \backslash (-\lambda_0, -\tilde{\Omega}(\gamma)].$$

There are two ways that this could be false: there could be either a multiple root or two roots with the same real part, but differing imaginary parts. As previously

noted, we will have a unique root of minimal real part for λ in the closed right half-plane, so we need only concern ourselves with $\Re\lambda < 0$.

We will identify a portion of the domain $\Re\lambda < 0$ for which there are neither multiple roots nor complex roots with the same real part. Note that this is a stricter condition than is needed, as the polynomial could have a double root for some λ, where the third root of $P(\mu)$ has a smaller real part than the multiple root.

Note that $P(\pm 1)$ never vanishes, and hence $P(\mu) = 0$ is equivalent to $R(\mu) = \lambda$, where

$$R(\mu) = c\mu + \frac{n\mu}{\mu^2 - 1}. \tag{C.1}$$

C.1. Roots of order greater than one. We start with the possibility of a double or triple root, as this is very easy to rule out. If μ is a multiple root, then in addition to $R(\mu) = \lambda$, we will also have

$$\frac{dR}{d\mu} = c - n\frac{1+\mu^2}{(1-\mu^2)^2} = 0,$$

which has solutions

$$\mu = -\sqrt{\frac{2c + n \pm \sqrt{8cn + n^2}}{2c}}.$$

We have ignored the roots with a + sign in front, as these will correspond to positive λ. Note that they are all real, and hence λ will also be real.

The λ one gets from the root

$$\mu_+ = -\sqrt{\frac{2c + n + \sqrt{8cn + n^2}}{2c}} \leq -\frac{3\sqrt{3}}{2}n$$

is decreasing in c. The other root,

$$\mu_- = -\sqrt{\frac{2c + n - \sqrt{8cn + n^2}}{2c}},$$

will map to λ values

$$\lambda(\mu_-) = -\sqrt{\frac{1}{8}}\sqrt{8c^2 + 20cn - n^2 - 8c\sqrt{n^2 + 8cn} - n\sqrt{n^2 + 8cn}}. \tag{C.2}$$

It can be checked that for $c > n > 1$, $R(\mu_+) < R(\mu_-)$. Therefore, for $\lambda > -\widetilde{\Omega}(c)$, with

$$\widetilde{\Omega}(c) = \sqrt{\frac{1}{8}}\sqrt{8c^2 + 20cn - n^2 - 8c\sqrt{n^2 + 8cn} - n\sqrt{n^2 + 8cn}}, \tag{C.3}$$

$P(\mu)$ cannot have a multiple root.

C.2. Roots of differing imaginary part. If $\mu_1 = \alpha + \imath\beta_1$ and $\mu_2 = \alpha + \imath\beta_2$ are two roots of P, then

$$R(\alpha + \imath\beta_1) = R(\alpha + \imath\beta_2).$$

After matching real and imaginary parts in this expression, the three unknowns, α, β_1, β_2, must satisfy the two equations

$$\begin{aligned}\Re\lambda &= c\alpha + \frac{n\alpha(-1+\alpha^2+\beta_1^2)}{(-1+\alpha^2)^2+2(1+\alpha^2)\beta_1^2+\beta_1^4} \\ &= c\alpha + \frac{n\alpha(-1+\alpha^2+\beta_2^2)}{(-1+\alpha^2)^2+2(1+\alpha^2)\beta_2^2+\beta_2^4},\end{aligned} \tag{C.4}$$

$$\begin{aligned}\Im\lambda &= c\beta_1 - \frac{n\beta_1(1+\alpha^2+\beta_1^2)}{(-1+\alpha^2)^2+2(1+\alpha^2)\beta_1^2+\beta_1^4} \\ &= c\beta_2 - \frac{n\beta_2(1+\alpha^2+\beta_2^2)}{(-1+\alpha^2)^2+2(1+\alpha^2)\beta_2^2+\beta_2^4}.\end{aligned} \tag{C.5}$$

Solving the (C.4) for β_2^2 in terms of α and β_1, there are two families of solutions:

$$\beta_2^2 = \beta_1^2, \tag{C.6}$$

$$\beta_2^2 = \frac{\left(1-\alpha^2\right)\left(3+\beta_1^2+\alpha^2\right)}{\alpha^2+\beta_1^2-1}. \tag{C.7}$$

Without loss of generality, we assume $\beta_1 \neq 0$.

Recall that λ is imaginary if and only if $P(\mu)$ has a purely imaginary root; hence the condition $\Re\lambda < 0$ ensures $\alpha \neq 0$. Then (C.7) implies $0 < |\alpha| \leq 1$. Furthermore, if $|\alpha| = 1$, then either the roots are conjugate or $\beta_2 = 0$. But if $\beta_2 = 0$, then $\mu_2 = 1$, which we know is not a root of $P(\mu)$.

C.2.1. Complex conjugates. When $\beta_1 = -\beta_2 = \beta$, (C.5) implies that λ is real and

$$c - \frac{n(1+\alpha^2+\beta^2)}{(-1+\alpha^2)^2+2(1+\alpha^2)\beta^2+\beta^4} = 0, \tag{C.8}$$

which has roots β^2,

$$\beta^2 = -1-\alpha^2+\frac{n}{2c} \pm \frac{\sqrt{n^2+16c^2\alpha^2}}{2c}. \tag{C.9}$$

We may immediately rule out the negative root for β^2. For $\beta^2 > 0$, α^2 must satisfy

$$1+\frac{n}{2c}-\frac{\sqrt{n^2+24c}}{2c} < \alpha^2 < 1+\frac{n}{2c}+\frac{\sqrt{n^2+24c}}{2c}. \tag{C.10}$$

Using (C.4),

$$\lambda(\alpha) = \frac{n+8c\alpha^2-\sqrt{n^2+16c^2\alpha^2}}{4\alpha}.$$

Since we are concerned only with $\lambda < 0$ here, α must, in addition to (C.10), satisfy

$$\frac{n+8c\alpha^2-\sqrt{n^2+16c^2\alpha^2}}{4\alpha} < 0. \tag{C.11}$$

When $\alpha > 0$, (C.11) requires

$$0 < \alpha^2 < \frac{c-n}{4c}.$$

But $\frac{1}{4}(c-n)/c < 1 + n/(2c) - \sqrt{n^2+24c}/(2c)$, so complex conjugate roots with $\alpha > 0$ are not possible with λ in the left half-plane.

When $\alpha < 0$, (C.11) requires

$$\alpha < -\frac{1}{2}\sqrt{\frac{c-n}{c}}$$

to satisfy (C.11). Consider α in the interval

$$\left[-\sqrt{1+\frac{n}{2c}-\frac{\sqrt{n^2+8cn}}{2c}}, -\frac{1}{2}\sqrt{\frac{c-n}{c}}\right).$$

In this interval, there will not be complex conjugate roots, as it violates (C.10).

λ, as a function of α, is negative and increasing on this interval. For λ in the image of this interval, we may completely rule out complex conjugate roots. The image of this interval is

$$\lambda\left(\left[-\sqrt{1+\frac{n}{2c}-\frac{\sqrt{n^2+8cn}}{2c}}, -\frac{1}{2}\sqrt{\frac{c-n}{c}}\right)\right) = \left[-\widetilde{\Omega}(c), 0\right).$$

Hence, for $0 > \lambda > -\tilde{\Omega}$, one may rule out both multiple roots and complex conjugates.

C.2.2. Nonconjugate complex roots. Consider the case of complex roots with the same real part, but imaginary parts such that $|\beta_1| \neq |\beta_2|$. Squaring both sides of (C.5) and plugging in (C.7) for β_2^2, we get a sixth order polynomial in β_1^2. The roots, as functions of α, are

$$\beta_1^2 = -1 - 2\alpha - \alpha^2, \tag{C.12}$$

$$\beta_1^2 = -1 + 2\alpha - \alpha^2, \tag{C.13}$$

$$\beta_1^2 = 1 - \alpha^2 - 2\sqrt{1-\alpha^2}, \tag{C.14}$$

$$\beta_1^2 = 1 - \alpha^2 + 2\sqrt{1-\alpha^2}, \tag{C.15}$$

$$\beta_1^2 = \frac{(n/c)^2 + 4(n/c)(1-\alpha^2) - 8(1-\alpha^4)}{8(1-\alpha^2)} - \frac{\sqrt{((n/c)^2 - 16(1-\alpha^2))((n/c)^2 - 8(1-\alpha^2)(2\alpha^2 - (n/c)))}}{8(1-\alpha^2)}, \tag{C.16}$$

$$\beta_1^2 = \frac{(n/c)^2 + 4(n/c)(1-\alpha^2) - 8(1-\alpha^4)}{8(1-\alpha^2)} + \frac{\sqrt{((n/c)^2 - 16(1-\alpha^2))((n/c)^2 - 8(1-\alpha^2)(2\alpha^2 - (n/c)))}}{8(1-\alpha^2)}. \tag{C.17}$$

(C.12) and (C.13) force β_1 to be imaginary, and hence they can be ruled out. Using (C.7), if β_1 is either (C.14) or (C.15), then $\beta_2 = \pm\beta_1$, conjugate roots.

In the last two cases, if β_1^2 is to be real, then

$$((n/c)^2 - 16(1-\alpha^2))((n/c)^2 - 8(1-\alpha^2)(2\alpha^2 - (n/c))) \geq 0. \tag{C.18}$$

If we can find for λ in the left half-plane sufficiently close enough to the imaginary axis that it is negative, we will be done. (C.18) is negative at $\alpha = 0$, so there exists a neighborhood of the imaginary axis, such that complex nonconjugate roots may be ruled out.

The roots on the left-hand side of (C.18) are

$$\alpha^2 = 1 - \frac{1}{16}\left(\frac{n}{c}\right)^2 = \frac{1}{16}(5-\gamma^2)(3+\gamma^2), \tag{C.19}$$

$$\alpha^2 = \frac{1}{2} + \frac{1}{4}\frac{n}{c} + \frac{1}{2}\sqrt{1-\frac{n}{c}} = \frac{1}{4}(3-\gamma)(1+\gamma), \tag{C.20}$$

$$\alpha^2 = \frac{1}{2} + \frac{1}{4}\frac{n}{c} - \frac{1}{2}\sqrt{1-\frac{n}{c}} = \frac{1}{4}(1-\gamma)(3+\gamma). \tag{C.21}$$

These are positive for $\gamma \in [0,1]$. It may be checked that (C.21) is the smallest for all γ. Hence the root of (C.18) such that μ will be closest to the imaginary axis is

$$\alpha = -\frac{1}{2}\sqrt{(1-\gamma)(3+\gamma)}. \tag{C.22}$$

α larger than (C.22) and less than zero will yield a λ that does not have nonconjugate complex roots.

Given λ, if $\mathcal{P}(\mu;\lambda)$ is to have two roots of same real part, α, but differing imaginary part, then, by trying any of the last four roots for β_1 (both positive and negative square roots of (C.16) and (C.17)), in (C.4) the real part of λ and α are related by

$$\Re\lambda = c\alpha\left(2 - \frac{(n/c)}{4(1-\alpha^2)+(n/c)}\right). \tag{C.23}$$

Note that

$$\frac{d\Re\lambda}{d\alpha} = c\left(\frac{32(1-\alpha^2)^2 + 12(n/c) - 20\alpha^2(n/c) + (n/c)^2}{\left(4(1-\alpha^2)+(c/n)\right)^2}\right) \tag{C.24}$$

and the derivative has one negative root with $|\alpha| < 1$ at

$$\alpha = -\sqrt{1 + \frac{5}{16}\frac{n}{c} - \frac{1}{16}\sqrt{64\frac{n}{c} + 17\left(\frac{n}{c}\right)^2}}. \tag{C.25}$$

Comparing (C.22) with (C.25), at $\gamma = 1$

$$\text{(C.25)} = -1 < 0 = \text{(C.22)},$$

and at $\gamma = 0$

$$\text{(C.25)} = -\sqrt{\frac{3}{4}} = \text{(C.22)}.$$

In addition, one may check that (C.25) is increasing in $\gamma \in [0,1]$, while (C.22) is decreasing on the same interval. Therefore

$$-\sqrt{1 + \frac{5}{16}\frac{n}{c} - \frac{1}{16}\sqrt{64\frac{n}{c} + 17\left(\frac{n}{c}\right)^2}} \le -\frac{1}{2}\sqrt{(1-\gamma)(3+\gamma)}$$

for all $\gamma \in [0,1]$.

For $\Re\mu = \alpha$ in

$$\left(-\frac{1}{2}\sqrt{(1-\gamma)(3+\gamma)}, 0\right) \tag{C.26}$$

(C.23) will be an increasing function in α, and (C.24) is positive at $\alpha = 0$. On the interval (C.26), the mapping is invertible and its image is

$$\left(-\frac{1}{4}c\sqrt{1-\gamma}\,(\gamma+3)^{3/2}, 0\right). \tag{C.27}$$

Therefore if λ has real part in the interval (C.27), and $\mathcal{P}(\mu;\lambda)$ is to have nonconjugate complex roots, α must lie in (C.26). But such an α violates (C.18), and we may conclude that there are no such roots. Letting $-\lambda_0$ denote the $\Re\lambda$ value at the left end point of (C.27),

$$\lambda_0 = \frac{c}{4}\sqrt{1-\gamma}\,(\gamma+3)^{3/2} = \frac{n}{4(1-\gamma^2)}\sqrt{1-\gamma}\,(\gamma+3)^{3/2}. \tag{C.28}$$

Hence for $\Re\lambda > -\lambda_0$, nonconjugate complex roots are not possible for $P(\mu)$.

Appendix D. The zero eigenvalue. In section 3.5, $\lambda = 0$ was identified as an eigenvalue of multiplicity at least two. Using the Evans function, the order of this eigenvalue may be related to the slope with respect to c of the invariant functional $\mathcal{N}[\phi_c]$.

Using the framework from section 3.3.2, set

$$\mathbf{y}^+ = \begin{pmatrix} \phi_c^{-m}\partial_x\phi_c \\ \partial_x\left(\phi_c^{-m}\partial_x\phi_c\right) \\ 0 \end{pmatrix}, \tag{D.1}$$

$$\mathbf{z}^- = \left(c\partial_x\left(\phi_c^{-m}\partial_x\phi_c\right), \quad -c\phi_c^{-m}\partial_x\phi_c, \quad -\int_{-\infty}^{x}\phi_c^{-n-m}\partial_x\phi_c\right). \tag{D.2}$$

These are solutions to the dynamical systems

$$\dot{\mathbf{y}} = B(x,\lambda=0,\gamma)\mathbf{y} \quad \text{and} \quad \dot{\mathbf{z}} = -\mathbf{z}B(x,\lambda=0,\gamma).$$

Here, $\mu_1 = -\gamma$. Employing the notation and formulation of the Evans function of [27], by [27, Proposition 1.6, parts 2 and 3], since

$$\mathbf{y}^+ = O(e^{-\gamma x}) \quad \text{as } x \to +\infty \quad \text{and} \quad \mathbf{z}^- = O(e^{\gamma x}) \quad \text{as } x \to -\infty,$$

$\mathbf{y}^+$ and $\mathbf{z}^-$ are scalar multiples of ζ^+ and η^-, respectively. ζ^+ and η^- are the solutions of the dynamical systems satisfying

$$\zeta^+ e^{\gamma x} \to \mathbf{v}^+ \quad \text{as } x \to +\infty \quad \text{and} \quad \eta^- e^{-\gamma x} \to \mathbf{w}^- \quad \text{as } x \to -\infty,$$

with

$$B^\infty\mathbf{v}^+ = -\gamma\mathbf{v}^+, \qquad \mathbf{w}^-B^\infty = -\gamma\mathbf{w}^-,$$

$$\mathbf{v}^+ = (1, \quad -\gamma, \quad 0)^T, \qquad \mathbf{w}^- = \left(2c\gamma^2\right)^{-1}\left(c\gamma^2, \quad -c\gamma, \quad -1\right),$$

allowing us to define the Evans function as

$$D(\lambda = 0) = \eta^-(x, \lambda = 0) \cdot \zeta^+(x, \lambda = 0).$$

From the properties of the solitary waves, discussed in section 2.2, there exists $\beta > 0$ such that

$$\phi_c^{-m} \partial_x \phi_c e^{\gamma x} \to -\beta \text{ as } x \to +\infty,$$

and hence

$$\zeta^+ = \frac{1}{-\beta} \mathbf{y}^+ \quad \text{and} \quad \eta^- = \frac{1}{-2\gamma c \beta} \mathbf{z}^-.$$

$D(0) = 0$ by inspection. From [27], the derivative of the Evans function for a system akin to ours is[2]

$$\partial_\lambda D(\lambda) = -\int_{-\infty}^{\infty} \eta^-(x, \lambda) \partial_\lambda \left[B(x, \lambda) - \mu_1(\lambda) I \right] \zeta^+(x, \lambda) dx. \tag{D.3}$$

It is then trivial to compute that $\partial_\lambda D(0) = 0$, since the integrand is an odd function.

Taking the derivative of (D.3) at $\lambda = 0$ gives an equation for $\partial_\lambda^2 D(0)$,

$$\partial_\lambda^2 D(0) = -\int_{-\infty}^{\infty} \eta_\lambda^- B_\lambda \zeta^+ dx - \int_{-\infty}^{\infty} \eta^- B_\lambda \zeta_\lambda^+ dx. \tag{D.4}$$

ζ_λ^+ and η_λ^- satisfy the ODEs

$$\dot{\mathbf{y}}_\lambda = B \mathbf{y}_\lambda + B_\lambda \mathbf{y} \quad \text{and} \quad \dot{\mathbf{z}}_\lambda = -\mathbf{z}_\lambda B - \mathbf{z} B_\lambda.$$

These problems are associated with the derivatives with respect to λ of (3.24) and (3.30) at $\lambda = 0$,

$$\partial_x L_c Y_\lambda = \left[I - \partial_x \left(\phi_c^n \partial_x \left(\phi_c^{-m} \cdot \right) \right) \right] Y, \tag{D.5}$$

$$-L_c^\star \partial_x Z_\lambda = \left[I - \phi_c^{-m} \partial_x \left(\phi_c^n \partial_x \cdot \right) \right] Z, \tag{D.6}$$

through the identifications

$$\begin{aligned} y_\lambda^{(1)} &= \phi_c^{-m} Y_\lambda, & z_\lambda^{(1)} &= c \partial_x \left(\phi_c^n \partial_x Z_\lambda \right), \\ y_\lambda^{(2)} &= \partial_x \left(\phi_c^{-m} Y_\lambda \right), & z_\lambda^{(2)} &= -c \phi_c^n \partial_x Z_\lambda, \\ y_\lambda^{(3)} &= L_c Y_\lambda + \phi_c^n \partial_x \left(\phi_c^{-m} Y \right), & z_\lambda^{(3)} &= -Z_\lambda. \end{aligned}$$

For $\lambda = 0$, (3.24), (3.30), (D.5), and (D.6) are related to the generalized kernels of A and $A^\star$:

$$\begin{aligned} Y &= \partial_x \phi_c, & Y_\lambda &= -\partial_c \phi_c, \\ Z &= \int_{-\infty}^{x} \frac{\partial_x \phi_c}{\phi_c^{n+m}} dx, & Z_\lambda &= -\int_{-\infty}^{x} \left(L_c^\star \right)^{-1} \left[I - \phi_c^{-m} \partial_x \left(\phi_c^n \partial_x \cdot \right) \right] \int_{-\infty}^{x} \frac{\partial_x \phi_c}{\phi_c^{n+m}} dx. \end{aligned}$$

[2] In particular, a system for which the matrix B has the same limit at $\pm\infty$.

Then

$$\zeta_\lambda^+ = \frac{1}{-\beta}\left(-\phi_c^{-m}\partial_c\phi_c, \quad -\partial_x\left(\phi_c^{-m}\partial_c\phi_c\right), \quad -L_c\partial_c\phi_c + \phi_c^n\partial_x\left(\phi_c^{-m}\partial_x\phi_c\right)\right)^T,$$

$$\eta_\lambda^- = \frac{1}{2\gamma c\beta}\left(c\partial_x\left(\phi_c^n\partial_x Z_\lambda\right), \quad -c\phi_c^n\partial_x Z_\lambda, \quad -Z_\lambda\right).$$

Finally, we compute

$$\partial_\lambda^2 D(0) = \frac{1}{c\gamma\beta^2}\partial_c\mathcal{N}[\phi_c]. \tag{D.7}$$

Acknowledgments. We thank Marc Spiegelman for his helpful comments and support, in addition to his contributions appearing in [38]. We have also benefited from discussions with Professor J. L. Bona and Professor P. Rosenau.

REFERENCES

[1] E. Aharonov, J. A. Whitehead, P. B. Kelemen, and M. Spiegelman, *Channeling instability of upwelling melt in the mantle*, J. Geophys. Res., 100 (1995), pp. 20433–20450.
[2] V. Barcilon and O. M. Lovera, *Solitary waves in magma dynamics*, J. Fluid Mech., 204 (1989), pp. 121–133.
[3] V. Barcilon and F. M. Richter, *Non-linear waves in compacting media*, J. Fluid Mech., 165 (1986), pp. 429–448.
[4] T. B. Benjamin, *The stability of solitary waves*, Proc. Roy. Soc. London Ser. A, 328 (1972), pp. 153–183.
[5] D. Bercovici, Y. Ricard, and G. Schubert, *A two-phase model for compaction and damage 1. General theory*, J. Geophys. Res., 106 (2001), pp. 8887–8906.
[6] M. S. Berger, *Nonlinearity and Functional Analysis*, Academic Press, New York, 1977.
[7] J. L. Bona, *personal correspondence*.
[8] J. Bona, *On the stability theory of solitary waves*, Proc. Roy. Soc. London Ser. A, 344 (1975), pp. 363–374.
[9] J. L. Bona and Y. A. Li, *Decay and analyticity of solitary waves*, J. Math. Pures Appl., 76 (1997), pp. 377–430.
[10] D. E. Edmunds and W. D. Evans, *Spectral Theory and Differential Operators*, Oxford Math. Monogr., Clarendon Press, Oxford University Press, New York, 1987.
[11] K. El Dika, *Asymptotic stability of solitary waves for the Benjamin-Bona-Mahony equation*, C. R. Math. Acad. Sci. Paris, 337 (2003), pp. 649–652.
[12] G. Hirth and D. L. Kohlstedt, *Experimental constraints on the dynamics of the partially molten upper mantle* 2: *Deformation in the dislocation creep regime*, J. Geophys. Res., 100 (1995), pp. 15441–15052.
[13] G. Hirth and D. L. Kohlstedt, *Experimental constraints on the dynamics of the partially molten upper mantle: Deformation in the diffusion creep regime*, J. Geophys. Res., 100 (1995), pp. 1981–2002.
[14] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1995.
[15] R. F. Katz, M. Spiegelman, and B. Holtzman, *The dynamics of melt and shear localization in partially molten aggregates*, Nature, 442 (2006), pp. 676–679.
[16] K. Knopp, *Theory of Functions: Part* I, Dover, New York, 1996.
[17] K. Knopp, *Theory of Functions: Part* II, Dover, New York, 1996.
[18] Y. Martel and F. Merle, *A Liouville theorem for the critical generalized Korteweg-de Vries equation*, J. Math. Pures Appl., 79 (2000), pp. 339–425.
[19] Y. Martel and F. Merle, *Asymptotic stability of solitons of the subcritical gKdV equations revisited*, Nonlinearity, 18 (2005), pp. 55–80.
[20] D. McKenzie, *The generation and compaction of partially molten rock*, J. Petrology, 25 (1984), pp. 713–765.
[21] J. R. Miller and M. I. Weinstein, *Asymptotic stability of solitary waves for the regularized long-wave equation*, Comm. Pure Appl. Math., 49 (1996), pp. 399–441.
[22] T. Mizumachi, *Asymptotic stability of solitary wave solutions to the regularized long-wave equation*, J. Differential Equations, 200 (2004), pp. 312–341.

[23] M. Nakayama and D. P. Mason, *Rarefactive solitary waves in two-phase fluid flow of compacting media*, Wave Motion, 15 (1992), pp. 357–392.
[24] P. Olson and U. Christensen, *Solitary wave propagation in a fluid conduit within a viscous matrix*, J. Geophys. Res., 91 (1986), pp. 6367–6374.
[25] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
[26] R. L. Pego, P. Smereka, and M. I. Weinstein, *Oscillatory instability of traveling waves for a KdV-Burgers equation*, Phys. D, 67 (1993), pp. 45–65.
[27] R. L. Pego and M. I. Weinstein, *Eigenvalues, and instabilities of solitary waves*, Philos. Trans. Roy. Soc. London Ser. A, 340 (1992), pp. 47–94.
[28] R. L. Pego and M. I. Weinstein, *Asymptotic stability of solitary waves*, Comm. Math. Phys., 164 (1994), pp. 305–349.
[29] R. L. Pego and M. I. Weinstein, *Convective linear stability of solitary waves for Boussinesq equations*, Stud. Appl. Math., 99 (1997), pp. 311–375.
[30] J. Prüss, *On the spectrum of C_0-semigroups*, Trans. Amer. Math. Soc., 284 (1984), pp. 847–857.
[31] J. Renner, K. Viskupic, G. Hirth, and B. Evans, *Melt extraction from partially molten peridotites*, Geochemistry, Geophysics, Geosystems, 4 (2003), article 8606; doi:10.1029/2002GC000369.
[32] P. Rosenau, *On a model equation of traveling and stationary compactons*, Phys. Lett. A, 356 (2006), pp. 44–50.
[33] M. Schechter, *Spectra of Partial Differential Operators*, North-Holland, Amsterdam, 1971.
[34] M. Schechter, *Principles of Functional Analysis*, Grad. Stud. Math. 36, AMS, Providence, RI, 2001.
[35] D. R. Scott and D. J. Stevenson, *Magma solitons*, Geophys. Res. Lett., 11 (1984), pp. 1161–1164.
[36] D. R. Scott and D. J. Stevenson, *Magma ascent by porous flow*, J. Geophys. Res., 91 (1986), pp. 9283–9296.
[37] D. R. Scott, D. J. Stevenson, and J. A. Whitehead, *Observations of solitary waves in a viscously deformable pipe*, Nature, 319 (1986), pp. 759–761.
[38] G. Simpson, M. Spiegelman, and M. I. Weinstein, *Degenerate dispersive equations arising in the study of magma dynamics*, Nonlinearity, 20 (2007), pp. 21–49.
[39] G. Simpson, M. Spiegelman, and M. I. Weinstein, *Magma transport: Dynamics of fluid flow in viscously deformable porous media*, in preparation.
[40] G. Simpson, M. I. Weinstein, and P. Rosenau, *On a Hamiltonian PDE arising in magma dynamics*, Discrete Contin. Dyn. Syst. Ser. B, 10 (2008), pp. 903–924.
[41] M. Spiegelman, *Flow in deformable porous media. Part 1: Simple analysis*, J. Fluid Mech., 247 (1993), pp. 17–38.
[42] M. Spiegelman, *Flow in deformable porous media. Part 2: Numerical analysis*, J. Fluid Mech., 247 (1993), pp. 39–63.
[43] M. Spiegelman, *Linear analysis of melt band formation by simple shear*, Geochemistry, Geophysics, Geosystems, 4 (2003), pp. 1525–2027.
[44] M. Spiegelman, P. B. Kelemen, and E. Aharonov, *Causes and consequences of flow organization during melt transport: The reaction infiltration instability in compactible media*, J. Geophys. Res., 106 (2001), pp. 2061–2077; available online from http://www.ldeo.columbia.edu/∼mspieg/SolFlow/.
[45] D. A. Wark and E. B. Watson, *Grain-scale permeabilities of texturally equilibrated, monomineralic rocks*, Earth and Planetary Sci. Lett., 164 (1998), pp. 591–605.
[46] D. A. Wark, C. A. Williams, E. B. Watson, and J. D. Price, *Reassessment of pore shapes in microstructurally equilibrated rocks, with implications for permeability of the upper mantle*, J. Geophys. Res., 108 (2003).
[47] M. I. Weinstein, *Modulational stability of ground states of nonlinear Schrödinger equations*, SIAM J. Math. Anal., 16 (1985), pp. 472–491.
[48] M. I. Weinstein, *Lyapunov stability of ground states of nonlinear dispersive evolution equations*, Comm. Pure Appl. Math., 39 (1986), pp. 51–68.
[49] J. A. Whitehead, *A laboratory demonstration of solitons using a vertical watery conduit in syrup*, Amer. J. Phys., 55 (1987), pp. 998–1003.
[50] J. A. Whitehead and K. R. Helfrich, *The Korteweg-de Vries equation from laboratory conduit and magma migration equations*, Geophys. Res. Lett., 13 (1986), pp. 545–546.
[51] C. Wiggins and M. Spiegelman, *Magma migration and magmatic solitary waves in 3-d*, Geophys. Res. Lett., 22 (1995), pp. 1289–1292.
[52] W. Zhu and G. Hirth, *A network model for permeability in partially molten rocks*, Earth and Planetary Sci. Lett., 212 (2003), pp. 407–416.

SIAM J. MATH. ANAL.
Vol. 40, No. 4, pp. 1392–1418

ENERGY TRANSPORT BY ACOUSTIC MODES OF HARMONIC LATTICES*

LISA HARRIS†, JANI LUKKARINEN‡, STEFAN TEUFEL§, AND FLORIAN THEIL†

Abstract. We study the large scale evolution of a scalar lattice excitation u which satisfies a discrete wave equation in three dimensions, $\ddot{u}_t(\gamma) = -\sum_{\gamma'} \alpha(\gamma-\gamma')u_t(\gamma')$, where $\gamma, \gamma' \in \mathbb{Z}^3$ are lattice sites. We assume that the dispersion relation ω associated to the elastic coupling constants $\alpha(\gamma-\gamma')$ is acoustic; i.e., it has a singularity of the type $|k|$ near the vanishing wave vector, $k=0$. To derive equations describing the macroscopic energy transport, we employ a related multiscale Wigner transform and a scale parameter $\varepsilon > 0$. The spatial and temporal scales of the Wigner transform are related to the corresponding lattice parameters via a scaling by ε. In the continuum limit, which is achieved by sending the parameter ε to 0, the Wigner transform disintegrates into three different limit objects: the Wigner transform of a rescaled weak-L^2 limit, an H-measure, and a Wigner measure. The first two provide the finer resolution of the energy concentrating at $k=0$ so that a set of closed evolution equations may arise. We demonstrate that these three limit objects satisfy a set of decoupled transport equations: a wave equation for the weak limit, a geometric optics transport equation for the H-measure limit, and a dispersive transport equation for the standard limiting Wigner measure. This yields a complete characterization of macroscopic energy transport in harmonic lattices with regular acoustic dispersion relations.

Key words. microlocal analysis, multiscale methods, homogenization, discrete wave equations

AMS subject classifications. 70J30, 74Q15, 37K60

DOI. 10.1137/070699184

1. Introduction.

The energy transport by atomic oscillations in crystalline solids is a central problem in solid state physics. To the first order approximation, the oscillations can be described by a discrete wave equation

$$\ddot{u}_t(\gamma) = -\sum_{\gamma'} \alpha(\gamma-\gamma')u_t(\gamma') \tag{1.1}$$

where $u(\gamma) \in \mathbb{R}$ is composed of the displacements of the crystal atoms from their equilibrium position, as will be discussed in section 1.1. To analyze physically relevant properties of the crystal, such as its thermal conductivity, we first need to understand how energy is transported within the crystal via purely harmonic vibrations. Such transport properties are determined by the dispersion relation ω of the crystal, here $\omega(k) = \sqrt{\hat{\alpha}(k)}$, the "hat" denoting a discrete Fourier transform. If ω is not smooth, then depending on the wavelength different types of continuum energy transport equations can arise. The different contributions to the continuum energy are described by different limit objects, so-called Wigner measures, or microlocal defect measures.

*Received by the editors August 3, 2007; accepted for publication (in revised form) June 24, 2008; published electronically November 5, 2008.
http://www.siam.org/journals/sima/40-4/69918.html

†Mathematics Institute, Warwick University, Coventry, CV4 7AL, UK (l.c.harris@warwick.ac.uk, f.theil@warwick.ac.uk).

‡Zentrum Mathematik, Technische Universität München, Boltzmannstr. 3, D-85747 Garching, Germany, and Department of Mathematics and Statistics, University of Helsinki, P.O. Box 68, FI-00014 Helsingin yliopisto, Finland (jani.lukkarinen@helsinki.fi). This author was supported by the Deutsche Forschungsgemeinschaft (DFG) projects Sp 181/19-1 and Sp 181/19-2 and by the Academy of Finland.

§Mathematisches Institut, Auf der Morgenstelle 10, 72076 Tübingen, Germany (stefan.teufel@uni-tuebingen.de).

These concepts were introduced by Tartar [23] and Gérard [10] to analyze the weak limits of certain nonlinear quantities, the energy density being one of them. Two-scale Wigner measures of the type we use here were first introduced independently by Fermanian Kammerer [5, 6] and by Nier [18].

Our main interest is to characterize the macroscopic evolution of the energy density. The starting point of our mathematical analysis is the *Wigner transform*, which can be interpreted as a "wavenumber resolved" energy density. Let us leave the details for section 2.1, and only summarize the main findings here. Let $v_t = \dot{u}_t$. We can then construct a complex field $\psi_t \in \ell_2(\mathbb{Z}^3)$, corresponding to a normal mode of the oscillations, out of the fields u_t and v_t. On the other hand, for any given $\psi \in \ell_2(\mathbb{Z}^3)$ and $\varepsilon > 0$, the lattice Wigner transform $W^\varepsilon = W^\varepsilon[\psi]$ allows defining a corresponding "energy" density, $e^\varepsilon = e^\varepsilon[\psi]$, by the formula

$$e^\varepsilon(x) = \int_{\mathbb{T}^3} \mathrm{d}k\, W^\varepsilon(x,k), \tag{1.2}$$

where $\varepsilon > 0$ denotes the "lattice spacing" and $x \in \mathbb{R}^3$ is a variable which interpolates between the points of the scaled lattice $\varepsilon\mathbb{Z}^3$. Here $\mathbb{T}^3$ denotes the 3-torus, and we identify $\mathbb{T}^3 = \mathbb{R}^3/\mathbb{Z}^3$. Then $e^\varepsilon(x,t) = e^\varepsilon[\psi_{t/\varepsilon}](x)$ can be considered to be a proper energy density at a macroscopic time t, as it satisfies $\int_{\mathbb{R}^3} \mathrm{d}x\, e^\varepsilon(x,t) = H(u_{t/\varepsilon}, v_{t/\varepsilon})$, where H denotes the Hamiltonian related to (1.1).

In general, a limit of a sequence of scaled Wigner transforms, $(W^\varepsilon[\psi^\varepsilon])$ with ε tending to 0, is given by a nonnegative Radon measure $\mu \in M_+(\mathbb{R}^3 \times \mathbb{T}^3)$. In the setup considered here, we have for a given time t a sequence of normal mode fields $\psi_{t/\varepsilon}$ which would then yield a family of nonnegative Radon measures μ_t as limit points of the corresponding Wigner transforms. However, for these limit measures to form a useful approximation of the original dynamical system, μ_t should also satisfy an autonomous evolution equation. This is typically not possible if the initial measure concentrates on the singular set of ω, i.e., to the points where ω is not smooth.

Here we will augment the standard Wigner transform scheme to encompass the most common type of singularity encountered in solid state physics: acoustic singularities, with $\omega(k)$ behaving like $|k|$ near $k = 0$. Such modes occur in general within crystal models with short range interactions, and they are particularly important as they are responsible for sound propagation in the crystal. With some effort, in the spirit of the results proven in [6], it is likely that these results could be extended to cover more complicated singular sets of ω, but we do not consider such generalizations here. Our result can also be seen as a generalization of the analysis in [17]. There it is shown that μ_t can be computed from μ_0 by solving a dispersive linear transport equation, provided that μ_0 does not concentrate on wavenumbers k where the dispersion relation ω is not $C^{(1)}$.

For acoustic modes, the dispersion relation fails to be $C^{(1)}$ at $k = 0$, and it is necessary to resolve finer details of the solutions near the singular point, whenever a concentration at $k = 0$ is possible. This will be accomplished here by introducing a second scale to the lattice Wigner transform. The multiscale Wigner transform, defined in section 2.1, maps a field ψ into a distribution $W^\varepsilon[\psi](x,k,q)$ where the new parameter $q \in \mathbb{R}^3$ resolves the energy density for wavenumbers of the order of ε. We consider a sequence of initial conditions such that the corresponding initial fields ψ_0^ε are bounded and tight in $\ell_2(\varepsilon\mathbb{Z}^3)$. We prove in Theorem 3.2 that then for all $t \in \mathbb{R}$ there are two measures, a Wigner measure μ_t on $\mathbb{R}^3 \times \mathbb{T}^3_*$, $\mathbb{T}^3_* = \mathbb{T}^3 \setminus \{0\}$, and an H-measure μ_t^{H} on $\mathbb{R}^3 \times S^2$, as well as an L^2-function ϕ_t such that $W^\varepsilon[\psi^\varepsilon_{t/\varepsilon}]$ converges

along a subsequence to a limit determined by $(\mu_t, \mu_t^{\mathrm{H}}, \phi_t)$. The subsequence can be chosen independently of t, and it will only be relevant for determining the limit of the initial data, that is, $(\mu_0, \mu_0^{\mathrm{H}}, \phi_0)$. For all other times $t \in \mathbb{R}$, the measures $\mu_t, \mu_t^{\mathrm{H}}$ and the L^2-function ϕ_t can be determined using the transport equations

$$\partial_t \mu_t(x,k) + \tfrac{1}{2\pi} \nabla\omega(k) \cdot \nabla_x \mu_t(x,k) = 0, \qquad k \in \mathbb{T}^3_*\,, \tag{1.3}$$

$$\partial_t \mu_t^{\mathrm{H}}(x,q) + \tfrac{1}{2\pi} \nabla\omega_0(q) \cdot \nabla_x \mu_t^{\mathrm{H}}(x,q) = 0, \quad q \in S^2\,, \tag{1.4}$$

$$\partial_t^2 \phi_t(x) = \mathrm{div}(\tfrac{1}{(2\pi)^2} A_0 \nabla \phi_t(x)), \tag{1.5}$$

where $x \in \mathbb{R}^3$, together with the initial conditions

$$\mu_t|_{t=0} = \mu_0, \quad \mu_t^{\mathrm{H}}\big|_{t=0} = \mu_0^{\mathrm{H}}, \quad \phi_t|_{t=0} = \phi_0, \quad \partial_t \hat{\phi}_t(p)|_{t=0} = -i\omega_0(p)\hat{\phi}_0(p), \quad p \in \mathbb{R}^3\,, \tag{1.6}$$

given by the limit objects $(\mu_0, \mu_0^{\mathrm{H}}, \phi_0)$ at $t = 0$. Here the constant matrix A_0 and the function ω_0 are determined by the Hessian of the square of the dispersion relation ω (defined in (2.6)) at $k = 0$, explicitly

$$A_0 = \frac{1}{2} D^2 \omega^2(0), \qquad \omega_0(p) = \sqrt{p \cdot A_0 p}\,. \tag{1.7}$$

It also follows from our analysis that the sum of the energies related to μ_t, μ_t^{H}, and ϕ_t is a constant of motion and equals the limiting value of the total energy of the initial excitations. This shows that no energy is lost in the taking of the continuum limit.

Equation (1.3) describes the propagation of energy along the harmonic lattice with the group velocity $\nabla\omega(k)/(2\pi)$. Equation (1.5) is a wave equation describing the evolution of macroscopic fluctuations. Equation (1.4) is usually known under the name "geometric optics" and describes the evolution of macroscopic fluctuations whose wavelength is much longer than the lattice spacing ε and much smaller than 1, the wavelength of the fluctuations resolved by ϕ_t. The surprising feature about acoustic modes is the separate evolution equation for the macroscopic perturbations ϕ_t. As will be apparent in Theorem 3.2, but as also indicated by the fact that the evolution equation for ϕ_t involves only the Hessian of ω^2 at $k = 0$, its evolution is determined by the second, fine-resolution scale introduced to the Wigner transform. This is in contradistinction to the case of ω which is smooth, as then the evolution of the energy density would not depend on the second scaling parameter.

The Wigner transform, or the Wigner function, was originally introduced to study semiclassical behavior in quantum mechanics, but it has been proven to be a useful tool in studying large scale behavior of wave equations as well [12, 19]. In particular, the method of calculating continuous, macroscopic energy by finding the limit object of a sequence of energies e^{ε} on rescaled lattice models is one that has been widely used and justified, for example, in [9, 14, 17]. In [17], the Wigner transform of the normal modes is employed in solving the macroscopic transport of energy in the above harmonic systems for deterministic initial data. The same system is considered in [3] with random initial data and in a larger function space, excluding, however, the type of concentration effects we study here. In [6] two-scale Wigner measures of the same type we use here were introduced for the study of concentration effects near shock hypersurfaces for the heat equation. The precise connection to the results in [5, 6] will be explained in Remark 3.6 in section 3.

In [8] spatially homogeneous Wigner measures are employed to study energy asymptotics in static systems. Both the method (which was developed before the

theory of 2-microlocal measures became available) and the results are in spirit very similar to ours. Using specific profile functions, three limit objects are constructed which capture the asymptotic behavior of the energy along bounded sequences. Only Fourier methods are used in [8] and therefore the spatial distribution of the energy cannot be characterized.

The main use of the Wigner transform is that, unlike the energy density e^ε itself, it contains enough information to satisfy a closed evolution equation in the limit $\varepsilon \to 0$. Indeed, it was shown in [17] that as long as there is no concentration on the singular set of the dispersion relation faster than $\varepsilon^{1/2}$, the Wigner transform of the time-evolved state vector $\psi_{t/\varepsilon}$ converges to a limit measure $\tilde{\mu}_t$ on $\mathbb{R} \times \mathbb{K}$. Here $\mathbb{K}$ is a suitable compactification of $\mathbb{T}^3 \setminus \mathbb{S}$, $\mathbb{S}$ being the singular set, which allows a continuous extension of the group velocity $\nabla\omega$. The measure $\tilde{\mu}_t$ is then proven to satisfy the transport equation

$$\partial_t \tilde{\mu}_t(x,k) + \tfrac{1}{2\pi} \nabla\omega(k) \cdot \nabla_x \tilde{\mu}_t(x,k) = 0. \tag{1.8}$$

However, the above assumption about the rate of convergence excludes macroscopic variations of the initial data in the case of acoustic singularities, and enforces $\phi_t = 0$ in the present terminology. Separately, it was also shown in [17] that if the weak limit of the rescaled initial data exists, then the limit satisfies a continuum wave equation. However, a combination of these results was not possible, leaving open the exact effect the continuum wave equation will have on the energy density. In [15] it is shown that the Wigner measures generated by sequences of solutions of discrete wave equations satisfy transport equations; energy conservation along the sequence is not discussed.

In this paper we will show how to overcome the above difficulties in a physically relevant class of models with a singular dispersion relation. Our main additional contribution to the results of the above-mentioned references is to introduce a lattice version of the multiscale Wigner transform and to employ this in solving the macroscopic evolution of the energy density in the presence of an acoustic singularity. In section 2 we will first present the microscopic dynamical model in detail. In section 2.1 we will define the Wigner transform, and discuss its relation to the energy of the microscopic lattice model. The main results will be presented in section 3.

1.1. Relation with solid state physics. A crystal in solid state physics is a state of matter in which the atoms retain a nearly perfect periodic structure over macroscopic times. The Hamiltonian model used for the time evolution in such a crystal is, to the first order accuracy, harmonic. If we assume that each periodic cell of the idealized perfectly periodic crystal structure contains n atoms, then we can form a vector $q(\gamma) \in \mathbb{R}^{3n}$ out of the displacements of the atoms in the periodic cell labeled by $\gamma \in \mathbb{Z}^3$. The (classical) Hamiltonian equations of motion of this harmonic model are then

$$\dot{q}_i(\gamma,t) = \frac{1}{m_i} p_i(\gamma,t), \qquad \dot{p}_i(\gamma,t) = -\sum_{\gamma',i'} \mathbb{A}(\gamma-\gamma')_{i,i'} q_{i'}(\gamma',t), \tag{1.9}$$

where $\gamma \in \mathbb{Z}^3$, $i = 1,\dots,3n$, and m_i denotes the mass of the atom whose displacement q_i measures.

By the change of variables to $\tilde{q}_i(\gamma) = m_i^{\frac{1}{2}} q_i(\gamma)$, $\tilde{p}_i(\gamma) = m_i^{-\frac{1}{2}} p_i(\gamma)$, these equations can be transformed into a standard form whose force matrix is given by $\tilde{\mathbb{A}}(\gamma)_{i,i'} = m_i^{-1/2} \mathbb{A}(\gamma)_{i,i'} m_{i'}^{-1/2}$. The standard form equations can then be solved by Fourier transform, and a diagonalization of the remaining multiplicative evolution equations

decomposes the $3n$ vector degrees of freedom into independent *normal modes*, called *phonons* in solid state physics. Each normal mode is a complex scalar field on the crystal lattice, and its time evolution is unitary and uniquely determined by the corresponding dispersion relation $\omega_i(k)$ on $\mathbb{T}^3$. More details on the related mathematical issues can be found in [3, 17].

The complete decoupling of the evolution equations of the normal modes allows us to study a scalar, single-mode model and still retain a straightforward applicability of the results in the above, physically more relevant, vector models. As the above procedure is fairly standard and well covered in the references, we will not present the decomposition in any more detail here. However, to illustrate the matter, we have performed the equivalent steps for the scalar model (1.1) in the beginning of section 2.

In solid state physics, the normal modes are divided into *optical* and *acoustic* depending on their regularity at $k = 0$: if the dispersion relation is regular at $k = 0$, then the mode is called optical, and if it behaves as $|k|$ at $k = 0$, the mode is called acoustic. The latter name arises as these modes are believed to be responsible for the propagation of sound waves in the crystal. Acoustic dispersion relations arise generally from atomistic Hamiltonians with short range interactions, and crystal models are typically expected to have three of them, related to the translation invariance of the microscopic forces.

Finally, let us remark that the discrete *linear* wave equation alone does not suffice to determine the physically relevant properties of the crystal, such as its thermal conductivity. However, it forms the basis for perturbative treatments which can address such questions. We refer to [1, 20, 21, 22, 25] for further details on the physical aspects of the topic, and to [2] for a review about related open problems.

2. The microscopic model. There are two mathematically equivalent descriptions of harmonic crystals. On the one hand, one can work with the Hamiltonian equations of motion and analyze the properties of the solutions. Anharmonic crystals can then be discussed in the same manner. On the other hand, one can employ the Fourier transform and the linearity to condense the Hamiltonian into the dispersion relation. Although the second approach leads to an immediate solution of the harmonic system, it cannot be used directly to analyze nonlinear models.

We will show in this chapter how the first approach reduces to the second one for harmonic lattices. Then we demonstrate how the Wigner transform can be employed in the analysis of the time evolution of the energy density of the Hamiltonian description.

We assume that the scalar excitation $u_t(\gamma)$, $\gamma \in \mathbb{Z}^3$, satisfies the discrete wave equation

$$\frac{\partial^2}{\partial t^2} u_t(\gamma) = -\sum_{\gamma' \in \mathbb{Z}^3} \alpha(\gamma - \gamma') u_t(\gamma') \tag{2.1}$$

with initial data $(u_t|_{t=0}, v_t|_{t=0}) \in X = \ell_2 \times \ell_2$, where v_t denotes the velocity field, $v_t = \partial_t u_t$. The numbers $\alpha(\gamma - \gamma')$ are the elastic coupling constants between the sites γ and γ'. We assume that α is real and symmetric ($\alpha(-\gamma) = \alpha(\gamma)$). Clearly, system (2.1) can be written in a Hamiltonian form and the energy

$$H(u, v) = \frac{1}{2}\Big(\sum_{\gamma \in \mathbb{Z}^3} |v(\gamma)|^2 + \sum_{\gamma \in \mathbb{Z}^3} \Big[\sum_{\gamma' \in \mathbb{Z}^3} u(\gamma)\alpha(\gamma - \gamma')u(\gamma') \Big] \Big) \tag{2.2}$$

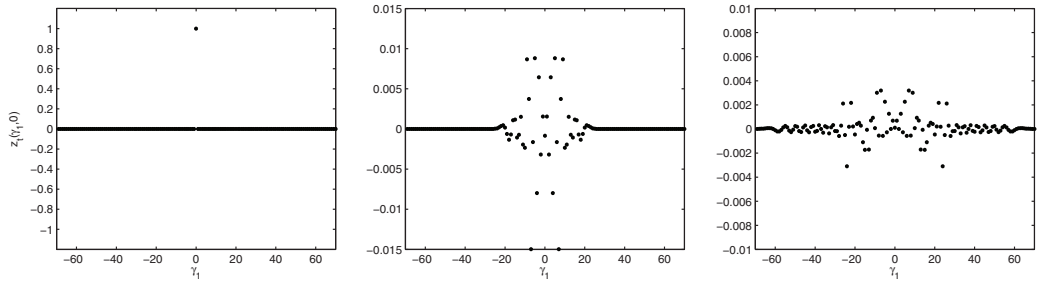


FIG. 2.1. *Values of $u_t(\gamma)$ along the axis $\gamma_2 = \gamma_3 = 0$ for $t = 0$, $t = 0.1/\varepsilon$, $t = 0.9/\varepsilon$ with $\varepsilon = \frac{1}{7} * 10^{-1}$. The evolution is given by* (2.1) *with the nearest neighbor elastic couplings* (2.13)*, and the initial conditions are $u_{t=0}(0) = 1$, $u_{t=0}(\gamma) = 0$ for all $\gamma \neq 0$, and $\dot{u}_{t=0} \equiv 0$.*

is constant along solutions. Depending on the initial conditions, the solutions of system (2.1) may develop large scale oscillations which carry a finite amount of energy; cf. Figure 2.1, where snapshots of u at several times are plotted.

Since system (2.1) is linear and invariant under discrete translations, we can write the solutions in a closed form using the Fourier transform.

DEFINITION 2.1. *We define the Fourier transform $\ell_2(\mathbb{Z}^3) \to L^2(\mathbb{T}^3)$ by extending*

$$(\mathcal{F}_{\gamma\to k}\psi)(k) = \hat{\psi}(k) = \sum_{\gamma\in\mathbb{Z}^3} \mathrm{e}^{-2\pi \mathrm{i} k\cdot\gamma}\psi(\gamma) \tag{2.3}$$

from ψ with finite support to all of $\ell_2(\mathbb{Z}^3)$. Here $\mathbb{T}^3 = \mathbb{R}^3/\mathbb{Z}^3$ denotes the unit 3-torus. The inverse transform is pointwise convergently defined by the integral

$$(\mathcal{F}_{k\to\gamma}\hat{\psi})(\gamma) = \int_{\mathbb{T}^3} \mathrm{d}k\, \mathrm{e}^{2\pi \mathrm{i} k\cdot\gamma}\hat{\psi}(k) = \psi(\gamma), \tag{2.4}$$

where the measure $\mathrm{d}k$ is induced by the Lebesgue measure on the parameterization T^3 of $\mathbb{T}^3$, where $T = (-\frac{1}{2}, \frac{1}{2}]$. In particular, $\int_{\mathbb{T}^3} \mathrm{d}k = 1$.

From now on the notation T^3 will refer to the above explicit parameterization of the torus $\mathbb{T}^3$. Occasionally, integration over $k \in \mathbb{T}^3$ will be applied to functions $f(k)$ which are not periodic in k. In these cases, using the parameterization $k \in T^3$ is implicitly understood. The parameterization mapping $\mathbb{T}^3 \to T^3$ will be denoted by $k \mapsto (k \bmod T^3)$ whenever the use of the parameterization needs to be explicitly stressed.

If one applies the Fourier transform to the Hamiltonian equations of motion determined by $H(u, v)$, one obtains a simpler system equivalent to (2.1):

$$\frac{\partial}{\partial t}\begin{pmatrix}\hat{u}_t(k)\\ \hat{v}_t(k)\end{pmatrix} = \begin{pmatrix}0 & 1\\ -\omega^2(k) & 0\end{pmatrix}\begin{pmatrix}\hat{u}_t(k)\\ \hat{v}_t(k)\end{pmatrix}, \quad k \in \mathbb{T}^3\,. \tag{2.5}$$

The function $\omega : \mathbb{T}^3 \to \mathbb{R}$ is the dispersion relation and it is related to the Hamiltonian via the following formula:

$$\omega(k) = \sqrt{\hat{\alpha}(k)} = \sqrt{\sum_{\gamma\in\mathbb{Z}^3} \alpha(\gamma)\cos(2\pi\gamma\cdot k)}, \tag{2.6}$$

where we have employed the assumption $\alpha(\gamma) = \alpha(-\gamma)$. Since α is real and satisfies the above symmetry property, we find that ω is also real and symmetric, i.e., $\omega(k) = \omega(-k)$.

A diagonalization of the matrix on the right-hand side of (2.5) motivates combining the real scalar fields u and v into the two complex fields $\psi_\pm = \psi_\pm[u,v] \in \ell_2(\mathbb{Z}^3,\mathbb{C})$ defined by the formula

$$\hat{\psi}_\sigma(k) = \frac{1}{\sqrt{2}}\left(\omega(k)\hat{u}(k) + \mathrm{i}\sigma\hat{v}(k)\right), \quad \sigma \in \{\pm 1\},\ k \in \mathbb{T}^3\,. \tag{2.7}$$

For all $(u,v) \in X$, we clearly have $\hat{\psi}_\sigma \in L^2(\mathbb{T}^3)$, and thus $\psi_\sigma \in \ell_2(\mathbb{Z}^3)$. In addition, since $\omega(-k) = \omega(k)$, we also have $\psi_-(\gamma) = \overline{\psi_+(\gamma)}$ for all γ. The transformation can always be inverted by applying

$$\hat{u} = \frac{1}{\omega\sqrt{2}}(\hat{\psi}_+ + \hat{\psi}_-)\,, \quad \hat{v} = -\frac{\mathrm{i}}{\sqrt{2}}(\hat{\psi}_+ - \hat{\psi}_-)\,. \tag{2.8}$$

Given a solution u_t, $v_t = \dot{u}_t$ to (2.1), we then define the complex normal mode fields by the formula

$$\psi_\sigma(\gamma,t) = \psi_\sigma[u_t,v_t](\gamma), \quad \sigma \in \{\pm 1\},\ \gamma \in \mathbb{Z}^3,\ t \in \mathbb{R}\,. \tag{2.9}$$

To see that these fields are indeed normal modes of the harmonic system, we apply (2.5) and find the evolution equations

$$\frac{\partial}{\partial t}\begin{pmatrix}\hat{\psi}_+(k,t)\\ \hat{\psi}_-(k,t)\end{pmatrix} = -\mathrm{i}\begin{pmatrix}\omega(k) & 0\\ 0 & -\omega(k)\end{pmatrix}\begin{pmatrix}\hat{\psi}_+(k,t)\\ \hat{\psi}_-(k,t)\end{pmatrix}, \tag{2.10}$$

which are readily solved to yield for all $t \in \mathbb{R}$, $k \in \mathbb{R}^3$

$$\hat{\psi}_\pm(k,t) = \mathrm{e}^{\mp\mathrm{i}\omega(k)t}\hat{\psi}_\pm(k,0)\,. \tag{2.11}$$

These are exactly the two evolution equations corresponding to a "phonon" mode with a dispersion relation ω. The fields ψ_t mentioned in the introduction can now be identified with $\psi_+(\cdot,t)$ in the present notation.

After these reduction steps it is obvious that the dispersion relation ω fully determines the properties of the solutions. We will assume throughout this paper that ω is of the acoustic type in the following precise sense.

DEFINITION 2.2. *We call $\omega \in C(\mathbb{T}^3,[0,\infty))$ an* acoustic dispersion relation *if $\lambda = \omega^2$ satisfies the following:*

1. *$\lambda \in C^{(3)}(\mathbb{T}^3,[0,\infty))$.*
2. *$\lambda(0) = 0$, and the Hessian of λ is invertible at 0.*

A dispersion relation is called regular acoustic *if it is acoustic and $\lambda(k) > 0$ for $k \neq 0$. The 3×3-matrix A_0 is the Hessian of $\frac{1}{2}\lambda$ at $k = 0$ and $\omega_0(q) = \sqrt{q\cdot A_0 q}$.*

Let us briefly motivate why these assumptions should be satisfied quite generally. For instance, if α is exponentially decaying, $\lambda = \hat{\alpha}$ will be analytic and thus satisfies the first requirement. The stability condition $\lambda \geq 0$ also needs to be satisfied everywhere; otherwise, the harmonic system has exponentially increasing solutions and is a poor model for oscillations in a crystal. If $\lambda(0) \neq 0$, the dispersion relation is not singular at $k = 0$. Thus in order to have a singular dispersion relation, the only real restriction for exponentially decaying stable interactions is the assumption about invertibility of the Hessian. Our analysis could likely be extended to cover also more degenerate instances of λ, although this would require some additional effort. By analogous reasoning, acoustic normal modes—in the above sense—are seen to appear commonly also in the vector models discussed in section 1.1, at least for exponentially

decaying interactions. However, in the vector models level crossings can lead to additional singularities in the dispersion relations, which would require a separate study. Finally, let us remark that exponential decay is not essential, since only sufficient regularity of the Hessian of λ will be required in the proof. For instance, if there are $C, \delta > 0$ such that $|\alpha(\gamma)| \leq C|\gamma|^{-6-\delta}$, then $\lambda = \hat{\alpha}$ satisfies item 1 in Definition 2.2. This permits us to deal also with situations where α is obtained via linearization around an equilibrium of a nonlinear system with potentials decaying slightly faster than a Lennard–Jones potential.

A prototype for the kind of dispersion relations considered here is the dispersion relation of the nearest neighbor square lattice,

$$\omega_{\mathrm{nn}}(k) = \Big[\sum_{\nu=1}^{3} 2(1-\cos(2\pi k^{\nu}))\Big]^{\frac{1}{2}}. \tag{2.12}$$

This is clearly a regular acoustic dispersion relation, in the sense of Definition 2.2, and for it A_0 is proportional to a unit matrix and $\omega_0(q) = 2\pi|q|$. The corresponding elastic couplings are given by $\alpha_{\mathrm{nn}}(\gamma' - \gamma) = -\Delta_{\gamma'\gamma}$, where Δ is the discrete Laplacian of the square lattice. Explicitly,

$$\alpha_{\mathrm{nn}}(\gamma) = \begin{cases} 6 & \text{if } \gamma = 0, \\ -1 & \text{if } |\gamma| = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{2.13}$$

To allow the creation of macroscopic oscillations we work with sequences of initial conditions that depend on the scaling parameter $\varepsilon > 0$ and consider the asymptotic behavior of the solutions as ε tends to 0.

2.1. Energy density and the lattice Wigner transform. From now on we will focus on analyzing asymptotic behavior of the fields $\psi_{\pm}$ as ε tends to 0. As is carefully discussed in [17], generalizing the definitions of the energy density and of the Wigner transform to the discrete setting is not completely obvious. In an attempt to minimize unnecessary repetition of certain basic results related to Wigner transforms, we will resort here to the definitions used in [14], which will allow us to rely on the properties proven in Appendix B of that reference. However, we wish to keep in mind that this choice might not be optimal for all purposes, and we refer the interested reader to the discussion and references found in [16, 17, 24] for further possibilities.

We employ here the definition that for any state $(u, v) \in X$, its energy density, $e^{\varepsilon} = e^{\varepsilon}[u, v]$, scaled to a lattice spacing $\varepsilon > 0$, is the following tempered distribution defined via the complex fields $\psi_{\sigma} = \psi_{\sigma}[u, v]$ in (2.7):

$$e^{\varepsilon}(x) = \sum_{\gamma \in \mathbb{Z}^3} \delta(x - \varepsilon\gamma) \frac{1}{2} \sum_{\sigma=\pm 1} |\psi_{\sigma}(\gamma)|^2, \tag{2.14}$$

where δ denotes the Dirac delta-distribution. This is a manifestly positive distribution which is identifiable with a measure whose total mass equals the total energy,

$$\begin{aligned} \int \mathrm{d}x\, e^{\varepsilon}[u, v](x) &= \sum_{\gamma \in \mathbb{Z}^3} \frac{1}{2} \sum_{\sigma=\pm 1} |\psi_{\sigma}(\gamma)|^2 = \frac{1}{2} \sum_{\sigma=\pm 1} \|\psi_{\sigma}\|^2 \\ &= \frac{1}{4} \sum_{\sigma=\pm 1} \int \mathrm{d}k\, |\omega(k)\hat{u}(k) + \mathrm{i}\sigma\hat{v}(k)|^2 = H(u, v) < \infty. \end{aligned} \tag{2.15}$$

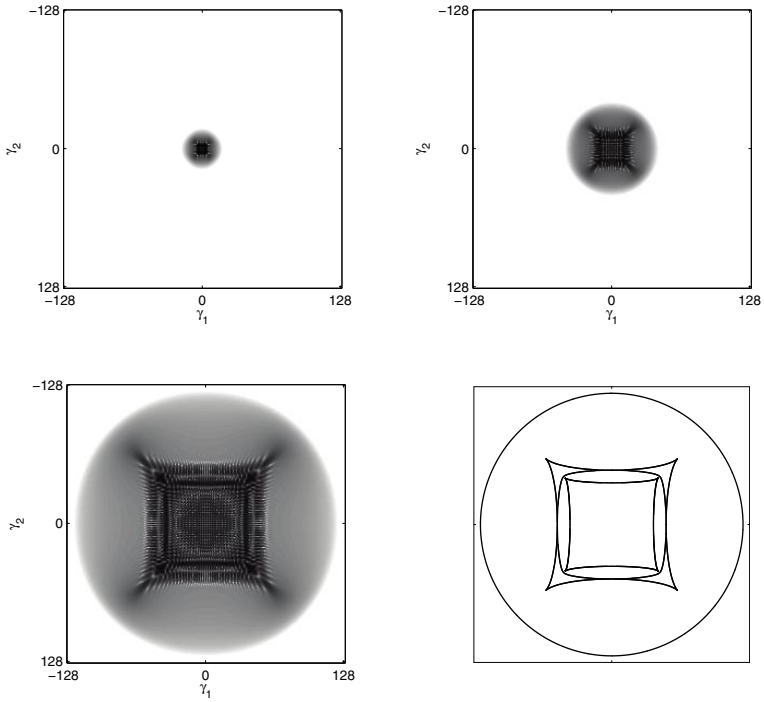


FIG. 2.2. *First three panels: Snapshots of the energy density* $|\psi_+(\gamma,t)|^2$ *in the plane* $\gamma_3 = 0$ *at* $t = 0.1/\varepsilon$, $t = 0.3/\varepsilon$, *and* $t = 0.95/\varepsilon$, $\varepsilon = \frac{1}{128}$, *with initial data and elastic constants as in Figure* 2.1. *The plot is of the logarithm of the density, with all values less than a fixed cutoff shown white. Last panel: Plot of the restriction to the plane* $x_3 = 0$ *of the singular set of the energy of the solution to the transport equation* (1.3), *using a corresponding choice of macroscopic initial data. As the solution is scale-invariant, no explicit length scale has been denoted.*

This justifies calling $e^\varepsilon[u,v]$ an energy density: it defines a distribution of the positive total energy between the lattice sites. The symmetry of ω implies that $|\psi_-(\gamma)| = |\psi_+(\gamma)|$, and thus we can also identify the energy density directly with the norm-density of $\psi_+[u,v]$,

$$e^\varepsilon(x) = \sum_{\gamma\in\mathbb{Z}^3} \delta(x-\varepsilon\gamma)|\psi_+(\gamma)|^2. \tag{2.16}$$

Then also for all $\varepsilon > 0$,

$$H(u,v) = \int \mathrm{d}x\, e^\varepsilon[u,v](x) = \|\psi_+\|^2_{\ell_2(\mathbb{Z}^3)} = \|\hat\psi_+\|^2_{L^2(\mathbb{T}^3)} = \|\hat\psi_-\|^2_{L^2(\mathbb{T}^3)}, \tag{2.17}$$

which is independent of t for any $(u,v) = (u_t, v_t)$.

Let us now concentrate on the case when a solution u_t to (2.1) has been given, and define for any t the distribution $e^\varepsilon(x,t)$ by setting

$$e^\varepsilon(x,t) = e^\varepsilon[u_{t/\varepsilon}, v_{t/\varepsilon}](x), \qquad x\in\mathbb{R}^3,\ t\in\mathbb{R}\,. \tag{2.18}$$

We have given an example of the time evolution of the energy density in Figure 2.2. The last panel in the figure contains the most obvious features which are implied

by the corresponding macroscopic evolution equation: the points of discontinuity of the solution. The macroscopic initial data is given by $\mathrm{d}\mu_0(x,k) = \delta(x)\frac{1}{2}|\omega(k)|^2\mathrm{d}x\mathrm{d}k$, which has no concentration at $k=0$. Thus only (1.3) is relevant. It is readily solved to yield as the energy density (defined in this case simply by integrating μ_t over the k-variable)

$$e(x,t) = \int_{\mathbb{T}^3} \mathrm{d}k\, \delta(x - t\tfrac{1}{2\pi}\nabla\omega(k)) = t^{-3}\int_{\mathbb{T}^3} \mathrm{d}k\, \delta(\tfrac{x}{t} - \tfrac{1}{2\pi}\nabla\omega(k))\,. \tag{2.19}$$

Evaluation of such integrals has been considered, for instance, in section 6.4 of [17]. $|\nabla\omega(k)|$ has its maximum near the point of discontinuity of the gradient, at $k=0$. This defines the outer circle outside which the solution must be zero. Inside the circle, the solution has a finite density, apart from points which correspond to values of k for which the Hessian of ω is not invertible. We have computed the positions of such points using *Mathematica* and plotted the result in the last panel in Figure 2.2. For a reader interested in the details of the computation, we point out that considering the case $x_3 = 0$ simplifies the problem, as it implies that either $k_3 = 0$ or $k_3 = \frac{1}{2}$.

We are interested in the limiting behavior of $e^\varepsilon(x,t)$ as ε tends to 0. Since the velocity of the waves with wave vector k depends on k, it is necessary to work with an object that encodes the density of waves with wave vector $k \in \mathbb{T}^3$ at $x \in \mathbb{R}^3$. This job is conveniently done by the Wigner transform. In order to avoid certain technical difficulties, we are going to define our Wigner transform only in the sense of distributions, i.e., via a duality principle.

First we introduce the space of Schwartz functions.

DEFINITION 2.3. *Let $\mathcal{S}_d = \mathcal{S}(\mathbb{R}^d)$ denote the Schwartz space and $\|\cdot\|_{\mathcal{S}_d,N}$ the corresponding Nth Schwartz norm. Explicitly, with α denoting an arbitrary multi-index and with $\langle x\rangle = \sqrt{1+x^2}$, then*

$$\|f\|_{\mathcal{S}_d,N} = \sup_{x\in\mathbb{R}^d}\max_{|\alpha|\le N} |\langle x\rangle^N \partial^\alpha f(x)|. \tag{2.20}$$

We also employ the shorthand notation $\mathcal{S} = \mathcal{S}_3$.

To extract the relevant weak limits as ε tends to 0, we have to specify a space of suitable test functions. We need to track the evolution of three different kinds of lattice vibrations (short, medium, and long wavelength), all done compatibly with the periodicity in the Fourier variable. This leads to the following, somewhat involved notion of multiscale test functions.

DEFINITION 2.4. *We call a test function $a \in C^\infty(\mathbb{R}^3\times\mathbb{T}^3\times\mathbb{R}^3)$* admissible *if it satisfies the following properties:*

1. $\sup_{k,q,|\alpha|\le N} \|\partial^\alpha_{k,q} a(\cdot,k,q)\|_{\mathcal{S},N} < \infty$ *for all* $N \ge 0$.
2. $q \mapsto a(x,k,q)$ *is constant for all* $x \in \mathbb{R}^3$ *and* k *such that* $\max_i |k_i \bmod T| \ge \frac{1}{4}$.
3. *There is a function* $b \in C^\infty(\mathbb{R}^3\times\mathbb{T}^3\times S^2)$ *such that for any* $N \ge 0$

$$\sup_{|q|\ge R, k\in\mathbb{T}^3} \|a(\cdot,k,q) - b(\cdot,k,\tfrac{q}{|q|})\|_{\mathcal{S},N} \to 0, \quad \text{when } R\to\infty. \tag{2.21}$$

The first and third conditions are the most important and can be summarized as follows: we assume the test functions to be Schwartz in x and smooth with bounded derivatives in k and q and to have a radial limit b in q which is approached uniformly in any of the Schwartz norms. The above requirements are not minimal. The second condition is only needed in order to guarantee that $k \mapsto a(x,k,(k \bmod T^3)/\varepsilon)$ is always

smooth on $\mathbb{T}^3$. Also, taking arbitrarily large N in the last step is not necessary; most likely $N = d + 3 = 6$ would suffice.

Having the notion of admissible test functions at our disposal we can define the central object of this paper: the multiscale Wigner transform.

DEFINITION 2.5. *Let $\psi \in \ell_2(\mathbb{Z}^3)$. We define the multiscale lattice Wigner transform $W^\varepsilon[\psi]$ at scale $\varepsilon > 0$ by*

$$\langle a, W^\varepsilon[\psi]\rangle = \int_{\mathbb{R}^3} \mathrm{d}p \int_{T^3} \mathrm{d}k\, \overline{\hat{a}(p, k, \tfrac{k}{\varepsilon})}\, \overline{\hat{\psi}(k - \varepsilon\tfrac{p}{2})}\hat{\psi}(k + \varepsilon\tfrac{p}{2}), \tag{2.22}$$

where a is an admissible test function and $\hat{a} = \mathcal{F}_{x\to p}a$, i.e.,

$$\hat{a}(p, k, q) = \int_{\mathbb{R}^3} \mathrm{d}x\, \mathrm{e}^{-2\pi \mathrm{i} p\cdot x} a(x, k, q). \tag{2.23}$$

The L^2-Wigner transform $W^{(\varepsilon)}_{cont}[\phi]$ of a function $\phi \in L^2(\mathbb{R}^3)$ at the scale $\varepsilon > 0$ is given by the distribution

$$b \mapsto \langle b, W^{(\varepsilon)}_{cont}[\phi]\rangle = \int_{\mathbb{R}^3\times\mathbb{R}^3} \mathrm{d}p\, \mathrm{d}q\, \overline{\hat{b}(p, q)}\, \overline{\hat{\phi}(q - \varepsilon\tfrac{p}{2})}\hat{\phi}(q + \varepsilon\tfrac{p}{2}) \tag{2.24}$$

for all $b \in \mathcal{S}(\mathbb{R}^3, C^\infty(\mathbb{R}^3))$, and with $\hat{b} = \mathcal{F}_{x\to p}b$.

The Wigner transform $W^{(\varepsilon)}_{\mathrm{cont}}$ is a rescaled version of the standard Wigner transform in L^2. The test-function space $\mathcal{S}(\mathbb{R}^3, C^\infty(\mathbb{R}^3))$ used above in its definition is obtained via the family of seminorms $p_N(b) = \sup_{|\alpha|,|x|,|q|\le N} |\langle x\rangle^N \partial^\alpha_{x,q} b(x, q)|$ with $b \in C^\infty(\mathbb{R}^3 \times \mathbb{R}^3)$. This is a Fréchet space, and $W^{(\varepsilon)}_{\mathrm{cont}}[\phi]$ is a continuous functional on it for any $\phi \in L^2(\mathbb{R}^3)$, as the following estimate reveals:

$$|\langle b, W^{(\varepsilon)}_{\mathrm{cont}}[\phi]\rangle| \le \sup_{p,q} |\langle p\rangle^4 \hat{b}(p, q)|\, \|\phi\|^2 \int_{\mathbb{R}^3} \mathrm{d}p\, \langle p\rangle^{-4}. \tag{2.25}$$

Although $\mathcal{S}_6$ is not dense in this test-function space, it is nevertheless enough to know how $W^{(\varepsilon)}_{\mathrm{cont}}$ acts on it. More precisely, if $W_i = W^{(\varepsilon)}_{\mathrm{cont}}[\phi_i]$, $i = 1, 2$, and $\langle b, W_1\rangle = \langle b, W_2\rangle$ for all $b \in \mathcal{S}_6$, then $W_1 = W_2$. This follows straightforwardly from an estimate similar to (2.25) using smooth cutoff functions to cut out the infinity of the q-variable. In addition, we will also need the property that if $b(x, q) = f(x)$, with $f \in \mathcal{S}_3$, then

$$\langle b, W^{(\varepsilon)}_{\mathrm{cont}}[\phi]\rangle = \int_{\mathbb{R}^3} \mathrm{d}x \overline{f(x)} |\phi(x)|^2. \tag{2.26}$$

That is, $\int_{\mathbb{R}^3} \mathrm{d}q\, W^{(\varepsilon)}_{\mathrm{cont}}[\phi](x, q) = |\phi(x)|^2$. For a more careful analysis of the properties of the standard Wigner transform, see [12].

In [14] the Wigner transform of a lattice state was defined as a distribution $W^\varepsilon_{\mathrm{latt}} \in \mathcal{S}'(\mathbb{R}^3 \times \mathbb{T}^3)$. The above definition is simply a refinement of this definition: formally for any ψ

$$W^\varepsilon(x, k, q) = \delta\big(q - \tfrac{k}{\varepsilon}\big) W^\varepsilon_{\mathrm{latt}}(x, k), \qquad x \in \mathbb{R}^3,\ k \in T^3,\ q \in \mathbb{R}^3\,. \tag{2.27}$$

This follows immediately from equation (B.6) of [14], after one realizes that if a is an admissible test function, then $(x, k) \mapsto a(x, k, (k \bmod T^3)/\varepsilon)$ belongs to $\mathcal{S}(\mathbb{R}^3 \times \mathbb{T}^3)$ for any $\varepsilon > 0$. This identification immediately allows us to use the results in [14] and

to prove that many of the basic properties of the usual Wigner transform carry over to the multiscale Wigner transform. Particularly important for us is the following relation.

PROPOSITION 2.6. *For any $f \in \mathcal{S}(\mathbb{R}^3)$, the test function $a_f(x,k,q) = f(x)$ is admissible, and for all $\varepsilon > 0$ and for all $(u,v) \in X$,*

$$\langle f, e^\varepsilon \rangle = \langle a_f, W^\varepsilon[\psi] \rangle, \tag{2.28}$$

where $e^\varepsilon = e^\varepsilon[u,v]$ and $\psi = \psi_+[u,v]$.

Proof. By inspection, we find that a is a well-defined test function and $\psi \in \ell_2$. Then, by the above-mentioned relation, $\langle a_f, W^\varepsilon[\psi] \rangle = \langle J_f, W^\varepsilon_{\mathrm{latt}}[\psi] \rangle$ with $J_f(x,k) = f(x)$. On the other hand, it follows directly from the definition of $W^\varepsilon_{\mathrm{latt}}$ (equation (B.2) in [14]) that

$$\begin{aligned} \langle J_f, W^\varepsilon_{\mathrm{latt}}[\psi] \rangle &= \sum_{\gamma,\gamma' \in \mathbb{Z}^3} \psi(\gamma)\overline{\psi(\gamma')} \int_{\mathbb{T}^3} \mathrm{d}k\, \mathrm{e}^{2\pi \mathrm{i} k \cdot (\gamma' - \gamma)} \overline{f(\varepsilon(\gamma' + \gamma)/2)} \\ &= \sum_{\gamma \in \mathbb{Z}^3} \overline{f(\varepsilon\gamma)} |\psi(\gamma)|^2 = \langle f, e^\varepsilon \rangle. \end{aligned} \tag{2.29}$$

This proves (2.28). ∎

The above proposition can be formally summarized by the formula

$$e^\varepsilon(x) = \int_{\mathbb{T}^3} \mathrm{d}k \int_{\mathbb{R}^3} \mathrm{d}q\, W^\varepsilon[\psi](x,k,q), \tag{2.30}$$

which implies also (in the sense of choosing any suitable test-function sequence approaching 1)

$$\int_{\mathbb{R}^3 \times \mathbb{T}^3 \times \mathbb{R}^3} \mathrm{d}x\, \mathrm{d}k\, \mathrm{d}q\, W^\varepsilon[\psi](x,k,q) = \|\psi\|^2_{\ell_2(\mathbb{Z}^3)}. \tag{2.31}$$

As noted earlier, analogous results hold for the L^2-Wigner transform.

Several definitions have been used in the literature to study homogenization limits of lattice systems. We follow [14] mainly for convenience. The definition in [14] is based on the "Weyl quantization rule" of symbols, and a similar definition of the Wigner transform as distributions with a "classical quantization rule" was proposed earlier in [16]. Other choices include using Husimi transforms or considering $L^2(\mathbb{T}^3) = L^2(T^3)$ as a subspace of $L^2(\mathbb{R}^3)$ and then relying on the standard Wigner transform [17]. Finally, one can consider also interpolations of the fields between lattice sites and then using the Wigner series defined in [12]. The relations between the various definitions are discussed in [11, 16, 17, 24]. As is apparent from the results presented in these references, the various 1-microlocal definitions tend to lead to the same limit Wigner measures. This is not the case for the above 2-microlocal measures, since at least the information about the quantization rule will be carried over to the homogenization limits. However, a more systematic study would be required to settle the issue.

3. Main results. The macroscopic evolution is obtained by sending ε to 0. Our objective is to characterize the asymptotic behavior of the Wigner transform $W^\varepsilon[\psi^\varepsilon_{t/\varepsilon}]$ of the solution $\psi^\varepsilon_{t/\varepsilon}$ of (2.10). The limit strongly depends on the dispersion relation ω. We will consider here regular acoustic dispersion relations, keeping in

mind that, for instance, in the case of a scalar field and nearest neighbor interactions in $\mathbb{Z}^3$ the dispersion relation ω is given by (2.12), which is regular acoustic with $\omega_0(q) = 2\pi|q|$. The main achievement of this paper as compared to [17] is that complicated assumptions concerning the concentrations of the Wigner transform W^ε in wavenumber space as ε tends to 0 are no longer needed. The only remaining requirements are boundedness and tightness of the sequence of initial excitations.

ASSUMPTION 3.1. *We consider a sequence of values $\varepsilon > 0$ such that $\varepsilon \to 0$. For each ε in the sequence we assume that there is given an initial data vector $\psi_0^\varepsilon \in \ell_2(\mathbb{Z}^3)$ such that*

1. $\sup_\varepsilon \|\psi_0^\varepsilon\| < \infty$;
2. *the sequence ψ_0^ε is tight on the scale ε^{-1}:*

$$\lim_{R\to\infty} \limsup_{\varepsilon\to 0} \sum_{|\gamma|>R/\varepsilon} |\psi_0^\varepsilon(\gamma)|^2 = 0. \tag{3.1}$$

After these preparations we are in a position to state our result. The main point is that if ω is a regular acoustic dispersion relation, the asymptotic behavior of the energy density is characterized by precisely three different objects: a Wigner transform of a weak limit (macroscopic waves), an H-measure (short macroscopic waves), and a Wigner measure (microscopic waves). No assumptions concerning energy concentrations except those stated in Assumption 3.1 are required.

THEOREM 3.2. *Let $\psi_0^\varepsilon \in \ell_2(\mathbb{Z}^3)$ be a sequence which satisfies Assumption* 3.1. *Let ω be a regular acoustic dispersion relation and define $\psi_t^\varepsilon \in \ell_2(\mathbb{Z}^3)$ for all $t \in \mathbb{R}$ by the formula*

$$\hat{\psi}_t^\varepsilon(k) = \mathrm{e}^{-\mathrm{i}t\omega(k)} \hat{\psi}_0^\varepsilon(k). \tag{3.2}$$

Let also $\mathbb{T}^3_* = \mathbb{T}^3 \setminus \{0\}$. Then there are positive, bounded Radon measures $\mu_0, \mu_0^{\mathrm{H}}$ on $\mathbb{R}^3 \times \mathbb{T}^3_*$ and $\mathbb{R}^3 \times S^2$, respectively, a function $\phi_0 \in L^2(\mathbb{R}^3)$, and a subsequence (not relabeled) such that for all admissible test functions a and $t \in \mathbb{R}$,

$$\begin{aligned} \lim_{\varepsilon\to 0} \langle a, W^\varepsilon[\psi^\varepsilon_{t/\varepsilon}]\rangle &= \int_{\mathbb{R}^3\times\mathbb{T}^3_*} \mathrm{d}\mu_t(x,k)\, \overline{b(x,k,\tfrac{k}{|k|})} \\ &\quad + \int_{\mathbb{R}^3\times S^2} \mathrm{d}\mu_t^{\mathrm{H}}(x,q)\, \overline{b(x,0,q)} + \langle a_0, W^{(1)}_{cont}[\phi_t]\rangle\,, \end{aligned} \tag{3.3}$$

where $b(x,k,q) = \lim_{R\to\infty} a(x,k,Rq)$ for $|q| = 1$, $a_0(x,q) = a(x,0,q)$, and ϕ_t, μ_t and μ_t^{H} are determined for any $f \in C(\mathbb{R}^3 \times \mathbb{T}^3_*)$ and $g \in C(\mathbb{R}^3 \times S^2)$ by

$$\hat{\phi}_t(q) := \mathrm{e}^{-\mathrm{i}t\omega_0(q)} \hat{\phi}_0(q), \tag{3.4}$$

$$\int_{\mathbb{R}^3\times\mathbb{T}^3_*} f(x,k)\,\mathrm{d}\mu_t(x,k) := \int_{\mathbb{R}^3\times\mathbb{T}^3_*} f(x + t\tfrac{1}{2\pi}\nabla\omega(k), k)\,\mathrm{d}\mu_0(x,k), \tag{3.5}$$

$$\int_{\mathbb{R}^3\times S^2} g(x,q)\,\mathrm{d}\mu_t^{\mathrm{H}}(x,q) := \int_{\mathbb{R}^3\times S^2} g(x + t\tfrac{1}{2\pi}\nabla\omega_0(q), q)\,\mathrm{d}\mu_0^{\mathrm{H}}(x,q)\,. \tag{3.6}$$

Moreover, for all t the energy equality holds:

$$\lim_{\varepsilon\to 0} \|\psi_0^\varepsilon\|^2 = \mu_t(\mathbb{R}^3\times\mathbb{T}^3_*) + \mu_t^{\mathrm{H}}(\mathbb{R}^3\times S^2) + \|\phi_t\|^2_{L^2(\mathbb{R}^3)}\,. \tag{3.7}$$

Remark 3.3. It is immediate from the definition that ϕ, μ, and μ^{H} are weak solutions of the set of decoupled linear transport equations (1.3)–(1.5).

The proof of Theorem 3.2 also shows that the subsequences which are extracted in the statement of the theorem can be characterized by a simple condition. In particular, the initial state of the wave equation, ϕ_0, is determined as the weak-$L^2(\mathbb{R}^3)$ limit of the sequence of the functions (ϕ_0^ε) with Fourier transforms

$$\hat{\phi}_0^\varepsilon(q) = \begin{cases} \varepsilon^{\frac{3}{2}} \hat{\psi}_0^\varepsilon(\varepsilon q) & \text{if } \|q\|_\infty \leq \frac{1}{2\varepsilon}, \\ 0, & \text{otherwise.} \end{cases} \tag{3.8}$$

The exact characterization is contained in the following corollary, whose proof will be given in section 5.

COROLLARY 3.4. *Let (ψ_0^ε) be a sequence which satisfies Assumption* 3.1. *Suppose that ϕ_0^ε converges weakly to ϕ_0, and that $\lim_{\varepsilon\to 0}\langle a, W^\varepsilon[\psi_0^\varepsilon]\rangle$ exists for every admissible test function a. Then there are unique positive, bounded Radon measures $\mu_0, \mu_0^{\mathrm{H}}$ on $\mathbb{R}^3 \times \mathbb{T}^3_*$ and $\mathbb{R}^3 \times S^2$, respectively, such that for every admissible test function a*

$$\begin{aligned} \lim_{\varepsilon\to 0}\langle a, W^\varepsilon[\psi_0^\varepsilon]\rangle = &\int_{\mathbb{R}^3\times\mathbb{T}^3_*} \mathrm{d}\mu_0(x,k)\, \overline{b(x,k,\tfrac{k}{|k|})} \\ &+ \int_{\mathbb{R}^3\times S^2} \mathrm{d}\mu_0^{\mathrm{H}}(x,q)\, \overline{b(x,0,q)} + \langle a_0, W^{(1)}_{cont}[\phi_0]\rangle\,, \end{aligned} \tag{3.9}$$

where a_0 and b are defined as in Theorem 3.2. *In addition, then* (3.3) *holds for all $t \in \mathbb{R}$ along the original sequence ε with the initial macroscopic data determined by the triplet $(\mu_0, \mu_0^{\mathrm{H}}, \phi_0)$.*

Remark 3.5. According to (2.27) for $a \in \mathcal{S}(\mathbb{R}^3 \times \mathbb{T}^3)$ the lattice Wigner transform $W^\varepsilon_{\mathrm{latt}}[\psi_0^\varepsilon]$ satisfies

$$\langle a, W^\varepsilon_{\mathrm{latt}}[\psi_0^\varepsilon]\rangle = \langle a, W^\varepsilon[\psi_0^\varepsilon]\rangle\,. \tag{3.10}$$

It is well known (see, e.g., [14]) that $W^\varepsilon_{\mathrm{latt}}[\psi_0^\varepsilon]$ has a limit μ_{latt},

$$\lim_{\varepsilon\to 0}\langle a, W^\varepsilon_{\mathrm{latt}}[\psi_0^\varepsilon]\rangle = \int_{\mathbb{R}^3\times\mathbb{T}^3} \mathrm{d}\mu_{\mathrm{latt}}(x,k)\, \overline{a(x,k)}\,, \tag{3.11}$$

which is a positive Radon measure on $\mathbb{R}^3 \times \mathbb{T}^3$. Evaluating (3.9) for $a \in \mathcal{S}(\mathbb{R}^3 \times \mathbb{T}^3)$ shows that the "standard Wigner measure" μ_{latt} can be expressed in terms of the two-scale Wigner measure according to

$$\mathrm{d}\mu_{\mathrm{latt}}(x,k) = \mathrm{d}\mu_{\mathrm{latt}}(x,k)\, \mathbb{1}(k \neq 0) + \int_{q\in S^2} \mathrm{d}\mu_0^{\mathrm{H}}(x,q) \otimes \delta(k)\mathrm{d}k + |\phi_0(x)|^2 \mathrm{d}x \otimes \delta(k)\mathrm{d}k\,, \tag{3.12}$$

where $\mathrm{d}x$ denotes Lebesgue measure on $\mathbb{R}^3$, δ the Dirac δ-function, and $\mathbb{1}$ the characteristic function.

Remark 3.6. Our proof of the existence of two-scale Wigner measures is similar to the strategy proposed by Fermanian Kammerer in [4, 5]. For a general result on existence of two-scale Wigner measures concentrating on hypersurfaces, see [6]. (See [6, Theorem 1.4] for a simple formulation in one dimension and [6, Theorem 1.6] for the general case.) Concerning the existence part the only novelty in our result is thus the use of the lattice Wigner transform.

However, in our approach we obtain the existence of the measures together with their time evolution. While the strategy for obtaining transport equations for two-scale Wigner measures outlined in section 6 of [7] is not directly applicable in our

case due to the singular nature of the dispersion relation ω, it could certainly be adapted. The existence part and the transport part of the result could thus be separated. However, our proof is technically simpler and we do not rely on any advanced pseudodifferential calculus.

4. Proof of Theorem 3.2. Let a be an admissible test function and consider a fixed $t \in \mathbb{R}$ when we need to inspect the $\varepsilon \to 0$ limit of

$$\langle a, W^\varepsilon[\psi^\varepsilon_{t/\varepsilon}]\rangle = \int_{\mathbb{R}^3} \mathrm{d}p \int_{\mathbb{T}^3} \mathrm{d}k\, \overline{\hat{a}(p,k,\tfrac{k}{\varepsilon})}\, \mathrm{e}^{-\mathrm{i}\frac{t}{\varepsilon}\left(\omega(k+\varepsilon\frac{p}{2})-\omega(k-\varepsilon\frac{p}{2})\right)}\, \overline{\hat{\psi}^\varepsilon_0(k-\varepsilon\tfrac{p}{2})}\, \hat{\psi}^\varepsilon_0(k+\varepsilon\tfrac{p}{2})\,. \tag{4.1}$$

First we identify the function ϕ_0 which contains the contributions of the long-wave excitations. Let $\hat{\phi}^\varepsilon_0$ be defined by (3.8). Since $\limsup_{\varepsilon\to 0}\|\hat{\phi}^\varepsilon_0\|_{L^2} = \limsup_{\varepsilon\to 0}\|\psi^\varepsilon_0\|_{\ell_2}$ is bounded by Assumption 3.1 there exist a subsequence and a function $\phi_0 \in L^2(\mathbb{R}^3)$ such that ϕ^ε_0 converges weakly to ϕ_0 in $L^2(\mathbb{R}^3)$. We then define ϕ_t by (3.4).

Using a localization function χ, which will be specified later, we now split the integral I on the right-hand side of (4.1) into three parts and an error term so that the contributions of short-, medium-, and long-wave excitation can be analyzed separately. We define

$$\begin{aligned}
I_> &= \int_{\mathbb{R}^3} \mathrm{d}p \int_{\mathbb{T}^3} \mathrm{d}k\, \overline{\hat{a}(p,k,\tfrac{k}{\varepsilon})}\, \mathrm{e}^{-\mathrm{i}\frac{t}{\varepsilon}\left(\omega(k+\varepsilon\frac{p}{2})-\omega(k-\varepsilon\frac{p}{2})\right)}\, \overline{\hat{\psi}^\varepsilon_0(k-\varepsilon\tfrac{p}{2})}\, \hat{\psi}^\varepsilon_0(k+\varepsilon\tfrac{p}{2})(1-\chi),\\
I^{\mathrm{H}}_< &= \int_{\mathbb{R}^3} \mathrm{d}p \int_{\mathbb{T}^3} \mathrm{d}k\, \overline{\hat{a}(p,k,\tfrac{k}{\varepsilon})}\, \mathrm{e}^{-\mathrm{i}\frac{t}{\varepsilon}\left(\omega(k+\varepsilon\frac{p}{2})-\omega(k-\varepsilon\frac{p}{2})\right)}\, \overline{(\hat{\psi}^\varepsilon_0(k-\varepsilon\tfrac{p}{2}) - \varepsilon^{-\frac{3}{2}}\hat{\phi}_0(\tfrac{k}{\varepsilon}-\tfrac{p}{2}))}\\
&\qquad \times (\hat{\psi}^\varepsilon_0(k+\varepsilon\tfrac{p}{2}) - \varepsilon^{-\frac{3}{2}}\hat{\phi}_0(\tfrac{k}{\varepsilon}+\tfrac{p}{2}))\chi,\\
I^{\mathrm{wv}}_< &= \varepsilon^{-3}\int_{\mathbb{R}^3} \mathrm{d}p \int_{\mathbb{T}^3} \mathrm{d}k\, \overline{\hat{a}(p,k,\tfrac{k}{\varepsilon})}\, \mathrm{e}^{-\mathrm{i}\frac{t}{\varepsilon}\left(\omega(k+\varepsilon\frac{p}{2})-\omega(k-\varepsilon\frac{p}{2})\right)}\, \overline{\hat{\phi}_0(\tfrac{k}{\varepsilon}-\tfrac{p}{2})}\, \hat{\phi}_0(\tfrac{k}{\varepsilon}+\tfrac{p}{2})\chi,\\
R &= I - I_> - I^{\mathrm{H}}_< - I^{\mathrm{wv}}_<.
\end{aligned}$$

The definition of R implies that

$$\langle a, W^\varepsilon[\psi^\varepsilon_{t/\varepsilon}]\rangle = I_> + I^{\mathrm{H}}_< + I^{\mathrm{wv}}_< + R. \tag{4.2}$$

To localize the oscillations in Fourier space we need smooth cutoff functions.

DEFINITION 4.1. *Let $f \in C^\infty(\mathbb{R},[0,1])$ denote a fixed function which is symmetric, $f(-x) = f(x)$, strictly monotonically decreasing on $[1,2]$, and*

$$f(x) = \begin{cases} 1 & \text{if } |x| \le 1,\\ 0 & \text{if } |x| \ge 2. \end{cases} \tag{4.3}$$

We further define $\varphi \in C^\infty(\mathbb{R}^3,[0,1])$ by $\varphi(k) = f(|k|)$.

Let $0 < \rho \le \frac{1}{4}$ be arbitrary and set $\chi = \chi(k,p;\varepsilon,\rho) = \varphi(\frac{k_+}{\rho})\varphi(\frac{k_-}{\rho})$, where $k_\pm = k \pm \varepsilon\frac{p}{2}$. We will continue to use this shorthand notation in the following under the tacit assumption that $k_\pm$ is always really a function of both k and εp. Let us now also point out that all four terms in the decomposition (4.2) depend on ε and ρ, via χ, even though we have not denoted this dependence explicitly.

The first term containing $1-\chi$ in (4.2) is zero if $|k_\pm| \le \rho$, while the remainder is zero if $|k_+|$ or $|k_-| \ge 2\rho$. Thus the chosen decomposition splits the integration over k and p into "large," "intermediate," and "small" wavenumbers. We will demonstrate

that the following convergences hold: there is a sequence of ρ, a subsequence of ε, and measures $\mu_t, \mu_t^{\mathrm{H}}$ satisfying the statements made in the theorem, such that

$$\lim_{\rho\to 0}\lim_{\varepsilon\to 0} I_> = \int_{\mathbb{R}^3\times\mathbb{T}^3_*} \mathrm{d}\mu_t(x,k)\,\overline{b(x,k,\tfrac{k}{|k|})}, \tag{4.4}$$

$$\lim_{\rho\to 0}\lim_{\varepsilon\to 0} I_<^{\mathrm{H}} = \int_{\mathbb{R}^3\times S^2} \mathrm{d}\mu_t^{\mathrm{H}}(x,q)\,\overline{b(x,0,q)}, \tag{4.5}$$

$$\lim_{\rho\to 0}\limsup_{\varepsilon\to 0}\left| I_<^{\mathrm{wv}} - \langle a_0, W^{(1)}_{\mathrm{cont}}[\phi_t]\rangle\right| = 0, \tag{4.6}$$

$$\lim_{\rho\to 0}\limsup_{\varepsilon\to 0}|R| = 0. \tag{4.7}$$

Clearly, (4.4)–(4.7) then imply (3.3).

Large wavenumbers. We split $I_>$ further into two parts using

$$\begin{aligned} 1-\varphi\left(\tfrac{k_-}{\rho}\right)\varphi\left(\tfrac{k_+}{\rho}\right) &= 1-\varphi\left(\tfrac{k}{\rho}\right)^2 + \left(\varphi\left(\tfrac{k}{\rho}\right)-\varphi\left(\tfrac{k_+}{\rho}\right)\right)\varphi\left(\tfrac{k}{\rho}\right) \\ &\quad + \left(\varphi\left(\tfrac{k}{\rho}\right)-\varphi\left(\tfrac{k_-}{\rho}\right)\right)\varphi\left(\tfrac{k_+}{\rho}\right). \end{aligned} \tag{4.8}$$

Let $h_\rho(k) = 1-\varphi(\frac{k}{\rho})^2$, which is a smooth function. The integral then becomes $I_> = I_>^1 + R_1$, where

$$I_>^1 = \int_{\mathbb{R}^3}\mathrm{d}p\int_{\mathbb{T}^3}\mathrm{d}k\,\overline{\hat a(p,k,\tfrac{k}{\varepsilon})}h_\rho(k)\,\mathrm{e}^{-\mathrm{i}\frac{t}{\varepsilon}\left(\omega(k+\varepsilon\frac{p}{2})-\omega(k-\varepsilon\frac{p}{2})\right)}\,\overline{\hat\psi_0^\varepsilon(k-\varepsilon\tfrac{p}{2})}\,\hat\psi_0^\varepsilon(k+\varepsilon\tfrac{p}{2}). \tag{4.9}$$

The remainder R_1 can be estimated using $|\varphi|\le 1$ and

$$\varphi\left(\tfrac{k_\pm}{\rho}\right)-\varphi\left(\tfrac{k}{\rho}\right) = \pm\tfrac{\varepsilon}{2\rho}\int_0^1 \mathrm{d}s\, p\cdot\nabla\varphi\Big(\tfrac{1}{\rho}(k\pm s\tfrac{\varepsilon}{2}p)\Big), \tag{4.10}$$

which yield the bound, with a universal constant C,

$$|R_1| \le C\tfrac{\varepsilon}{\rho}\sup_{k,q}\|a(\cdot,k,q)\|_{\mathcal{S},d+2}\|\hat\psi_0^\varepsilon\|^2\|\nabla\varphi\|_\infty. \tag{4.11}$$

Therefore, there is a constant c' such that $|R_1|\le c'\varepsilon/\rho$ and thus $R_1\to 0$ when $\varepsilon\to 0$ for all ρ.

We then consider $I_>^1$. The presence of h_ρ guarantees that the integrand is zero unless $|k|\ge\rho$. Thus we can change $\hat a(p,k,k/\varepsilon)$ to $\hat b(p,k,k/|k|)$ in the integrand with an error R_2 bounded by

$$|R_2| \le C\sup_{k,|q|\ge\rho/\varepsilon}\|a(\cdot,k,q)-b(\cdot,k,\tfrac{q}{|q|})\|_{\mathcal{S},d+1}\|\hat\psi_0^\varepsilon\|^2, \tag{4.12}$$

where C is a universal constant. Therefore, the assumptions imply that $R_2\to 0$ when $\varepsilon\to 0$ for all ρ. On the other hand, for $|k|\ge\rho$ and $|p|<\rho/\varepsilon$, inequality (A.2) derived in the appendix implies

$$\left|\tfrac{1}{\varepsilon}\left(\omega(k+\varepsilon\tfrac{p}{2})-\omega(k-\varepsilon\tfrac{p}{2})\right)-p\cdot\nabla\omega(k)\right| \le C_3\varepsilon\tfrac{|p|^2}{|k|}\le C_3\tfrac{\varepsilon}{\rho}|p|^2. \tag{4.13}$$

Therefore, using the estimate $|\mathrm{e}^{\mathrm{i}x} - \mathrm{e}^{\mathrm{i}y}| \leq \min(|x-y|, 2)$, valid for all $x, y \in \mathbb{R}$, we find that we can further change the t-dependent exponential in the integrand to $\mathrm{e}^{-\mathrm{i}tp\cdot\nabla\omega(k)}$ with an error R_3 satisfying

$$|R_3| \leq C \sup_{k,q} \|b(\cdot,k,q)\|_{\mathcal{S},d+3} \Big(\tfrac{\varepsilon}{\rho}|t| + \int_{|p|\geq\rho/\varepsilon} \mathrm{d}p \langle p\rangle^{-d-3} \Big) \tag{4.14}$$

for some constant C. Therefore, also $\lim_{\varepsilon\to 0} R_3 = 0$ for all ρ. In summary, $I^1_> = I^2_> + R_2 + R_3$, where R_2, R_3 are negligible, and

$$I^2_> = I^2_>(\varepsilon,\rho) = \int_{\mathbb{R}^3} \mathrm{d}p \int_{\mathbb{T}^3} \mathrm{d}k\, \overline{\hat{b}(p,k,\tfrac{k}{|k|})} h_\rho(k) \mathrm{e}^{-\mathrm{i}tp\cdot\nabla\omega(k)}\, \overline{\hat{\psi}^\varepsilon_0(k-\varepsilon\tfrac{p}{2})}\, \hat{\psi}^\varepsilon_0(k+\varepsilon\tfrac{p}{2})\,. \tag{4.15}$$

Let us for a moment consider the lattice Wigner transform $W^\varepsilon_{\mathrm{latt}}$ of ψ^ε_0 as defined in [14]. As pointed out after Definition 2.5, then for any test function $f \in \mathcal{S}(\mathbb{R}^3\times\mathbb{T}^3)$, we get an admissible test function by the formula $a_f(x,k,q) = f(x,k)$ and $\langle f, W^\varepsilon_{\mathrm{latt}}\rangle = \langle a_f, W^\varepsilon[\psi^\varepsilon_0]\rangle$. Since ψ^ε_0 is a norm-bounded sequence, the sequence $W^\varepsilon_{\mathrm{latt}}$ is weak-$*$ bounded, and thus there are $W^0_{\mathrm{latt}} \in \mathcal{S}'(\mathbb{R}^3\times\mathbb{T}^3)$ and a subsequence along which $W^\varepsilon_{\mathrm{latt}} \overset{*}{\rightharpoonup} W^0_{\mathrm{latt}}$.

Since the sequence ψ^ε_0 is by assumption also tight on the scale ε^{-1}, we can then apply Theorems B.4 and B.5 of [14] and conclude that W^0_{latt} is given by a positive, bounded Radon measure μ on $\mathbb{R}^3\times\mathbb{T}^3$ such that for all *continuous* functions $f \in C(\mathbb{T}^3)$ and $p \in \mathbb{R}^3$,

$$\lim_{\varepsilon\to 0} \int_{\mathbb{T}^3} \mathrm{d}k\, f(k) \overline{\hat{\psi}^\varepsilon_0(k-\varepsilon\tfrac{p}{2})}\, \hat{\psi}^\varepsilon_0(k+\varepsilon\tfrac{p}{2}) = \int_{\mathbb{R}^3\times\mathbb{T}^3} \mathrm{d}\mu(x,k)\, f(k) \mathrm{e}^{-2\pi\mathrm{i}p\cdot x}. \tag{4.16}$$

Because $\nabla\omega(k)$ and $\frac{k}{|k|}$ are continuous apart from $k=0$, the function $k \mapsto h_\rho(k)\overline{\hat{b}(p,k,\frac{k}{|k|})}\mathrm{e}^{-\mathrm{i}tp\cdot\nabla\omega(k)}$ is everywhere continuous for all $\rho > 0$ and $p \in \mathbb{R}^3$. Therefore, by the dominated convergence theorem, for all ρ we find

$$\begin{aligned} \lim_{\varepsilon\to 0} I^2_> &= \int_{\mathbb{R}^3} \mathrm{d}p \int_{\mathbb{R}^3\times\mathbb{T}^3} \mathrm{d}\mu(x,k)\, \overline{\hat{b}(p,k,\tfrac{k}{|k|})} h_\rho(k) \mathrm{e}^{-2\pi\mathrm{i}p\cdot(x+t\nabla\omega(k)(2\pi)^{-1})} \\ &= \int_{\mathbb{R}^3\times\mathbb{T}^3} \mathrm{d}\mu(x,k)\, h_\rho(k) \overline{b(x+t\tfrac{1}{2\pi}\nabla\omega(k),k,\tfrac{k}{|k|})}\,. \end{aligned} \tag{4.17}$$

When $\rho\to 0$, the integrand approaches pointwise $\overline{b(x+t\frac{1}{2\pi}\nabla\omega(k),k,\frac{k}{|k|})}$ apart from $k=0$ when the limit is 0. Therefore, by the dominated convergence theorem

$$\begin{aligned} &\lim_{\rho\to 0} \int_{\mathbb{R}^3\times\mathbb{T}^3} \mathrm{d}\mu(x,k)\, h_\rho(k) \overline{b(x+t\tfrac{1}{2\pi}\nabla\omega(k),k,\tfrac{k}{|k|})} \\ &= \int_{\mathbb{R}^3\times\mathbb{T}^3_*} \mathrm{d}\mu(x,k)\, \overline{b(x+t\tfrac{1}{2\pi}\nabla\omega(k),k,\tfrac{k}{|k|})} = \int_{\mathbb{R}^3\times\mathbb{T}^3_*} \mathrm{d}\mu_t(x,k)\, \overline{b(x,k,\tfrac{k}{|k|})}, \end{aligned} \tag{4.18}$$

where we have defined the bounded, positive Radon measure μ_t using $\mu_0 = \mu|_{\mathbb{R}^3\times\mathbb{T}^3_*}$ in the formula (3.5). We have shown that (4.4) holds.

Small wavenumbers and the remainder. After a change of variables $q = \frac{k}{\varepsilon}$, one obtains that

$$I^{\mathrm{wv}}_< = \int_{\mathbb{R}^3} \mathrm{d}p \int_{T^3/\varepsilon} \mathrm{d}q\, \overline{\hat{a}(p,\varepsilon q,q)}\, \mathrm{e}^{-\mathrm{i}\frac{t}{\varepsilon}(\omega(\varepsilon q_+)-\omega(\varepsilon q_-))}\, \overline{\hat{\phi}_0(q_-)}\, \hat{\phi}_0(q_+)\, \varphi(\tfrac{\varepsilon q_+}{\rho})\, \varphi(\tfrac{\varepsilon q_-}{\rho})\,, \tag{4.19}$$

where $q_\pm = q \pm \frac{p}{2}$. We can immediately replace the integration region for the q-integral by $\mathbb{R}^3$. To see this, note that the integrand is zero, unless $|q \pm \frac{p}{2}| \leq \frac{2\rho}{\varepsilon}$ for both signs. Since $2q = q_+ + q_-$, this can happen only if also $|q| \leq \frac{2\rho}{\varepsilon} \leq \frac{1}{2\varepsilon}$, which implies that the integrand is zero if $\|q\|_\infty > \frac{1}{2\varepsilon}$.

However, if $|\varepsilon q| \leq 2\rho$, then

$$|\hat{a}(p, \varepsilon q, q) - \hat{a}(p, 0, q)| \leq \sup_k |\nabla_k \hat{a}(p, k, q)|\, 2\rho, \tag{4.20}$$

and thus we can replace in the integrand the function $\hat{a}(p, \varepsilon q, q)$ by $\hat{a}(p, 0, q)$, with an error R_1' which is bounded by $C\rho$ with a constant C independent of ε and ρ. Thus we only need to consider the integral

$$I_<^1 = \int_{\mathbb{R}^3} \mathrm{d}p \int_{\mathbb{R}^3} \mathrm{d}q\, \overline{\hat{a}(p, 0, q)}\, \mathrm{e}^{-\mathrm{i}\frac{t}{\varepsilon}(\omega(\varepsilon q_+) - \omega(\varepsilon q_-))}\, \overline{F^{\varepsilon,\rho}(q_-)} F^{\varepsilon,\rho}(q_+), \tag{4.21}$$

where $F^{\varepsilon,\rho}(q) = \hat{\phi}_0(q)\varphi(\frac{\varepsilon q}{\rho})$. Next we apply the estimate (A.3) derived in the appendix, proving that if now $|p| \leq \frac{1}{2\varepsilon}$, then

$$\left|\tfrac{1}{\varepsilon}(\omega(\varepsilon q_+) - \omega(\varepsilon q_-)) - \omega_0(q_+) + \omega_0(q_-)\right| \leq C_4 \varepsilon |p|\, |q|\,. \tag{4.22}$$

Following the same argument as earlier, we can then conclude that the t-dependent exponential can be changed to $\mathrm{e}^{-\mathrm{i}t(\omega_0(q_+) - \omega_0(q_-))}$, with an error R_2' which satisfies the estimate

$$|R_2'| \leq C \sup_q \|a(\cdot, 0, q)\|_{\mathcal{S}, d+2} \Big(|t|\rho + \int_{|p| \geq (2\varepsilon)^{-1}} \mathrm{d}p\, \langle p \rangle^{-d-2} \Big). \tag{4.23}$$

Thus $\lim_{\rho \to 0} \limsup_{\varepsilon \to 0} |R_2'| = 0$ for all t. Finally, we need to change $F^{\varepsilon,\rho}$ to $\hat{\phi}_0$, with an error R_3' which can be bounded by $C\|F^{\varepsilon,\rho} - \hat{\phi}_0\|$. Since the bound goes to zero when $\varepsilon \to 0$ and

$$I_<^{\mathrm{wv}} = \int_{\mathbb{R}^3} \mathrm{d}p \int_{\mathbb{R}^3} \mathrm{d}q\, \overline{\hat{a}(p, 0, q)}\, \mathrm{e}^{-\mathrm{i}t(\omega_0(q_+) - \omega_0(q_-))}\, \overline{\hat{\phi}_0(q_-)} \hat{\phi}_0(q_+) + R_1' + R_2' + R_3', \tag{4.24}$$

equation (4.6) has been established.

Similar estimates can be employed to demonstrate the vanishing of the remainder, (4.7). From the definition of R we get

$$\begin{aligned} R = &\int_{\mathbb{R}^3} \mathrm{d}p \int_{T^3/\varepsilon} \mathrm{d}q\, \overline{\hat{a}(p, \varepsilon q, q)}\, \mathrm{e}^{-\mathrm{i}\frac{t}{\varepsilon}(\omega(\varepsilon q_+) - \omega(\varepsilon q_-))}\, \overline{\Big(\hat{\phi}_0^\varepsilon(q_-) - \hat{\phi}_0(q_-)\Big)}\, \hat{\phi}_0(q_+) \varphi(\tfrac{\varepsilon q_+}{\rho})\, \varphi(\tfrac{\varepsilon q_-}{\rho}) \\ &+ \int_{\mathbb{R}^3} \mathrm{d}p \int_{T^3/\varepsilon} \mathrm{d}q \overline{\hat{a}(p, \varepsilon q, q)}\, \mathrm{e}^{-\mathrm{i}\frac{t}{\varepsilon}(\omega(\varepsilon q_+) - \omega(\varepsilon q_-))}\, \overline{\hat{\phi}_0(q_-)} \Big(\hat{\phi}_0^\varepsilon(q_+) - \hat{\phi}_0(q_+)\Big) \varphi(\tfrac{\varepsilon q_+}{\rho}) \varphi(\tfrac{\varepsilon q_-}{\rho}). \end{aligned} \tag{4.25}$$

We then apply the above estimates to remove the ε-dependence from all other terms in the integrands, apart from the differences $\hat{\phi}_0^\varepsilon - \hat{\phi}_0$. The error has a bound which vanishes when $\rho \to 0$. We are then left with

$$\begin{aligned} &\int_{\mathbb{R}^3} \mathrm{d}\eta \int_{\mathbb{R}^3} \mathrm{d}\xi\, \overline{\hat{a}(\eta - \xi, 0, \tfrac{1}{2}(\eta + \xi))}\, \mathrm{e}^{-\mathrm{i}t(\omega_0(\eta) - \omega_0(\xi))}\, \overline{\Big(\hat{\phi}_0^\varepsilon(\xi) - \hat{\phi}_0(\xi)\Big)}\, \hat{\phi}_0(\eta) \\ &+ \int_{\mathbb{R}^3} \mathrm{d}\eta \int_{\mathbb{R}^3} \mathrm{d}\xi\, \overline{\hat{a}(\eta - \xi, 0, \tfrac{1}{2}(\eta + \xi))}\, \mathrm{e}^{-\mathrm{i}t(\omega_0(\eta) - \omega_0(\xi))}\, \overline{\hat{\phi}_0(\xi)} \Big(\hat{\phi}_0^\varepsilon(\eta) - \hat{\phi}_0(\eta)\Big), \end{aligned} \tag{4.26}$$

which vanishes as $\varepsilon \to 0$, since ϕ_0^ε converges weakly to ϕ_0. This establishes (4.7).

Intermediate wavenumbers. Changing coordinates to $q = \frac{k}{\varepsilon}$ yields

$$I^{\mathrm{H}}_{<} = \int_{\mathbb{R}^3} \mathrm{d}p \int_{T^3/\varepsilon} \mathrm{d}q\, \overline{\hat{a}(p, \varepsilon q, q)}\, \mathrm{e}^{-\mathrm{i}\frac{t}{\varepsilon}(\omega(\varepsilon q_+)-\omega(\varepsilon q_-))}\, \overline{\hat{f}^{\varepsilon,\rho}(q_-)} \hat{f}^{\varepsilon,\rho}(q_+)\,,$$

where $\hat{f}^{\varepsilon,\rho}(q) = \varphi(\frac{\varepsilon q}{\rho})(\hat{\phi}^{\varepsilon}_0(q) - \hat{\phi}_0(q))$. Let $M > 0$ be arbitrary. We split off the values $|p| > M$ from the integral defining $I^3_<$. The difference $R'_4 = R'_4(\varepsilon, \rho, M)$ can be bounded by $C \int_{|p|>M} \mathrm{d}p\, \langle p\rangle^{-d-1}$ and thus $\lim_{M\to\infty} \sup_{\rho,\varepsilon} |R'_4| = 0$. We divide the remaining integral over $|p| \le M$ further into two parts using the identity

$$1 = \tilde{\varphi}\left(\tfrac{q_-}{2M}\right) \tilde{\varphi}\left(\tfrac{q_+}{2M}\right) + 1 - \tilde{\varphi}\left(\tfrac{q_-}{2M}\right) \tilde{\varphi}\left(\tfrac{q_+}{2M}\right), \tag{4.27}$$

where $\tilde{\varphi} = 1 - \varphi$. If $|q| \ge 5M$, then $|q_\pm| \ge 4M$ and the second part is zero. It can be checked by inspection that the sequence $(f^{\varepsilon,\rho})_\varepsilon$ is bounded and tight, and it has a weak limit zero. Of these properties only the tightness is nonobvious, but this can also be easily deduced from the formula

$$f^{\varepsilon,\rho}(x) = \rho^3 \varepsilon^{-3/2} \sum_{\gamma\in\mathbb{Z}^3} \psi^{\varepsilon}_0(\gamma) \hat{\varphi}(\rho(\gamma - \tfrac{1}{\varepsilon}x)) - \int_{\mathbb{R}^3} \mathrm{d}y\, \phi_0(x + \tfrac{\varepsilon}{\rho}y)\hat{\varphi}(y). \tag{4.28}$$

Therefore, by Lemma A.2, $\lim_{\varepsilon\to 0} \|\hat{f}^{\varepsilon,\rho}\|_{L^2(B_{6M})} = 0$ for all M, ρ. This implies that the contribution of the second term, denoted by R'_5, satisfies $\lim_{\varepsilon\to 0} |R'_5| = 0$ for all M, ρ.

We are thus left with

$$I^4_{<} = \int_{|p|\le M} \mathrm{d}p \int_{\mathbb{R}^3} \mathrm{d}q\, \overline{\hat{a}(p, 0, q)}\, \mathrm{e}^{-\mathrm{i}t(\omega_0(q_+)-\omega_0(q_-))}\, \overline{\hat{g}^{\varepsilon,\rho,M}(q_-)} \hat{g}^{\varepsilon,\rho,M}(q_+), \tag{4.29}$$

where $\hat{g}^{\varepsilon,\rho,M}(q) = \tilde{\varphi}\left(\frac{q}{2M}\right) \hat{f}^{\varepsilon,\rho}(q)$ and the integrand can be nonzero only for $|q| \ge M$. We thus only need to consider $|q| \ge M$ and $|p| \le M$. First we replace in the integrand $\hat{a}(p, 0, q)$ by $\hat{b}(p, 0, \hat{q})$, $\hat{q} = \frac{q}{|q|}$, with an error R'_6 which is bounded by

$$|R'_6| \le C \sup_{k,|q|\ge M} \|a(\cdot, k, q) - b(\cdot, k, \tfrac{q}{|q|})\|_{\mathcal{S},d+1} \tag{4.30}$$

for some constant C. Then we change $\mathrm{e}^{-\mathrm{i}t(\omega_0(q_+)-\omega_0(q_-))}$ to $\mathrm{e}^{-\mathrm{i}tp\cdot\nabla\omega_0(\hat{q})}$ with an error R'_7 which can be estimated using the inequality (A.4) of the appendix. This proves that there is a constant C such that

$$|R'_7| \le C\tfrac{|t|}{M} \sup_{k,q} \|b(\cdot, k, q)\|_{\mathcal{S},d+3}. \tag{4.31}$$

Therefore, $I^4_< = I^5_< + R'_6 + R'_7$, where

$$I^5_{<} = \int_{|p|\le M} \mathrm{d}p \int_{\mathbb{R}^3} \mathrm{d}q\, \overline{\hat{b}(p, 0, \hat{q})}\, \mathrm{e}^{-\mathrm{i}tp\cdot\nabla\omega_0(\hat{q})}\, \overline{\hat{g}^{\varepsilon,\rho,M}(q_-)} \hat{g}^{\varepsilon,\rho,M}(q_+) \tag{4.32}$$

and $\lim_{M\to\infty} \sup_{\rho,\varepsilon} |R'_6 + R'_7| = 0$ for all t.

Let

$$b_t(x, \hat{q}) = b(x + t\tfrac{1}{2\pi}\nabla\omega_0(\hat{q}), 0, \hat{q}). \tag{4.33}$$

Then $b_t \in \mathcal{S}(\mathbb{R}^3 \times S^2)$, and it has an extension to a function $J_t \in \mathcal{S}_6$; i.e., there is J_t such that $J_t(x,q) = b_t(x,q)$ for all $|q| = 1$. On the other hand, then

$$I^5_< = \int_{|p|\le M} \mathrm{d}p \int_{\mathbb{R}^3} \mathrm{d}q\, \overline{\hat{b}_t(p,\hat{q})}\, \overline{\hat{g}^{\varepsilon,\rho,M}(q_-)} \hat{g}^{\varepsilon,\rho,M}(q_+) = \langle J_t, \Lambda^{\varepsilon,\rho,M}\rangle, \tag{4.34}$$

where $\Lambda^{\varepsilon,\rho,M} \in \mathcal{S}_6'$ denotes the distribution

$$J \mapsto \langle J, \Lambda^{\varepsilon,\rho,M}\rangle = \int_{|p|\le M} \mathrm{d}p \int_{\mathbb{R}^3} \mathrm{d}q\, \overline{\hat{J}(p,\hat{q})}\, \overline{\hat{g}^{\varepsilon,\rho,M}(q_-)} \hat{g}^{\varepsilon,\rho,M}(q_+). \tag{4.35}$$

Clearly, each $\Lambda^{\varepsilon,\rho,M}$ has support in $\mathbb{R}^3 \times S^2$, and there is a constant C such that for all J, ε, ρ, M

$$|\langle J, \Lambda^{\varepsilon,\rho,M}\rangle| \le C \|\hat{g}^{\varepsilon,\rho,M}\|^2 \|J\|_{\mathcal{S},d+1}, \tag{4.36}$$

with $\|\hat{g}^{\varepsilon,\rho,M}\|$ denoting the $L^2(\mathbb{R}^3)$-norm. However, since we have $\|\hat{g}^{\varepsilon,\rho,M}\| \le \|\hat{f}^{\varepsilon,\rho}\|$, where $\|\hat{f}^{\varepsilon,\rho}\|$ is bounded in ε, the Banach–Alaoglu theorem implies that the family $(\Lambda^{\varepsilon,\rho,M})_{\varepsilon,\rho,M}$ belongs to a weak-$*$ sequentially compact set. Therefore, for every ρ, M there is a subsequence of (ε) and $\Lambda^{\rho,M}$ such that $\Lambda^{\varepsilon,\rho,M} \overset{*}{\rightharpoonup} \Lambda^{\rho,M}$ along this subsequence. In addition, for every ρ there is Λ^ρ and a sequence of integers M such that $\Lambda^{\rho,M} \overset{*}{\rightharpoonup} \Lambda^\rho$ along this sequence. Finally, there are $\Lambda \in \mathcal{S}_6'$ and a sequence of integers $N \ge 4$ such that for $\rho = \frac{1}{N}$, $\Lambda^\rho \overset{*}{\rightharpoonup} \Lambda$.

All of the above distributions clearly must have support on $\mathbb{R}^3 \times S^2$. We will soon prove that, in addition, for all $f \in \mathcal{S}_6$ and ρ

$$\langle |f|^2, \Lambda^\rho\rangle \ge 0. \tag{4.37}$$

This implies that then also $\langle |f|^2, \Lambda\rangle \ge 0$. Therefore, by the Bochner–Schwartz theorem, there is a positive Radon measure μ^{H} on $\mathbb{R}^6$ such that for all test functions J, $\langle J, \Lambda\rangle = \int \mu^{\mathrm{H}}(\mathrm{d}x, \mathrm{d}k)\, J(x,k)$. Since also μ^{H} must have support on $\mathbb{R}^3 \times S^2$, we can thus identify it with a positive Radon measure μ^{H}_0 on $\mathbb{R}^3 \times S^2$. By considering test functions $J(x,k) = \mathrm{e}^{-\delta^2 x^2}$ in the limit $\delta \to 0$, it is also clear that μ^{H}_0 must be bounded. We then define the positive, bounded Radon measures μ^{H}_t, $t \in \mathbb{R}$, by the formula (3.6). It follows from the construction of μ^{H}_t that (4.5) holds along the above sequences ρ, ε.

The main missing ingredient is provided by the following lemma.

LEMMA 4.1. *For $p, q \in \mathbb{R}^3$, let us denote $q_\pm = q \pm \frac{p}{2}$, $\hat{q}_\pm = q_\pm/|q_\pm|$, and $\hat{q} = q/|q|$. There is a constant C such that for all $q, q' \in \mathbb{R}^3$ and $f \in \mathcal{S}_6$*

$$\left| \int_{\mathbb{R}^3} \mathrm{d}x\, \mathrm{e}^{-2\pi \mathrm{i} p\cdot x} \overline{f(x,q')} f(x,q) \right| \le C \langle p\rangle^{-d-1} \|f\|^2_{\mathcal{S},d+1}. \tag{4.38}$$

If, in addition, $|q| \ge M$ and $|p| \le M$, then also

$$\left| \mathcal{F}_{x\to p}(|f|^2)(p,\hat{q}) - \int_{\mathbb{R}^3} \mathrm{d}x\, \mathrm{e}^{-2\pi \mathrm{i} p\cdot x} \overline{f(x,\hat{q}_+)} f(x,\hat{q}_-) \right| \le \tfrac{1}{M} C \langle p\rangle^{-d-1} \|f\|^2_{\mathcal{S},d+2}. \tag{4.39}$$

Before proving the lemma we demonstrate that it implies the inequality (4.37). Let $f \in \mathcal{S}_6$ be arbitrary. Define $q_\pm, \hat{q}_\pm$ as in Lemma 4.1, except that here also let $\hat{0} = 0$, and let

$$I^{\varepsilon,\rho,M,f} = \int_{\mathbb{R}^3} \mathrm{d}p \int_{\mathbb{R}^3} \mathrm{d}q \left[\int_{\mathbb{R}^3} \mathrm{d}x\, \mathrm{e}^{2\pi \mathrm{i} p\cdot x} f(x,\hat{q}_+) \overline{f(x,\hat{q}_-)} \right] \overline{\hat{g}^{\varepsilon,\rho,M}(q_-)} \hat{g}^{\varepsilon,\rho,M}(q_+). \tag{4.40}$$

Note that this integral is well-defined by (4.38). Then, by the estimate (4.39) and uniform boundedness of $g^{\varepsilon,\rho,M}$, there is a constant c such that

$$\big|\langle |f|^2, \Lambda^{\varepsilon,\rho,M}\rangle - I^{\varepsilon,\rho,M,f}\big| \leq c\|f\|^2_{\mathcal{S},d+2}\Big(\int_{|p|\geq M} \mathrm{d}p\, \langle p\rangle^{-d-1} + \tfrac{1}{M}\Big). \tag{4.41}$$

On the other hand, $I^{\varepsilon,\rho,M,f} \geq 0$ always. To see this, consider first the case when $g^{\varepsilon,\rho,R} \in \mathcal{S}$. Then $\hat{g}^{\varepsilon,\rho,R} \in \mathcal{S}$, and by changing variables from (q,p) to (q_+,q_-) in (4.40) and then using Fubini's theorem to reorder the integrals, we find that

$$I^{\varepsilon,\rho,M,f} = \int_{\mathbb{R}^3} \mathrm{d}x |G(x)|^2, \qquad \text{with } G(x) = \int_{\mathbb{R}^3} \mathrm{d}q\, \mathrm{e}^{2\pi \mathrm{i} q\cdot x} f(x,\hat{q})\hat{g}^{\varepsilon,\rho,R}(q) \in L^2. \tag{4.42}$$

Since $I^{\varepsilon,\rho,M,f}$ depends L^2-continuously on $g^{\varepsilon,\rho,R}$ this implies that also for general $g^{\varepsilon,\rho,R} \in L^2$ we have $I^{\varepsilon,\rho,M,f} \geq 0$. Since the right-hand side of (4.41) vanishes if first $\varepsilon \to 0$ and then $M \to \infty$, we must thus also have $\langle |f|^2, \Lambda^\rho\rangle \geq 0$. This proves (4.37).

The only remaining task is to prove Lemma 4.1. Consider first (4.38). If $|p| \leq 1$, we have trivially a bound

$$\int_{\mathbb{R}^3} \mathrm{d}x\, |f(x,q')|\, |f(x,q)| \leq \|f\|^2_{\mathcal{S},d+1} \int_{\mathbb{R}^3} \mathrm{d}x\, \langle x\rangle^{-2d-2} \leq c\|f\|^2_{\mathcal{S},d+1}. \tag{4.43}$$

If $|p| \geq 1$, we perform N partial integrations in the direction of p, i.e., in the direction $\hat{p} = \frac{p}{|p|}$, yielding

$$\int_{\mathbb{R}^3} \mathrm{d}x\, \mathrm{e}^{-2\pi\mathrm{i}p\cdot x}\overline{f(x,q')}f(x,q) = \tfrac{1}{(2\pi\mathrm{i}|p|)^N} \int_{\mathbb{R}^3} \mathrm{d}x\, \mathrm{e}^{-2\pi\mathrm{i}p\cdot x}(\hat{p}\cdot\nabla)^N\big(\overline{f(x,q')}f(x,q)\big). \tag{4.44}$$

By the Leibniz rule,

$$(\hat{p}\cdot\nabla)^N\left(\overline{f(x,q')}f(x,q)\right) = \sum_{n=0}^N \binom{N}{n} (\hat{p}\cdot\nabla)^n \overline{f(x,q')}(\hat{p}\cdot\nabla)^{N-n} f(x,q), \tag{4.45}$$

which is bounded by $c'_N\langle x\rangle^{-N}\|f\|^2_{\mathcal{S},N}$. Choosing $N = d+1$ then yields (4.38) for some constant. Adjusting the constant C so that the bound is true also for $|p| \leq 1$ proves that (4.38) is valid.

To prove (4.39), consider q,p as required in the second part of the lemma. Also let $q(s) = q + \frac{s}{2}p$. Then $|q(s)| \geq |q|/2 > 0$ for all $|s| \leq 1$, and $\hat{q}(s)$ is thus well-defined and smooth. Therefore, for any $g \in \mathcal{S}_6$ and $s_0 \in [-1,1]$,

$$\begin{aligned} g(x,\hat{q}(s_0)) - g(x,\hat{q}) &= \int_0^{s_0} \mathrm{d}s\, \tfrac{\mathrm{d}}{\mathrm{d}s} g(x,\hat{q}(s)) \\ &= \int_0^{s_0} \mathrm{d}s\, \Big(\tfrac{1}{2|q(s)|}p - \tfrac{p\cdot\hat{q}(s)}{2|q(s)|}\hat{q}(s)\Big)\cdot \nabla_q g(x,q)|_{q=\hat{q}(s)}, \end{aligned} \tag{4.46}$$

implying

$$|g(x,\hat{q}_\pm) - g(x,\hat{q})| \leq 2\tfrac{|p|}{|q|} \sup_{|q|=1} |\nabla_q g(x,q)|. \tag{4.47}$$

Also for all $g_1, g_2 \in \mathcal{S}_6$,

$$\begin{aligned}&\overline{g_1(x,\hat{q})}g_2(x,\hat{q}) - \overline{g_1(x,\hat{q}_-)}g_2(x,\hat{q}_+)\\ &\quad= \overline{(g_1(x,\hat{q}) - g_1(x,\hat{q}_-))}g_2(x,\hat{q}) + \overline{g_1(x,\hat{q}_-)}\,(g_2(x,\hat{q}) - g_2(x,\hat{q}_+))\,.\end{aligned} \tag{4.48}$$

Following the steps made in the first part of the proof, and replacing the earlier estimates with the above more accurate ones when necessary, we can conclude that the constant C can be adjusted so that for these values of q, p (4.39) also holds. This completes the proof of (4.5).

Energy equality. The energy equality (3.7) follows by considering a sequence of test functions $a^\delta = \mathrm{e}^{-\delta^2 x^2}$ and taking $\delta \to 0$. To see this, first note that the right-hand side of (3.7) is clearly independent of t, and thus it is enough to consider $t = 0$. Thanks to (2.28) and to the tightness of the sequence ψ_0^ε, we obtain for this particular test function that

$$\lim_{\delta\to 0}\lim_{\varepsilon\to 0}\langle a^\delta, W^\varepsilon[\psi_0^\varepsilon]\rangle = \lim_{\delta\to 0}\lim_{\varepsilon\to 0}\langle a^\delta, e^\varepsilon\rangle = \lim_{\varepsilon\to 0}\|\psi_0^\varepsilon\|^2.$$

Equation (3.3) implies that

$$\begin{aligned}\lim_{\varepsilon\to 0}\langle a^\delta, W^\varepsilon[\psi_0^\varepsilon]\rangle = \int_{\mathbb{R}^3}\mathrm{d}x\, a^\delta(x)|\phi_0(x)|^2 + \int_{\mathbb{R}^3}\int_{\mathbb{T}^3} a^\delta(x)\,\mathrm{d}\mu_0(x,k)\\ + \int_{\mathbb{R}^3}\int_{S^2} a^\delta(x)\,\mathrm{d}\mu_0^{\mathrm{H}}(x,k).\end{aligned}$$

Sending δ to 0 yields that

$$\lim_{\varepsilon\to 0}\|\hat{\psi}_0^\varepsilon\|^2 = \|\phi_0\|^2_{L^2(\mathbb{R}^3)} + \mu_0(\mathbb{R}^3\times\mathbb{T}^3) + \mu_0^{\mathrm{H}}(\mathbb{R}^3\times S^2),$$

and the energy equality has been established. This finishes the proof of Theorem 3.2.

5. Proof of Corollary 3.4. Let I_0 denote the original sequence of ε, and consider an arbitrary subsequence I of I_0. Since $(\psi_0^\varepsilon)_{\varepsilon\in I}$ then also satisfies Assumption 3.1, we can conclude from Theorem 3.2 that for every I there is a subsequence I' such that (3.3) holds for all t with the initial conditions given by some triplet $(\mu_I, \mu_I^{\mathrm{H}}, \phi_I)$. From the construction of the subsequence in the proof of Theorem 3.2, we know that ϕ_I can be chosen as the weak limit of ϕ_0^ε along the subsequence I'. The first assumption thus implies that we can always choose $\phi_I = \phi_0$. Let us also denote $\mu_0 = \mu_{I_0}$ and $\mu_0^{\mathrm{H}} = \mu_{I_0}^{\mathrm{H}}$, and to prove the stated uniqueness, we will prove that $\mu_I = \mu_0$ and $\mu_I^{\mathrm{H}} = \mu_0^{\mathrm{H}}$ for all I.

Consider first an arbitrary $\tilde{a} \in C_c^\infty(\mathbb{R}^3\times\mathbb{T}^3_*)$. Let $a(x,k,q) = \tilde{a}(x,k)$ for $k \neq 0$ and define $a(x,0,q) = 0$. Then a is an admissible test function with $a_0 = 0 = b(k,0,q)$, and for any subsequence I we thus obtain, using the second assumption,

$$\int_{\mathbb{R}^3\times\mathbb{T}^3_*}\mathrm{d}\mu_I(x,k)\,\overline{\tilde{a}(x,k)} = \lim_{\varepsilon\in I}\langle a, W^\varepsilon[\psi_0^\varepsilon]\rangle = \int_{\mathbb{R}^3\times\mathbb{T}^3_*}\mathrm{d}\mu_0(x,k)\,\overline{\tilde{a}(x,k)}. \tag{5.1}$$

Such $\tilde{a}$ are dense in $C_c(\mathbb{R}^3\times\mathbb{T}^3_*)$, and thus $\mu_I = \mu_0$ on $\mathbb{R}^3\times\mathbb{T}^3_*$.

Consider then an arbitrary $b \in C_c^\infty(\mathbb{R}\times S^2)$. Let φ be a smooth cutoff function as in Definition 4.1, and define $a(x,k,q) = b(x, q/|q|)\,(1-\varphi(2q))$. Then a is an admissible

test function and indeed $\lim_{R\to\infty} a(x,k,Rq) = b(x,q)$ for all $|q|=1$. By the already proven results we then find that for any subsequence I

$$\int_{\mathbb{R}^3\times S^2} \mathrm{d}\mu_I^{\mathrm{H}}(x,q)\,\overline{b(x,q)} = \lim_{\varepsilon\to 0}\langle a, W^\varepsilon[\psi_0^\varepsilon]\rangle - \int_{\mathbb{R}^3\times\mathbb{T}^3_*} \mathrm{d}\mu_0(x,k)\,\overline{b(x,\tfrac{k}{|k|})} - \langle a_0, W^{(1)}_{\mathrm{cont}}[\phi_0]\rangle$$

$$= \int_{\mathbb{R}^3\times S^2} \mathrm{d}\mu_0^{\mathrm{H}}(x,q)\,\overline{b(x,q)}\,. \tag{5.2}$$

Therefore, also $\mu_I^{\mathrm{H}} = \mu_0^{\mathrm{H}}$, which concludes the uniqueness part of the proof of the corollary.

Finally, define for $t\neq 0$ the triplet $(\mu_t, \mu_t^{\mathrm{H}}, \phi_t)$ using $(\mu_0, \mu_0^{\mathrm{H}}, \phi_0)$ as the initial data. By the uniqueness proved above, for any subsequence I, (3.3) holds along the subsubsequence I'. As the right-hand side is thus independent of I, this proves that the limit also holds along the original sequence I_0. This completes the proof of the corollary. ∎

Appendix. The proof of Theorem 3.2 uses two simple lemmas, which are provided here. The first lemma summarizes several properties of regular acoustic dispersion relations.

LEMMA A.1. *Let ω be a regular acoustic dispersion relation, as in Definition 2.2, and let λ, A_0, and ω_0 be related to ω as in the definition. Then the following assertions are true.*

1. $\omega \in C^{(3)}(\mathbb{T}^3\setminus\{0\},\mathbb{R})$.
2. $\nabla\lambda(0) = 0$ *and* $A_0 > 0$.
3. *There are constants $C_1, C_2 > 0$ such that for all $\|k\|_\infty \le \frac{3}{4}$,*

$$\omega(k) \ge C_1|k|, \qquad |\nabla\lambda(k)| \le C_2|k|\,. \tag{A.1}$$

   In addition, $\|\nabla\omega\|_\infty < \infty$.
4. *There is C_3 such that if $\varepsilon > 0$, $p\in\mathbb{R}^3$, and $\|k\|_\infty \le \frac{1}{2}$, with $|k| > \varepsilon|p|$, then*

$$|\omega(k+\tfrac{1}{2}\varepsilon p) - \omega(k-\tfrac{1}{2}\varepsilon p) - \varepsilon p\cdot\nabla\omega(k)| \le C_3\varepsilon^2\tfrac{|p|^2}{|k|}. \tag{A.2}$$

5. *There is C_4 such that if $\varepsilon > 0$ and $p,q\in\mathbb{R}^3$ with $|p|, \|q\|_\infty \le \frac{1}{2}\varepsilon^{-1}$, then for $q_\pm = q \pm \frac{1}{2}p$,*

$$|\omega(\varepsilon q_+) - \omega(\varepsilon q_-) - \omega_0(\varepsilon q_+) + \omega_0(\varepsilon q_-)| \le C_4\varepsilon^2|p|\,|q|\,. \tag{A.3}$$

6. *There is C_5 such that if $p,q\in\mathbb{R}^3$ with $q\neq 0$ and $|p|\le|q|$, then for $q_\pm = q\pm\frac{1}{2}p$,*

$$|\omega_0(q_+) - \omega_0(q_-) - p\cdot\nabla\omega_0(\tfrac{q}{|q|})| \le C_5\tfrac{|p|^2}{|q|}. \tag{A.4}$$

Proof. From now on, we consider ω, λ, A_0, and ω_0 satisfying Definition 2.2. The first item is then obvious, and the second one follows from the assumptions, since 0 is a minimum. The second inequality in (A.1) follows by using item 2 and $\|D^2\lambda\|_\infty < \infty$, when by Taylor expansion $|\nabla\lambda(k)| \le C_2|k|$ for all $k\in\mathbb{R}$. To prove the first inequality, we first note that, by continuity, also $\|D^3\lambda\|_\infty < \infty$. Thus there is c' such that for all $k\in\mathbb{R}^3$

$$|\lambda(k) - \lambda_0(k)| \le c'|k|^3, \tag{A.5}$$

where $\lambda_0(k) = \frac{1}{2}k \cdot A_0 k$. Since $A_0 > 0$, there is $c > 0$ such that $\lambda_0(k) \geq c^2 k^2$ for all k. Thus there is $\delta' > 0$ such that for all $|k| < \delta'$, we have $|1 - \lambda(k)/\lambda_0(k)| \leq \frac{3}{4}$, and for these k therefore also $\omega(k) = \sqrt{\lambda(k) - \lambda_0(k) + \lambda_0(k)} \geq \frac{1}{2}\sqrt{\lambda_0(k)} \geq \frac{c}{2}|k|$. Since λ has no zeroes in the complement set, the constant can then be adjusted so that (A.1) holds for all k with $\|k\|_\infty \leq \frac{3}{4}$.

We still need to prove the third property, boundedness of $\nabla\omega$. For later use, let us, more generally, consider a nonnegative $f \in C^{(2)}(\mathbb{R}^3)$, $q \in \mathbb{R}^3$, and k such that $f(k) \neq 0$. Then we have

$$(q \cdot \nabla)\sqrt{f(k)} = \frac{1}{2\sqrt{f(k)}} q \cdot \nabla f(k), \tag{A.6}$$

$$(q \cdot \nabla)^2\sqrt{f(k)} = -\frac{1}{4f(k)^{3/2}}(q \cdot \nabla f(k))^2 + \frac{1}{2\sqrt{f(k)}}(q \cdot \nabla)^2 f(k). \tag{A.7}$$

This implies that there is a constant C such that for all $q \in \mathbb{R}^3$ and $k \neq 0$, with $\|k\|_\infty \leq \frac{3}{4}$,

$$|q \cdot \nabla\omega(k)|, |q \cdot \nabla\omega_0(k)| \leq C|q|\,, \tag{A.8}$$

$$|(q \cdot \nabla)^2\omega(k)|, |(q \cdot \nabla)^2\omega_0(k)| \leq C\frac{|q|^2}{|k|}\,. \tag{A.9}$$

In particular, by periodicity therefore $\|\nabla\omega\|_\infty < \infty$.

To prove item 4, consider k, p, ε as in the claim, and define a function $f(s) = \omega(k_+) - \omega(k_-)$ with $k_\pm = k \pm s\frac{1}{2}p$ and $s \in [-\varepsilon, \varepsilon]$. Then $\|k_\pm\|_\infty \leq \frac{3}{2}\|k\|_\infty$ and $|k_\pm| \geq |k| - \varepsilon\frac{1}{2}p \geq \frac{1}{2}|k| > 0$, and thus f belongs to $C^{(3)}$. In particular, $f(0) = 0$, $f'(0) = p \cdot \nabla\omega(k)$, and

$$\begin{aligned} f''(s) = \frac{1}{2}\Big[&\frac{1}{\omega(k_+)}\Big(\frac{p}{2} \cdot \nabla\Big)^2\lambda(k_+) - \frac{1}{\omega(k_-)}\Big(\frac{p}{2} \cdot \nabla\Big)^2\lambda(k_-) \\ &- \frac{1}{2\omega(k_+)^3}\Big(\frac{p}{2} \cdot \nabla\lambda(k_+)\Big)^2 + \frac{1}{2\omega(k_-)^3}\Big(\frac{p}{2} \cdot \nabla\lambda(k_-)\Big)^2\Big]. \end{aligned} \tag{A.10}$$

Using item 3, we find that there is C such that this is uniformly bounded by $C|p|^2/|k|$. Then a Taylor expansion at the origin proves item 4.

Since for all $k \neq 0$

$$\omega(k) - \omega_0(k) = \sqrt{\lambda(k)} - \sqrt{\lambda_0(k)} = \frac{\lambda(k) - \lambda_0(k)}{\sqrt{\lambda(k)} + \sqrt{\lambda_0(k)}}\,, \tag{A.11}$$

we then also have

$$q \cdot \nabla(\omega(k) - \omega_0(k)) = \frac{q \cdot \nabla\lambda(k) - q \cdot \nabla\lambda_0(k)}{\omega(k) + \omega_0(k)} - (\lambda(k) - \lambda_0(k))\frac{q \cdot \nabla\omega(k) + q \cdot \nabla\omega_0(k)}{(\omega(k) + \omega_0(k))^2}\,. \tag{A.12}$$

By Taylor expansion at the origin, we find that there is C' such that for all $q \in \mathbb{R}^3$,

$$|\lambda(k) - \lambda_0(k)| \leq C'|k|^3, \qquad |q \cdot \nabla\lambda(k) - q \cdot \nabla\lambda_0(k)| \leq C'|k|^2|q|, \tag{A.13}$$

$$|(q \cdot \nabla)^2\lambda(k) - (q \cdot \nabla)^2\lambda_0(k)| \leq C'|k|\,|q|^2\,. \tag{A.14}$$

Thus there is also C such that for all $q \in \mathbb{R}^3$, and for $\|k\|_\infty \leq \frac{3}{4}$ with $k \neq 0$,

$$|q \cdot \nabla(\omega(k) - \omega_0(k))| \leq C|k|\,|q|\,, \qquad |(q \cdot \nabla)^2(\omega(k) - \omega_0(k))| \leq C|q|^2. \tag{A.15}$$

Let us then consider q, p, ε satisfying the assumptions made in the final item. Since then $\|\varepsilon q_\pm\|_\infty \leq \frac{3}{4}$, we can apply the previous estimates. If $p = 0$, then $q_- = q_+$ and the bound in (A.3) is trivially valid for any C_4. Consider thus $p \neq 0$, and assume first that p is not proportional to q. Since then the line segment $[0,1] \ni s \mapsto \varepsilon q + s\varepsilon p/2$ does not pass through the origin, the function $s \mapsto \omega(\varepsilon q + s\varepsilon p/2) - \omega_0(\varepsilon q + s\varepsilon p/2)$ is in $C^{(3)}([0,1])$. We make a Taylor expansion of this function at $s = 0$, yielding

$$\omega(\varepsilon q + \varepsilon\tfrac{p}{2}) - \omega_0(\varepsilon q + \varepsilon\tfrac{p}{2}) = \omega(\varepsilon q) - \omega_0(\varepsilon q) + \varepsilon\tfrac{p}{2} \cdot (\nabla\omega(\varepsilon q) - \nabla\omega_0(\varepsilon q)) + R\,. \tag{A.16}$$

Here (A.15) implies $|R| \leq C\varepsilon^2|p|^2$, since

$$R = \int_0^1 \mathrm{d}s\,(1-s)\Big(\varepsilon\tfrac{p}{2} \cdot \nabla\Big)^2 (\omega(\varepsilon q + s\varepsilon\tfrac{p}{2}) - \omega_0(\varepsilon q + s\varepsilon\tfrac{p}{2}))\,. \tag{A.17}$$

This proves that (A.3) holds in this case for some constant C_4. In the final case $p \propto q$, we choose a direction u orthogonal to q, and use the previous estimate with $p + \delta u$ instead of p for an arbitrary $0 < \delta \leq 1$. Since the left-hand side of (A.3) is continuous in δ, the bound must then hold also for $\delta = 0$, proving the validity of the estimate also in this case.

To prove (A.4), we use the fact that by assumption $|q_\pm| \geq \frac{1}{2}|q| > 0$, and thus denoting $\hat{q} = q/|q|$, we get

$$\omega_0(q_+) - \omega_0(q_-) = |q|\big(\omega_0\big(\hat{q} + \tfrac{p}{2|q|}\big) - \omega_0\big(\hat{q} - \tfrac{p}{2|q|}\big)\big) = p \cdot \nabla\omega_0(\hat{q}) + R, \tag{A.18}$$

where $|R| \leq C|p|^2/|q|$. We have thus completed the proof of the lemma. ∎

The second lemma recalls a well-known fact in Fourier analysis: the Fourier transform of a bounded and tight sequence of L^2 functions converges strongly on compact sets. For the convenience of the reader we give a proof.

LEMMA A.2. *Let $f^\varepsilon \in L^2(\mathbb{R}^d)$ be a bounded and tight sequence of functions such that $f^\varepsilon \rightharpoonup 0$ as $\varepsilon \to 0$. Then for every $\Omega \subset \mathbb{R}^d$ with a finite measure*

$$\lim_{\varepsilon\to 0} \|\hat{f}^\varepsilon\|_{L^2(\Omega)} = 0. \tag{A.19}$$

Proof. Let $\beta > 0$ be arbitrary. By tightness of f^ε there is $R_\beta > 0$ such that

$$\limsup_{\varepsilon\to 0} \int_{|x|\geq R_\beta} \mathrm{d}x\,|f^\varepsilon(x)|^2 \leq \beta^2. \tag{A.20}$$

Define now

$$f^{\varepsilon,\beta}(x) = \begin{cases} f^\varepsilon(x) & \text{if } |x| \leq R_\beta, \\ 0, & \text{otherwise.} \end{cases} \tag{A.21}$$

By the boundedness of $f^{\varepsilon,\beta}$ in $L^2(\mathbb{R}^d)$ there exist g^β and a subsequence (ε') such that $f^{\varepsilon',\beta} \rightharpoonup g^\beta$ in $L^2(\mathbb{R}^d)$ as $\varepsilon' \to 0$. Estimate (A.20) implies that $\limsup_{\varepsilon\to 0} \|f^\varepsilon - f^{\varepsilon,\beta}\|^2_{L^2(\mathbb{R}^3)} \leq \beta^2$. Since also $f^\varepsilon \rightharpoonup 0$, we have

$$\|g^\beta\|^2 = \lim_{\varepsilon'} |\langle g^\beta, f^{\varepsilon',\beta}\rangle| \leq \|g^\beta\| \limsup_{\varepsilon\to 0} \|f^\varepsilon - f^{\varepsilon,\beta}\| \tag{A.22}$$

and, therefore, $\|g^\beta\|_{L^2(\mathbb{R}^d)} \leq \beta$.

Since g^β has support in a ball of radius R_β, also $g_1^\beta(x) = -2\pi i x g^\beta(x) \in L^2(\mathbb{R}^d)$. Therefore, $\hat{g}^\beta \in H^1(\mathbb{R}^d)$ and $\nabla \hat{g}^\beta = \hat{g}_1^\beta$. Similarly, also $\hat{f}^{\varepsilon,\beta} \in H^1(\mathbb{R}^d)$ and it is straightforward to check that $\nabla \hat{f}^{\varepsilon',\beta} \rightharpoonup \nabla \hat{g}^\beta$ in L^2. Since Ω is assumed to be a set of finite measure, we then have $\lim_{\varepsilon'\to 0} \|\hat{f}^{\varepsilon',\beta} - \hat{g}^\beta\|_{L^2(\Omega)} = 0$ (for a proof, see, for instance, Theorem 8.6. in [13]). Collecting all the above estimates together proves that

$$\begin{aligned}&\limsup_{\varepsilon'\to 0} \|\hat{f}^{\varepsilon'}\|_{L^2(\Omega)} \\ &\quad \le \limsup_{\varepsilon'\to 0} \|\hat{f}^{\varepsilon'} - \hat{f}^{\varepsilon',\beta}\|_{L^2(\Omega)} + \limsup_{\varepsilon'\to 0} \|\hat{f}^{\varepsilon',\beta} - \hat{g}^\beta\|_{L^2(\Omega)} + \|\hat{g}^\beta\|_{L^2(\Omega)} \le 2\beta.\end{aligned}$$

Since β can be arbitrarily small, there must be a subsequence (ε'') such that $\lim_{\varepsilon''\to 0} \|\hat{f}^{\varepsilon''}\|_{L^2(\Omega)} = 0$.

Since the assumptions on the sequence f^ε are preserved for subsequences, we can consider an arbitrary subsequence and apply the above result to it. Then we can conclude that for every subsequence there is a subsubsequence along which (A.19) holds. This implies that the limit (A.19) actually holds also along the original sequence. ∎

Acknowledgments. The authors would like to thank Clotilde Fermanian Kammerer, Alexander Mielke, and Herbert Spohn for several instructive discussions. We are also grateful to the anonymous referees for a careful reading of the manuscript and for suggesting several useful references.

REFERENCES

[1] N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Holt, Rinehart and Winston, New York, 1976.

[2] F. Bonetto, J. L. Lebowitz, and L. Rey-Bellet, *Fourier's law: A challenge to theorists*, in Mathematical Physics 2000, A. Fokas, A. Grigoryan, T. Kibble, and B. Zegarlinski, eds., Imperial College Press, London, 2000, pp. 128–150.

[3] T. V. Dudnikova and H. Spohn, *Local stationarity for lattice dynamics in the harmonic approximation*, Markov Process. Related Fields, 12 (2006), pp. 645–678.

[4] C. Fermanian Kammerer, *Measure semi-classiques et équation de la chaleur*, Ph.D. thesis, Université Paris-Sud, Orsay, 1995.

[5] C. Fermanian Kammerer, *Mesures semi-classiques 2-microlocales*, C. R. Acad. Sci. Paris, 331 (2000), pp. 515–518.

[6] C. Fermanian Kammerer, *Propagation and absorption of concentration effects near shock hypersurfaces for the heat equation*, Asymptot. Anal., 24 (2000), pp. 107–141.

[7] C. Fermanian Kammerer, *Two scale analysis of a bounded family in L^2 on a submanifold of the phase space*, C. R. Acad. Sci. Paris, 340 (2005), pp. 269–274.

[8] N. B. Firoozye, *Homogenization on Lattices: Small Parameter Limits, H-Measures, and Discrete Wigner Measures*, IMA Preprint 1177, University of Minnesota, Minneapolis, MN, 1993; available online from http://www.ima.umn.edu/preprints/OCTOBER1993/1177.pdf.

[9] G. Francfort and P. Gérard, *The wave equation on a thin domain: Energy density and observability*, J. Hyperbolic Differential Equations, 1 (2004), pp. 351–366.

[10] P. Gérard, *Compacité par compensation et régularité 2-microlocale*, in Séminar Equation aux Dérivées Partielles, Exp. No. VII, Ecole Polytechnique, Palaiseau, 1988–1989.

[11] P. Gérard and E. Leichtnam, *Ergodic properties of eigenfunctions for the Dirichlet problem*, Duke Math. J., 71 (1993), pp. 559–607.

[12] P. Gérard, P. A. Markowich, N. J. Mauser, and F. Paupaud, *Homogenization limits and Wigner transforms*, Commun. Pure Appl. Math., 50 (1997), pp. 323–379.

[13] E. H. Lieb and M. Loss, *Analysis*, 2nd ed., AMS, Providence, RI, 2001.

[14] J. Lukkarinen and H. Spohn, *Kinetic limit for wave propagation in a random medium*, Arch. Ration. Mech. Anal., 183 (2007), pp. 93–162.

[15] F. Macià, *Propagatión y control de vibraciones en medios discretos y continuous*, Ph.D. thesis, Universidad Complutense de Madrid, Madrid, Spain, 2002.

[16] F. Macià, *Wigner measures in the discrete setting: High-frequency analysis of sampling and reconstruction operators*, SIAM J. Math. Anal., 36 (2004), pp. 347–383.
[17] A. Mielke, *Macroscopic behavior of microscopic oscillations in harmonic lattices via Wigner-Husimi transforms*, Arch. Ration. Mech. Anal., 181 (2006), pp. 401–448.
[18] F. Nier, *A semi-classical picture of quantum scattering*, Ann. Sci. École Norm. Sup., 29 (1996), pp. 149–183.
[19] L. Ryzhik, G. Papanicolaou, and J. B. Keller, *Transport equations for elastic and other waves in random media*, Wave Motion, 24 (1996), pp. 327–370.
[20] H. Spohn, *The phonon Boltzmann equation, properties and link to weakly anharmonic lattice dynamics*, J. Statist. Phys., 124 (2006), pp. 1041–1104.
[21] H. Spohn, *Erratum on "The phonon Boltzmann equation, properties and link to weakly anharmonic lattice dynamics,"* J. Statist. Phys., 123 (2006), p. 707.
[22] H. Spohn, *Collisional invariants for the phonon Boltzmann equation*, J. Statist. Phys., 124 (2006), pp. 1131–1135.
[23] L. Tartar, *H-measures, a new approach for studying homogenisation, oscillations and concentration effects in partial differential equations*, Proc. Roy. Soc. Edinburgh Sect. A, 115 (1990), pp. 193–230.
[24] S. Teufel and G. Panati, *Propagation of Wigner functions for the Schrödinger equation with a perturbed periodic potential*, in Multiscale Methods in Quantum Mechanics, Ph. Blanchard and G. Dell'Antonio, eds., Birkhäuser, Boston, 2004, pp. 207–220.
[25] J. M. Ziman, *Electrons and Phonons: The Theory of Transport Phenomena in Solids*, Oxford University Press, London, 1967.

SIAM J. MATH. ANAL.
Vol. 40, No. 4, pp. 1419–1440

CONCENTRATION PHENOMENA IN A NONLOCAL QUASI-LINEAR PROBLEM MODELLING PHYTOPLANKTON I: EXISTENCE*

YIHONG DU† AND SZE-BI HSU‡

Abstract. We study the positive steady state of a quasi-linear reaction-diffusion system in one space dimension introduced by Klausmeier and Litchman for the modelling of the distributions of phytoplankton biomass and its nutrient. The system has nonlocal dependence on the biomass function, and it has a biomass-dependent drifting term describing the active movement of the biomass towards the location of the optimal growth condition. We obtain complete descriptions of the profile of the solutions when the coefficient of the drifting term is large, rigorously proving the numerically observed phenomenon of concentration of biomass for this model. Our theoretical results reveal four critical numbers for the model not observed before and offer several further insights into the problem being modelled. This is Part I of a two-part series, where we obtain nearly optimal existence and nonexistence results. The asymptotic profile of the solutions is studied in the separate Part II.

Key words. quasi-linear, nonlocal dependence, phytoplankton, concentration phenomenon, reaction-diffusion equation

AMS subject classifications. 35J55, 35J65, 92D25

DOI. 10.1137/07070663X

1. Introduction.

In this paper, we study the problem

$$\begin{cases} -[d_1 u_x + \sigma c(x)u]_x = [g(x) - m]u, & 0 < x < 1, \\ -d_2 v_{xx} = -g(x)u, & 0 < x < 1, \\ d_1 u_x + \sigma c(x)u = 0, & x = 0, 1, \\ v_x(0) = 0, \quad v_x(1) = \beta[v_0 - v(1)], \end{cases} \tag{1.1}$$

where d_1, d_2, σ, m, v_0, and β are positive constants,

$$g(x) = f(\min\{\alpha v(x), w(x)\}), \quad f(s) = \frac{rs}{K_I + s},$$

and

$$w(x) = w_0 \exp\left[- A_0 x - A \int_0^x u(s)ds \right],$$

with α, r, K_I, w_0, A, and A_0 positive constants. We note that the right-hand sides of the differential equations in (1.1) depend on the unknown functions u and v in a nonlocal manner. Moreover, the positive function $c(x)$ is determined by u and v in a rather unconventional way to be explained below. We are interested in positive

*Received by the editors October 29, 2007; accepted for publication (in revised form) June 24, 2008; published electronically November 5, 2008. This research was partially supported by the Australian Research Council, the NCTS, and the National Science Council of Taiwan.
http://www.siam.org/journals/sima/40-4/70663.html
†Department of Mathematics, School of Science and Technology, University of New England, Armidale, NSW2351, Australia (ydu@turing.une.edu.au), and Department of Mathematics, Qufu Normal University, People's Republic of China.
‡Department of Mathematics, National Tsing-Hua University, Hsinchu, Taiwan 300, Republic of China (sbhsu@math.nthu.edu.tw).

solutions of (1.1), namely $u > 0$ and $v > 0$ in $[0, 1]$. From (1.1) it is easy to see that for any such solution, v is an increasing function. Clearly w is a decreasing function. The function $c(x)$ is defined by

$$c(x) = \frac{x - x_0}{\delta + |x - x_0|},$$

where $\delta > 0$ is a small constant and $x_0 \in [0, 1]$ is the intersection point of the functions $\alpha u(x)$ and $w(x)$ whenever such an intersection occurs in $[0, 1]$; if $\alpha u(x)$ and $v(x)$ do not intersect, then $x_0 = 0$ if $\alpha v > w$ on $[0, 1]$, and $x_0 = 1$ if $\alpha v < w$ on [0,1]. In other words, x_0 is given by the following description:

$$\min\{\alpha v(x), w(x)\} = \alpha v(x)\ \forall x \in [0, x_0); \quad \min\{\alpha v(x), w(x)\} = w(x)\ \forall x \in (x_0, 1].$$

This unconventional dependence of c on the unknown solution (u, v) makes (1.1) a very special quasi-linear problem.

Such a system arises in the mathematical modelling of phytoplankton in a one-dimensional water column, where $u(x)$ represents the distribution of phytoplankton biomass, $v(x)$ stands for the distribution of nutrient, and x denotes the depth in the water column, with $x = 0$ at the surface and $x = 1$ at the bottom. The term $\sigma c(x)$ is used to describe the active movement of the biomass towards the spatial location with the optimal growth condition. Klausmeier and Litchman [KL] propose using this model to study the concentration phenomenon widely observed for phytoplankton in lakes and oceans. Their numerical analysis in [KL] demonstrates that for large σ, the biomass function $u(x)$ concentrates at a certain level $x = x_*$ while the nutrient function $v(x)$ is close to a piecewise linear function. They then treat u as a constant multiple of the δ-function concentrating at x_* and propose a game theoretical model to determine the location of x_*.

In this paper, we rigorously prove the existence of such a concentration phenomenon and obtain accurate formulas for the determination of x_* and the total biomass. Our theoretical results offer several further insights into the model besides those obtained through numerical analysis in [KL]; for example, we show the existence of four critical values $v_{**} < v_* < v^* < v^{**}$ for v_0 (the nutrient level at the sediment), such that

(i) $x_* = 0$ when $v_0 \geq v^*$, $x_* \in (0, 1)$ when $v_0 \in (v_*, v^*)$, and $x_* = 1$ when $v_0 \leq v_*$;

(ii) the total biomass increases with v_0 in the range $v_{**} < v_0 < v^{**}$, but it stays constant for $v_0 \geq v^{**}$ or $v_0 \leq v_{**}$ (and with v_0 above a certain level so that the biomass can survive).

It turns out that the game theoretical model of [KL] is a simplified version of our equations governing x_* and the total phytoplankton biomass for the case $v_* \leq v_0 \leq v^*$. A more detailed description of these results is given in the introduction of Part II, with their biological interpretations given in section 4 there.

To explain this model more precisely, we start by a brief description of the background and motivation of this research. Phytoplankton, the generic name of microorganisms living in lakes and oceans, is the basis of the aquatic food chain. Its importance for the proper functioning of the aquatic ecosystem has long been recognized, and its behavior has been widely studied. The distribution of phytoplankton in lakes and oceans is highly heterogeneous. To better understand this property of the phytoplankton, various mathematical models have been proposed and numerically analyzed; see, for example, [EATSH, KL, PT, PTHS, HTKS]. However, little

rigorous mathematical analysis is available. In [YN], an ordinary differential equation model for the vertical distributions of phytoplankton is theoretically analyzed; see also [IT, BFH, BFHK] for earlier related research. It is our hope that the current paper may induce further rigorous mathematical research in this direction and that the techniques developed here may find more applications.

We now describe the model in more detail. In poorly mixed water columns, it has been observed that algae can be heterogeneously distributed, with thin layers of biomass on the surface, at depth, or on the sediment surface; examples for each of these cases can be found in [KL]. To model these phenomena, [KL] proposes a reaction-diffusion-taxis model of phytoplankton, nutrients, and light, based on the principle of light and nutrient competition. They use the following system to describe the distribution of phytoplankton in a one-dimensional water column, with depth represented by $0 \le z \le z_b$; $z = 0$ at the surface and $z = z_b$ at the bottom:

$$\text{(1.2)}\qquad \begin{cases} b_t = D_b b_{zz} + [\nu(g^0_z)b]_z + [g^0 - m]b, & 0 < z < z_b, t > 0, \\ R_t = D_R R_{zz} - bg^0/Y + \epsilon mb/Y, & 0 < z < z_b, t > 0, \\ D_b b_z + \nu(g^0_z)b = 0, & z = 0, z_b, t > 0, \\ R_z(t,0) = 0, R_z(t,z_b) = h[R_{in} - R(t,z_b)], & t > 0, \\ I(t,z) = I_{in} \exp\Big(- \displaystyle\int_0^z [ab(t,s) + a_{bg}]ds\Big), & 0 < z < z_b, t > 0, \end{cases}$$

where a, a_{bg}, h, D_b, D_R, m, I_{in}, R_{in}, and Y are positive constants, $\epsilon \in [0,1)$, $\nu(s)$ is an odd decreasing function that approaches $\nu_{max} > 0$ as $s \to -\infty$, and

$$g^0(t,z) = \min\{f_I(I(t,z)), f_R(R(t,z))\},$$

with

$$f_I(s) = r\frac{s}{s+K_I}, \quad f_R(s) = r\frac{s}{s+K_R}, \quad r, K_I, K_R > 0.$$

In (1.2), $b(t,z)$ denotes the distribution of the phytoplankton biomass, $R(t,z)$ represents the nutrient distribution, and $I(t,z)$ stands for the distribution of light. The constant I_{in} is the light distribution at the surface, and, by the Lambert–Beer law, light at depth z is given by

$$I(t,z) = I_{in} \exp\Big(- \int_0^z [ab(t,s) + a_{bg}]ds\Big),$$

where a and a_{bg} are, respectively, the phytoplankton and background attenuation coefficients. In this model, it is assumed that the change in phytoplankton biomass at depth z results from three processes: growth, loss, and movement. The functions $f_I(I)$ and $f_R(R)$ are the phytoplankton growth rate when only one of the resources I and R is limited (the other being regarded as sufficient). By Liebig's law of the minimum for essential resources, the gross phytoplankton growth rate is given by $g^0(t,z) = \min\{f_I(I(t,z)), f_R(R(t,z))\}$. Biomass is lost at density-dependent rate m, representing respiration, death, and grazing. D_b is the passive diffusion rate of the biomass, while $[\nu(g^0_z)b]_z$ describes active movement of the biomass towards a spatial location (i.e., depth) with a better growth condition. The no-flux boundary condition for b means that no phytoplankton enters or leaves the water column at $z = 0$ and $z = z_b$. The equation for R is based on the assumption that nutrients in

the water column are mixed with eddy diffusion with diffusion coefficient D_R and are consumed by phytoplankton at the rate $-bg^0/Y$, and the term $\epsilon mb/Y$ means that ϵ proportion of the nutrients in dead phytoplankton is immediately recycled. Here Y describes the yield of phytoplankton biomass per unit nutrient consumed. The boundary condition for R means that nutrients do not leave or enter the top of the water column but are supplied at the bottom, with nutrients in the sediments fixed at constant concentration R_{in}, which diffuse across the sediment-water interface at a rate proportional to the concentration difference across the interface; the parameter h describes the permeability of the interface.

In [KL], taking $\epsilon = 0$ and $\nu(s) = \nu_0(s) := -\nu_{max}\mathrm{sgn}(s)$ (where $\mathrm{sgn}(s)$ is the sign function, which equals 1, -1, or 0 according to whether $s > 0$, $s < 0$, or $s = 0$), the equilibrium distributions of b, R, and I are calculated numerically for various parameter values (see Table 1 and Figure 1 in [KL]). The numerical simulation in [KL] shows that as ν_{max} increases, the biomass distribution concentrates at a certain depth $z = z^*$. Further, based on intuition and formal analysis, a game theoretical approach is proposed in [KL], which can be used to calculate z^*. Though the connection between (1.2) and the simplified game theoretical approach is not rigorously established, the predictions deduced from the game theoretical model agree well with the numerical results based on (1.2); see details in [KL].

In this paper, we theoretically analyze the equilibrium solutions of (1.2). So $b = b(z)$, $R = R(z)$, and $I = I(z)$. Naturally, only positive solutions are of interest to us.

As in [KL], we assume that $\epsilon = 0$. We denote $f(s) = rs/(s+K_I)$ and $\alpha = K_I/K_R$. Then

$$f_I(s) = f(s), \quad f_R(s) = f(\alpha s).$$

Since $f'(s) > 0$ we find that

$$g^0(z) = \min\{f(I(z)), \quad f(\alpha R(z))\} = f(\min\{I(z), \alpha R(z)\}).$$

Clearly $I(z)$ is a decreasing function. Since $D_R R'' = bg^0/Y > 0$ and $R'(0) = 0$, we find that $R'(z) > 0$ for $z \in (0, z_b]$. Therefore $R(z)$ is increasing, and we can always find a unique $z_0 \in [0, z_b]$ such that

$$g^0(z) = f(\alpha R(z)) \ \forall z \in [0, z_0), \quad g^0(z) = f(I(z)) \ \forall z \in (z_0, z_b].$$

Evidently z_0 depends on I and R, and $z_0 = 0$ if $\alpha R(z) \geq I(z)$ on $[0, z_b]$, and $z_0 = z_b$ when $\alpha R(z) \leq I(z)$ on $[0, z_b]$.

In view of the above discussions for $g^0(z)$, we see that

$$\nu_0(g^0_z(z)) = -\nu_{max}\mathrm{sgn}(g^0_z(z)) = -\nu_{max} \ \forall z \in [0, z_0),$$
$$\nu_0(g^0_z(z)) = -\nu_{max}\mathrm{sgn}(g^0_z(z)) = \nu_{max} \ \forall z \in (z_0, z_b].$$

In this paper, we use a continuous approximation of the above step function used in [KL]; namely, we take

$$\nu(g^0_z(z)) = \nu_{\delta'}(z) := \nu_{max}\frac{z - z_0}{\delta' + |z - z_0|}.$$

It is easily seen that $\nu_{\delta'}(z) \to \nu_0(g^0_z(z))$ as $\delta' \to 0$. We stress again that $\nu_{\delta'}$ depends on I and R through the definition of z_0.

Next we normalize the functions in (1.2) by

$$u(x) = b(z_b x)/Y, \quad v(x) = R(z_b x), \quad w(x) = I(z_b x), \quad 0 \le x \le 1,$$

and define

$$d_1 = z_b^2 D_b, \quad d_2 = z_b^2 D_R, \quad \sigma = \nu_{max} z_b, \quad \delta = \delta'/z_b,$$

$$A = a z_b Y, \quad A_0 = a_{bg} z_b, \quad \beta = h z_b, v_0 = R_{in}, \quad w_0 = I_{in}.$$

We denote

$$c(x) = c_{v,w}(x) = \frac{x - x_0}{\delta + |x - x_0|}, \quad x_0 = \frac{z_0}{z_b}.$$

Then after some simple calculations we find that the steady-state version of (1.2) becomes (1.1), or, written in a more comprehensive form,

$$\text{(1.3)} \qquad \begin{cases} -[d_1 u_x + \sigma c(x) u]_x = [f(\min\{\alpha v, w\}) - m]u, & 0 < x < 1, \\ -d_2 v_{xx} = -f(\min\{\alpha v, w\})u, & 0 < x < 1, \\ d_1 u_x(0) + \sigma c(0)u(0) = d_1 u_x(1) + \sigma c(1)u(1) = 0, & \\ v_x(0) = 0, \quad v_x(1) = \beta[v_0 - v(1)], & \\ w(x) = w_0 \exp\left[- A_0 x - A \displaystyle\int_0^x u(s)ds \right], & 0 \le x \le 1. \end{cases}$$

Let us note that from the equation for u and the strong maximum principle, if u is nonnegative on $[0, 1]$, then it is either identically 0 or positive everywhere in $[0, 1]$. It is also easy to see that whenever u is positive, $0 < v < v_0$ in $[0, 1]$.

Since this paper is very long, and the techniques used in the first half of the paper are rather different from those in the second half, we divide it into two separate parts. Part I here is mainly concerned with the existence and nonexistence problem, and Part II studies the asymptotic behavior of the positive solutions as $\sigma \to \infty$.

In section 2, treating m as a parameter, we make use of a bifurcation argument to obtain two critical numbers $0 < m_* \le m^*$ such that (1.3) has no positive solution when $m \ge m^*$ and it has at least one positive solution when $0 < m < m_*$. We also show that as m (the death rate of the biomass) decreases to 0, the biomass blows up everywhere; the exact limiting profiles of the biomass function and the nutrient function as $m \to 0$ are also obtained. In section 3, we show that as $\sigma \to \infty$, m_* and m^* converge to the same limit $f(\min\{\alpha v_0, w_0\})$. This demonstrates that our existence and nonexistence results are sharp for large σ.

The asymptotic behavior of the positive solutions when $\sigma \to \infty$ is investigated separately in Part II (see [DH]), where we fix $0 < m < f(\min\{\alpha v_0, w_0\})$ and study the asymptotic behavior of a positive solution (u_n, v_n) of (1.3) with $\sigma = \sigma_n \to \infty$.

2. Existence and nonexistence results. The function $c(x)$ appearing in (1.3) plays a very important role in our analysis. From the definition of $c(x) = c_{v,w}(x)$ we find that it is well defined if $v(x)$ is increasing in $[0, 1]$ and $w(x)$ is decreasing in $[0, 1]$, and

$$c(x) = \frac{x - x_0}{\delta + |x - x_0|},$$

where $x_0 \in [0,1]$ is uniquely determined by the following:

$$\min\{\alpha v(x), w(x)\} = \alpha v(x)\ \forall x \in [0, x_0); \quad \min\{\alpha v(x), w(x)\} = w(x)\ \forall x \in (x_0, 1].$$

It is easily seen that for the definition of $c_{v,w}(x)$, the requirement that v is increasing and w is decreasing can be relaxed; we can allow one (but not both) of the following:

$$\text{(i) } v \text{ is nondecreasing, (ii) } w \text{ is nonincreasing.}$$

Let us also observe that $c(x)$ is a C^1 function, with $c'(x) = \delta(\delta + |x - x_0|)^{-2}$.

With this in mind, we find that $(u, v, w) = (0, v_0, w_*)$ solves (1.3), where $w_*(x) = w_0 e^{-A_0 x}$. We will call this the trivial solution. To find nontrivial solutions, we now treat m as a parameter and look for special values of m so that positive solutions of (1.3) may bifurcate from this trivial solution. If $m_* \geq 0$ is such a value, then there exist $m_n \to m_*$ and (u_n, v_n, w_n) solving (1.3) with $m = m_n$ such that $u_n > 0$, $v_n > 0$, and $u_n \to 0$, $v_n \to v_0$, and $w_n \to w_*$ and in $C^1[0,1]$ as $n \to \infty$. Now $v_n(x)$ is increasing and $w_n(x)$ is decreasing for $x \in [0,1]$; therefore $c_{v_n,w_n}(x)$ is well defined. Moreover, it is easily checked that $c_{v_n,w_n} \to c_{v_0,w_*}$ in $C([0,1])$ as $n \to \infty$. To simplify the notation, we write $c^n(x) = c_{v_n,w_n}(x)$ and $c^0(x) = c_{v_0,w_*}(x)$. Therefore u_n satisfies

$$\begin{cases} -[d_1 u_n' + \sigma c^n(x) u_n]' = [f(\min\{\alpha v_n, w_n\}) - m_n] u_n & \text{in } (0,1), \\ d_1 u_n' + \sigma c^n(x) u_n = 0 & \text{for } x = 0, 1. \end{cases}$$

Here and in what follows, we use the notation $u' = u_x$, etc. To determine the value of m_*, we first deduce a useful equation from the equation for u_n. So we define $\hat{u}_n = u_n / \|u_n\|_\infty$. Then we have

$$\text{(2.1)} \qquad \begin{cases} -(d_1 \hat{u}_n' + \sigma c^n \hat{u}_n)' + m_n \hat{u}_n = f(\min\{\alpha v_n, w_n\}) \hat{u}_n & \text{in } (0,1), \\ d_1 \hat{u}_n' + \sigma c^n \hat{u}_n = 0 & \text{for } x = 0, 1. \end{cases}$$

Since the right-hand side of the first equation in (2.1) and $\{\hat{u}_n\}$ are both bounded in $L^\infty([0,1])$, and since m_n, c^n, $(c^n)'$ are bounded in $L^\infty([0,1])$, we can use standard L^p theory for elliptic operators (see [GT]) to conclude that $\{\hat{u}_n\}$ is a bounded sequence in $W^{2,p}([0,1])$ for any $p > 1$. By the Sobolev embedding theorem, we see that $\{\hat{u}_n\}$ is compact in $C^1([0,1])$. By passing to a subsequence, we may assume that $\hat{u}_n \to \hat{u}$ in $C^1([0,1])$, and then we easily see that $\hat{u}$ satisfies (in the weak sense)

$$\text{(2.2)} \qquad \begin{cases} -(d_1 \hat{u}' + \sigma c^0 \hat{u})' + m_* \hat{u} = f(\min\{\alpha v_0, w_*\}) \hat{u} & \text{in } (0,1), \\ d_1 \hat{u}' + \sigma c^0 \hat{u} = 0 & \text{for } x = 0, 1. \end{cases}$$

Since $\hat{u} \geq 0$ and $\|\hat{u}\|_\infty = 1$, we necessarily have, by applying the strong maximum principle to (2.2), that $\hat{u} > 0$. This implies that $-m_*$ is the principal eigenvalue of the problem

$$\text{(2.3)} \qquad \begin{cases} -(d_1 u' + \sigma c^0 u)' - f(\min\{\alpha v_0, w_*\}) u = \lambda u & \text{in } (0,1), \\ d_1 u' + \sigma c^0 u = 0 & \text{for } x = 0, 1. \end{cases}$$

One easily checks that 0 is the first eigenvalue of the problem

$$\begin{cases} -(d_1 u' + \sigma c^0 u)' = \lambda u & \text{in } (0,1), \\ d_1 u' + \sigma c^0 u = 0 & \text{for } x = 0, 1. \end{cases}$$

Since $-f(\min\{\alpha v_0, w_*\}) < 0$, by the characterization of the first eigenvalues (see, for example, Theorems 2.4 and 2.8 of [D]), the first eigenvalue of (2.3) is less than 0, and hence $m_* > 0$. On the other hand, if (u, v, w) is a positive solution to (1.3), then rewriting the equation for u in the form

$$\begin{cases} -(d_1 u' + \sigma c u)' - f(\min\{\alpha v, w\})u = -mu & \text{in } (0,1), \\ d_1 u' + \sigma c u = 0 & \text{for } x = 0, 1, \end{cases}$$

we find that $-m$ is the first eigenvalue of the problem

$$\begin{cases} -(d_1 u' + \sigma c u)' - f(\min\{\alpha v, w\})u = \lambda u & \text{in } (0,1), \\ d_1 u' + \sigma c u = 0 & \text{for } x = 0, 1, \end{cases}$$

which we denote by $\lambda_1(c, v, w)$. Clearly $v < v_0$ and $w < w_*$ in $(0,1)$. It follows that $-f(\min\{\alpha v, w\}) > -f(\min\{\alpha v_0, w_*\})$ in $(0,1)$ and hence

$$\lambda_1(c, v, w) > \lambda_1(c, v_0, w_*). \tag{2.4}$$

In this notation, clearly $-m_* = \lambda_1(c^0, v_0, w_*)$. Since $c(x) = (x - x_0)/(\delta + |x - x_0|)$ is determined completely by x_0, it is convenient to introduce the notation

$$C_{x_0} = \frac{x - x_0}{\delta + |x - x_0|}$$

and

$$m^* = -\inf_{x_0 \in [0,1]} \lambda_1(C_{x_0}, v_0, w_*).$$

It is easy to show that m^* is achieved by some $x_0 \in [0,1]$, and $m^* \geq m_*$. Moreover, from the above discussion, we have the following result.

PROPOSITION 2.1. *If* (1.3) *has a positive solution, then necessarily* $m < m^*$.

We will show that (1.3) has a positive solution for every $0 < m < m_*$. Before that we briefly discuss some further simple estimates for the values of m so that (1.3) has a positive solution. So suppose that (1.3) has a positive solution (u, v, w). Integrating the equation for u over $[0,1]$, we obtain

$$\int_0^1 \Big[f(\min\{\alpha v, w\}) - m \Big] u dx = 0.$$

Since $u > 0$, $[f(\min\{\alpha v, w\}) - m]$ must change sign over $(0,1)$, and therefore

$$\min_{[0,1]} f(\min\{\alpha v(x), w(x)\}) < m < \max_{[0,1]} f(\min\{\alpha v(x), w(x)\}).$$

It follows that

$$f(\min\{\alpha v(0), w(1)\}) < m < f(\min\{\alpha v_0, w_0\}). \tag{2.5}$$

From a similar consideration, we have

$$m^*, m_* \in \big(f(\min\{\alpha v_0, w_*(1)\}), f(\min\{\alpha v_0, w_0\}) \big). \tag{2.6}$$

We now use a global bifurcation argument to show that (1.3) has a positive solution for every $m \in (0, m_*)$. First, we transform (1.3) into an abstract nonlinear

equation. Due to its unconventional nature, we cannot use a simple inverse operator trick to do this. In fact, to cope with the rather implicit dependence of $c_{v,w}$ on (v,w), in the following, we have to choose the function spaces for the abstract setting very carefully and then analyze the properties of the abstract operator mostly by definitions.

Fix $\gamma \in (0,1)$ and set

$$K := \{\phi \in C^{1,\gamma}([0,1]) : \phi \text{ is nondecreasing in } [0,1]\},$$
$$P := \{\phi \in C^{1,\gamma}([0,1]) : \phi \text{ is nonnegative in } [0,1]\}.$$

Clearly they are closed convex sets in $C^{1,\gamma}([0,1])$, and, moreover, P is a positive cone.

For given $(u,v) \in P \times K$ and $m \geq 0$, we define

$$w = w_0 \exp\Big[- A_0 x - A \int_0^x u(s)ds\Big],$$
$$c(x) = c_{v,w}(x), \quad v_+ = \max\{v, 0\}$$

and will use the solutions of the following problems to define an abstract operator $T(m,u,v)$ such that $T(m,u,v) = (u,v)$ for $(u,v) \in P \times K$ if and only if (u,v) is a nonnegative solution of (1.3). So we consider the problems

$$\text{(2.7)} \quad \begin{cases} -(d_1\phi' + \sigma c\phi)' + (m+1)\phi = f(\min\{\alpha v_+, w\})u + u, & 0 < x < 1, \\ d_1\phi' + \sigma c\phi = 0, & x = 0, 1, \end{cases}$$

and

$$\text{(2.8)} \quad \begin{cases} -d_2\psi'' = -f(\min\{\alpha v_+, w\})u, & 0 < x < 1, \\ \psi'(0) = 0, \quad \psi'(1) = \beta[v_0 - \psi(1)]. \end{cases}$$

Clearly (2.7) has a unique solution ϕ, and it is nonnegative. Let $\zeta(x) = v_0 - \psi(x)$. Then (2.8) becomes

$$\text{(2.9)} \quad \begin{cases} -d_2\zeta'' = f(\min\{\alpha v_+, w\})u, & 0 < x < 1, \\ \zeta'(0) = 0, \quad \zeta'(1) + \beta\zeta(1) = 0. \end{cases}$$

It is easily seen that (2.9) has a unique solution ζ, and it is nonnegative. Moreover, from $\zeta'' \leq 0$, and $\zeta'(0) = 0$ we deduce that ζ is nonincreasing. Hence ψ is nondecreasing and $\psi(x) \leq v_0$ in $[0,1]$. Thus the solution operator $(\phi,\psi) = T(m,u,v)$ is well defined for $(u,v) \in P \times K$ and $m \geq 0$, and $T(m,\cdot,\cdot)$ maps $P \times K$ into itself.

We show next that T is continuous. Suppose that $(m_n, u_n, v_n) \in [0,\infty) \times P \times K$, $m_n \to m$, and $(u_n, v_n) \to (u,v)$ in $C^{1,\gamma}([0,1]) \times C^{1,\gamma}([0,1])$. Then it is easily checked that $c^n := c_{v_n,w_n} \to c := c_{v,w}$ in $C^1([0,1])$. Denote $(\phi_n, \psi_n) = T(m_n, u_n, v_n)$ and $\zeta_n = v_0 - \psi_n$. We have

(2.10)

$$\begin{cases} -(d_1\phi_n' + \sigma c^n\phi_n)' + (m_n+1)\phi_n = f(\min\{\alpha(v_n)_+, w_n\})u_n + u_n, & 0 < x < 1, \\ d_1\phi_n' + \sigma c^n\phi_n = 0, & x = 0, 1, \end{cases}$$

and

$$\begin{cases} -d_2\zeta_n'' = f(\min\{\alpha(v_n)_+, w_n\})u_n, & 0 < x < 1, \\ \zeta_n'(0) = 0, \quad \zeta_n'(1) + \beta\zeta_n(1) = 0. \end{cases} \tag{2.11}$$

Applying standard L^p theory to both (2.10) and (2.11), we find that $\{\phi_n\}$ and $\{\zeta_n\}$ are bounded in $W^{2,p}([0,1])$ for all $p > 1$. Hence they are precompact in $C^{1,\gamma}([0,1])$. This implies that by passing to a subsequence, $\phi_n \to \phi$ and $\psi_n \to \psi$ in $C^{1,\gamma}([0,1])$. Moreover, letting $n \to \infty$ in (2.10) and (2.11) we find that necessarily $(\phi, \psi) = T(m, u, v)$. Therefore the entire original sequence converges with limit (ϕ, ψ). This proves the continuity of T.

We further show that T is compact. Suppose that $\{(m_n, u_n, v_n)\} \subset [0,\infty) \times P \times K$ is bounded. Then along some subsequence n_k, (u_n, v_n) converges in the $C^1([0,1]) \times C^1([0,1])$ norm to some (u, v), and, by passing to a further subsequence, we may assume that $m_{n_k} \to m$. We may now repeat the arguments in the above continuity proof to conclude that $T(m_{n_k}, u_{n_k}, v_{n_k}) \to T(m, u, v)$ in $C^{1,\gamma}([0,1])$. Therefore T is a compact operator on $[0,\infty) \times P \times K$.

Suppose that $m \geq 0$ and $(u, v) \in P \times K$ satisfies $(u, v) = T(m, u, v)$. We claim that $v \geq 0$ in $[0,1]$. Otherwise, due to the monotonicity of v there exists $x_0 \in (0,1]$ such that $v < 0$ in $[0, x_0)$ and $v \geq 0$ in $(x_0, 1]$. Therefore, by (2.8),

$$-v'' = 0 \text{ in } (0, x_0).$$

Since $v'(0) = 0$, we deduce that $v'(x) = 0$ in $(0, x_0)$, and hence v is a negative constant in $(0, x_0)$, say $v = -c$. This is possible only if $x_0 = 1$ (otherwise, v is discontinuous at $x = x_0$ since $v \geq 0$ in $(x_0, 1]$), but then from $v'(1) = \beta[v_0 - v(1)]$ we deduce that $-c = v_0 > 0$, a contradiction. Therefore we must have $v \geq 0$ in $[0,1]$, as claimed. Thus (u, v) is a nonnegative solution of (1.3) if and only if $(u, v) \in P \times K$ and $T(m, u, v) = (u, v)$.

In order to apply the global bifurcation theory to the operator equation

$$(u, v) - T(m, u, v) = 0,$$

we now calculate the Fréchet derivative of T with respect to (u, v) at $(m, 0, v_0)$, in the convex set $P \times K$ of $C^{1,\gamma}([0,1]) \times C^{1,\gamma}([0,1])$, where $m \geq 0$. From (2.7) and (2.8) we easily see that $T(m, 0, v_0) = (0, v_0)$. For $(u, v) \in C([0,1]) \times C([0,1])$, we define $(\xi, \eta) = L_m(u, v)$ to be the unique solution of the following linear problems:

$$\begin{cases} -(d_1\xi' + \sigma c^0\xi)' + (m+1)\xi = f(\min\{\alpha v_0, w_*\})u + u, & 0 < x < 1, \\ d_1\xi' + \sigma c^0\xi = 0, & x = 0, 1, \end{cases} \tag{2.12}$$

$$\begin{cases} -d_2\eta'' = -f(\min\{\alpha v_0, w_*\})u, & 0 < x < 1, \\ \eta'(0) = 0, \quad \eta'(1) + \beta\eta(1) = 0, \end{cases} \tag{2.13}$$

where c^0 and w_* are defined as at the beginning of this section.

Suppose $(u_n, v_n) \to (0, v_0)$ in $P \times K$. Denote

$$(\phi_n, \psi_n) = T(m, u_n, v_n) \text{ and } (\tau_n, \theta_n) = L_m(u_n, v_n - v_0).$$

We want to show that

$$\|(\phi_n, \psi_n) - (0, v_0) - (\tau_n, \theta_n)\| = o(\|(u_n, v_n - v_0)\|), \tag{2.14}$$

where $\|(u,v)\| = \max\{\|u\|, \|v\|\}$, and $\|u\| = \|u\|_{C^{1,\gamma}([0,1])}$. This would imply that the Fréchet derivative of T with respect to (u,v) at $(m,0,v_0)$, in the convex set $P \times K$, is the linear operator L_m.

Suppose $(u_n, v_n) \to (0, v_0)$ in $P \times K$. Without loss of generality we assume that $u_n \not\equiv 0$. We define w_n and $c^n = c_{v_n,w_n}$ as before, and let $\hat{\phi}_n = \phi_n/\|u_n\|$, $\hat{u}_n = u_n/\|u_n\|$. Then

(2.15)
$$\begin{cases} -(d_1\hat{\phi}_n' + \sigma c^n \hat{\phi}_n)' + (m+1)\hat{\phi}_n = f(\min\{\alpha(v_n)_+, w_n\})\hat{u}_n + \hat{u}_n, & 0 < x < 1, \\ d_1\hat{\phi}_n' + \sigma c^n \hat{\phi}_n = 0, & x = 0, 1. \end{cases}$$

Since the right-hand side of the first equation in (2.15) is bounded in $L^\infty([0,1])$, much as before we deduce from the L^p theory and the Sobolev imbedding theorem that there exists some positive constant C independent of n such that

$$\|\hat{\phi}_n\| \le C \ \forall n \ge 1.$$

We now define $\Phi_n = (\phi_n - \tau_n)/\|u_n\|$, and from the equations for ϕ_n and τ_n we obtain

$$\begin{cases} -(d_1\Phi_n' + \sigma c^0 \Phi_n)' + (m+1)\Phi_n = f_n, & 0 < x < 1, \\ d_1\Phi_n' + \sigma c^0 \hat{\Phi}_n = g_n, & x = 0, 1, \end{cases} \tag{2.16}$$

where

$$f_n := [f(\min\{\alpha(v_n)_+, w_n\}) - f(\min\{\alpha v_0, w_*\})]\hat{u}_n + [\sigma(c^n - c^0)\hat{\phi}_n]'$$

and

$$g_n = \sigma(c^0 - c^n)\hat{\phi}_n.$$

It is easy to check that f_n converges to 0 in $L^\infty([0,1])$, and g_n converges to 0 in $C^1([0,1])$. Since $(m+1) \ge 1$, we may apply the L^p estimate to (2.16) to conclude that $\|\Phi_n\| \to 0$ as $n \to \infty$. Hence

$$\|\phi_n - \tau_n\| = o(\|u_n\|).$$

Define $\Psi_n = [(v_0 - \psi_n) + \theta_n]/\|u_n\|$. Then from the equations for ψ_n and θ_n we deduce that

$$\begin{cases} -d_2\Psi_n'' = [f(\min\{\alpha(v_n)_+, w_n\}) - f(\min\{\alpha v_0, w_*\})]\hat{u}_n, & 0 < x < 1, \\ \Psi_n'(0) = 0, \quad \Psi_n'(1) + \beta\Psi_n(1) = 0. \end{cases} \tag{2.17}$$

Since the right-hand side of the first equation in (2.17) converges to 0 in $L^\infty([0,1])$, we can apply the L^p theory to (2.17) to conclude that $\|\Psi_n\| \to 0$ as $n \to \infty$. Therefore

$$\|\psi_n - v_0 - \theta_n\| = o(\|u_n\|).$$

Thus we have

$$\|(\phi_n, \psi_n) - (0, v_0) - (\tau_n, \theta_n)\| = o(\|(u_n, v_n - v_0)\|).$$

Summarizing the above discussions, we have the following result.

PROPOSITION 2.2. *The operator $T : [0,\infty) \times P \times K \to P \times K$ is completely continuous, and it is Fréchet differentiable at $(m, 0, v_0)$ with respect to (u, v) in the convex set $P \times K$, with derivative operator L_m. Moreover, $(u, v) = T(m, u, v)$ implies that $v \in P$; (u, v) is a nonnegative solution of (1.3) if and only if $(u, v) = T(m, u, v)$.*

We are now ready to prove the main result of this section.

THEOREM 2.3. *For every $m \in (0, m_*)$, problem (1.3) has at least one positive solution. Moreover, if m_n decreases to 0 and (u_n, v_n) is a positive solution of (1.3) with $m = m_n$, then $u_n \to \infty$ uniformly in $[0, 1]$ and there exists a unique $\tau \in \big(0, \frac{v_0}{1+\beta^{-1}}\big)$ (determined by (2.27) below) such that*

$$\frac{u_n(x)}{\|u_n\|_\infty} \to \Big(1 + \frac{x}{\delta}\Big)^{\frac{\sigma}{d_1}\delta} e^{-\frac{\sigma}{d_1}x}, \quad v_n(x) \to \tau x + v_0 - \tau(1 + \beta^{-1})$$

uniformly in $[0, 1]$. Furthermore, for each $m \in (0, m_*)$, there is a positive solution (m, u, v) lying on the global bifurcation branch, $\Gamma = \{(m, u, v)\} \subset (0,\infty) \times C^{1,\gamma}([0, 1]) \times C^{1,\gamma}([0, 1])$, bifurcating from the trivial solution branch $\Gamma_0 := \{(m, 0, v_0) : m \in (-\infty,\infty)\}$ at $m = m_*$.

Proof. Since the proof is rather long, we divide it into several steps.

Step 1: *Existence of an unbounded global solution branch.*

We observe that 1 is an eigenvalue of L_{m_*} with eigenvector (ϕ_1, ψ_1) satisfying $\phi_1 > 0$ and $\psi_1 < 0$. Indeed, $\phi_1 > 0$ is a principal eigenfunction of (2.3) with $\lambda = -m_*$, and ψ_1 is the unique solution of (2.13) with $u = \phi_1$, and hence $\psi_1 < 0$. In order to apply the abstract global bifurcation theory in positive cones, we define $S : [0,\infty) \times P \times (-K) \to P \times (-K)$ by

$$S(m, u, \xi) = (\phi, \zeta) \text{ if and only if } T(m, u, v_0 - \xi) = (\phi, v_0 - \zeta).$$

Then from the properties of T we find that S is completely continuous. Moreover, if we denote by $DS(m, 0, 0)$ the Fréchet derivative of S with respect to (u, ξ) in $P \times (-K)$ at $(u, \xi) = (0, 0)$, then 1 is an eigenvalue of $DS(m_*, 0, 0)$ with eigenvector $(\phi_1, -\psi_1)$, where ϕ_1 and ψ_1 are as given above. Let us denote by P_0 the nonnegative functions in $(-K)$. Clearly P_0 is a cone in $C^{1,\gamma}([0, 1])$, and hence $P \times P_0$ is a cone in $C^{1,\gamma}([0, 1]) \times C^{1,\gamma}([0, 1])$. Moreover, it is easy to check through the definition of T that $S(m, \cdot, \cdot)$ maps $P \times P_0$ into itself, and 1 is the only eigenvalue of $DS(m_*, 0, 0)$ with an eigenvector in $P \times P_0$, and for any $m \geq 0$, $m \neq m_*$, 1 is not an eigenvalue of $DS(m, 0, 0)$ corresponding to an eigenvector in $P \times P_0$. These properties allow us to apply Corollary 18.4 of [A] to conclude that there exists a global unbounded branch of solutions of $(u, \zeta) = S(m, u, \zeta)$ in $R^1 \times \big(P \times P_0 \setminus \{(0, 0)\}\big)$. We denote this global branch by $\tilde{\Gamma} = \{(m, u, \zeta)\}$ and define

$$\Gamma := \{(m, u, v_0 - \zeta) : (m, u, \zeta) \in \tilde{\Gamma}\}.$$

Clearly Γ is a global branch of solutions to $(u, v) = T(m, u, v)$ with $u \geq 0$ and $v = v_0 - \zeta \leq v_0$. We claim that $u \neq 0$. Otherwise, $u = 0$ and we deduce that $v = v_0 - \zeta = v_0$. Hence $\zeta = 0$, contradicting the fact that $(m, u, \zeta) \in R^1 \times \big(P \times P_0 \setminus \{(0, 0)\}\big)$. Hence $u \geq 0$ and $u \neq 0$. It then follows from Proposition 2.2 that $v_0 - \zeta \geq 0$. But $u \neq 0$ implies that $v_0 - \zeta \neq 0$, and hence $(u, v_0 - \zeta)$ is a positive solution of (1.3). Thus we have proved that Γ is an unbounded branch of positive solutions of (1.3).

Step 2: *We show that the m-range of Γ covers $(0, m_*)$.*

If $(m,u,v)\in\Gamma$, then from Proposition 2.1 and (2.5) we deduce that $0<m<m^*$. Therefore we can find a sequence $(m_n,u_n,v_n)\in\Gamma$ such that $m_n\to m_0\in[0,m^*]$ and $\|(u_n,v_n)\|\to\infty$. Note that $c_{v_n,w_n}=C_{x_n}$ for some $x_n\in[0,1]$ uniquely determined by v_n and w_n. By passing to a subsequence we may assume that $x_n\to x_0\in[0,1]$. Then it is easily seen that $C_{x_n}\to C_{x_0}$ in $C^1([0,1])$. We necessarily have, by passing to a subsequence, that $\|u_n\|_\infty\to\infty$, for otherwise from the equation for u_n we can deduce that $\|u_n\|$ is bounded, which in turn implies that $\|v_n\|$ is bounded, contradicting our assumption that $\|(u_n,v_n)\|\to\infty$. Therefore we may assume that $\|u_n\|_\infty\to\infty$. Denote $\hat{u}_n=u_n/\|u_n\|_\infty$. Then we can use the L^p estimate to the equation for $\hat{u}_n$ to deduce that $\{\hat{u}_n\}$ is precompact in $C^{1,\gamma}([0,1])$. Hence we may assume that $\hat{u}_n\to\hat{u}$ in $C^{1,\gamma}([0,1])$. Since $0\le f(\min\{\alpha v_n,w_n\})\le f(w_*)$, we may assume that $f(\min\{\alpha v_n,w_n\})$ converges to f_0 weakly in $L^2([0,1])$. Clearly we also have $0\le f_0\le f(w_*)$. Passing to the weak limit in the equation for $\hat{u}_n$ we deduce that $\hat{u}$ is a weak solution of

$$\begin{cases} -(d_1\hat{u}'+\sigma C_{x_0}\hat{u})'=(f_0-m_0)\hat{u}, & 0<x<1,\\ d_1\hat{u}'+\sigma C_{x_0}\hat{u}=0, & x=0,1. \end{cases} \tag{2.18}$$

Since $\hat{u}\ge 0$ and $\|\hat{u}\|_\infty=1$, we can apply the Harnack inequality and the strong maximum principle to (2.18) to conclude that $\hat{u}>0$ in $[0,1]$. This implies that $u_n=\|u_n\|_\infty\hat{u}_n\to\infty$ uniformly in $[0,1]$. Therefore $w_n\to 0$ uniformly on any compact subset of $(0,1]$. This implies that $f_0=0$, and hence we deduce from (2.18) that $-m_0$ is the first eigenvalue of

$$\begin{cases} -(d_1u'+\sigma C_{x_0}u)'=\lambda u, & 0<x<1,\\ d_1u'+\sigma C_{x_0}u=0, & x=0,1. \end{cases} \tag{2.19}$$

Hence $m_0=0$ and

$$\hat{u}=\exp\Big[-\frac{\sigma}{d_1}\int_0^x C_{x_0}(s)ds\Big].$$

This implies that the entire original sequence $\{m_n\}$ converges to 0. By the connectedness of Γ, we can conclude that for every $m\in(0,m_*)$, (1.3) has at least one positive solution lying on Γ. Moreover, when $x_n\to x_0$, we have

$$\frac{u_n}{\|u_n\|_\infty}\to\exp\Big[-\frac{\sigma}{d_1}\int_0^x C_{x_0}(s)ds\Big].$$

Step 3: *The limiting profile of* u_n.

We will show in a moment that $x_0=0$ and hence the entire original sequence $u_n/\|u_n\|_\infty$ converges in $C^{1,\gamma}([0,1])$ to $\exp\big[-\frac{\sigma}{d_1}\int_0^x C_0(s)ds\big]$.

Let $\zeta_n=v_0-v_n$. Then

$$\begin{cases} -d_2\zeta_n''=f(\min\{\alpha v_n,w_n\})u_n, & 0<x<1,\\ \zeta_n'(0)=0, \quad \zeta_n'(1)+\beta\zeta_n(1)=0. \end{cases} \tag{2.20}$$

We have

$$0\le f(\min\{\alpha v_n,w_n\})u_n\le f(w_n(x))u_n(x).$$

Moreover,

$$\begin{aligned} f(w_n(x))u_n(x) &\leq (r/K_I)w_n(x)u_n(x) \\ &= (r/K_I)e^{-A_0x}e^{-A\int_0^x u_n(s)ds}u_n(x) \\ &\leq Ce^{-A\|u_n\|_\infty \int_0^x \hat{u}_n(s)ds}\|u_n\|_\infty \hat{u}_n(x). \end{aligned}$$

Since $\hat{u}_n \to \hat{u} > 0$ uniformly in $[0,1]$, there exist $c_1, c_2 > 0$ such that $c_1 \leq \hat{u}_n \leq c_2$ and hence

$$\begin{aligned} g_n &:= Ce^{-A\|u_n\|_\infty \int_0^x \hat{u}_n(s)ds}\|u_n\|_\infty \hat{u}_n(x) \\ &\leq Cc_2e^{-Ac_1\|u_n\|_\infty x}\|u_n\|_\infty. \end{aligned}$$

One easily sees from the above inequality that $g_n \to 0$ uniformly in compact subsets of $(0,1]$. It follows that

$$f_n := f(\min\{\alpha v_n, w_n\})u_n \to 0$$

uniformly in compact subsets of $(0,1]$.

We can now prove that $x_0 = 0$. Otherwise, $x_0 \in (0,1]$ and thus $x_n > x_0/2$ for all large n. Since $v_n(x)$ is increasing in x, and $\alpha v_n(x) < w_n(x)$ in $[0, x_n)$, it follows that

$$f(\min\{\alpha v_n, w_n\})u_n \leq f(\alpha v_n(x_0/2))u_n \leq f(w_n(x_0/2))u_n \text{ in } [0, x_0/2].$$

By our earlier estimates for g_n, we find that

$$f(w_n(x_0/2))u_n \leq Cc_2e^{-Ac_1\|u_n\|_\infty (x_0/2)}\|u_n\|_\infty \to 0.$$

Hence $f_n \to 0$ uniformly in $[0,1]$, which implies, by (2.20), that $\zeta_n \to 0$ in $C^{1,\gamma}([0,1])$. In particular, $\zeta_n \to 0$ uniformly in $[0,1]$; but this leads to a contradiction:

$$v_0 - \zeta_n = v_n \leq w_n/\alpha \leq w_n(x_0/2)/\alpha \to 0 \text{ uniformly in } [x_0/2, x_n].$$

Hence $x_0 = 0$. Therefore we have

$$x_n \to 0 \text{ and } u_n/\|u_n\|_\infty \to \phi_0, \tag{2.21}$$

where

$$\phi_0(x) = \exp\left[-\frac{\sigma}{d_1}\int_0^x C_0(s)ds\right] = e^{-\frac{\sigma}{d_1}x}\Big(1+\frac{x}{\delta}\Big)^{\frac{\sigma}{d_1}\delta}.$$

Step 4: *The limiting profile of* v_n.

Since $v_n \geq 0$ we have $0 \leq \zeta_n \leq v_0$. Moreover, due to (2.20) and the fact that $f_n \to 0$ uniformly in compact subsets of $(0,1]$, we can use standard elliptic regularity theory and a diagonal process to find a subsequence of $\{\zeta_n\}$, still denoted by ζ_n, such that $\zeta_n \to \zeta$ in $C^1([\epsilon,1])$ for every $\epsilon \in (0,1)$, and ζ satisfies

$$-d_2\zeta'' = 0 \text{ in } (0,1], \quad \zeta'(1) + \beta\zeta(1) = 0, \quad 0 \leq \zeta \leq v_0.$$

It follows that

$\zeta(x) = \tau(1 + \beta^{-1} - x)$ for some $\tau \geq 0$ to be determined below.

On the other hand, ζ_n can be explicitly expressed as

$$\zeta_n(x) = d_2^{-1}(1 + \beta^{-1} - x) \int_0^1 f_n(s)ds + d_2^{-1} \int_x^1 (x - s) f_n(s)ds.$$

Since

$$\int_0^1 f_n(s)ds = d_2\beta\zeta_n(1) \in (0, d_2\beta v_0(1)],$$

and $f_n \to 0$ uniformly on any compact subset of $(0, 1]$, one easily sees that

$$\int_x^1 (x - s) f_n(s)ds \to 0$$

uniformly for $x \in [0, 1]$. Thus

$$\tau = \beta\zeta(1) = \lim_{n\to\infty} \beta\zeta_n(1) = d_2^{-1} \lim_{n\to\infty} \int_0^1 f_n(s)ds = d_2^{-1} \lim_{n\to\infty} \int_0^\epsilon f_n(s)ds \tag{2.22}$$

for any $\epsilon \in (0, 1)$.

Since $v_n(x)$ is monotone increasing in x and $v_n''(x) \geq 0$, by an elementary argument we see that the fact that $v_n \to v_0 - \zeta$ in $C^1([\epsilon, 1])$ for every $\epsilon \in (0, 1)$ implies that $v_n \to v_0 - \zeta$ uniformly in $[0, 1]$.

We now show that $\tau > 0$. Suppose $\tau = 0$. Then $\zeta_n \to 0$ and hence $v_n \to v_0$ in $C^1([0, 1])$. Therefore, due to

$$w_n(0) = w_0 \text{ and } w_n \to 0 \text{ uniformly in } [\epsilon, 1] \ \forall \text{ small } \epsilon > 0,$$

when $w_0 > \alpha v_0$, we have $0 < x_n < 1$ for all large n, and

$$w_n(x_n) = \alpha v_n(x_n) \to \alpha v_0 \text{ as } n \to \infty;$$

if $w_0 < \alpha v_0$, then $x_n = 0$ and $w_n(x_n) = w_0$ for all large n; if $w_0 = \alpha v_0$, then either $0 < x_n < 1$ and $w_n(x_n) = \alpha v_n(x_n)$, or $x_n = 0$ and $w_n(x_n) = w_0$; in either case we can conclude that $w_n(x_n) \to w_0$ as $n \to \infty$.

Summarizing, we find that we always have

$$w_0 e^{-A_0 x_n} e^{-A\int_0^{x_n} u_n(s)ds} = w_n(x_n) \to \sigma_0 := \min\{w_0, \alpha v_0\}.$$

Since $x_n \to 0$ and $\hat{u}_n \to \phi_0$, we have $e^{-A_0 x_n} = 1 + o(1)$, and

$$\int_0^{x_n} u_n(s)ds = \|u_n\|_\infty \int_0^{x_n} \hat{u}_n(s)ds = \|u_n\|_\infty x_n \phi_0(0)[1 + o(1)] = \|u_n\|_\infty x_n[1 + o(1)].$$

Here $o(1)$ denotes a generic sequence converging to 0 as $n \to \infty$. This implies that

$$w_0 e^{-A\|u_n\|_\infty x_n} \to \sigma_0.$$

Hence

$$\|u_n\|_\infty x_n \to \tau_0 := A^{-1} \ln\Big(\frac{w_0}{\sigma_0}\Big). \tag{2.23}$$

Since we now assume that $\tau = 0$, by (2.22) we must have $\lim_{n\to\infty} \int_0^\epsilon f_n(x)dx = 0$. On the other hand, making use of $x_n \to 0$, $\|u_n\|_\infty x_n \to \tau_0$, $\hat{u}_n \to \phi_0$, and $\phi_0(0) = 1$, we have, for all large n and small ϵ,

$$\begin{aligned}
\int_0^\epsilon f_n(x)dx &\geq \int_{x_n}^\epsilon f(w_n(x))u_n(x)dx \\
&\geq \int_{x_n}^\epsilon f\Big(w_0 e^{-A\epsilon} \exp[-A\|u_n\|_\infty 2\phi_0(0)x]\Big)\frac{\phi_0(0)}{2}\|u_n\|_\infty dx \\
&= \int_{\|u_n\|_\infty x_n}^{\|u_n\|_\infty \epsilon} f\Big(w_0 e^{-A\epsilon} e^{-2Ay}\Big)\Big(\frac{1}{2}\Big)dy \\
&\to \Big(\frac{1}{2}\Big)\int_{\tau_0}^\infty f\Big(w_0 e^{-A\epsilon} e^{-2Ay}\Big)dy > 0.
\end{aligned}$$

This contradiction shows that we must have $\tau > 0$.

In order to find the asymptotic limit of the entire sequence $\{v_n\}$, we need to determine the value of τ. Recall that by passing to a subsequence, $v_n \to v_0 - \zeta$ uniformly in $[0, 1]$. Since $x_n \to 0$, we have either

$$w_n(x_n) = \alpha v_n(x_n) \to \alpha[v_0 - \zeta(0)] = \alpha[v_0 - \tau(1 + \beta^{-1})] := \xi_\tau,$$

which is the case when $\xi_\tau < w_0$, or $w_n(x_n) \to w_0$ when $\xi_\tau \geq w_0$. By the expression of $w_n(x_n)$, as before, we deduce that

$$w_0 e^{-A\|u_n\|_\infty x_n} \to \tilde{\xi}_\tau := \min\{w_0, \xi_\tau\}, \quad \|u_n\|_\infty x_n \to \sigma_\tau := A^{-1} \ln\big(w_0/\tilde{\xi}_\tau\big). \tag{2.24}$$

For fixed $\epsilon \in (0, 1)$, we have

$$\int_0^\epsilon f_n(x)dx = \int_0^{x_n} f_n(x)dx + \int_{x_n}^\epsilon f(w_n(x))u_n(x)dx.$$

It is easy to check that, in every possible case, we have

$$\int_0^{x_n} f_n(x)dx = f(\tilde{\xi}_\tau)\phi_0(0)[1 + o(1)]\|u_n\|_\infty x_n = f(\tilde{\xi}_\tau)\sigma_\tau[1 + o(1)].$$

Since we already know that $f(w_n(x))u_n(x) \to 0$ uniformly on any compact subset of $(0, 1]$, we find that, for any fixed $\epsilon_1 \in (0, \epsilon)$ and all large n,

(2.25)

$$
\begin{aligned}
&\int_{x_n}^{\epsilon} f\big(w_n(x)\big)u_n(x)dx\\
&= \int_{x_n}^{\epsilon_1} f\big(w_n(x)\big)u_n(x)dx + o_n(1)\\
&= \int_{x_n}^{\epsilon_1} f\big(w_0 e^{-A_0 x} e^{-A\int_0^x u_n(s)ds}\big)u_n(x)dx + o_n(1)\\
&= \int_{x_n}^{\epsilon_1} f\big(w_0[1+o_{\epsilon_1}(1)]e^{-A\|u_n\|_\infty \hat{u}_n(0)x[1+o_{\epsilon_1}(1)]}\big)\|u_n\|_\infty \hat{u}_n(0)[1+o_{\epsilon_1}(1)]dx + o_n(1)\\
&= [1+o_{\epsilon_1}(1)]\int_{x_n}^{\epsilon_1} f\big(w_0 e^{-A\|u_n\|_\infty \hat{u}_n(0)x}\big)\|u_n\|_\infty \hat{u}_n(0)dx + o_n(1)\\
&= [1+o_{\epsilon_1}(1)]\int_{\|u_n\|_\infty \hat{u}_n(0)x_n}^{\epsilon_1\|u_n\|_\infty \hat{u}_n(0)} f\big(w_0 e^{-Ay}\big)dy + o_n(1)\\
&= [1+o_{\epsilon_1}(1)]\bigg[\int_{\sigma_\tau}^{\infty} f\big(w_0 e^{-Ay}\big)dy + o_n(1)\bigg] + o_n(1),
\end{aligned}
$$

where $o_n(1) \to 0$ as $n \to \infty$ for fixed ϵ_1, and $o_{\epsilon_1}(1) \to 0$ as $\epsilon_1 \to 0$ uniformly in n. For arbitrary $\epsilon_1 \in (0, \epsilon)$, we first let $n \to \infty$ and then let $\epsilon_1 \to 0$, and we obtain from (2.25) that

$$
\lim_{n\to\infty}\int_{x_n}^{\epsilon} f\big(w_n(x)\big)u_n(x)dx = \int_{\sigma_\tau}^{\infty} f\big(w_0 e^{-Ay}\big)dy.
$$

Therefore

$$
\lim_{n\to\infty}\int_{0}^{\epsilon} f_n(x)dx = f(\tilde{\xi}_\tau)\sigma_\tau + \int_{\sigma_\tau}^{\infty} f\big(w_0 e^{-Ay}\big)dy. \tag{2.26}
$$

Making use of (2.24) and (2.26), we can rewrite (2.22) as

$$
d_2\tau = f(w_0 e^{-A\sigma_\tau})\sigma_\tau + \int_{\sigma_\tau}^{\infty} f\big(w_0 e^{-Ay}\big)dy. \tag{2.27}
$$

It can be easily checked that the function

$$
F(\theta) := f(w_0 e^{-A\theta})\theta + \int_{\theta}^{\infty} f\big(w_0 e^{-Ay}\big)dy
$$

satisfies $F'(\theta) < 0$ and hence it is decreasing in $[0, \infty)$, with

$$
F(0) = \int_0^{\infty} f\big(w_0 e^{-Ay}\big)dy = \frac{r}{A}\ln\bigg(\frac{w_0 + K_I}{K_I}\bigg), \quad F(\infty) = 0.
$$

From the definition of σ_τ, we find that $\tau \to \sigma_\tau$ is nondecreasing, with

$$
\sigma_0 = A^{-1}\ln\bigg(\frac{w_0}{\min\{w_0, \alpha v_0\}}\bigg), \quad \sigma_{v_0/(1+\beta^{-1})} = \infty.
$$

Therefore

$\tau \to F(\sigma_\tau)$ is nonincreasing in $[0, v_0/(1+\beta^{-1}))$ with $F(\sigma_0) > 0, F(\sigma_{v_0/(1+\beta^{-1})}) = 0$.

This implies that (2.27) has a unique solution $\tau \in \big(0, v_0/(1+\beta^{-1}\big)$. Thus,

$$v_n \to v_0 - \tau(1+\beta^{-1}-x) \text{ uniformly in } [0,1],$$

with τ uniquely determined by (2.27). Since $\tau > 0$ is uniquely determined, the above convergence is true for the entire original sequence $\{v_n\}$.

The proof is complete. □

3. The limit of m_* and m^* as $\sigma \to \infty$. In order to investigate the asymptotic behavior of the positive solutions of (1.3) as $\sigma \to \infty$, we need to first understand the limits of m_* and m^* as $\sigma \to \infty$. To stress their dependence on σ, we write $m_* = m_*(\sigma)$ and $m^* = m^*(\sigma)$. Let us recall that (2.6) holds; that is, $m_*(\sigma)$ and $m^*(\sigma)$ are always between the positive numbers $f(\min\{\alpha v_0, w_*(1)\})$ and $f(\min\{\alpha v_0, w_0\})$. We now prove the following result.

THEOREM 3.1.

$$\lim_{\sigma\to\infty} m_*(\sigma) = \lim_{\sigma\to\infty} m^*(\sigma) = f(\min\{\alpha v_0, w_0\}). \tag{3.1}$$

Proof. Since

$$m_*(\sigma) \le m^*(\sigma) \le f(\min\{\alpha v_0, w_0\}),$$

we need only show that

$$\lim_{\sigma\to\infty} m_*(\sigma) = f(\min\{\alpha v_0, w_0\}).$$

Moreover, it suffices to prove this along an arbitrary sequence of positive numbers increasing to ∞. Let $\{\sigma_n\}$ be such a sequence, and denote $m_n = m_*(\sigma_n)$. By definition, there exists $u_n > 0$ in $[0,1]$ such that

$$\begin{cases} -[d_1 u_n' + \sigma_n c^0(x) u_n]' = [f(\min\{\alpha v_0, w_*\}) - m_n] u_n & \text{in } (0,1), \\ d_1 u_n' + \sigma_n c^0(x) u_n = 0 & \text{for } x = 0, 1. \end{cases} \tag{3.2}$$

To simplify the notation, we will write

$$f_0(x) = f(\min\{\alpha v_0, w_*(x)\}).$$

Moreover, we define x_0^* by

$$c^0(x) = C_{x_0^*}(x) = \frac{x - x_0^*}{\delta + |x - x_0^*|}.$$

If $x_0^* = 1$, i.e., $\alpha v_0 \le w_*(x)$ in $[0,1]$ and hence $f_0(x) \equiv f(\alpha v_0)$, then clearly

$$m_n \equiv f(\alpha v_0) = f(\min\{\alpha v_0, w_0\}).$$

So (3.1) holds trivially in this case.

Suppose from now on that $x_0^* \in [0,1)$. Therefore $f_0(x)$ is a constant in $[0, x_0^*)$ and is decreasing in $[x_0^*, 1]$. As before, integrating the first equation of (3.2) we find that $(f_0(x) - m_n)$ must change sign in $(0,1)$, and therefore there exists a unique $x_n \in (0,1)$ such that $f_0(x) > m_n$ in $[0, x_n)$, and $f_0(x) < m_n$ in $(x_n, 1]$.

By (2.6), $\{m_n\}$ is a bounded sequence, and, by passing to a subsequence, we may assume that

$$m_n \to m_0 \le f(\min\{(\alpha v_0, w_0\}) = f_0(x_0^*).$$

To determine the value of m_0, we use several steps.

Step 1: *Change of variables.*

Let

$$\phi_n(x) := \exp\left[-\frac{\sigma_n}{d_1}\int_{x_0^*}^{x} c^0(x)dx\right].$$

Clearly

$$d_1\phi_n' + \sigma_n c^0\phi_n = 0, \quad \phi_n(x_0^*) = 1, \quad 0 < \phi_n(x) \le 1 \text{ in } [0,1],$$

and $\phi_n \to 0$ uniformly on compact subsets of $[0,1]\setminus\{x_0^*\}$.

Define

$$\psi_n(x) = u_n(x)/\phi_n(x).$$

Then (3.2) becomes

$$\begin{cases} -(d_1\phi_n\psi_n')' = [f_0(x) - m_n]\phi_n\psi_n & \text{in } (0,1), \\ \psi_n' = 0 & \text{for } x = 0, 1. \end{cases} \tag{3.3}$$

Define

$$\xi_n(x) = \sigma_n^{-1/2}\int_0^x \frac{1}{\phi_n(s)}ds.$$

Then ξ_n is an increasing function in $[0,1]$, with $\xi_n(0) = 0$ and

$$\xi_n(1) = y_n := \sigma_n^{-1/2}\int_0^1 \frac{1}{\phi_n(x)}dx \to \infty \text{ as } n \to \infty.$$

Let $\eta_n : [0, y_n] \to [0,1]$ be the inverse function of $\xi_n(x)$, and define

$$U_n(y) = \psi_n(\eta_n(y)) = \psi_n(x).$$

From (3.3) a simple calculation shows that

$$\begin{cases} -U_n'' = d_1^{-1}\sigma_n\phi_n^2(\eta_n(y))[f_0(\eta_n(y)) - m_n]U_n & \text{in } (0, y_n), \\ U_n'(0) = U_n'(y_n) = 0. \end{cases} \tag{3.4}$$

Step 2: *Estimates of* $\sigma_n\phi_n^2(\eta_n(y))$.

For our later estimates, we need to analyze the function $\tilde{\phi}_n(y) := \sigma_n\phi_n^2(\eta_n(y))$. To this end, for some $\tau > 0$ small to be determined later, we define $\hat{y}_n$ and Δ_n by

$$\hat{y}_n := \xi_n(x_0^*), \quad \sigma_n^{-\tau} := \sigma_n^{-1/2}\int_{x_0^*}^{x_0^*+\Delta_n} \frac{1}{\phi_n(x)}dx,$$

so that

$$\hat{y}_n \pm \sigma_n^{-\tau} = \xi_n(x_0^* \pm \Delta_n), \quad x_0^* \pm \Delta_n = \eta_n(\hat{y}_n \pm \sigma_n^{-\tau}).$$

We will show that $\tilde{\phi}_n(y)$ behaves like a δ-function concentrating at $y = \hat{y}_n$. For definiteness, we assume that $x_0^* > 0$; the case $x_0^* = 0$ can be treated by a simple modification of the arguments below. Then it is easily seen that as $n \to \infty$,

$$\hat{y}_n \to \infty, \quad y_n - \hat{y}_n \to \infty, \quad \Delta_n \to 0.$$

Using $c^0(x) = \frac{x-x_0^*}{\delta+|x-x_0^*|}$, we can easily check that, for any given small $\epsilon > 0$, there exists $\delta_0 = \delta_0(\epsilon) > 0$ small so that, when $|x - x_0^*| \le \delta_0$,

$$\exp\left[-\frac{\sigma_n}{2\delta d_1}(x-x_0^*)^2\right] \le \phi_n(x) \le \exp\left[-\frac{\sigma_n(1-\epsilon)}{2\delta d_1}(x-x_0^*)^2\right]. \tag{3.5}$$

Therefore, for all large n,

$$\begin{aligned}
\sigma_n^{1/2-\tau} &= \int_{x_0^*}^{x_0^*+\Delta_n} \frac{1}{\phi_n(x)}dx \\
&\le \int_{x_0^*}^{x_0^*+\Delta_n} \exp\left[\frac{\sigma_n}{2\delta d_1}(x-x_0^*)^2\right]dx \\
&= \int_0^{\Delta_n} \exp\left(\frac{\sigma_n}{2\delta d_1}x^2\right)dx \\
&\le \Delta_n \exp\left(\frac{\sigma_n\Delta_n^2}{2\delta d_1}\right),
\end{aligned}$$

$$\begin{aligned}
\sigma_n^{1/2-\tau} &= \int_{x_0^*}^{x_0^*+\Delta_n} \frac{1}{\phi_n(x)}dx \\
&\ge \int_{x_0^*}^{x_0^*+\Delta_n} \exp\left[\frac{\sigma_n(1-\epsilon)}{2\delta d_1}(x-x_0^*)^2\right]dx \\
&= \int_0^{\Delta_n} \exp\left[\frac{\sigma_n(1-\epsilon)}{2\delta d_1}x^2\right]dx \\
&\ge \int_{(1-\epsilon)\Delta_n}^{\Delta_n} \exp\left[\frac{\sigma_n(1-\epsilon)}{2\delta d_1}x^2\right]dx \\
&\ge \epsilon\Delta_n \exp\left[\frac{\sigma_n\Delta_n^2(1-\epsilon)^3}{2\delta d_1}\right].
\end{aligned}$$

It follows that $\sigma_n\Delta_n^2 \to \infty$, and

$$\exp\left(-\frac{\sigma_n\Delta_n^2}{2\delta d_1}\right) \le \Delta_n\sigma_n^{\tau-1/2}, \tag{3.6}$$

and

$$\sigma_n^{1/2}\Delta_n \exp\left[\frac{\sigma_n\Delta_n^2(1-\epsilon)^3}{2\delta d_1}\right] \le \epsilon^{-1}\sigma_n^{1-\tau},$$

which gives

$$\left(\frac{1}{2}\right)\ln(\sigma_n\Delta_n^2) + \frac{(1-\epsilon)^3}{2\delta d_1}\sigma_n\Delta_n^2 \le \ln\left(\epsilon^{-1}\sigma_n^{1-\tau}\right).$$

Since $\ln(\sigma_n\Delta_n^2) = o(\sigma_n\Delta_n^2)$, and

$$\ln(\epsilon^{-1}\sigma_n^{1-\tau}) = (1-\tau)\ln\sigma_n + o(\ln\sigma_n),$$

the last inequality above implies that

$$\sigma_n\Delta_n^2 \le C_\epsilon \ln\sigma_n \tag{3.7}$$

for some $C_\epsilon > 0$ and all large n.

As a consequence of (3.5), (3.6), and (3.7) we have

$$\begin{aligned}
\tilde{\phi}_n(\hat{y}_n \pm \sigma_n^{-\tau}) &= \sigma_n\phi_n^2(x_0^* \pm \Delta_n)\\
&\le \sigma_n \exp\Big[-\frac{\sigma_n(1-\epsilon)}{\delta d_1}\Delta_n^2\Big]\\
&\le \sigma_n\big[\Delta_n^2\sigma_n^{2\tau-1}\big]^{(1-\epsilon)}\\
&\le \big[C_\epsilon\ln\sigma_n\big]^{(1-\epsilon)}\sigma_n^{1+2(\tau-1)(1-\epsilon)}\\
&\to 0 \text{ as } n\to\infty,
\end{aligned}$$

provided that τ is chosen in the interval $(0,1/2)$ and $\epsilon > 0$ is small enough.

From the property of $\phi_n(x)$, we see that the above estimates imply that

$$\tilde{\phi}_n(y) \to 0 \text{ uniformly in } [0,y_n]\setminus[\hat{y}_n - \sigma_n^{-\tau}, \hat{y}_n + \sigma_n^{-\tau}], \tag{3.8}$$

and for any $M > 0$,

$$\begin{aligned}
\int_{\hat{y}_n - M}^{\hat{y}_n + M} \tilde{\phi}_n(y)dy &= \int_{\hat{y}_n-\sigma_n^{-\tau}}^{\hat{y}_n+\sigma_n^{-\tau}} \tilde{\phi}_n(y)dy + o(1)\\
&= \int_{\eta_n(\hat{y}_n-\sigma_n^{-\tau})}^{\eta_n(\hat{y}_n+\sigma_n^{-\tau})} \sigma_n\phi_n^2(x)\xi_n'(x)dx + o(1)\\
&= \int_{x_0^*-\Delta_n}^{x_0^*+\Delta_n} \sigma_n^{1/2}\phi_n(x)dx + o(1)\\
&= 2\sigma_n^{1/2}\int_0^{\Delta_n} \exp\Big(-\frac{\sigma_n x^2}{2\delta d_1}[1+o(1)]\Big)dx + o(1)\\
&= [2+o(1)]\int_0^{\sigma_n^{1/2}\Delta_n} \exp\Big(-\frac{x^2}{2\delta d_1}\Big)dx + o(1)\\
&= 2\int_0^{\infty} \exp\Big(-\frac{x^2}{2\delta d_1}\Big)dx + o(1).
\end{aligned}$$

In other words, for any $M > 0$,

$$\lim_{n\to\infty}\int_{\hat{y}_n-M}^{\hat{y}_n+M} \tilde{\phi}_n(y)dy = c_0 := 2\int_0^{\infty}\exp\Big(-\frac{x^2}{2\delta d_1}\Big)dx. \tag{3.9}$$

Step 3: *The limiting profile of* $U_n(y)$.

We now define

$$\begin{aligned}
\hat{U}_n(y) &= U_n(y+\hat{y}_n-1)/U_n(\hat{y}_n-1),\\
\hat{f}_n &= d_1^{-1}\tilde{\phi}_n(y+\hat{y}_n-1)\big[f_0\big(\eta_n(y+\hat{y}_n-1)\big) - m_n\big].
\end{aligned}$$

Clearly $\hat{U}_n(0) = 1$, and by (3.4) we have

$$\begin{cases} -\hat{U}_n'' = \hat{f}_n \hat{U}_n \text{ in } (1-\hat{y}_n, y_n + 1 - \hat{y}_n), \\ \hat{U}_n'(1-\hat{y}_n) = \hat{U}_n'(y_n + 1 - \hat{y}_n) = 0. \end{cases} \tag{3.10}$$

From (3.8) we see that

$$\hat{f}_n(y) \to 0 \text{ uniformly in } [1-\hat{y}_n, 1-\sigma_n^{-\tau}] \cup [1+\sigma_n^{-\tau}, y_n + 1 - \hat{y}_n].$$

Since $\hat{U}_n(0) = 1$, the boundedness of $\hat{f}_n$ over $[1-\hat{y}_n, 1-\sigma_n^{-\tau}]$ allows us to apply the Harnack inequality to conclude that $\hat{U}_n$ has a bound C_J independent of n over any bounded interval $J \subset [1-\hat{y}_n, 1-\sigma_n^{-\tau}]$ with $0 \in J$. We can now apply to (3.10) the L^p theory, the Sobolev imbedding theorem, and a standard diagonal argument, to obtain a subsequence of $\{\hat{U}_n\}$, still denoted by $\hat{U}_n$, such that $\hat{U}_n \to \hat{U}$ in $C^1(J)$ for any bounded interval $J \subset (-\infty, 1)$, and $\hat{U}$ satisfies

$$\hat{U}'' = 0 \text{ in } (-\infty, 1), \quad \hat{U}(0) = 1. \tag{3.11}$$

Since $\hat{U}$ is nonnegative in $(-\infty, 1)$, we deduce from (3.11) that

$$\hat{U}(y) = 1 + ay, \quad a \in [-1, 0].$$

Now consider the sequence $\{\hat{U}_n(2)\}$. We claim that this is a bounded sequence. Indeed, from our earlier observation for the sign of $[f_0(x) - m_n]$, we know that the right-hand side of the first equation in (3.4) changes sign from positive to negative when y increases across $\tilde{y}_n := \xi_n(x_n)$. It follows that $U_n''(y)$ changes sign from negative to positive as y increases across $\tilde{y}_n$. Since $U_n'(0) = U_n'(y_n) = 0$, we find that $U_n' \le 0$ in $[0, y_n]$ and hence $U_n(y)$ is nonincreasing in y, which implies that $\hat{U}_n(y)$ is nonincreasing in y and hence $0 \le \hat{U}_n(2) \le \hat{U}_n(0) = 1$. We can now use the fact that $\hat{f}_n \to 0$ uniformly in $[1+\sigma_n^{-\tau}, y_n+1-\hat{y}_n]$, as above, to conclude that, subject to passing to a further subsequence, $\hat{U}_n \to \hat{U}_*$ in $C^1(J)$ for any bounded interval $J \subset (1, \infty)$, and $\hat{U}_*$ satisfies

$$\hat{U}_*'' = 0, \quad 0 \le \hat{U}_* \le 1 \text{ in } (1, \infty).$$

Therefore $\hat{U}_*$ must be a constant, say $\hat{U}_* \equiv b$.

Using (3.9) we find that $\hat{f}_n$ is a bounded sequence in $L^1([0,2])$. By (3.10), we have

$$\hat{U}_n'(y) = \hat{U}_n'(0) - \int_0^y \hat{f}_n(y)\hat{U}_n(y)dy \quad \forall y \in [0,2].$$

Since $\hat{U}_n'(0) \to \hat{U}'(0) = a$ and $0 \le \hat{U}_n(y) \le \hat{U}_n(0) = 1$, the above identity implies that $|\hat{U}_n'(y)| \le C$ for some $C > 0$ and all $n \ge 1$ and $y \in [0,2]$. Therefore $\{\hat{U}_n(y)\}$ is equicontinuous in $[0,2]$. It follows that $\hat{U}_*$ must be a continuous extension of $\hat{U}$. Therefore $b = 1 + a$.

Step 4: *We show that* $m_0 = f(x_0^*)$.

We are now ready to determine the value of m_0 by using our estimates for $\tilde{\phi}_n$ and $\hat{U}_n$. We note that (3.8) and (3.9) imply that, for large n, $\tilde{\phi}_n(y + \hat{y}_n - 1)$ behaves like the δ-function concentrating at $y = 1$. We now use these properties of $\tilde{\phi}_n$ and (3.10) to obtain

$$\hat{U}_n'(2) = \hat{U}_n'(0) - \int_0^2 \hat{f}_n(y)\hat{U}_n(y)dy \to a - [f_0(x_0^*) - m_0](1+a)c_0.$$

But since $\hat{U}_*$ is a constant function over $(1,\infty)$, we have $\hat{U}_n'(2) \to \hat{U}_*'(2) = 0$. Therefore

$$a = [f_0(x_0^*) - m_0](1+a)c_0.$$

The right-hand side of the above identity is nonnegative, but $a \le 0$. Therefore we must have $a = 0$ and $m_0 = f_0(x_0^*)$. This implies that the entire original sequence $\{m_n\}$ converges to $f_0(x_0^*)$. Hence (3.1) holds, and the proof is complete. □

If we fix m such that $0 < m < f(\min\{\alpha v_0, w_0\})$ and let σ_n be an increasing sequence of positive numbers converging to ∞, then by Theorems 2.3 and 3.1, for all large n, (1.3) with $\sigma = \sigma_n$ has at least one positive solution. Suppose that (u_n, v_n) is such a solution. We will analyze the behavior of (u_n, v_n) as $n \to \infty$. This will be done in Part II; see [DH].

Acknowledgments. We thank the referees for their useful suggestions and comments. One referee pointed us to extra references ([IT] and [HTKS]) and provided detailed suggestions on improving the exposition of the paper (for both Part I and Part II); these have been very helpful. Y. Du thanks NCTS for the hospitality during his visit, when a major part of this paper was being written.

REFERENCES

[A] H. Amann, *Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces*, SIAM Rev., 18 (1976), pp. 620–709.

[BFH] E. Beretta, A. Fasano, and Y. Hosono, *Equilibrium of a phytoplankton population in a laboratory controlled system*, Surveys Math. Indust., 1 (1992), pp. 283–336.

[BFHK] E. Beretta, A. Fasano, Y. Hosono, and V. B. Kolmanovskiĭ, *Stability analysis of the phytoplankton vertical steady states in a laboratory test tube*, Math. Methods Appl. Sci., 17 (1994), pp. 551–575.

[D] Y. Du, *Order Structure and Topological Methods in Nonlinear Partial Differential Equations*: *Maximal Principles and Applications*, Vol. 1, World Scientific, Hackensack, NJ, 2006.

[DH] Y. Du and S.-B. Hsu, *Concentration phenomena in a nonlocal quasi-linear problem modelling phytoplankton* II: *Limiting profile*, SIAM J. Math. Anal., 40 (2008), pp. 1441–1470.

[EATSH] U. Ebert, M. Arrayas, N. M. Temme, B. P. Sommeijer, and J. Huisman, *Critical conditions for phytoplankton blooms*, Bull. Math. Biol., 63 (2001), pp. 1095–1124.

[GT] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.

[HTKS] J. Huisman, N. N. P. Thi, D. M. Karl, and B. Sommeijer, *Reduced mixing generates oscillations and chaos in the oceanic deep chlorophyll maximum*, Nature, 439 (2006), pp. 322–325.

[IT] H. Ishii and I. Takagi, *Global stability of stationary solutions to a nonlinear diffusion equation in phytoplankton dynamics*, J. Math. Biol., 16 (1982), pp. 1–24.

[KL] C. A. Klausmeier and E. Litchman, *Algal games: The vertical distribution of phytoplankton in poorly mixed water columns*, Limnol. Oceanogr., 46 (2001), pp. 1998–2007.

[PT] N. N. Pham Thi, *On positive solutions in a phytoplankton-nutrient model*, J. Comput. Appl. Math., 177 (2005), pp. 467–473.

[PTHS] N. N. Pham Thi, J. Huisman, and B. P. Someijer, *Simulation of three-dimensional phytoplankton dynamics: Competition in light-limited environments*, J. Comput. Appl. Math., 174 (2005), pp. 57–77.

[YN] K. Yoshiyama and H. Nakajima, *Catastrophic shifts in vertical distributions of phytoplankton, the existence of a bifurcation set*, J. Math. Biol., 52 (2006), pp. 235–276.

SIAM J. MATH. ANAL.
Vol. 40, No. 4, pp. 1441–1470

CONCENTRATION PHENOMENA IN A NONLOCAL QUASI-LINEAR PROBLEM MODELLING PHYTOPLANKTON II: LIMITING PROFILE*

YIHONG DU† AND SZE-BI HSU‡

Abstract. This is Part II of our study on the positive steady state of a quasi-linear reaction-diffusion system in one space dimension introduced by Klausmeier and Litchman for the modelling of the distributions of phytoplankton biomass and its nutrient. In Part I, we proved nearly optimal existence and nonexistence results. In Part II, we obtain complete descriptions of the profile of the solutions when the coefficient of the drifting term is large, rigorously proving the numerically observed phenomenon of concentration of biomass for this model. Moreover, we reveal four critical numbers for the model and provide further insights to the problem being modelled.

Key words. quasi-linear, nonlocal dependence, phytoplankton, concentration phenomenon, reaction-diffusion equation

AMS subject classifications. 35J55, 35J65, 92D25

DOI. 10.1137/070706641

1. Introduction.

We continue our investigation in [DH] on the problem

$$\begin{cases} -[d_1 u_x + \sigma c(x) u]_x = [g(x) - m]u, & 0 < x < 1, \\ -d_2 v_{xx} = -g(x)u, & 0 < x < 1, \\ d_1 u_x + \sigma c(x) u = 0, & x = 0,\ 1, \\ v_x(0) = 0,\ v_x(1) = \beta[v_0 - v(1)], \end{cases} \tag{1.1}$$

where d_1, d_2, σ, m, v_0, and β are positive constants,

$$g(x) = f(\min\{\alpha v(x), w(x)\}), \qquad f(s) = \frac{rs}{K_I + s},$$

and

$$w(x) = w_0 \exp\left[-A_0 x - A \int_0^x u(s)ds\right],$$

with α, r, K_I, w_0, A, and A_0 positive constants. We are interested in positive solutions of (1.1), namely, $u > 0$ and $v > 0$ in $[0, 1]$. From (1.1) it is easy to see that for any such solution v is an increasing function. Clearly w is a decreasing function. The function $c(x)$ is defined by

$$c(x) = \frac{x - x_0}{\delta + |x - x_0|},$$

*Received by the editors October 29, 2007; accepted for publication (in revised form) July 2, 2008; published electronically November 5, 2008. This research was partially supported by the Australian Research Council, the NCTS, and the National Science Council of Taiwan.
http://www.siam.org/journals/sima/40-4/70664.html
†Department of Mathematics, School of Science and Technology, University of New England, Armidale, NSW2351, Australia (ydu@turing.une.edu.au), and Department of Mathematics, Qufu Normal University, Qufu, Shandong 273165, People's Republic of China.
‡Department of Mathematics, National Tsing-Hua University, Hsinchu, Taiwan 300, Republic of China (sbhsu@math.nthu.edu.tw).

where $\delta > 0$ is a small constant and $x_0 \in [0,1]$ is uniquely determined by the following description:

$$\min\{\alpha v(x), w(x)\} = \alpha v(x) \quad \forall x \in [0, x_0); \quad \min\{\alpha v(x), w(x)\} = w(x) \quad \forall x \in (x_0, 1].$$

(Due to the monotonicity of $v(x)$ and $w(x)$, such x_0 always exists.)

Problem (1.1) is a rescaled version of a model proposed by Klausmeier and Litchman in [KL] for the study of phytoplankton in a one-dimensional water column, where $u(x)$ represents the distribution of phytoplankton biomass, $v(x)$ stands for the distribution of nutrient, and x denotes the depth in the water column, with $x = 0$ at the surface and $x = 1$ at the bottom. The term $\sigma c(x)$ is used to describe the active movement of the biomass towards spatial location with optimal growth condition. Klausmeier and Litchman [KL] use this system to model the concentration phenomenon of phytoplankton in lakes and oceans, and the numerical analysis in [KL] demonstrates that, for large σ, the biomass function $u(x)$ concentrates at a certain level $x = x_*$, while the nutrient function $v(x)$ is close to a piecewise linear function. They then treat u as a constant multiple of the δ-function concentrating at x_* and propose a game theoretical model to determine the location of x_*. We refer the reader to Part I [DH] for further details regarding the background of (1.1).

Here we rigorously prove the existence of such a concentration phenomenon and obtain exact formulas for the determination of x_* and the total biomass. In doing so, we reveal the existence of four critical values $v_{**} < v_* < v^* < v^{**}$ for v_0 (the nutrient level at the sediment) such that

(i) $x_* = 0$ when $v_0 \geq v^*$, $x_* \in (0,1)$ when $v_0 \in (v_*, v^*)$, and $x_* = 1$ when $v_0 \leq v_*$;

(ii) the total biomass increases with v_0 in the range $v_{**} < v_0 < v^{**}$, but stays constant for $v_0 \geq v^{**}$ or $v_0 \leq v_{**}$ (but with v_0 above a certain level so that the biomass can survive).

In order to give a more detailed description of these results, we first recall the main results of Part I [DH], where we proved the following two theorems.

THEOREM 1.1. *There exist* $0 < m_* \leq m^* < \infty$ *such that* (1.1) *has a positive solution for* $m \in (0, m_*)$ *and has no positive solution for* $m > m^*$.

The values of m_* and m^* depend on the parameters in (1.1). To stress their dependence on σ, we write $m_* = m_*(\sigma)$, $m^* = m^*(\sigma)$.

THEOREM 1.2.

$$\lim_{\sigma\to\infty} m_*(\sigma) = \lim_{\sigma\to\infty} m^*(\sigma) = f(\min\{\alpha v_0, w_0\}).$$

To investigate the limiting profile of the positive solutions of (1.1) as $\sigma \to \infty$, we will fix m such that $0 < m < f(\min\{\alpha v_0, w_0\})$ and let σ_n be an increasing sequence of positive numbers converging to ∞. By Theorems 1.1 and 1.2, for all large n, (1.1) with $\sigma = \sigma_n$ has at least one positive solution. Suppose that (u_n, v_n) is such a solution. We will analyze the behavior of (u_n, v_n) as $n \to \infty$. This will be done in the following two sections.

In section 2, we find all the possible limiting profiles that a subsequence of $\{(u_n, v_n)\}$ can have; in particular, we find the limiting equations governing these possible limiting profiles. More precisely, let $x_n \in [0,1]$ be uniquely determined by

$$\begin{cases} \min\{\alpha v_n(x), w_n(x)\} = \alpha v_n(x) & \text{for } x \in [0, x_n), \\ \min\{\alpha v_n(x), w_n(x)\} = w_n(x) & \text{for } x \in (x_n, 1], \end{cases}$$

where

$$w_n(x) = w_0 \exp\Big[-A_0 x - A\int_0^x u_n(s)ds\Big].$$

We consider the following possibilities:

$$\text{(i) } x_n \to x_* \in (0,1), \qquad \text{(ii) } x_n \to 0, \qquad \text{(iii) } x_n \to 1.$$

The cases (ii) and (iii) are each further divided into two subcases, namely, for case (ii),

$$\text{(a1) } \sigma_n^{1/2} x_n \to \infty, \qquad \text{(a2) } \sigma_n^{1/2} x_n \to a_* \in [0,\infty);$$

for case (iii),

$$\text{(b1) } \sigma_n^{1/2}(1-x_n) \to \infty, \qquad \text{(b2) } \sigma_n^{1/2}(1-x_n) \to b_* \in [0,\infty).$$

One easily sees that, subject to a subsequence, the above are all the possible behaviors of the sequence $\{x_n\}$. Eventually we will show in section 3 that the limit of the entire sequence $\{x_n\}$ always exists and that this limit is completely determined by the value of v_0, which in turn allows us to completely determine the profiles of u_n and v_n for large n. But in order to prove these facts, we need to first find all the possible limiting profiles of $\{(u_n, v_n)\}$ and the limiting equations that govern these profiles for each of the above listed cases. The main results of section 2 are summarized below.

If case (i) occurs, we show (see Lemma 2.3) that as $n \to \infty$, subject to a subsequence, $u_n \to 0$ uniformly in $[0, x_* - \epsilon] \cup [x_* + \epsilon, 1]$ for any small $\epsilon > 0$ and

$$\int_0^1 u_n(x)dx \to \tau_* C_0, \qquad C_0 := \int_{-\infty}^{\infty} e^{-x^2/(2\delta d_1)}dx = \sqrt{2\delta d_1 \pi},$$

$$v_n \to v_0 - \frac{\tau_*}{d_2} m C_0 (1 + \beta^{-1} - \max\{x, x_*\})$$

uniformly in $[0,1]$, where $x_* \in (0,1)$ and $\tau_* > 0$ are determined by

$$\text{(1.2)} \qquad \begin{cases} w_0 e^{-A_0 x_* - A\tau_*(C_0/2)} = \alpha\Big[v_0 - \dfrac{\tau_*}{d_2} m C_0 (1 + \beta^{-1} - x_*)\Big], \\ m = \displaystyle\int_0^1 f\big(w_0 e^{-A_0 x_* - A\tau_* \max\{C_0/2, C_0 y\}}\big)dy. \end{cases}$$

If case (ii)(a1) occurs, we show (see Lemma 2.4) that the above conclusions hold with $x_* = 0$; in particular,

$$\text{(1.3)} \qquad w_0 e^{-A\tau_*(C_0/2)} = \alpha\Big[v_0 - \frac{\tau_*}{d_2} m C_0 (1 + \beta^{-1})\Big],$$

and

$$\text{(1.4)} \qquad m = \int_0^1 f\big(w_0 e^{-A\tau_* \max\{C_0/2, C_0 y\}}\big)dy.$$

Since (1.4) uniquely determines $\tau_* > 0$, we can substitute this τ_* into (1.3) to obtain a special value for v_0, say $v_0 = v_*$.

Similarly, if case (iii)(b1) occurs, we can show (Lemma 2.5) that the conclusions of case (i) hold except that $x_* = 1$; in particular,

$$w_0 e^{-A_0 - A\tau_*(C_0/2)} = \alpha\left[v_0 - \frac{\tau_*}{d_2} mC_0(1+\beta^{-1}-1)\right], \tag{1.5}$$

and

$$m = \int_0^1 f\big(w_0 e^{-A_0 - A\tau_* \max\{C_0/2, C_0 y\}}\big)dy. \tag{1.6}$$

Analogously, $\tau_* > 0$ is uniquely determined by (1.6), and one can then use (1.5) to obtain a special value for v_0, say $v_0 = v^*$.

If case (ii)(a2) occurs, we show (Lemma 2.4) that as $n \to \infty$, subject to a subsequence, $u_n \to 0$ uniformly in $[\epsilon, 1]$ for any small $\epsilon > 0$,

$$\int_0^1 u_n(x)dx \to \tau_* C(a_*), \qquad C(a_*) = \int_{-a_*}^{\infty} e^{-x^2/(2\delta d_1)}dx,$$

$$v_n \to v_0 - \frac{\tau_*}{d_2} mC(a_*)(1+\beta^{-1}-x)$$

uniformly in $[0,1]$, where $a_* \in [0,\infty)$ and $\tau_* > 0$ are determined by

$$m = \int_0^1 f\big(w_0 e^{-A\tau_* \max\{C(a_*) - C_0/2, C(a_*)y\}}\big)dy, \tag{1.7}$$

and

$$\alpha\Big(v_0 - \frac{\tau_*}{d_2} mC(a_*)(1+\beta^{-1})\Big) = w_0 e^{-A\tau_*[C(a_*) - C_0/2]} \quad \text{if } a_* > 0, \tag{1.8}$$

$$\alpha\left(v_0 - \frac{\tau_*}{d_2} m\left(\frac{C_0}{2}\right)(1+\beta^{-1})\right) \geq w_0 \quad \text{if } a_* = 0. \tag{1.9}$$

If case (iii)(b2) occurs, we show (Lemma 2.5) that as $n \to \infty$, subject to a subsequence, $u_n \to 0$ uniformly in $[0, 1-\epsilon]$ for any small $\epsilon > 0$,

$$\int_0^1 u_n(x)dx \to \tau_* \tilde{C}(b_*), \qquad \tilde{C}(b_*) = \int_{-\infty}^{b_*} e^{-x^2/(2\delta d_1)}dx = C(-b_*),$$

and

$$v_n \to v_0 - \frac{\tau_*}{d_2}\beta^{-1}\tilde{C}(b_*),$$

where $b_* \in [0,\infty)$ and $\tau_* > 0$ are determined by

$$m = \int_0^1 f\big(w_0 e^{-A_0 - A\tau_* \max\{C_0/2, \tilde{C}(b_*)y\}}\big)dy \tag{1.10}$$

and

$$\alpha\Big(v_0 - \frac{\tau_*}{d_2\beta} C(b_*)\Big) = w_0 e^{-A_0 - A\tau_* C_0/2} \quad \text{if } b_* > 0, \tag{1.11}$$

$$\alpha\left(v_0 - \frac{\tau_*}{d_2\beta}\left(\frac{C_0}{2}\right)\right) \leq w_0 e^{-A_0 - A\tau_* C_0/2} \quad \text{if } b_* = 0. \tag{1.12}$$

In section 3, through careful analysis of the limiting equations (1.2)–(1.12), we show that the entire sequence $\{x_n\}$ always converges to a point $x_* \in [0,1]$, that exactly one of the cases considered in section 2 occurs, and that in each case the limit of the entire sequence in the conclusion exists. More precisely, if $v_* < v_0 < v^*$ (recall that v_* and v^* are defined above in cases (ii)(a1) and (iii)(b1), respectively), then case (i) must occur, and (1.2) uniquely determines x_* and τ_*. If $v_0 = v_*$, then case (ii)(a1) occurs; if $v_0 = v^*$, then case (iii)(b1) occurs. If $v_0 > v^*$, then case (ii)(a2) must happen, and if $v_0 < v_*$, then case (iii)(b2) must happen. Moreover, our analysis on the limiting total biomass $\lim_{n\to\infty} \int_0^1 u_n(x)dx$ reveals two further critical values of v_0, $v_{**} < v_*$ and $v^{**} > v^*$ such that this limiting total biomass is strictly increasing with v_0 for v_0 in the range $v_{**} \le v_0 \le v^{**}$ but remains constant (i.e., no longer changes with v_0) when $v_0 \ge v^{**}$ or when $v_0 \le v_{**}$. See Theorems 3.1–3.3 for more accurate descriptions of these results.

In section 4 we give biological interpretations of our main results and compare our rigorous limiting equations with the game theoretical model of [KL].

Though the proofs are rather involved, they consist mainly of elementary mathematical analysis; most of the proofs in section 2 and all of the arguments in section 3 can be understood with sound knowledge of calculus and real analysis.

2. The limiting equations. We will keep using the notation of Part I [DH]. It turns out that the techniques used in the proof of Theorem 3.1 in Part I are not quite suitable for our purpose here. We will introduce some different techniques.

Suppose that $0 < m < f(\min\{\alpha v_0, w_0\})$ and σ_n, (u_n, v_n) are as given in the introduction above. Suppose $c_{v_n,w_n}(x) = C_{x_n}(x)$, $x_n \in [0,1]$. By passing to a subsequence we may assume that $x_n \to x_* \in [0,1]$. Then

$$C_{x_n} = \frac{x - x_n}{\delta + |x - x_n|} \to C_{x_*}$$

in $C^1([0,1])$.

In order to obtain useful equations to determine the profiles of u_n and v_n, we need to stretch the variable x appropriately. We define

$$\Phi_n(x) = \exp\left[-\frac{\sigma_n}{2d_1}\int_{x_n}^{x} C_{x_n}(s)ds\right]$$

and

$$\Psi_n(x) = u_n(x)/\Phi_n(x).$$

By a direct computation we obtain

$$\begin{cases} -d_1\Psi_n'' + \sigma_n\Gamma_n(x)\Psi_n = \big[f(\min\{\alpha v_n, w_n\}) - m\big]\Psi_n, & x \in (0,1), \\ d_1\Psi_n' + (\sigma_n/2)C_{x_n}\Psi_n = 0, & x = 0,1, \end{cases}$$

where

$$\Gamma_n(x) := \frac{\sigma_n(x - x_n)^2 - 2d_1\delta}{4d_1(\delta + |x - x_n|)^2}.$$

Let us introduce the stretched variable $y = \sigma^{1/2}(x - x_n)$ and define

$$V_n(y) := \Psi_n\big(\sigma_n^{-1/2}y + x_n\big), \qquad C_n(y) := \sigma_n^{1/2}C_{x_n}\big(\sigma_n^{-1/2}y + x_n\big) = \frac{y}{\delta + \sigma_n^{-1/2}|y|},$$

$$a_n := -\sigma_n^{1/2}x_n, \qquad b_n := \sigma_n^{1/2}(1 - x_n),$$

and

$$F_n(y) := f\big(\min\{\alpha v_n(\sigma_n^{-1/2}y + x_n), w_n(\sigma_n^{-1/2}y + x_n)\}\big).$$

Then

$$\text{(2.1)}\quad \begin{cases} -d_1 V_n'' + \dfrac{y^2 - 2d_1\delta}{4d_1(\delta + \sigma_n^{-1/2}|y|)^2} V_n = \sigma_n^{-1}\big[F_n(y) - m\big]V_n, & y \in (a_n, b_n), \\ d_1 V_n' + (1/2)C_n V_n = 0, & y = a_n, b_n. \end{cases}$$

In the discussions below, we will consider the cases $x_* \in (0,1)$, $x_* = 0$, and $x_* = 1$ separately.

LEMMA 2.1. *Suppose $x_n \to x_* \in (0,1)$, and set $\tilde{V}_n(y) = V_n(y)/\|V_n\|_{L^\infty([a_n,b_n])}$. Then*

$$\tilde{V}_n \to V_0 \quad \text{in } C^1(J) \text{ for any finite interval } J \subset (-\infty, \infty),$$

where $V_0(y) = \exp[-\frac{y^2}{4d_1\delta}]$ is the unique solution of

$$-d_1 V'' = \frac{2d_1\delta - y^2}{4d_1\delta^2} V, \quad 0 < V \le 1,\ V(0) = 1,\ V'(0) = 0.$$

Proof. Since $x_* \in (0,1)$, we have $a_n \to -\infty$ and $b_n \to \infty$ as $n \to \infty$. Let us note that, for $y \in [a_n, -(2d_1\delta)^{1/2} - \epsilon]$ with $\epsilon > 0$ sufficiently small and all large n, the first equation in (2.1) implies that $V_n''(y) > 0$. Since $d_1 V_n'(a_n) = -(1/2)C_n(a_n)V_n(a_n) \ge 0$, we deduce that $V_n'(y) > 0$ in $(a_n, -(2d_1\delta)^{1/2} - \epsilon]$ for all large n. Hence V_n is increasing in this range. Similarly, we can see that $V_n(y)$ is decreasing in the range $y \in [(2d_1\delta)^{1/2} + \epsilon, b_n]$ for all large n. Therefore $\max V_n = V_n(y_n)$ for some $y_n \in [-(2d_1\delta)^{1/2} - \epsilon, (2d_1\delta)^{1/2} + \epsilon]$, and $\tilde{V}_n(y) = V_n(y)/V_n(y_n)$. We may assume that $y_n \to y^*$ as $n \to \infty$. We now define

$$\tilde{F}_n(y) := \frac{2d_1\delta - y^2}{4d_1(\delta + \sigma_n^{-1/2}|y|)^2} + \sigma_n^{-1}\big[F_n(y) - m\big].$$

Then $\tilde{V}_n(y_n) = 1$, and

$$\text{(2.2)}\quad \begin{cases} -d_1 \tilde{V}_n'' = \tilde{F}_n \tilde{V}_n, \quad 0 < \tilde{V}_n \le 1, & y \in (a_n, b_n), \\ d_1 \tilde{V}_n' + (1/2)C_n \tilde{V}_n = 0, & y = a_n,\ b_n. \end{cases}$$

Since $\{\tilde{F}_n\}$ is uniformly bounded over any bounded interval and $0 \le \tilde{V}_n \le 1$, we may apply the interior L^p theory (see [GT]) to (2.2) and use the Sobolev imbedding theorem and a standard diagonal argument to conclude that, by passing to a subsequence, $\tilde{V}_n \to \tilde{V}$ in $C^1(J)$ for any bounded interval J, and $\tilde{V}$ satisfies

$$\text{(2.3)}\quad -d_1 \tilde{V}'' = \frac{2d_1\delta - y^2}{4d_1\delta^2}\tilde{V}, \quad 0 < \tilde{V} \le 1 \text{ in } (-\infty, \infty),\ \tilde{V}(y^*) = 1,\ \tilde{V}'(y^*) = 0.$$

By the monotonicity property of $V_n(y)$ observed earlier, we know that $\tilde{V}(y)$ is nondecreasing in $(-\infty, -(2d_1\delta)^{1/2})$ and is nonincreasing in $((2d_1\delta)^{1/2}, \infty)$. We can now use (2.3) to conclude that $\tilde{V}'(y)$ is positive and increasing in $(-\infty, -(2d_1\delta)^{1/2})$, reaching a positive maximum at $y = -(2d_1\delta)^{1/2}$; then is decreasing in $(-(2d_1\delta)^{1/2},$

$(2d_1\delta)^{1/2}$), reaching a negative minimum at $y = (2d_1\delta)^{1/2}$; and for $y > (2d_1\delta)^{1/2}$, is increasing and stays negative. Therefore $V'(y)$ has a unique zero at some $y_0 \in \big(-(2d_1\delta)^{1/2}, (2d_1\delta)^{1/2}\big)$, which is the unique maximum point of $\tilde{V}$. Thus $y_0 = y^*$. In other words, $\tilde{V}(y)$ is increasing in $(-\infty, y^*)$ and is decreasing in $(y^*, 0)$. It then follows from an elementary analysis that $\tilde{V}$ decays to 0 as $|y| \to \infty$, and there exists $C_1, C_2 > 0$ such that

$$\tilde{V}(y),\ |\tilde{V}'(y)| \le C_1 e^{-C_2|y|} \quad \forall y \in (-\infty, \infty).$$

We now multiply $\tilde{V}(-y)$ to (2.3), integrate over $[y^*, \infty)$, and then apply integration by parts. Since $\tilde{V}(-y)$ satisfies the differential equation in (2.3), we deduce

$$\tilde{V}'(-y^*)\tilde{V}(y^*) + \tilde{V}'(y^*)\tilde{V}(-y^*) = 0.$$

It follows that $\tilde{V}'(-y^*) = 0$. Since y^* is the only zero of $\tilde{V}'$, we must have $y^* = -y^*$, that is, $y^* = 0$. By the uniqueness theorem of initial value problems of ordinary differential equations, we must have $\tilde{V} = V_0$, the unique solution of (2.3) with $y^* = 0$. A simple calculation confirms that the function $\exp[-\frac{y^2}{4d_1\delta}]$ solves the equation for V_0. Hence, by uniqueness,

$$V_0(y) = \exp\left[-\frac{y^2}{4d_1\delta}\right].$$

Since V_0 is uniquely determined, it follows that the entire original sequence $\{\tilde{V}_n\}$ converges to V_0. □

Using the monotonicity of $\tilde{V}_n$ and the fact that $V_0(y) \to 0$ as $|y| \to \infty$, we see that Lemma 2.1 implies

$$\big\|\Psi_n(\cdot)/\|\Psi_n\|_\infty - V_0\big(\sigma_n^{1/2}(\cdot - x_n)\big)\big\|_{L^\infty([0,1])} \to 0 \quad \text{as } n \to \infty, \tag{2.4}$$

since for large n the function

$$\Psi_n(x)/\|\Psi_n\|_\infty - V_0\big(\sigma_n^{1/2}(x - x_n)\big)$$

is uniformly small at those values of $x \in [0,1]$ such that $\sigma_n^{1/2}(x - x_n)$ stays bounded (by Lemma 2.1), and, by the properties of $\tilde{V}_n$ and V_0, the values of the function at the remaining $x \in [0,1]$ are also small.

We now denote $\tilde{\Psi}_n(x) = \Psi_n(x)/\|\Psi_n\|_\infty$ and consider the function

$$\tilde{u}_n(x) := \sigma_n^{1/2}\Phi_n(x)\tilde{\Psi}_n(x) = \left(\frac{\sigma_n^{1/2}}{\|\Psi_n\|_\infty}\right)u_n.$$

We will show that, for large n, $\tilde{u}_n$ behaves like the δ-function concentrating at x_*. Indeed, we have the following result.

LEMMA 2.2. *For any given small $\epsilon > 0$, $|x - x_n| \ge \epsilon$ implies*

$$0 < \tilde{u}_n(x) \le \sigma_n^{1/2} \exp\left[-\frac{\sigma_n}{4(\delta+1)d_1}\epsilon^2\right] \to 0. \tag{2.5}$$

Moreover, when $x_n \to x_* \in (0,1)$,

$$\lim_{n\to\infty} \int_0^1 \tilde{u}_n(x)dx = C_0 := \int_{-\infty}^{\infty} e^{-x^2/(2\delta d_1)}dx = \sqrt{2\delta d_1 \pi}. \tag{2.6}$$

Proof. For any given small $\epsilon > 0$, there exists $\delta_0 = \delta_0(\epsilon) > 0$ small so that, when $|x - x_n| \le \delta_0$,

$$\exp\Big[-\frac{\sigma_n}{4\delta d_1}(x-x_n)^2\Big] \le \Phi_n(x) \le \exp\Big[-\frac{\sigma_n(1-\epsilon)}{4\delta d_1}(x-x_n)^2\Big].$$

For any $x \in [0,1]$, we have

$$\exp\Big[-\frac{\sigma_n}{4\delta d_1}(x-x_n)^2\Big] \le \Phi_n(x) \le \exp\Big[-\frac{\sigma_n}{4(\delta+1) d_1}(x-x_n)^2\Big].$$

Since $\tilde{\Psi}_n \le 1$, for $|x - x_n| \ge \epsilon$, we have

$$\tilde{u}_n(x) \le \sigma_n^{1/2} \exp\Big[-\frac{\sigma_n}{4(\delta+1)d_1}\epsilon^2\Big] \to 0.$$

This proves (2.5). Moreover, we have

$$\begin{aligned}
\int_0^1 \tilde{u}_n(x)dx &= \int_{x_n-\epsilon}^{x_n+\epsilon} \sigma_n^{1/2}\Phi_n(x)\tilde{\Psi}_n(x)dx + o(1)\\
&= \int_{-\epsilon\sigma_n^{1/2}}^{\epsilon\sigma_n^{1/2}} \Phi_n(x_n+\sigma_n^{-1/2}y)\tilde{V}_n(y)dy + o(1)\\
&= \int_{-\infty}^{\infty} \exp\Big[-\frac{y^2}{4d_1\delta}\Big]V_0(y)dy + o(1)\\
&= \int_{-\infty}^{\infty} \exp\Big[-\frac{y^2}{2d_1\delta}\Big]dy + o(1).
\end{aligned}$$

Hence (2.6) holds. For later application, let us also note from the above argument that

$$\lim_{n\to\infty}\int_{x_n}^1 \tilde{u}_n(x)dx = \lim_{n\to\infty}\int_0^{x_n} \tilde{u}_n(x)dx = C_0/2. \qquad \square \tag{2.7}$$

Denote $\tau_n := \|\Psi_n\|_\infty \sigma_n^{-1/2}$. We find that

$$u_n(x) = \tau_n \tilde{u}_n(x).$$

LEMMA 2.3. *Suppose that $x_n \to x_* \in (0,1)$. Then $\{\tau_n\}$ has a subsequence, still denoted by itself, such that $\tau_n \to \tau_* > 0$. Moreover, τ_* and x_* must satisfy*

$$w_0 e^{-A_0 x_* - A\tau_*(C_0/2)} = \alpha\Big[v_0 - \frac{\tau_*}{d_2} m C_0(1+\beta^{-1}-x_*)\Big] \tag{2.8}$$

and

$$m = \int_0^1 f\big(w_0 e^{-A_0 x_* - A\tau_* \max\{C_0/2, C_0 y\}}\big)dy. \tag{2.9}$$

Furthermore, by possibly passing to a further subsequence, $u_n \to 0$ in $C([0,1]\setminus[x_*-\epsilon, x_*+\epsilon])$, for all $\epsilon > 0$, and

$$v_n(x) \to v_0 - \frac{\tau_*}{d_2} m C_0(1+\beta^{-1} - \max\{x, x_*\}) \tag{2.10}$$

uniformly in $[0,1]$.

Proof. By passing to a subsequence, we have two possible cases:

$$\text{(i)}\ \ \tau_n \to \infty, \quad \text{(ii)}\ \ \tau_n \to \tau_* \in [0,\infty).$$

Step 1. Case (i) cannot happen.

Suppose $\tau_n \to \infty$; we are going to derive a contradiction. Denote

$$f_n = f(\min\{\alpha v_n, w_n\}).$$

Since

$$w_n(x_n) \le w_0 e^{-A\tau_n \int_0^{x_n} \tilde{u}_n(s)ds},$$

and by (2.7)

$$\int_0^{x_n} \tilde{u}_n(s)ds \to C_0/2 > 0,$$

we easily see that $w_n(x_n) \to 0$. It follows that

$$\|f_n\|_\infty = f_n(x_n) = f(w_n(x_n)) \to 0.$$

This implies that

$$\int_0^1 f_n \tilde{u}_n dx \to 0.$$

On the other hand, we may integrate the equation for u_n to obtain

$$\int_0^1 [f_n(x) - m]u_n dx = 0,$$

which implies that

$$\int_0^1 [f_n(x) - m]\tilde{u}_n dx = 0.$$

Letting $n \to \infty$ and using (2.6), we obtain

$$mC_0 = \lim_{n\to\infty} \int_0^1 f_n \tilde{u}_n dx = 0,$$

which contradicts our assumption that $m > 0$. Therefore case (i) cannot happen.

Step 2. The limiting profile of u_n and v_n.

We next consider case (ii), namely, $\tau_n \to \tau_* \in [0,\infty)$. In this case, due to (2.5), $u_n = \tau_n \tilde{u}_n \to 0$ in $C([0,1] \setminus [x_* - \epsilon, x_* + \epsilon])$, for all $\epsilon > 0$, and hence

$$\tau_n f_n \tilde{u}_n \to 0 \quad \text{uniformly in } [0, x_* - \epsilon] \cup [x_* + \epsilon, 1] \quad \forall \epsilon > 0. \tag{2.11}$$

Let $\zeta_n = v_0 - v_n$. Then

$$-d_2 \zeta_n'' = \tau_n f_n \tilde{u}_n \text{ in } (0,1), \quad \zeta_n'(0) = 0, \quad \zeta_n'(1) + \beta \zeta_n(1) = 0. \tag{2.12}$$

Since $v_n \geq 0$, we have $\zeta_n \leq v_0$. Since $\tau_n f_n \tilde{u}_n > 0$, from (2.12) and the maximum principle, we deduce that $\zeta_n > 0$. Hence we always have $0 < \zeta_n \leq v_0$. Therefore we can integrate (2.12) to obtain

$$\eta_n := \tau_n \int_0^1 f_n \tilde{u}_n dx = d_2[\zeta_n'(0) - \zeta_n'(1)] = d_2\beta\zeta_n(1) \in [0, d_2\beta v_0].$$

This implies that, by passing to a subsequence, we may assume that $\eta_n \to \eta_* \in [0, d_2\beta v_0]$.

Moreover, using (2.11), (2.12), and $\eta_n \to \eta_*$, we find that

$$\begin{aligned}
&\{\zeta_n'\} \text{ is a bounded sequence in } L^\infty([0,1]),\\
&\zeta_n'(x) \to 0 \text{ uniformly in } [0, x_* - \epsilon] \quad \forall \epsilon > 0,\\
&\zeta_n'(x) \to -\eta_*/d_2 \text{ uniformly in } [x_* + \epsilon, 1] \quad \forall \epsilon > 0.
\end{aligned}$$

Since, moreover, $0 \leq \zeta_n \leq v_0$, we conclude that $\{\zeta_n\}$ is precompact in $C([0,1])$. Hence, by passing to a subsequence, we may assume that $\zeta_n \to \zeta$ in $C([0,1])$.

On the other hand, we may apply the L^p theory to (2.12) and the Sobolev imbedding theorem to find a further subsequence, still denoted by ζ_n, such that $\zeta_n \to \tilde{\zeta}$ in $C^1(J)$ for any compact interval $J \subset [0, x_*) \cup (x_*, 1]$, and $\tilde{\zeta}$ satisfies (in the weak sense)

$$-d_2\tilde{\zeta}'' = 0 \text{ in } [0, x_*) \cup (x_*, 1], \qquad \tilde{\zeta}'(0) = 0, \qquad \tilde{\zeta}'(1) + \beta\tilde{\zeta}(1) = 0.$$

Clearly we must have $\tilde{\zeta} = \zeta$. Moreover, our earlier analysis on ζ_n implies that $\zeta'(x) = 0$ in $[0, x_*)$ and $\zeta'(x) = -\eta_*/d_2$ in $(x_*, 1]$. These properties uniquely determine ζ:

$$\zeta(x) = (\eta_*/d_2)(1 + \beta^{-1} - \max\{x_*, x\}). \tag{2.13}$$

Step 3. $\tau_* > 0$.

Otherwise, $\tau_* = 0$ and hence $\eta_* = 0$. It follows that $\zeta = 0$ and $v_n \to v_0$ uniformly in $[0,1]$, and that

$$w_n(x) = w_0 e^{-A_0 x} e^{-A\tau_n \int_0^x \tilde{u}_n(s)ds} \to w_0 e^{-A_0 x} = w_*(x)$$

uniformly in $[0,1]$. This implies that

$$x_* = x_0^* \quad \text{and} \quad f_n(x) \to f_0(x) := f(\min\{\alpha v_0, w_*\}) \text{ uniformly in } [0,1].$$

We may now integrate the equation for u_n to obtain, as before,

$$\int_0^1 [f_n(x) - m]\tilde{u}_n dx = 0.$$

Letting $n \to \infty$, we deduce

$$[f_0(x_0^*) - m]C_0 = 0,$$

which contradicts our assumption that $m < f(\min\{\alpha v_0, w_0\}) = f_0(x_0^*)$. Hence $\tau_* > 0$.

Step 4. The equations for x_* and τ_*.

We now set out to find the equations that determine x_* and τ_*. By (2.7),

$$w_n(x_n) = w_0 e^{-A_0 x_n} e^{-A\tau_n \int_0^{x_n} \tilde{u}_n(s)ds} \to w_0 e^{-A_0 x_*} e^{-A\tau_*(C_0/2)}.$$

On the other hand,

$$w_n(x_n) = \alpha v_n(x_n) \to \alpha\big[v_0 - \zeta(x_*)\big].$$

Thus we necessarily have

$$w_0 e^{-A_0 x_* - A\tau_*(C_0/2)} = \alpha\big[v_0 - \zeta(x_*)\big] = \alpha\big[v_0 - (\eta_*/d_2)(1+\beta^{-1} - x_*)\big]. \tag{2.14}$$

Moreover, using (2.5), (2.7), and the fact that $\alpha v_n \to \alpha(v_0 - \zeta)$ uniformly in $[0,1]$, we deduce

$$\int_0^{x_n} f(\alpha v_n)\tilde{u}_n dx \to (C_0/2) f\big(\alpha v_0 - \alpha\zeta(x_*)\big). \tag{2.15}$$

Using

$$w_n(x) = w_0 e^{-A_0 x} e^{-A\tau_n \int_0^x \tilde{u}_n(s)ds}$$

and the property of $\tilde{u}_n$, we obtain, for any small $\epsilon > 0$,

$$\begin{aligned}
&\int_{x_n}^1 f(w_n)\tilde{u}_n dx \\
&\quad = \int_{x_n}^1 f\big(w_0 e^{-A_0 x - A\tau_n \int_0^{x_n} \tilde{u}_n(s)ds - A\tau_n \int_{x_n}^x \tilde{u}_n(s)ds}\big)\tilde{u}_n dx \\
&\quad = \int_{x_n}^{x_*+\epsilon} f\big(w_0 e^{-A_0 x - A\tau_n(C_0/2) e^{-A\tau_n \int_{x_n}^x \tilde{u}_n(s)ds}}\big)\tilde{u}_n dx + o(1) \\
&\quad = [1+o_\epsilon(1)] \int_{x_n}^{x_*+\epsilon} f\big(w_0 e^{-A_0 x_* - A\tau_*(C_0/2) e^{-A\tau_n \int_{x_n}^x \tilde{u}_n(s)ds}}\big)\tilde{u}_n dx + o(1) \\
&\quad = [1+o_\epsilon(1)] \int_0^{[\int_{x_n}^1 \tilde{u}_n(s)ds]} f\big(w_0 e^{-A_0 x_* - A\tau_*(C_0/2) e^{-A\tau_n y}}\big)dy + o(1) \\
&\quad = [1+o_\epsilon(1)] \int_0^{C_0/2} f\big(w_0 e^{-A_0 x_* - A\tau_*(C_0/2) e^{-A\tau_* y}}\big)dy + o(1),
\end{aligned}$$

where $o_\epsilon(1)$ represents a quantity that converges to 0 as $\epsilon \to 0$.

Thus

$$\int_{x_n}^1 f(w_n)\tilde{u}_n(x)dx \to \int_0^{C_0/2} f\big(w_0 e^{-A_0 x_* - A\tau_*(C_0/2) e^{-A\tau_* y}}\big)dy \tag{2.16}$$

as $n \to \infty$.

Combining (2.15) and (2.16), we obtain

$$\begin{aligned}
\eta_* &= \lim_{n\to\infty} \tau_n \int_0^1 f_n \tilde{u}_n dx \\
&= \tau_* \left[(C_0/2) f\big(\alpha v_0 - \alpha\zeta(x_*)\big) + \int_0^{C_0/2} f\big(w_0 e^{-A_0 x_* - A\tau_*(C_0/2) e^{-A\tau_* y}}\big)dy \right].
\end{aligned} \tag{2.17}$$

Moreover, we may integrate the equation for u_n to obtain

$$\int_0^1 [f_n(x) - m]\tilde{u}_n dx = 0.$$

Letting $n \to \infty$ and using (2.15), (2.16), we obtain

$$mC_0 = (C_0/2)f(\alpha v_0 - \alpha\zeta(x_*)) + \int_0^{C_0/2} f\big(w_0 e^{-A_0 x_* - A\tau_*(C_0/2)} e^{-A\tau_* y}\big)dy.$$

This combined with (2.17) yields

$$\eta_* = \tau_* mC_0 \tag{2.18}$$

and combined with (2.14) gives

$$\begin{aligned} m &= (1/2)f(w_0 e^{-A_0 x_* - A\tau_*(C_0/2)}) + C_0^{-1}\int_0^{C_0/2} f\big(w_0 e^{-A_0 x_* - A\tau_*(C_0/2)} e^{-A\tau_* y}\big)dy \\ &= C_0^{-1}\int_0^{C_0} f\big(w_0 e^{-A_0 x_* - A\tau_* \max\{C_0/2, y\}}\big)dy \\ &= \int_0^1 f\big(w_0 e^{-A_0 x_* - A\tau_* \max\{C_0/2, C_0 y\}}\big)dy; \end{aligned}$$

thus (2.9) is proved. Equation (2.8) and (2.10) clearly follow from (2.13), (2.14), and (2.18). □

We now consider the case $x_* = 0$. By passing to a subsequence, we have two subcases:

$$\text{(a1) } a_n := \sigma_n^{1/2} x_n \to \infty, \qquad \text{(a2) } a_n \to a_* \in [0, \infty).$$

LEMMA 2.4. *In subcase* (a1)*, all of the conclusions in Lemmas* 2.2 *and* 2.3 *hold. In subcase* (a2)*, $\{\tau_n\}$ has a subsequence, still denoted by itself, such that $\tau_n \to \tau_* > 0$. Moreover, τ_* and a_* must satisfy*

$$m = \int_0^1 f\big(w_0 e^{-A\tau_* \max\{C(a_*) - C_0/2, C(a_*)y\}}\big)dy \tag{2.19}$$

and

$$\alpha\Big(v_0 - \frac{\tau_*}{d_2} mC(a_*)(1+\beta^{-1})\Big) = w_0 e^{-A\tau_*[C(a_*) - C_0/2]} \qquad \text{if } a_* > 0, \tag{2.20}$$

$$\alpha\left(v_0 - \frac{\tau_*}{d_2} m\left(\frac{C_0}{2}\right)(1+\beta^{-1})\right) \geq w_0 \qquad \text{if } a_* = 0, \tag{2.21}$$

where

$$C(a_*) := \int_{-a_*}^{\infty} \exp\left[-\frac{y^2}{2d_1\delta}\right]dy.$$

Furthermore, by possibly passing to a further subsequence, $u_n \to 0$ in $C([\epsilon, 1])$, for all $\epsilon \in (0,1)$,

$$\lim_{n\to\infty}\int_0^{x_n} \tilde{u}_n(x)dx = C(a_*) - C_0/2, \qquad \lim_{n\to\infty}\int_0^1 \tilde{u}_n(x)dx = C(a_*), \tag{2.22}$$

and

$$v_n(x) \to v_0 - \frac{\tau_*}{d_2} mC(a_*)(1 + \beta^{-1} - x) \tag{2.23}$$

uniformly in $[0,1]$.

Proof. In subcase (a1), we may repeat the arguments used for the case $x_* \in (0,1)$ above to see that all the conclusions there (with x_* replaced by 0) remain valid; the proofs carry over with minor modifications.

Consider now subcase (a2). In this case, we may use interior and boundary L^p estimates and the Sobolev imbedding theorem to conclude that, by passing to a subsequence, $\|\tilde{V}_n - \tilde{V}\|_{C^1([a_n,M])} \to 0$ for all $M > 0$, where $\tilde{V}$ satisfies, instead of (2.3),

$$\begin{cases} -d_1\tilde{V}'' = \dfrac{2d_1\delta - y^2}{4d_1\delta^2}\tilde{V}, \quad 0 < \tilde{V} \le 1 \text{ in } (-a_*, \infty), \\ d_1\tilde{V}'(-a_*) - \dfrac{a_*}{2\delta}\tilde{V}(-a_*) = 0, \quad \tilde{V}(y^*) = 1,\ \tilde{V}'(y^*) = 0. \end{cases} \tag{2.24}$$

Note that as before $\tilde{V}$ is decreasing in $[(2d_1\delta)^{1/2}, \infty)$. This and (2.24) imply that $\tilde{V}$ converges to 0 as $y \to \infty$. Moreover, an elementary consideration shows that

$$|\tilde{V}'(y)|, \tilde{V}(y) \le C_1 e^{-C_2 y}$$

for some C_1, $C_2 > 0$, and all $y > 0$.

We will show that $y^* = 0$ and $\tilde{V}$ is again the unique solution of (2.3) with $y^* = 0$, namely V_0. Since V_0 and $|V_0'|$ are bounded from above by a function of the form $C_1 e^{-C_2|y|}$, we can multiply the first equation in (2.24) by V_0, integrate over $[y^*, \infty)$, and use integration by parts to deduce

$$d_1[\tilde{V}V_0' - \tilde{V}'V_0]\big|_{y^*}^{\infty} = 0.$$

It follows that $V_0'(y^*) = 0$, which implies that $y^* = 0$. Therefore, by the uniqueness of initial value problems of the ordinary differential equations, we deduce $\tilde{V} \equiv V_0$. Let us note that a direct calculation shows

$$d_1 V_0'(y) + \frac{y}{2\delta}V_0(y) = 0 \quad \text{for every } y \in (-\infty, \infty).$$

Therefore (2.24) does not introduce any restriction for a_*.

Since now $\sigma_n^{1/2} x_n \to a_*$, instead of (2.6), we have

$$\lim_{n\to\infty} \int_0^{x_n} \tilde{u}_n(x)dx = C(a_*) - C_0/2, \qquad \lim_{n\to\infty} \int_0^{1} \tilde{u}_n(x)dx = C(a_*), \tag{2.25}$$

where

$$C(a_*) := \int_{-a_*}^{\infty} \exp\left[-\frac{y^2}{4d_1\delta}\right] V_0(y)dy = \int_{-a_*}^{\infty} \exp\left[-\frac{y^2}{2d_1\delta}\right] dy.$$

We proceed as in the case $x_* \in (0,1)$ and have two possibilities for τ_n as before. We show that, in the current case, we still cannot have $\tau_n \to \infty$. Arguing indirectly, we assume that $\tau_n \to \infty$.

Then in the case $a_* > 0$, we have $C(a_*) - C_0/2 > 0$, and hence

$$w_n(x_n) \le w_0 e^{-A\tau_n \int_0^{x_n} \tilde{u}_n(s)ds} \to 0.$$

It follows that

$$\|f_n\|_\infty = f_n(x_n) = f(w_n(x_n)) \to 0$$

and

$$\int_0^1 f_n \tilde{u}_n dx \to 0.$$

If $a_* = 0$, then $C(a_*) - C_0/2 = 0$ and

$$\begin{aligned}
\int_0^1 f_n(x)\tilde{u}_n(x)dx &= \int_{x_n}^1 f(w_n(x))\tilde{u}_n(x)dx + o(1)\\
&\le \int_{x_n}^1 f\big(w_0 e^{-A\tau_n \int_{x_n}^x \tilde{u}_n(s)ds}\big)\tilde{u}_n dx + o(1)\\
&= \int_0^{[\int_{x_n}^1 \tilde{u}_n(s)ds]} f\big(w_0 e^{-A\tau_n y}\big)dy + o(1)\\
&\le \epsilon f(w_0) + \int_\epsilon^{C_0/2} f\big(w_0 e^{-A\tau_n y}\big)dy + o(1)\\
&= \epsilon f(w_0) + o(1) \quad \forall \epsilon \in (0, C_0/2).
\end{aligned}$$

Therefore we always have

$$\int_0^1 f_n \tilde{u}_n dx \to 0 \text{ as } n \to \infty.$$

As before, we may integrate the equation for u_n to obtain

$$\int_0^1 [f_n(x) - m]\tilde{u}_n dx = 0.$$

Letting $n \to \infty$ and using the above estimate, we deduce

$$-mC(a_*) = 0,$$

a contradiction to our assumption that $m > 0$. Therefore we cannot have $\tau_n \to \infty$.

Thus we can only have the case $\tau_n \to \tau_*$. Then much as before we deduce $u_n \to 0$ in $C([\epsilon, 1])$ for all $\epsilon \in (0, 1)$, and

$$\zeta_n \to \zeta := (\eta_*/d_2)(1 + \beta^{-1} - x)$$

in $C([0,1]) \cap C^1([\epsilon, 1])$ for all $\epsilon \in (0,1)$. If $\tau_* = 0$, we can deduce as before that $m = f_0(x_0^*)$, a contradiction to our initial assumption on m. Therefore $\tau_* > 0$.

If $a_* = 0$, we first choose $y_n \in (x_n, 1)$ such that $y_n \to 0$ and $\int_{y_n}^1 \tilde{u}_n(x)dx \to 0$, and then we have

$$\begin{aligned}
\int_0^1 f_n(x)\tilde{u}_n(x)dx &= \int_{x_n}^{y_n} f(w_n(x))\tilde{u}_n(x)dx + o(1)\\
&= \int_{x_n}^{y_n} f\big(w_0 e^{-A_0 x - A\tau_n \int_0^{x_n} \tilde{u}_n(s)ds - A\tau_n \int_{x_n}^x \tilde{u}_n(s)ds}\big)\tilde{u}_n dx + o(1)\\
&= \int_{x_n}^{y_n} f\big(w_0 e^{-A\tau_n \int_{x_n}^x \tilde{u}_n(s)ds}\big)\tilde{u}_n dx + o(1)\\
&= \int_{x_n}^1 f\big(w_0 e^{-A\tau_n \int_{x_n}^x \tilde{u}_n(s)ds}\big)\tilde{u}_n dx + o(1)\\
&= \int_0^{C_0/2} f\big(w_0 e^{-A\tau_* y}\big)dy + o(1).
\end{aligned}$$

If $a_* > 0$, then $x_n > 0$ and $w_n(x_n) = \alpha v_n(x_n)$. From

$$v_n(x_n) \to v_0 - \zeta(0)$$

and

$$w_n(x_n) = w_0 e^{-A_0 x_n - A\tau_n \int_0^{x_n} \tilde{u}_n dx} \to w_0 e^{-A\tau_*[C(a_*) - C_0/2]}$$

we obtain

$$\alpha[v_0 - \zeta(0)] = w_0 e^{-A\tau_*[C(a_*) - C_0/2]}.$$

Moreover, similar to the above,

$$\begin{aligned}\int_{x_n}^1 f_n(x)\tilde{u}_n(x) &= \int_{x_n}^{y_n} f(w_n(x))\tilde{u}_n(x)dx + o(1)\\ &= \int_{x_n}^{y_n} f\big(w_0 e^{-A\tau_n \int_0^{x_n} \tilde{u}_n(s)ds - A\tau_n \int_{x_n}^{x} \tilde{u}_n(s)ds}\big)\tilde{u}_n dx + o(1)\\ &= \int_0^{C_0/2} f\big(w_0 e^{-A\tau_*[C(a_*) - C_0/2] - A\tau_* y}\big)dy + o(1),\end{aligned}$$

and

$$\begin{aligned}\int_0^{x_n} f_n(x)\tilde{u}_n(x)dx &= \int_0^{x_n} f(\alpha v_n(x))\tilde{u}_n(x)dx\\ &= f\big(\alpha[v_0 - \zeta(0)]\big)[C(a_*) - C_0/2] + o(1)\\ &= [C(a_*) - C_0/2] f\big(w_0 e^{-A\tau_*[C(a_*) - C_0/2]}\big) + o(1).\end{aligned}$$

Therefore we always have

$$\int_0^1 f_n \tilde{u}_n dx \to \int_0^{C(a_*)} f\big(w_0 e^{-A\tau_* \max\{[C(a_*) - C_0/2], y\}}\big)dy. \tag{2.26}$$

We may now use

$$\int_0^1 [f_n(x) - m]\tilde{u}_n dx = 0$$

to obtain

$$mC(a_*) = \int_0^{C(a_*)} f\big(w_0 e^{-A\tau_* \max\{C(a_*) - C_0/2, y\}}\big)dy.$$

Therefore

$$m = \int_0^1 f\big(w_0 e^{-A\tau_* \max\{C(a_*) - C_0/2, C(a_*)y\}}\big)dy,$$

and (2.19) is proved.

We thus obtain

$$\eta_* = \tau_* \lim_{n\to\infty} \int_0^1 f_n \tilde{u}_n dx = \tau_* m C(a_*).$$

Therefore,

$$v_n(x) \to v_0 - \zeta = v_0 - \frac{\tau_*}{d_2} mC(a_*)(1+\beta^{-1}-x)$$

uniformly in $[0,1]$; that is, (2.23) holds.

Let us note that (2.22) was already proved in (2.25). So it remains to prove (2.20) and (2.21). If $a_* > 0$, then $x_n > 0$, and we necessarily have $\alpha v_n(x_n) = w_n(x_n)$. Recall that

$$w_n(x_n) \to w_0 e^{-A\tau_*[C(a_*)-C_0/2]}, \qquad v_n(x_n) \to v_0 - \zeta(0).$$

Hence

$$\alpha\Big(v_0 - \frac{\tau_*}{d_2} mC(a_*)(1+\beta^{-1})\Big) = w_0 e^{-A\tau_*[C(a_*)-C_0/2]}.$$

If $a_* = 0$, then $x_n = 0$ is possible, and so we have $\alpha v_n(x_n) \ge w_n(x_n)$ in general, and instead of the above identity we should have

$$\alpha\left(v_0 - \frac{\tau_*}{d_2} m\left(\frac{C_0}{2}\right)(1+\beta^{-1})\right) \ge w_0.$$

Thus (2.20) and (2.21) are established. The proof is now complete. □

Finally we consider the case $x_* = 1$. By passing to a subsequence, we have two subcases:

$$\text{(b1) } b_n := \sigma_n^{1/2}(1-x_n) \to \infty, \qquad \text{(b2) } b_n \to b_* \in [0,\infty).$$

LEMMA 2.5. *In subcase* (b1)*, all of the conclusions in Lemmas* 2.2 *and* 2.3 *hold. In subcase* (b2)*, $\{\tau_n\}$ has a subsequence, still denoted by itself, such that $\tau_n \to \tau_* > 0$. Moreover, τ_* and b_* must satisfy*

$$m = \int_0^1 f\big(w_0 e^{-A_0 - A\tau_* \max\{C_0/2, \tilde{C}(b_*)y\}}\big)dy \tag{2.27}$$

and

$$\alpha\Big(v_0 - \frac{\tau_*}{d_2\beta}\tilde{C}(b_*)\Big) = w_0 e^{-A_0 - A\tau_* C_0/2} \quad \text{if } b_* > 0, \tag{2.28}$$

$$\alpha\left(v_0 - \frac{\tau_*}{d_2\beta}\left(\frac{C_0}{2}\right)\right) \le w_0 e^{-A_0 - A\tau_* C_0/2} \quad \text{if } b_* = 0, \tag{2.29}$$

where

$$\tilde{C}(b_*) := \int_{-\infty}^{b_*} \exp\Big[-\frac{y^2}{2d_1\delta}\Big]dy = C(-b_*).$$

Furthermore, by possibly passing to a further subsequence, $u_n \to 0$ in $C([0,1-\epsilon])$ for every $\epsilon \in (0,1)$,

$$\lim_{n\to\infty}\int_{x_n}^1 \tilde{u}_n(x)dx = \tilde{C}(b_*) - C_0/2, \qquad \lim_{n\to\infty}\int_0^1 \tilde{u}_n(x)dx = \tilde{C}(b_*), \tag{2.30}$$

$$v_n(x) \to v_0 - \zeta = v_0 - \frac{\tau_*}{d_2\beta}\tilde{C}(b_*) \tag{2.31}$$

uniformly in $[0,1]$.

Proof. In subcase (b1), we may repeat the arguments used in Lemmas 2.2 and 2.3 for the case $x_* \in (0,1)$ to see that all the conclusions there (with x_* replaced by 1) remain valid; the proofs need only minor modifications.

We now consider subcase (b2). Then instead of (2.3) we have

$$\begin{cases} -d_1\tilde{V}'' = \dfrac{2d_1\delta - y^2}{4d_1\delta}\tilde{V}, \quad 0 < \tilde{V} \le 1 \text{ in } (-\infty, b_*), \\ \tilde{V}'(b_*) + \dfrac{b_*}{2\delta}\tilde{V}(b_*) = 0, \quad \tilde{V}(y^*) = 1,\ \tilde{V}'(y^*) = 0. \end{cases} \tag{2.32}$$

Note that as before $\tilde{V}$ is increasing in $(-\infty, -(2d_1\delta)^{1/2}]$. This and (2.32) imply that $\tilde{V}$ converges to 0 as $y \to -\infty$. Moreover, an elementary consideration shows that

$$|\tilde{V}'(y)|, \tilde{V}(y) \le C_1 e^{-C_2|y|}$$

for some C_1, $C_2 > 0$, and all $y < 0$.

As in the case for (2.24), we can similarly show that $y^* = 0$ and $\tilde{V} \equiv V_0$, the unique solution of (2.3) with $y^* = 0$. Moreover, (2.32) introduces no restriction for b_*.

Since $\sigma_n^{1/2}(1 - x_n) \to b_*$, instead of (2.6), we have

$$\lim_{n\to\infty} \int_{x_n}^1 \tilde{u}_n(x)dx = \tilde{C}(b_*) - C_0/2, \qquad \lim_{n\to\infty} \int_0^1 \tilde{u}_n(x)dx = \tilde{C}(b_*),$$

where

$$\tilde{C}(b_*) := \int_{-\infty}^{b_*} \exp\left[-\frac{y^2}{4d_1\delta}\right] V_0(y)dy = C(-b_*).$$

This establishes (2.30).

We proceed as in the case $x_* \in (0,1)$ and have two possibilities for τ_n as before. We show that in the current case, we still cannot have $\tau_n \to \infty$. Arguing indirectly, we assume that $\tau_n \to \infty$.

Since $\int_0^{x_n} \tilde{u}_n dx \to C_0/2$, we have

$$w_n(x_n) \le w_0 e^{-A\tau_n \int_0^{x_n} \tilde{u}_n(s)ds} \to 0.$$

It follows that

$$\|f_n\|_\infty = f_n(x_n) \le f(w_n(x_n)) \to 0,$$

and

$$\int_0^1 f_n \tilde{u}_n dx \to 0.$$

As before, we may integrate the equation for u_n to obtain

$$\int_0^1 [f_n(x) - m]\tilde{u}_n dx = 0.$$

Letting $n \to \infty$ and using the above estimate, we deduce

$$-m\tilde{C}(b_*) = 0,$$

a contradiction to our assumption that $m > 0$. Therefore we cannot have $\tau_n \to \infty$.

Thus we can have only the case $\tau_n \to \tau_*$. Then much as before we deduce $u_n \to 0$ in $C([0, 1-\epsilon])$ for each $\epsilon \in (0,1)$ and $\zeta_n \to \zeta$ in $C([0,1]) \cap C^1([0,1-\epsilon])$, for all $\epsilon \in (0,1)$, with ζ satisfying

$$\zeta'' = 0 \text{ in } [0,1), \qquad \zeta' = 0 \text{ in } [0,1).$$

Hence ζ is a constant. To determine its value, we use

$$-d_2\zeta_n'(1) = \int_0^1 \tau_n f_n \tilde{u}_n dx \to \tau_* \tilde{C}(b_*)$$

and

$$\zeta_n'(1) + \beta\zeta_n(1) = 0$$

to deduce

$$-\frac{\tau_*}{d_2}\tilde{C}(b_*) + \beta\zeta = 0,$$

and hence

$$\zeta = \frac{\tau_*}{d_2\beta}\tilde{C}(b_*). \tag{2.33}$$

If $\tau_* = 0$, then $\zeta \equiv 0$, and hence $v_n \to v_0$ uniformly in $[0,1]$ and

$$w_n(x) = w_0 e^{-A_0 x - A\tau_n \int_0^x \tilde{u}_n dx} \to w_0 e^{-A_0 x}$$

uniformly in $[0,1]$. Then we can deduce as before that $m = f_0(x_0^*)$, a contradiction to our initial assumption on m. Therefore $\tau_* > 0$.

We have

$$\begin{aligned}\int_0^{x_n} f_n(x)\tilde{u}_n(x)dx &= \int_0^{x_n} f(\alpha v_n(x))\tilde{u}_n(x)dx \\ &= (C_0/2) f\big(\alpha(v_0 - \zeta)\big) + o(1).\end{aligned}$$

If $b_* = 0$, then

$$\int_{x_n}^1 f_n(x)\tilde{u}_n(x)dx = o(1).$$

If $b_* > 0$, then $x_n > 0$ and $w_n(x_n) = \alpha v_n(x_n)$. From

$$v_n(x_n) \to v_0 - \zeta = v_0 - \frac{\tau_*}{d_2\beta}\tilde{C}(b_*)$$

and

$$w_n(x_n) = w_0 e^{-A_0 x_n - A\tau_n \int_0^{x_n} \tilde{u}_n dx} \to w_0 e^{-A_0 - A\tau_* C_0/2},$$

we obtain

$$\alpha\Big(v_0 - \frac{\tau_*}{d_2\beta}\tilde{C}(b_*)\Big) = w_0 e^{-A_0 - A\tau_* C_0/2}. \tag{2.34}$$

Moreover,

$$\begin{aligned}\int_{x_n}^1 f_n(x)\tilde{u}_n(x)dx &= \int_{x_n}^1 f(w_n(x))\tilde{u}_n(x)dx\\ &= \int_{x_n}^1 f\big(w_0 e^{-A_0 x - A\tau_n\int_0^{x_n}\tilde{u}_n(s)ds - A\tau_n\int_{x_n}^x \tilde{u}_n(s)ds}\big)\tilde{u}_n dx\\ &= \int_0^{\tilde{C}(b_*)-C_0/2} f\big(w_0 e^{-A_0 - A\tau_* C_0/2 - A\tau_* y}\big)dy + o(1).\end{aligned}$$

Therefore, whether $b_* = 0$ or $b_* > 0$, we always have

$$\int_0^1 f_n(x)\tilde{u}_n(x)dx \to \int_0^{\tilde{C}(b_*)} f\big(w_0 e^{-A_0 - A\tau_* \max\{C_0/2, y\}}\big)dy. \tag{2.35}$$

We may now use

$$\int_0^1 [f_n(x) - m]\tilde{u}_n dx = 0$$

to obtain

$$m\tilde{C}(b_*) = \int_0^{\tilde{C}(b_*)} f\big(w_0 e^{-A_0 - A\tau_* \max\{C_0/2, y\}}\big)dy,$$

which gives (2.27).

Note that if $b_* = 0$, then $x_n = 1$ is possible, and we have only $w_n(x_n) \geq \alpha v(x_n)$, so instead of (2.27), we should have

$$\alpha\Big(v_0 - \frac{\tau_*}{d_2\beta}\tilde{C}(b_*)\Big) \leq w_0 e^{-A_0 - A\tau_* C_0/2}.$$

Thus we have established (2.28) and (2.29). Clearly (2.31) follows from (2.33) and the fact that $v_n \to v_0 - \zeta$ uniformly in $[0,1]$. The proof is complete. □

3. Limiting profile of the positive solutions. We are now ready to state and prove our main results. We will show that the limiting equations obtained in the previous section uniquely determine x_* and τ_*, and the value of v_0 determines which set of limiting equations should be used for calculating x_* and τ_*. In this way, the asymptotic behavior of the positive solutions is completely determined.

Let us recall that m is fixed such that

$$0 < m < f(\min\{\alpha v_0, w_0\}), \tag{3.1}$$

and $\sigma_n \to \infty$ is a sequence of positive numbers. Therefore by Theorems 1.1 and 1.2, problem (1.1) with $\sigma = \sigma_n$ has a positive solution (u_n, v_n) for all large n. Recall that $C_0 > 0$ is given in (2.6), which is completely determined by δ and d_1. Due to (3.1) there exists a unique $\tau_0^* > 0$ such that

$$m = \int_0^1 f\big(w_0 e^{-A\tau_0^* \max\{C_0/2, C_0 y\}}\big)dy. \tag{3.2}$$

Let us then define

$$v^* = v^*(m) := \frac{w_0}{\alpha}e^{-A\tau_0^* C_0/2} + \frac{\tau_0^*}{d_2}mC_0(1+\beta^{-1}). \tag{3.3}$$

Let $\underline{v}(m) > 0$ be uniquely determined by

$$m = f(\alpha \underline{v}(m)).$$

By (3.1), we always have $v_0 > \underline{v}(m)$.

When $m < f(w_0 e^{-A_0})$, we can find a unique $\tau_1^* > 0$ such that

$$m = \int_0^1 f\big(w_0 e^{-A_0 - A\tau_1^* \max\{C_0/2, C_0 y\}}\big) dy. \tag{3.4}$$

We now define

$$v_* = v_*(m) := \begin{cases} \dfrac{w_0}{\alpha} e^{-A_0 - A\tau_1^* C_0/2} + \dfrac{\tau_1^*}{d_2} m C_0 \beta^{-1} & \text{if } m < f(w_0 e^{-A_0}), \\ \underline{v}(m) & \text{if } f(w_0 e^{-A_0}) \le m < f(w_0). \end{cases} \tag{3.5}$$

It is easily seen that $v_*(m)$ is continuous in m.

As we will see below, to completely determine the asymptotic profile of (u_n, v_n), it is necessary to distinguish the cases $v_0 \in [v_*(m), v^*(m)]$, $v_0 > v^*(m)$, and $v_0 < v_*(m)$.

THEOREM 3.1. *Suppose that* $v_0 > \underline{v}(m)$ *and*

$$v_*(m) \le v_0 \le v^*(m). \tag{3.6}$$

Then the system (2.8) *and* (2.9), *namely,*

$$\begin{cases} w_0 e^{-A_0 x_* - A\tau_*(C_0/2)} = \alpha \left[v_0 - \dfrac{\tau_*}{d_2} m C_0 (1 + \beta^{-1} - x_*) \right], \\ m = \displaystyle\int_0^1 f\big(w_0 e^{-A_0 x_* - A\tau_* \max\{C_0/2, C_0 y\}}\big) dy, \end{cases}$$

has a unique solution pair (x_*, τ_*) *satisfying* $x_* \in [0,1]$ *and* $\tau_* > 0$. *Moreover,*

$$u_n \to 0 \text{ in } C([0,1] \setminus [x_* - \epsilon, x_* + \epsilon]) \quad \forall\, \epsilon > 0, \qquad \int_0^1 u_n dx \to \tau_* C_0,$$

$$v_n(x) \to v_0 - \frac{\tau_*}{d_2} m C_0 (1 + \beta^{-1} - \max\{x, x_*\}) \quad \text{uniformly in } [0,1].$$

Furthermore, $x_* = 0$ *if* $v_0 = v^*(m)$, $x_* \in (0,1)$ *if* $v_*(m) < v_0 < v^*(m)$, *and* $x_* = 1$ *if* $v_0 = v_*(m)$.

Proof. Using the notation of the previous section, by passing to a subsequence, $x_n \to x_* \in [0,1]$. By possibly passing to a further subsequence, the behavior of (u_n, v_n) as $n \to \infty$ is then determined by Lemmas 2.2, 2.3 (if $x_* \in (0,1)$), Lemma 2.4 (if $x_* = 0$ and subcases (a1) and (a2) occur), and Lemma 2.5 (if $x_* = 1$ and subcases (b1) and (b2) happen).

If we can show that x_* and τ_* are uniquely determined by the value of v_0, then the corresponding results in the previous section would hold not only for a subsequence, but for the entire original sequence, and hence the behavior of (u_n, v_n) as $n \to \infty$ would be completely determined.

The rather long proof below is broken into several steps.

Step 1. Subcases (a2) and (b2) do not happen

First we observe that subcase (a2) does not happen. Indeed, if this case occurs, then since $C(a_*) < C_0/2$, we see (as explained below) from a careful comparison of (2.19) and (3.2) that

$$\tau_* > \tau_0^*, \quad \tau_* C(a_*) < \tau_0^* C_0/2, \quad \tau_*[C_0/2 + C(a_*)] > \tau_0^* C_0.$$

In the comparison, we can deduce these inequalities one at a time, in the above order, and the previous inequalities are used for obtaining the next inequality. For example, to deduce $\tau_* C(a_*) < \tau_0^* C_0/2$ from $\tau_* > \tau_0^*$, we observe that $\tau_* C(a_*) \geq \tau_0^* C_0/2$ and $\tau_* > \tau_0^*$ would imply

$$\begin{aligned}\tau_* \max\{C(a_*), [C_0/2 + C(a_*)]y\} &\geq \max\{\tau_0^* C_0/2, [\tau_* C_0/2 + \tau_0^* C_0/2]y\} \\ &\geq \tau_0^* \max\{C_0/2, C_0 y\}\end{aligned}$$

with strict inequality holding in the last step for $y \in [1/2, 1]$, which is impossible when one compares (2.19) with (3.2).

It then follows from (2.20) and (2.21) that $v_0 > v^*(m)$, contradicting (3.6).

Similarly, if subcase (b2) happens, then from (2.27) we deduce

$$\tau_* > \tau_1^* \quad \text{and} \quad \tau_*[C_0/2 + \tilde{C}(b_*)] < \tau_1^* C_0,$$

which imply, by (2.28) and (2.29), that $v_0 < v_*(m)$, again contradicting (3.6). Therefore subcase (b2) cannot happen.

Thus, by our discussion in the previous section, we have the cases where (2.8) and (2.9) hold. To show that (2.8) and (2.9) have a unique solution (x_*, τ_*) satisfying $x_* \in [0, 1]$ and $\tau_* > 0$, we establish a procedure to uniquely find x_* and τ_*. In the discussion below, we will treat $v_0 > 0$ as a varying parameter.

Step 2. A procedure to solve (2.8) and (2.9).

It is useful to use the new variable

$$\lambda = A_0 x_* + A\tau_* C_0/2.$$

Then

$$x_* = (\lambda - A\tau_* C_0/2)/A_0,$$

and (2.8) can be rewritten as

$$\frac{w_0}{\alpha} e^{-\lambda} = v_0 - \frac{\tau_*}{d_2} m C_0 \left(1 + \beta^{-1} - \frac{\lambda - A\tau_* C_0/2}{A_0}\right)$$

or

$$\frac{mC_0}{d_2 A_0} \tau_* \big[(1 + \beta^{-1})A_0 - \lambda + A(C_0/2)\tau_*\big] = v_0 - \frac{w_0}{\alpha} e^{-\lambda}.$$

We now consider the quadratic equation of τ:

$$\frac{mC_0}{d_2 A_0} \tau \big[(1 + \beta^{-1})A_0 - \lambda + A(C_0/2)\tau\big] = v_0 - \frac{w_0}{\alpha} e^{-\lambda}. \tag{3.7}$$

For each $v_0 > 0$, let $\lambda_0(v_0)$ denote the minimal nonnegative λ such that $v_0 - \frac{w_0}{\alpha} e^{-\lambda} \geq 0$. Clearly

$$\lambda_0(v_0) = 0 \text{ if } v_0 \geq w_0/\alpha, \ \lambda_0(v_0) \text{ is decreasing in } (0, w_0/\alpha], \ \lim_{v_0 \to 0} \lambda_0(v_0) = \infty. \tag{3.8}$$

For each $v_0 > 0$ and $\lambda \geq \lambda_0(v_0)$, the quadratic equation (3.7) has a maximal zero, which we denote by $\tau(\lambda, v_0)$. It is easily seen that $\tau(\lambda, v_0) \geq 0$ and

$$\text{when } v_0 \leq w_0/\alpha,\ \tau(\lambda_0(v_0), v_0) = \max\left\{0, \frac{\lambda_0(v_0) - A_0(1+\beta^{-1})}{AC_0/2}\right\}, \tag{3.9}$$

$$\tau(\lambda, v_0) \text{ is increasing in } \lambda \text{ and in } v_0, \quad \lim_{v_0\to\infty} \tau(\lambda, v_0) = \infty \text{ for fixed } \lambda \geq 0. \tag{3.10}$$

Since $\lambda_0(w_0/\alpha) = 0$, by (3.9), $\tau(\lambda_0(w_0/\alpha), w_0/\alpha) = 0$. Let us consider the continuous function

$$M(v_0) = \int_0^1 f\big(w_0 e^{-\lambda_0(v_0) - A\tau(\lambda_0(v_0), v_0)\max\{0, C_0 y - C_0/2\}}\big)dy.$$

The above observation shows that $M(w_0/\alpha) = f(w_0) > m$. By (3.8), we have $M(v_0) \to 0$ as $v_0 \to 0$. By (3.10), we deduce $M(v_0) \to 0$ as $v_0 \to \infty$. Hence from the continuity of $M(v_0)$ we can find v_{min} and v_{max} such that

$$0 < v_{min} < w_0/\alpha < v_{max} < \infty,$$

$$M(v_0) > m \quad \forall v_0 \in (v_{\min}, v_{max}), \qquad M(v_{min}) = M(v_{max}) = m.$$

Now for each $v_0 \in (v_{min}, v_{max})$,

$$m < \int_0^1 f\big(w_0 e^{-\lambda_0(v_0) - A\tau(\lambda_0(v_0), v_0)\max\{0, C_0 y - C_0/2\}}\big)dy.$$

This and the monotonicity of $\tau(\lambda, v_0)$ in λ imply that for such v_0 we can find a unique $\lambda_* = \lambda_*(v_0) > \lambda_0(v_0)$ such that

$$m = \int_0^1 f\big(w_0 e^{-\lambda_* - A\tau(\lambda_*, v_0)\max\{0, C_0 y - C_0/2\}}\big)dy.$$

Clearly $v_0 \to \lambda_*(v_0)$ is continuous and

$$\lambda_*(v_{min} + 0) = \lambda_0(v_{min}), \qquad \lambda_*(v_{max} - 0) = \lambda_0(v_{max}).$$

So we may define

$$\lambda_*(v_{min}) = \lambda_0(v_{min}), \qquad \lambda_*(v_{max}) = \lambda_0(v_{max}).$$

We claim that the function $T(v_0) := \tau(\lambda_*(v_0), v_0)$ is increasing in $[v_{min}, v_{max}]$. Otherwise, we can find $v_{min} \leq s_1 < s_2 \leq v_{max}$ such that $T(s_1) \geq T(s_2)$. Since

$$\int_0^1 f\big(w_0 e^{-\lambda_*(s_1) - AT(s_1)\max\{0, C_0 y - C_0/2\}}\big) = \int_0^1 f\big(w_0 e^{-\lambda_*(s_2) - AT(s_2)\max\{0, C_0 y - C_0/2\}}\big),$$

$T(s_1) \geq T(s_2)$ implies that $\lambda_*(s_1) \leq \lambda_*(s_2)$, which implies, by the monotonicity of $\tau(\lambda, v_0)$,

$$T(s_1) = \tau(\lambda_*(s_1), s_1) < \tau(\lambda_*(s_2), s_2) = T(s_2).$$

This contradiction proves the claimed monotonicity of $T(v_0)$.

We show next that $T(v_{max}) > \tau_0^*$. Since $v_{max} > w_0/\alpha$, we have $\lambda_0(v_{max}) = 0$ and hence

$$m = M(v_{max}) = \int_0^1 f\big(w_0 e^{-A\tau(0,v_{max})\max\{0,C_0y-C_0/2\}}\big)dy.$$

By (3.2),

$$m = \int_0^1 f\big(w_0 e^{-A\tau_0^* C_0/2 - A\tau_0^*\max\{0,C_0y-C_0/2\}}\big)dy.$$

Comparing the above two expressions, we obtain $\tau(0, v_{max}) > \tau_0^*$. Hence

$$T(v_{max}) = \tau(\lambda_*(v_{max}), v_{max}) = \tau(\lambda_0(v_{max}), v_{max}) = \tau(0, v_{max}) > \tau_0^*,$$

as we wanted.

We now consider $T(v_{min})$. We have two different cases: $m < f(w_0e^{-A_0})$ and $m \geq f(w_0e^{-A_0})$. First consider the case $m < f(w_0e^{-A_0})$. We show that $T(v_{min}) < \tau_1^*$ in this case. Since $\lambda_*(v_{min}) = \lambda_0(v_{min})$ we have

$$T(v_{min}) = \tau(\lambda_0(v_{min}), v_{min}).$$

Hence, by (3.9),

$$T(v_{min}) = \max\left\{0, \frac{\lambda_0(v_{min}) - A_0(1+\beta^{-1})}{AC_0/2}\right\}.$$

If $\frac{\lambda_0(v_{min})-A_0(1+\beta^{-1})}{AC_0/2} \leq 0$, then $T(v_{min}) = 0 < \tau_1^*$. If $\frac{\lambda_0(v_{min})-A_0(1+\beta^{-1})}{AC_0/2} > 0$, then

$$T(v_{min}) = \frac{\lambda_0(v_{min}) - A_0(1+\beta^{-1})}{AC_0/2},$$

and hence

$$\begin{aligned} m &= \int_0^1 f\big(w_0 e^{-\lambda_0(v_{min})-AT(v_{min})\max\{0,C_0y-C_0/2\}}\big)dy \\ &= \int_0^1 f\big(w_0 e^{-A_0(1+\beta^{-1})-AT(v_{min})C_0/2-AT(v_{min})\max\{0,C_0y-C_0/2\}}\big)dy \\ &= \int_0^1 f\big(w_0 e^{-A_0(1+\beta^{-1})-AT(v_{min})\max\{C_0/2,C_0y\}}\big)dy. \end{aligned}$$

Comparing this with (3.4), we find that $T(v_{min}) < \tau_1^*$.

With the above properties of $T(v_0)$, we can uniquely determine v_* and v^* with $v_{min} < v_* < v^* < v_{max}$ such that

$$T(v^*) = \tau_0^*, \qquad T(v_*) = \tau_1^*.$$

We claim that $v^* = v^*(m)$ and $v_* = v_*(m)$. Indeed, from

$$m = \int_0^1 f\big(w_0 e^{-\lambda_*(v^*)-AT(v^*)\max\{0,C_0y-C_0/2\}}\big)dy$$

and $T(v^*) = \tau_0^*$, we easily see by comparing with (3.2) that $\lambda_*(v^*) = \tau_0^* AC_0/2$. Hence

$$\tau_0^* = T(v^*) = \tau(\lambda_*(v^*), v^*) = \tau(\tau_0^* AC_0/2, v^*).$$

By the definition of $\tau(\lambda, v_0)$, the above identity means that $\tau = \tau_0^*$ solves (3.7) with $\lambda = \tau_0^* AC_0/2$ and $v_0 = v^*$. Therefore we may compare with (3.3) to deduce $v^* = v^*(m)$. Similarly, we can show that $v_* = v_*(m)$.

Since T is monotone, for each $v_0 \in [v_*(m), v^*(m)]$, $T(v_0) \in [\tau_1^*, \tau_0^*]$. Hence we can compare (3.2) and (3.4) with

$$m = \int_0^1 f\big(w_0 e^{-\lambda_*(v_0) - AT(v_0)\max\{0, C_0 y - C_0/2\}}\big) dy$$

to find that, for such v_0, we necessarily have

$$AT(v_0)C_0/2 \le \lambda_*(v_0) \le A_0 + AT(v_0)C_0/2;$$

otherwise we would arrive at contradictions to $T(v_0) \in [\tau_1^*, \tau_0^*]$. This implies that there exists a unique $x_* \in [0, 1]$ such that

$$\lambda_*(v_0) = A_0 x_* + AT(v_0)C_0/2.$$

Let $\tau_* = T(v_0)$; we find that (x_*, τ_*) solves (2.8) and (2.9).

We next consider the case $m \ge f(w_0 e^{-A_0})$. In this case, $v_*(m) = \underline{v}(m)$; moreover, we show that

$$T(v_{min}) = 0, \quad v_{min} = \underline{v}(m).$$

Indeed, from

$$T(v_{min}) = \tau(\lambda_*(v_{min}), v_{min}) = \tau(\lambda_0(v_{min}), v_{min}),$$

we obtain

$$\begin{aligned} m &= \int_0^1 f\big(w_0 e^{-\lambda_0(v_{min}) - A\tau(\lambda_0(v_{min}), v_{min})\max\{0, C_0 y - C_0/2\}}\big) dy \\ &\le f(w_0 e^{-\lambda_0(v_{min})}). \end{aligned}$$

It follows that $\lambda_0(v_{min}) \le A_0 < A_0(1+\beta^{-1})$. By (3.9), we deduce $\tau(\lambda_0(v_{min}), v_{min}) = 0$, that is, $T(v_{min}) = 0$. This gives

$$m = \int_0^1 f(w_0 e^{-\lambda_0(v_{min})}) dy = f(w_0 e^{\lambda_0(v_{min})}).$$

Hence

$$\alpha \underline{v}(m) = w_0 e^{-\lambda_0(v_{min})}.$$

On the other hand, since $v_{min} < w_0/\alpha$, by the definition of the function λ_0,

$$v_{min} - \frac{w_0}{\alpha} e^{-\lambda_0(v_{min})} = 0.$$

Therefore we have $v_{min} = \underline{v}(m)$.

We can now conclude that there exists a unique $v^* \in (v_{min}, v_{max})$ such that $T(v^*) = \tau_0^*$. We may then prove $v^* = v^*(m)$ as before. Since T is monotone, for each $v_0 \in (v_*(m), v^*(m)] = (\underline{v}(m), v^*(m)]$, $T(v_0) \in (0, \tau_0^*]$. Hence we can compare (3.2) and $m \geq f(w_0 e^{-A_0})$ with (3.1) to deduce

$$AT(v_0)C_0/2 \leq \lambda_*(v_0) < A_0 + AT(v_0)C_0/2,$$

and there exists a unique $x_* \in [0, 1)$ such that

$$\lambda_*(v_0) = A_0 x_* + AT(v_0)C_0/2.$$

Let $\tau_* = T(v_0)$; we find that (x_*, τ_*) solves (2.8) and (2.9).

The above discussion shows that when (3.6) holds, (2.8) and (2.9) have at least one solution (x_*, τ_*) satisfying $x_* \in [0, 1]$ and $\tau_* > 0$, and such a solution can be found by following the above procedure.

Step 3. Uniqueness of (x_*, τ_*) and completion of the proof.

We next show that when (3.6) holds, (2.8) and (2.9) have a unique solution (x_*, τ_*) satisfying $x_* \in [0, 1]$ and $\tau_* > 0$. So let (x_*, τ_*) be an arbitrary solution of (2.8) and (2.9) with $v_0 \in [v_*(m), v^*(m)] \cap (\underline{v}(m), v^*(m)]$ and $x_* \in [0, 1]$, $\tau_* > 0$. Then τ_* must be the maximal zero of (3.7) with $\lambda = A_0 x_* + A\tau_* C_0/2 > 0$; this is the case because $v_0 - \frac{w_0}{\alpha} e^{-\lambda} > 0$, and thus the two zeros of (3.7) are of opposite sign. Therefore, using our earlier notation,

$$\tau_* = \tau(\lambda, v_0), \quad \lambda > \lambda_0(v_0).$$

Then (2.9) yields

$$\begin{aligned} m &= \int_0^1 f(w_0 e^{-A_0 x_* - A\tau_* \max\{C_0/2, C_0 y\}}) dy \\ &= \int_0^1 f(w_0 e^{-\lambda - A\tau(\lambda, v_0) \max\{0, C_0 y - C_0/2\}}) dy. \end{aligned}$$

Since $v_0 \in [v_*(m), v^*(m)] \cap (\underline{v}(m), v^*(m)] \subset (v_{min}, v_{max})$, in view of the above identity, our definition of $\lambda_*(v_0)$ implies that $\lambda = \lambda_*(v_0)$ and hence $\tau(\lambda, v_0) = T(v_0)$; that is, $\tau_* = \tau(\lambda, v_0) = T(v_0)$. This implies that the solution pair (x_*, τ_*) is the same as the one obtained through our procedure introduced above for solving (2.8) and (2.9). Hence there is a unique solution.

With τ_* and x_* uniquely determined now, it is easily seen that our conclusions for u_n and v_n follow from Lemmas 2.2, 2.3, 2.4, and 2.5.

Moreover, from the above procedure for finding (x_*, τ_*), we easily see that $x_* = 0$ if $v_0 = v^*(m)$, $x_* \in (0, 1)$ if $v_*(m) < v_0 < v^*(m)$, and $x_* = 1$ if $v_0 = v_*(m)$.

The proof of the theorem is now complete. □

Next we consider the case that $v_0 > v^*(m)$. Let $0 < \lambda_0 < \lambda^0$ be uniquely determined by

$$m = f(w_0 e^{-A\lambda_0}) = \int_0^1 f\big(w_0 e^{-A\lambda^0 y}\big) dy. \tag{3.11}$$

For each $\lambda \in [0, \lambda_0]$, we can find a unique $\Gamma = \Gamma(\lambda)$ such that

$$m = \int_0^1 f\big(w_0 e^{-A \max\{\lambda, \Gamma y\}}\big) dy. \tag{3.12}$$

Moreover, it is easily seen that $\lambda \mapsto \Gamma(\lambda)$ is a continuous decreasing function with

$$\Gamma(\lambda_0) = \lambda_0, \qquad \Gamma(0) = \lambda^0.$$

Therefore we can find a unique $\lambda_*^0 \in (0, \lambda_0)$ such that

$$\Gamma(\lambda_*^0) = 2\lambda_*^0.$$

Comparing with (3.2), we find that actually

$$\lambda_*^0 = \tau_0^* C_0/2. \tag{3.13}$$

We define

$$\Lambda(\lambda) := \frac{w_0}{\alpha} e^{-A\lambda} + \frac{\Gamma(\lambda)}{d_2} m(1+\beta^{-1}).$$

Clearly $\Lambda(\lambda)$ is a decreasing function on $[0, \lambda_0]$ with

$$\Lambda(0) = \frac{w_0}{\alpha} + \frac{\lambda^0}{d_2} m(1+\beta^{-1}), \qquad \Lambda(\lambda_*^0) = \frac{w_0}{\alpha} e^{-A\lambda_*^0} + \frac{2\lambda_*^0}{d_2} m(1+\beta^{-1}).$$

Due to (3.13), we find that

$$\Lambda(\lambda_*^0) = v^*(m).$$

THEOREM 3.2. *Suppose that*

$$v_0 > v^*(m) = \Lambda(\lambda_*^0). \tag{3.14}$$

If $v_0 < \Lambda(0)$ *and* $\lambda^* \in (0, \lambda_*^0)$ *is uniquely determined by* $v_0 = \Lambda(\lambda^*)$*, then*

$$u_n \to 0 \text{ in } C([\epsilon, 1]) \quad \forall \epsilon \in (0,1), \qquad \int_0^1 u_n dx \to \Gamma(\lambda^*),$$

$$v_n(x) \to v_0 - \frac{\Gamma(\lambda^*)}{d_2} m(1+\beta^{-1}-x) \text{ uniformly in } [0,1].$$

If $v_0 \geq \Lambda(0)$*, then the above conclusions hold with* $\lambda^* = 0$.

Proof. We first show that case (a2) happens. Let us start by observing that the cases leading to (2.8) and (2.9) (namely, cases (i), (ii)(a1), and (iii)(b1)) cannot happen. Indeed, in these cases, (x_*, τ_*) solves (2.8) and (2.9) with $x_* \in [0,1]$ and $\tau_* > 0$. As in Step 3 of the proof of Theorem 3.1, denoting $\lambda = A_0 x_* + A\tau_* C_0/2$, we must have $\tau_* = \tau(\lambda, v_0)$ and $\lambda > \lambda_0(v_0)$. Then (2.9) gives

$$m = \int_0^1 f(w_0 e^{-\lambda - A\tau(\lambda, v_0)\max\{0, C_0 y - C_0/2\}}) dy.$$

Since $v_0 > v^*(m)$, we have either

$$v_0 > v_{max} \quad \text{or} \quad v_0 \in (v^*(m), v_{max}].$$

If $v_0 \in (v^*(m), v_{max}] \subset (v_{min}, v_{max})$, then the above identity implies that $\lambda = \lambda_*(v_0)$ and hence $\tau(\lambda, v_0) = T(v_0)$. From $v_0 > v^*(m)$ we now deduce $\tau_* = T(v_0) > \tau_0^*$, and hence we can compare (2.9) with (3.2) to deduce $x_* < 0$, a contradiction.

If $v_0 > v_{max}$, then by the monotonicity of $\tau(\cdot,\cdot)$, we deduce

$$\tau(\lambda, v_0) > \tau(\lambda, v_{max}) > \tau(0, v_{max}).$$

Therefore, recalling $\lambda_*(v_{max}) = \lambda_0(v_{\max}) = 0$, we obtain

$$\begin{aligned} m &= \int_0^1 f(w_0 e^{-\lambda - A\tau(\lambda,v_0)\max\{0, C_0 y - C_0/2\}})dy \\ &< \int_0^1 f(w_0 e^{-A\tau(0,v_{max})\max\{0, C_0 y - C_0/2\}})dy \\ &= m, \end{aligned}$$

again a contradiction. Therefore none of the cases that lead to (2.8) and (2.9) can happen. This implies that either (a2) or (b2) happens.

Next we show that case (b2) cannot happen. Otherwise, by (2.27) we obtain

$$m < f(w_0 e^{-A_0}).$$

Hence $\tau_1^* > 0$ is defined. Moreover, comparing (2.27) with (3.4), we obtain

$$\tau_* > \tau_1^*, \qquad \tau_* \tilde{C}(b_*) < \tau_1^* C_0,$$

which imply, by (2.28) and (2.29), that $v_0 < v_*(m) < v^*(m)$, contradicting (3.14).

Therefore we necessarily have case (a2). We now introduce the notation

$$\lambda = \tau_*[C(a_*) - C_0/2], \quad \Gamma = \tau_* C(a_*).$$

From (2.19), (2.20), and (2.21) we find that

$$m = \int_0^1 f\big(w_0 e^{-A\max\{\lambda, \Gamma y\}}\big)dy, \tag{3.15}$$

$$v_0 \geq \frac{w_0}{\alpha} e^{-A\lambda} + \frac{\Gamma}{d_2} m(1 + \beta^{-1}), \tag{3.16}$$

with equality holding if $a_* > 0$.

Suppose now that $v_0 \geq \Lambda(0)$. We claim that in this case we have $\lambda = 0$ and hence, by (3.15), $\Gamma = \Gamma(0) = \lambda^0$. Suppose for the sake of contradiction that $\lambda > 0$. From (3.15) and (3.11) we easily see that $\lambda \leq \lambda_0$. Now $C(a_*) - C_0/2 > 0$, and hence $a_* > 0$. Thus equality in (3.16) holds. By (3.15) we deduce $\Gamma = \Gamma(\lambda)$, and hence it follows from (3.16) that $v_0 = \Lambda(\lambda) < \Lambda(0)$, contradicting our assumption on v_0 above. Hence in this case, we have $\lambda = 0$ and thus

$$C(a_*) - C_0/2 = 0, \qquad \tau_* = \Gamma(0)/(C_0/2).$$

Next we suppose that $v^*(m) < v_0 < \Lambda(0)$. From (3.15) we deduce $\Gamma = \Gamma(\lambda)$ for some $\lambda \in [0, \lambda_0]$. We must have $\lambda > 0$ for otherwise, from (3.15) and (3.16), we deduce $\Gamma = \Gamma(0)$ and $v_0 \geq \Lambda(0)$, contradicting our current assumption on v_0. Therefore $\lambda > 0$ and hence $a_* > 0$, implying that equality in (3.16) holds. Recalling $\Gamma = \Gamma(\lambda)$, we thus obtain $v_0 = \Lambda(\lambda)$ and $\lambda = \lambda^*$. It follows that τ_* and a_* in Lemma 2.4 are uniquely determined by

$$\tau_*[C(a_*) - C_0/2] = \lambda^*, \qquad \tau_* C(a_*) = \Gamma(\lambda^*),$$

namely,

$$\tau_* = \frac{\Gamma(\lambda^*) - \lambda^*}{C_0/2}, \qquad a_* = C^{-1}(\lambda^*/\tau_* + C_0/2).$$

The rest of the proof now follows from Lemma 2.4. ∎

We now consider the remaining case that $\underline{v}(m) < v_0 < v_*(m)$, which can happen only if $m < f(w_0 e^{-A_0})$. Suppose that λ_0, λ^0, λ^0_*, and $\Gamma(\lambda)$ are as in Theorem 3.2 but with w_0 there replaced by $w_0 e^{-A_0}$, and we denote them by $\tilde{\lambda}_0$, $\tilde{\lambda}^0$, $\tilde{\lambda}^0_*$, and $\tilde{\Gamma}(\lambda)$, respectively. Define

$$\Delta(\lambda) := \frac{w_0}{\alpha} e^{-A_0 - A\lambda} + \frac{\tilde{\Gamma}(\lambda)}{d_2} m\beta^{-1}.$$

Then $\Delta(\lambda)$ is a decreasing function over $[0, \tilde{\lambda}_0]$ with

$$\Delta(0) = \frac{w_0}{\alpha} e^{-A_0} + \frac{\tilde{\lambda}^0}{d_2} m\beta^{-1}, \qquad \Delta(\tilde{\lambda}^0_*) = \frac{w_0}{\alpha} e^{-A_0 - A\tilde{\lambda}^0_*} + \frac{\tilde{\lambda}^0_*}{d_2} m\beta^{-1} = v_*(m).$$

THEOREM 3.3. *Suppose that* $m < f(w_0 e^{-A_0})$ *and*

$$\underline{v}(m) < v_0 < v_*(m) = \Delta(\tilde{\lambda}^0_*). \tag{3.17}$$

If $v_0 > \Delta(0)$ *and* $\lambda_* \in (0, \tilde{\lambda}^0_*)$ *is uniquely determined by* $v_0 = \Delta(\lambda_*)$, *then*

$$u_n \to 0 \text{ in } C([0, 1-\epsilon])\ \forall \epsilon \in (0,1), \quad \int_0^1 u_n dx \to \tilde{\Gamma}(\lambda_*),$$

$$v_n(x) \to v_0 - \frac{\tilde{\Gamma}(\lambda_*)}{d_2} m\beta^{-1} \text{ uniformly in } [0,1].$$

If $\underline{v}(m) < v_0 \le \Delta(0)$, *then the above conclusions hold with* $\lambda_* = 0$.

Proof. This is similar to the proof of Theorem 3.2. Here we can show that case (b2) must happen, and then we use Lemma 2.5. We omit the details. ∎

Remark 3.4. Our results in this section reveal an interesting property for the limiting total biomass $\lim_{n\to\infty} \int_0^1 u_n(x)dx$. First consider the case $v_*(m) \le v_0 \le v^*(m)$. By Theorem 3.1, in this case the above limit has value $\tau_* C_0$. By the proof of Theorem 3.1, we know that $\tau_* = T(v_0)$ with $T(v_0)$ a strictly increasing function of v_0. Therefore the limit is strictly increasing with v_0 for $v_0 \in [v_*(m), v^*(m)]$.

If $v_0 \in (v^*(m), \Lambda(0))$, then by Theorem 3.2, $\lim_{n\to\infty} \int_0^1 u_n(x)dx = \Gamma(\lambda^*) = \Gamma \circ \Lambda^{-1}(v_0)$. Since $\Gamma(\cdot)$ and $\Lambda(\cdot)$ are both strictly decreasing functions, $\Gamma \circ \Lambda^{-1}(\cdot)$ is strictly increasing, and hence the limit is strictly increasing with v_0 for $v_0 \in (v^*(m), \Lambda(0))$. However, by Theorem 3.2, this limit takes the same value $\Gamma(0)$ for all $v_0 \ge \Lambda(0)$.

If $v_0 \in (\Delta(0), v_*(m))$, then by Theorem 3.3, the limit takes value $\tilde{\Gamma}(\lambda_*) = \tilde{\Gamma} \circ \Delta^{-1}(v_0)$, which is strictly increasing in v_0, but it takes the same value $\tilde{\Gamma}(0)$ for all $v_0 \in (\underline{v}(m), \Delta(0)]$.

Therefore if we denote

$$v_{**} = v_{**}(m) := \Delta(0), \qquad v^{**} = v^{**}(m) := \Lambda(0),$$

then $\lim_{n\to\infty} \int_0^1 u_n(x)dx$ is strictly increasing with v_0 as v_0 varies in the range $v_{**} \le v_0 \le v^{**}$, but is constant for $v_0 \ge v^{**}$ or for $v_0 \in (\underline{v}(m), v_{**}]$.

4. Biological interpretations. In this section we compare our results with the game theoretical model in [KL] and explain the predictions that our theoretical results offer for the phytoplankton problem being modelled.

Since $u(x)$ and $v(x)$ are, respectively, rescaled versions of the biomass function $b(x)$ and the nutrient function $R(x)$ used in the original model of [KL] (see Part I for details), we will interpret $u(x)$ and $v(x)$ as representing the (steady) distributions of the biomass and nutrient at depth x of a water column with surface at $x = 0$ and bottom at $x = 1$.

(i) First we note that, if we replace $\max\{C_0/2, C_0 y\}$ in (2.9) by C_0, then the system of equations for (x_*, τ_*) in Theorem 3.1 reduces to the game theoretical model of [KL], namely, equations (4) and (5) in [KL] with $\hat{B} = \tau_* C_0$. Thus the game theoretical model of [KL] is a simplified version of the rigorous limiting equations here. It captures the main properties of the limiting equations but only in the case that $v_* \le v_0 \le v^*$.

(ii) When $v_0 > v^* = v^*(m)$, from Theorem 3.2 and Step 1 of the proof of Theorem 3.1, we see that as $\sigma \to \infty$ the total biomass has limit

$$\Gamma(\lambda^*) = \tau_* C(a_*) > \tau_0^* C_0.$$

If we have simply used (2.8) and (2.9) with $x_* = 0$ to calculate the total biomass, we would have obtained the incorrect limit $\tau_0^* C_0$. Similarly, the limit of the total biomass in the case of Theorem 3.3 is less than the value one would have obtained by simply using (2.8) and (2.9) with $x_* = 1$.

(iii) Remark 3.4 shows that as $\sigma \to \infty$, the limit of the total biomass is strictly increasing with v_0 in the range $v_{**} \le v_0 \le v^{**}$. It is interesting to notice that the layer position of the biomass (in the limit) already reaches the surface at $v_0 = v^*$ (i.e., $x_* = 0$ when $v_0 = v_*$), but as the nutrient level at the sediment v_0 increases past the critical value v^*, though the layer remains at the surface, the total biomass keeps increasing until v_0 reaches a second critical value v^{**}, where the total biomass reaches a maximum, and then remains at this value even if v_0 is further increased. On the other hand, if v_0 is decreased to v_*, the layer of the biomass lowers to the bottom ($x_* = 1$), but as v_0 decreases past v_*, though the layer remains at the bottom, the total biomass keeps decreasing until v_0 reaches the critical value v_{**}, where the total biomass reaches its minimal value, and then remains at this minimal value until v_0 is so low ($v_0 \le \underline{v}(m)$) that the phytoplankton can no longer survive.

(iv) Our results support the important predictions in [KL] that depth-regulating phytoplankton can form a thin layer in a poorly mixed water column and that the concentration of the limiting nutrient should be low and constant above the phytoplankton layer and linearly increasing with depth below the layer. The predictions in item (iii) above seem to provide new insights to this problem.

(v) We could fix v_0 and use a different parameter in the model, say the surface light level w_0, as a varying parameter to interpret the phenomena represented in items (i)–(iii) above.

Acknowledgment. Y. Du thanks NCTS for the hospitality during his visit, when a major part of this paper was written.

REFERENCES

[DH] Y. Du and S.-B. Hsu, *Concentration phenomena in a nonlocal quasi-linear problem modelling phytoplankton* I: *Existence*, SIAM J. Math. Anal., 40 (2008), pp. 1419–1440.

[GT] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.

[KL] C. A. Klausmeier and E. Litchman, *Algal games: The vertical distribution of phytoplankton in poorly mixed water columns*, Limnol. Oceanogr., 46 (2001), pp. 1998–2007.

SIAM J. MATH. ANAL.
Vol. 40, No. 4, pp. 1471–1498

ON p-HARMONIC MAP HEAT FLOWS FOR $1 \le p < \infty$ AND THEIR FINITE ELEMENT APPROXIMATIONS*

JOHN W. BARRETT†, XIAOBING FENG‡, AND ANDREAS PROHL§

Abstract. Motivated by emerging applications from imaging processing, this paper studies the heat flow of a generalized p-harmonic map into spheres for the whole spectrum, $1 \le p < \infty$, in a unified framework. The existence of global weak solutions is established for the flow using the energy method together with a regularization and a penalization technique. In particular, a BV-solution concept is introduced and the existence of such a solution is proved for the 1-harmonic map heat flow. The main idea used to develop such a theory is to exploit the properties of measures of the forms $\mathcal{A} \cdot \nabla \mathbf{v}$ and $\mathcal{A} \wedge \nabla \mathbf{v}$, which pair a divergence-L^1, or a divergence-measure, tensor field $\mathcal{A}$ and a BV-vector field $\mathbf{v}$. Based on these analytical results, a practical fully discrete finite element method is then proposed for approximating weak solutions of the p-harmonic map heat flow, and the convergence of the proposed numerical method is also established.

Key words. p-harmonic maps, heat flow, penalization, energy method, color image denoising, finite element method

AMS subject classifications. 35K65, 58E20, 35Q80, 65M12, 65M60

DOI. 10.1137/070680825

1. Introduction.

Let $\Omega \subset \mathbf{R}^m$ be a bounded domain with smooth boundary $\partial\Omega$, and let S^{n-1} denote the unit sphere in $\mathbf{R}^n$. A map $\mathbf{u} \in C^1(\Omega, S^{n-1})$ is called a *p-harmonic map* if it is a critical point of the following p-energy:

$$E_p(\mathbf{v}) := \frac{1}{p} \int_\Omega |\nabla \mathbf{v}|^p \, \mathrm{d}x, \qquad 1 \le p < \infty. \tag{1.1}$$

It is well known [12, 21, 40] that the Euler–Lagrange equation of the p-energy is

$$-\Delta_p \mathbf{u} = |\nabla \mathbf{u}|^p \mathbf{u}, \tag{1.2}$$

where

$$\Delta_p \mathbf{u} := \mathrm{div}(|\nabla \mathbf{u}|^{p-2} \nabla \mathbf{u}). \tag{1.3}$$

Note that Δ_p is often called the p-Laplacian. It is easy to see that (1.2) is a *degenerate elliptic* equation for $p > 2$ and a *singular elliptic* equation for $1 \le p < 2$. These degeneracy and singular characteristics both disappear when $p = 2$.

We call a map $\mathbf{u} \in W^{1,p}(\Omega, S^{n-1})$ a *weakly p-harmonic map* if $\mathbf{u}$ satisfies (1.2) in the distributional sense. Here $W^{1,p}(\Omega, S^{n-1})$ denotes the Sobolev space

$$W^{1,p}(\Omega, S^{n-1}) := \left\{ \mathbf{u} \in W^{1,p}(\Omega, \mathbf{R}^n);\ \mathbf{u}(x) \in S^{n-1} \text{ for a.e. } x \in \Omega \right\}.$$

*Received by the editors January 23, 2007; accepted for publication (in revised form) June 10, 2008; published electronically November 7, 2008.
http://www.siam.org/journals/sima/40-4/68082.html
†Department of Mathematics, Imperial College London, London, SW7 2AZ, UK (jwb@ic.ac.uk).
‡Department of Mathematics, The University of Tennessee, Knoxville, TN 37996 (xfeng@math.utk.edu). The work of this author was partially supported by NSF grant DMS-0410266.
§Mathematisches Institut, Universität Tübingen, Auf der Morgenstelle 10, D-72076 Tübingen, Germany (prohl@na.uni-tuebingen.de)

One well-known method for looking for a weakly p-harmonic map is the homotopy method (or the gradient descent method), which then leads to considering the following gradient flow (or heat flow) for the p-energy functional E_p complemented with some given boundary and initial conditions:

$$\mathbf{u}_t - \Delta_p \mathbf{u} = |\nabla \mathbf{u}|^p \mathbf{u} \qquad \text{in } \Omega_T := \Omega \times (0,T), \tag{1.4}$$

$$|\mathbf{u}| = 1 \qquad \text{in } \Omega_T. \tag{1.5}$$

Clearly, (1.4) is a *degenerate parabolic* equation for $p > 2$ and a *singular parabolic* equation for $1 \le p < 2$. Again, these degeneracy and singular characteristics both disappear when $p = 2$.

We remark that p-harmonic maps and weakly p-harmonic maps between two Riemannian manifolds (M, g) and (N, h) and their heat flows can be defined in the same fashion (cf. [21, 43, 44, 46, 47]). In this paper, we shall only consider the case $M = \Omega$ and $N = S^{n-1}$, which is sufficient for the applications that we are interested in.

The p-harmonic map and its heat flow, in particular, the harmonic map and the harmonic map heat flow ($p = 2$), have been extensively studied in the past twenty years for $1 < p < \infty$ (cf. [9, 11, 12, 13, 16, 17, 18, 21, 22, 25, 30, 31, 32, 34, 38, 39, 40, 41, 46, 48] for $1 < p < \infty$; [26, 27] for $p = 1$). In the case when the target manifold is a sphere, the existence of a global weak solution for the harmonic flow was first proved by Chen in [11] using a penalization technique. The result and the penalization technique were extended to the p-harmonic flow for $p > 2$ by Chen, Hong, and Hungerbühler in [12]. The p-harmonic flow for $1 < p < 2$ was solved by Misawa in [41] using a time discretization technique (the method of Rothe) proposed in [31], and by Liu in [39] using a penalization technique similar to that of [12]. The p-harmonic flow ($1 < p < \infty$) from a unit ball in $\mathbf{R}^m$ into $S^1 \subset \mathbf{R}^2$ was studied by Courilleau and Demengel in [16]. In the case of general target manifolds the p-harmonic flow ($1 < p < \infty$) with small initial data was treated by Fardoun and Regbaoui in [22], and the conformal case of the p-harmonic flow was considered by Hungerbühler in [33]. Nonuniqueness of the p-harmonic flow was addressed in [15, 32]. Recently, Giga, Kashima, and Yamazaki [28] proved existence of strong local solutions for 1-harmonic map heat flow using nonlinear semigroup theory. Besides the great amount of mathematical interests in the p-harmonic map and the p-harmonic flow, research on these problems has been strongly motivated by applications of the harmonic map and its heat flow in liquid crystals and micromagnetism; we refer to [2, 5, 7, 8, 19, 29, 36, 37, 51] and the references therein for detailed exposition in this direction.

Another reason, which is the main motivation of this paper, for studying the p-harmonic map and its heat flow, in particular, for $1 \le p < 2$, arises from their emerging and intriguing applications to image processing for denoising color images (cf. [49, 50]). We recall that a color image is often expressed by the RGB color system, in which a vector $\mathbf{I}(x) = (r(x), g(x), b(x))$ for each pixel $x = (x_1, x_2)$ is used to represent the intensity of the three primary colors (red, green, and blue). The chromaticity and brightness of a color image are deduced from the RGB system by decomposing $\mathbf{I}(x)$ into two components,

$$\eta(x) := |\,\mathbf{I}(x)\,| \qquad \text{(brightness)},$$

$$\mathbf{g} := \frac{\mathbf{I}(x)}{|\,\mathbf{I}(x)\,|} \qquad \text{(chromaticity)},$$

where $|\mathbf{I}(x)|$ stands for the Euclidean norm of $\mathbf{I}(x)$. By definition, the chromaticity must lie on the unit sphere S^2. One of the main benefits of the chromaticity and brightness decomposition is that it allows one to denoise η and $\mathbf{g}$ separately using different methods. For example, one may denoise η by the well-known total variation (TV) model of Rudin, Osher, and Fatemi [42] (also see [24]), but denoise $\mathbf{g}$ by another model. One such model is to define the recovered chromaticity $\mathbf{u}$ as a (generalized) p-harmonic map [49, 50]

$$\mathbf{u} = \operatorname*{argmin}_{\mathbf{v}\in W^{1,p}(\Omega;S^2)} J_{p,\lambda}(\mathbf{v}) \qquad \text{for } p \geq 1, \tag{1.6}$$

where

$$J_{p,\lambda}(\mathbf{v}) := E_p(\mathbf{v}) + \frac{\lambda}{2}\int_\Omega |\mathbf{v}-\mathbf{g}|^2\,dx \qquad \text{for } \lambda > 0. \tag{1.7}$$

In particular, the cases $1 \leq p < 2$ are the most important and interesting since the recovered images keep geometric information such as edges and corners of the noisy color images. We shall call (1.6) the *p-harmonic model* for color image denoising. We also remark that the second term on the right-hand side of (1.7) is often called a fidelity term. As in the TV model [42, 24], the parameter λ controls the trade-off between goodness of fit-to-the-data and variability in $\mathbf{u}$.

Again, to find a solution for the p-harmonic map model (1.6), we consider the gradient flow (heat flow) for the energy functional $J_{p,\lambda}$, which is given by

$$\mathbf{u}_t - \Delta_p\mathbf{u} + \lambda(\mathbf{u}-\mathbf{g}) = \mu_{p,\lambda}\mathbf{u} \qquad \text{in } \Omega_T, \tag{1.8}$$

$$|\mathbf{u}| = 1 \qquad \text{in } \Omega_T, \tag{1.9}$$

$$\mathcal{B}_p\mathbf{n} = 0 \qquad \text{on } \partial\Omega_T := \partial\Omega\times(0,T), \tag{1.10}$$

$$\mathbf{u} = \mathbf{u}_0 \qquad \text{on } \Omega\times\{t=0\}, \tag{1.11}$$

where

$$\mathcal{B}_p := |\nabla\mathbf{u}|^{p-2}\nabla\mathbf{u}, \qquad \mu_{p,\lambda} := |\nabla\mathbf{u}|^p + \lambda\,(1-\mathbf{u}\cdot\mathbf{g}). \tag{1.12}$$

The goals of this paper are twofold. First, we shall present a general theory of weak solutions for the parabolic system (1.8)–(1.11) for the whole spectrum $1 \leq p < \infty$. To the best of our knowledge, there is no theory known in the literature for the 1-harmonic map and its heat flow, which on the other hand is the most important (and most difficult) case for the color image denoising application. So our theory and results fill this void. Furthermore, our theory handles the p-harmonic map heat flow (1.8)–(1.11) for all $1 \leq p < \infty$ in a unified fashion, rather than treating the system separately for different values of p and using different methods (cf. [11, 12, 13, 16, 18, 21, 22, 30, 31, 34, 41, 46, 48]). Second, based on the above theoretical work, we also develop and analyze a practical fully discrete finite element method for approximating the solutions of the p-harmonic map heat flow (1.8)–(1.11).

We now highlight the main ideas and key steps of our approach. Notice that there are two nonlinear terms in the p-harmonic flow: the p-Laplace term and the right-hand side due to the nonconvex constraint $|\mathbf{u}| = 1$, so the main difficulties are how to handle these two terms and how to pass to the limit in these two terms when a

compactness argument is employed. To handle the degeneracy of the p-Laplace term, we approximate the p-energy $E_p(\mathbf{v})$ by the following regularized energy

$$E_p^\varepsilon(\mathbf{v}) := \frac{b_p(\varepsilon)}{2}\int_\Omega |\nabla \mathbf{v}|^2\,dx + \frac{1}{p}\int_\Omega |\nabla \mathbf{v}|_\varepsilon^p\,dx \tag{1.13}$$

$$= \int_\Omega \Big\{ \frac{b_p(\varepsilon)}{2}|\nabla \mathbf{v}|^2 + \frac{1}{p}\big[\,|\nabla \mathbf{v}|^2 + a_p(\varepsilon)^2\,\big]^{\frac{p}{2}}\Big\} dx,$$

where $\varepsilon > 0$ and

$$a_p(\varepsilon) := \begin{cases} 0 & \text{if } 2 \le p < \infty, \\ \varepsilon & \text{if } 1 \le p < 2, \end{cases} \qquad b_p(\varepsilon) := \varepsilon^\alpha \quad \text{for } 1 \le p < \infty \tag{1.14}$$

for some $\alpha > 0$. Here and in the rest of this paper we adopt the shorthand notation

$$|\nabla \mathbf{v}|_\varepsilon := \sqrt{|\nabla \mathbf{v}|^2 + a_p(\varepsilon)^2}. \tag{1.15}$$

To handle the nonconvex constraint $|\mathbf{u}| = 1$, we approximate it by the well-known Ginzburg–Landau penalization [7], that is, we abandon the exact constraint, but enforce it approximately by adding a penalization term to the regularized p-energy E_p^ε. To this end, we replace the energy E_p^ε by

$$E_p^{\varepsilon,\delta}(\mathbf{v}) := E_p^\varepsilon(\mathbf{v}) + L^\delta(\mathbf{v}) \qquad \text{for } \varepsilon, \delta > 0, \tag{1.16}$$

where

$$L^\delta(\mathbf{v}) := \frac{1}{\delta}\int_\Omega F(\mathbf{v})\,dx, \qquad F(\mathbf{v}) := \frac{1}{4}\big(|\mathbf{v}|^2 - 1\big)^2 \qquad \forall \delta > 0,\ \mathbf{v} \in \mathbf{R}^n. \tag{1.17}$$

So the idea is, as δ gets smaller and smaller, the energy functional $E_p^{\varepsilon,\delta}$ becomes more and more favorable for maps $\mathbf{u}$ which take values close to the unit sphere S^{n-1}.

Consequently, the regularized model for the p-harmonic model (1.6) (with general m and n) reads

$$\mathbf{u}^{\varepsilon,\delta} = \operatorname*{argmin}_{\mathbf{v} \in W^{1,p}(\Omega;\mathbf{R}^n)} J_{p,\lambda}^{\varepsilon,\delta}(\mathbf{v}) \qquad \text{for } p \ge 1, \tag{1.18}$$

where

$$J_{p,\lambda}^{\varepsilon,\delta}(\mathbf{v}) := E_p^{\varepsilon,\delta}(\mathbf{v}) + \frac{\lambda}{2}\int_\Omega |\mathbf{v} - \mathbf{g}|^2\,dx. \tag{1.19}$$

In addition, the gradient flow for the regularized energy functional $J_{p,\lambda}^{\varepsilon,\delta}$ is given by

$$\mathbf{u}_t^{\varepsilon,\delta} - \Delta_p^\varepsilon \mathbf{u}^{\varepsilon,\delta} + \frac{1}{\delta}\big(|\mathbf{u}^{\varepsilon,\delta}|^2 - 1\big)\mathbf{u}^{\varepsilon,\delta} + \lambda(\mathbf{u}^{\varepsilon,\delta} - \mathbf{g}) = 0 \qquad \text{in } \Omega_T, \tag{1.20}$$

$$\mathcal{B}_p^{\varepsilon,\delta}\mathbf{n} = 0 \qquad \text{on } \partial\Omega_T, \tag{1.21}$$

$$\mathbf{u}^{\varepsilon,\delta} = \mathbf{u}_0 \qquad \text{on } \Omega \times \{t = 0\}, \tag{1.22}$$

which is an approximation to the original flow (1.8)–(1.11), where

$$\mathcal{B}_p^{\varepsilon,\delta} := b_p(\varepsilon)\nabla \mathbf{u}^{\varepsilon,\delta} + |\nabla \mathbf{u}^{\varepsilon,\delta}|_\varepsilon^{p-2}\nabla \mathbf{u}^{\varepsilon,\delta}, \qquad \Delta_p^\varepsilon \mathbf{u}^{\varepsilon,\delta} := \operatorname{div} \mathcal{B}_p^{\varepsilon,\delta}. \tag{1.23}$$

After having introduced the regularized flow (1.20)–(1.22), the next step is to analyze this regularized flow, in particular, to derive uniform (in ε and δ) a priori estimates. Finally, we pass to the limit in (1.20)–(1.22), first letting $\delta \to 0$ and then setting $\varepsilon \to 0$. As pointed out earlier, the main difficulty here is passing to the limit in two nonlinear terms on the left-hand side of (1.20). For $1 < p < \infty$, this will be done using a compactness technique and exploiting the symmetries of the unit sphere S^{n-1}, as done in [11, 12, 13, 21, 32, 41, 46]. However, for $p = 1$, since $L^1(\Omega)$ is not a reflexive Banach space, instead of working in the Sobolev space $W^{1,1}(\Omega)$, we are forced to work in $BV(\Omega)$, the space of functions of bounded variation, since solutions of the original 1-harmonic map heat flow (1.8)–(1.11) belong only to $L^\infty((0,T);[BV(\Omega)]^n)$ in general. This lack of regularity makes the analysis for $p = 1$ much more delicate than that for $1 < p < \infty$.

We note also that the regularized flow (1.20)–(1.22) not only plays an important role for proving existence of weak solutions for the flow (1.8)–(1.11), but also provides a practical and convenient formulation for approximating the solutions. The second goal of this paper is to develop a practical fully discrete finite element method for approximating solutions of the regularized flow (1.20)–(1.22); and hence, approximating solutions of the original p-harmonic flow (1.8)–(1.11) via the regularized flow. It is well known that explicitly enforcing the nonconvex constraint $|\mathbf{u}| = 1$ at the discrete level is hard to achieve. The penalization used in (1.20) allows one to get around this numerical difficulty at the expense of introducing an additional scale, δ. As expected, the numerical difficulty now is to control the dependence of the regularized solutions on δ and to establish scaling laws which relate the numerical scales (spatial and temporal mesh sizes) to the penalization scale δ for both stability and accuracy concerns. We refer to [6] and the references therein for more discussions in this direction and discussions on a related problem arising from liquid crystal applications. Borrowing a terminology from the phase transition of materials science, the regularized flow (1.20)–(1.22) may be regarded as a "diffuse interface" model for the original "sharp interface" model (1.8)–(1.11), and the "diffuse interface" is represented by the region $\{|\mathbf{u}^{\varepsilon,\delta}| < 1 - \delta\}$.

The remaining part of the paper is organized as follows. In section 2 we collect some known results and facts, which will be used in the later sections. In section 3 we present a complete analysis for the regularized flow (1.20)–(1.22), which includes proving its well-posedness, an energy law, a maximum principle, and uniform (in ε and δ) a priori estimates. As expected, these uniform a priori estimates serve as the basis for carrying out the energy method and the compactness arguments in the later sections. In section 4 we pass to the limit in (1.20)–(1.22) as $\delta \to 0$ for each *fixed* $\varepsilon > 0$. As in [13, 12, 32, 41], the main idea is to exploit the monotonicity of the operator Δ_p^ε and the symmetries of the unit sphere S^{n-1}. Sections 5 and 6 are devoted to passing to the limit as $\varepsilon \to 0$ in the ε-dependent limiting system obtained in section 4. For $1 < p < \infty$, this will be done by following the idea of section 4. However, for $p = 1$ the analysis becomes much more delicate because the nonreflexivity of $W^{1,1}(\Omega)$ forces us to work in the $BV(\Omega)$ space. The main idea used to develop a BV-solution concept is to exploit the properties of measures of the forms $\mathcal{A} \cdot \nabla \mathbf{v}$ and $\mathcal{A} \wedge \nabla \mathbf{v}$, which pair a divergence-L^1, or a divergence-measure, tensor field $\mathcal{A}$ and a BV-vector field $\mathbf{v}$. Finally, based on the theoretical results of sections 3–6, in section 7 we propose and analyze a practical fully discrete finite element method for approximating solutions of the p-harmonic flow (1.8)–(1.11) via the regularized flow (1.20)–(1.22). It is proved that the proposed numerical scheme

satisfies a discrete energy inequality, which mimics the differential energy inequality, and this leads to uniform (in ε and δ) a priori estimates and the convergence of the numerical approximations to the solutions of the flow (1.8)–(1.11) as the spatial and temporal mesh sizes, and the parameters ε and δ, all tend to zero.

2. Preliminaries. The standard notation for spaces is adopted in this paper (cf. [1, 3]). For example, $W^{k,p}(\Omega)$, $k \geq 0$ integer and $1 \leq p < \infty$, denotes the Sobolev spaces over the domain Ω and $\|\cdot\|_{W^{k,p}}$ denotes its norm. $W^{0,p}(\Omega) = L^p(\Omega)$ and $W^{k,2}(\Omega) = H^k(\Omega)$ are also used. $L^q((0,T); W^{k,p}(\Omega, \mathbf{R}^n))$ denotes the space of vector-valued functions (or maps), whose $W^{k,p}(\Omega)$-norm is L^q-integrable as a function of t over the interval $(0,T)$, and $\|\cdot\|_{L^q(W^{k,p})} := (\int_0^T \|\cdot\|^q_{W^{k,p}} dt)^{\frac{1}{q}}$ for $q \in [1,\infty)$ denotes its norm, with the standard modification for $q = \infty$. We use $\langle \cdot, \cdot \rangle$ to denote a generic dual product between elements of a Banach space X and its dual space X'.

In addition, $\mathcal{M}(\Omega)$ (resp., $[\mathcal{M}(\Omega)]^n$) denotes the space of real-valued (resp., $\mathbf{R}^n$-valued) finite Radon measures on Ω. Recall that $\mathcal{M}(\Omega)$ is the dual space of $C_0(\Omega)$ (cf. [3]). For a positive non-Lebesgue measure $\mu \in \mathcal{M}(\Omega)$, $L^p(\Omega, \mu)$ is used to denote the space of L^p-integrable functions with respect to the measure μ, and for $f \in L^1(\Omega, \mu)$, the measure $f\mu$ is defined by

$$f\mu(A) := \int_A f \, d\mu \qquad \text{for any Borel set } A \subset \Omega.$$

For any $\mu \in \mathcal{M}(\Omega)$, $\mu = \mu^a + \mu^s$ denotes its Radon–Nikodým decomposition, where μ^a and μ^s, respectively, denotes the absolute continuous part and the singular part of μ with respect to the Lebesgue measure $\mathcal{L}^n$.

Furthermore, $BV(\Omega)$ is used to denote the space of functions of bounded variation. Recall that a function $u \in L^1(\Omega)$ is called a function of *bounded variation* if all of its first order partial derivatives (in the distributional sense) are measures with finite total variations in Ω. Hence, the gradient of such a function u, denoted by Du, is a bounded vector-valued measure, with the finite total variation

$$|Du| \equiv |Du|(\Omega) := \sup\Big\{ \int_\Omega -u \operatorname{div} \mathbf{v} \, dx \,;\ \mathbf{v} \in [C_0^1(\Omega)]^n,\ \|\mathbf{v}\|_{L^\infty} \leq 1 \Big\}. \tag{2.1}$$

$BV(\Omega)$ is known to be a Banach space endowed with the norm

$$\|u\|_{BV} := \|u\|_{L^1} + |Du|. \tag{2.2}$$

For any $u \in BV(\Omega)$, $(Du)^a$ and $|Du|^a$ are used to denote, respectively, the absolute continuous part of Du and $|Du|$ with respect to the Lebesgue measure $\mathcal{L}^n$. We refer to [3] for detailed discussions about the space $BV(\Omega)$ and properties of BV functions.

For any vector $\mathbf{a} \in \mathbf{R}^n$ and any matrices $\mathcal{A}$, $\mathcal{B} \in \mathbf{R}^{n \times m}$ with jth column vectors $\mathcal{A}^{(j)}$ and $\mathcal{B}^{(j)}$, respectively, we define the following wedge products:

$$\mathcal{A} \wedge \mathbf{a} := [\mathcal{A}^{(1)} \wedge \mathbf{a}, \ldots, \mathcal{A}^{(m)} \wedge \mathbf{a}], \qquad \mathbf{a} \wedge \mathcal{A} := [\mathbf{a} \wedge \mathcal{A}^{(1)}, \ldots, \mathbf{a} \wedge \mathcal{A}^{(m)}], \tag{2.3}$$

$$\mathcal{A} \wedge \mathcal{B} := \mathcal{A}^{(1)} \wedge \mathcal{B}^{(1)} + \cdots + \mathcal{A}^{(m)} \wedge \mathcal{B}^{(m)}. \tag{2.4}$$

We point out that (2.4) defines $\mathcal{A} \wedge \mathcal{B}$ to be a vector instead of a matrix, which seems not to be natural. However, we shall see in section 6 that it turns out that this is a convenient and useful notation.

We conclude this section by citing some known results which will be used in the later sections. The first result is known as "the decisive monotonicity trick," and its proof can be found in [52].

LEMMA 2.1. *Let X be a reflexive Banach space, and X' denote the dual space of X. Suppose that an operator $\mathcal{F} : X \longrightarrow X'$ satisfies*

(i) *$\mathcal{F}$ is monotone on X, that is, $\langle \mathcal{F}u - \mathcal{F}v, u - v\rangle \geq 0$ for all $u, v \in X$;*

(ii) *$\mathcal{F}$ is hemicontinuous, that is, the function $t \mapsto \langle \mathcal{F}(u + tv), w\rangle$ is continuous on $[0, 1]$ for all $u, v, w \in X$.*

In addition, suppose that $u_k \to u$ and $\mathcal{F}u_k \to f$ weakly in X and X', respectively, as $n \to \infty$, and

$$\overline{\lim_{k\to\infty}} \langle \mathcal{F}u_k, u_k\rangle \leq \langle f, u\rangle.$$

Then

$$\mathcal{F}u = f.$$

The second lemma is a compactness result, it can be proved following the proof of Theorem 2.1 of [12] (also see Theorem 3 of Chapter 4 of [20]) using the fact that the operator Δ_p^ε is uniformly elliptic.

LEMMA 2.2. *For $1 \leq p < \infty$, let $p^* = \max\{2, p\}$. For a fixed $\varepsilon > 0$, let $\{\mathbf{w}^{\varepsilon,\delta}\}_{\delta>0}$ be bounded in $L^\infty((0,T); W^{1,p^*}(\Omega, \mathbf{R}^n))$, and let $\{\mathbf{f}^{\varepsilon,\delta}\}_{\delta>0}$ be bounded in $L^1((0,T); L^1(\Omega,\ \mathbf{R}^n))$, both uniformly in δ. Moreover, suppose that $\{\frac{\partial \mathbf{w}^{\varepsilon,\delta}}{\partial t}\}_{\delta>0}$ is bounded in $L^2((0,T);\ L^2(\Omega, \mathbf{R}^n))$ uniformly in δ and $\mathbf{w}^{\varepsilon,\delta}$ satisfies the equation*

$$\frac{\partial \mathbf{w}^{\varepsilon,\delta}}{\partial t} - \Delta_p^\varepsilon \mathbf{w}^{\varepsilon,\delta} = \mathbf{f}^{\varepsilon,\delta} \qquad \text{in } \Omega_T, \quad \delta > 0,$$

in the distributional sense. Then $\{\mathbf{w}^{\varepsilon,\delta}\}_{\delta>0}$ is precompact in $L^q((0,T); W^{1,q}(\Omega, \mathbf{R}^n))$ for all $1 \leq q < p^*$.

The last lemma is a variation of Lemma 2.2, and its proof can be found in [38] (also see Lemma 9 of [40]).

LEMMA 2.3. *For $1 < p < \infty$, let $\{\mathbf{w}^\varepsilon\}_{\varepsilon>0}$ be bounded in $L^\infty((0,T); W^{1,p}(\Omega, \mathbf{R}^n))$, and $\{\mathbf{f}^\varepsilon\}$ be bounded in $L^1((0,T);\ L^1(\Omega, \mathbf{R}^n))$, both uniformly in ε. Moreover, suppose that $\{\frac{\partial \mathbf{w}^\varepsilon}{\partial t}\}$ is bounded uniformly in ε in $L^2((0,T); L^2(\Omega, \mathbf{R}^n))$ and $\mathbf{w}^\varepsilon$ satisfies the equation*

$$\frac{\partial \mathbf{w}^\varepsilon}{\partial t} - \Delta_p^\varepsilon \mathbf{w}^\varepsilon = \mathbf{f}^\varepsilon \qquad \text{in } \Omega_T, \quad \varepsilon > 0,$$

in the distributional sense. Then $\{\mathbf{w}^\varepsilon\}_{\varepsilon>0}$ is precompact in $L^q((0,T); W^{1,q}(\Omega, \mathbf{R}^n))$ for all $1 \leq q < p$.

Remark 2.1. In Lemma 2.2, $\varepsilon > 0$ is a fixed parameter and the differential operator Δ_p^ε does not depend on the variable index δ. It can be shown that (cf. Theorem 2.1 of [12]) that the lemma still holds for $\varepsilon = 0$ when $p \geq 2$. On the other hand, ε is a variable index in Lemma 2.3, and the operator Δ_p^ε also depends on the variable index ε.

3. Well-posedness of the regularized flow (1.20)–(1.22). In this section we shall analyze the regularized flow (1.20)–(1.22) for each fixed pair of positive numbers (ε, δ). We establish an energy law, a maximum principle, uniform (in both ε and δ) a priori estimates, and existence and uniqueness of weak and classical solutions. We begin with a couple of definitions for solutions to (1.20)–(1.22).

DEFINITION 3.1. *For $1 \le p < \infty$, a map $\mathbf{u}^{\varepsilon,\delta} : \Omega_T \to \mathbf{R}^n$ is called a global* weak *solution to* (1.20)–(1.22) *if*

(i) $\mathbf{u}^{\varepsilon,\delta} \in L^\infty((0,T); W^{1,p^*}(\Omega,\mathbf{R}^n)) \cap H^1((0,T); L^2(\Omega,\mathbf{R}^n))$, *for* $p^* = \max\{2,p\}$;

(ii) $|\mathbf{u}^{\varepsilon,\delta}| \le 1$ *a.e. in* Ω_T;

(iii) $\mathbf{u}^{\varepsilon,\delta}$ *satisfies* (1.20)–(1.22) *in the distributional sense.*

DEFINITION 3.2. *A weak solution $\mathbf{u}^{\varepsilon,\delta}$ to* (1.20)–(1.22) *is called a* strong *solution if $\mathbf{u}^{\varepsilon,\delta} \in L^p((0,T);\, W^{2,p^*}(\Omega,\mathbf{R}^n)) \cap H^1((0,T); L^{p^*}(\Omega,\mathbf{R}^n))$. It is called a* regular *solution if in addition $\mathbf{u}^{\varepsilon,\delta} \in H^1((0,T);\, W^{1,p^*}(\Omega,\mathbf{R}^n))$.*

3.1. Energy law and a priori estimates. Since (1.20)–(1.22) is the gradient flow for the functional $J^{\varepsilon,\delta}_{p,\lambda}$, its regular solutions must satisfy a dissipative energy law. Indeed, we have the following lemma.

LEMMA 3.3. *Let $\mathbf{u}_0$ and $\mathbf{g}$ be sufficiently smooth, and suppose that $\mathbf{u}^{\varepsilon,\delta}$ is a regular solution to* (1.20)–(1.22). *Then $\mathbf{u}^{\varepsilon,\delta}$ satisfies the following energy law:*

$$J^{\varepsilon,\delta}_{p,\lambda}(\mathbf{u}^{\varepsilon,\delta}(s)) + \int_0^s \| \mathbf{u}^{\varepsilon,\delta}_t(t) \|^2_{L^2}\, dt = J^{\varepsilon,\delta}_{p,\lambda}(\mathbf{u}_0) \qquad \text{for a.e. } s \in [0,T]. \tag{3.1}$$

Proof. Testing (1.20) with $\mathbf{u}^{\varepsilon,\delta}_t$ we get

$$\| \mathbf{u}^{\varepsilon,\delta}_t(t) \|^2_{L^2} + \frac{d}{dt} \int_\Omega \Big\{ \frac{b_p(\varepsilon)}{2} |\nabla \mathbf{u}^{\varepsilon,\delta}(t)|^2 + \frac{1}{p} |\nabla \mathbf{u}^{\varepsilon,\delta}(t)|^p_\varepsilon + \frac{1}{\delta} F(\mathbf{u}^{\varepsilon,\delta}(t)) + \frac{\lambda}{2} |\mathbf{u}^{\varepsilon,\delta}(t) - \mathbf{g}|^2 \Big\} dx = 0.$$

The desired identity (3.1) then follows from integrating the above equation in t over the interval $[0,s]$ and using the definition of $J^{\varepsilon,\delta}_{p,\lambda}$. ∎

The above energy law immediately implies the following uniform (in ε and δ) a priori estimates.

COROLLARY 3.4. *Suppose that $\mathbf{u}_0$ and $\mathbf{g}$ satisfy*

$$J^{\varepsilon,\delta}_{p,\lambda}(\mathbf{u}_0) \le c_0 \tag{3.2}$$

for some positive constant c_0 independent of ε and δ. Then there exists another positive constant $C := C(p,\lambda,c_0)$ which is also independent of ε and δ such that

$$\| \mathbf{u}^{\varepsilon,\delta} \|_{L^\infty(W^{1,p})} + \| \mathbf{u}^{\varepsilon,\delta} \|_{H^1(L^2)} \le C \qquad \text{for } 1 \le p < \infty, \tag{3.3}$$

$$\delta^{-\frac{1}{2}} \| \, |\mathbf{u}^{\varepsilon,\delta}|^2 - 1 \, \|_{L^\infty(L^2)} \le C \qquad \text{for } 1 \le p < \infty, \tag{3.4}$$

$$\| \, |\nabla \mathbf{u}^{\varepsilon,\delta}|^{p-2}_\varepsilon \nabla \mathbf{u}^{\varepsilon,\delta} \, \|_{L^\infty(L^{p'})} \le C \qquad \text{for } p' = \frac{p}{p-1}, \quad 1 \le p < \infty, \tag{3.5}$$

$$\sqrt{b_p(\varepsilon)} \, \| \nabla \mathbf{u}^{\varepsilon,\delta} \|_{L^\infty(L^2)} \le C \qquad \text{for } 1 \le p < \infty. \tag{3.6}$$

Next, using a test function technique of [12], we show that the modulus $|\mathbf{u}^{\varepsilon,\delta}|$ of every weak solution $\mathbf{u}^{\varepsilon,\delta}$ to (1.20)–(1.22) satisfies a maximum principle.

LEMMA 3.5. *Suppose that $|\mathbf{g}| \le 1$ and $|\mathbf{u}_0| \le 1$ a.e. in Ω. Then weak solutions to* (1.20)–(1.22) *satisfy $|\mathbf{u}^{\varepsilon,\delta}| \le 1$ a.e. in Ω_T.*

Proof. Define the function

$$\chi(z) := \frac{(z-1)_+}{z} = \begin{cases} 0 & \text{for } 0 \le z \le 1, \\ \dfrac{z-1}{z} & \text{for } z > 1. \end{cases} \tag{3.7}$$

It is easy to check that χ is a nonnegative monotone increasing function on the interval $[0,\infty)$.

Now testing (1.20) with $\mathbf{v} := \mathbf{u}^{\varepsilon,\delta}\chi(|\mathbf{u}^{\varepsilon,\delta}|)$ we get

$$\frac{1}{2}\frac{d}{dt}\int_{\{|\mathbf{u}^{\varepsilon,\delta}|>1\}}\big(|\mathbf{u}^{\varepsilon,\delta}|-1\big)^2\,dx+\int_{\{|\mathbf{u}^{\varepsilon,\delta}|>1\}}\Big\{\,b_p(\varepsilon)|\nabla\mathbf{u}^{\varepsilon,\delta}|^2+|\nabla\mathbf{u}^{\varepsilon,\delta}|_\varepsilon^{p-2}\,|\nabla\mathbf{u}^{\varepsilon,\delta}|^2$$
$$+\frac{1}{\delta}\big(|\mathbf{u}^{\varepsilon,\delta}|^2-1\big)|\mathbf{u}^{\varepsilon,\delta}|^2+\lambda\big(|\mathbf{u}^{\varepsilon,\delta}|^2-\mathbf{u}^{\varepsilon,\delta}\cdot\mathbf{g}\big)\,\Big\}\,\chi(|\mathbf{u}^{\varepsilon,\delta}|)\,dx$$
$$+\frac{1}{4}\int_{\{|\mathbf{u}^{\varepsilon,\delta}|>1\}}\frac{[b_p(\varepsilon)+|\nabla\mathbf{u}^{\varepsilon,\delta}|_\varepsilon^{p-2}]\,|\nabla|\mathbf{u}^{\varepsilon,\delta}|^2|^2}{|\mathbf{u}^{\varepsilon,\delta}|^3}\,dx=0.$$

Since $\mathbf{u}^{\varepsilon,\delta}\cdot\mathbf{g}\le|\mathbf{u}^{\varepsilon,\delta}|\cdot|\mathbf{g}|\le|\mathbf{u}^{\varepsilon,\delta}|$, the second integral is nonnegative. The assertion then follows from integrating the above inequality and using the assumption $|\mathbf{u}_0|\le 1$ a.e. in Ω. ∎

3.2. Existence of global weak and classical solutions. We now state and prove the existence of global weak and classical solutions to the regularized flow (1.20)–(1.22) for each fixed pair of positive numbers (ε,δ). Since $-\Delta_p^\varepsilon$ is uniformly elliptic for all $1\le p<\infty$, the existence of classical solutions follows immediately from the classical theory of parabolic partial differential equations (cf. [35]).

THEOREM 3.6. *Let $\Omega\subset\mathbf{R}^m$ be a bounded domain with a smooth boundary. Suppose that $\mathbf{u}_0$ and $\mathbf{g}$ are sufficiently smooth functions (say, $\mathbf{u}_0,\mathbf{g}\in[C^3(\overline{\Omega})]^n$) and satisfy* (3.2). *Then for each fixed pair of positive numbers (ε,δ), the regularized flow* (1.20)–(1.22) *possesses a unique global classical solution $\mathbf{u}^{\varepsilon,\delta}$. Moreover, $\mathbf{u}^{\varepsilon,\delta}$ satisfies the following energy law:*

$$J_{p,\lambda}^{\varepsilon,\delta}(\mathbf{u}^{\varepsilon,\delta}(s))+\int_0^s\|\,\mathbf{u}_t^{\varepsilon,\delta}(t)\,\|_{L^2}^2\,dt=J_{p,\lambda}^{\varepsilon,\delta}(\mathbf{u}_0)\qquad\text{for a.e. } s\in[0,T]. \tag{3.8}$$

Proof. The existence and uniqueness follow from an application of the standard results for parabolic systems; see Chapter 5 of [35]. Equation (3.8) follows from Lemma 3.3. ∎

For less regular functions $\mathbf{u}_0$ and $\mathbf{g}$, we have the following weaker result.

THEOREM 3.7. *Let $\Omega\subset\mathbf{R}^m$ be a bounded domain with smooth boundary. Suppose that $|\mathbf{u}_0|\le 1$ and $|\mathbf{g}|\le 1$ a.e. in Ω and $J_{p,\lambda}^{\varepsilon,\delta}(\mathbf{u}_0)<\infty$. Then the regularized flow* (1.20)–(1.22) *has a unique global weak solution $\mathbf{u}^{\varepsilon,\delta}$ in the sense of Definition* 3.1. *Moreover, $\mathbf{u}^{\varepsilon,\delta}$ satisfies the following energy inequality:*

$$J_{p,\lambda}^{\varepsilon,\delta}(\mathbf{u}^{\varepsilon,\delta}(s))+\int_0^s\|\,\mathbf{u}_t^{\varepsilon,\delta}(t)\,\|_{L^2}^2\,dt\le J_{p,\lambda}^{\varepsilon,\delta}(\mathbf{u}_0)\qquad\forall s\in[0,T]. \tag{3.9}$$

Proof. Let η_ρ denote any well-known mollifier (cf. [35]), and let $\mathbf{u}_{0,\rho}:=\eta_\rho*\mathbf{u}_0$ and $\mathbf{g}_\rho:=\eta_\rho*\mathbf{g}$ denote the mollifications of $\mathbf{u}_0$ and $\mathbf{g}$, respectively. Let $\mathbf{u}_\rho^{\varepsilon,\delta}$ denote the classical solution to (1.20)–(1.22) corresponding to the smoothed datum functions $\mathbf{u}_{0,\rho}$ and $\mathbf{g}_\rho$. Hence, $\mathbf{u}_\rho^{\varepsilon,\delta}$ satisfies the energy law (3.8) with $\mathbf{u}_{0,\rho}$ and $\mathbf{g}_\rho$ in the place of $\mathbf{u}_0$ and $\mathbf{g}$.

From Lemma 3.5 we know that $|\mathbf{u}_\rho^{\varepsilon,\delta}|$ satisfies the maximum principle, thus,

$$\max_{(x,t)\in\overline{\Omega}_T}|\mathbf{u}_\rho^{\varepsilon,\delta}(x,t)|\le 1.$$

Next, since

$$\lim_{\rho\to 0} J^{\varepsilon,\delta}_{p,\lambda}(\mathbf{u}^{\varepsilon}_{0,\rho}) = J^{\varepsilon,\delta}_{p,\lambda}(\mathbf{u}_0) < \infty,$$

the energy law for $\mathbf{u}^{\varepsilon,\delta}_{\rho}$ immediately implies that $\mathbf{u}^{\varepsilon,\delta}_{\rho}$ satisfies estimates (3.3)–(3.6), *uniformly* in ρ and δ.

The remainder of the proof is to extract a convergent subsequence of $\{\mathbf{u}^{\varepsilon,\delta}_{\rho}\}_{\rho>0}$ and to pass to the limit as $\rho \to 0$ in the weak formulation of (1.20). This can be done easily in all terms of (1.20) except the nonlinear term in Δ^{ε}_{p} (see (1.23)). To overcome the difficulty, we appeal to Minty's trick or the "decisive monotonicity trick" as described in Lemma 2.1. Since this part of proof is the same as that of Theorem 1.5 of [12], we omit it and refer to [12] for the details.

The uniqueness of weak solutions follows from a standard perturbation argument and using the fact that Δ^{ε}_{p} is a monotone operator. Finally, (3.9) follows from setting $\rho \to 0$ in the energy law (3.8) for $\mathbf{u}^{\varepsilon,\delta}_{\rho}$ and from using the lower semicontinuity of $J^{\varepsilon,\delta}_{p,\lambda}$ and the L^2-norm with respect to L^2 weak convergence. ∎

We conclude this section with some remarks.

Remark 3.1. (a) Since $W^{1,1}(\Omega)$ is not a *reflexive* Banach space, Minty's trick would not apply in the case $p = 1$ without the help of the $b_p(\varepsilon)\Delta$ term in the operator Δ^{ε}_{p} (see (1.23)). On the other hand, in the presence of this term, Minty's trick does apply since we now deal with the Sobolev space $H^1(\Omega)$ instead of $W^{1,1}(\Omega)$. In fact, if $b_p(\varepsilon) = 0$ in (1.20), BV solutions are what one can only expect in general for the regularized flow (1.20)–(1.22) in the case $p = 1$ (cf. [24]).

(b) We also point out that using nonzero parameter $b_p(\varepsilon)$ is not necessary in the case $1 < p < \infty$. For example, the conclusion of Theorem 3.7 still holds if we replace the above $b_p(\varepsilon)$ and p^* by

$$\widehat{b}_p(\varepsilon) = \begin{cases} 0 & \text{if } 1 < p < \infty, \\ \varepsilon^{\alpha} & \text{if } p = 1 \quad (\alpha > 0) \end{cases} \qquad \text{and} \quad \widehat{p^*} = \begin{cases} p & \text{if } 1 < p < \infty, \\ 2 & \text{if } p = 1. \end{cases}$$

On the other hand, the conclusion of Theorem 3.6 may no longer be true after this modification.

(c) Theorem 3.7 also holds when Ω is a bounded Lipschitz domain. One way to prove this assertion is to use the Galerkin method as done in [12], or to use the finite element method to be introduced in section 7.

4. Passing to the limit as $\delta \to 0$. The goal of this section is to derive the limiting flow of (1.20)–(1.22) as $\delta \to 0$ for each *fixed* $\varepsilon > 0$. As in [11, 12, 13, 31, 41, 46], the key ideas for passing to the limit are to use the compactness result of Lemma 2.2 and to exploit the symmetries of the unit sphere S^{n-1}. Our main result of this section is the following existence theorem.

THEOREM 4.1. *Let $p^* := \max\{2, p\}$ and let $\Omega \subset \mathbf{R}^m$ be a bounded Lipschitz domain. For $1 \le p < \infty$, suppose that $\mathbf{u}_0 \in W^{1,p^*}(\Omega, \mathbf{R}^n)$, $|\mathbf{u}_0| = 1$ and $|\mathbf{g}| \le 1$ in Ω. Then there exists a map $\mathbf{u}^{\varepsilon} \in L^{\infty}((0,T); W^{1,p^*}(\Omega, \mathbf{R}^n)) \cap H^1((0,T); L^2(\Omega, \mathbf{R}^n))$ such that*

$$|\mathbf{u}^{\varepsilon}| = 1 \qquad \text{a.e. in } \Omega_T, \tag{4.1}$$

and $\mathbf{u}^\varepsilon$ *is a weak solution (in the distributional sense; see* (4.29)*) to the problem*

$$\mathbf{u}_t^\varepsilon - \Delta_p^\varepsilon \mathbf{u}^\varepsilon + \lambda(\mathbf{u}^\varepsilon - \mathbf{g}) = \mu_{p,\lambda}^\varepsilon \mathbf{u}^\varepsilon \qquad \text{in } \Omega_T, \tag{4.2}$$

$$\mathcal{B}_p^\varepsilon \mathbf{n} = 0 \qquad \text{on } \partial\Omega_T, \tag{4.3}$$

$$\mathbf{u}^\varepsilon = \mathbf{u}_0 \qquad \text{on } \Omega \times \{t = 0\}, \tag{4.4}$$

where

$$\mu_{p,\lambda}^\varepsilon := b_p(\varepsilon)|\nabla \mathbf{u}^\varepsilon|^2 + |\nabla \mathbf{u}^\varepsilon|_\varepsilon^{p-2}\,|\nabla \mathbf{u}^\varepsilon|^2 + \lambda(1 - \mathbf{u}^\varepsilon \cdot \mathbf{g}). \tag{4.5}$$

Moreover, $\mathbf{u}^\varepsilon$ *satisfies the energy inequality*

$$J_{p,\lambda}^\varepsilon(\mathbf{u}^\varepsilon(s)) + \int_0^s \| \mathbf{u}_t^\varepsilon(t) \|_{L^2}^2 \, dt \le J_{p,\lambda}^\varepsilon(\mathbf{u}_0) \le J_{p,\lambda}^1(\mathbf{u}_0) \qquad \text{for a.e. } s \in [0,T] \tag{4.6}$$

and the additional estimates

$$\| \, |\nabla \mathbf{u}^\varepsilon|_\varepsilon^{p-2} \nabla \mathbf{u}^\varepsilon \, \|_{L^\infty(L^{p'})} \le C \qquad \text{for } p' = \frac{p}{p-1}, \quad 1 \le p < \infty, \tag{4.7}$$

$$\| \operatorname{div} \mathcal{B}_p^\varepsilon \|_{L^2(L^1)} \le C \qquad \text{for } 1 \le p < \infty, \tag{4.8}$$

$$\| \operatorname{div} (\mathcal{B}_p^\varepsilon \wedge \mathbf{u}^\varepsilon) \|_{L^2(L^2)} \le C \qquad \text{for } 1 \le p < \infty, \tag{4.9}$$

$$\sqrt{b_p(\varepsilon)} \| \nabla \mathbf{u}^\varepsilon \|_{L^\infty(L^2)} \le C \tag{4.10}$$

for some positive ε*-independent constant* C*. Here* $\mathcal{B}_p^\varepsilon$ *is the* $n \times m$ *matrix*

$$\mathcal{B}_p^\varepsilon := b_p(\varepsilon)\nabla \mathbf{u}^\varepsilon + |\nabla \mathbf{u}^\varepsilon|_\varepsilon^{p-2} \nabla \mathbf{u}^\varepsilon \tag{4.11}$$

and

$$J_{p,\lambda}^\varepsilon(\mathbf{v}) := E_p^\varepsilon(\mathbf{v}) + \frac{\lambda}{2} \int_\Omega |\mathbf{v} - \mathbf{g}|^2 \, dx. \tag{4.12}$$

Proof. In light of Remark 3.1(c) and the density argument, without loss of the generality we assume that Ω is a bounded smooth domain, and $\mathbf{u}_0$ and $\mathbf{g}$ are smooth functions. On noting that $|\mathbf{u}_0| = 1$ implies that $L^\delta(\mathbf{u}_0) = 0$, then $E_p^{\varepsilon,\delta}(\mathbf{u}_0) = E_p^\varepsilon(\mathbf{u}_0)$ in (1.16). Since $E_p^\varepsilon(\mathbf{u}_0) \le E_p^1(\mathbf{u}_0)$, hence the assumptions on $\mathbf{u}_0$ and $\mathbf{g}$ ensure (3.2) holds. Since the proof is long, we divide it into three steps.

Step 1: *Extracting a convergent subsequence.* Let $\mathbf{u}^{\varepsilon,\delta}$ be the weak solution solution to (1.20)–(1.22) whose existence is established in Theorem 3.7. Since $E_p^{\varepsilon,\delta}(\mathbf{u}_0) \le E_p^1(\mathbf{u}_0)$, then $J_{p,\lambda}^{\varepsilon,\delta}(\mathbf{u}_0)$ is uniformly bounded with respect to ε and δ. Hence, (3.9) implies that $\mathbf{u}^{\varepsilon,\delta}$ satisfies the uniform (in both ε and δ) estimates (3.3)–(3.6) and the maximum principle $|\mathbf{u}^{\varepsilon,\delta}| \le 1$ on $\overline{\Omega}_T$.

By the weak compactness of $W^{1,p}(\Omega)$ and Sobolev embedding (cf. [1, 45]), there exists a subsequence of $\{\mathbf{u}^{\varepsilon,\delta}\}_{\delta>0}$ (still denoted by the same notation) and a map

$\mathbf{u}^{\varepsilon} \in L^{\infty}((0,T);W^{1,p^*}(\Omega,\mathbf{R}^n)) \cap H^1((0,T);L^2(\Omega,\mathbf{R}^n))$ such that as $\delta \to 0$

$$\mathbf{u}^{\varepsilon,\delta} \longrightarrow \mathbf{u}^{\varepsilon} \qquad \text{weakly* in } L^{\infty}((0,T);W^{1,p^*}(\Omega,\mathbf{R}^n)), \tag{4.13}$$

$$\text{strongly in } L^2((0,T);L^2(\Omega,\mathbf{R}^n)), \tag{4.14}$$

$$\text{a.e. in } \Omega_T, \tag{4.15}$$

$$\mathbf{u}_t^{\varepsilon,\delta} \longrightarrow \mathbf{u}_t^{\varepsilon} \qquad \text{weakly in } L^2((0,T);L^2(\Omega,\mathbf{R}^n)), \tag{4.16}$$

$$|\mathbf{u}^{\varepsilon,\delta}| \longrightarrow 1 \qquad \text{strongly in } L^2((0,T);L^2(\Omega)). \tag{4.17}$$

It follows immediately from (4.15) and (4.17) that

$$|\mathbf{u}^{\varepsilon}| = 1 \qquad \text{a.e. in } \Omega_T. \tag{4.18}$$

Step 2: *Wedge product technique and passing to the limit.* Since $|\mathbf{u}^{\varepsilon,\delta}| \leq 1$ in $\overline{\Omega}_T$, an application of the Lebesgue-dominated convergence theorem yields, on noting (4.15), that

$$\mathbf{u}^{\varepsilon,\delta} \overset{\delta \searrow 0}{\longrightarrow} \mathbf{u}^{\varepsilon} \qquad \text{strongly in } L^r((0,T);L^r(\Omega,\mathbf{R}^n)) \quad \forall r \in [1,\infty). \tag{4.19}$$

Let

$$\mathbf{f}^{\varepsilon,\delta} := \frac{1}{\delta}\big(1 - |\mathbf{u}^{\varepsilon,\delta}|^2\big)\mathbf{u}^{\varepsilon,\delta} - \lambda\big(\mathbf{u}^{\varepsilon,\delta} - \mathbf{g}\big).$$

It follows from the estimate $|\mathbf{u}^{\varepsilon,\delta}| \leq 1$ and the inequality $|\mathbf{u}^{\varepsilon,\delta}| \leq \frac{1}{2}\big(1 + |\mathbf{u}^{\varepsilon,\delta}|^2\big)$ that

$$\frac{1}{\delta}\int_0^T \int_\Omega \big(1 - |\mathbf{u}^{\varepsilon,\delta}|^2\big)\,|\mathbf{u}^{\varepsilon,\delta}|\,dxdt \leq \frac{1}{\delta}\int_0^T \int_\Omega \big(1 - |\mathbf{u}^{\varepsilon,\delta}|^2\big)^2\,dxdt + \frac{1}{\delta}\int_0^T \int_\Omega \big(1 - |\mathbf{u}^{\varepsilon,\delta}|^2\big)\,|\mathbf{u}^{\varepsilon,\delta}|^2\,dxdt.$$

Inequality (3.4) immediately implies that the first term on the right-hand side is uniformly bounded in δ. Testing (1.20) by $\mathbf{u}^{\varepsilon,\delta}$ and using estimates (3.3), (3.5), and (3.6), we conclude that the second term on the right-hand side is also uniformly bounded in δ. Hence, $\mathbf{f}^{\varepsilon,\delta}$ is uniformly bounded with respect to δ in $L^1((0,T);L^1(\Omega,\mathbf{R}^n))$. By Lemma 2.2 we have that

$$\nabla\mathbf{u}^{\varepsilon,\delta} \overset{\delta \searrow 0}{\longrightarrow} \nabla\mathbf{u}^{\varepsilon} \qquad \text{strongly in } L^q((0,T);L^q(\Omega,\mathbf{R}^{n\times m})) \quad \forall q \in [1,p^*). \tag{4.20}$$

This and (3.5) imply that

$$|\nabla\mathbf{u}^{\varepsilon,\delta}|_{\varepsilon}^{p-2}\nabla\mathbf{u}^{\varepsilon,\delta} \overset{\delta \searrow 0}{\longrightarrow} |\nabla\mathbf{u}^{\varepsilon}|_{\varepsilon}^{p-2}\nabla\mathbf{u}^{\varepsilon} \tag{4.21}$$

weakly* in $L^{\infty}((0,T);L^{p'}(\Omega,\mathbf{R}^{n\times m}))$ with $p' = \frac{p}{p-1}$ if $p \neq 1$ and weakly* in $L^{\infty}((0,T);L^{\infty}(\Omega,\mathbf{R}^{n\times m}))$ if $p = 1$.

Next, taking the wedge product of (1.20) with $\mathbf{u}^{\varepsilon,\delta}$ yields

$$\mathbf{u}_t^{\varepsilon,\delta} \wedge \mathbf{u}^{\varepsilon,\delta} - \operatorname{div}\big(\mathcal{B}_p^{\varepsilon,\delta} \wedge \mathbf{u}^{\varepsilon,\delta}\big) - \lambda\,\mathbf{g} \wedge \mathbf{u}^{\varepsilon,\delta} = 0, \tag{4.22}$$

where $\mathcal{B}_p^{\varepsilon,\delta}$ is defined in (1.23).

Testing (4.22) with any $\mathbf{w} \in L^\infty((0,T);W^{1,p^*}(\Omega,\mathbf{R}^n))$ we get

$$\int_0^T \int_\Omega \Big\{ \big(\mathbf{u}_t^{\varepsilon,\delta} \wedge \mathbf{u}^{\varepsilon,\delta}\big) \cdot \mathbf{w} + \big(\mathcal{B}_p^{\varepsilon,\delta} \wedge \mathbf{u}^{\varepsilon,\delta}\big) \cdot \nabla \mathbf{w} - \lambda \big(\mathbf{g} \wedge \mathbf{u}^{\varepsilon,\delta}\big) \cdot \mathbf{w} \Big\}\, dxdt = 0. \tag{4.23}$$

It follows from setting $\delta \to 0$ in (4.23) and using (4.16), (4.19), and (4.21) that

$$\int_0^T \int_\Omega \Big\{ \big(\mathbf{u}_t^{\varepsilon} \wedge \mathbf{u}^{\varepsilon}\big) \cdot \mathbf{w} + \big(\mathcal{B}_p^{\varepsilon} \wedge \mathbf{u}^{\varepsilon}\big) \cdot \nabla \mathbf{w} - \lambda \big(\mathbf{g} \wedge \mathbf{u}^{\varepsilon}\big) \cdot \mathbf{w} \Big\}\, dxdt = 0, \tag{4.24}$$

where $\mathcal{B}_p^{\varepsilon}$ is given by (4.11).

Note that (4.18) implies

$$\mathbf{u}_t^{\varepsilon} \cdot \mathbf{u}^{\varepsilon} = 0, \qquad \big(\mathcal{B}_p^{\varepsilon}\big)^T \mathbf{u}^{\varepsilon} = 0 \qquad \text{a.e. in } \Omega_T. \tag{4.25}$$

This in turn yields the following identity

$$\int_0^T \int_\Omega \Big\{ \mathbf{u}_t^{\varepsilon} \cdot \mathbf{u}^{\varepsilon} \varphi + \mathcal{B}_p^{\varepsilon} \cdot \nabla (\mathbf{u}^{\varepsilon} \varphi) + \lambda \big(\mathbf{u}^{\varepsilon} - \mathbf{g}\big) \cdot \mathbf{u}^{\varepsilon} \varphi \Big\} dxdt = \int_0^T \int_\Omega \mu_{p,\lambda}^{\varepsilon}\, \varphi\, dxdt \tag{4.26}$$

for any $\varphi \in L^\infty(\Omega_T) \cap L^\infty((0,T);W^{1,p^*}(\Omega))$, where $\mu_{p,\lambda}^{\varepsilon}$ is defined by (4.5).

Finally, for any $\mathbf{v} \in [C^1(\overline{\Omega}_T)]^n$, taking $\mathbf{w} = \mathbf{u}^{\varepsilon} \wedge \mathbf{v}$ in (4.24), $\varphi = \mathbf{u}^{\varepsilon} \cdot \mathbf{v}$ in (4.26), and using the formula $\mathbf{a} \cdot (\mathbf{b} \wedge \mathbf{c}) = (\mathbf{a} \wedge \mathbf{b}) \cdot \mathbf{c}$ yield

$$\begin{aligned} \int_0^T \int_\Omega \Big\{ & \mathbf{u}_t^{\varepsilon} \cdot \big(\mathbf{u}^{\varepsilon} \wedge (\mathbf{u}^{\varepsilon} \wedge \mathbf{v})\big) + \mathcal{B}_p^{\varepsilon} \cdot \nabla (\mathbf{u}^{\varepsilon} \wedge (\mathbf{u}^{\varepsilon} \wedge \mathbf{v})) \\ & - \lambda\, \mathbf{g} \cdot \big(\mathbf{u}^{\varepsilon} \wedge (\mathbf{u}^{\varepsilon} \wedge \mathbf{v})\big) \Big\} dxdt = 0, \end{aligned} \tag{4.27}$$

$$\begin{aligned} \int_0^T \int_\Omega \Big\{ & \mathbf{u}_t^{\varepsilon} \cdot \mathbf{u}^{\varepsilon} (\mathbf{u}^{\varepsilon} \cdot \mathbf{v}) + \mathcal{B}_p^{\varepsilon} \cdot \nabla (\mathbf{u}^{\varepsilon} (\mathbf{u}^{\varepsilon} \cdot \mathbf{v})) \\ & + \lambda \big(\mathbf{u}^{\varepsilon} - \mathbf{g}\big) \cdot \mathbf{u}^{\varepsilon} (\mathbf{u}^{\varepsilon} \cdot \mathbf{v}) \Big\}\, dxdt = \int_0^T \int_\Omega \mu_{p,\lambda}^{\varepsilon} \mathbf{u}^{\varepsilon} \cdot \mathbf{v}\, dxdt. \end{aligned} \tag{4.28}$$

Subtracting (4.27) from (4.28) and using the identity

$$\mathbf{v} = (\mathbf{u}^{\varepsilon} \cdot \mathbf{v})\, \mathbf{u}^{\varepsilon} - \mathbf{u}^{\varepsilon} \wedge (\mathbf{u}^{\varepsilon} \wedge \mathbf{v}),$$

we obtain that

$$\int_0^T \int_\Omega \Big\{ \mathbf{u}_t^{\varepsilon} \cdot \mathbf{v} + \mathcal{B}_p^{\varepsilon} \cdot \nabla \mathbf{v} + \lambda \big(\mathbf{u}^{\varepsilon} - \mathbf{g}\big) \cdot \mathbf{v} \Big\}\, dxdt = \int_0^T \int_\Omega \mu_{p,\lambda}^{\varepsilon} \mathbf{u}^{\varepsilon} \cdot \mathbf{v}\, dxdt \tag{4.29}$$

for any $\mathbf{v} \in [C^1(\overline{\Omega}_T)]^n$. This is equivalent to saying that $\mathbf{u}^{\varepsilon}$ is a weak solution (in the distributional sense) to (4.2)–(4.4).

Step 3: *Wrapping up.* We conclude the proof by showing the estimates (4.6)–(4.10). First, (4.6) follows immediately from letting $\delta \to 0$ in (3.9), appealing to

Fatou's lemma and the lower semicontinuity of L^2- and L^{p^*}-norms with respect to L^2- and L^{p^*}-weak convergence. We emphasize again that this is possible in the case $p = 1$, because the uniform (in δ) estimate (3.6) implies that $u^\varepsilon \in L^\infty((0,T);H^1(\Omega,\mathbf{R}^n))$. Inequalities (4.7) and (4.10) are direct consequences of (4.6). Finally, the bounds (4.8) and (4.9) follow immediately from (4.29) and (4.24), respectively. Hence the proof is complete. □

Remark 4.1. If $b_p(\varepsilon) = 0$ is used in the regularization, the solutions to (4.1)–(4.4) are only expected to belong to $L^\infty((0,T);[BV(\Omega)]^n)$ in general when $p = 1$.

5. Passing to the limit as $\varepsilon \to 0$: The case $1 < p < \infty$. In this section, we shall pass to the limit as $\varepsilon \to 0$ in (4.1)–(4.4) and show that the limit map is a weak solution to (1.8)–(1.11). Since the analysis and techniques for passing the limit for $1 < p < \infty$ and $p = 1$ are quite different, we shall first consider the case $1 < p < \infty$ in this section and leave the case $p = 1$ to the next section. We begin with a definition of weak solutions to (1.8)–(1.11) in the case $1 < p < \infty$.

DEFINITION 5.1. *For $1 < p < \infty$, a map $\mathbf{u} : \Omega_T \to \mathbf{R}^n$ is called a global* weak *solution to* (1.8)–(1.11) *if*

(i) $\mathbf{u} \in L^\infty((0,T);W^{1,p}(\Omega,\mathbf{R}^n)) \cap H^1((0,T);L^2(\Omega,\mathbf{R}^n))$;

(ii) $|\,bu| = 1$ *a.e. on* Ω_T;

(iii) $\mathbf{u}$ *satisfies* (1.8)–(1.11) *in the distributional sense.*

Our main result of this section is the following existence theorem.

THEOREM 5.2. *Let $1 < p < \infty$, and suppose that the assumptions on $\mathbf{u}_0$ and $\mathbf{g}$ in Theorem* 4.1 *still hold. Then problem* (1.8)–(1.11) *has a weak solution $\mathbf{u}$ in the sense of Definition* 5.1. *Moreover, $\mathbf{u}$ satisfies the energy inequality*

$$J_{p,\lambda}(\mathbf{u}(s)) + \int_0^s \|\,\mathbf{u}_t(t)\,\|_{L^2}^2\,dt \le J_{p,\lambda}(\mathbf{u}_0) \qquad \text{for a.e. } s \in [0,T], \tag{5.1}$$

where $J_{p,\lambda}$ is defined by (1.7).

Proof. We divide the proof into three steps.

Step 1: *Extracting a convergent subsequence.* Let $\mathbf{u}^\varepsilon$ denote the solution of (4.1)–(4.4) constructed in Theorem 4.1. From (4.1), (4.6)–(4.10), the weak compactness of $W^{1,p}(\Omega)$, and Sobolev embedding (cf. [1, 45]), there exists a subsequence of $\{\mathbf{u}^\varepsilon\}_{\varepsilon>0}$ (still denoted by the same notation) and a map $\mathbf{u} \in L^\infty((0,T);W^{1,p}(\Omega,\mathbf{R}^n)) \cap H^1((0,T);L^2(\Omega,\mathbf{R}^n))$ such that as $\varepsilon \to 0$

$$\mathbf{u}^\varepsilon \longrightarrow \mathbf{u} \qquad \text{weakly* in } L^\infty((0,T);W^{1,p}(\Omega,\mathbf{R}^n)), \tag{5.2}$$

$$\text{strongly in } L^2((0,T);L^p(\Omega,\mathbf{R}^n)), \tag{5.3}$$

$$\text{a.e. in } \Omega_T, \tag{5.4}$$

$$b_p(\varepsilon)\nabla\mathbf{u}^\varepsilon \longrightarrow 0 \qquad \text{weakly in } L^2((0,T);L^2(\Omega,\mathbf{R}^{n\times m})), \tag{5.5}$$

$$\mathbf{u}_t^\varepsilon \longrightarrow \mathbf{u}_t \qquad \text{weakly in } L^2((0,T);L^2(\Omega,\mathbf{R}^n)). \tag{5.6}$$

It follows immediately from (5.4) and (4.18) that

$$|\mathbf{u}| = 1 \qquad \text{a.e. in } \Omega_T. \tag{5.7}$$

Step 2: *Passing to the limit.* First, by (5.4) and the Lebesgue-dominated convergence theorem we have that

$$\mathbf{u}^\varepsilon \overset{\varepsilon\searrow 0}{\longrightarrow} \mathbf{u} \qquad \text{strongly in } L^r((0,T);L^r(\Omega,\mathbf{R}^n)) \quad \forall r \in [1,\infty). \tag{5.8}$$

Next, let $\mathbf{f}^\varepsilon := \mu_{p,\lambda}^\varepsilon \mathbf{u}^\varepsilon - \lambda(\mathbf{u}^\varepsilon - \mathbf{g})$. Clearly, $\mathbf{f}^\varepsilon \in L^1((0,T);L^1(\Omega;\mathbf{R}^n))$ and is uniformly bounded, on noting (4.5)–(4.6). By Lemma 2.3 we get

$$\nabla \mathbf{u}^\varepsilon \overset{\varepsilon\searrow 0}{\longrightarrow} \nabla \mathbf{u} \qquad \text{strongly in } L^q((0,T);L^q(\Omega,\mathbf{R}^{n\times m})) \quad \forall q\in[1,p), \tag{5.9}$$

which, using (4.7) and (5.5), implies that

$$|\nabla \mathbf{u}^\varepsilon|_\varepsilon^{p-2}\nabla \mathbf{u}^\varepsilon \overset{\varepsilon\searrow 0}{\longrightarrow} |\nabla \mathbf{u}|^{p-2}\nabla \mathbf{u} \quad \text{weakly* in } L^\infty((0,T);L^{p'}(\Omega,\mathbf{R}^{n\times m})), \tag{5.10}$$

$$\mathcal{B}_p^\varepsilon \overset{\varepsilon\searrow 0}{\longrightarrow} \mathcal{B}_p := |\nabla \mathbf{u}|^{p-2}\nabla \mathbf{u} \quad \text{weakly in } L^2((0,T);L^{p'_*}(\Omega,\mathbf{R}^{n\times m})), \tag{5.11}$$

where $p' = \frac{p}{p-1}$ and $p'_* := \min\{2,p'\}$.

It then follows from taking $\varepsilon\to 0$ in (4.24) and using (5.6), (5.8), and (5.11) that

$$\int_0^T\int_\Omega \Big\{ \big(\mathbf{u}_t \wedge \mathbf{u}\big)\cdot \mathbf{w} + \big(\mathcal{B}_p \wedge \mathbf{u}\big)\cdot \nabla \mathbf{w} - \lambda\big(\mathbf{g}\wedge \mathbf{u}\big)\cdot \mathbf{w}\Big\}\, dxdt = 0 \tag{5.12}$$

for any $\mathbf{w}\in C^1(\overline{\Omega}_T)$. Since $\mathcal{B}_p \in L^\infty((0,T);L^{p'}(\Omega,\mathbf{R}^{n\times m}))$, by the standard density argument one can show that (5.12) also holds for all $\mathbf{w}\in L^\infty((0,T);W^{1,p}(\Omega,\mathbf{R}^n))\cap L^\infty(\Omega_T)$.

Since $|\mathbf{u}| = 1$ a.e. in Ω_T, there holds the following identity, which is analogous to (4.26):

$$\int_0^T\int_\Omega \Big\{ \mathbf{u}_t\cdot \mathbf{u}\varphi + \mathcal{B}_p\cdot\nabla(\mathbf{u}\varphi) + \lambda\big(\mathbf{u}-\mathbf{g}\big)\cdot \mathbf{u}\varphi\Big\}\, dxdt = \int_0^T\int_\Omega \mu_{p,\lambda}\varphi\, dxdt \tag{5.13}$$

for any $\varphi\in L^\infty(\Omega_T)\cap L^\infty((0,T);W^{1,p}(\Omega,\mathbf{R}^n))$, where $\mu_{p,\lambda}$ is defined by (1.12).

Finally, for any $\mathbf{v}\in [C^1(\overline{\Omega}_T)]^n$, on choosing $\mathbf{w} = \mathbf{u}\wedge\mathbf{v}$ in (5.12) and $\varphi = \mathbf{u}\cdot\mathbf{v}$ in (5.13), subtracting the resulting equations, and using the identity

$$\mathbf{v} = (\mathbf{u}\cdot\mathbf{v})\,\mathbf{u} - \mathbf{u}\wedge(\mathbf{u}\wedge\mathbf{v}),$$

we obtain that

$$\int_0^T\int_\Omega \Big\{ \mathbf{u}_t\cdot\mathbf{v} + \mathcal{B}_p\cdot\nabla\mathbf{v} + \lambda\big(\mathbf{u}-\mathbf{g}\big)\cdot\mathbf{v}\Big\}\, dxdt = \int_0^T\int_\Omega \mu_{p,\lambda}\mathbf{u}\cdot\mathbf{v}\, dxdt \tag{5.14}$$

for any $\mathbf{v}\in[C^1(\overline{\Omega}_T)]^n$. Hence, $\mathbf{u}$ is a weak solution to (1.8)–(1.11).

Step 3: *Wrapping up.* We conclude the proof by showing the energy inequality (5.1). First, notice that (4.6) implies that

$$\int_\Omega \Big\{\frac{1}{p}|\nabla \mathbf{u}^\varepsilon(s)|^p + \frac{\lambda}{2}|\mathbf{u}^\varepsilon(s)-\mathbf{g}|^2\Big\}dx + \int_0^s \|\,\mathbf{u}_t^\varepsilon(t)\,\|_{L^2}^2\, dt \le J_{p,\lambda}^\varepsilon(\mathbf{u}_0) \quad \forall s\in[0,T]. \tag{5.15}$$

Then (5.1) follows from taking $\varepsilon\to 0$ in (5.15) and using Fatou's lemma and the lower semicontinuity of the L^p-norm with respect to L^p-weak convergence. The proof is complete. □

Remark 5.1. (a) We remark that it was proved in [15, 32, 41] that weak solutions to (1.8)–(1.11) are not unique in general.

(b) Although the above proof is carried out for any $\alpha>0$ in the definition of $b_p(\varepsilon)$ (cf. (1.15)), α should be chosen large enough so that the error due to the perturbation

term $b_p(\varepsilon)\Delta$ is much smaller than the error due to other regularization terms in numerical simulations.

(c) The existence result of Theorem 5.2 is established under the assumption $\mathbf{u}_0 \in W^{1,p*}(\Omega,\mathbf{R}^n)$ with $p* = \max\{p,2\}$. This condition can be weakened to $\mathbf{u}_0 \in W^{1,p}(\Omega,\mathbf{R}^n)$ in the case $1 < p < 2$ by a smoothing technique.

6. Passing to the limit as $\varepsilon \to 0$: The case $p = 1$. In this section, we consider the case $p = 1$ and establish the existence of global weak solutions for the 1-harmonic map heat flow (1.8)–(1.11) by passing to the limit as $\varepsilon \to 0$ in (4.2)–(4.4). There are two main difficulties which prevent one from repeating the analysis and techniques of the previous section. First, the compactness result of Lemma 2.3 no longer holds when $p = 1$. Second, since the sequence $\{\mathbf{u}^\varepsilon\}_{\varepsilon>0}$ is uniformly bounded only in $L^\infty((0,T);[W^{1,1}(\Omega)\cap L^\infty(\Omega)]^n)$, and $W^{1,1}(\Omega,\mathbf{R}^n)$ is not a reflexive Banach space, hence the limiting map $\mathbf{u}$ now belongs to $L^\infty((0,T);[BV(\Omega)\cap L^\infty(\Omega)]^n)$; i.e., $\mathbf{u}(t)$ is only a map of bounded variation. As expected, these two difficulties make the passage to the limit as $\varepsilon \to 0$ more difficult and delicate.

6.1. Technical tools and lemmas. In this subsection, we shall cite some technical tools and lemmas in order to develop a weak solution concept to be given in the next subsection for the 1-harmonic map heat flow. Specifically, we need the pairings $\mathcal{A}\cdot D\mathbf{v}$ and $\mathcal{A}\wedge D\mathbf{v}$ between a tensor field $\mathcal{A}$ and a BV-vector field $\mathbf{v}$ as a generalization of the pairing $\mathbf{b}\cdot Dv$ between a vector field $\mathbf{b}$ and a BV-function v developed in [4, 10]. The proofs of the key results in achieving this can be found in [23], where an extension of the theory for the 1-harmonic case studied in this section is developed for a general linear growth functional on the gradient.

We recall from [23] the space of *divergence-L^q tensors*

$$\mathcal{Y}(\Omega)_q := \{\mathcal{A}\in L^\infty(\Omega;\mathbf{R}^{n\times m});\ \mathrm{div}\mathcal{A}\in L^q(\Omega;\mathbf{R}^n)\}\qquad \text{for } 1\le q<\infty, \tag{6.1}$$

and that $\mathcal{A}\cdot D\mathbf{v}$ and $\mathcal{A}\wedge D\mathbf{v}$ are defined as follows.

Definition 6.1. *For any $\mathcal{A}\in\mathcal{Y}(\Omega)_1$ and $\mathbf{v}\in[BV(\Omega)\cap L^\infty(\Omega)]^n$, we define $\mathcal{A}\cdot D\mathbf{v}$ and $\mathcal{A}\wedge D\mathbf{v}$ to be the functionals on $C_0^\infty(\Omega)$ and $[C_0^\infty(\Omega)]^n$, respectively, by*

$$\langle \mathcal{A}\cdot D\mathbf{v},\psi\rangle := -\int_\Omega (\mathcal{A}^T\mathbf{v})\cdot\nabla\psi\,dx - \int_\Omega (\mathrm{div}\mathcal{A}\cdot\mathbf{v})\psi\,dx \quad \forall\psi\in C_0^\infty(\Omega), \tag{6.2}$$

$$\langle \mathcal{A}\wedge D\mathbf{v},\mathbf{w}\rangle := -\int_\Omega (\mathcal{A}\wedge\mathbf{v})\cdot\nabla\mathbf{w}\,dx - \int_\Omega (\mathrm{div}\mathcal{A}\wedge\mathbf{v})\cdot\mathbf{w}\,dx \quad \forall\mathbf{w}\in[C_0^\infty(\Omega)]^n, \tag{6.3}$$

where $\mathcal{A}^T$ stands for the matrix transpose of $\mathcal{A}$ and the notation (2.4) *is used in* (6.3).

We now list some properties of the pairings $\mathcal{A}\cdot D\mathbf{v}$ and $\mathcal{A}\wedge D\mathbf{v}$ and refer to section 2 of [23] for their proofs. The first lemma declares that $\mathcal{A}\cdot D\mathbf{v}$ and $\mathcal{A}\wedge D\mathbf{v}$ are Radon measures in Ω.

Lemma 6.2. *For any Borel set $E\subset\Omega$, there hold*

$$\big|\langle\mathcal{A}\cdot D\mathbf{v},\psi\rangle\big| \le \max_{x\in E}|\psi(x)|\cdot\|\,\mathcal{A}\,\|_{L^\infty(E,\mathbf{R}^{n\times m})}\cdot|D\mathbf{v}|(E)\quad\forall\psi\in C_0(E),$$

$$\big|\langle\mathcal{A}\wedge D\mathbf{v},\mathbf{w}\rangle\big| \le \max_{x\in E}|\mathbf{w}(x)|\cdot\|\,\mathcal{A}\,\|_{L^\infty(E,\mathbf{R}^{n\times m})}\cdot|D\mathbf{v}|(E)\quad\forall\mathbf{w}\in[C_0(E)]^n.$$

Hence, it follows from the Riesz theorem (cf. Theorem 1.54 *of* [3]*) that both functionals $\mathcal{A}\cdot D\mathbf{v}$ and $\mathcal{A}\wedge D\mathbf{v}$ are Radon measures in Ω.*

COROLLARY 6.3. *The measures $\mathcal{A}\cdot D\mathbf{v}$, $|\mathcal{A}\cdot D\mathbf{v}|$, $\mathcal{A}\wedge D\mathbf{v}$, and $|\mathcal{A}\wedge D\mathbf{v}|$ all are absolutely continuous with respect to the measure $|D\mathbf{v}|$ in Ω. Moreover, there hold inequalities*

$$\big|(\mathcal{A}\cdot D\mathbf{v})(E)\big| \le \big|\mathcal{A}\cdot D\mathbf{v}\big|(E) \le \|\,\mathcal{A}\,\|_{L^\infty(E',\mathbf{R}^{n\times m})}\cdot|D\mathbf{v}|(E), \tag{6.4}$$

$$\big|(\mathcal{A}\wedge D\mathbf{v})(E)\big| \le \big|\mathcal{A}\wedge D\mathbf{v}\big|(E) \le \|\,\mathcal{A}\,\|_{L^\infty(E',\mathbf{R}^{n\times m})}\cdot|D\mathbf{v}|(E) \tag{6.5}$$

for all Borel sets E and for all open sets E' such that $E\subset E'\subset\Omega$.

Hence, by the Radon–Nikodým theorem (cf. Theorem 1.28 *of* [3]*), there exist $|D\mathbf{v}|$-measurable functions $\Theta:=\Theta(\mathcal{A},D\mathbf{v},x):\Omega\to\mathbf{R}$ and $\Lambda:=\Lambda(\mathcal{A},D\mathbf{v},x):\Omega\to\mathbf{R}^n$ such that*

$$(\mathcal{A}\cdot D\mathbf{v})(E)=\int_E\Theta\,d|D\mathbf{v}| \quad \textit{and} \quad \|\,\Theta\,\|_{L^\infty(\Omega,|D\mathbf{v}|)}\le\|\,\mathcal{A}\,\|_{L^\infty(\Omega,\mathbf{R}^{n\times m})}, \tag{6.6}$$

$$(\mathcal{A}\wedge D\mathbf{v})(E)=\int_E\Lambda\,d|D\mathbf{v}| \quad \textit{and} \quad \|\,\Lambda\,\|_{L^\infty(\Omega,|D\mathbf{v}|)}\le\|\,\mathcal{A}\,\|_{L^\infty(\Omega,\mathbf{R}^{n\times m})} \tag{6.7}$$

for all Borel sets $E\subset\Omega$.

The second lemma declares that every $\mathcal{A}\in\mathcal{Y}(\Omega)_1$ has a well-behaved *traction* $\mathcal{A}\mathbf{n}$ on the boundary of a Lipschitz domain Ω.

LEMMA 6.4. *Let Ω be a bounded domain with a Lipschitz continuous boundary $\partial\Omega$ in $\mathbf{R}^m$. Then there exists a linear operator $\beta:\mathcal{Y}(\Omega)_1\to L^\infty(\partial\Omega;\mathbf{R}^n)$ such that*

$$\|\,\beta(\mathcal{A})\,\|_{L^\infty(\partial\Omega,\mathbf{R}^n)}\le\|\,\mathcal{A}\,\|_{L^\infty(\Omega,\mathbf{R}^{n\times m})}, \tag{6.8}$$

$$\langle\mathcal{A},\mathbf{v}\rangle_{\partial\Omega}=\int_{\partial\Omega}\beta(\mathcal{A})(x)\,\mathbf{v}(x)\,d\mathcal{H}^{m-1} \quad \forall\mathbf{v}\in[BV(\Omega)\cap L^\infty(\Omega)]^n, \tag{6.9}$$

$$\beta(\mathcal{A})(x)=\mathcal{A}(x)\mathbf{n}(x) \qquad \forall x\in\partial\Omega,\ \mathcal{A}\in C^1(\overline{\Omega},\mathbf{R}^{n\times m}). \tag{6.10}$$

Remark 6.1. Since $\beta(\mathcal{A})$ is a weakly defined traction of $\mathcal{A}$ on $\partial\Omega$, hence we shall use $\mathcal{A}\mathbf{n}$ to denote $\beta(\mathcal{A})$ in the rest of this section.

The third lemma declares that the following hold.

LEMMA 6.5. *Let Ω be a bounded domain with a Lipschitz continuous boundary $\partial\Omega$ in $\mathbf{R}^m$. Then for any $\mathcal{A}\in\mathcal{Y}(\Omega)_1$ and $\mathbf{v}\in[BV(\Omega)\cap L^\infty(\Omega)]^n$ there hold the identities*

$$\int_\Omega \mathrm{div}\mathcal{A}\cdot\mathbf{v}\,dx+(\mathcal{A}\cdot D\mathbf{v})(\Omega)=\int_{\partial\Omega}\mathcal{A}\mathbf{n}\cdot\mathbf{v}\,d\mathcal{H}^{m-1}, \tag{6.11}$$

$$\int_\Omega \mathrm{div}\mathcal{A}\wedge\mathbf{v}\,dx+(\mathcal{A}\wedge D\mathbf{v})(\Omega)=\int_{\partial\Omega}\mathcal{A}\mathbf{n}\wedge\mathbf{v}\,d\mathcal{H}^{m-1}. \tag{6.12}$$

Lemmas 6.6 and 6.7 state continuity results for the measure $\mathcal{A}\cdot D\mathbf{v}$ and $\mathcal{A}\wedge D\mathbf{v}$ with respect to $\mathcal{A}$ and $\mathbf{v}$, respectively.

LEMMA 6.6. *Let $\mathcal{A}_j,\mathcal{A}\in\mathcal{Y}(\Omega)_1$ and suppose that*

$$\mathcal{A}_j\longrightarrow\mathcal{A} \quad \textit{weakly}* \textit{ in } L^\infty(E),$$

$$\mathrm{div}\mathcal{A}_j\longrightarrow\mathrm{div}\mathcal{A} \quad \textit{weakly in } L^1(E)$$

for all open sets $E \subset\subset \Omega$. Then for all $\mathbf{v} \in [BV(\Omega) \cap L^\infty(\Omega)]^n$ the following hold:

$$\mathcal{A}_j \cdot D\mathbf{v} \longrightarrow \mathcal{A} \cdot D\mathbf{v} \quad \textit{weakly}* \textit{ in } \mathcal{M}(\Omega), \tag{6.13}$$

$$\mathcal{A}_j \wedge D\mathbf{v} \longrightarrow \mathcal{A} \wedge D\mathbf{v} \quad \textit{weakly}* \textit{ in } [\mathcal{M}(\Omega)]^n \tag{6.14}$$

and

$$\Theta(\mathcal{A}_j, D\mathbf{v}, \cdot) \longrightarrow \Theta(\mathcal{A}, D\mathbf{v}, \cdot) \quad \textit{weakly}* \textit{ in } L^\infty(E)\, \forall E \subset\subset \Omega, \tag{6.15}$$

$$\Lambda(\mathcal{A}_j, D\mathbf{v}, \cdot) \longrightarrow \Lambda(\mathcal{A}, D\mathbf{v}, \cdot) \quad \textit{weakly}* \textit{ in } [L^\infty(E)]^n\, \forall E \subset\subset \Omega. \tag{6.16}$$

Here "$E \subset\subset \Omega$" means that E is compactly contained in Ω; that is, $E \subset \overline{E} \subset \Omega$ and $\overline{E}$ is compact.

LEMMA 6.7. *Let $\mathcal{A} \in \mathcal{Y}(\Omega)_1$ and $\mathbf{v} \in [BV(\Omega) \cap L^\infty(\Omega)]^n$. Suppose that $\{\mathbf{v}_j\} \subset [C^\infty(\Omega) \cap BV(\Omega)]^n$ strictly converges to $\mathbf{v}$ (cf. Definition* 3.14 *of* [3]*). Then*

$$\mathcal{A} \cdot D\mathbf{v}_j \longrightarrow \mathcal{A} \cdot D\mathbf{v} \quad \textit{weakly}* \textit{ in } \mathcal{M}(\Omega), \tag{6.17}$$

$$\mathcal{A} \wedge D\mathbf{v}_j \longrightarrow \mathcal{A} \wedge D\mathbf{v} \quad \textit{weakly}* \textit{ in } [\mathcal{M}(\Omega)]^n. \tag{6.18}$$

Moreover,

$$\int_\Omega \mathcal{A} \cdot D\mathbf{v}_j \, dx \longrightarrow \int_\Omega \mathcal{A} \cdot D\mathbf{v}, \tag{6.19}$$

$$\int_\Omega \mathcal{A} \wedge D\mathbf{v}_j \, dx \longrightarrow \int_\Omega \mathcal{A} \wedge D\mathbf{v}. \tag{6.20}$$

Lemma 6.8 gives the precise representations for the density functions Θ and Λ defined in Corollary 6.3.

LEMMA 6.8. (i) *If $\mathcal{A} \in \mathcal{Y}(\Omega)_1 \cap C(\Omega, \mathbf{R}^{n\times m})$ and $\mathbf{v} \in [BV(\Omega) \cap L^\infty(\Omega)]^n$, then there hold*

$$\Theta(\mathcal{A}, D\mathbf{v}, x) = \mathcal{A}(x) \cdot \frac{D\mathbf{v}}{|D\mathbf{v}|}(x), \qquad \Lambda(\mathcal{A}, D\mathbf{v}, x) = \mathcal{A}(x) \wedge \frac{D\mathbf{v}}{|D\mathbf{v}|}(x) \tag{6.21}$$

for $|D\mathbf{v}|$ a.e. in Ω.

(ii) *If $\mathcal{A} \in \mathcal{Y}(\Omega)_1$ and $\mathbf{v} \in [BV(\Omega) \cap L^\infty(\Omega)]^n$, then there hold*

$$\Theta(\mathcal{A}, D\mathbf{v}, x) = \mathcal{A}(x) \cdot \frac{D\mathbf{v}}{|D\mathbf{v}|}(x), \qquad \Lambda(\mathcal{A}, D\mathbf{v}, x) = \mathcal{A}(x) \wedge \frac{D\mathbf{v}}{|D\mathbf{v}|}(x) \tag{6.22}$$

for $|D\mathbf{v}|^a$ a.e. in Ω, where $\frac{D\mathbf{v}}{|D\mathbf{v}|}$ denotes the density function of the measure $D\mathbf{v}$ with respect to the measure $|D\mathbf{v}|$, and $|D\mathbf{v}|^a$ denotes the absolute continuous part of the measure $|D\mathbf{v}|$ with respect to the Lebesgue measure $\mathcal{L}^n$.

Next, we recall from [23] the space of *divergence-measure tensors*

$$\mathcal{DT}(\Omega) := \left\{ \mathcal{A} \in L^\infty(\Omega, \mathbf{R}^{n\times m}); \operatorname{div}\mathcal{A} \in [\mathcal{M}(\Omega)]^n \right\} \tag{6.23}$$

and briefly discuss two of its important properties. First, as in the case of the space of the divergence-L^1 tensors $\mathcal{Y}(\Omega)_1$, Definition 6.1 is still valid for $\mathcal{A} \in \mathcal{DT}(\Omega)$ and $\mathbf{v} \in [BV(\Omega) \cap L^\infty(\Omega) \cap C(\Omega)]^n$ (cf. [4]). Second, there is a well-behaved traction $\mathcal{A}\mathbf{n}$ for every $\mathcal{A} \in \mathcal{DT}(\Omega)$.

LEMMA 6.9. *Let Ω be a bounded domain with Lipschitz continuous boundary $\partial\Omega$ in $\mathbf{R}^m$. Then there exists a linear operator $\alpha : \mathcal{DT}(\Omega) \to L^\infty(\partial\Omega; \mathbf{R}^n)$ such that*

$$\| \alpha(\mathcal{A}) \|_{L^\infty(\partial\Omega, \mathbf{R}^n)} \le \| \mathcal{A} \|_{L^\infty(\Omega, \mathbf{R}^{n\times m})}, \tag{6.24}$$

$$\langle \mathcal{A}, \mathbf{v} \rangle_{\partial\Omega} = \int_{\partial\Omega} \alpha(\mathcal{A})(x)\, \mathbf{v}(x)\, d\mathcal{H}^{m-1} \qquad \forall \mathbf{v} \in [BV(\Omega) \cap L^\infty(\Omega) \cap C(\Omega)]^n, \tag{6.25}$$

$$\alpha(\mathcal{A})(x) = \mathcal{A}(x)\mathbf{n}(x) \qquad \forall x \in \partial\Omega,\ \mathcal{A} \in C^1(\overline{\Omega}, \mathbf{R}^{n\times m}). \tag{6.26}$$

Moreover, for any $\mathcal{A} \in \mathcal{DT}(\Omega)$ and $\mathbf{v} \in [BV(\Omega) \cap L^\infty(\Omega) \cap C(\Omega)]^n$, let $\mathcal{A}\mathbf{n} := \alpha(\mathcal{A})$ on $\partial\Omega$. Then there hold the following Green's formulas:

$$(\mathrm{div}\mathcal{A} \cdot \mathbf{v})(\Omega) + (\mathcal{A} \cdot D\mathbf{v})(\Omega) = \int_{\partial\Omega} \mathcal{A}\mathbf{n} \cdot \mathbf{v}\, d\mathcal{H}^{m-1}, \tag{6.27}$$

$$(\mathrm{div}\mathcal{A} \wedge \mathbf{v})(\Omega) + (\mathcal{A} \wedge D\mathbf{v})(\Omega) = \int_{\partial\Omega} \mathcal{A}\mathbf{n} \wedge \mathbf{v}\, d\mathcal{H}^{m-1}. \tag{6.28}$$

Third, the following product rule holds.

LEMMA 6.10. *For any $\mathcal{A} \in \mathcal{DT}(\Omega)$ and $\mathbf{v} \in [BV(\Omega) \cap L^\infty(\Omega)]^n$, the identities*

$$\mathrm{div}(\mathcal{A}^T\mathbf{v}) = (\mathrm{div}\mathcal{A}) \cdot \overline{\mathbf{v}} + \overline{\mathcal{A} \cdot D\mathbf{v}}, \tag{6.29}$$

$$\mathrm{div}(\mathcal{A} \wedge \mathbf{v}) = (\mathrm{div}\mathcal{A}) \wedge \overline{\mathbf{v}} + \overline{\mathcal{A} \wedge D\mathbf{v}} \tag{6.30}$$

hold in the sense of Radon measures in Ω, where $\overline{\mathbf{v}}$ denotes the limit of a mollified sequence for $\mathbf{v}$ through a positive symmetric mollifier, and $\overline{\mathcal{A} \cdot D\mathbf{v}}$ (resp., $\overline{\mathcal{A} \wedge D\mathbf{v}}$) is a Radon measure which is absolutely continuous with respect to the measure $|D\mathbf{v}|$, and whose absolutely continuous part $(\overline{\mathcal{A} \cdot D\mathbf{v}})^a$ (resp., $(\overline{\mathcal{A} \wedge D\mathbf{v}})^a$) with respect to the Lebesgue measure $\mathcal{L}^m$ in Ω coincides with $\mathcal{A} \cdot (\nabla\mathbf{v})^a$ (resp., $\mathcal{A} \wedge (\nabla\mathbf{v})^a$) a.e. in Ω, that is, $(\overline{\mathcal{A} \cdot D\mathbf{v}})^a = \mathcal{A} \cdot (\nabla\mathbf{v})^a$ (resp., $(\overline{\mathcal{A} \wedge D\mathbf{v}})^a = \mathcal{A} \wedge (\nabla\mathbf{v})^a$) for $\mathcal{L}^m$ a.e. in Ω.

Remark 6.2. Equations (6.29) and (6.30) hold without all the overbars if either $\mathbf{v} \in [BV(\Omega) \cap L^\infty(\Omega) \cap C(\Omega)]^n$ or $\mathcal{A} \in \mathcal{Y}(\Omega)_1$.

Remark 6.3. It should be noted that the results of Lemmas 6.6–6.8 also hold for the tensor fields in $\mathcal{DT}(\Omega)$ and the vector fields in $[BV(\Omega) \cap L^\infty(\Omega) \cap C(\Omega)]^n$.

6.2. Existence of weak solutions of 1-harmonic map heat flow. Throughout the rest of this paper we let $B_1(\mathbf{R}^{n\times m})$ denote the unit ball in the Euclidean space $\mathbf{R}^{n\times m}$; that is,

$$B_1(\mathbf{R}^{n\times m}) = \Big\{ \mathcal{A} \in \mathbf{R}^{n\times m};\ |\mathcal{A}| := \Big(\sum_{j=1}^{m} \sum_{i=1}^{n} \mathcal{A}_{ij}^2 \Big)^{\frac{1}{2}} \le 1 \Big\}.$$

In addition, let

$$\sigma_\lambda^\varepsilon := \mu_{1,\lambda}^\varepsilon \mathbf{u}^\varepsilon, \qquad \mathcal{B}^\varepsilon := \frac{\nabla\mathbf{u}^\varepsilon}{|\nabla\mathbf{u}^\varepsilon|_\varepsilon}.$$

Hence $\mathcal{B}_1^\varepsilon = b_1(\varepsilon)\nabla\mathbf{u}^\varepsilon + \mathcal{B}^\varepsilon$ (cf. (4.11)).

We now give a definition of weak solutions to (1.8)–(1.11) in the case $p = 1$.

DEFINITION 6.11. *For $p = 1$, a map $\mathbf{u} : \Omega_T \to \mathbf{R}^n$ is called a global* weak *solution to* (1.8)–(1.11) *if there exists a tensor (or matrix-valued function) $\mathcal{B}$ such that*

(i) $\mathbf{u} \in L^\infty((0,T);[BV(\Omega)\cap L^\infty(\Omega)]^n)\cap H^1((0,T);L^2(\Omega,\mathbf{R}^n))$;
(ii) $|\mathbf{u}| = 1$ *for* $\mathcal{L}^{m+1}$ *a.e. in* Ω_T;
(iii) $\mathcal{B} \in L^\infty((0,T);L^\infty(\Omega,B_1(\mathbf{R}^{n\times m})))\cap L^2((0,T);\mathcal{DT}(\Omega))$;
(iv) $\mathbf{u}$ *and* $\mathcal{B}$ *satisfy* $\mathcal{B}\wedge\mathbf{u}\in L^2((0,T);\mathcal{Y}(\Omega)_2)$, $\mathcal{B}^T\mathbf{u}=0$, *and* $\mathcal{B}\cdot(D\mathbf{u})^a = |D\mathbf{u}|^a$ *for* $\mathcal{L}^{m+1}$ *a.e. in* Ω_T;
(v) $\mathcal{B}\mathbf{n} = 0$ *on* $\partial\Omega_T$ *in the sense of Lemma* 6.9;
(vi) *there holds the identity*

$$\int_0^T\int_\Omega \Big\{\mathbf{u}_t\cdot\mathbf{v} + \mathcal{B}\cdot\nabla\mathbf{v} + \lambda(\mathbf{u}-\mathbf{g})\cdot\mathbf{v}\Big\}\,dxdt = \int_0^T\Big(\int_\Omega \mathbf{v}\,d\sigma_\lambda\Big)\,dt$$

for any $\mathbf{v}\in[C^1(\overline{\Omega}_T)]^n$, *where* σ_λ *denotes the vector-valued Radon measure*

$$\sigma_\lambda = (\mathcal{B}\wedge\mathbf{u})\wedge D\mathbf{u} + \lambda\,(1-\mathbf{g}\cdot\mathbf{u})\,\mathbf{u}. \tag{6.31}$$

Moreover, the Radon measure $(\mathcal{B}\wedge\mathbf{u})\wedge D\mathbf{u}$ *is absolutely continuous with respect to the measure* $|D\mathbf{u}|$, *and for* $\mathcal{L}^1$ *a.e.* $t\in(0,T)$ *there exists a function* $\Phi(\mathcal{B},D\mathbf{u},x,t):\Omega\to\mathbf{R}$ *such that*

$$((\mathcal{B}\wedge\mathbf{u})\wedge D\mathbf{u})(E) = \int_E \Phi(\mathcal{B}\wedge\mathbf{u},D\mathbf{u},x,t)\,d|D\mathbf{u}| \quad \textit{for all Borel sets } E\subset\Omega,$$

$$\|\,\Phi(t)\,\|_{L^\infty(\Omega,|D\mathbf{u}|)} \le \|\,\mathcal{B}(t)\,\|_{L^\infty(\Omega,\mathbf{R}^{n\times m})} \quad \textit{for } \mathcal{L}^1 \textit{ a.e. } t\in(0,T),$$

$$\Phi(\mathcal{B}\wedge\mathbf{u},D\mathbf{u},x,t) = (\mathcal{B}(t)\wedge\mathbf{u})\wedge\frac{D\mathbf{u}}{|D\mathbf{u}|} \quad \textit{for } |D\mathbf{u}|^a \textit{ a.e. in } \Omega,\ \mathcal{L}^1 \textit{ a.e. } t\in(0,T),$$

$$((\mathcal{B}\wedge\mathbf{u})\wedge D\mathbf{u})^a = (\mathcal{B}\wedge\mathbf{u})\wedge(D\mathbf{u})^a = |D\mathbf{u}|^a\mathbf{u} \quad \textit{for } \mathcal{L}^{m+1} \textit{ a.e. in } \Omega_T.$$

Remark 6.4. (a) The tensor field $\mathcal{B}$ extends $\frac{D\mathbf{u}}{|D\mathbf{u}|}$ as a calibration across the gulfs.

(b) If $\mathbf{u}(t)\in W^{1,1}_{\rm loc}(\Omega)$ and $\mathcal{B}(t)\in\mathcal{Y}(\Omega)_1$, using (ii) and (iv) and the identity $(\mathbf{a}\wedge\mathbf{b})\wedge\mathbf{c} = (\mathbf{a}\cdot\mathbf{c})\mathbf{b} - (\mathbf{b}\cdot\mathbf{c})\mathbf{a}$ we get that $(\mathcal{B}\wedge\mathbf{u})\wedge D\mathbf{u} = |\nabla\mathbf{u}|\,\mathbf{u}$. Hence, (6.31) can be rewritten as

$$\sigma_\lambda = |\nabla\mathbf{u}|\,\mathbf{u} + \lambda\,(1-\mathbf{g}\cdot\mathbf{u})\,\mathbf{u}. \tag{6.32}$$

Thus, (6.31) is a weak form of (6.32) since $|D\mathbf{u}|\,\mathbf{u}$ may not be defined for $\mathbf{u}(t)\in[BV(\Omega)\cap L^\infty(\Omega)]^n$ and $\mathcal{B}(t)\in\mathcal{DT}(\Omega)$.

Our main result of this section is the following existence theorem.

THEOREM 6.12. *Let* $p=1$, *and suppose that* $\mathbf{u}_0\in H^1(\Omega,\mathbf{R}^n)$ *and* $\mathbf{g}\in L^2(\Omega,\mathbf{R}^n)$ *with* $|\mathbf{u}_0| = 1$ *and* $|\mathbf{g}|\le 1$ *a.e. in* Ω. *Then problem* (1.8)–(1.11) *has a global weak solution* $\mathbf{u}$ *in the sense of Definition* 6.11. *Moreover,* $\mathbf{u}$ *satisfies the energy inequality*

$$I_\lambda(\mathbf{u}(s)) + \int_0^s \|\,\mathbf{u}_t(t)\,\|_{L^2}^2\,dt \le I_\lambda(\mathbf{u}_0) \qquad \textit{for a.e. } s\in[0,T], \tag{6.33}$$

where

$$I_\lambda(\mathbf{u}) := |D\mathbf{u}|(\Omega) + \frac{\lambda}{2}\int_\Omega |\mathbf{u}-\mathbf{g}|^2\,dx. \tag{6.34}$$

Proof. The proof is divided into four steps.

Step 1: *Extracting a convergent subsequence and passing to the limit.* Let $\mathbf{u}^\varepsilon$ denote the solution of (4.1)–(4.4) constructed in Theorem 4.1. It follows from (4.1), (4.6), (4.7), and (4.8) that $\{\mathbf{u}^\varepsilon\}_{\varepsilon>0}$ is uniformly bounded in $L^\infty((0,T);[W^{1,1}(\Omega)\cap L^\infty(\Omega)]^n)\cap H^1((0,T);L^2(\Omega,\mathbf{R}^n))$, $\{\mathcal{B}^\varepsilon\}_{\varepsilon>0}$ in $L^\infty((0,T);L^\infty(\Omega,B_1(\mathbf{R}^{n\times m})))$, and $\{\mathrm{div}\mathcal{B}_1^\varepsilon\}_{\varepsilon>0}$ in $L^2((0,T);L^1(\Omega;\mathbf{R}^n))$, and $\{\sigma_\lambda^\varepsilon\}_{\varepsilon>0}$ is uniformly bounded in $L^\infty((0,T);L^1(\Omega,\mathbf{R}^n))$. Since $L^1(\Omega)\subset\mathcal{M}(\Omega)$ and $W^{1,1}(\Omega)\subset BV(\Omega)$, by the weak compactness of $\mathcal{M}(\Omega)$ and $BV(\Omega)$ (cf. [3]) we have that there exist subsequences of $\{\mathbf{u}^\varepsilon\}_{\varepsilon>0}$, $\{\mathcal{B}^\varepsilon\}_{\varepsilon>0}$, $\{\mathcal{B}_1^\varepsilon\}_{\varepsilon>0}$, and $\{\sigma_\lambda^\varepsilon\}_{\varepsilon>0}$ (still denoted by the same notation), respectively, and maps $\mathbf{u}\in L^\infty((0,T);[BV(\Omega)\cap L^\infty(\Omega)]^n)\cap H^1((0,T);L^2(\Omega,\mathbf{R}^n))$, $\mathcal{B}\in L^\infty((0,T);L^\infty(\Omega,B_1(\mathbf{R}^{n\times m})))$, $\nu\in L^2((0,T);[\mathcal{M}(\Omega)]^n)$, and $\sigma_\lambda\in L^\infty((0,T);[\mathcal{M}(\Omega)]^n)$, respectively, such that as $\varepsilon\to 0$

$$\mathbf{u}^\varepsilon \longrightarrow \mathbf{u} \qquad \text{weakly* in } L^\infty((0,T);[BV(\Omega)\cap L^\infty(\Omega)]^n), \tag{6.35}$$

$$\text{strongly in } L^2((0,T);L^1(\Omega,\mathbf{R}^n)) \tag{6.36}$$

$$\text{a.e. in } \Omega_T, \tag{6.37}$$

$$b_1(\varepsilon)\nabla\mathbf{u}^\varepsilon \longrightarrow 0 \qquad \text{weakly in } L^2((0,T);L^2(\Omega,\mathbf{R}^{n\times m})), \tag{6.38}$$

$$\mathbf{u}_t^\varepsilon \longrightarrow \mathbf{u}_t \qquad \text{weakly in } L^2((0,T);L^2(\Omega,\mathbf{R}^n)), \tag{6.39}$$

$$\mathrm{div}\mathcal{B}_1^\varepsilon \longrightarrow \nu \qquad \text{weakly* in } L^2((0,T);[\mathcal{M}(\Omega)]^n), \tag{6.40}$$

$$\mathcal{B}^\varepsilon \longrightarrow \mathcal{B} \qquad \text{weakly* in } L^\infty((0,T);L^\infty(\Omega,B_1(\mathbf{R}^{n\times m}))), \tag{6.41}$$

$$\sigma_\lambda^\varepsilon \longrightarrow \sigma_\lambda \qquad \text{weakly* in } L^\infty((0,T);[\mathcal{M}(\Omega)]^n). \tag{6.42}$$

It follows immediately from (6.37) and (4.18) that

$$|\mathbf{u}| = 1 \qquad \text{a.e. in } \Omega_T, \tag{6.43}$$

and an application of the Lebesgue-dominated convergence theorem yields that

$$\mathbf{u}^\varepsilon \overset{\varepsilon\searrow 0}{\longrightarrow} \mathbf{u} \qquad \text{strongly in } L^r((0,T);L^r(\Omega,\mathbf{R}^n)) \quad \forall r\in[1,\infty). \tag{6.44}$$

Now, taking $\varepsilon\to 0$ in (4.24), and (4.29) (with $p=1$) we have for any $\mathbf{v}\in[C^1(\overline{\Omega}_T)]^n$ and $\mathbf{w}\in L^\infty((0,T);H^1(\Omega;\mathbf{R}^n))$

$$\int_0^T\int_\Omega \Big\{\big(\mathbf{u}_t\wedge\mathbf{u}\big)\cdot\mathbf{w}+\big(\mathcal{B}\wedge\mathbf{u}\big)\cdot\nabla\mathbf{w}-\lambda\big(\mathbf{g}\wedge\mathbf{u}\big)\cdot\mathbf{w}\Big\}\,dxdt=0, \tag{6.45}$$

$$\int_0^T\int_\Omega \Big\{\mathbf{u}_t\cdot\mathbf{v}+\mathcal{B}\cdot\nabla\mathbf{v}+\lambda\big(\mathbf{u}-\mathbf{g}\big)\cdot\mathbf{v}\Big\}\,dxdt=\int_0^T\Big(\int_\Omega \mathbf{v}\,d\sigma_\lambda\Big)\,dt. \tag{6.46}$$

Step 2: *Identifying ν and σ_λ.* First, it follows from the identity $(\mathcal{B}^\varepsilon)^T\mathbf{u}^\varepsilon=0$ (cf. (4.25)), (6.41), and (6.44) that

$$\mathcal{B}^T\mathbf{u}=0 \qquad \text{a.e. in } \Omega_T. \tag{6.47}$$

Second, for any $\mathbf{v}\in[C^1(\overline{\Omega}_T)]^n$, it follows from (6.38), (6.40), and (6.41) that

$$\int_{\Omega_T}\mathbf{v}\cdot d\nu=\lim_{\varepsilon\to0}\int_{\Omega_T}\mathbf{v}\cdot\mathrm{div}\,\mathcal{B}_1^\varepsilon\,dxdt=-\lim_{\varepsilon\to0}\int_{\Omega_T}\nabla\mathbf{v}\cdot\mathcal{B}_1^\varepsilon\,dxdt=-\int_{\Omega_T}\nabla\mathbf{v}\cdot\mathcal{B}\,dxdt.$$

Hence, div$\mathcal{B}$ exists and

$$\operatorname{div}\mathcal{B} = \nu, \qquad \text{and therefore} \quad \mathcal{B} \in L^2((0,T);\mathcal{DT}(\Omega)). \tag{6.48}$$

It then follows from (6.11) that

$$\mathcal{B}\mathbf{n} = 0 \qquad \text{on } \partial\Omega_T.$$

Third, notice that (6.45) immediately implies that

$$\operatorname{div}(\mathcal{B}\wedge\mathbf{u}) \in L^2(\Omega_T,\mathbf{R}^n), \qquad \text{and hence} \quad \mathcal{B}\wedge\mathbf{u} \in L^2((0,T);\mathcal{Y}(\Omega)_2). \tag{6.49}$$

Let $\{\mathbf{u}_\rho\}$ denote the smooth approximation sequence of $\mathbf{u}$ as constructed in Theorem 3.9 of [3]. For any $\mathbf{v} \in [C_0^1(\Omega_T)]^n$, setting $\mathbf{w} = \mathbf{u}_\rho \wedge \mathbf{v}$ in (6.45) we get

$$\int_0^T \int_\Omega \Big\{ \mathbf{u}_t \cdot \big(\mathbf{u}\wedge(\mathbf{u}_\rho\wedge\mathbf{v})\big) + \big(\mathcal{B}\wedge\mathbf{u}\big)\cdot\nabla(\mathbf{u}_\rho\wedge\mathbf{v}) - \lambda\,\mathbf{g}\cdot\big(\mathbf{u}\wedge(\mathbf{u}_\rho\wedge\mathbf{v})\big)\Big\}\,dxdt = 0. \tag{6.50}$$

It follows from (6.43), the convergence property of $\mathbf{u}_\rho$ (cf. [3]), and the identity $\mathbf{u}\wedge(\mathbf{u}\wedge\mathbf{v}) = \mathbf{u}(\mathbf{u}\cdot\mathbf{v}) - \mathbf{v}$ that

$$\begin{aligned}
\lim_{\rho\to 0}\int_\Omega \mathbf{u}_t\cdot\big(\mathbf{u}\wedge(\mathbf{u}_\rho\wedge\mathbf{v})\big)\,dx &= \int_\Omega \mathbf{u}_t\cdot\big(\mathbf{u}\wedge(\mathbf{u}\wedge\mathbf{v})\big)\,dx = \int_\Omega \mathbf{u}_t\cdot\big(\mathbf{u}(\mathbf{u}\cdot\mathbf{v})-\mathbf{v}\big)\,dx \\
&= -\int_\Omega \mathbf{u}_t\cdot\mathbf{v}\,dx, \\
\lim_{\rho\to 0}\int_\Omega \mathbf{g}\cdot\big(\mathbf{u}\wedge(\mathbf{u}_\rho\wedge\mathbf{v})\big)\,dx &= \int_\Omega \mathbf{g}\cdot\big(\mathbf{u}\wedge(\mathbf{u}\wedge\mathbf{v})\big)\,dx \\
&= \int_\Omega \Big\{\big(\mathbf{u}-\mathbf{g}\big)\cdot\mathbf{v} - (1-\mathbf{g}\cdot\mathbf{u})\mathbf{u}\cdot\mathbf{v}\Big\}dx.
\end{aligned}$$

It follows from Theorem 3.9 of [3], Lemma 6.7, and the identity

$$\int_\Omega \big(\mathcal{B}\wedge\mathbf{u}\big)\cdot\nabla(\mathbf{u}_\rho\wedge\mathbf{v})\,dx = \int_\Omega \Big\{\big((\mathcal{B}\wedge\mathbf{u})\wedge\mathbf{u}_\rho\big)\cdot\nabla\mathbf{v} + \big((\mathcal{B}\wedge\mathbf{u})\wedge\nabla\mathbf{u}_\rho\big)\cdot\mathbf{v}\Big\}dx$$

that

$$\begin{aligned}
\lim_{\rho\to 0}\int_\Omega \big(\mathcal{B}\wedge\mathbf{u}\big)\cdot\nabla(\mathbf{u}_\rho\wedge\mathbf{v})\,dx &= \int_\Omega \big((\mathcal{B}\wedge\mathbf{u})\wedge\mathbf{u}\big)\cdot\nabla\mathbf{v}\,dx + \langle(\mathcal{B}\wedge\mathbf{u})\wedge D\mathbf{u},\mathbf{v}\rangle \\
&= -\int_\Omega \mathcal{B}\cdot\nabla\mathbf{v}\,dx + \langle(\mathcal{B}\wedge\mathbf{u})\wedge D\mathbf{u},\mathbf{v}\rangle.
\end{aligned}$$

Here we have used the fact that $(\mathcal{B}\wedge\mathbf{u})\wedge\mathbf{u} = -\mathcal{B}$ in light of (6.43) and (6.47), and the measure $(\mathcal{B}\wedge\mathbf{u})\wedge D\mathbf{u}$ is defined by (6.3) with $\mathcal{A} = \mathcal{B}\wedge\mathbf{u}$ and $\mathbf{v} = \mathbf{u}$.

Finally, substituting the above three equations into (6.50) and multiplying the equation by (-1) we get

$$\begin{aligned}
&\int_0^T\int_\Omega \Big\{\mathbf{u}_t\cdot\mathbf{v} + \mathcal{B}\cdot\nabla\mathbf{v} + \lambda\,\big(\mathbf{u}-\mathbf{g}\big)\cdot\mathbf{v}\Big\}\,dxdt \\
&\qquad = \int_0^T \big\langle(\mathcal{B}\wedge\mathbf{u})\wedge D\mathbf{u} + \lambda(1-\mathbf{g}\cdot\mathbf{u})\mathbf{u},\mathbf{v}\big\rangle\,dt
\end{aligned} \tag{6.51}$$

for any $\mathbf{v} \in [C_0^1(\Omega_T)]^n$. This and (6.46) imply that

$$\sigma_\lambda = (\mathcal{B} \wedge \mathbf{u}) \wedge D\mathbf{u} + \lambda(1 - \mathbf{g} \cdot \mathbf{u})\mathbf{u}.$$

Step 3: *Identifying* $\mathcal{B}$. First, since $\{\mathcal{B}^\varepsilon \cdot \nabla \mathbf{u}^\varepsilon\}$ is uniformly bounded in $L^2((0,T); L^1(\Omega))$, then there exists a subsequence (still denoted by the same notation) and $\mu \in L^2((0,T); \mathcal{M}(\Omega))$ such that

$$\mathcal{B}^\varepsilon \cdot \nabla \mathbf{u}^\varepsilon \longrightarrow \mu \quad \text{weakly* in } L^2((0,T); \mathcal{M}(\Omega)). \tag{6.52}$$

Let $\mathbf{u}_\rho^\varepsilon$ and $\mathbf{u}_\rho$ denote mollified sequences for $\mathbf{u}^\varepsilon$ and $\mathbf{u}$, respectively, through a positive symmetric mollifier. For any open set $E \subset \Omega$ we have

$$\begin{aligned}
\lim_{\varepsilon\to 0} \int_0^T \int_E \mathcal{B}^\varepsilon \cdot \nabla \mathbf{u}^\varepsilon \, dxdt &= \lim_{\varepsilon\to 0} \lim_{\rho\to 0} \int_0^T \int_E \mathcal{B}^\varepsilon \cdot D\mathbf{u}_\rho^\varepsilon \, dxdt \\
&= \lim_{\rho\to 0} \lim_{\varepsilon\to 0} \int_0^T \int_E \mathcal{B}^\varepsilon \cdot D\mathbf{u}_\rho^\varepsilon \, dxdt \\
&= \lim_{\rho\to 0} \int_0^T (\mathcal{B} \cdot D\mathbf{u}_\rho)(E)\, dt \quad \text{(by (6.41), (6.44))} \\
&= \int_0^T \overline{\mathcal{B} \cdot D\mathbf{u}}\,(E)\, dt \quad \text{(by (6.29))},
\end{aligned} \tag{6.53}$$

where $\overline{\mathcal{B} \cdot D\mathbf{u}}$ is defined in Lemma 6.10.

Hence, it follows from (6.52), (6.53), and Lemma 6.10 that

$$\mu = \overline{\mathcal{B} \cdot D\mathbf{u}} << |D\mathbf{u}|. \tag{6.54}$$

We refer the reader to Definition 1.24 of [3] for the notation "<<."

On the other hand, a direct calculation yields that

$$\mathcal{B}^\varepsilon \cdot \nabla \mathbf{u}^\varepsilon = \frac{|\nabla \mathbf{u}^\varepsilon|^2}{|\nabla \mathbf{u}^\varepsilon|_\varepsilon} \geq |\nabla \mathbf{u}^\varepsilon|_\varepsilon - \varepsilon.$$

Setting $\varepsilon \to 0$ and by the lower semicontinuity of the BV-norm (cf. [3]) we obtain

$$\mu >> |D\mathbf{u}|,$$

which together with (6.54) and Lemma 6.10 yield that

$$|D\mathbf{u}| = \mu = \overline{\mathcal{B} \cdot D\mathbf{u}}, \tag{6.55}$$

and hence

$$|D\mathbf{u}|^a = (\overline{\mathcal{B} \cdot D\mathbf{u}})^a = \mathcal{B} \cdot (D\mathbf{u})^a. \tag{6.56}$$

Finally, we note that all other properties of $\mathcal{B}$ and the measure $(\mathcal{B} \wedge \mathbf{u}) \wedge D\mathbf{u}$ listed in (vi) of Definition 6.11 are immediate consequences of (6.55) and Lemmas 6.2, 6.4–6.10, and Corollary 6.3.

Step 4: *Finishing up.* We now conclude the proof of the theorem by showing the energy inequality (6.33). First, notice that (4.6) implies that for a.e. $s \in [0,T]$

$$I_\lambda(\mathbf{u}^\varepsilon(s)) + \int_0^s \| \mathbf{u}_t^\varepsilon(t) \|_{L^2}^2 \, dt \le J_{1,\lambda}^\varepsilon(\mathbf{u}_0) \tag{6.57}$$

$$\le \frac{b_1(\varepsilon)}{2} \| \nabla \mathbf{u}_0 \|_{L^2}^2 + a_1(\varepsilon)|\Omega| + I_\lambda(\mathbf{u}_0).$$

Then (6.33) follows from taking $\varepsilon \to 0$ in (6.57), using the lower semicontinuity of the BV-seminorm and the L^2-norm with respect to L^2-weak convergence. The proof is complete. ∎

Remark 6.5. (a) Since weak solutions to (1.8)–(1.11) are not unique in general for $1 < p < \infty$ (cf. [15, 32, 41]), we expect that this nonuniqueness also holds for the case $p = 1$.

(b) The existence result in Theorem 6.12 is proved under the assumption $\mathbf{u}_0 \in H^1(\Omega, \mathbf{R}^n)$. This assumption can be weakened to $\mathbf{u}_0 \in [BV(\Omega) \cap L^\infty(\Omega)]^n$ using a smoothing technique.

7. Fully discrete finite element approximations.

7.1. Formulation of fully discrete finite element methods. For ease of exposition, we assume Ω is a polytope in this section. Let $\mathcal{T}_h$ be a quasi-uniform "triangulation" of the domain Ω of mesh size $0 < h < 1$ and $\overline{\Omega} = \bigcup_{K \in \mathcal{T}_h} \overline{K}$ ($K \in \mathcal{T}_h$ are tetrahedrons in the case $m = 3$). Let $J_\tau := \{t_k\}_{k=0}^L$ be a uniform partition of $[0,T]$ with mesh size $\tau := \frac{T}{L}$, and $\partial_t \mathbf{v}^k := (\mathbf{v}^k - \mathbf{v}^{k-1})/\tau$. For an integer $r \ge 1$, let $P_r(K)$ denote the space of polynomials of degree less than or equal to r on K. We introduce the finite element space

$$\mathbf{V}^h = \big\{ \mathbf{v}_h \in C(\overline{\Omega}, \mathbf{R}^n) \cap H^1(\Omega, \mathbf{R}^n);\ \mathbf{v}_h|_K \in [P_r(K)]^n \, \forall K \in \mathcal{T}_h \big\}.$$

Notice that the density function F defined in (1.17) is not a convex function. On the other hand, there exist two convex functions W_+ and W_- such that

$$F(\mathbf{v}) = W_+(\mathbf{v}) - W_-(\mathbf{v}). \tag{7.1}$$

One such an example is $W_+(\mathbf{v}) = \frac{|\mathbf{v}|^4}{4}$ and $W_-(\mathbf{v}) = \frac{|\mathbf{v}|^2}{2} - \frac{1}{4}$. Clearly, the above decomposition is not unique.

We are now ready to introduce our fully discrete finite element discretizations for the initial boundary value problem (1.20)–(1.22). Find $\mathbf{u}_h^k \in \mathbf{V}^h$ for $k = 1, 2, \ldots, L$ such that

$$\int_\Omega \Big\{ \partial_t \mathbf{u}_h^k \cdot \mathbf{v}_h + \mathcal{B}_h^k \cdot \nabla \mathbf{v}_h + \lambda \left(\mathbf{u}_h^k - \mathbf{g} \right) \cdot \mathbf{v}_h + \frac{1}{\delta} W_+'(\mathbf{u}_h^k) \cdot \mathbf{v}_h \Big\} \, dx \tag{7.2}$$

$$= \frac{1}{\delta} \int_\Omega W_-'(\mathbf{u}_h^{k-1}) \cdot \mathbf{v}_h \, dx \qquad \forall \mathbf{v} \in \mathbf{V}^h,$$

where

$$\mathcal{B}_h^k = \big[\, b_p(\varepsilon) + |\nabla \mathbf{u}_h^k|_\varepsilon^{p-2} \, \big] \nabla \mathbf{u}_h^k, \tag{7.3}$$

with some starting value $\mathbf{u}_h^0 \in \mathbf{V}^h$ to be specified later. Note that for notational brevity we have omitted the indices ε, δ, and p on $\mathbf{u}_h^k$ and $\mathcal{B}_h^k$

For each k, (7.2) is a nonlinear equation in $\mathbf{u}_h^k$. Hence, the above numerical method is an implicit scheme, and its well-posedness is ensured by the following theorem.

THEOREM 7.1. *For each fixed $k \geq 1$, suppose that $\mathbf{u}_h^{k-1} \in \mathbf{V}^h$ is known. Then there exists a unique solution $\mathbf{u}_h^k \in \mathbf{V}^h$ to* (7.2)–(7.3). *Moreover, $\{\mathbf{u}_h^k\}_{k=0}^L$ satisfies the following energy estimate:*

$$\frac{\tau}{2}\sum_{k=1}^{\ell} \| \partial_t \mathbf{u}_h^k \|_{L^2}^2 + J_{p,\lambda}^{\varepsilon,\delta}(\mathbf{u}_h^k) \leq J_{p,\lambda}^{\varepsilon,\delta}(\mathbf{u}_h^0) \qquad \text{for } 1 \leq \ell \leq L. \tag{7.4}$$

Here $J_{p,\lambda}^{\varepsilon,\delta}$ is defined by (1.19).

Proof. For each fixed $k \geq 1$, it is easy to check that (7.2)–(7.3) is the Euler–Lagrange equation of the following functional over $\mathbf{V}^h$:

$$G_k(\mathbf{v}) := \int_\Omega \Big\{ \frac{1}{2\tau}|\mathbf{v} - \mathbf{u}_h^{k-1}|^2 + \frac{b_p(\varepsilon)}{2}\,|\nabla \mathbf{v}|^2 + \frac{1}{p}\,|\nabla \mathbf{v}|_\varepsilon^p + \frac{\lambda}{2}|\mathbf{v} - \mathbf{g}|^2 + \frac{1}{\delta} W_+(\mathbf{v}) \Big\}\, dx - \frac{1}{\delta}\int_\Omega W_-'(\mathbf{u}_h^{k-1}) \cdot \mathbf{v}\, dx. \tag{7.5}$$

Since G_k is a convex, coercive, and differentiable functional, then it has a unique minimizer $\mathbf{u}_h^k \in \mathbf{V}^h$ (cf. [47]), and hence (7.2)–(7.3) has a unique solution.

Since $\mathbf{u}_h^k$ is the minimizer of G_k over $\mathbf{V}^h$, we have that

$$G_k(\mathbf{u}_h^k) \leq G_k(\mathbf{u}_h^{k-1}). \tag{7.6}$$

It follows from the convexity of W_- that

$$W_-'(\mathbf{u}_h^{k-1})\big(\mathbf{u}_h^k - \mathbf{u}_h^{k-1}\big) \leq W_-(\mathbf{u}_h^k) - W_-(\mathbf{u}_h^{k-1}).$$

This and (7.6) imply that

$$\frac{1}{2}\| \partial_t \mathbf{u}_h^k \|_{L^2}^2 + \frac{J_{p,\lambda}^{\varepsilon,\delta}(\mathbf{u}_h^k) - J_{p,\lambda}^{\varepsilon,\delta}(\mathbf{u}_h^{k-1})}{\tau} \leq 0.$$

The bound (7.4) then follows from applying the summation operator $\tau \sum_{k=1}^{\ell}$ ($1 \leq \ell \leq L$) to the last inequality. Hence the proof is complete. ∎

7.2. Convergence analysis. The goal of this subsection is to show that the numerical solution of (7.2)–(7.3) converges to the unique weak solution of (1.20)–(1.22) as $h, \tau \to 0$. There are two approaches to reach this goal. The first approach assumes the existence of the solution of (1.20)–(1.22), which in fact has been proved in Theorem 3.7, and then proves that $\mathbf{u}_h^k$ converges to that solution. The other approach shows the convergence *without* assuming the existence of the solution of (1.20)–(1.22). This can be done by applying the energy method and compactness argument used in the proof of Theorem 3.7 to the finite element solution $\{\mathbf{u}_h^k\}$. In the following, we shall go with the latter approach since this will also provide an alternative proof for Theorem 3.7 as alluded to in Remark 3.1(c).

For the fully discrete finite element solution $\{\mathbf{u}_h^k\}$, we define its linear interpolation in t as follows:

$$\mathbf{U}^{\varepsilon,\delta,h,\tau}(\cdot,t) := \frac{t - t_{k-1}}{\tau}\mathbf{u}_h^k(\cdot) + \frac{t_k - t}{\tau}\mathbf{u}_h^{k-1}(\cdot) \quad \forall\, t \in [t_{k-1}, t_k], \quad 1 \leq k \leq L. \tag{7.7}$$

Clearly, $\mathbf{U}^{\varepsilon,\delta,h,\tau}$ is continuous in both x and t.

The main result of this section is the following convergence theorem.

THEOREM 7.2. *For $1 \le p < \infty$, suppose that $\mathbf{u}_0 \in W^{1,p^*}(\Omega,\mathbf{R}^n)$ with $p^* = \max\{2,p\}$, $|\mathbf{u}_0| = 1$, and $|\mathbf{g}| \le 1$ in Ω. For each pair of positive numbers (ε,δ), let $\mathbf{U}^{\varepsilon,\delta,h,\tau}$ be defined by* (7.7)*. Then there exists $\mathbf{u}^{\varepsilon,\delta} \in L^\infty(\Omega_T)$ such that*

$$\lim_{h,\tau\to 0} \| \mathbf{u}^{\varepsilon,\delta} - \mathbf{U}^{\varepsilon,\delta,h,\tau} \|_{L^q(\Omega_T)} = 0 \qquad \forall q \in [1,\infty), \tag{7.8}$$

provided that

$$\lim_{h\to 0} \| \mathbf{u}_0 - \mathbf{u}_h^0 \|_{W^{1,p^*}(\Omega)} = 0.$$

Moreover, $\mathbf{u}^{\varepsilon,\delta}$ solves (1.20)–(1.22) *in the sense of Definition* 3.1.

Proof. The proof follows along the same lines as that of Theorem 1.5 of [12], where the convergence of a general Galerkin approximation was proved for the case $p \ge 2$. Since the finite element approximation is a special Galerkin approximation, the proof of Theorem 1.5 of [12] can easily be adapted to the finite element approximation $\mathbf{U}^{\varepsilon,\delta,h,\tau}$ for $p \ge 2$ thanks to the discrete energy estimate (7.4) and the facts that $\partial_t \mathbf{u}_h^k = \mathbf{U}_t^{\varepsilon,\delta,h,\tau}$ and $\tau \sum_{k=1}^L \| \partial_t \mathbf{u}_h^k \|_{L^2}^2 = \| \mathbf{U}_t^{\varepsilon,\delta,h,\tau} \|_{L^2(L^2)}$.

Since the operator $-\Delta_p^\varepsilon$ is *uniformly* elliptic, as a result, the compactness of Lemma 2.2 holds not only for $p \ge 2$ but also for $1 \le p < 2$. In addition, note that $\mathbf{U}^{\varepsilon,\delta,h,\tau}$ is uniformly (in h and τ) bounded in $L^\infty((0,T);H^1(\Omega,\mathbf{R}^n))$ for $1 \le p < 2$. Hence, the proof of Theorem 1.5 of [12] can be adapted with slight modifications to prove (7.8) for the case $1 \le p < 2$. ∎

Remark 7.1. Several practical choices of $\mathbf{u}_h^0$ are possible. For instance, both the L^2-projection of $\mathbf{u}_0$ and the Clemént finite element interpolation of $\mathbf{u}_0$ into $\mathbf{V}^h$ (cf. [14]) are qualified candidates for $\mathbf{u}_h^0$.

An immediate consequence of Theorems 4.1, 5.2, 6.12, 7.1, and 7.2 is the following convergence theorem.

THEOREM 7.3. *Let $1 \le p < \infty$, let $\mathbf{U}^{\varepsilon,\delta,h,\tau}$ be defined by* (7.7)*, and assume that the assumptions of Theorems* 7.1, 7.2, 4.1, 5.2, *and* 6.12 *hold. Then there exists a subsequence of $\{\mathbf{U}^{\varepsilon,\delta,h,\tau}\}$ (still denoted by the same notation) and a weak solution $\mathbf{u}$ of* (1.8)–(1.11) *such that*

$$\lim_{\varepsilon,\delta\to 0,\ h,\tau\to 0,} \| \mathbf{u} - \mathbf{U}^{\varepsilon,\delta,h,\tau} \|_{L^q(\Omega_T)} = 0 \qquad \forall q \in [1,\infty). \tag{7.9}$$

Acknowledgments. The authors would like to thank the Mathematisches Forschungsinstitut Oberwolfach for the kind hospitality and opportunity of its "Research in Pairs" program. The second author would like to thank Professor Fanghua Lin for a helpful discussion and the Institute of Mathematics and Its Applications (IMA) of the University of Minnesota for its support and hospitality during the author's recent visit to the IMA. The authors would also like to thank two anonymous referees for carefully reading the paper and for their valuable suggestions.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] F. ALOUGES, *A new algorithm for computing liquid crystal stable configurations: The harmonic mapping case*, SIAM J. Numer. Anal., 34 (1997), pp. 1708–1726.
[3] L. AMBROSIO, N. FUSCO, AND D. PALLARA, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford University Press, New York, 2000.

[4] G. Anzellotti, *Pairing between measures and bounded functions and compensated compactness*, Ann. Mat. Pura Appl. (4), 135 (1983), pp. 293–318.
[5] J. W. Barrett, X. Feng, and A. Prohl, *Convergence of a fully discrete finite element method for a degenerate parabolic system modeling nematic liquid crystals with variable degree of orientation*, M2AN Math. Model. Numer. Anal., 40 (2006), pp. 175–199.
[6] R. Becker, X. Feng, and A. Prohl, *Finite element approximations of the Ericksen–Leslie model for nematic liquid crystal flow*, SIAM J. Numer. Anal., 46 (2008), pp. 1704–1731.
[7] F. Bethuel, H. Brezis, and F. Hélein, *Ginzburg-Landau Vertices*, Birkhäuser, New York, 1994.
[8] H. Brezis, *Operateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North-Holland, Amsterdam, The Netherlands, 1973.
[9] K. C. Chang, W. D. Ding, and R. Ye, *Finite-time blow-up of the heat flow of harmonic maps from surfaces*, J. Differential Geom., 36 (1992), pp. 507–515.
[10] G.-Q. Chen and H. Frid, *Divergence-measure fields and hyperbolic conservation laws*, Arch. Rational Mech. Anal., 147 (1999), pp. 89–118.
[11] Y. Chen, *The weak solutions to the evolution problem of harmonic maps*, Math. Z., 201 (1989), pp. 69–74.
[12] Y. Chen, M.-H. Hong, and N. Hungerbühler, *Heat flow of p-harmonic maps with values into spheres*, Math. Z., 215 (1994), pp. 25–35.
[13] Y. M. Chen and M. Struwe, *Regularity for the heat flow for harmonic maps*, Math. Z., 201 (1989), pp. 83–103.
[14] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, The Netherlands, 1978.
[15] J.-M. Coron, *Nonuniqueness for the heat flow of harmonic maps*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 335–344.
[16] P. Courilleau and F. Demengel, *Heat flow for p-harmonic maps with values in the circle*, Nonlinear Anal., 41 (2000), pp. 689–700.
[17] F. Duzaar and M. Fuchs, *Existence and regularity of functions which minimize certain energies in homotopy classes of mappings*, Asymptot. Anal., 5 (1991), pp. 129–144.
[18] J. Eells and J. H. Sampson, *Harmonic mappings of Riemannian manifolds*, Amer. J. Math., 86 (1964), pp. 109–169.
[19] J. Ericksen and D. Kinderlehrer, *Theory and applications of liquid crystals*, IMA Vol. Math. Appl. 5, Springer-Verlag, New York, 1997, pp. 99–122.
[20] L. C. Evans, *Weak Convergence Methods for Nonlinear Partial Differential Equations*, AMS, Providence, RI, 1990.
[21] A. Fardoun and R. Regbaoui, *Heat flow for p-harmonic maps between compact Riemannian manifolds*, Indiana Math. J., 40 (2002), pp. 1305–1320.
[22] A. Fardoun and R. Regbaoui, *Heat flow for p-harmonic maps with small initial data*, Calc. Var. Partial Differential Equations, 16 (2003), pp. 1–16.
[23] X. Feng, *Divergence-L^q and divergence-measure tensors fields and gradient flow for linear growth functionals of maps into spheres*, Calc. Var. Partial Differential Equations, submitted.
[24] X. Feng and A. Prohl, *Analysis of total variation flow and its finite element approximations*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 533–556.
[25] A. Freire, *Uniqueness for the harmonic map flow in two dimensions*, Calc. Var. Partial Differential Equations, 1 (1995), pp. 95–105.
[26] M. Giaquinta, G. Modica, and J. Soucek, *Variational problems for maps of bounded variation with values in S^1*, Calc. Var. Partial Differential Equations, 1 (1993), pp. 87–121.
[27] M. Giaquinta, G. Modica, and J. Soucek, *Cartesian Currents in the Calculus Variations*, II: *Variational Integrals*, Springer-Verlag, New York, 1998.
[28] Y. Giga, Y. Kashima, and N. Yamazaki, *Local solvability of a constrained gradient system of total variation*, Abstr. Appl. Anal., 8 (2004), pp. 651–682.
[29] B. Guo and M. C. Hong, *The Landau-Lifshitz equation of the ferromagnetic spin chain and harmonic maps*, Calc. Var. Partial Differential Equations, 1 (1994), pp. 311–334.
[30] R. Hardt and F. H. Lin, *Mappings minimizing the L^p-norm of the gradient*, Comm. Pure Appl. Math., 15 (1987), pp. 555–588.
[31] N. Hungerbühler, *Global weak solutions of the p-harmonic flow into homogeneous spaces*, Indiana Math. J., 45 (1996), pp. 275–288.
[32] N. Hungerbühler, *Non-uniqueness for the p-harmonic flow*, Canad. Math. Bull, 40 (1997), pp. 793–798.
[33] N. Hungerbühler, *m-harmonic flow*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 24 (1997), pp. 593–631.

[34] N. Hungerbühler, *Heat flow into spheres for a class of energies*, in Variational Problems in Riemannian Geometry, Progr. Nonlinear Differential Equations Appl. 59, Birkhäuser, Basel, 2004, pp. 45–65.
[35] O. A. Ladyženskaja, V. A. Solonnikov, and N. N. Uarlceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1967.
[36] L. D. Landau and E. M. Lifshitz, *Electrodynamics of Continuous Media*, Pergamon, Oxford, 1960.
[37] F. H. Lin, *Static and moving defects in liquid crystals*, in Proceedings of the International Congress of Mathematicians (Kyoto, 1990), Math. Soc. Japan, Tokyo, 1991, pp. 1165–1171.
[38] X.-G. Liu, *A note on heat flow of p-harmonic mappings*, Kexue Tongbao, 42 (1997), pp. 15–18 (in Chinese).
[39] X.-G. Liu, *A remark on p-harmonic heat flows*, Chinese Sci. Bull., 42 (1997), pp. 441–444.
[40] M. Misawa, *Approximation of p-harmonic maps by the penalized equation* Nonlinear Anal., 47 (2001), pp. 1069–1080.
[41] M. Misawa, *On the p-harmonic flow into spheres in the singular case*, Nonlinear Anal., 50 (2002), pp. 485–494.
[42] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.
[43] R. Schoen and K. Uhlenbeck, *A regularity theorem for harmonic maps*, J. Differential Geom., 17 (1982), pp. 307–335.
[44] R. Schoen and K. Uhlenbeck, *Regularity of minimizing harmonic maps into the sphere*, Invent. Math., 78 (1984), pp. 89–100.
[45] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–96.
[46] M. Struwe, *On the evolution of harmonic maps of Riemannian surfaces*, Comment. Math. Helv., 60 (1985), pp. 558–581.
[47] M. Struwe, *Variational Methods*, Springer-Verlag, New York, 1990.
[48] M. Struwe, *Geometric evolution problems*, in Nonlinear Partial Differential Equations in Differential Geometry (Park City, UT, 1992), IAS/Park City Math. Ser. 2, AMS, Providence, RI, 1996, pp. 257–339.
[49] B. Tang, G. Sapiro, and V. Caselles, *Diffusion of general data on non-flat manifolds via harmonic maps theory: The direction diffusion case*, Int. J. Comput. Vision, 36 (2000), pp. 149–161.
[50] L. A. Vese and S. J. Osher, *Numerical methods for p-harmonic flows and applications to image processing*, SIAM J. Numer. Anal., 40 (2002), pp. 2085–2104.
[51] E. Virga, *Variational Theories for Liquid Crystals*, Chapman and Hall, London, 1994.
[52] E. Zeidler, *Nonlinear Functional Analysis and Its Applications*, Vol. II/B, Springer-Verlag, New York, 1990.

SIAM J. MATH. ANAL.
Vol. 40, No. 4, pp. 1499–1515

GLOBAL EXISTENCE OF SOME INFINITE ENERGY SOLUTIONS FOR A PERFECT INCOMPRESSIBLE FLUID*

RALPH SAXTON† AND FERIDE TIĞLAY†

Abstract. This paper provides results on local and global existence for a class of solutions to the Euler equations for an incompressible, inviscid fluid. By considering a class of solutions which exhibits a characteristic growth at infinity we obtain an initial value problem for a nonlocal equation. We establish local well-posedness in all dimensions and persistence in time of these solutions for three and higher dimensions. We also examine a weaker class of global solutions.

Key words. incompressible fluid, three dimensions, stagnation point, global existence, Cauchy problem for periodic initial data, infinite energy

AMS subject classifications. 35Q35, 76B03

DOI. 10.1137/080713768

1. Introduction. A fundamental question in the study of fluids concerns the possibility of finite time blow up of solutions to the Euler equations for a perfect, incompressible fluid. It is well known that blow up cannot take place in the two-dimensional case for solutions defined over a bounded domain subject to Dirichlet boundary conditions (see, for example, the work of Wolibner [17] and Ebin [7]), since smooth data in this case lead to solutions remaining smooth for all time. However, the question remains open in higher dimensions.

A separate class of solutions consists of those having "stagnation-point" form, which attracted early attention by Weyl [16] and Lin [13], and provides a set of equations which depend on only a single spatial variable and time. The resulting equations, once solved, provide exact solutions to the full Euler equations. However, the associated growth of the full solutions in certain directions means that the flows possess, at best, only locally finite kinetic energy. Nevertheless, one may discuss such questions as finite time blow up for this class, and it has been shown by Stuart in [15] that this can take place when the reduced equations are defined over the real line and those solutions decay at infinity. In this case the corresponding spatial domain for the full equations is $\mathbb{R}^n$ for $n = 2$ and $n = 3$, with the full set of solutions growing linearly in the other direction(s).

The evolution of two-dimensional solutions, which can blow up on the unbounded domain $\mathbb{R}^2$, therefore differs significantly from the class of globally defined solutions which exists for bounded subdomains of $\mathbb{R}^2$. So the consequences of higher-dimensional stagnation-point solutions blowing up might be thought to result simply from their behavior at infinity rather than bearing on the question of singularity formation. This view is in some sense strengthened by the results of Childress et al. in [3] (see also a related result by Cox in [5]), where solutions defined over a two-dimensional, infinite strip were examined. Since blow up was still found over this, smaller, semi-infinite domain, stagnation-point solutions defined over such domains

*Received by the editors January 18, 2008; accepted for publication (in revised form) May 27, 2008; published electronically November 19, 2008.
http://www.siam.org/journals/sima/40-4/71376.html
†Department of Mathematics, University of New Orleans, New Orleans, LA 70148 (rsaxton@math.uno.edu, ftigley@uno.edu). The second author carried out part of this work at the Department of Mathematics at the University of Notre Dame.

would give the appearance of behaving, generally, much as those defined on the full space.

A similar approach is implemented by Constantin in [4] to reduce the three-dimensional Euler equations, periodic in two directions, to a nonlocal Riccati equation and prove the blow up in finite time by solving these equations on characteristics.

In this paper we consider a stagnation-point class of solutions defined over $\mathbb{R}^n$ which is spatially periodic in one coordinate direction. In two dimensions, the equations reduce to those of [3] and, although we examine slightly different boundary conditions, the same blow up results essentially apply. In three and higher dimensions, however, we find a "regularizing" effect not present in solutions which decay to zero at infinity in the same coordinate direction, and this leads to the existence for all time of all such solutions, stemming from sufficiently smooth initial data.

Section 2 sets out the fundamental field equations, which in a basic sense date back to [13], [16]. Section 3 is devoted to local well-posedness (existence, uniqueness, and continuous dependence on initial data) of classical solutions to the initial value problem for the pseudodifferential equation derived in section 2. We establish this result by rewriting the problem on the topological group $\mathcal{D}$ of C^1 class diffeomorphisms as an initial value problem for an ordinary differential equation. Section 4 provides a priori estimates for more regular classes of solutions, leading to global existence of such solutions in three and higher dimensions.

In section 5, we reconsider a class of piecewise affine solutions previously mentioned in [3]. These solutions are less regular than those arising in our existence results. It is found that they exist globally, independently of the underlying dimension.

2. The *n*-dimensional equation. Consider the n-dimensional Euler equations for an ideal, inviscid and incompressible fluid

$$\partial_t \mathbf{u} + \mathbf{u}.\nabla \mathbf{u} + \nabla p = 0, \tag{2.1a}$$

$$\nabla.\mathbf{u} = 0, \tag{2.1b}$$

where $\mathbf{x} = (x_1, \dots x_n) \equiv (x_1, \mathbf{x}')$. Denoting x_1 by x, $\mathbf{u}(x, \mathbf{x}', t)$ represents the spatial velocity field of the fluid and $p(x, \mathbf{x}', t)$ its pressure. We impose the ansatz

$$\mathbf{u}(x, \mathbf{x}', t) = (u(x,t), -\partial_x u(x,t)\mathbf{v}(\mathbf{x}', t)), \tag{2.2}$$

where the $(n-1)$-dimensional vector field, $\mathbf{v}$, will be chosen below. As a consequence of (2.2), (2.1a) may be written as

$$\partial_t u + u\partial_x u + \partial_x p = 0, \tag{2.3}$$

together with

$$\partial_t \partial_x u\, \mathbf{v} + \partial_x u\, \partial_t \mathbf{v} + u\partial_x^2 u\, \mathbf{v} - (\partial_x u)^2 \mathbf{v} \cdot \nabla' \mathbf{v} - \nabla' p = \mathbf{0}, \tag{2.4}$$

where the primed operators refer to the variable $\mathbf{x}'$. Using (2.2) and (2.3), one sees that $\nabla' \partial_x p = 0$. Hence, differentiating (2.4) in x eliminates the pressure term to give

$$\partial_x(\partial_t \partial_x u + u\partial_x^2 u)\, \mathbf{v} + \partial_x^2 u\, \partial_t \mathbf{v} - \partial_x((\partial_x u)^2)\mathbf{v} \cdot \nabla' \mathbf{v} = \mathbf{0}. \tag{2.5}$$

Applying the $\nabla' \cdot$ operator to (2.5) and using (2.1b) with (2.2) to find that $\nabla' \cdot \mathbf{v} = 1$ shows

$$\partial_x(\partial_t \partial_x u + u\partial_x^2 u) - \partial_x((\partial_x u)^2)\ \nabla' \mathbf{v} : \nabla' \mathbf{v} = 0, \tag{2.6}$$

where $\nabla' \mathbf{v} : \nabla' \mathbf{v} = \mathrm{tr}(\nabla' \mathbf{v})^2 = \partial_j v_k \, \partial_k v_j$ (summing over j, k from 2 to n). For compatibility, we must choose $\mathbf{v}$ such that $\nabla' \mathbf{v} : \nabla' \mathbf{v}$ is independent of $\mathbf{x}'$. This can be done, for instance, by choosing $\mathbf{v} = \frac{1}{n-1} \mathbf{x}'$, in which case $\nabla' \mathbf{v} : \nabla' \mathbf{v} = \frac{1}{n-1}$ and (2.4) takes the form

$$\left(\partial_t \partial_x u + u \partial_x^2 u - \frac{1}{n-1} (\partial_x u)^2 \right) \mathbf{v} - \nabla' p = \mathbf{0}. \tag{2.7}$$

We note that this particular choice of $\mathbf{v}$ corresponds to the stagnation-point form solution referred to in the introduction. As a result, the solution becomes unbounded in the $\mathbf{x}'$ direction, and hence has infinite energy when considered over the entire n-dimensional domain. We next examine the periodic, initial-boundary value problem, with boundary conditions for $x \in \mathbb{T} \simeq \mathbb{R}/\mathbb{Z}, t \geq 0$, given by

$$\mathbf{u}(0, \mathbf{x}', t) = \mathbf{u}(1, \mathbf{x}', t) \tag{2.8}$$

and

$$p(0, \mathbf{x}', t) = p(1, \mathbf{x}', t). \tag{2.9}$$

Since (2.6) now becomes

$$\partial_x (\partial_t \partial_x u + u \partial_x^2 u) - \frac{1}{n-1} \partial_x ((\partial_x u)^2)) = 0, \tag{2.10}$$

which we remark happens to be the x-derivative of a Calogero-class equation (see [2])

$$\partial_x \partial_t u + u \partial_x^2 u - \Phi(\partial_x u) = 0,$$

we obtain the equation

$$\partial_t \partial_x u + u \partial_x^2 u - \frac{1}{n-1} (\partial_x u)^2 = f \tag{2.11}$$

with f purely a function of time. This implies, by (2.7), that

$$\nabla' p = \frac{f}{n-1} \ \mathbf{x}', \tag{2.12}$$

while by (2.3), $-\partial_x^2 p = \frac{n}{n-1} (\partial_x u)^2 + f$ and so $\Delta p = -\frac{n}{n-1} (\partial_x u)^2$. Finally, for sufficiently smooth functions $u(x, t)$, using (2.8) and integrating (2.11) we have

$$f = -\frac{n}{n-1} \int_{\mathbb{T}} (\partial_x u)^2 dx, \tag{2.13}$$

while (2.3), (2.8), and (2.9) imply that

$$\frac{d}{dt} \int_{\mathbb{T}} u \, dx = 0. \tag{2.14}$$

Let us introduce the operator ∂_x^{-1} defined by

$$\partial_x^{-1} \phi(x, t) = \int_{x_0}^{x} \phi(y, t) dy - \int_{\mathbb{T}} \int_{x_0}^{x} \phi(y, t) \, dy dx.$$

We make the observation that ∂_x and ∂_x^{-1} generally do not commute since $[\partial_x, \partial_x^{-1}]\phi = \int_{\mathbb{T}} \phi \, dx$, where $[P, Q] = PQ - QP$. Consider (2.11), written in the form

$$\partial_x(\partial_t u + u\partial_x u) = \frac{n}{n-1}(\partial_x u)^2 + f(t) \tag{2.15}$$

with $n > 1$. As a result of (2.13) and the fact that $\partial_x^{-1}\partial_x\phi = \phi - \int_{\mathbb{T}} \phi \, dx$, we may write (2.15) in a nonlocal form as

$$\partial_x(\partial_t u + u\partial_x u) = \frac{n}{n-1}\,\partial_x^{-1}\partial_x((\partial_x u)^2) \tag{2.16}$$

and then, using the periodicity of u, we obtain

$$\partial_t u + u\partial_x u = \frac{n}{n-1}\,\partial_x^{-2}\partial_x((\partial_x u)^2). \tag{2.17}$$

Okamoto and Zhu [14] previously established local existence for (2.10) with $u \in H^2$, using a method introduced by Kato and Lai. Their approach requires showing uniqueness separately and then using uniqueness to prove continuous dependence on initial data. Here we instead derive a local well-posedness result in C^1 which follows from Picard iteration after rewriting the equation as an ordinary differential equation on an infinite-dimensional Banach space. This method can also be used to prove well-posedness in Sobolev spaces $H^s(\mathbb{T})$ for $s > 3/2$ (see [10] and [12], for example, for a similar result for the Camassa–Holm equation).

3. Local existence of classical solutions. In [1], Arnold observed that the initial value problem for the classical Euler equations of a perfect fluid can be stated as a geometric problem of finding geodesics on the group of volume preserving diffeomorphisms. Following this observation, Ebin and Marsden [8] developed the functional analytic tools to establish sharp local well-posedness results for the Euler equations. This method has since been used for other equations with similar geometric interpretations; for example, Misiołek [12] obtained local well-posedness in $C^1(\mathbb{T})$ for the Camassa–Holm equation, which is the equation for geodesics of the H^1 metric on the Virasoro group.

In this section we develop an appropriate analytic framework for (2.17), using a similar approach to prove the following theorem.

THEOREM 3.1. *Suppose that $n > 1$. Then there exists a unique solution*

$$u \in C^0([0,T), C^1(\mathbb{T})) \cap C^1([0,T), C^0(\mathbb{T}))$$

to the Cauchy problem for (2.17) *with initial data $u_0 \in C^1(\mathbb{T})$ for some $T > 0$, and the solution depends continuously on the initial data.*

Let γ be the flow generated by u, that is,

$$\frac{d\gamma}{dt}(x,t) = u(\gamma(x,t),t),$$

or $u = \dot{\gamma} \circ \gamma^{-1}$. Then we obtain the equation

$$\ddot{\gamma} = \frac{n}{n-1}\,\partial_x^{-2}\partial_x\left((\partial_x(\dot{\gamma}\circ\gamma^{-1}))^2\right)\circ\gamma \tag{3.1}$$

from (2.17). Therefore it is sufficient to prove that

$$F(X,\gamma) = \frac{n}{n-1}\,\left(\partial_x^{-2}\partial_x\left((\partial_x(X\circ\gamma^{-1}))^2\right)\right)\circ\gamma$$

defines a continuously differentiable vector field in a neighborhood of the identity on the topological group $\mathcal{D}$ of C^1 class diffeomorphisms. Then Theorem 3.1 follows by Picard iteration over Banach spaces.

We remark that the smooth dependence on initial data for (3.1) implies only continuous dependence on initial data for (2.17). The map $\gamma \to \gamma^{-1}$ is continuous but not locally Lipschitz [8], and this prevents obtaining more regularity for the initial data to solution map by this method. The question of whether the regularity of the solution map $u_0 \to u(t)$ can be improved, or not, is open. It is known, for instance, that it is not possible to improve the regularity of this map for the Camassa–Holm equation in Sobolev spaces [11].

In the remainder of this section, C_γ will represent a generic constant depending only on the C^1 norms of γ and γ^{-1}.

Proof of Theorem 3.1. Let us denote by P_γ the operator given by conjugation

$$P_\gamma(\phi) := P(\phi \circ \gamma^{-1}) \circ \gamma$$

for any $\gamma \in \mathcal{D}$ and pseudodifferential operator P. Using this notation we write

$$F(X,\gamma) = \frac{n}{n-1}\,(\partial_x^{-2}\partial_x)_\gamma((\partial_x)_\gamma X)^2.$$

Next we compute the directional derivative $\partial_\gamma F_{(X,\gamma)}$ and prove that it is a bounded linear map.

Note that $(\partial_x^{-2}\partial_x)f = \partial_x^{-1}\{f - \int_0^1 f\,dx\}$ is a bounded operator from $C^0(\mathbb{T})$ into $C^1(\mathbb{T})$. Furthermore we abuse the notation slightly and denote by $\mathcal{D}$ the connected component of orientation preserving C^1 diffeomorphisms of $\mathbb{T}$.

Let $s \to \gamma_s$ be a smooth curve in $\mathcal{D}$ such that $\gamma_0 = id$ and $\partial_s\gamma_s|_{s=0} = W$ for $W \in C^1(\mathbb{T})$. Then we have

$$\partial_\gamma F_{(X,\gamma)}(W) = \frac{n}{n-1}\{(\partial_\varepsilon G_\varepsilon)|_{\varepsilon=0} \circ \gamma + W(\partial_x G_\varepsilon)_{\varepsilon=0} \circ \gamma\}, \tag{3.2}$$

where $G_\varepsilon = \partial_x^{-2}\partial_x\{(\partial_x(X \circ \gamma_\varepsilon^{-1}))^2\}$. We know that $\partial_x\partial_x^{-1}$ gives the identity; hence the second summand on the right in (3.2) can be written as

$$W\partial_x G_\varepsilon|_{\varepsilon=0} \circ \gamma = \left\{((\partial_x)_\gamma X)^2 - \int_0^1 (\partial_x(X \circ \gamma^{-1}))^2 dx\right\} W. \tag{3.3}$$

Moreover, the computation of the first summand on the right in (3.2) is reduced by

$$(\partial_\varepsilon G_\varepsilon)|_{\varepsilon=0} = \partial_x^{-2}\partial_x(\partial_\varepsilon H_\varepsilon|_{\varepsilon=0}) \tag{3.4}$$

to determine $\partial_\varepsilon H_\varepsilon|_{\varepsilon=0}$, where $H_\varepsilon = (\partial_x(X \circ \gamma_\varepsilon^{-1}))^2$. Here a straightforward computation leads to

$$\partial_\varepsilon H_\varepsilon|_{\varepsilon=0} = -\left(\partial_x(X \circ \gamma^{-1})\right)^2 \partial_x(W \circ \gamma^{-1}) - \partial_x\{(W \circ \gamma^{-1})\left(\partial_x(X \circ \gamma^{-1})\right)^2\}. \tag{3.5}$$

Then, after an integration by parts, we obtain

$$\begin{aligned}\partial_\varepsilon G_\varepsilon|_{\varepsilon=0} = &-\left(\partial_x(X \circ \gamma^{-1})\right)^2 (W \circ \gamma^{-1}) + \textstyle\int_0^1 \left(\partial_x(X \circ \gamma^{-1})\right)^2 (W \circ \gamma^{-1})dx \\ &- \partial_x^{-2}\partial_x\{(\partial_x(X \circ \gamma^{-1}))^2\partial_x(W \circ \gamma^{-1})\} \\ &+ \left(x - \frac{1}{2}\right)\{\left(\partial_x(X \circ \gamma^{-1})\right)^2 (W \circ \gamma^{-1})\}|_0^1.\end{aligned}$$

The last term on the left-hand side of the above inequality vanishes since X and W are periodic functions and γ is an orientation preserving diffeomorphism. Therefore we have

$$\begin{aligned}\partial_\gamma F_{(X,\gamma)}(W) = \frac{n}{n-1}\Big\{ &- \partial_x^{-2}\partial_x\{(\partial_x(X\circ\gamma^{-1}))^2\partial_x(W\circ\gamma^{-1})\}\circ\gamma \\ &+ \int_0^1 (\partial_x(X\circ\gamma^{-1}))^2 W\circ\gamma^{-1}dx \\ &- W\int_0^1 (\partial_x(X\circ\gamma^{-1}))^2 dx\Big\}. \end{aligned} \tag{3.6}$$

The linearity of the map $W \to \partial_\gamma F_{(X,\gamma)}(W)$ is clear. Thus we proceed to show that it is bounded. It is sufficient to estimate the C^1 norms of all three summands on the right in (3.6). The second and third terms are both bounded by $C_\gamma\|W\|_{C^0}\|X\|_{C^1}^2$. For the first term on the right in (3.6), we have

$$\begin{aligned}\|(\partial_x^{-2}\partial_x)_\gamma\{((\partial_x)_\gamma X)^2(\partial_x)_\gamma W\}\|_{C^1} \le &\|(\partial_x^{-2}\partial_x)_\gamma\{((\partial_x)_\gamma X)^2(\partial_x)_\gamma W\}\|_{C^0} \\ &+ \|((\partial_x)_\gamma X)^2(\partial_x)_\gamma W\|_{C^0}\|\gamma\|_{C^1},\end{aligned} \tag{3.7}$$

which is bounded by $C_\gamma\|X\|_{C^1}^2\|W\|_{C^1}$.

In the direction of X, the Gâteaux derivative of F is given by

$$\partial_X F_{(X,\gamma)}(W) = \frac{2n}{n-1}(\partial_x^{-2}\partial_x)_\gamma\left((\partial_x)_\gamma X(\partial_x)_\gamma W\right),$$

and this is a bounded map since

$$\begin{aligned}\|\partial_X F_{(X,\gamma)}(W)\|_{C^1} &\le C_n\|(\partial_x^{-2}\partial_x)_\gamma\left((\partial_x)_\gamma X(\partial_x)_\gamma W\right)\|_{C^0} \\ &\quad + \left\|(\partial_x)_\gamma X(\partial_x)_\gamma W - \int_{\mathbb{T}}(\partial_x)_\gamma X(\partial_x)_\gamma W\, dx\right\|_{C^0}\|\gamma\|_{C^1} \\ &\le C_{n,\gamma}\|X\|_{C^1}\|W\|_{C^1},\end{aligned}$$

where $C_{n,\gamma}$ depends only on n and C^1 norms of γ and γ^{-1}.

In order to complete the proof of Theorem 3.1 it is sufficient to show that F is Fréchet differentiable; i.e., both directional derivatives $\partial_\gamma F$ and $\partial_X F$ are continuous maps.

Continuity of $(X,\gamma) \to \partial_\gamma F_{(X,\gamma)}(W)$. The following inequality reduces the proof of continuity of $\partial_\gamma F_{(X,\gamma)}(W)$ to estimating the two summands on the right-hand side:

$$\begin{aligned}\|\partial_\gamma F_{(X_1,\gamma_1)}(W) - \partial_\gamma F_{(X_2,\gamma_2)}(W)\|_{C^1} \le\ &\|\partial_\gamma F_{(X_1,\gamma_1)}(W) - \partial_\gamma F_{(X_2,\gamma_1)}(W)\|_{C^1} \\ &+ \|\partial_\gamma F_{(X_2,\gamma_1)}(W) - \partial_\gamma F_{(X_2,\gamma_2)}(W)\|_{C^1}\end{aligned}$$

where the inequality holds up to a constant depending on n. We rewrite the C^1 norm that we wish to estimate to show continuity in X as

$$\begin{aligned}\|\partial_\gamma F_{(X_1,\gamma)} - \partial_\gamma F_{(X_2,\gamma)}\|_{C^1} \leq & \|(\partial_x^{-2}\partial_x)_\gamma \left\{(((\partial_x)_\gamma X_1)^2 - ((\partial_x)_\gamma X_2)^2)(\partial_x)_\gamma W\right\}\|_{C^1} \\ & + \left|\int_0^1 \left(\partial_x(X_1\circ\gamma - X_2\circ\gamma)\right)^2 W\circ\gamma^{-1}dx\right| \\ & + |W|\left|\int_0^1 \left(\partial_x(X_1\circ\gamma - X_2\circ\gamma)\right)^2 dx\right|. \end{aligned} \tag{3.8}$$

The last two summands in (3.8) are bounded by $C_\gamma\|X_1 - X_2\|_{C^1}^2\|W\|_{C^0}$. For the remaining term, we have

$$\begin{aligned}&\|(\partial_x^{-2}\partial_x)_\gamma \left\{(((\partial_x)_\gamma X_1)^2 - ((\partial_x)_\gamma X_2)^2)(\partial_x)_\gamma W\right\}\|_{C^1} \\ &\quad \leq \|(\partial_x^{-2}\partial_x)_\gamma \left\{(((\partial_x)_\gamma X_1)^2 - ((\partial_x)_\gamma X_2)^2)(\partial_x)_\gamma W\right\}\|_{C^0} \\ &\qquad + \|\left((\partial_x(X_1\circ\gamma^{-1}))^2 - (\partial_x(X_2\circ\gamma^{-1}))^2\right)\partial_x(W\circ\gamma^{-1})\|_{C^0}\|\partial_x\gamma\|_{C^0},\end{aligned}$$

which is bounded by $C_\gamma\|X_1 - X_2\|_{C^1}\|X_1 + X_2\|_{C^1}\|W\|_{C^1}$.

Our next estimate establishes continuity of $\gamma \to \partial_\gamma F_{(X,\gamma)}(W)$. Note that it is sufficient to consider

$$\begin{aligned}&\|\partial_\gamma F_{(X,\gamma)}(W) - \partial_\gamma F_{(X,id)}(W)\|_{C^1} \\ &\quad \leq \|(\partial_x^{-2}\partial_x)_\gamma \left\{((\partial_x)_\gamma X)^2(\partial_x)_\gamma W\right\} - \partial_x^{-2}\partial_x((\partial_x X)^2\partial_x W)\|_{C^1} \qquad (3.9)\\ &\qquad + \left|\int_0^1 \left\{(\partial_x(X\circ\gamma^{-1}))^2(W\circ\gamma^{-1}) - (\partial_x X)^2 W\right\}dx\right| \qquad (3.10)\\ &\qquad + \|W\|_{C^0}\left|\int_0^1 \left\{(\partial_x(X\circ\gamma^{-1}))^2 - (\partial_x X)^2\right\}dx\right|, \qquad (3.11)\end{aligned}$$

where the inequality is up to a constant depending on n. After adding and subtracting the appropriate terms, (3.10) is bounded by

$$\|W\|_{C^0}\|\partial_x X\|_{C^0}\|(\partial_x X\circ\gamma^{-1})\partial_x\gamma^{-1} - \partial_x X\|_{C^0} + \|\partial_x X\|_{C^0}^2\|W\|_{C^1}\|\gamma - id\|_{C^0}, \tag{3.12}$$

which is bounded (up to a constant C_γ) by $\|X\|_{C^1}^2\|W\|_{C^1}\|\gamma - id\|_{C^1}$. The term in (3.11) is estimated similarly. Hence in order to establish continuity in γ it is sufficient to bound

$$\begin{aligned}&\|(\partial_x^{-2}\partial_x)_\gamma \left\{((\partial_x)_\gamma X)^2(\partial_x)_\gamma W\right\} - \partial_x^{-2}\partial_x((\partial_x X)^2\partial_x W)\|_{C^1} \\ &\quad \leq \|(\partial_x^{-2}\partial_x)_\gamma \left\{((\partial_x)_\gamma X)^2(\partial_x)_\gamma W\right\} - \partial_x^{-2}\partial_x((\partial_x X)^2\partial_x W)\|_{C^0} \qquad (3.13)\\ &\qquad + \|\left\{(\partial_x(X\circ\gamma^{-1})^2\partial_x(W\circ\gamma^{-1})\right\}\circ\gamma\partial_x\gamma - (\partial_x X)^2\partial_x W\|_{C^0}. \qquad (3.14)\end{aligned}$$

The norm in (3.14) is equal to

$$\|\partial_x W(\partial_x X)^2\left\{(\partial_x\gamma^{-1})^2\circ\gamma - 1\right\}\|_{C_0}, \tag{3.15}$$

which is bounded by $C_\gamma\|\partial_x X\|_{C^0}^2\|\partial_x W\|_{C^0}\|\gamma - id\|_{C^1}$. For (3.13) it is sufficient to estimate

$$\|(\partial_x^{-2}\partial_x)_\gamma S - \partial_x^{-2}\partial_x S\|_{C^0} + \|\partial_x^{-2}\partial_x S - \partial_x^{-2}\partial_x((\partial_x X)^2\partial_x W)\|_{C^0}, \tag{3.16}$$

where

$$S = S(X, \gamma, W) = (\partial_x(X \circ \gamma^{-1}))^2 \circ \gamma \partial_x(W \circ \gamma^{-1}) \circ \gamma.$$

After a change of variables we regroup the terms in the first summand of (3.16) to obtain

(3.17)
$$(\partial_x^{-2}\partial_x)_\gamma S - \partial_x^{-2}\partial_x S = \int_0^1 \int_{\gamma^{-1}(x)}^x S(y)\partial_x\gamma(y)dydx + \int_{x_0}^x S(y)(\partial_x\gamma(y) - 1)dy$$

(3.18)
$$- \int_0^1 \int_{x_0}^x S(y)(\partial_x\gamma(y) - 1)dydx$$

(3.19)
$$- \left(\gamma(x) - \frac{1}{2}\right) \int_0^1 S \circ \gamma^{-1}(y)dy + \left(x - \frac{1}{2}\right) \int_0^1 S(y)dy.$$

The right-hand side of this equality can be simplified further as

(3.20)
$$(\partial_x^{-2}\partial_x)_\gamma S - \partial_x^{-2}\partial_x S = \int_0^1 \int_{\gamma^{-1}(x)}^x S(y)\partial_x\gamma(y)dydx$$

(3.21)
$$+ \partial_x^{-2}\partial_x(S(\partial_x\gamma - 1)) - (\gamma(x) - x) \int_0^1 S \circ \gamma^{-1}(y)dy.$$

Clearly the sup norms of all three summands in (3.20)–(3.21) vanish in the limit as γ goes to id in C^1. The second summand in (3.16) is equal to the C^0 norm of

(3.22)
$$\partial_x^{-2}\partial_x \left(\partial_x W(\partial_x X)^2 \left\{(\partial_x\gamma^{-1})^3 \circ \gamma - 1\right\}\right),$$

which is bounded by $C_\gamma \|W\|_{C^1} \|X\|_{C^1}^2 \|\gamma - id\|_{C^1}$. Hence the continuity of $\gamma \to \partial F_{(X,\gamma)}$ follows.

The continuity of $(X, \gamma) \to \partial_X F_{(X,\gamma)}$ can be shown analogously. Therefore $F(X, \gamma)$ defines a continuously differentiable map in a neighborhood of (id, u_0). This completes the proof of Theorem 3.1. □

4. Global existence for $n \geq 3$. In this section we investigate the persistence of solutions of the initial value problem for (2.17) and show that, unlike the two-dimensional case where solutions may blow up in finite time (see [3], [5]), they persist for $n \geq 3$ in the appropriate function spaces.

For the theorem below let us use the following notation:

$$X_n(\mathbb{T}) = \begin{cases} W^{2,\infty}(\mathbb{T}), & n = 3, \\ W^{2,\frac{n-1}{n-3}}(\mathbb{T}), & n > 3, \end{cases}$$

and

$$Y_n(\mathbb{T}) = \begin{cases} W^{1,\infty}(\mathbb{T}), & n = 3, \\ W^{1,\frac{n-1}{n-3}}(\mathbb{T}), & n > 3. \end{cases}$$

THEOREM 4.1. *Let $n \geq 3$ and assume that $u_0(x) \in X_n(\mathbb{T})$. Then the solution $u(x,t)$ from Theorem 3.1 has a unique extension to all $T < \infty$ such that $u(x,t) \in C^0([0,T], X_n(\mathbb{T})) \cap C^1([0,T], Y_n(\mathbb{T}))$.*

Proof. Consider (2.10), expressed in the form

$$\partial_t \partial_x^2 u + u\partial_x^3 u + \frac{n-3}{n-1}\,\partial_x u \partial_x^2 u = 0. \tag{4.1}$$

Using the flow γ of u ($\dot\gamma = u \circ \gamma$) in (4.1), we first solve for $\partial_x^2 u$:

$$\partial_x^2 u \circ \gamma(t) = u_0'' \exp\left(-\frac{n-3}{n-1}\int_0^t \partial_x u \circ \gamma(s) ds\right). \tag{4.2}$$

By the identity $\partial_x \dot\gamma = \partial_x u \circ \gamma\, \partial_x \gamma$, we also have

$$\partial_x \gamma(t) = \exp\left(\int_0^t \partial_x u \circ \gamma(s) ds\right). \tag{4.3}$$

Therefore

$$\partial_x^2 u \circ \gamma(t) (\partial_x \gamma(t))^{\frac{n-3}{n-1}} = u_0''. \tag{4.4}$$

Note that Theorem 3.1 implies that $\gamma \in C^1$ locally in time and it follows, given $\partial_x \gamma(0) = 1$, that there exists an interval, $t \in [0, \tau(\varepsilon))$, over which $0 < \varepsilon \leq \inf_{x\in\mathbb{T}} \partial_x\gamma(t) \leq \sup_{x\in\mathbb{T}} \partial_x\gamma(t) \leq \varepsilon^{-1}$. Equation (4.4) then implies that, locally, $u \in X_n(\mathbb{T})$, since γ maps $\mathbb{T}$ diffeomorphically to itself. In turn, (2.16) shows that $\partial_t u \in Y_n(\mathbb{T})$ over the same time interval. (With additional assumptions on the data, further regularity can also be bootstrapped to higher derivatives.)

Assuming then that sufficient smoothness holds locally in time, we find on multiplying (4.1) by $|\partial_x^2 u|^{p-2}\,\partial_x^2 u$ that

$$\partial_t |\partial_x^2 u|^p + u\partial_x |\partial_x^2 u|^p + p\,\frac{n-3}{n-1}\,\partial_x u |\partial_x^2 u|^p = 0. \tag{4.5}$$

Since (2.2) and (2.8) imply that both u and $\partial_x u$ are periodic functions of x, the same is true of $\partial_x^2 u$, by (2.11). One therefore obtains, on integrating (4.5) over $\mathbb{T}$,

$$\frac{d}{dt}\int_{\mathbb{T}} |\partial_x^2 u|^p dx + \left(p\,\frac{n-3}{n-1} - 1\right)\int_{\mathbb{T}} \partial_x u |\partial_x^2 u|^p dx = 0, \tag{4.6}$$

from which it follows that the $L^{\frac{n-1}{n-3}}(\mathbb{T})$ norm of $\partial_x^2 u$ is uniformly conserved in time for $n > 3$. The case $n = 3$ can either be considered as the limit $n \to 3$ with $p \to \infty$ in (4.6), or directly using (4.1) which shows that $\partial_x^2 u$ is constant along characteristics and hence its $L^\infty(\mathbb{T})$ norm is uniformly conserved.

Periodicity of $u(x,t)$ in x implies there exists a zero for $\partial_x u$, say at $x = x_0(t)$, and so for $x, x_0 \in \mathbb{T}$,

$$\partial_x u(x,t) = \int_{x_0}^x \partial_y^2 u(y,t) dy.$$

For $n > 3$, we therefore have the estimate

$$|\partial_x u(x,t)| \leq |x - x_0|^{\frac{2}{n-1}} ||u_0''||_{\frac{n-1}{n-3}} \leq ||u_0''||_{\frac{n-1}{n-3}}$$

using Hölder's inequality, and so

$$||\partial_x u||_\infty \leq ||u_0''||_{\frac{n-1}{n-3}}.$$

If $n = 3$, then

$$||\partial_x^2 u||_\infty = ||u_0''||_\infty,$$

which means

$$|\partial_x u(x,t)| \leq |x - x_0|||u_0''||_\infty \leq ||u_0''||_\infty$$

for all $x \in \mathbb{T}$, and so

$$||\partial_x u||_\infty \leq ||u_0''||_\infty$$

for all $t > 0$.

Further, since $u(x,t) - u_0(x)$ has mean zero by (2.14), there exists $x = x_1(t)$, where $u(x_1,t) = u_0(x_1)$ and, for $x, x_1 \in \mathbb{T}$, we have

$$u(x,t) = u_0(x) + \int_{x_1}^{x} (\partial_y u(y,t) - u_0'(y))dy.$$

It follows that

$$|u(x,t)| \leq ||u_0||_\infty + |x - x_1|(||\partial_x u||_\infty + ||u_0'||_\infty)$$

for all $x \in \mathbb{T}$, which gives the inequality

$$||u||_\infty \leq ||\partial_x u||_\infty + ||u_0||_{C^1}.$$

Combining the results of the previous two paragraphs shows that

$$||u||_{C^1} \leq ||u_0||_{C^2} \quad \text{for} \quad n = 3, \tag{4.7}$$

and

$$||u||_{C^1} \leq ||u_0||_{C^1} + ||u_0''||_{\frac{n-1}{n-3}} \quad \text{for} \quad n > 3. \tag{4.8}$$

Finally, on using the properties of the operators $\partial_x^{-1}\partial_x$ and $\partial_x^{-2}\partial_x$ in (2.16) and (2.17) together with the above estimates, it is seen that $||\partial_t u||_\infty$ and $||\partial_t \partial_x u||_\infty$ are majorized by a function of $||u_0||_{C^2}$ for $n = 3$, while $||\partial_t u||_\infty$ and $||\partial_t \partial_x u||_{\frac{n-1}{n-3}}$ are majorized by a function of $||u_0||_{W^{2,\frac{n-1}{n-3}}}$ for $n > 3$.

In both of these cases it follows that the $C^1(\mathbb{T})$ norm of u and the $C^0(\mathbb{T})$ norm of $\partial_t u$ remain uniformly bounded in time over any interval of local existence and, by bootstrapping the arguments of Theorem 3.1, the solution can be continued, globally, in time. □

As a remark, we note here how a blow up argument made in [3], which involves a nontrivial class of separable solutions to (2.10) for $n = 2$, fails to apply in the case $n > 3$. In particular, the possible appearance of a $(\tau - t)^{-1}$ factor, $\tau > 0$, in the two-dimensional case no longer exists in higher dimensions.

Given the solution form $u(x,t) = X(x)T(t)$, (2.10) reduces to

$$\lambda X''(x) + \frac{n-3}{n-1} X'(x)X''(x) + X(x)X'''(x) = 0, \ x \in \mathbb{T}, \tag{4.9}$$

where

$$\dot{T}(t) - \lambda T(t)^2 = 0,\ t \geq 0, \tag{4.10}$$

and λ is a constant. Multiplying (4.9) by $|X''(x)|^{\frac{5-n}{n-3}}X''(x)$ now gives

$$\lambda|X''(x)|^{\frac{n-1}{n-3}} + \frac{n-3}{n-1}(X'(x)|X''(x)|^{\frac{n-1}{n-3}} + X(x)(|X''(x)|^{\frac{n-1}{n-3}})') = 0. \tag{4.11}$$

By using periodicity, an integration of (4.11) over $\mathbb{T}$ for $n > 3$ therefore shows

$$\lambda \int_{\mathbb{T}} |X''(x)|^{\frac{n-1}{n-3}} dx = 0 \tag{4.12}$$

and the result then follows. We note that there are in general one or more points of inflection in nontrivial, periodic solutions, which prevents this argument from holding in two dimensions.

5. Weak solutions. In this section, we construct a basic, piecewise differentiable class of weak solutions $u(x,t) \in C^0([0,T), PC^1(\mathbb{T})) \cap C^1([0,T), PC^0(\mathbb{T}))$ to (2.11), which are found to exist for all $T > 0$, regardless of the underlying dimension.

For every vector field $\Phi(\mathbf{x},t) \in C_0^\infty([0,T) \times \mathbb{T} \times \mathbb{R}^{n-1}; \mathbb{R}^n)$ such that $\nabla \cdot \Phi = 0$, and for every scalar function $\theta(\mathbf{x},t) \in C_0^\infty([0,T) \times \mathbb{T} \times \mathbb{R}^{n-1}; \mathbb{R})$, the velocity field $\mathbf{u}(\mathbf{x},t)$ in (2.1b) satisfies

$$\int_Q \partial_t \Phi \cdot \mathbf{u} + (\nabla \Phi\, \mathbf{u}) \cdot \mathbf{u}\, d\mathbf{x}\, dt + \int_{\mathbb{T}\times\mathbb{R}^{n-1}} \Phi(\mathbf{x},0) \cdot \mathbf{u}(\mathbf{x},0)\, d\mathbf{x} = 0, \tag{5.1}$$

$$\int_{\mathbb{R}^n} \nabla\theta \cdot \mathbf{u}\, d\mathbf{x} = 0, \tag{5.2}$$

where $Q = [0,T) \times \mathbb{T} \times \mathbb{R}^{n-1}$.

In terms of (2.2), (5.1) and (5.2) reduce to

$$\begin{aligned} &\int_Q \partial_t\phi\, u - \partial_t\Phi' \cdot \mathbf{v}\, \partial_x u\, d\mathbf{x}\, dt \\ &+ \int_Q \partial_x\phi\, u^2 - (\partial_x\Phi' + \nabla'\phi) \cdot \mathbf{v}\, u\partial_x u + (\nabla'\Phi'\, \mathbf{v}) \cdot \mathbf{v}\, (\partial_x u)^2\, d\mathbf{x}\, dt \\ &+ \int_{\mathbb{T}\times\mathbb{R}^{n-1}} \phi(\mathbf{x},0)u(x,0) - \Phi'(\mathbf{x},0) \cdot \mathbf{v}(\mathbf{x},0)\partial_x u(x,0)\, d\mathbf{x} = 0, \end{aligned} \tag{5.3}$$

$$\int_{\mathbb{T}\times\mathbb{R}^{n-1}} \partial_x\theta\, u - \nabla'\theta \cdot \mathbf{v}\, \partial_x u\, d\mathbf{x} = 0, \tag{5.4}$$

in which we have used the notation $\Phi = (\phi, \Phi')$ to distinguish the first component from the remaining $n-1$ components of Φ. Denoting by $[\mathbf{u}] = \mathbf{u}_+ - \mathbf{u}_-$ the jump in $\mathbf{u}$ across any smooth surface of discontinuity, S, and considering test functions whose support crosses S, (5.2) shows that

$$[\mathbf{u}] \cdot \mathbf{n} = 0, \tag{5.5}$$

where $\mathbf{n} = (\mathfrak{n}, \mathbf{n}')$ is normal to S. Then, by (5.4), we have

$$[u]\mathfrak{n} - [\partial_x u]\mathbf{v}\cdot\mathbf{n}' = 0, \text{ where } \mathbf{v} = \frac{1}{n-1}\mathbf{x}'. \tag{5.6}$$

In examining weak, frontlike, piecewise continuous solutions for which $[u] = 0$ and $[\partial_x u] \neq 0$ (see [6]), it follows that these discontinuities propagate so that $\mathbf{n}' \cdot \mathbf{x}' = 0$. A weak formulation specific to such discontinuities may be derived by means of appropriate choice of test functions from (5.3), or by observing that (2.11) may be written in conservation form as

$$\partial_t((\partial_x u)^{1-n}) + \partial_x(u(\partial_x u)^{1-n}) + (n-1)(\partial_x u)^{-n} f = 0, \ x \in \mathbb{T}. \tag{5.7}$$

We will admit weak solutions, $u(x,t)$, which satisfy the relation

$$\int_{\mathfrak{Q}} \partial_t\varphi\,(\partial_x u)^{1-n} + \partial_x\varphi\, u(\partial_x u)^{1-n} - (n-1)f\varphi(\partial_x u)^{-n}\,dx\,dt \tag{5.8}$$

$$+ \int_{\mathbb{T}} \varphi(x,0)(\partial_x u)^{1-n}(x,0)\,dx = 0 \tag{5.9}$$

for all $\varphi(x,t) \in C_0^\infty(\mathfrak{Q})$, where $\mathfrak{Q} = [0,T) \times \mathbb{T}$. Using standard Rankine–Hugoniot-type arguments [6], discontinuities in $\partial_x u$ that jump across a curve $x = \psi(t)$ are seen to satisfy

$$(-\dot{\psi} + u(\psi,t))[(\partial_x u)^{1-n}] = 0, \tag{5.10}$$

and such discontinuities therefore propagate with the flow of (2.11); i.e., $\psi(t)$ is a member of the characteristic family $\dot{\gamma} = u \circ \gamma$.

5.1. Piecewise affine solutions. We begin by commenting on the general case of periodic, N-phase, piecewise affine solutions. Given that both $\partial_x u$ and $\partial_t u$ may be discontinuous across the curves $x = \psi_i(t), 1 \leq i \leq N-1$, in order for u to remain continuous there we must have $[u](\gamma(t),t) = 0$, and so $\frac{d}{dt}[u](\gamma(t),t) = 0$. As a result, $\partial_t[u] + u\partial_x[u] = 0$, and first derivative jumps are seen to satisfy the relations

$$[\partial_t u] + u[\partial_x u] = 0 \text{ and } [\partial_x p] = 0 \tag{5.11}$$

from (2.3). Under these conditions on u, the expression for $f(t)$, which was obtained in (2.13) by integrating (2.11) for $u \in C^1(\mathbb{T})$, remains unchanged:

$$f = -\frac{n}{n-1}\int_{\mathbb{T}} (\partial_x u)^2 dx. \tag{5.12}$$

In the special case, $N = 2$, which we consider here, our form of solution is given by the periodic extension of the function

$$u(x,t) = \alpha + \begin{cases} x\,p, & x \in (0,\tilde{\phi}), \\ \tilde{\phi}p + (x-\tilde{\phi})\,q, & x \in (\tilde{\phi},\tilde{\psi}), \\ \tilde{\phi}p + (\tilde{\psi}-\tilde{\phi})\,q + (x-\tilde{\psi})\,p, & x \in (\tilde{\psi},1), \end{cases} \tag{5.13}$$

where $\tilde{\phi} = \phi - [[\phi]], \tilde{\psi} = \psi - [[\psi]]$, with $[[.]]$ denoting the "integer part" of the argument. The functions $\tilde{\phi}(t)$ and $\tilde{\psi}(t)$ are the representatives in $[0,1]$ of the phase

curves $\phi(t), \psi(t) \in (-\infty, \infty)$, which start out from $\phi(0), \psi(0) \in [0,1]$ and separate those regions where $\partial_x u(x,t)$ periodically takes on values of $p(t)$ or $q(t)$.

Proceeding heuristically for the moment, periodicity of $u(x,t)$ requires that

$$\mathcal{N}(t) = \phi(t)p(t) + (\psi(t) - \phi(t))q(t) + (1 - \psi(t))p(t) = 0. \tag{5.14}$$

Given the spatial periodicity in pressure (see (2.9)), we recall that integration of (2.3) over one period showed the integral $\int_0^1 u(x,t)dx$ to be independent of time. This allows (5.13) to be used to give an expression for $\alpha(t)$. The result may be written, for instance, in terms of $p, \tilde{\phi}$, and $\tilde{\psi}$, as

$$\alpha + \frac{p}{2}(\tilde{\phi} + \tilde{\psi} - 1) = c, \tag{5.15}$$

where we have set $c = \int_0^1 u(x,0)\,dx = \alpha(0) + \frac{p(0)-q(0)}{2}(\psi(0) - \phi(0))(\phi(0) + \psi(0) - 1)$. We will assume that $0 < \phi(0) < \psi(0) < 1$. Choosing, for convenience, $c = 0$, the "average characteristic" must propagate with speed zero and we will see, consequently, that both $\phi(t)$ and $\psi(t)$ remain in $[0,1]$ for all time. The distinctions between $\tilde{\phi}$ and ϕ, $\tilde{\psi}$, and ψ will therefore not be made further here.

Combining (2.11), (5.12), and (5.13) now gives

$$\dot{p} = \frac{1}{n-1}p^2 + f, \quad \dot{q} = \frac{1}{n-1}q^2 + f, \tag{5.16}$$

where

$$f = -\frac{n}{n-1}(\phi p^2 + (\psi - \phi)q^2 + (1 - \psi)p^2). \tag{5.17}$$

Also, by (5.10),

$$\dot{\phi} = \alpha + \phi p \tag{5.18}$$

and

$$\dot{\psi} = \alpha + \phi p + (\psi - \phi)q. \tag{5.19}$$

We next verify that (5.14) follows from the system of equations (5.16)–(5.19). Differentiation and some simplification gives

$$\begin{aligned}\dot{\mathcal{N}} &= -(\psi - \phi)(p - q)q + (\psi - \phi)q^2 + (\phi + (1 - \psi))\left(\frac{p^2}{n-1} + f\right) + (\psi - \phi)\left(\frac{q^2}{n-1} + f\right) \\ &= -(\psi - \phi)(p - q)q - ((\phi + (1 - \psi))p^2 + (\psi - \phi)q^2) \\ &= -p(q(\psi - \phi) + p(\phi + (1 - \psi))) \\ &= -p\mathcal{N},\end{aligned}$$

and so $\mathcal{N}(t) = \mathcal{N}(0)e^{-\int_0^t p(s)ds}$. In particular, taking periodic data, $\mathcal{N}(0) = 0$, means that $\mathcal{N}(t) = 0$, $t > 0$. We may therefore use (5.14) to write (5.19) as

$$(1 - \psi)^{\cdot} = -\alpha + (1 - \psi)p, \tag{5.20}$$

and again employ (5.14) to express the following phase fractions as functions of p and q:

$$\psi - \phi = \frac{p}{p-q}, \quad \phi + (1-\psi) = \frac{-q}{p-q}\,. \tag{5.21}$$

Using these relations in (5.17) leads to

$$f = \frac{n}{n-1} pq \tag{5.22}$$

from which (5.16) reduces to the autonomous system

$$\dot{p} = \frac{1}{n-1}(p^2 + n\,pq), \tag{5.23}$$

$$\dot{q} = \frac{1}{n-1}(q^2 + n\,pq). \tag{5.24}$$

Subtracting (5.18) from (5.19) implies

$$\psi(t) - \phi(t) = (\psi(0) - \phi(0)) \exp\left(\int_0^t q(s)ds\right), \tag{5.25}$$

which means that the center phase fraction does not collapse as long as $\int_0^t q(s)ds$ remains bounded away from $-\infty$. Similarly, adding (5.18) and (5.20) gives

$$\phi(t) + (1-\psi(t)) = (\phi(0) + (1-\psi(0)) \exp\left(\int_0^t p(s)ds\right), \tag{5.26}$$

and the outer phase fraction exists as long as $\int_0^t p(s)ds > -\infty$. Comparing (5.25) and (5.26) shows also that periodicity imposes the following requirement on $p(t)$ and $q(t)$ in terms of their initial phase fractions:

$$(\phi(0) + (1-\psi(0)) \exp\left(\int_0^t p(s)ds\right) + (\psi(0) - \phi(0)) \exp\left(\int_0^t q(s)ds\right) = 1. \tag{5.27}$$

Using (5.16) to compute $p - q$ next gives

$$p(t) - q(t) = (p(0) - q(0)) \exp\left(\frac{1}{n-1}\int_0^t p(s) + q(s)\,ds\right), \tag{5.28}$$

which shows $p(t) - q(t)$ cannot change sign. By (5.21), $p(t)$ and $q(t)$ consequently keep their signs as long as neither phase fraction collapses. This can alternatively be seen by considering a sketch of u and observing that p and q have opposite signs and can vanish only simultaneously. Thus, without loss of generality, we subsequently assume $p(t) > 0 > q(t)$, for at least some $t \geq 0$.

In the following lemma, we solve the system (5.23), (5.24) implicitly in order to show that two-phase solutions of the type (5.13) exist for all time.

LEMMA 5.1. *Let* $(p(t), q(t))$ *satisfy* (5.23), (5.24) *with initial data* $p(0) > 0 > q(0)$ *(respectively,* $p(0) < 0 < q(0)$*). Then* $(p(t), q(t))$ *exists for all* $t \in (-\infty, \infty)$ *and satisfies* $p(t) > 0 > q(t)$ *(respectively,* $p(t) < 0 < q(t)$*). Further,* $(p(t), q(t)) \to (0, 0)$ *as* $t \to \pm\infty$*, and* $||u(.,t)||_{C^1} + ||\partial_t u(.,t)||_{C^0} \to 0$ *as* $t \to \pm\infty$.

Proof. Writing (5.23) and (5.24) using polar variables $p(t) = r(t)\cos\theta(t)$, $q(t) = r(t)\sin\theta(t)$, $r(t) \geq 0$, leads to the following:

$$\dot{r}(t) = \frac{r^2}{n-1}(\cos^3(\theta(t)) + n(\cos\theta(t) + \sin\theta(t))\cos\theta(t)\sin\theta(t) + \sin^3(\theta(t))), \tag{5.29}$$

and

$$\dot{\theta}(t) = r(\cos\theta(t) - \sin\theta(t))\cos\theta(t)\sin\theta(t). \tag{5.30}$$

In the case $p(0) > 0 > q(0)$, $\theta(0) \in (-\pi/2, 0)$, so by (5.30) $\dot{\theta}(0) < 0$ and, as long as $\theta(t) \in (-\pi/2, 0)$, $\dot{\theta}(t) < 0$.

We will show that $\theta(t) \to -\pi/2$ as $t \to \infty$ ($\theta(t) \to 0$ as $t \to -\infty$) by using (5.29) and (5.30). Integrating the resulting expression for $\frac{d\ln r}{d\theta}$ gives

$$r(\theta) = c|\cos\theta\sin\theta|^{\frac{1}{n-1}}|\cos\theta - \sin\theta|^{-\frac{n+1}{n-1}}, \tag{5.31}$$

where $c > 0$ denotes a generic constant. Inserting this expression for $r(\theta)$ in (5.30) results in

$$\begin{aligned}\dot{\theta}(t) &= c|\cos\theta\sin\theta|^{\frac{1}{n-1}}|\cos\theta - \sin\theta|^{-\frac{n+1}{n-1}}(\cos\theta - \sin\theta)\cos\theta\sin\theta \\ &= -c|\cos\theta\sin\theta|^{\frac{n}{n-1}}|\cos\theta - \sin\theta|^{-\frac{2}{n-1}} \qquad (5.32)\end{aligned}$$

for $\theta(t) \in (-\pi/2, 0)$, and so

$$\int_{\theta(0)}^{\theta(t)} \frac{|\cos\theta - \sin\theta|^{\frac{2}{n-1}}}{|\cos\theta\sin\theta|^{\frac{n}{n-1}}}d\theta = -ct, \ -\frac{\pi}{2} < \theta(0) < 0, \ c > 0. \tag{5.33}$$

Since $\frac{n}{n-1} > 1$, the integral expression diverges, both as $\theta(t) \to -\pi/2$ ($t \to +\infty$) and as $\theta(t) \to 0$ ($t \to -\infty$). Thus $\theta(t)$ and, from (5.31), $r(t)$, are bounded, continuous functions of time, with $\theta(t) \in (-\pi/2, 0)$. By (5.31) then, $r(t) \to 0$ as $t \to \pm\infty$, and, noting that from (5.14), (5.15),

$$\alpha(t) = -\frac{p(t) - q(t)}{2}(\psi(t) - \phi(t))(\phi(0) + \psi(0) - 1)\,,$$

it follows from (5.35) and (5.36) that

$$p(t), q(t), \alpha(t), \text{ and } u(x,t) \to 0 \text{ as } t \to \pm\infty.$$

The remaining conclusions are easily obtained. ∎

Finally, we show that for $c = 0$ the phases remain in the interval $[0, 1]$ for all time.

THEOREM 5.1. *Suppose $c = 0$ and $0 < \phi(0) < \psi(0) < 1$. Then the phases ϕ, ψ stay in $[0, 1]$. In particular*

$$\phi(t) \in [\phi(0), \tfrac{1}{2}(\phi(0) + \psi(0)) \ \textit{and} \ \psi(t) \in (\tfrac{1}{2}(\phi(0) + \psi(0)), \psi(0)]$$

for all $t > 0$. Further, the time asymptotic behavior satisfies

$$\lim_{t\to\infty} \phi(t) = \tfrac{1}{2}(\phi(0) + \psi(0)) = \lim_{t\to\infty} \psi(t).$$

Proof. With (5.15) giving α (for $c = 0$), substituting into (5.18) and (5.20) shows that, by (5.26),

(5.34)

$$\dot{\phi}(t) = (1-\psi(t))^{\cdot} = \frac{p}{2}(\phi + (1-\psi)) = \frac{1}{2}(\phi(0) + (1-\psi(0)))p(t)\exp\left(\int_0^t p(s)ds\right),$$

and the individual phases therefore satisfy

$$\phi(t) = \phi(0) + \frac{1}{2}(\phi(0) + (1-\psi(0)))\left(\exp\left(\int_0^t p(s)ds\right) - 1\right) \tag{5.35}$$

and

$$\psi(t) = \psi(0) - \frac{1}{2}(\phi(0) + (1-\psi(0)))\left(\exp\left(\int_0^t p(s)ds\right) - 1\right). \tag{5.36}$$

Now we examine (5.27). Assuming $q < 0 < p$, the first term is monotonically increasing in time and must converge, for $0 < \phi(0) < \psi(0) < 1$, to a positive limit. The second term is positive but monotonically decreasing, so it may converge, as $t \to \infty$, either to a positive limit or to zero. Formally, setting $0 < \int_0^\infty p(t)dt = L < \infty$ and $-\infty \le \int_0^\infty q(t)dt = M < 0$, (5.27) and (5.28) imply

$$(\phi(0) + (1-\psi(0)))e^L + (\psi(0) - \phi(0))e^M = 1 \tag{5.37}$$

and

$$\lim_{t\to\infty}(p(t) - q(t)) = (p(0) - q(0))e^{\frac{L+M}{n+1}}. \tag{5.38}$$

If M is finite and $p(0) \neq q(0)$, then (5.38), together with the necessity for $p(t)$ to approach zero as $t \to \infty$, means that $\lim_{t\to\infty} q(t) \neq 0$. However, this implies that $M = -\infty$, a contradiction. On the other hand, if $M = -\infty$, then, since $L < \infty$, (5.38) requires that $\lim_{t\to\infty} q(t) = 0$, which is permitted. We conclude that, for smooth solutions $p(t), q(t)$ to exist with $p(0) > 0 > q(0)$, we require that $L = \int_0^\infty p(t)dt < \infty$ and $M = \int_0^\infty q(t)dt = -\infty$. Thus, by (5.37),

$$e^L = \frac{1}{\phi(0) + 1 - \psi(0)}. \tag{5.39}$$

Writing (5.35) in the form

$$\phi(t) = \frac{1}{2}(\phi(0) - (1-\psi(0))) + \frac{1}{2}(\phi(0) + (1-\psi(0)))\exp\int_0^t p(s)ds, \tag{5.40}$$

it follows that

$$\lim_{t\to\infty}\phi(t) = \frac{1}{2}(\phi(0) + \psi(0)) \tag{5.41}$$

and, similarly,

$$\lim_{t\to\infty}\psi(t) = \frac{1}{2}(\phi(0) + \psi(0)). \tag{5.42}$$

By the monoticity in time of $\int_0^t p(s)ds$, therefore $\phi(t) \in [\phi(0), \frac{1}{2}(\phi(0) + \psi(0))$ and $\psi(t) \in (\frac{1}{2}(\phi(0) + \psi(0)), \psi(0)]$ for all $t > 0$.

Acknowledgment. The first author would like to thank Barbara Keyfitz, colleagues, and staff for their support at the Fields Institute, Toronto, where part of this work was done.

REFERENCES

[1] V. Arnold, *Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits*, Ann. Inst. Fourier (Grenoble), 16 (1966), pp. 319–361.
[2] F. Calogero, *A solvable nonlinear wave equation*, Stud. Appl. Math., 70 (1984), pp. 189–199.
[3] S. Childress, G. R. Ierley, E. A. Spiegel, and W. R. Young, *Blow-up of unsteady two-dimensional Euler and Navier-Stokes solutions having stagnation-point form*, J. Fluid Mech., 203 (1989), pp. 1–22.
[4] P. Constantin, *The Euler equations and nonlocal conservative Riccati equations*, Int. Math. Res. Not., 9 (2000), pp. 455–465.
[5] S. M. Cox, *Two-dimensional flow of a viscous fluid in a channel with porous walls*, J. Fluid Mech., 227 (1991), pp. 1–33.
[6] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Grundlehren Math. Wiss. 325, Springer-Verlag, Berlin, 2000.
[7] D. G. Ebin, *A concise presentation of the Euler equations of hydrodynamics*, Comm. Partial Differential Equations, 9 (1984), pp. 539–559.
[8] D. G. Ebin and J. Marsden, *Groups of diffeomorphisms and the motion of an incompressible fluid*, Ann. of Math. (2), 92 (1970), pp. 102–163.
[9] J. Eells, *A setting for global analysis*, Bull. Amer. Math. Soc., 72 (1966), pp. 751–807.
[10] A. A. Himonas and G. Misiołek, *The Cauchy problem for an integrable shallow-water equation*, Differential Integral Equations, 14 (2001), pp. 821–831.
[11] A. A. Himonas and G. Misiołek, *High frequency smooth solutions and well-posedness of the Camassa-Holm equation*, Int. Math. Res. Not., 51 (2005), pp. 3135–3151.
[12] G. Misiołek, *Classical solutions of the periodic Camassa-Holm equation*, Geom. Funct. Anal., 12 (2002), pp. 1080–1104.
[13] C. C. Lin, *Note on a class of exact solutions in magnetohydrodynamics*, Arch. Rational Mech. Anal., 1 (1958), pp. 391–395.
[14] H. Okamoto and J. Zhu, *Some similarity solutions of the Navier-Stokes equations and related topics*, Taiwanese J. Math., 4 (2000), pp. 65–103.
[15] J. T. Stuart, *Nonlinear Euler partial differential equations: Singularities in their solution*, in Applied Mathematics, Fluid Mechanics, Astrophysics, World Scientific, Singapore, 1988, pp. 81–95.
[16] H. Weyl, *On the differential equations of the simplest boundary-layer problems*, Ann. of Math. (2), 43 (1942), pp. 381–407.
[17] W. Wolibner, *Un theorème sur l'existence du mouvement plan d'un fluide parfait, homogene, incompressible, pendant un temps infiniment long*, Math. Z., 37 (1933), pp. 698–726.

SIAM J. MATH. ANAL.
Vol. 40, No. 4, pp. 1516–1549

BREAKING OF SYMMETRICAL PERIODIC SOLUTIONS IN A SINGULARLY PERTURBED KDV MODEL*

ALEXANDER TOVBIS†

Abstract. There are several recent developments in the well-known problem of breaking of homoclinic orbits (splitting of separatrices) of a system that undergoes a singular perturbation. First, survival of a homoclinic orbit is an exceptional situation that can be linked to triviality of the Stokes phenomenon of the underlying "truncated" equation. Second, homoclinic connections to exponentially small periodic orbits survive the perturbation in the generic case. In this paper we consider a different problem: we study deformations of "genuine" periodic orbits of the second order equation $y'' = y + y^2$ that undergoes the singular perturbation $\varepsilon^2 y'''' + (1-\varepsilon^2)y'' = y + y^2$, where $\varepsilon > 0$ is a small parameter. We prove that if the period and the constant of motion do not change too rapidly (in ε), a genuine (nontrivial) periodic solution does not survive the perturbation.

Key words. singular perturbations, periodic solutions, exponentially small phenomena

AMS subject classifications. 34E, 34M, 34D15, 37J

DOI. 10.1137/070694053

1. Introduction.

1.1. Breaking of homoclinic connections. Let $y = 0$ be a hyperbolic stationary point of an n-dimensional differential equation $y' = f(y)$, and let W_s, W_u be the stable and unstable manifolds at $y = 0$. It is said that the stationary point $y = 0$ has a homoclinic connection if there exists a phase trajectory, originating and ending at $y = 0$, that lies on $W_s \cap W_u$. For example, the (bounded) separatrix solution

$$y(x) = -\frac{3/2}{\cosh^2(x/2)} \tag{1}$$

to

$$y''(x) = y(x) + y^2(x) \tag{2}$$

corresponds to the homoclinic connection of the stationary point $(0, 0)$ in the phase plane of (2). There exists a general problem to describe how a singular perturbation of the original equation affects the homoclinic (or a heteroclinic) connection. A large number of particular singularly perturbed equations, originating from the corresponding physical or computational problems, have been discussed in the literature (see, for example, [STL] and the references therein). In particular, the singular perturbation

$$\varepsilon^2 y''''(x) + (1-\varepsilon^2)y''(x) = y(x) + y^2(x) \tag{3}$$

of (2) is related to the traveling wave reduction of a fifth order KdV equation that models gravity water waves. Existence or nonexistence of homoclinic connections of (3) has been studied in a number of papers; see, for example, [HM], [PRG], [GJ]. A rigorous proof of nonexistence of such connections was obtained in [AM] (see also

*Received by the editors June 8, 2007; accepted for publication (in revised form) June 4, 2008; published electronically November 19, 2008. This work was supported by NSF grant DMS 0508779.
http://www.siam.org/journals/sima/40-4/69405.html
†Department of Mathematics, University of Central Florida, Orlando, FL 32816 (atovbis@pegasus.cc.ucf.edu).

[Eck]) and later on in [To4], where the "asymptotic beyond all orders" approach, first suggested in [KS] for a simple crystal growth model, was put on a rigorous basis. This approach is, in a sense, a natural way to study exponentially small phenomena, such as the mismatch between the stable and the unstable manifolds of (3). The goal of the present paper is to show that a similar argument can be used to prove nonexistence of symmetrical periodic solutions to (3) (see below), subject to some additional requirements.

The approach of [KS] is to rescale (3), which is called the outer equation, to the inner equation

$$v''''(z) + (1-\varepsilon^2)v''(z) = \varepsilon^2 v(z) + v^2(z), \tag{4}$$

where $x = \varepsilon z$ and $v = \varepsilon^2 y$. The main advantage of (4) versus (3) is that the (exponentially small) difference between the stable and unstable solutions is detectable even in the leading order part

$$v''''(z) + v''(z) = v^2(z), \tag{5}$$

of (4), which is called the truncated equation (4). This means that the difference between the stable and unstable solutions to (3) can be studied through the Stokes phenomenon of (5). In fact, in [TP] we considered the family of singular perturbations

$$\varepsilon^2 y^{(\mathrm{iv})} + (1-\varepsilon^2)y'' - y = y^2 + \varepsilon^2\gamma\left(2yy'' + (y')^2\right) \tag{6}$$

of (2), parametrized by $\gamma \in \mathbb{R}$, and found exact conditions for persistence of the homoclinic connection (under the singular perturbation) in terms of the Stokes constant for the corresponding truncated inner equation. The technique developed in [To4] can be modified to consider singularly perturbed problems in other settings. For example, it was used in [To5] to show that the discretized equation (3) does not have a homoclinic connection.

Notice that persistence of the homoclinic connection under singular perturbations (3), (6) is equivalent to existence of a symmetrical (even) stable solution to the corresponding equation. Since the family of stable solutions to (3) (or to (6)) is one-dimensional (translation), we can arrange for $y'(0) = 0$; then the symmetry of the stable solution $y(x)$ is equivalent to $y'''(0) = 0$.

1.2. Deformation of periodic solutions under singular perturbations. The problem of deformations of periodic solutions under singular perturbations has not yet received as much attention as the problem of homoclinic connections. However, it is known that there exist symmetrical periodic solutions to (3) that are exponentially small in ε; see [Lo] and later papers [IL1], [IL2]. We consider such periodic solutions as deformations of a constant solution $y \equiv 0$ of the unperturbed equation (2). Existence of even periodic solutions that are deformations of a trivial solution was also proved in [BPBA] for a certain class of reversible fourth order Hamiltonian systems with zero Hamiltonian. Discussion of even periodic solutions for some fourth order ODEs with nonzero Hamiltonian and cubic nonlinearity ((3) has a quadratic nonlinearity) can be found in [PT].

In the present paper we extended the technique of [To4] to study deformations of a genuine (nonconstant) periodic solution of (2). We limit our attention to periodic solutions inside the potential well of (2). Any such solution is given by the Weierstrass elliptic $\wp$-function with invariants g_2, g_3,

$$y(x) = 6\wp_{g_2,g_3}(x - x_0) - \frac{1}{2}, \tag{7}$$

and has an integral of motion

$$y'^2 = y^2 + \frac{2}{3}y^3 + C, \tag{8}$$

where $x_0 \in \mathbb{C}$ is a translation and $g_2 = \frac{1}{12}$, $g_3 = -\frac{1}{36}(C + \frac{1}{6})$ (see section 1.3). The requirement $C \in (-\frac{1}{3}, 0)$ on the constant of motion (energy) C guarantees that solution $y(x)$ is inside the potential well. (The value $C = 0$ corresponds to the trivial solution and to the separatrix solution of (2), whereas $C = -\frac{1}{3}$ corresponds to the stationary solution $y(x) \equiv -1$.) Solution (7) is a periodic function with finite basic real and purely imaginary periods

$$2\omega_1 = \int_{e_1}^{\infty} \frac{du}{\sqrt{4u^3 - \frac{1}{12}u - g_3}}, \quad 2\omega_3 = i\int_{-\infty}^{e_3} \frac{du}{\sqrt{4u^3 - \frac{1}{12}u - g_3}}, \tag{9}$$

where e_1, e_3 are the largest and the smallest real roots of $4u^3 - \frac{1}{12}u - g_3 = 0$ (see below), respectively. It is also an even function provided that x_0 is an integer linear combination of ω_1, ω_3. It will be convenient for us to choose

$$x_0 = \omega_3, \tag{10}$$

so that the unperturbed solution (7) is bounded, real-valued, symmetrical, and periodic with the period $2\omega_1$ on $x \in \mathbb{R}$. Equivalently, we can consider solution (7) with $x_0 = 0$ along the horizontal line $\omega_3 + \mathbb{R}$ in the complex x-plane.

The perturbed equation (3) also has an integral of motion

$$\varepsilon^2(2y'''y' - y''^2) + (1 - \varepsilon^2)y'^2 - y^2 - \frac{2}{3}y^3 = C(\varepsilon), \tag{11}$$

where $C(\varepsilon)$ is the constant of motion (energy). Since we are interested in deformations of periodic solutions satisfying (8), we require $C(0) = C \in (-\frac{1}{3}, 0)$.

Let $\alpha \geq 1$. We say that a solution $y(x, \varepsilon)$ to the perturbed equation (3) is a C^α-deformation of a periodic solution (7) (in the potential well) to the unperturbed equation (2) under the singular perturbation (3) on interval S if

$$Y(x, \varepsilon) = Y(x, 0) + \varepsilon^\alpha \tilde{Y}(x, \varepsilon), \tag{12}$$

where vector $\tilde{Y}(x, \varepsilon)$ is continuous in ε, uniformly on interval S of the x-axis. Here $Y(x, \varepsilon) = \text{Col}\ (y(x, \varepsilon), y'(x, \varepsilon), y''(x, \varepsilon), y'''(x, \varepsilon))$. Note that a C^α-deformation of a periodic solution (7) has constant of motion (11) satisfying

$$C(\varepsilon) = C + \varepsilon^\alpha \tilde{C}(\varepsilon), \tag{13}$$

where $\tilde{C}(\varepsilon)$ is a continuous function.

Our main result is the following theorem.

THEOREM 1.1. *Let $y(x, 0) = 6\wp_{\frac{1}{12}, g_3}(x - \omega_3) - \frac{1}{2}$, where $|g_3| < 6^{-3}$, be a periodic solution (inside the potential well) of the nonperturbed equation (2). Let $n \in \mathbb{N}$ and $y(x, \varepsilon)$ be a C^α-deformation, $\alpha \geq 1$, of $y(x, 0)$ under the perturbation (3) on some open interval S that contains the segment $[\omega_1, n\omega_1]$. Then the deformation $y(x, \varepsilon)$ does not contain a sequence of symmetric periodic solutions $y(x, \varepsilon_m)$ of (3) with periods $2k\omega_1(\varepsilon_m)$, where $k = 1, 2, \ldots, n$, $\lim_{m\to\infty} \varepsilon_m = 0$, and $\omega_1(\varepsilon)$ is subject to*

$$\omega_1(\varepsilon) - \omega_1 = \varepsilon^{\frac{\alpha}{2}}\tau(\varepsilon), \tag{14}$$

where $\tau(\varepsilon)$ is a continuous function for small $\varepsilon \geq 0$.

Similarly to the homoclinic connection problem, the central idea of our proof here is uniform (in small $\varepsilon \geq 0$) control of an iterative solution of inner equation (4). In a certain sense, though, the homoclinic connection problem is simpler, because the manifolds of stable and unstable solutions to (3) are one-dimensional, whereas deformations of a periodic solution to (2) under the perturbation (3) form a manifold of the full dimension. We rescale the outer equation (3) to the inner equation (4) by $x = \varepsilon z$, $v = \varepsilon^2 y$ and approach the problem through the following sequence of steps: (a) we construct by iterations a two-parameter family $\mathcal{F}_\alpha$ of deformations of the rescaled periodic solution (7), (10), satisfying (4); (b) we show that solutions $v(z, \varepsilon)$ to the inner equation (4) that corresponds to a C^α-deformation $y(x, \varepsilon)$ of a periodic solution (7) with the period $2\omega_1$ belong to the family $\mathcal{F}_\alpha$; and (c) for any $k \in \mathbb{N}$ we prove that the family $\mathcal{F}_\alpha$ does not contain a sequence $\{v(z, \varepsilon_m)\}_1^\infty$ of symmetrical and periodic (with the period $2k\frac{\omega_1(\varepsilon_m)}{\varepsilon_m}$) solutions if $\omega_1(\varepsilon)$ is subject to (14).

The paper is organized in the following way. In the remaining part of this section we describe the Stokes phenomenon for the truncated inner equation (5) and some connections between deformations of homoclinic and periodic orbits of (3). Solution of (3) by iterations, which yields the two-parameter family of solutions $\mathcal{F}_\alpha$, is obtained in sections 2–3. Finally, the proof of Theorem 1.1 is completed in section 4.

1.3. Periodic solutions of the unperturbed equation. It is easy to check that the values C from the interval $(-\frac{1}{3}, 0)$ define periodic solutions of the unperturbed equation (2) within the well of the potential $-y^2 - \frac{2}{3}y^3$. After the change of variables $y = 6X - \frac{1}{2}$, (8) is reduced to

$$X'^2 = 4X^3 - \frac{1}{12}X - g_3, \tag{15}$$

where $g_3 = -\frac{1}{36}(C + \frac{1}{6})$. The solution to this equation is given by the Weierstrass elliptic function $\wp(x) = \wp_{g_2,g_3}(x)$, where the invariant $g_2 = \frac{1}{12}$ and g_3 is defined above. Thus, periodic solutions to the unperturbed equation (2) are given by

$$y(x) = 6\wp_{g_2,g_3}(x) - \frac{1}{2}. \tag{16}$$

The number $\Delta = g_2^3 - 27g_3^2$ is called the discriminant of $\wp_{g_2,g_3}(x)$. The condition $\Delta > 0$ is equivalent to $|g_3| < 6^{-3}$, i.e., to $C \in (-\frac{1}{3}, 0)$. Under this condition the Weierstrass function $\wp$ has real period $2\omega_1$ defined by (9) and purely imaginary period $2\omega_3$, where $\omega_1, -i\omega_3$ are positive numbers (the latter requires correct choice of the branch of the square root in (9)). Using the standard notation, we denote $\omega_2 = \omega_1 + \omega_3$ and $e_j = \wp(\omega_j)$, where $j = 1, 2, 3$. It is well known that (see, for example, [GR]) e_j are the roots of the cubic polynomial in (15),

$$\wp'(\omega_j) = 0,\ j = 1, 2, 3, \quad e_1 + e_2 + e_3 = 0, \qquad \text{and} \qquad e_3 \leq e_2 \leq e_1. \tag{17}$$

When x varies along the real axis, the value of $\wp$ varies between e_1 and $+\infty$ (as $\wp(x)$ is an even function that has a second order pole at the origin with the principle part $\frac{1}{x^2}$); see Figure 1.

This corresponds to the periodic solution of (2), defined by (16), with the range from $\tilde{e}_1$ to infinity outside the potential well, i.e., to the unbounded branch of the periodic orbit (the energy of the motion is fixed). Here $\tilde{e}_j = 6e_j - \frac{1}{2}$, $j = 1, 2, 3$; see Figure 2.

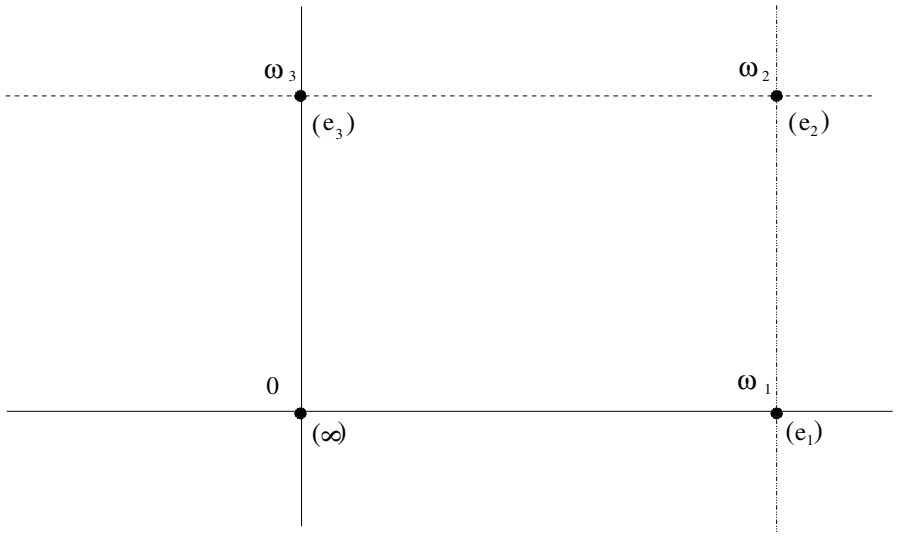


FIG. 1. *Part of the parallelogram of periods* $0\omega_1\omega_2\omega_3$ *for the Weierstrass functions* $\wp(x)$ *in the complex* x*-plane together with the corresponding values of* $\wp$ *shown in the brackets.*

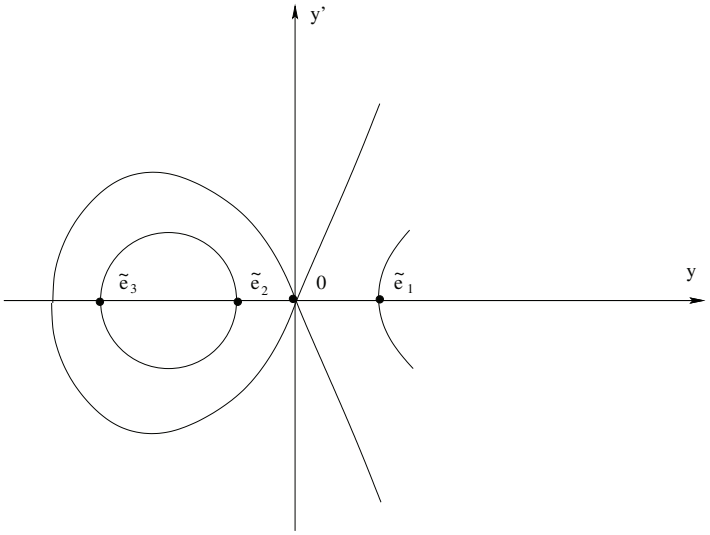


FIG. 2. *Phase portrait of* (2) *showing periodic solutions* ($\Delta > 0$) *of the same period inside (bounded) and outside (unbounded) the potential well. Here* $\tilde{e}_j = 6e_j - \frac{1}{2}$, $j = 1, 2, 3$.

When x varies along the horizontal line $\Im x = \omega_3$, the value of $\wp$ varies between e_3 and e_2. That corresponds to the periodic solution of (2), defined by (16), with the range from $\tilde{e}_3$ to $\tilde{e}_2$ inside the potential well, i.e., to the bounded branch of the periodic orbit. When x varies along the vertical line $\Re x = \omega_1$, the value of $\wp$ varies between e_1 and e_2. That corresponds to the tunneling between the unbounded and bounded branches of the periodic orbit defined by the solution (16); see Figures 1 and 2. For $\Delta < 0$ the periods ω_1 and ω_3 are complex conjugated so that ω_2 is real. In this case the value of $\wp$ varies between e_2 and $+\infty$, which corresponds to infinite periodic motion outside the potential well; see Figure 3.

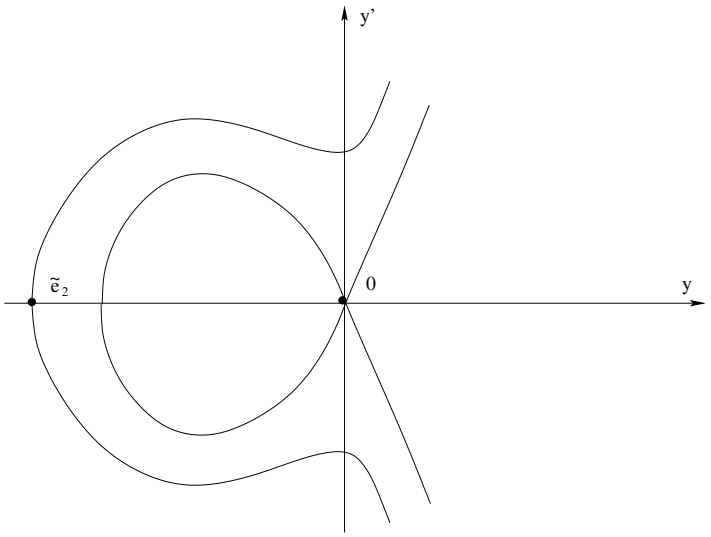


FIG. 3. *Phase portrait of* (2) *showing an unbounded periodic solution* ($\Delta < 0$) *outside the potential well. Here* $\tilde{e}_2 = 6e_2 - \frac{1}{2}$.

Here and henceforth it will be convenient for us to consider x_0 from (7) to be zero. Then we are interested in the behavior of $y(x)$ along the horizontal line $\omega_3(\varepsilon) + \mathbb{R}$.

1.4. Solutions to the perturbed equation. Considering perturbations of the separatrix solution (1) of the original equation, we focus our attention on the stable and unstable solutions of the perturbed equation (3) (those are solutions corresponding to the one-dimensional manifolds W_s and W_u in the phase space of (3)). We define the unstable and stable solutions to (3) as

$$y_{u,s}(x,\varepsilon) = \sum_{k=1}^{\infty} y_k(\varepsilon) e^{\pm kx}, \tag{18}$$

respectively, where $y_1(\varepsilon)$ as an arbitrary continuous positive function and all y_k, $k > 1$, are uniquely defined through y_1 [To4]. The series (18) is convergent in properly chosen left- and right-half planes of the complex x-plane, respectively. It is easy to check that for $\varepsilon = 0$ the choice of $y_1(0) = 6$ yields the unbounded separatrix solution

$$y(x) = \frac{3/2}{\sinh^2(x/2)}, \tag{19}$$

which is related to the bounded separatrix solution (1) through the shift $x \mapsto x + i\pi$. Considering x as a complex variable, now and henceforth we refer to (19) as the separatrix solution of (3).

Perturbations of periodic solutions to the original equation (2) that we are interested in are not as clearly identifiable as solutions (18) in the separatrix case. We start with considering the separatrix solution of (3) as the limit of periodic solutions (16) as $C \to 0^-$. In this limit the real half-period $\omega_1 \to \infty$, and the imaginary half-period $\omega_3 \to i\pi$. Thus, the point $\omega_2 \to i\pi + \infty$ as $C \to 0^-$.

Note that the stable and unstable solutions (18) are symmetrical with respect to $x = 0$ since $y_s(x,\varepsilon) = y_u(-x,\varepsilon)$. Alternatively, we can consider $y_u(x)$ as the

symmetrical continuation of $y_s(x)$ through $x = \infty$ on the compactification of $\mathbb{R}$. The same observation is correct for the horizontal line $\Im x = \pi$ and $x = i\pi + \infty$. The analogue of the point $x = i\pi + \infty$ in the periodic case is $x = \omega_2$. Therefore, by analogy with the separatrix case, we will consider perturbations of the periodic solution (16) that *are symmetrical with respect to the point* $\omega_2(\varepsilon)$. Equivalently, that means that the first and third derivatives of $y(x)$ turn zero at $x_0 = \omega_2$. The remaining two of the total of four conditions to define $y(x, \varepsilon)$ are $y'(\omega_3(\varepsilon), \varepsilon) = 0$ and the integral of motion (11). Thus, $y(x, \varepsilon)$ satisfies third order differential equation (11) with the prescribed energy $C(\varepsilon)$ together with three boundary conditions

$$y'(\omega_3(\varepsilon), \varepsilon) = 0, \quad y'(\omega_2(\varepsilon), \varepsilon) = y'''(\omega_2(\varepsilon), \varepsilon) = 0, \tag{20}$$

where $C(0) = C$ and $2\omega_1(\varepsilon), 2\omega_3(\varepsilon)$ are the fundamental periods of the Weierstrass function $\wp$ in the solution (16) of (8).

It is easy to see that if $f(x)$ is an even (symmetrical at $x = 0$) periodic function with a period $2\omega_1$, then $f(x)$ is symmetrical at any point $x = k\omega_1$, $k \in \mathbb{N}$. Since a solution $y(x, \varepsilon)$ to the BVP (11), (20) is symmetrical at $x = \omega_2(\varepsilon)$ on the horizontal line $\Im x = \Im\omega_3(\varepsilon)$, $y(x, \varepsilon)$ is an even periodic solution if and only if

$$y'''(\omega_3(\varepsilon), \varepsilon) = 0. \tag{21}$$

1.5. The Stokes phenomenon for the inner equation. Periodic solutions (16) to (2) have second order poles at $x = 0$. The asymptotic beyond all orders approach of [KS] suggests blowing up this singularity by rescaling

$$x = \varepsilon z, \quad y(x, \varepsilon) = \varepsilon^{-2} v(z, \varepsilon), \tag{22}$$

where z is a new independent complex variable. The rescaled equation (inner equation) is given by (4), which we can rewrite as

$$(D_z^2 + 1)(D_z^2 - \varepsilon^2)v = v^2. \tag{23}$$

Let $v_+(z, \varepsilon)$ and $v_-(z, \varepsilon)$ denote, respectively, the rescaled solutions to the BVP (11), (20) and to its "symmetrical" problem, where conditions at $\omega_2(\varepsilon)$ are replaced by the same conditions at $-\bar{\omega}_2(\varepsilon) = -\omega_1(\varepsilon) + \omega_3(\varepsilon)$, i.e., $y'(-\bar{\omega}_2(\varepsilon), \varepsilon) = y'''(-\bar{\omega}_2(\varepsilon), \varepsilon) = 0$. Our observation (proved below) is that, similarly to the case of the separatrix solution, the leading order (in ε) of the difference between $v_+(z, \varepsilon)$ and $v_-(z, \varepsilon)$ can be derived from the truncated equation

$$(D_z^2 + 1)D_z^2 v = v^2. \tag{24}$$

Note that $z = \infty$ is the irregular singular point of (24) and that this equation has a unique formal power series solution

$$\hat{v}(z) = \sum_{k=1}^{\infty} \frac{v_k}{z^{2k}}, \tag{25}$$

where $v_1 = 6$.

The inverse Laplace transform $\mathcal{L}^{-1}$ converts the truncated inner equation (24) into the convolution equation

$$(p^4 + p^2)V(p) = V(p) * V(p), \tag{26}$$

where p is a dual variable (called Borel variable), $V(p) = [\mathcal{L}^{-1}v](p)$, and $F(p)*G(p) = \int_0^p F(p-\tau)G(\tau)d\tau$. It is well known that the asymptotic expansion (25) of $v(z)$ yields the corresponding asymptotic expansion in powers of p of $V(p)$ at $p = 0$.

THEOREM 1.2. *Equation* (26) *admits a unique nontrivial power series solution in odd powers of* p. *This solution defines a function* $V(p)$ *that is analytic at the whole* p*-plane except for two vertical rays: from* $p = i$ *upward and from* $p = -i$ *downward. The function* $V(p)$ *is of exponential order* 1 *along any nonvertical ray in this cut* p*-plane.*

This theorem follows from a more general statement (Main Theorem) of [To2].

COROLLARY 1.3. *Let* $v_\pm(z)$ *be defined by*

$$v_\pm(z) = \int_0^{\pm\infty} e^{-zp}V(p)dp. \tag{27}$$

These functions are the only analytic solutions of the truncated inner equation (24) *that satisfy*

$$v_\pm(z) \sim \hat{v}(z) \quad \text{as } z \to \infty, \quad z \in S_\pm, \tag{28}$$

where $S_\pm$ *are sectors*

$$S_+ = \{z : \ |\arg z| < \pi\}, \qquad S_- = \{z : \ |\arg z - \pi| < \pi\}. \tag{29}$$

Moreover,

$$v_+(z) - v_-(z) = -2\pi i s e^{iz}(1 + o(1)) \qquad \text{as} \qquad z \to \infty, \quad 0 < \arg z < \pi, \tag{30}$$

where the constant s *is determined through*

$$s = \lim_{p \to -i} (p + i)V(p). \tag{31}$$

Proof. The Taylor expansion of $V(p)$ at $p = 0$ can be obtained by applying the inverse Laplace transform (Borel transform) to the formal series $\hat{v}(z)$. The function $V(p)$ is analytic on a Riemann surface that has possible branch points of the logarithmic type at $p = ik$, where $k \in \mathbb{Z} \setminus \{0\}$. The statement on the behavior of $V(p)$ at singularities $p = \pm i$ follows from Theorem 2.2 in [To2]. The uniqueness of solutions $v_\pm(z)$ satisfying (28) follows from Theorem 1.2 and properties of the Laplace transform. □

DEFINITION 1.4. *The constant* s *in* (31) *is called the Stokes constant for the truncated inner equation* (24).

PROPOSITION 1.5. *The following three conditions are equivalent:* (i) *the formal power series solution* (25) *has a positive radius of convergence;* (ii) $s = 0$*; and* (iii) $v_+(z) \equiv v_-(z)$.

Proof. The fact that (i) implies (iii) is obvious. The inverse statement follows from the fact that $v(z) = v_+(z) \equiv v_-(z)$ implies that the function $v(z)$ is single-valued near infinity and has asymptotic expansion $\hat{v}(z)$ in the full neighborhood of infinity (see [Wa]). It is also clear that (i) implies (ii), since in this case $\mathcal{L}^{-1}\hat{v}(z)$ is an entire function. Suppose now that $s = 0$. Since $V(p)$ is real-analytic on $\mathbb{R}$, we obtain that $V(p) = o(\frac{1}{p-(\pm i)})$ as $p \to \pm i$. That means, according to Corollary 2.2 in [To2], that $v_+(z)$ coincides with $v_-(z)$ in both the lower and upper z half-planes. Thus (ii) implies (iii). □

Divergence (zero radius of convergence) of the formal power series (25) was proved in [To2, p. 247]. According to Proposition 1.5, this implies that $v_+(z) \not\equiv v_-(z)$. (In fact, nonzero numerical values of s for equations, equivalent to (24), were computed in [PRG], [GJ] by means of formal Borel summation.) This implies, at least at the formal level, that solution $v_+(z,\varepsilon)$ to (23) does not coincide with solution $v_-(z,\varepsilon)$ to (23). To make this statement rigorous, it is sufficient to show that (a)

$$v_\pm(z,\varepsilon) \to v_\pm(z) \tag{32}$$

as $\varepsilon \to 0$ in some regions $R_\pm$ uniformly in $z \in R_\pm$, and (b) intersection $R_+ \cap R_-$ contains some segment of a positive length.

2. Inner equation. In the following theorem (Theorem 2.1) we construct a two-parameter family of solutions to (23) that are symmetrical at $z = \tilde{\omega}_2(\varepsilon)$ and that converge to $v_+(z)$, or to its translation $v_+(z-h)$, $h \in \mathbb{R}$, as $\varepsilon \to 0$ uniformly in the corresponding region (see below) of the complex z-plane. Here $\tilde{\omega}_j(\varepsilon) = \frac{\omega_j(\varepsilon)}{\varepsilon}$, $j = 1, 2, 3$.

This result constitutes an essential part of the method of [KS]; convergence of $v_+(z,\varepsilon)$ to $v_+(z)$ was mentioned only briefly in the original paper [KS] (with regard to a third order equation considered there), but explicit formulations and proofs were omitted. In other papers, connection between solutions of (23) and (24) was considered only on the formal level. The proof of convergence of $v_+(z,\varepsilon)$ to $v_+(z)$ in the separatrix case was given in [To4] (see also [To5] for the discretized equation (2), as well as [TP] for a more general family of singular perturbations of (2)). The major difficulty there was based on the fact that solutions of the full and of the truncated inner equations have different rates of convergence to 0 as $z \to \infty$ (in proper directions): $v_+(z,\varepsilon)$, $\varepsilon > 0$, approaches 0 exponentially fast (in z), while $v_+(z)$ has only power order convergence. One needs to find, however, some uniform majorization that will cover both cases. In order to find such a majorization, we suggested [To4], [To5] linearizing (23) around the unperturbed solution $6q^2$, where $q(z,\varepsilon) = \frac{\varepsilon}{\sinh\frac{\varepsilon z}{2}}$, and applying the contraction mapping principle to the obtained equation. A similar approach will be developed below with the natural choice

$$q(z,\varepsilon) = \varepsilon\sqrt{\wp_{\frac{1}{12},g_3(\varepsilon)}(\varepsilon z) - \frac{1}{12}}, \tag{33}$$

where $g_3 = g_3(\varepsilon)$ is determined through $\omega_1 = \omega_1(\varepsilon)$ by (9).

2.1. Linearized equation. The substitution

$$v(z,\varepsilon) = u(z,\varepsilon) + 6c(\varepsilon)q^2(z,\varepsilon), \tag{34}$$

where $q(z,\varepsilon)$ is given by (33), reduces (23) to

$$(D_z^2+1)(D_z^2-\varepsilon^2)u = u^2 + 12c(\varepsilon)q^2u + f(q). \tag{35}$$

Here $c = c(\varepsilon)$ is a continuous function satisfying $c(0) = 1$. Calculations of f after some algebra yield

$$f(q) = c\left[36(c-1-4\varepsilon^2)q^4 - 6!q^6 - 2\varepsilon^6 J(\varepsilon)\right], \tag{36}$$

where

$$J(\varepsilon) = -36g_3(\varepsilon) - \frac{1}{6} \tag{37}$$

is determined through $\omega_1 = \omega_1(\varepsilon)$. (To simplify notation, we omit the ε dependence in ω_j or $\tilde{\omega}_j$, $j = 1, 2, 3$, provided that such omission cannot cause any misunderstanding.) Note that $\lim_{\varepsilon\to 0} 6c(\varepsilon)q^2(z,\varepsilon) = \frac{6}{z^2}$. This is the leading term of the asymptotic expansion (25) of $v_+(z)$. So, substitution (34) linearizes (23) with respect to the leading term for both positive and zero values of ε.

Let

$$w = (D^2 + 1)u. \tag{38}$$

Then (35) becomes

$$(D^2 - \varepsilon^2)w = \left[(D^2 + 1)^{-1}w\right]^2 + 12cq^2(D^2 + 1)^{-1}w + f(q) \tag{39}$$

or

$$\begin{aligned}[D^2 - (\varepsilon^2 + 12q^2)]w = {} & \left[(D^2 + 1)^{-1}w\right]^2 \\ & + 12cq^2[(D^2 + 1)^{-1} - 1]w - 12q^2(1 - c)w + f(q).\end{aligned} \tag{40}$$

Using the identity

$$(D^2 + 1)^{-1} - 1 = -(D^2 + 1)^{-1}D^2 \tag{41}$$

and solving (39) for D^2w, we finally get

$$\begin{aligned}[D^2 - (\varepsilon^2 + 12q^2)]w = {} & [1 - 12cq^2(D^2 + 1)^{-1}]\left(f(q) + [(D^2 + 1)^{-1}w]^2\right) \\ & - 12q^2(1 - c)w - 12cq^2(D^2 + 1)^{-1}\left[\varepsilon^2 + 12cq^2(D^2 + 1)^{-1}\right]w.\end{aligned} \tag{42}$$

This integrodifferential equation is equivalent to (35), (38). We solve (42) by the contraction mapping principle, considering it as a "perturbation" of the nonhomogeneous linear ODE

$$[D^2 - (\varepsilon^2 + 12q^2)]w = [1 - 12cq^2(D^2 + 1)^{-1}]f(q). \tag{43}$$

The corresponding homogeneous equation

$$[D^2 - (\varepsilon^2 + 12q^2)]w = 0 \tag{44}$$

has two independent solutions $v_1(z)$ and $v_2(z)$, where $v_2(z)$ is symmetric (even) and $v_1(z)$ is antisymmetric (odd) at the origin. Indeed, $v = 6q^2$ is a solution to $(D^2-\varepsilon^2)v = v^2$, which is the rescaled equation (2). Differentiating both sides of the latter equation, we get $(D^2 - \varepsilon^2)v' = 2vv'$. This equation coincides with (44), where $w = v'$. Thus, we obtain

$$v_1(z) = 6\frac{d}{dz}q^2 = 6\varepsilon^3\wp'(\varepsilon z). \tag{45}$$

The second linearly independent solution can be taken as

$$v_2(z) = v_1(z)\int_0^z \frac{d\xi}{v_1^2(\xi)}. \tag{46}$$

Note that $\wp(\varepsilon z)$ is symmetrical and, thus, v_1 is an antisymmetrical (odd) function with respect to any integer combination of half periods $\tilde{\omega}_1$ and $\tilde{\omega}_3$. It will be shown below (section 2.3) that solutions

$$v_{2j}(z) = v_2(z) - \frac{1}{6\varepsilon^7\Delta}\left[\frac{3g_3}{2}\omega_j - g_2\zeta(\omega_j)\right]v_1(z), \qquad j = 1,2,3, \tag{47}$$

to (44), where

$$\zeta(x) = \frac{1}{x} - \int_0^x \left(\wp(y) - \frac{1}{y^2}\right)dy \tag{48}$$

is a "negative antiderivative" of $\wp$, are symmetrical at the points $z = \tilde{\omega}_j$, respectively.

2.2. Formulation of the theorem. Let $g(z)$ be a function, symmetrical (even) with respect to the point $\tilde{\omega}_2(\varepsilon)$, and let $\mathcal{I}_2 = (D^2+1)^{-1}$ denote the operator, inverse to D^2+1 and such that $\mathcal{I}_2 g(z)$ is symmetrical with respect to $\tilde{\omega}_2(\varepsilon)$. The construction of $\mathcal{I}_2$ will be given in section 3.1.

Equation (42) can be written in the operator form as $w = \mathcal{N}w$, where

(49)

$$\mathcal{N}w = \mathcal{I}_1 \circ \left[(1 - 12cq^2\mathcal{I}_2)[f + (\mathcal{I}_2 w)^2] - 12q^2(1-c)w + 12cq^2\mathcal{I}_2(\varepsilon^2 + 12cq^2\mathcal{I}_2)w\right]$$

and

$$\mathcal{I}_1 g(z) = -v_1(z)\int_{\tilde{\omega}_2}^z v_2(\xi)g(\xi)d\xi + v_2(z)\int_{\tilde{\omega}_2}^z v_1(\xi)g(\xi)d\xi. \tag{50}$$

We define iterations

$$w_0 = 0, \quad w_1 = \mathcal{N}w_0 + \varepsilon^6 b(\varepsilon)v_{22}(z), \quad w_n = \mathcal{N}w_{n-1},\ n = 2,3,\dots, \tag{51}$$

and $\Delta w_n = w_n - w_{n-1}$, $n = 1,2,\dots$, where $b(\varepsilon)$ is a continuous function in a vicinity of $\varepsilon = 0$. The following theorem proves uniform convergence of the series $\sum_{n=1}^{\infty}\Delta w_n$ in a region $R_{z_0} \subset \mathbb{C}$, which is defined below. Thus, we obtain a family of solutions to (42), parametrized by $c(\varepsilon)$ and $b(\varepsilon)$, which are symmetrical at $z = \tilde{\omega}_2$.

Consider the parallelogram of periods of the elliptic function $q(z,\varepsilon)$ (the periods are $2\tilde{\omega}_1$ and $2\tilde{\omega}_3$) centered at the origin with the square cut that has vertices $\pm z_0(1+i)$ and $\pm z_0(1-i)$, where $z_0 > 0$ is a positive constant. We require $\varepsilon \in [0, \varepsilon_0]$, where

$$\varepsilon_0 z_0 < \min\{\omega_1, |\omega_3|\}. \tag{52}$$

By R_{z_0} we denote a quarter of this figure, i.e., of the parallelogram of periods with the square cut, that lies in the first quadrant; see Figure 4.

THEOREM 2.1. *Let E be a closed subinterval of $(-\frac{1}{3}, 0)$, and let $J(\varepsilon)$ from (37) be a function, defined on a neighborhood of $\varepsilon = 0$ with the range in E. Let continuous functions $b(\varepsilon), c(\varepsilon)$ satisfy*

$$|b(\varepsilon)| \le l\varepsilon^{\alpha} \qquad \text{and} \qquad |c(\varepsilon) - 1| \le l\varepsilon^{\alpha} \tag{53}$$

for some $l > 0$, $\alpha \ge 1$, and for all nonnegative ε from a vicinity of $\varepsilon = 0$. Then there exist $\varepsilon_0 > 0$, $z_0 > 0$ satisfying (52), which depend only on E and l, and such that the series

$$w(z,\varepsilon) = \sum_{n=1}^{\infty}\Delta w_n(z,\varepsilon) \tag{54}$$

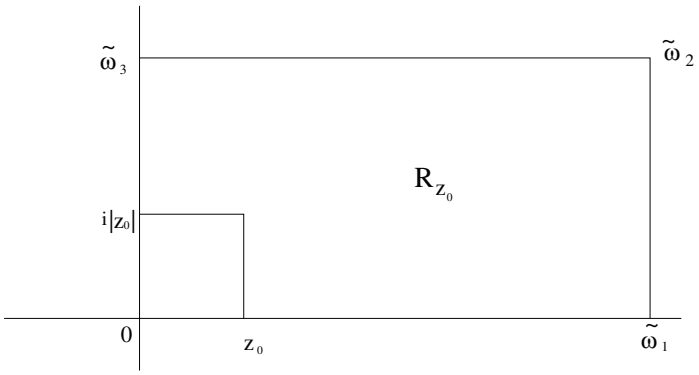


FIG. 4. *Region* R_{z_0}.

converges uniformly in $R_{z_0} \times [0, \varepsilon_0]$.

Remark 2.2. The rectangular "cut" in R_{z_0} can be replaced by a cut of any other shape so that the distance between R_{z_0} and the origin is $O(z_0)$. In fact, it will be replaced by a more convenient triangular "cut" in the course of the proof.

The proof of the theorem is given in section 3.4.

2.3. The second fundamental solution to the linearized equation. Note that, similarly to the separatrix case, the solution v_1 to (44) is an odd function, while the solution v_2 is an even function; see (45)–(46). Moreover, $\wp(\varepsilon z)$ is symmetrical, and, thus, v_1 is an antisymmetrical (odd) function with respect to any integer combination of half-periods $\tilde{\omega}_1$ and $\tilde{\omega}_3$.

Our computation of v_2 is based on the well-known fact (see, for example, [WW]) that the difference between two elliptic functions with the same periods and same principal parts at each singular point is a constant. The only zeros of the function $\wp'(\varepsilon z)$ within the parallelogram of periods are simple zeros at the points $\tilde{\omega}_j$, $j = 1, 2, 3$. Therefore

$$\frac{1}{\wp'^2(\varepsilon z)} = A_0 + \sum_{j=1}^{3} A_j \wp(\varepsilon z - \omega_j) \tag{55}$$

for all z and

$$\int_0^z \frac{d\xi}{\wp'^2(\varepsilon \xi)} = A_0 z - \frac{1}{\varepsilon} \sum_{j=1}^{3} A_j [\zeta(\varepsilon z - \omega_j) + \zeta(\omega_j)], \tag{56}$$

where the constants A_k, $k = 0, 1, 2, 3$, will be discussed below and $\zeta(x)$, defined by (48), is an odd function. The following arguments use some standard facts about the Weierstrass $\wp$-function that can be found, for example, in [WW], [Ha].

Using the "addition theorem" for ζ-function

$$\zeta(\varepsilon z - \omega_j) - \zeta(-\omega_j) = \zeta(\varepsilon z) + \frac{1}{2} \frac{\wp'(\varepsilon z)}{\wp(\varepsilon z) - e_j}, \quad j = 1, 2, 3, \tag{57}$$

we get

$$\int_0^z \frac{d\xi}{\wp'^2(\varepsilon \xi)} = \frac{1}{\varepsilon} \left[A_0 \varepsilon z - \zeta(\varepsilon z) \sum_{j=1}^{3} A_j - \frac{1}{2} \wp'(\varepsilon z) \sum_{j=1}^{3} \frac{A_j}{\wp(\varepsilon z) - e_j} \right]. \tag{58}$$

Constants A_k, $k=0,1,2,3$, are calculated in Lemma 2.3 below. Differential equations

$$\wp'^2 = 4\prod_{j=1}^{3}(\wp - e_j) \quad \text{and} \quad \wp'' = 6\wp^2 - \frac{g_2}{2} \tag{59}$$

for the Weierstrass $\wp$-function and identities

$$e_1e_2 + e_1e_3 + e_2e_3 = -\frac{g_2}{4} \quad \text{and} \quad e_1e_2e_3 = \frac{g_3}{4} \tag{60}$$

are also utilized in this lemma.

LEMMA 2.3.

(61)

$$A_0 = \frac{9g_3}{\Delta}, \quad A_j = \left[\frac{1}{2(e_j-e_i)(e_j-e_k)}\right]^2, \quad j=1,2,3, \quad \text{and} \quad A_+ = \sum_{j=1}^{3} A_j = \frac{6g_2}{\Delta},$$

where the indices i,j,k are a permutation of $1,2,3$.

Proof. The Taylor expansion of $\wp$ at ω_j, $j=1,2,3$, yields

$$\wp(x) = e_j + a_j(x-\omega_j)^2 + O(x-\omega_j)^4. \tag{62}$$

Therefore the principal part of $(\wp')^{-2}$ at ω_j is $\frac{1}{4a_j^2(x-\omega_j)^2}$. According to (59), $a_j = \frac{1}{2}\wp''(\omega_j) = 3e_j^2 - \frac{g_2}{4}$. The principal part of $\wp(x)$ at the origin is $\frac{1}{x^2}$. Thus

$$A_j = \frac{1}{(6e_j^2 - \frac{g_2}{2})^2} = \left[\frac{1}{2(e_j-e_i)(e_j-e_k)}\right]^2, \tag{63}$$

where the indices i,j,k are a permutation of $1,2,3$. The latter expression follows from

$$3e_j^2 - \frac{g_2}{4} = 2e_j^2 + e_j(e_j+e_i) + e_je_k + e_ie_k = 2e_j^2 + e_ie_k = 2(e_i+e_k)^2 + e_ie_k \tag{64}$$

$$= (2e_i+e_k)(2e_k+e_i) = (e_i-e_j)(e_k-e_j),$$

where (17) and (60) were taken into account.

Direct computations, based on (17) and (60), show that

$$\sum_{j=1}^{3} e_j^2 = \frac{g_2}{2}, \quad \sum_{i<k} e_i^2e_k^2 = \frac{g_2^2}{16}, \quad \text{and} \quad \sum_{j=1}^{3} e_j^3 = \frac{3}{4}g_3, \tag{65}$$

where summation in the second expression is taken over all possible pairs i,k such that $1 \le i < k \le 3$. Then

$$\sum_{j=1}^{3} A_j = \frac{\sum_{i<k}(e_i-e_k)^2}{4\prod_{i<k}(e_i-e_k)^2} = \frac{3g_2}{8\prod_{i<k}(e_i-e_k)^2}. \tag{66}$$

Next we compute the denominator, using (60) and (63):

$$-8\prod_{i<k}(e_i-e_k)^2 = \prod_{j=1}^{3}\left(6e_j^2 - \frac{g_2}{2}\right) = -\frac{g_2^3}{8} + 6\frac{g_2^2}{4}\sum_{j=1}^{3} e_j^2 - 18g_2\sum_{i<k} e_i^2e_k^2 \tag{67}$$

$$+ 6^3\prod_{j=1}^{3} e_j^2 = \frac{1}{2}(27g_3^2 - g_2^3) = -\frac{\Delta}{2}.$$

Now the second statement of the lemma follows from (66).

In order to evaluate A_0 we put $z = 0$ at (55). Since $\wp'$ has a pole at the origin, (55) yields

$$A_0 = -\sum_{j=1}^{3} A_j e_j = \frac{\sum_{j=1}^{3} e_j(e_i - e_k)^2}{-4\prod_{i<k}(e_i - e_k)^2}. \tag{68}$$

Using (17) and (65), we calculate the numerator in the latter expression as $-\frac{9}{4}g_3$. Then the first statement of the lemma follows from (67)–(68). □

Now, the combination of (45)–(46) and of (58)–(59) yields

$$v_2(z) = \frac{1}{6\varepsilon^4}\left[[A_0 x - A_+\zeta(x)]\wp'(x) - 2\sum_{j=1}^{3} A_j(\wp(x) - e_i)(\wp(x) - e_k)\right], \tag{69}$$

where for convenience we use the notation $x = \varepsilon z$. The latter sum is a quadratic polynomial in $\wp$, which, according to (17) and (68), can be written as $-2A_+\wp^2 + 2A_0\wp + B$, where the constant $B = -2\sum_{j=1}^{3} A_j e_i e_k$. In order to determine B, note that (46) implies that v_2 is an even function, which has zero of order four at the origin. Using the Laurent expansions of $\wp$, $\wp'$, and ζ at the origin (see [GR]) and equating the constant term of (69) to be zero, we obtain

$$B = -2\sum_{j=1}^{3} A_j e_i e_k = A_+\left(\frac{g_2}{10} + \frac{g_2}{30}\right) + 2A_+\frac{g_2}{10} = \frac{2g_2^2}{\Delta}. \tag{70}$$

Combining the latter equation with (69) and Lemma 2.3 yields

$$\begin{aligned} v_2(z) &= \frac{1}{\varepsilon^4\Delta}\left(\left[\frac{3g_3}{2}\varepsilon z - g_2\zeta(\varepsilon z)\right]\wp'(\varepsilon z) - 2g_2\wp^2(\varepsilon z) + 3g_3\wp(\varepsilon z) + \frac{g_2^2}{3}\right) \\ &= \frac{1}{\varepsilon^4\Delta}\left(\frac{3g_3}{2}\left[\varepsilon z\wp'(\varepsilon z) + 2\wp(\varepsilon z)\right] - \frac{g_2}{3}\left[3\zeta(\varepsilon z)\wp'(\varepsilon z) + 6\wp^2(\varepsilon z) - g_2\right]\right). \end{aligned} \tag{71}$$

Note that $v_2(z)$ is an elliptic function if and only if $\frac{3g_3}{2}x - g_2\zeta(x)$ is doubly periodic with fundamental periods $2\omega_1, 2\omega_3$. Using $\zeta(x + 2\omega_j) = \zeta(x) + 2\zeta(\omega_j)$ for all x in the domain of ζ and $j = 1, 2, 3$ [GR], we can write the condition that v_2 is an elliptic function as

$$3g_3\omega_j - 2g_2\zeta(\omega_j) = 0, \quad j = 1, 3. \tag{72}$$

Considering these equations as a linear system for unknowns $3g_3, -2g_2$, we come to the conclusion that it cannot have nontrivial solutions since its determinant $\omega_1\zeta(\omega_3) - \omega_3\zeta(\omega_1) \equiv -\frac{i\pi}{2}$ [GR]. Thus v_2 is not an elliptic function for any energy $C \in (-\frac{1}{3}, 0)$. Note that in the case of the separatrix solution v_2 is also not a hyperbolic function (see [To4]).

2.4. Estimates for $v_1(z), v_2(z)$. Let T denote trapezoid $ABCD$ in $\mathbb{C}$ with the vertices $A(\tilde{\omega}_1 - \tilde{\omega}_3), B(\tilde{\omega}_1 + 2\tilde{\omega}_3), C(\frac{3}{2}\tilde{\omega}_3), D(-\frac{1}{2}\tilde{\omega}_3)$. We start with the estimate

$$|q^2(z, \varepsilon)| \le \frac{Q}{|z|^2} \tag{73}$$

with some $Q > 0$, which is valid for any $\varepsilon \geq 0$ and for any z in the trapezoid T. The value of the constant Q can be chosen as $Q = 1 + |\omega_2|^2\tilde{Q}$, where $\tilde{Q} = \frac{1}{12} + \max_{x\in\tilde{T}} |\wp(x) - \frac{1}{x^2}|$. Here $\tilde{T}$ denotes the trapezoid in the x-plane, $x = \varepsilon z$, that corresponds to T. Note that ω_2 is defined by the energy $J(\varepsilon)$ of the periodic solution (33). However, we can choose Q such that (73) holds for all $J(\varepsilon) \in E$.

LEMMA 2.4. *There exist positive constants A and B such that the estimates*

$$|v_1(z)| \leq \frac{A}{|z|^3} \qquad \text{and} \qquad |v_2(z)| \leq B|z|^4 \tag{74}$$

are valid in T for any $\varepsilon \geq 0$ and any $J(\varepsilon) \in E$.

Proof. The first estimate follows from (45) and the inequality

$$\left|\wp'(x) + \frac{2}{x^3}\right| \leq A^3 \tag{75}$$

in $\tilde{T}$, similarly to (73).

The second inequality (74) follows from the fact that the expression in the square brackets in the right-hand side of (69) is bounded by $6|x|^4B$ in $\tilde{T}$ with some positive constant B that does not depend on ε and on $J(\varepsilon) \in E$. ∎

3. Solution by iterations.

3.1. Operator $(D^2 + 1)^{-1}$. Let $\omega \in \mathbb{C}$, $a > 0$, and let an analytic function $g(z)$ be real-valued on $\Re z = \Re\omega$ and symmetrical with respect to ω. Note that under these assumptions $g(z)$ is also real-valued along $\Im z = \Im\omega$.

PROPOSITION 3.1. *The operator I_a^ω defined by*

$$I_a^\omega g(z) = \frac{1}{2i}\left[e^{iz}\int_{\omega-ia}^{z} e^{-i\xi}g(\xi)d\xi - e^{-iz}\int_{\omega+ia}^{z} e^{i\xi}g(\xi)d\xi\right] \tag{76}$$

is inverse to $D^2 + 1$, and $I_a^\omega g(z)$ is real-valued on $\Re z = \Re\omega$ and symmetrical with respect to ω.

Proof. The first statement of the proposition is obvious. In order to prove the second statement, we decompose

$$I_a^\omega g(z) = I^\omega g(z) + \frac{1}{2i}\left[e^{iz}\int_{\omega-ia}^{\omega} e^{-i\xi}g(\xi)d\xi - e^{-iz}\int_{\omega+ia}^{\omega} e^{i\xi}g(\xi)d\xi\right], \tag{77}$$

where $I^\omega = I_0^\omega$. Let $ib = \int_\omega^{\omega+ia} e^{i(\xi-\omega)}g(\xi)d\xi$. Since g is real-valued along the path of integration, $b \in \mathbb{R}$ and $-ib = \int_\omega^{\omega-ia} e^{-i(\xi-\omega)}g(\xi)d\xi$. Thus, the second term in (77) is symmetrical and real-valued as it is

$$\frac{1}{2i}[ibe^{i(z-\omega)} + ibe^{-i(z-\omega)}] = b\cos(z-\omega). \tag{78}$$

To complete the proof we observe that

$$\begin{aligned} I^\omega &= \frac{1}{2i}\left[e^{iz}\int_{\omega}^{z} e^{-i\xi}g(\xi)d\xi - e^{-iz}\int_{\omega}^{z} e^{i\xi}g(\xi)d\xi\right] = \int_\omega^z \sin(z-\xi)g(\xi)d\xi \\ &= \sin(z-\omega)\int_\omega^z \cos(\xi-\omega)g(\xi)d\xi - \cos(z-\omega)\int_\omega^z \sin(\xi-\omega)g(\xi)d\xi \end{aligned} \tag{79}$$

is also symmetrical with respect to ω and real-valued. ∎

Using the reflection principle, it is easy to show that the function $g(z)$ of Proposition 3.1 is periodic with real or purely imaginary period $2P$ if and only if it is symmetrical with respect to $\omega + P$. Then Proposition 3.1 shows that there is a (real) one-parameter family of operators I_a^ω, preserving the symmetry of g at ω. The following proposition shows that symmetry can be preserved at both ω and $\omega + P$ if P is not a multiple of π. If P is a multiple of π, one can prove that a symmetrical and periodic $(D^2+1)^{-1}g(z)$ exists if and only if the Fourier expansion of $g(z)$ does not contain a $\cos z$ term.

PROPOSITION 3.2. *Let $g(z)$ be a function, symmetrical with respect to ω and $\omega + P$ and real-valued on $[\omega, \omega + P]$, where P is either real or purely imaginary. If $P \neq k\pi$, $k \in \mathbb{N}$, then there exists an operator I that is inverse to $D^2 + 1$ and such that Ig inherits the abovementioned properties of g.*

Proof. To satisfy the requirements at ω, the operator I has to be of the form $Ig = I^\omega g + d\cos(z-\omega)$, where $d \in \mathbb{R}$ (and depends on g). Then

(80)

$$Ig = I^{\omega+P}g + d\cos(z-\omega) + \frac{1}{2i}\left[e^{i(z-\omega-P)}\int_\omega^{\omega+P} e^{-i(\xi-\omega-P)}g(\xi)d\xi - e^{-i(z-\omega-P)}\int_\omega^{\omega+P} e^{i(\xi-\omega-P)}g(\xi)d\xi\right].$$

Let $A_{1,2}$ denote the (two) terms in the square brackets, respectively, and let $A = \frac{1}{2}(A_1 + A_2)$, $B = \frac{1}{2}(A_1 - A_2)$. Then the square bracket expression in (80) becomes $A\sin(z-\omega-P) - iB\cos(z-\omega-P)$. Therefore

$$Ig = I^{\omega+P}g + (d\cos P - iB)\cos(z-\omega-P) + (A - d\sin P)\sin(z-\omega-P). \tag{81}$$

Using (79) we see that the first two terms in the right-hand side of (81) are symmetrical at $\omega + P$. Thus Ig is symmetrical with respect to $\omega + P$ if and only if

$$d = \frac{2\int_\omega^{\omega+P}\cos(\xi-\omega-P)g(\xi)d\xi}{\sin P}. \tag{82}$$

Equation (82) can be satisfied for any $P \neq k\pi$, $k \in \mathbb{Z}$. Noting that d and iB are real for both real and purely imaginary P concludes the proof. ∎

3.2. Estimates for the operator $\mathcal{I}_2$. Consider the trapezoid T introduced in section 2.4. Given a number $z_1 \in (0, \frac{1}{4}\tilde{\omega}_1)$, which is independent of ε, we cut a triangle near the origin with vertices z_1, z_2, z_3, where z_2, z_3 are the points on the intersection of the imaginary axis and the lines through the points z_1, A and the points z_1, B, respectively. The obtained region is denoted by T_{z_1}. Clearly, for a given $z_0 > 0$ we have $R_{z_0} \subset T_{z_1}$ if z_0 is large enough (i.e., according to (52), if ε_0 is small enough). We define the operator

$$\mathcal{I}_2 = I^{\tilde{\omega}_2}_{-2i\tilde{\omega}_3}. \tag{83}$$

For a given $z \in T_{z_1}$, $\gamma_\pm(z)$ denote contours of integration $[z, B]$ and $[z, A]$, respectively; see Figure 5.

PROPOSITION 3.3. *The inequality*

$$\left|\frac{\xi}{z}\right| > \frac{\omega_1}{\sqrt{\omega_1^2 - \frac{25}{4}\omega_3^2}} \tag{84}$$

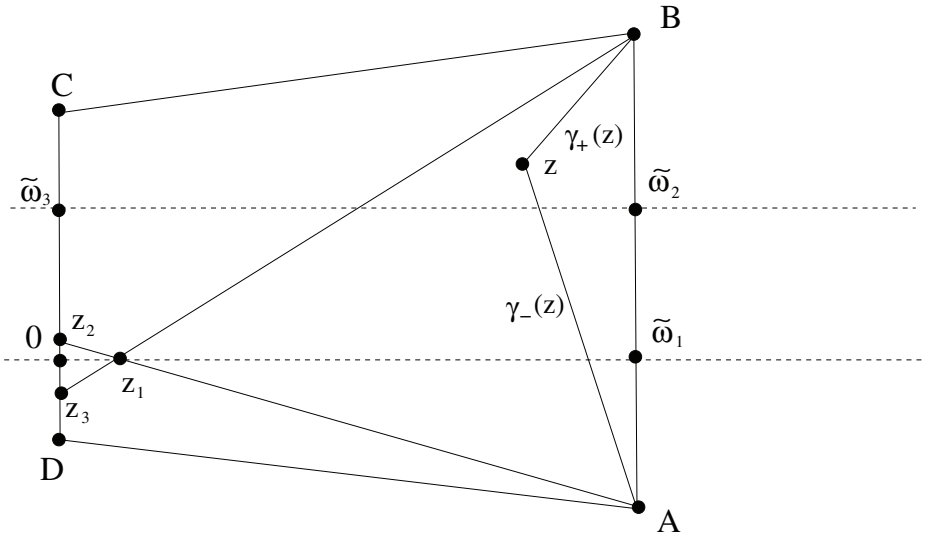


FIG. 5. *Region T_{z_1} and contours $\gamma_\pm(z)$.*

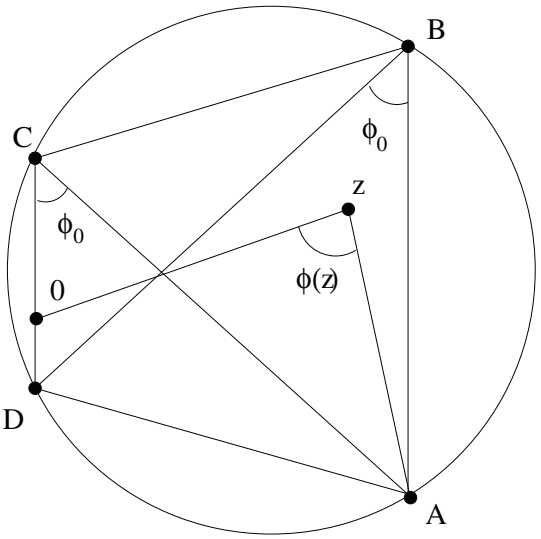


FIG. 6. *Trapezoid $ABCD$ and angles $\phi(z)$, ϕ_0.*

holds for all $z \in T_{z_1}$ and all $\xi \in \gamma_+(z) \cap \gamma_-(z)$.

Proof. For an arbitrary $z \in T_{z_1}$, let $\phi(z)$ denote the angle between the segments $\gamma_-(z)$ and $[0, z]$ (see Figure 6), and let $\alpha(z) = \min_{\xi\in\gamma_-(z)} \frac{|\xi|}{|z|}$. We choose the acute angle for $\phi(\tilde{\omega}_1 + 2\tilde{\omega}_3)$ (at the vertex B) and then define $\phi(z)$ in T_{z_1} by continuity. Clearly $\alpha(z) = 1$ if $\phi(z) \geq \frac{\pi}{2}$ and $\alpha(z) = \sin\phi(z)$ if $\phi(z) < \frac{\pi}{2}$. The latter case can happen only if $\Im z \geq 0$; hence $\alpha(z) = 1$ if $\Im z < 0$. Let ϕ_0 be the acute angle between AB and DB. We show that $\phi(z) \geq \phi_0$ for every $z \in T_{z_1}$.

Indeed, let us inscribe the right trapezoid $ABCD$ into a circle. Then (at the vertex C) $\phi(\frac{3}{2}\tilde{\omega}_3) = \phi_0$. Simple geometrical considerations (see Figure 6) imply that

$\min_{z\in T_{z_1}}\phi(z)$ is attained at $z=\frac{3}{2}\tilde{\omega}_3$. Thus, $\alpha(z)\geq\sin\phi_0$. Therefore

$$\alpha(z)\geq\sin\phi_0=\frac{\omega_1}{\sqrt{\omega_1^2-\frac{25}{4}\omega_3^2}} \tag{85}$$

for any $z\in T_{z_1}$. The same estimate holds for the contour $\gamma_+(z)$. The proof is completed. □

PROPOSITION 3.4. *For any* $z\in T_{z_1}$

$$|\mathcal{I}_2 f(z)|\leq\sqrt{1-4\frac{\omega_1^2}{\omega_3^2}}\max_{\xi\in\gamma_+(z)\cup\gamma_-(z)}|f(\xi)|, \tag{86}$$

where f *is a continuous function on* T_{z_1}.

Proof. According to the construction of γ_+,γ_-,

$$\Re[i(\xi-z)]\leq-\frac{|\omega_3|}{\sqrt{4\omega_1^2-\omega_3^2}}|z-\xi|\quad\text{and}\quad\Re[i(z-\xi)]\leq-\frac{|\omega_3|}{\sqrt{4\omega_1^2-\omega_3^2}}|z-\xi| \tag{87}$$

for all $\xi\in\gamma_\pm(z)$, respectively. Therefore

$$\begin{aligned}|\mathcal{I}_2 f(z)|&\leq\frac{1}{2}\left|\int_{\gamma_-(z)}e^{i(z-\xi)}f(\xi)d\xi+\int_{\gamma_-(z)}e^{i(z-\xi)}f(\xi)d\xi\right|\\&\leq\max_{\xi\in\gamma_+(z)\cup\gamma_-(z)}|f(\xi)|\int_0^\infty e^{-\left(|\omega_3|/\sqrt{4\omega_1^2-\omega_3^2}\right)\lambda}d\lambda,\end{aligned} \tag{88}$$

which implies (86). □

LEMMA 3.5. *If*

$$|f(z)|\leq\frac{1}{|z|^a} \tag{89}$$

in T_{z_1} *with* $a\geq0$, *then*

$$|\mathcal{I}_2 f(z)|\leq\frac{\beta}{|z|^a}, \tag{90}$$

where $\beta=(1-\frac{25\omega_3^2}{4\omega_1^2})^{\frac{1}{2}}(1-4\frac{\omega_1^2}{\omega_3^2})^{\frac{a}{2}}$.

This lemma is a direct consequence of Propositions 3.3 and 3.4.

3.3. Estimates for the operator $\mathcal{I}_1$. Solution $\mathcal{I}_1 g$ to linear differential equation (43), where g denotes the right-hand side of (43), that we consider in the paper is given by (50). In order to estimate $\mathcal{I}_1$, we need the following simple statement.

PROPOSITION 3.6. *Let a ray* λ *on the complex* z*-plane lie outside the disk* $|z|<a$, $a>0$. *Then for any* $\alpha\geq2$

$$\left|\int_\lambda\frac{dz}{z^\alpha}\right|\leq\frac{\pi}{a^{\alpha-1}}. \tag{91}$$

Proof. Let m denote the line containing the ray λ. Suppose m is tangent to the disk $|z|<a$ at the point ω. If $z\in m$, then $|z|^2=a^2+x^2$, where $x=|z-\omega|$. Then

$$\left|\int_\lambda\frac{dz}{z^\alpha}\right|\leq\int_{-\infty}^\infty\frac{dx}{(a^2+x^2)^{\frac{\alpha}{2}}}\leq\frac{2}{a^{\alpha-2}}\int_0^\infty\frac{dx}{(a^2+x^2)}=\frac{\pi}{a^{\alpha-1}}. \tag{92}$$

This estimate will definitely hold if m does not intersect the closed disk $|z| \le a$.

Suppose now that m intersects the disk $|z| < a$. Let ω denote the vertex of the ray λ, and let ζ denote one of the two points which lie on the distance a from both ω and m. Then $|\xi - \zeta| < |\xi|$ for any $\xi \in \lambda$, and we can apply the estimate (92) again, placing the origin at ζ. □

Notice that estimates of Proposition 3.3 are applicable to the contours in (50). Then, using Lemma 2.4 and Proposition 3.6, we can easily establish the following estimates.

LEMMA 3.7. *If $|g(z)| \le K|z|^{-n}$, $n \ge 6$, or if $|g(z)| \le K|z|^{-4}$ in T_{z_1}, then*

$$|\mathcal{I}_1 g(z)| \le \frac{M_1 K}{|z^{n-2}|} \quad \text{and} \quad |\mathcal{I}_1 g(z)| \le \frac{M_2 K}{\varepsilon |z^{-3}|}, \tag{93}$$

respectively, where $M_1 = (1 - \frac{25}{4}\frac{\omega_3^2}{\omega_1^2})^{\frac{n-5}{2}} \pi AB$ and $M_2 = \max\{2|\omega_3|, \sqrt{\omega_1^2 - \frac{9}{4}\omega_3^2}\} AB$.

3.4. Proof of Theorem 2.1. *Proof.* We start with the obvious observation that

$$\varepsilon \le \frac{\sqrt{\omega_1^2 - 4\omega_3^2}}{|z|} \tag{94}$$

if $z \in T_{z_1}$. Using the first inequality in (53), we obtain that $f(q)$ from (36) satisfies

$$|f| \le \tilde{K}_0 |z|^{-6} + \varepsilon \tilde{K}_1 |z|^{-4} \tag{95}$$

in T_{z_1} with some positive constants $\tilde{K}_0, \tilde{K}_1$ that do not depend on ε and $J(\varepsilon) \in E$. Thus, according to (49) and the estimates for $\mathcal{I}_{1,2}$,

$$|\mathcal{N} w_0| \le \frac{K}{2} \left(\frac{\varepsilon^{\alpha - 1}}{|z|^3} + \frac{1}{|z|^4} \right) \tag{96}$$

in T_{z_1} for some constant $K > 0$ that does not depend on ε, $J(\varepsilon)$, and z_1.

Let us now prove by induction that by choosing a sufficiently large $z_1 > 0$ we can find $\delta > 0$, such that the estimate

$$\Delta w_n \le (\delta M)^{n-1} \frac{K}{|z|^3}, \quad \text{where} \quad \delta M \le \frac{1}{2}, \tag{97}$$

holds in T_{z_1} for all $n = 1, 2, \dots$, all $J(\varepsilon) \in E$, and all sufficiently small ε. Here $M = \max\{M_1, M_2\}$, where M_1 is given in Lemma 3.7 with $n = 6$.

According to Lemma 2.4, the estimate for v_{22} for all $z \in T_{z_1}$ and $J(\varepsilon) \in E$ is

$$|v_{22}(z)| \le B|z|^4 + \frac{L}{\varepsilon^7 |z|^3} \tag{98}$$

for some $L > 0$. Then, according to (94), (96), and (53), we can choose $K > 0$ to be so large that $w_1 = \mathcal{N} w_0 + \varepsilon^6 b(\varepsilon) v_{22}$ satisfies

$$|w_1| \le \frac{K}{|z|^3} \tag{99}$$

in T_{z_1} for all $J(\varepsilon) \in E$ and all sufficiently small ε. Thus, for $n = 1$, (97) has been established. Assume that this estimate is true for $k = 1, 2, \dots, n$, and let us establish it for $k = n + 1$. First, we can represent

$$\begin{aligned} \Delta w_{n+1} = \mathcal{I}_1 \circ \big[& (1 - 12cq^2 \mathcal{I}_2)[\mathcal{I}_2(w_n + w_{n-1}) \cdot \mathcal{I}_2 \Delta w_n] - 12q^2(1 - c)\Delta w_n \\ & + 12cq^2 \mathcal{I}_2(\varepsilon^2 + 12cq^2 \mathcal{I}_2)\Delta w_n \big]. \end{aligned} \tag{100}$$

Taking into account (99) and Lemma 3.5, we obtain the estimates

(101)
$$\frac{12Ql\varepsilon^{\alpha}}{|z|}\left|\frac{\Delta w_n}{z}\right|,\ 12cQ\beta\left[\frac{\varepsilon}{|z|}\left|\frac{\varepsilon\Delta w_n}{z}\right|+\frac{12cQ\beta}{|z|}\left|\frac{\Delta w_n}{z^3}\right|\right],\ 4K\beta^2\left(1+\frac{12cQ\beta}{|z|^2}\right)\left|\frac{\Delta w_n}{z^3}\right|$$

for the second, third, and first (nonlinear) terms in the square brackets in (100), respectively. For the latter estimate, observe that both w_n and w_{n-1} are bounded by $\frac{2K}{|z|^3}$ according to (97). (Note that the constants β, K, M are independent of z_1.) Applying operator $\mathcal{I}_1$ to these three terms and utilizing Lemma 3.7, we obtain the estimates

(102)
$$\frac{12Ql}{|z|}\varepsilon^{\alpha-1}M\left|\Delta w_n\right|,\ 12cQ\beta\left[\varepsilon+\frac{12cQ\beta}{|z|}\right]\frac{M}{|z|}\left|\Delta w_n\right|,\ 4K\beta^2\left(1+\frac{12cQ\beta}{|z|^2}\right)\left|\frac{M\Delta w_n}{z}\right|.$$

Factoring $M|\Delta w_n|$, we obtain $|\Delta w_{n+1}|\le\delta M|\Delta w_n|$, where

$$\delta=\frac{12Ql\varepsilon^{\alpha-1}+12cQ\beta\varepsilon}{|z|}+\frac{12^2c^2Q^2\beta^2}{|z|^2}+\frac{4K\beta^2}{|z|}\left(1+\frac{12cQ\beta}{|z|^2}\right). \tag{103}$$

By choosing a sufficiently large $z_1>0$ and sufficiently small $\varepsilon_0>0$, we can guarantee the condition $\delta M\le\frac{1}{2}$ in T_{z_1} for all $\varepsilon\le\varepsilon_0$ and for all $J(\varepsilon)\in E$. Thus, (97) holds for $k=n+1$. We can choose some $z_0\ge z_1$ so that $R_{z_0}\subset T_{z_1}$ and $\varepsilon_0 z_0\le\max\{\omega_1,|\omega_3|\}$ (taking smaller ε_0 if necessary). Thus, inequalities (97) prove convergence of iterations to a solution of (42) that is uniform in $R_{z_0}\times(0,\varepsilon_0]$ for all $J(\varepsilon)\in E$.

In the case $\varepsilon=0$ the region T_{z_1} becomes the right half-plane with the appropriate cut, $q^2=\frac{6}{z^2}$, $f(q)=-\frac{6!}{z^6}$, and the fundamental solutions $v_{1,2}$ to the homogeneous equations (43) become $v_1(z)=\frac{-12}{z^3}$ and $v_2(z)=\frac{-z^4}{84}$. Thus, for the case $\varepsilon=0$ Lemma 2.4 becomes trivial, and the proof of convergence of iterations $w=\sum_1^\infty\Delta w_n$ that solve (42) holds for the case $\varepsilon=0$. (Note that in this case the contours of integration in the integral operators $\mathcal{I}_{1,2}$ are the rays, emanating from z, with the slope $\frac{\Im w}{\Re w}$, where w is the beginning of the corresponding contour in the case $\varepsilon>0$; in other words, $\arg w$ does not depend on ε.) Thus, the proof of Theorem 2.1 is completed. □

Remark 3.8. It is easy to verify that $\alpha>1$ in Theorem 2.1 implies that there exist some $L_n>0$ such that for all $n\in\mathbb{N}$

$$|\Delta w_n|\le L_n|z|^{-n(1+\tilde\alpha)-2} \tag{104}$$

in R_{z_0}, where $\tilde\alpha=\min\{1,\alpha-1\}$. Moreover, for all $n\in\mathbb{N}$ there exist some $K_n>0$ such that

$$\left|\sum_{j=n}^{\infty}\Delta w_n\right|\le K_n|z|^{-n(1+\tilde\alpha)-2}. \tag{105}$$

Let $P\subset T$ denote the triangle with vertices $\tilde\omega_3,\tilde\omega_1$, and $\tilde\omega_1+2\tilde\omega_3$. It is clear that $z\in P$ implies that all the contours of integration in operators $\mathcal{I}_{1,2}$ are in P. Notice that in P

$$\min\{\omega_1,|\omega_3|\}\le\varepsilon|z|\le\sqrt{\omega_1^2-4\omega_3^2}. \tag{106}$$

The latter condition allows us to prove convergence of iterations in P without the requirement (53) in Theorem 2.1.

COROLLARY 3.9. *If instead of* (53) *in Theorem* 2.1 *we require only*

$$\lim_{\varepsilon\to 0} b(\varepsilon) = \lim_{\varepsilon\to 0}[c(\varepsilon)-1] = 0, \tag{107}$$

the statement of Theorem 2.1 *is still valid in* $P\times[0,\varepsilon_0]$*; i.e., there exists some* $\varepsilon_0>0$ *such that the series* (54) *is uniformly convergent in* $P\times[0,\varepsilon_0]$.

Proof. The function

$$h(\varepsilon) = \max\{|b(\varepsilon)|, |c(\varepsilon)-1|\} \tag{108}$$

is continuous and nonnegative, and $\lim_{\varepsilon\to 0} h(\varepsilon) = 0$. Then inequalities (95) and (96) become

$$|f| \le \tilde{K}_0|z|^{-6} + h(\varepsilon)\tilde{K}_1|z|^{-4} \quad \text{and} \quad |\mathcal{N}w_0| \le \frac{K}{2}\left(\frac{h(\varepsilon)}{\varepsilon|z|^3} + \frac{1}{|z|^4}\right), \tag{109}$$

respectively. Then, as in Theorem 2.1, we can choose $K>0$ so large that

$$|w_1| \le \frac{\tilde{h}(\varepsilon)K}{|z|^2} \tag{110}$$

in P, where $\tilde{h}(\varepsilon) = \max\{h(\varepsilon), \varepsilon^2\}$.

As in Theorem 2.1, we will prove convergence of (54) by induction if we show that by choosing a sufficiently small $\varepsilon_0>0$ we can find $\delta>0$, such that the estimate

$$\Delta w_n \le (\delta M)^{n-1}\frac{\tilde{h}(\varepsilon)K}{|z|^2}, \quad \text{where} \quad \delta M \le \frac{1}{2}, \tag{111}$$

holds in P for all $n=1,2,\dots$ and all $J(\varepsilon)\in E$. Taking into account (111) and Lemma 3.5, we obtain the estimates

$$12Ql\tilde{h}(\varepsilon)\left|\frac{\Delta w_n}{z^2}\right|, \tag{112}$$

$$12cQ\beta\left[\left|\frac{\varepsilon^2\Delta w_n}{z^2}\right| + \frac{12cQ\beta}{|z|^2}\left|\frac{\Delta w_n}{z^2}\right|\right], \quad 4\tilde{h}(\varepsilon)K\beta^2\left(1+\frac{12cQ\beta}{|z|^2}\right)\left|\frac{\Delta w_n}{z^2}\right|$$

for the second, third, and first (nonlinear) terms in the square brackets in (100), respectively. Making sure that $\tilde{h}(\varepsilon)$ and $|z^{-2}|$ are sufficiently small by choosing a sufficiently small ε_0 and using Lemma 3.7, we can repeat the arguments of Theorem 2.1 in order to complete the proof of convergence of iterations (54). ∎

Remark 3.10. In fact, we can prove convergence of iterations (54) in larger domains than those stated in Theorem 2.1 and Corollary 3.9. Fix some $n\in\mathbb{N}$. For Theorem 2.1, we replace the trapezoid T with vertices $ABCD$ by the (concave) hexagon $\mathcal{T}$ with vertices $AONBCD$, where $O=\tilde{\omega}_1$, $N=n\tilde{\omega}_1+\tilde{\omega}_3$, and other vertices are the same as in T; see Figure 7. For $z\in\mathcal{T}$, contours of integration in operators $\mathcal{I}_{1,2}$ are the same as before, except for contour $\gamma_-(z)$ that is now the union of segments $[z,O]$ and $[OA]$. For Corollary 3.9, we replace the triangle P with vertices $\tilde{\omega}_3$, $\tilde{\omega}_1$, $\tilde{\omega}_1+2\tilde{\omega}_3$ by a quadrilateral $\mathcal{P}$ with vertices $\tilde{\omega}_1$, $n\tilde{\omega}_1+\tilde{\omega}_3$, $\tilde{\omega}_1+2\tilde{\omega}_3$, $\tilde{\omega}_3$. For $z\in\mathcal{P}$, contours of integration in operators $\mathcal{I}_{1,2}$ are the same as before. It is easy to verify that the estimates for operators $\mathcal{I}_{1,2}$ from Lemmas 3.7 and 3.5 hold in domains $\mathcal{T}$ and $\mathcal{P}$, possibly with larger constants that depend on n. Now proofs of Theorem 2.1 and Corollary 3.9 can be extended to the domains $\mathcal{T}$ and $\mathcal{P}$, respectively.

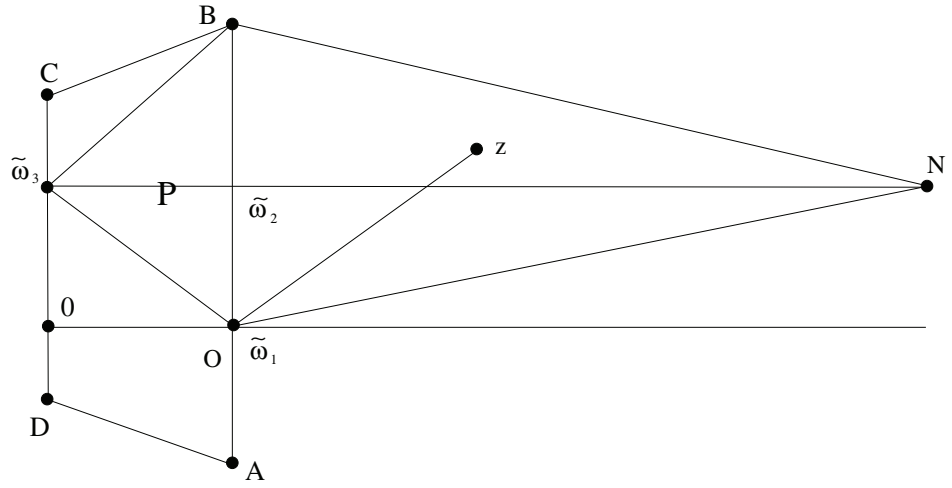


FIG. 7. *Hexagon $\mathcal{T}$ with vertices $AONBCD$. The contour $\gamma_-(z)$ is the union of segments $[z, O]$ and $[OA]$.*

3.5. Symmetry at $\tilde{\omega}_2$. Here we will show that any solution to inner equation (23), obtained through iterations (54), is symmetrical with respect to $z = \tilde{\omega}_2$. Consequently, corresponding solutions to the outer equation (1), connected with (23) via (22), are symmetric at $x = \omega_2$.

The operator $\mathcal{I}_1$ can be represented in the following form.

PROPOSITION 3.11. *If a function $g(z)$ is analytic and bounded in T_{z_1}, then*

$$\mathcal{I}_1 g = v_1(z) \int_{\tilde{\omega}_2}^{z} \frac{\int_{\tilde{\omega}_2}^{t} v_1(\tau) g(\tau) d\tau}{v_1^2(t)} dt. \tag{113}$$

Proof. Using (46) and integration by parts, one gets

$$\begin{aligned} \mathcal{I}_1 g &= -v_1(z) \int_{\tilde{\omega}_2}^{z} \left(v_1(t) g(t) \int_0^t \frac{d\tau}{v_1^2(\tau)} \right) dt + v_2(z) \int_{\tilde{\omega}_2}^{z} v_1(t) g(t) dt \\ &= v_1(z) \times \left[- \int_{\tilde{\omega}_2}^{t} v_1(\tau) g(\tau) d\tau \cdot \int_0^t \frac{d\tau}{v_1^2(\tau)} \Big|_{\tilde{\omega}_2}^{z} \right. \\ &\qquad \left. + \int_{\tilde{\omega}_2}^{z} \frac{\int_{\tilde{\omega}_2}^{t} v_1(\tau) g(\tau) d\tau}{v_1^2(\tau)} dt + \int_0^z \frac{dt}{v_1^2(\tau)} \cdot \int_{\gamma(z)} v_1(t) g(t) dt \right] \\ &= v_1(z) \lim_{\zeta \to \tilde{\omega}_2} \left[\int_{\zeta}^{z} v_1(t) g(t) dt \cdot \int_0^{\zeta} \frac{dt}{v_1^2(\tau)} \right] + v_1(z) \int_{\tilde{\omega}_2}^{z} \frac{\int_{\tilde{\omega}_2}^{t} v_1(\tau) g(\tau) d\tau}{v_1^2(\tau)} dt. \end{aligned} \tag{114}$$

Note that the limit in the latter expression is zero because $\int_{\gamma(\zeta)} v_1(t) g(t) dt$ has a zero of order two at $\tilde{\omega}_2$, whereas $\int_0^{\zeta} \frac{dt}{v_1^2(\tau)}$ has a simple pole there. ∎

Based on (45) and (73), solutions to the homogeneous equation (44) preserving the symmetry at the points $\tilde{\omega}_j$, $j = 1, 2, 3$, are given by

$$v_{2j}(z) = v_2(z) - \frac{1}{6\varepsilon^7 \Delta} \left[\frac{3 g_3}{2} \omega_j - g_2 \zeta(\omega_j) \right] v_1(z), \qquad j = 1, 2, 3, \tag{115}$$

respectively. Indeed, direct calculations show that $D_z v_{2j}(\tilde\omega_j) = 0$. Then, according to (44),

$$D_z^3 v_{2j} = (\varepsilon^2 + 12q^2) D_z v_{2j} + 12(q^2)' v_{2j} = 0 \tag{116}$$

at $z = \tilde\omega_j$. By continuing this argument for higher derivatives, we prove symmetry of v_{2j} at $z = \tilde\omega_j$. Also, note that $v_{2j}(\tilde\omega_j) \in \mathbb{R}$. Then, according to the differential equation for v_{2j}, all even derivatives of v_{2j} at $z = \tilde\omega_j$ are also real.

The fact that operator $\mathcal{I}_1$ preserves the symmetry at the point $\tilde\omega_2$ and real-valuedness along the line $\Im z = \Im\tilde\omega_2$ is a direct consequence of Proposition 3.11. Thus, according to (36), Δw_1 is symmetric with respect to $\tilde\omega_2$ and real-valued on the line $\Im z = \Im\tilde\omega_2$. Then, according to (100) and to Proposition 3.1, the same is true for $w = \sum_1^\infty \Delta w_n$. Now, according to (34), the corresponding solution $v(z,\varepsilon)$ to (23) satisfies

$$v'(\tilde\omega_2, \varepsilon) = v'''(\tilde\omega_2, \varepsilon) = 0 \tag{117}$$

and is real-valued on $\Im z = \Im\tilde\omega_2$.

Thus, in Theorem 2.1 and Corollary 3.9 we have constructed a two-parameter family of solutions to (23), where $b(\varepsilon)$ and $c(\varepsilon)$ are the parameters, that are real along $\Im z = \Im\tilde\omega_2$ and satisfy (117).

3.6. Calculation of $\mathcal{I}_1 f(q)$. In order to calculate $\mathcal{I}_1 f(q)$, we use (36) and (45) to evaluate the integral

$$\begin{aligned}\int_{\tilde\omega_2}^{t} v_1(\tau) f(q(\tau)) d\tau &= \int f(q) d(6q^2) \\ &= c\left[\frac{c-1-4\varepsilon^2}{3}(6q^2)^3 - 5/6(6q^2)^4 - 2\varepsilon^6 C(\varepsilon)(6q^2)\right]\Big|_{\tilde\omega_2}^{z}\end{aligned} \tag{118}$$

in (113). Switching to the original variable $x = \varepsilon z$ and using (113), we obtain

$$\begin{aligned}&\mathcal{I}_1 f = 6\varepsilon^3 c\wp'(x) \\ &\times \int_{\omega_2}^{x} \frac{(6\wp(t) - \frac12)\left[\frac{c-1-4m^2}{3\varepsilon}(6\wp(s) - \frac12)^2 - \frac{5\varepsilon}{6}(6\wp(s) - \frac12)^3 - 2\varepsilon C(\varepsilon)\right]\Big|_{s=\omega_2}^{s=t} dt}{144\prod_1^3(\wp(t) - e_j)}.\end{aligned} \tag{119}$$

The integrand in (119) is nonsingular at $t = \omega_2$ since the numerator has zero at this point. Using this fact, we can cancel the factor $(\wp - e_2)$ in both the numerator and denominator. Then direct calculations show that the integrand in (119) becomes

$$\begin{aligned}&\frac{c-1-4\varepsilon^2}{3\varepsilon} \cdot \frac{\wp^2 + (e_2 - \frac14)\wp + e_2^2 - \frac14 e_2 + \frac{1}{48}}{4(\wp - e_1)(\wp - e_3)} \\ &\qquad - \frac{\frac{5\varepsilon}{6}(6\wp(t) - \frac12)^4}{144\prod_1^3(\wp(t) - e_j)} - \frac{2\varepsilon C(\varepsilon)(6\wp(t) - \frac12)}{144\prod_1^3(\wp(t) - e_j)}.\end{aligned} \tag{120}$$

Let us focus on the first term of (120). Separating the integer part, we can rewrite it as

$$\frac{c-1-4\varepsilon^2}{3\cdot 4\varepsilon}\left[1-\frac{1}{4}\frac{\wp+e_2-\frac{1}{6}}{(\wp-e_1)(\wp-e_3)}\right] = \frac{c-1-4\varepsilon^2}{3\cdot 4\varepsilon}\left[1-\frac{B_0+B_1\wp(t-\omega_1)+B_3\wp(t-\omega_3)}{4}\right], \tag{121}$$

where

$$B_1=-\frac{e_3+\frac{1}{6}}{(e_1-e_3)^2(e_1-e_2)},\quad B_3=\frac{e_1+\frac{1}{6}}{(e_1-e_3)^2(e_2-e_3)},\quad \text{and}\quad B_0=-B_1e_1-B_3e_3. \tag{122}$$

The expressions for B_1, B_3, and B_0 were obtained by evaluating the principal parts of the left-hand side of (122) at $t=\omega_1$, $t=\omega_3$, and $t=0$, respectively.

Using similar calculations, we obtain expressions

$$-\frac{15\varepsilon}{2}\wp(t)+\frac{5\varepsilon}{2}-\frac{5\varepsilon}{24}\left[E_0+E_1\wp(t-\omega_1)+E_3\wp(t-\omega_3)\right]\quad\text{and} \tag{123}$$

$$-\frac{\varepsilon C(\varepsilon)}{12}\left[D_0+D_1\wp(t-\omega_1)+D_3\wp(t-\omega_3)\right]$$

for the second and the third terms of (120), respectively, where

$$D_1=\frac{1}{(e_1-e_3)^2(e_1-e_2)},\quad D_3=-\frac{1}{(e_1-e_3)^2(e_2-e_3)},\quad D_0=-D_1e_1-D_3e_3 \tag{124}$$

and

$$E_3=-\frac{\frac{9}{4}e_3+\frac{3}{2}e_2+36e_2^3-\frac{1}{3}}{(e_1-e_3)^2(e_2-e_3)},\qquad E_0=-E_1e_1-E_3e_3. \tag{125}$$

(The constant E_0 could be calculated similarly to E_3, but we do not need its explicit value.)

Substituting (121)–(125) into (119), we obtain

$$\mathcal{I}_1 f = c\varepsilon^3\wp'(x)\left\{45\varepsilon[\zeta(x)-\zeta(\omega_2)]+6G_1[\zeta(x-\omega_1)-\zeta(\omega_1)]\right. \tag{126}$$

$$\left.+6G_3[\zeta(x-\omega_3)-\zeta(\omega_3)]+\left[-15\varepsilon+(c-1-4\varepsilon^2)(2\varepsilon)^{-1}-6G_1e_1-6G_3e_3\right](x-\omega_2)\right\},$$

where $x=\varepsilon z$ and

$$G_j=-\frac{B_j(c-1-4\varepsilon^2)}{48\varepsilon}-\frac{5\varepsilon}{24}E_j-\frac{\varepsilon C(\varepsilon)}{12}D_j,\qquad j=1,3. \tag{127}$$

3.7. The limit of iterations as $\varepsilon\to 0$. In this subsection we assume that conditions of Theorem 2.1 hold. In the $\varepsilon=0$ case, direct calculation, based on (50) and (49), shows that

$$\Delta w_1(z,0)=-\frac{90}{z^4}+\frac{12\cdot 6!}{7}\left(-z^{-3}\int_\infty^z t^2\mathcal{I}_2[t^{-6}]dt+z^4\int_\infty^z t^{-5}\mathcal{I}_2[t^{-6}]dt\right). \tag{128}$$

It is easy to check that Δw_n, $n = 2, 3, \dots$, is a function, analytic in T_{z_1} and possessing an asymptotic expansion in powers of z^{-2} there, such that

$$\Delta w_n(z) \sim O(z^{-2(n+1)}), \qquad z \to \infty, \ |\arg z| < \frac{\pi}{2}. \tag{129}$$

Then, according to Corollary 1.3, the solution (34) to (23) in the case $\varepsilon = 0$ coincides with $v_+(z)$, i.e., $v(z, 0) = v_+(z)$.

Let $w(z, \varepsilon)$ denote the solution (54) from Theorem 2.1. In order to prove the continuity of $w(z, \varepsilon)$ in ε at $\varepsilon = 0$, we need the following statement.

PROPOSITION 3.12. *Let $g(z, \varepsilon)$ be analytic in z in T_{z_1} and continuous in ε for all sufficiently small $\varepsilon \geq 0$. Moreover, let $g(z, 0) \sim O(z^{-b})$ and $\Delta_\varepsilon g(z, \varepsilon) \sim O(z^{-a})$ in T_{z_1}, the latter uniformly in ε, where $a, b \geq 0$ and $\Delta_\varepsilon g(z, \varepsilon) = g(z, \varepsilon) - g(z, 0)$. Then*

$$\Delta_\varepsilon \mathcal{I}_2 g(z, \varepsilon) = O(z^{-a}) + O(\varepsilon^b), \tag{130}$$

and, if $b > 5$,

$$\Delta_\varepsilon \mathcal{I}_1 g(z, \varepsilon) = O(z^{2-a}) + O(\varepsilon^{b-5} z^{-3}) + O(z^{-b+4}\varepsilon^2), \tag{131}$$

where both (130) *and* (131) *are uniform in T_{z_1} and in small $\varepsilon \geq 0$.*

Proof. Since $g(z, \varepsilon)$ is analytic in T_{z_1},

$$\begin{aligned} \Delta_\varepsilon \mathcal{I}_2 g(z, \varepsilon) = \frac{1}{2i} \left[\int_{\tilde{\omega}_1 - \tilde{\omega}_3}^{z} e^{i(z-\xi)} \Delta_\varepsilon g(\xi, \varepsilon) d\xi - \int_{\tilde{\omega}_1 + 2\tilde{\omega}_3}^{z} e^{-i(z-\xi)} \Delta_\varepsilon g(\xi, \varepsilon) d\xi \right] \\ + \frac{1}{2i} \left[\int_{\infty}^{\tilde{\omega}_1 - \tilde{\omega}_3} e^{i(z-\xi)} g(\xi, 0) d\xi - \int_{\infty}^{\tilde{\omega}_1 + 2\tilde{\omega}_3} e^{-i(z-\xi)} g(\xi, 0) d\xi \right]. \end{aligned} \tag{132}$$

According to Lemma 3.5, estimate (130) follows directly (132).

To prove (131), we first consider $\Delta_\varepsilon q^2(z, \varepsilon)$. Representing $\wp(x) = [\wp(x) - \frac{1}{x^2}] + \frac{1}{x^2} = \Psi(x) + \frac{1}{x^2}$, we note that $\Psi(x)$ is analytic in the trapezoidal region $\tilde{T}$ (vertices at $\omega_1 - \omega_3, \omega_1 + 2\omega_3, \frac{3}{2}\omega_3, -\frac{1}{2}\omega_3$). (The trapezoid $\tilde{T}$ is the image of the trapezoid T under the scaling $x = \varepsilon z$.) Since $q^2(z, 0) = \frac{1}{z^2}$ for any $z \in T$, we have

$$\Delta_\varepsilon q^2(z, \varepsilon) = \varepsilon^2 \left[\Psi(\varepsilon z) - \frac{1}{12} \right] = O(\varepsilon^2) \tag{133}$$

uniformly in T. Similarly, using (45), (46), and (71), we obtain

$$v_1(z, \varepsilon) - v_1(z, 0) = O(\varepsilon^3) \tag{134}$$

and

$$v_2(z, \varepsilon) - v_2(z, 0) = O(\varepsilon^2 z^6) \tag{135}$$

uniformly in T.

Now,

(136)

$$\begin{aligned}\Delta_\varepsilon \mathcal{I}_1 g(z,\varepsilon) = &-\left[v_1(z,\varepsilon)\int_{\tilde\omega_2}^{z} v_2(\xi,\varepsilon)g(\xi,\varepsilon)d\xi - v_1(z,0)\int_{\tilde\omega_2}^{z} v_2(\xi,0)g(\xi,0)d\xi\right]\\ &+\left[v_2(z,\varepsilon)\int_{\tilde\omega_2}^{z} v_1(\xi,\varepsilon)g(\xi,\varepsilon)d\xi - v_2(z,0)\int_{\tilde\omega_2}^{z} v_1(\xi,0)g(\xi,0)d\xi\right]\\ &+\left[-v_1(z,0)\int_{\infty}^{\tilde\omega_2} v_2(\xi,0)g(\xi,0)d\xi + v_2(z,0)\int_{\infty}^{\tilde\omega_2} v_1(\xi,0)g(\xi,0)d\xi\right].\end{aligned}$$

The third term in the right-hand side of (136) is of the order $O(\varepsilon^{b-2})$ as $\varepsilon \to 0$, while the first term can be represented as

(137)

$$\begin{aligned}&- v_1(z,\varepsilon)\int_{\tilde\omega_2}^{z}\left[v_2(\xi,\varepsilon)\Delta_\varepsilon g(\xi,\varepsilon) + g(\xi,0)\Delta_\varepsilon v_2(\xi,\varepsilon)\right]d\xi\\ &- \Delta_\varepsilon v_1(z,\varepsilon)\int_{\tilde\omega_2}^{z} v_2(\xi,0)g(\xi,0)d\xi.\end{aligned}$$

Using Lemma 2.4 together with (134), (135), we obtain estimates

(138)

$$O(z^{2-a}) + O(\varepsilon^2 z^{4-b}) + O(\varepsilon^3 z^{5-b})$$

for the three terms of (137) that are uniform in T and in small $\varepsilon \geq 0$. The second term in the right-hand side of (136) yields the same result. Now (131) follows from (138) and (94). □

THEOREM 3.13. (1) *If in condition* (53) *of Theorem* 2.1 *we require* $\alpha > 1$, *then*

(139)

$$\lim_{\varepsilon\to 0} v(z,\varepsilon) = v_+(z)$$

uniformly on compact subsets of T_{z_1}.

(2) *If condition* (53) *of Theorem* 2.1 *is replaced by*

(140)

$$c(\varepsilon) = 1 + \varepsilon^\alpha \tilde c(\varepsilon) \qquad \text{and} \qquad b(\varepsilon) = \varepsilon^\alpha \tilde b(\varepsilon),$$

where $\alpha = 1$ *and functions* $\tilde c(\varepsilon), \tilde b(\varepsilon)$ *are continuous for small* $\varepsilon \geq 0$, *then*

(141)

$$\lim_{\varepsilon\to 0} v(z,\varepsilon) = v_+(z - z_*)$$

uniformly on compact subsets of T_{z_1}, *where the translation* z_* *depends on* $\tilde b(0), \tilde c(0)$.

Proof. (1) As a consequence of (133), we obtain that (139) is equivalent to

(142)

$$\lim_{\varepsilon\to 0} w(z,\varepsilon) = w(z,0)$$

as $\varepsilon \to 0$ uniformly on compact subsets of T_{z_1}. Since $w(z,\varepsilon) = \sum_{n=1}^{\infty} \Delta w_n(z,\varepsilon)$, where the series converges uniformly for all sufficiently small $\varepsilon \geq 0$ and all $z \in T_{z_1}$, it is sufficient to prove the continuity of each $\Delta w_n(z,\varepsilon)$ in $\varepsilon \geq 0$ in order to prove the continuity of $w(z,\varepsilon)$.

According to (36) and (94),

$$f(z,0) = -6!z^{-6} \quad \text{and} \quad \Delta_\varepsilon f(z,\varepsilon) = O(\varepsilon^\alpha z^{-4}). \tag{143}$$

Then, according to Proposition 3.12,

$$\Delta_\varepsilon \mathcal{I}_1 f(z,\varepsilon) = O(\varepsilon^{\tilde\alpha} z^{-3}), \tag{144}$$

where $\tilde\alpha = \min\{\alpha - 1, 1\}$. Using

$$\Delta_\varepsilon[h(z,\varepsilon)g(z,\varepsilon)] = h(z,\varepsilon)\Delta_\varepsilon g(z,\varepsilon) + \Delta_\varepsilon h(z,\varepsilon) g(z,0) \tag{145}$$

together with (36) and (130), we prove that $\Delta w_1(z,\varepsilon) = w_1(z,\varepsilon)$ also satisfies (144).

Now $\lim_{\varepsilon\to 0} \Delta w_n(z,\varepsilon) = \Delta w_n(z,0)$, $n = 2,3,\dots$, can be proved by induction. Assume that $\Delta_\varepsilon(\Delta w_k(z,\varepsilon))$ satisfies (144) for every $k \in \mathbb{N}$, $k \le n$. We want to prove the statement for $k = n+1$. According to (94), (129), and (130), $\Delta_\varepsilon g(z,\varepsilon) = O(\varepsilon^{\tilde\alpha} z^{-6})$, where $g(z,\varepsilon)$ denotes the argument of $\mathcal{I}_1$ in (100). It is easy to see that $g(z,0) = O(z^{-2(n+3)})$. Then the fact that $\Delta_\varepsilon[\Delta w_{n+1}(z,\varepsilon)]$ satisfies (144) follows from (131). The proof of part (1) is completed.

(2) According to (126), for a given $z \in T_{z_1}$

$$\lim_{\varepsilon\to 0} \mathcal{I}_1 f = -\frac{90}{z^4} + \frac{\tilde c(0)\left\{\sum_{j=1,3} B_j[2\zeta(\omega_j) - \omega_2 e_j] - 4\omega_2\right\}}{4z^3}, \tag{146}$$

where B_j are defined by (122). Similarly,

$$\lim_{\varepsilon\to 0} \varepsilon^7 \tilde b(\varepsilon) v_{22} = -\tilde b(0)\frac{\frac{3g_3}{2}\omega_2 - g_2\zeta(\omega_2)}{3\Delta z^3}. \tag{147}$$

Thus,

$$\lim_{\varepsilon\to 0} w_1(z,\varepsilon) = -\frac{90}{z^4} + \frac{12 z_*}{z^3}, \tag{148}$$

where the constant z_*, which depend on $\tilde b(0), \tilde c(0)$, can be derived from (146), (147).

Now we can repeat the argument of part (1) to show that $\Delta w_n(z,\varepsilon)$ is continuous in ε. It is pretty straightforward to show that $\Delta w_n(z,0) = O(z^{-n-2})$ as $z \to \infty$ and that $\Delta w_n(z,0)$ has asymptotic expansion in z^{-1}. Due to uniform convergence of series (54) in Theorem 2.1, we proved that $\lim_{\varepsilon\to 0} v(z,\varepsilon)$ is a solution of the truncated equation (24) that has asymptotic expansion in z^{-1} as $z \to \infty$ with the leading terms $\frac{6}{z^2} + \frac{12 z_*}{z^3} + \cdots$. Then (141) follows from Corollary 1.3. □

4. Nonexistence of symmetric periodic solutions to (3). Let S be an interval of $\mathbb{R}$ that contains the point $\omega_1 = \omega_1(0)$, and let $y(x,\varepsilon)$ be a C^α-deformation of $y(x,0) = 6\wp_{\frac{1}{12},g_3}(x-\omega_3) - \frac{1}{2}$, where $|g_3| < 6^{-3}$, and where ω_1, ω_3, and g_3 are related through (9). As mentioned earlier, it is more convenient for us to consider $y(x,\varepsilon)$ as a C^α-deformation of $y(x,0) = 6\wp_{\frac{1}{12},g_3}(x) - \frac{1}{2}$ on the interval $S + \omega_3$. Assume that $y(x,\varepsilon)$ contains a sequence $\{y(x,\varepsilon_m)\}_1^\infty$ of solutions that are symmetrical at $x = \omega_3(\varepsilon_m)$ (after a proper translation $\beta(\varepsilon)$) and periodic (along the horizontal line $x = \omega_3(\varepsilon_m) + \mathbb{R}$) with the period $2\omega_1(\varepsilon_m)$, where $\lim_{m\to\infty} \varepsilon_m = 0$ and $\omega_1(\varepsilon_m)$ satisfy (14). Here $\omega_1(\varepsilon), \omega_3(\varepsilon)$, and $g_3(\varepsilon)$ are related through (9).

Let us show that $\beta(\varepsilon_m) = O(\varepsilon_m^\alpha)$. Indeed, since $y(x,\varepsilon)$ is a C^α-deformation of $y(x,0)$, we know that $y'(\omega_2(\varepsilon)+\beta(\varepsilon),\varepsilon) - y'(\omega_2(\varepsilon)+\beta(\varepsilon),0) = O(\varepsilon^\alpha)$ and $\lim_{\varepsilon\to 0}\beta(\varepsilon) =$

0. But solutions $y(x+\beta(\varepsilon_m),\varepsilon_m)$ are symmetrical at $\omega_3(\varepsilon_m)$ and have period $2\omega_1(\varepsilon_m)$, $m\in\mathbb{N}$, so $y'(\omega_2(\varepsilon_m)+\beta(\varepsilon_m),\varepsilon_m)=0$. Thus,

$$y'(\omega_2(\varepsilon_m)+\beta(\varepsilon_m),0)=O(\varepsilon_m^\alpha). \tag{149}$$

Since $y'(\omega_2,0)=0$, then $y'(\omega_2(\varepsilon_m)+\beta(\varepsilon_m),0)=\beta(\varepsilon_m)y''(\omega_2(\varepsilon_m)+\theta(\varepsilon_m)\beta(\varepsilon_m),0)$ by the mean value theorem, where $\theta(\varepsilon_m)\in(0,1)$. Combining this with (149) and the fact that $\wp''$ is separated from zero in a vicinity of ω_2 (see (64)), we obtain $\beta(\varepsilon_m)=O(\varepsilon_m^\alpha)$. Since derivatives of $y(x,0)$ are bounded on the line $\omega_3+\mathbb{R}$, it is easy to show that $y(x+\beta(\varepsilon_m),\varepsilon_m)$ are also C^α-deformations of $y(x,0)$. Thus, without any loss of generality, we will consider only such deformations $y(x,\varepsilon)$ of $y(x,0)$ that are symmetrical at $\omega_2(\varepsilon)$. That means that solutions $v(z,\varepsilon)$ to the inner equation (4), that correspond to such $y(x,\varepsilon)$, are symmetrical at $\tilde\omega_2(\varepsilon)$.

Let us denote by $\mathcal{F}_\alpha$, $\alpha\ge1$, the two-parameter family of solutions constructed in Theorem 2.1 that satisfy an additional assumption: condition (53) in Theorem 2.1 is replaced by (140), where $\alpha\ge1$. The proof of Theorem 1.1 is divided into the following two steps: (1) if $v(z,\varepsilon)$ is the inner solution, symmetrical at $z=\tilde\omega_2(\varepsilon)$, that corresponds to a C^α-deformation $y(x,\varepsilon)$ and if $\omega_1(\varepsilon)$ satisfies (14), then $v(z,\varepsilon)\in\mathcal{F}_\alpha$; (2) if $\omega_1(\varepsilon)$ satisfies (14), then there is no symmetrical and $2\omega_1(\varepsilon)$ periodic solution $v(z,\varepsilon)\in\mathcal{F}_\alpha$. Here and below we use $\tilde\omega_j$ to denote $\tilde\omega_j(\varepsilon)$, $j=1,2,3$, wherever such notation is clear. Additionally, we prove that the BVP

$$2v'''v'-v''^2+(1-\varepsilon^2)v'^2-\varepsilon^2v^2-\frac{2}{3}v^3=\varepsilon^6C(\varepsilon), \tag{150}$$

$$v'(\tilde\omega_3,\varepsilon)=v'(\tilde\omega_2,\varepsilon)=v'''(\tilde\omega_2,\varepsilon)=0$$

has a unique solution in $\mathcal{F}_\alpha$.

If $y(x,\varepsilon)$ is a C^α-deformation of solution (16), then the constant of motion $C(\varepsilon)$ of $y(x,\varepsilon)$, given by (11), satisfies (13). Then the corresponding inner solution $v(z,\varepsilon)$ satisfies (150) and the two latter boundary conditions of (150). It is easy to see that $v(z,\varepsilon)$ is periodic with the period $2\omega_1(\varepsilon)$ and symmetrical if and only if, additionally, $v'(\tilde\omega_3,\varepsilon)=v'''(\tilde\omega_3,\varepsilon)=0$. In other words, a solution $v(z,\varepsilon)$ to the BVP (150) is symmetric and periodic with the period $2\omega_1(\varepsilon)$ if and only if $v'''(\tilde\omega_3,\varepsilon)=0$.

4.1. Solution of BVP (150). Here we prove existence and uniqueness of solution to BVP (150) within the family $\mathcal{F}_\alpha$ using the implicit function theorem.

If $v(z,\varepsilon)$ is a solution to the BVP (150), then the integral of motion (150) evaluated at the points $z=\tilde\omega_2$, $z=\tilde\omega_3$ becomes

$$\varepsilon^2v^2(\tilde\omega_j)+\frac{2}{3}v^3(\tilde\omega_j)+v''^2(\tilde\omega_j)+\varepsilon^6C(\varepsilon)=0, \tag{151}$$

where $j=2,3$. We want to show that for sufficiently small ε there exists a unique $v\in\mathcal{F}_\alpha$ satisfying (151). To this end, we can restrict our attention on the triangle $P\subset T_{z_1}$, which was introduced in section 3.4.

According to (34), (49),

$$v=6q^2+\tilde v=6q^2+6\varepsilon^\alpha\tilde cq^2+\mathcal{I}_2\left[\varepsilon^{6+\alpha}\tilde bv_{22}+\mathcal{I}_1f(q)+\tilde w\right], \tag{152}$$

where

$$\tilde w=-12c\mathcal{I}_1q^2\mathcal{I}_2f(q)+\sum_{n=2}^{\infty}\Delta w_n. \tag{153}$$

It is easy to see that $v(z,\varepsilon) = O(\varepsilon^2)$ uniformly in P as $\varepsilon \to 0$. According to (106), (140), and Remark 3.8, we have $w(z,\varepsilon) = O(\varepsilon^{3+\tilde{\alpha}})$ uniformly in P, where $\tilde{\alpha} = \min\{1, \alpha - 1\}$ as $\varepsilon \to 0$. Then, according to (152) and Lemma 2.4,

$$\tilde{v} \sim O(\varepsilon^{3+\tilde{\alpha}}) \quad \text{and} \quad \tilde{w} \sim O(\varepsilon^{4+2\tilde{\alpha}}) \tag{154}$$

uniformly in P as $\varepsilon \to 0$. Now combining (154) and (39), we see that

$$w'' = \varepsilon^2 w + [\mathcal{I}_2 w]^2 + 12cq^2\mathcal{I}_2 w + f(q) \sim O(\varepsilon^{5+\tilde{\alpha}}) \quad \text{as} \quad \varepsilon \to 0, \tag{155}$$

uniformly on P. Then

$$w'(z,\varepsilon) = \int_{\tilde{\omega}_2}^{z} w''(t,\varepsilon)dt = O(\varepsilon^{4+\tilde{\alpha}}) \tag{156}$$

uniformly in $z \in P$. To estimate v'' in (151) we need the following statement.

PROPOSITION 4.1. *Let $g(z,\varepsilon)$ be a differentiable in $z \in T_{z_1}$ function. If there exists a constant $M > 0$ such that for all small $\varepsilon > 0$ both $|g(z,\varepsilon)|$ and $|g'(z,\varepsilon)|$ are bounded by M in T_{z_1}, then there exists some $\delta_0 > 0$ such that*

$$\mathcal{I}_2 g(z,\varepsilon) = g(z,\varepsilon) - \mathcal{I}_2 g''(z,\varepsilon) + O\left(e^{-\frac{\delta_0}{\varepsilon}\rho(z)}\right) \tag{157}$$

as $\varepsilon \to 0$ uniformly in P, where $\rho(z) = \min\{|z - (\tilde{\omega}_1 + 3\tilde{\omega}_3)|, |z - (\tilde{\omega}_1 - \tilde{\omega}_3)|\}$.

Proof. Integrating by parts $\mathcal{I}_2 g(z,\varepsilon)$, expressed by (76) and (83), twice, we obtain

$$\begin{aligned}\mathcal{I}_2 g(z,\varepsilon) = g(z,\varepsilon) - \mathcal{I}_2 g''(z,\varepsilon) - \frac{1}{2}\, e^{i(z-\xi)}\,[g(\xi,\varepsilon) - ig'(\xi,\varepsilon)]\Big|_{\xi=\tilde{\omega}_1-\tilde{\omega}_3} \\ - \frac{1}{2}\, e^{-i(z-\xi)}\,[g(\xi,\varepsilon) + ig'(\xi,\varepsilon)]\Big|_{\xi=\tilde{\omega}_1+3\tilde{\omega}_3}.\end{aligned} \tag{158}$$

The last two terms of (158) are exponentially small in ε^{-1} according to the construction of T_{z_1}. ∎

Under the assumptions of Proposition 4.1,

$$D^2\mathcal{I}_2 g = g - \mathcal{I}_2 g = \mathcal{I}_2 D^2 g + O\left(e^{-\frac{\delta_0}{\varepsilon}\rho(z)}\right). \tag{159}$$

According to (156), Proposition 4.1 is applicable for $g(z,\varepsilon) = w(z,\varepsilon)$. Thus,

$$v'' = 6c(\varepsilon)(q^2)'' + \mathcal{I}_2 w'' = O(\varepsilon^4) \quad \text{as} \quad \varepsilon \to 0, \tag{160}$$

uniformly on $[\tilde{\omega}_3, \tilde{\omega}_2]$.

Note that $y = \frac{6q^2}{\varepsilon^2}$ is the solution of the unperturbed equation (2) with the constant of motion $J(\varepsilon)$; see (37). Thus,

$$\left(\frac{6q^2}{\varepsilon^2}\right)^2 + \frac{2}{3}\left(\frac{6q^2}{\varepsilon^2}\right)^3 + J(\varepsilon) = 0 \tag{161}$$

at $x = \varepsilon z = \omega_j$, $j = 1, 2, 3$. Notice that (9) and (14) imply that

$$J(\varepsilon) = C + \varepsilon^{\alpha}\tilde{J}(\varepsilon), \tag{162}$$

where $\tilde{J}(\varepsilon)$ is a continuous function. Now substituting $v = 6q^2 + \tilde{v}$ into (151), dividing both sides by $\varepsilon^{7+\tilde{\alpha}}$, and taking into account (161), we obtain

$$2\left(\frac{6q^2}{\varepsilon^2}\right)\breve{v} + \varepsilon^{1+\tilde{\alpha}}\breve{v}^2 + 2\left(\frac{6q^2}{\varepsilon^2}\right)^2\breve{v} + 2\varepsilon^{1+\tilde{\alpha}}\left(\frac{6q^2}{\varepsilon^2}\right)\breve{v}^2 + \frac{2}{3}\varepsilon^{2+2\tilde{\alpha}}\breve{v}^3 + \frac{v''^2}{\varepsilon^{7+\tilde{\alpha}}} + \tilde{C}(\varepsilon) - \tilde{J}(\varepsilon) = 0 \tag{163}$$

at $z = \tilde{\omega}_2, \tilde{\omega}_3$, where

$$\breve{v}(z,\varepsilon) = \frac{\tilde{v}(z,\varepsilon)}{\varepsilon^{3+\tilde{\alpha}}} = \tilde{c}(\varepsilon)\left(\frac{6q^2}{\varepsilon^2}\right) + \varepsilon^4\tilde{b}(\varepsilon)\mathcal{I}_2 v_{22} + \mathcal{I}_2\frac{\mathcal{I}_1 f(q) + \tilde{w}}{\varepsilon^{3+\tilde{\alpha}}}. \tag{164}$$

Equation (163) evaluated at the points $z = \tilde{\omega}_2, \tilde{\omega}_3$ forms a system that we denote by

$$F(\varepsilon, \tilde{b}, \tilde{c}) = 0. \tag{165}$$

LEMMA 4.2. *There exist some $\varepsilon_1 \in (0, \varepsilon_0]$ and functions $\tilde{b}(\varepsilon), \tilde{c}(\varepsilon)$, continuous on $[0, \varepsilon_1]$, such that the system* (165) *holds identically on $[0, \varepsilon_1]$. Moreover, the solution $\tilde{b}(\varepsilon), \tilde{c}(\varepsilon)$ to* (165) *is unique.*

Proof. The proof is based on the implicit function theorem (see, for example, [MB, p. 122]). According to this theorem, we have to show that (a) $F(0, \tilde{b}(0), \tilde{c}(0)) = 0$ for some $\tilde{b}(0), \tilde{c}(0)$; (b) matrix

$$\operatorname{Col}\left(\frac{\partial F}{\partial \tilde{b}}, \frac{\partial F}{\partial \tilde{c}}\right)\Big|_{\varepsilon=0} \tag{166}$$

is not singular; and (c) F and $\frac{\partial F}{\partial \tilde{c}}$, $\frac{\partial F}{\partial \tilde{b}}$ are continuous in all variables in a vicinity of $(0, \tilde{b}(0), \tilde{c}(0))$.

According to (156), we can apply Proposition 4.1 to $\breve{v}(z,\varepsilon) - \tilde{c}(\varepsilon)\frac{6q^2}{\varepsilon^2} = \frac{\mathcal{I}_2 w}{\varepsilon^{3+\tilde{\alpha}}}$. Then

$$\lim_{\varepsilon\to 0}\breve{v}(z,\varepsilon) = \tilde{c}(0)\left(6\wp(\varepsilon z) - \frac{1}{2}\right) + \tilde{b}(0)\lim_{\varepsilon\to 0}\varepsilon^4 v_{22}(z,\varepsilon) + \lim_{\varepsilon\to 0}\frac{\mathcal{I}_1 f(q)}{\varepsilon^{3+\tilde{\alpha}}} + \lim_{\varepsilon\to 0}\frac{\tilde{w}}{\varepsilon^{3+\tilde{\alpha}}} \tag{167}$$

uniformly on $[\tilde{\omega}_3, \tilde{\omega}_2]$. Due to Remark 3.8, the last limit in (167) is zero.

Using (63), (69), (47), and the fact that $v_1(\tilde{\omega}_j) = 0$, we have

$$\lim_{\varepsilon\to 0}\varepsilon^4 v_{22}(\tilde{\omega}_j, \varepsilon) = -\frac{1}{12}\frac{1}{(e_j - e_i)(e_j - e_k)}. \tag{168}$$

Note that $\mathcal{I}_1 f(q)(\tilde{\omega}_2) = 0$. To calculate $\lim_{\varepsilon\to 0}\varepsilon^{-3-\tilde{\alpha}}\mathcal{I}_1 f(q)(\tilde{\omega}_3)$, we notice that

$$\wp'(x - \omega_j) = \left(6e_j^2 - \frac{g_2}{2}\right)(x - \omega_j) + O((x - \omega_j)^2) = 2(e_i - e_j)(e_k - e_j)(x - \omega_j) + O((x - \omega_j)^2) \tag{169}$$

as $x \to \omega_j$, $j = 1, 2, 3$. It follows then from (126) that

$$\varepsilon^{-3}\mathcal{I}_1 f(q)(\omega_3, \varepsilon) = 12c(\varepsilon)(e_1 - e_3)(e_2 - e_3)G_3 = \frac{c(\varepsilon)}{e_1 - e_3}\left\{-\frac{c(\varepsilon) - 1 - 4\varepsilon^2}{4\varepsilon}\left(e_1 + \frac{1}{6}\right) - \frac{5\varepsilon}{2}\left(\frac{9}{4}e_3 + \frac{3}{2}e_2 + 36e_2^3 - \frac{1}{3}\right) + \varepsilon C(\varepsilon)\right\}. \tag{170}$$

Using (126), we obtain

$$\lim_{\varepsilon\to 0}\varepsilon^{-3-\tilde\alpha}\mathcal{I}_1 f(q)(\tilde\omega_3 = -\frac{\tilde c(\varepsilon)(e_1+\frac16)}{4(e_1-e_3)} - \delta_{\tilde\alpha,1}\frac{e_1+\frac16+\frac52\left(\frac94 e_3+\frac32 e_2+36e_2^3-\frac13\right)+36g_3+\frac16}{e_1-e_3} \tag{171}$$

at $z=\tilde\omega_3$, where $\delta_{\tilde\alpha,1}=1$ if $\tilde\alpha=1$ and $\delta_{\tilde\alpha,1}=0$ otherwise. Combining (167)–(171), we obtain

$$\lim_{\varepsilon\to 0}\breve v(\tilde\omega_2) = \left(6e_2-\frac12\right)\tilde c(0)+\frac{\tilde b(0)}{12(e_1-e_2)(e_2-e_3)}, \tag{172}$$

$$\lim_{\varepsilon\to 0}\breve v(\tilde\omega_3) = \left[\left(6e_3-\frac12\right)-\frac{e_1+\frac16}{4(e_1-e_3)}\right]\tilde c(0)-\frac{\tilde b(0)}{12(e_1-e_3)(e_2-e_3)}-\delta_{\tilde\alpha,1}\tilde D,$$

where $\tilde D$ denotes the last fraction in (171).

According to (154) and (160), in the limit $\varepsilon\to 0$ equations (163) become

$$2\left[36\wp^2(\omega_j)-\frac14\right]\lim_{\varepsilon\to 0}\breve v(\tilde\omega_j)+\tilde C(0)-\tilde J(0)=0, \tag{173}$$

where $j=2,3$. Clearly, (172)–(173) form a linear system of equations for the unknowns $\tilde b(0),\tilde c(0)$. Note that the first factor in (173) is different from zero since, otherwise, according to (64), we would have either $C=0$ or $C=-\frac13$, which contradicts our assumption $C\in(-\frac13,0)$. After some algebra, we calculate the determinant of the matrix of linear system (172)–(173) as

(174)

$$\begin{vmatrix} 6e_2-\frac12 & \frac{1}{12(e_1-e_2)(e_2-e_3)}\\ 6e_3-\frac12-\frac{e_1+\frac16}{4(e_1-e_3)} & -\frac{1}{12(e_1-e_3)(e_2-e_3)}\end{vmatrix} = \frac{1}{3\sqrt{\Delta}}\begin{vmatrix}(6e_2-\frac12)(e_1-e_2) & 1\\ (6e_3-\frac12)(e_1-e_3)-\frac{e_1+\frac16}{4} & -1\end{vmatrix}$$

$$= \frac{7(e_1+\frac16)}{12\sqrt{\Delta}},$$

where Δ, defined by (67), is different from zero. This determinant is different from zero since $e_1>0$. Thus, we have established parts (a) and (b) of the lemma; i.e., we have established that $F(0,\tilde b(0),\tilde c(0))=0$ at $z=\tilde\omega_2,\tilde\omega_3$ for some $\tilde b(0),\tilde c(0)$ and that matrix (166) is nonsingular. Continuity of F follows from Theorem 3.13 and the fact that iterations $\Delta w_n(z,\varepsilon)$ in the solution (54) are polynomials in $\tilde b(0),\tilde c(0)$ of degree not exceeding n. Estimates of Theorem 2.1 can be readily adjusted to prove convergence of the series $\sum_{n=1}^\infty\frac{\partial\Delta w_n(z,\varepsilon)}{\partial\tilde c}$ and $\sum_{n=1}^\infty\frac{\partial\Delta w_n(z,\varepsilon)}{\partial\tilde b}$ at $z=\tilde\omega_2$, $z=\tilde\omega_3$. Part (c) and the whole proof are completed. □

THEOREM 4.3. *If $C(\varepsilon)$ and $\omega_1(\varepsilon)$ satisfy* (13) *and* (14)*, respectively, where $\alpha\ge 1$, then the BVP* (150) *has a unique solution in $\mathcal{F}_\alpha$.*

Proof. According to Lemma 4.2, it is sufficient to show that (151) at $z=\tilde\omega_3$ and the rescaled integral of motion (150) imply $v'(\tilde\omega_3)=0$. Indeed, addition of these two equations yields

$$v'(\tilde\omega_3)[2v'''(\tilde\omega_3)+(1-\varepsilon^2)v'(\tilde\omega_3)]=0. \tag{175}$$

That means that $v'(\tilde{\omega}_3) = 0$ or $2v'''(\tilde{\omega}_3) + (1-\varepsilon^2)v'(\tilde{\omega}_3) = 0$. According to (34), and since operators D and D^2+1 commute, this implies $u'(\tilde{\omega}_3) = 0$ or $2w'(\tilde{\omega}_3) = (1+\varepsilon^2)u'(\tilde{\omega}_3)$. If $u'(\tilde{\omega}_3) = 0$, then $v'(\tilde{\omega}_3) = 0$, and the proof is completed. The assumption $u'(\tilde{\omega}_3) \neq 0$ leads to a contradiction. Indeed, on one hand, we have

$$u'(\tilde{\omega}_3) = \frac{2}{1+\varepsilon^2} w'(\tilde{\omega}_3), \tag{176}$$

whereas, on the other hand, according to Proposition 4.1,

$$u'(\tilde{\omega}_3) = \mathcal{I}_2 w'|_{z=\tilde{\omega}_3} = w'(\tilde{\omega}_3) - \mathcal{I}_2 w'''|_{z=\tilde{\omega}_3} + O\left(e^{-\frac{\delta_0}{\varepsilon}}\right). \tag{177}$$

Proposition 4.1 is applicable to $\mathcal{I}_2 w'$ according to (39) and (156). Moreover, differentiating (39), we obtain $w''' = O(\varepsilon^{6+\tilde{\alpha}})$ uniformly on $[\tilde{\omega}_3, \tilde{\omega}_2]$. Thus, (177) contradicts (176). □

4.2. Proof of Theorem 1.1.

Proof. As mentioned in the beginning of section 4, our proof consists of two steps. Step (1) is to show that the inner solution $v(z,\varepsilon) = \varepsilon^2 y(\varepsilon z, \varepsilon)$, corresponding to a C^α-deformation $y(x,\varepsilon)$ of the periodic solution $y(x,0)$, is an $\mathcal{F}_\alpha$ solution. Let us consider first the case $n = 1$.

As discussed above, we can assume without any loss of generality that $v(z,\varepsilon_m)$ is symmetrical with respect to $z = \tilde{\omega}_3(\varepsilon_m)$ and $z = \tilde{\omega}_2(\varepsilon_m)$, where $g_3(\varepsilon_m)$ and $\omega_3(\varepsilon_m)$ are defined by $\omega_1(\varepsilon_m)$ through (9). Then $v'(\tilde{\omega}_2(\varepsilon_m),\varepsilon_m) = v'''(\tilde{\omega}_2(\varepsilon_m),\varepsilon_m) = 0$, so that $v(z,\varepsilon_m)$ is defined by initial conditions $v(\tilde{\omega}_2(\varepsilon_m),\varepsilon_m)$ and $v''(\tilde{\omega}_2(\varepsilon_m),\varepsilon_m)$. In order to show that there is a solution in $\mathcal{F}_\alpha$ with the abovementioned initial conditions at $z = \tilde{\omega}_2(\varepsilon_m)$, we first establish that

$$v(\tilde{\omega}_2(e_n),\varepsilon_m) - \varepsilon_m^2\left(6e_2(0) - \frac{1}{2}\right) = \varepsilon_m^{2+\alpha}\tilde{v}(\varepsilon_m), \tag{178}$$

$$v''(\tilde{\omega}_2(e_n),\varepsilon_m) - \varepsilon_m^4\left(3e_2^2(0) - \frac{1}{48}\right) = \varepsilon_m^{4+\alpha}\tilde{v}''(\varepsilon_m),$$

where $\tilde{v}(\varepsilon)$, $\tilde{v}''(\varepsilon)$ are continuous functions. Indeed, according to our assumptions and taking into account (14),

$$\begin{aligned} y(\omega_2(\varepsilon_m),\varepsilon_m) - y(\omega_2,0) \\ = [y(\omega_2(\varepsilon_m),\varepsilon_m) - y(\omega_2(\varepsilon_m),0)] + [y(\omega_2(\varepsilon_m),0) - y(\omega_2,0)] \\ = \varepsilon_m^\alpha \tilde{y}(\omega_2(\varepsilon_m),\varepsilon_m) + \frac{1}{2}\varepsilon_m^\alpha \tau^2(\varepsilon_m)\wp''(\omega_2(0) + \theta(\omega_1(\varepsilon_m) - \omega_1)), \end{aligned} \tag{179}$$

where $\theta \in (0,1)$. Here we used the mean value theorem and the fact that $\wp'(\omega_2(0)) = 0$. A similar estimate holds for $y''(\omega_2(\varepsilon_m),\varepsilon_m) - y''(\omega_2,0)$. Then, (178) follows from (179).

According to (164), equations (178) in the leading order can be written as

$$\tilde{c}(\varepsilon)\left(\frac{6q^2}{\varepsilon^2}\right)+\varepsilon^4\tilde{b}(\varepsilon)\mathcal{I}_2v_{22}+\mathcal{I}_2\frac{\mathcal{I}_1f(q)+\tilde{w}}{\varepsilon^{3+\tilde{\alpha}}} \tag{180}$$

$$=\tilde{v}(\varepsilon)-\frac{1}{2}\tau^2(\varepsilon)\wp''(\omega_2(0)+\theta(\omega_1(\varepsilon)-\omega_1(0))),$$

$$\tilde{c}(\varepsilon)D_z^2\left(\frac{6q^2}{\varepsilon^2}\right)+\varepsilon^4\tilde{b}(\varepsilon)D_z^2\mathcal{I}_2v_{22}+D_z^2\mathcal{I}_2\frac{\mathcal{I}_1f(q)+\tilde{w}}{\varepsilon^{3+\tilde{\alpha}}}$$

$$=\tilde{v}''(\varepsilon)-\frac{3}{2}(e_2(\varepsilon)+e_2(0))\tau^2(\varepsilon)\wp''(\omega_2(0)+\theta(\omega_1(\varepsilon)-\omega_1(0))).$$

According to (167) and Proposition 4.1, in the limit $\varepsilon\to 0$ equations (180) become

$$\tilde{c}(0)\lim_{\varepsilon\to0}\frac{6q^2}{\varepsilon^2}+\tilde{b}(0)\lim_{\varepsilon\to0}\varepsilon^4v_{22}(\tilde{\omega}_2(\varepsilon),\varepsilon)=\tilde{v}(0)-\frac{1}{2}\tau^2\wp''(\omega_2(0)), \tag{181}$$

$$\tilde{c}(0)\lim_{\varepsilon\to0}\left[\frac{6q^2}{\varepsilon^2}+\frac{36q^4}{\varepsilon^4}\right]+\tilde{b}(0)\lim_{\varepsilon\to0}\left\{\varepsilon^4v_{22}(\tilde{\omega}_2(\varepsilon),\varepsilon)\left[1+\frac{12q^2}{\varepsilon^2}\right]\right\}=\tilde{v}''(0)$$

$$-3e_2(0)\tau^2(0)\wp''(\omega_2(0)).$$

Here we used the fact that $\frac{6q^2}{\varepsilon^2}$ and v_{22} satisfy differential equations (2) and (44), respectively. Calculation of the determinant of the latter system yields

$$\lim_{\varepsilon\to0}\begin{vmatrix}\frac{6q^2}{\varepsilon^2} & \varepsilon^4v_{22}\\ \frac{6q^2}{\varepsilon^2}+\frac{36q^4}{\varepsilon^4} & \left[1+\frac{12q^2}{\varepsilon^2}\right]\varepsilon^4v_{22}\end{vmatrix}=\lim_{\varepsilon\to0}\begin{vmatrix}\frac{6q^2}{\varepsilon^2} & \varepsilon^4v_{22}\\ \frac{36q^4}{\varepsilon^4} & 12q^2\varepsilon^2v_{22}\end{vmatrix} \tag{182}$$

$$=36\lim_{\varepsilon\to0}q^4v_{22}=\frac{3(e_2(0)-\frac{1}{2})^2}{(e_1(0)-e_2(0))(e_2(0)-e_3(0))}\neq0.$$

Following the arguments of Lemma 4.2, we can now use the implicit function theorem to show that the system (180) has a unique solution $\tilde{c}(\varepsilon),\tilde{b}(\varepsilon)$, where $\tilde{c}(\varepsilon),\tilde{b}(\varepsilon)$ are continuous functions. Thus, $v(z,\varepsilon)\in\mathcal{F}_\alpha$. To prove step (1) for the case for general n we simply have to consider the region $\mathcal{P}$ from Remark 3.10 instead of the triangle P and the point $n\tilde{\omega}_1(\varepsilon)+\tilde{\omega}_3(\varepsilon)$ instead of $\tilde{\omega}_2(\varepsilon)$.

To prove step (2), we assume that there is a family of solutions $v(z,\varepsilon_m)\subset\mathcal{F}_\alpha$ to (4), where $\omega_1(\varepsilon)$ satisfies (14), that is symmetrical at $\tilde{\omega}_3(\varepsilon_m)$. However, according to Theorem 3.13,

$$\lim_{m\to\infty}v(z,\varepsilon_m)=v_+(z) \tag{183}$$

uniformly in z on any closed segment of the imaginary axis that belongs to T_{z_1}. Due to the symmetry with respect to the imaginary axis, we also have

$$\lim_{m\to\infty}v(z,\varepsilon_m)=v_-(z) \tag{184}$$

uniformly in z on the same segment of the imaginary axis. But, according to Corollary 1.3, $v_+(z)\not\equiv v_-(z)$. The obtained contradiction proves nonexistence of a family of symmetric periodic solutions $v(z,\varepsilon_m)\subset\mathcal{F}_\alpha$, $\alpha>1$, where $\varepsilon_m\to0$. The same result, according to Theorem 3.13, part (2), holds for the case $\alpha=1$. □

REFERENCES

[AM] C. J. Amick and J. B. McLeod, *A singular perturbation problem in water waves*, Stability Appl. Anal. Contin. Media, 1 (1991), pp. 127–148.

[BPBA] R. E. Beardmore, M. A. Peletier, C. J. Budd, and M. Ahmer Wadee, *Bifurcations of periodic solutions satisfying the zero-Hamiltonian constraint in reversible differential equations*, SIAM J. Math. Anal., 36 (2005), pp. 1461–1488.

[Eck] W. Eckhaus, *Singular perturbations of homoclinic orbits in* $\mathbb{R}^4$, SIAM J. Math. Anal., 23 (1992), pp. 1269–1290.

[GR] I. S. Gradsteyn and I. W. Ryzhik, *Tables of Integrals, Series and Products*, Academic Press, New York, 1965.

[GJ] R. Grimshaw and N. Joshi, *Weakly nonlocal solitary waves in a singularly perturbed Korteweg–de Vries equation*, SIAM J. Appl. Math., 55 (1995), pp. 124–135.

[Ha] H. Hancock, *Theory of Elliptic Functions*, Dover, New York, 1958.

[HM] J. M. Hammersley and G. Mazzarino, *Computational aspects of some autonomous differential equations*, Proc. Roy. Soc. London Ser. A, 424 (1989), pp. 19–37.

[IL1] G. Iooss and E. Lombardi, *Normal forms with exponentially small remainder: Application to homoclinic connections for the reversible* $0^{2+}i\omega$ *resonance*, C. R. Math. Acad. Sci. Paris, 339 (2004), pp. 831–838.

[IL2] G. Iooss and E. Lombardi, *Normal forms with exponentially small remainder for analytic vector fields*, J. Differential Equations, 212 (2005), pp. 1–61.

[KS] M. D. Kruskal and H. Segur, *Asymptotics beyond all orders in a model of crystal growth*, Stud. Appl. Math., 85 (1991), pp. 129–181.

[Lo] E. Lombardi, *Oscillatory Integrals and Phenomena Beyond All Algebraic Orders. With Applications to Homoclinic Orbits in Reversible Systems*, Lecture Notes in Math. 1741, Springer-Verlag, Berlin, 2000.

[MB] E. J. McShane and T. A. Botts, *Real Analysis*, Van Nostrand Company, New York, 1959.

[PT] L. A. Peletier and W. C. Troy, *Spatial Patterns: Higher Order Models in Physics and Mechanics*, Birkhäuser Boston, Boston, 2001.

[PRG] Y. Pomeau, A. Ramani, and B. Grammaticos, *Structural stability of the Korteweg-de Vries solutions under a singular perturbation*, Phys. D, 31 (1988), pp. 127–134.

[STL] H. Segur, S. Tanveer, and H. Levine, eds., *Asymptotics beyond All Orders*, NATO ASI Ser. B. Phys. 284, Plenum, New York, 1991.

[To2] A. Tovbis, *Asymptotics beyond all orders and analytic properties of inverse Laplace transforms of solutions*, Comm. Math. Phys., 163 (1994), pp. 245–255.

[To4] A. Tovbis, *Breaking homoclinic connections for a singularly perturbed differential equation and the Stokes phenomenon*, Stud. Appl. Math., 104 (2000), pp. 353–386.

[To5] A. Tovbis, *On approximation of stable and unstable manifolds and the Stokes phenomenon*, in Nonlinear PDE's, Dynamics and Continuum Physics, Contemp. Math. 255, AMS, Providence, RI, 2000, pp. 199–228.

[TP] A. Tovbis and D. Pelinovsky, *Exact conditions for existence of homoclinic orbits in the fifth-order KdV model*, Nonlinearity, 19 (2006), pp. 2277–2312.

[WW] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, Cambridge University Press, Cambridge, UK, 1927.

[Wa] W. Wasow, *Asymptotic Expansions for Ordinary Differential Equations*, Robert E. Krieger Publishing, Huntington, NY, 1976.

SIAM J. MATH. ANAL.
Vol. 40, No. 4, pp. 1550–1584

RISE OF CORRELATIONS OF TRANSFORMATION STRAINS IN RANDOM POLYCRYSTALS*

LEONID BERLYAND†, OSCAR BRUNO‡, AND ALEXEI NOVIKOV§

Abstract. We investigate the statistics of the transformation strains that arise in random martensitic polycrystals as boundary conditions cause its component crystallites to undergo martensitic phase transitions. In our laminated polycrystal model the orientation of the n grains (crystallites) is given by an uncorrelated random array of the orientation angles θ_i, $i = 1, \ldots, n$. Under imposed boundary conditions the polycrystal grains may undergo a martensitic transformation. The associated transformation strains $\boldsymbol{\varepsilon_i}$, $i = 1, \ldots, n$ depend on the array of orientation angles, and they can be obtained as a solution to a nonlinear optimization problem. While the random variables θ_i, $i = 1, \ldots, n$ are uncorrelated, the random variables $\boldsymbol{\varepsilon_i}$, $i = 1, \ldots, n$ may be correlated. This issue is central in our considerations. We investigate it in following three different scaling limits: (i) Infinitely long grains (laminated polycrystal of height $L = \infty$); (ii) Grains of finite but large height ($L \gg 1$); and (iii) Chain of short grains ($L = l_0/(2n)$, $l_0 \ll 1$). With references to de Finetti's theorem, Riesz' rearrangement inequality, and near neighbor approximations, our analyses establish that under the scaling limits (i), (ii), and (iii) the arrays of transformation strains arising from given boundary conditions exhibit no correlations, long-range correlations, and exponentially decaying short-range correlations, respectively.

Key words. polycrystals, misfit, phase transitions, correlations, De Finetti's theorem, Riesz' rearrangement inequality

AMS subject classifications. 35J20, 74N15, 82B44

DOI. 10.1137/070679685

1. Introduction.

We investigate the statistics of the transformation strains (misfits) that arise in random martensitic polycrystals as boundary conditions cause its component crystallites to undergo solid-to-solid (martensitic) phase transitions. Martensitic transformations are shape-deforming phase transitions that can be induced in certain alloys as a result of changes in the imposed strains, stresses, or temperatures. These transitions occur when a crystalline solid transforms between its parent phase (austenite) and any of a number of variants of the product phase (martensite). We focus on a setting that, while sufficiently simple to allow for a complete analytical treatment, provides significant insights on the problem: We study laminated polycrystals that consist of sequences of n of grains of rectangular cross-section—of base $1/n$ and height $L = L(n)$, so that a complete polycrystal is an infinite parallelepiped with rectangular cross-section of base 1 and height L. The goal of this work is to provide a rigorous probabilistic theory for the misfit statistics in such polycrystals and, in particular, to provide a rationale for the approximations implicit in

*Received by the editors January 9, 2007; accepted for publication (in revised form) June 24, 2008; published electronically November 19, 2008.
http://www.siam.org/journals/sima/40-4/67968.html

†Department of Mathematics, Pennsylvania State University, University Park, PA 16802 (berlyand@math.psu.edu). This author was supported by NSF grants DMS-0204637 and DMS-0708324.

‡California Institute of Technology, Applied & Computational Mathematics, 1200 E. California Boulevard, MC 217-50, Pasadena, CA 91125 (bruno@acm.caltech.edu). This author acknowledges support from NSF grant DMS-0408040, AFOSR grant FA9550-05-1-0006, and NRC-JPL award 1263315.

§Department of Mathematics, Pennsylvania State University, University Park, PA 16802 (anovikov@math.psu.edu). This author was supported by NSF grant DMS-0604600.

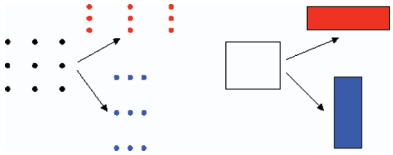


FIG. 1.1. *A reference crystallite undergoes stress-free transformations: Atomic view (left) and macroscopic view (right).*

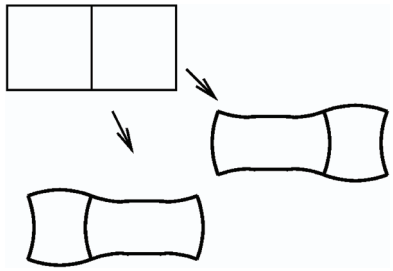


FIG. 1.2. *One of two grains undergoes a stress-free transformation.*

the numerical algorithms [6, 7, 8] for polycrytalline phase transitions in two- and three-dimensional space.

The microstructure in a laminated polycrystal is described by a sequence of the orientation angles θ_i, $i = 1, \dots, n$; θ_i represents the orientation of the two-dimensional lattice structure in the ith grain. We assume θ_i is a sequence of n independent identically distributed (i.i.d.) random variables. The transformation in the ith grain gives rise to a strain tensor, the transformation strain $\boldsymbol{\varepsilon}_i^{\mathrm{T}}$, $(i = 1, \dots, n)$, which is constant, and it takes one of three *admissible* values: No deformation (the original square lattice remains square), or deformation into one of two rectangular crystalline lattices parallel to the original square lattice. The phase transition in the polycrystal gives rise to a sequence of transformation strains $\boldsymbol{\varepsilon}_i^{\mathrm{T}}$, $i = 1, \dots, n$ obtained by the minimization of the elastic *misfit energy* among all admissible configurations.

We briefly explain the concept of *misfit* using a simple example of a polycrystal with two grains. Assume a rectangular single-crystalline grain, considered separately, can undergo a stress-free (two-dimensional version of the) cubic-to-orthorhombic phase deformation into shapes depicted in Figure 1.1. A polycrystal with two square grains can undergo deformations as depicted in Figure 1.2. The elastic energy of the former transformation is zero, because it is stress-free. In contrast, the latter transformation requires some elastic *misfit energy* that arises because when two crystallites are combined in a polycrystal, their boundaries must remain coherent after the transformation. In general, minimization of misfit energy leads to interactions amongst all of the grains in a polycrystal. Our probabilistic setup allows us to provide a rigorous description of this phenomenology.

The main results of this paper characterize the probability distributions of the random variables $\boldsymbol{\varepsilon}_i^{\mathrm{T}}$ that arise as minimizers of the overall elastic energy for a given i.i.d. distribution of the angle sequence θ_i. Such results are provided in three different cases according to whether the grains are (1) infinitely long ($L = \infty$), (2) of finite but large height ($L = L \gg 1$), and (3) short height ($L = l_0/(2n)$, $l_0 \ll 1$). In case (1) our treatment applies to arbitrary i.i.d. probability measures ρ defining the distribution of angles, and in cases (2) and (3), in turn, we restrict consideration to

i.i.d. distribution of angles with Bernoulli probability measures ρ. Our main results can be briefly described as follows:

1. *Infinitely long grains. Theorem* 5.2. For an arbitrary i.i.d. distribution of angles θ_i, $i = 1, 2, \ldots, n$, under certain technical assumptions, in the limit $n \to \infty$, the transformation strains $\boldsymbol{\varepsilon}_i^{\mathrm{T}}$, $i = 1, 2 \ldots, n$ are also i.i.d. with probability measure μ, where the measure μ is the minimizer of a certain functional ((5.8) below). In particular, in the case of infinitely long grains there are no correlations between transformation strains of any two grains.
2. *Long finite grains. $L \gg 1$. Theorem* 6.4. If θ_i, $i = 1, \ldots, n$ are Bernoulli random variables (4.3), then in the limit $n \to \infty$, $\boldsymbol{\varepsilon}_i^{\mathrm{T}}$, $i = 1, 2 \ldots, n$ have long-range but no short-range correlations.
3. *Short grains. $L = l_0/(2n)$, $l_0 \ll 1$. Theorems* 7.4 *and* 7.5. If θ_i, $i = 1, \ldots, n$ are Bernoulli random variables (4.3), then in the limit $n \to \infty$, $\boldsymbol{\varepsilon}_i^{\mathrm{T}}$, $i = 1, 2 \ldots, n$ have short-range but no long-range correlations.

Results 2 and 3 can be explained as follows. The cornerstone of our study is the maximization of an integral energy functional (see (3.4) below) of the form $\int K_L(x - t) f(x) f(t) dx dt$. Its integral kernel $K_L(x)$ decays on different length scales for long and short grains. For long grains it decays on the length scale of the composite (on $O(1)$ scale), while for short grains it decays on the length scale of a grain (on $O(1/n)$ scale). Maximization with respect to this integral kernel leads to long-range and short-range correlations for long and short grains, respectively. Formally, correlations arise because grains that undergo the stress-free transformation tend to "group together" on the scale of the decay of the integral kernel. We justify this heuristic idea in the case of long grains (see section 6) by applying a randomized version of the Riesz rearrangement inequality. In the case of short grains (see section 7) we show the transforming grains group together—by applying an isoperimetric inequality.

The paper is organized as follows: After describing in section 2 our model of the polycrystal, in section 3 we solve an auxiliary linear elasticity problem, and we obtain an explicit expression for the stored elastic energy for a fixed admissible array of transformation strains. In section 4 we describe our probabilistic model. Our main results are then established in the next three sections, where the nonlinear minimization problem for a random polycrystal is solved. The cases concerning infinitely long grains, finitely long grains, and short grains are studied in sections 5, 6, and 7, respectively.

2. Formulation.

Stress-free transformation. A two-dimensional polycrystal is a collection of grains. In our model, each grain is a single crystal (a crystallite) which can undergo a shape-deforming phase transition that results in a transformation strain. An untransformed grain with a horizontal-vertical square lattice (angle $\theta = 0$) may either elongate in the horizontal direction and remain unchanged in the vertical direction (the upper-right state in Figure 1.1); it may elongate in the vertical direction and remain unchanged in the horizontal direction (the lower right state in Figure 1.1); or, finally, it may not transform at all and thus have its size unchanged (the left state in Figure 1.1). These states correspond to the transformation strains:

$$\boldsymbol{\varepsilon}_0^1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \ \boldsymbol{\varepsilon}_0^2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \ \boldsymbol{\varepsilon}_0^0 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}. \tag{2.1}$$

The first and the second state correspond to a nontrivial transformation. The null strain $\boldsymbol{\varepsilon}_0^0$ corresponds to absence of transformation.

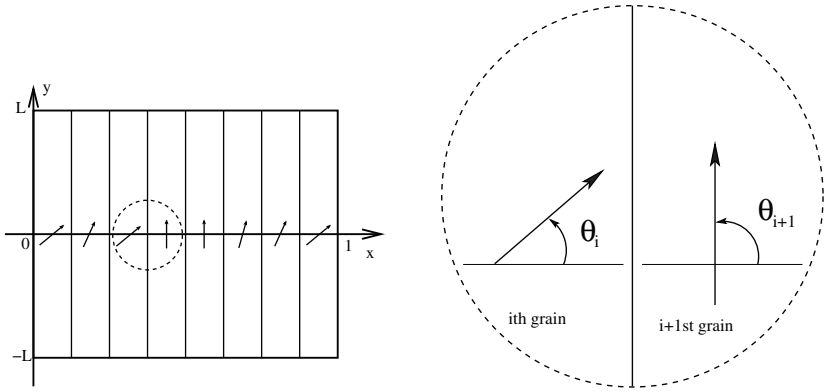


FIG. 2.1. *The laminated polycrystal.*

Mathematical model of a laminated polycrystal. Grains in a polycrystal have a varying orientation of the crystalline lattices. We consider a rectangular polycrystal $\Pi_L = [0,1] \times [-L, L]$ (see Figure 2.1) partitioned into n vertical rectangular layers (the *grains*) of width $1/n$ and height $2L$,

$$\Pi_L = \cup_{i=1}^n \Pi_L^i,\ \Pi_L^i = \left[\frac{i-1}{n}, \frac{i}{n}\right] \times [-L, L].$$

Each grain Π_L^i is occupied by a crystallite obtained by rotation by the *orientation angle*

$$\theta_i,\ 0 \le \theta_i \le \pi/2,$$

of the reference crystallite (see Figure 2.1).

The array of crystallites' orientations is completely determined by the vector of the orientation angles

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n). \tag{2.2}$$

Using the matrix of rotation by an angle θ

$$R = R(\theta) = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix},$$

we see that the stress-free transformation strain for the grain Π_L^i must lie in the set

$$\mathcal{S}_{\theta_i} = \left\{\boldsymbol{\varepsilon}^1(\theta_i), \boldsymbol{\varepsilon}^2(\theta_i), \boldsymbol{\varepsilon}^0(\theta_i)\right\},\ 0 \le \theta_i \le \pi/2, \tag{2.3}$$

where

$$\begin{aligned} \boldsymbol{\varepsilon}^1(\theta) &= R\boldsymbol{\varepsilon}^1(0)R^t = \begin{pmatrix} \cos^2(\theta) & \sin(\theta)\cos(\theta) \\ \sin(\theta)\cos(\theta) & \sin^2(\theta) \end{pmatrix}, \\ \boldsymbol{\varepsilon}^2(\theta) &= R\boldsymbol{\varepsilon}^2(0)R^t = \begin{pmatrix} \sin^2(\theta) & -\sin(\theta)\cos(\theta) \\ -\sin(\theta)\cos(\theta) & \cos^2(\theta) \end{pmatrix}, \\ \boldsymbol{\varepsilon}^0(\theta) &= R\boldsymbol{\varepsilon}^0(0)R^t = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}. \end{aligned} \tag{2.4}$$

The superscript t stands for the matrix transpose. The set of all sequences of strains that are admissible for some sequence of angles is denoted by $\tilde{\Omega}_n$:

$$\tilde{\Omega}_n = \left\{ \boldsymbol{\varepsilon}^{\mathrm{T}} | \boldsymbol{\varepsilon}^{\mathrm{T}} = (\boldsymbol{\varepsilon}_1^{\mathrm{T}}, \boldsymbol{\varepsilon}_2^{\mathrm{T}}, \boldsymbol{\varepsilon}_3^{\mathrm{T}}, \ldots, \boldsymbol{\varepsilon}_n^{\mathrm{T}}),\ \boldsymbol{\varepsilon}_i^{\mathrm{T}} \in \mathcal{S}_{\theta_i} \text{ for some } \theta_i \in [0, \pi/2] \right\}. \tag{2.5}$$

The set of all sequences of strains that are admissible for a given sequence $\boldsymbol{\theta}$ will be denoted by

$$\tilde{\Omega}_n(\boldsymbol{\theta}) = \left\{ \boldsymbol{\varepsilon}^{\mathrm{T}} | \boldsymbol{\varepsilon}^{\mathrm{T}} \in \tilde{\Omega}_n \text{ such that } \boldsymbol{\varepsilon}_i^{\mathrm{T}} \in \mathcal{S}_{\theta_i} \right\}. \tag{2.6}$$

Linear elasticity equations for given transformation strains. For a given sequence of the orientation angles $\boldsymbol{\theta} = \{\theta_i, i = 1, \ldots, n\}$ there are up to 3^n corresponding sequences $\boldsymbol{\varepsilon}^{\mathrm{T}} = \{\boldsymbol{\varepsilon}_i^{\mathrm{T}}, i = 1, \ldots, n\}$ in the class $\tilde{\Omega}_n(\boldsymbol{\theta})$ defined in (2.6). Here we introduce the relevant elasticity PDEs on the domain Π_L for a *given* such $\boldsymbol{\varepsilon}^{\mathrm{T}}$. We assume that each grain can be described by isotropic elasticity equations with elastic moduli given by

$$c_{ijkl} = \lambda \delta_{ij} \delta_{kl} + \mathrm{G}(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}), \tag{2.7}$$

where λ and G are the Lamé constants [19].

As an applied displacement is imposed, our polycrystal may acquire microscopic strains $\boldsymbol{\varepsilon}$ which contain combined contributions of elastic and stress-free transformations (see [9]):

$$\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^{\text{elastic}} + \boldsymbol{\varepsilon}^{\mathrm{T}}. \tag{2.8}$$

Then Hooke's law $\sigma_{ij} = c_{ijkl}\varepsilon_{kl}^{\text{elastic}}$ yields the stress-strain relation of *linear elasticity* under a given transformation strain $\boldsymbol{\varepsilon}^{\mathrm{T}}$

$$\sigma_{ij} = c_{ijkl}(\varepsilon_{kl} - \varepsilon_{kl}^{\mathrm{T}}). \tag{2.9}$$

Here the strain tensor ε_{kl} is determined by the displacement vector $\boldsymbol{u} = (u_1(x, y), u_2(x, y))$:

$$\varepsilon_{ij} = \frac{1}{2}\left(\partial_i u_j + \partial_j u_i\right) \text{ where } \partial_1 = \frac{\partial}{\partial x},\ \partial_2 = \frac{\partial}{\partial y}. \tag{2.10}$$

The stress tensor satisfies the elasticity equations

$$\partial_j \sigma_{ij} = 0 \text{ for } i = 1, 2. \tag{2.11}$$

The above equation is to be understood in the distributional sense, and thus the traction must be continuous across the interfaces between grains:

$$[\sigma_{i1}](x, y) = 0, \text{ for } i = 1, 2 \text{ and } x = m/n,\ m = 1, 2, \ldots, n-1. \tag{2.12}$$

For such a *given* admissible configuration $\boldsymbol{\varepsilon}_i^{\mathrm{T}}$, we assume a *given* imposed displacement that is chosen in the direction transversal to the laminates:[1]

$$u_1(0, y) = 0,\ u_1(1, y) = U,\ u_2(0, 0) = 0, \tag{2.13}$$

together with the zero-traction boundary conditions

$$\sigma_{12}(0, y) = \sigma_{12}(1, y) = 0,\ \sigma_{i2}(x, \pm L) = 0. \tag{2.14}$$

[1] Other boundary conditions (e.g., shears) could be treated similarly.

It is easy to check that, for a fixed admissible configuration $\boldsymbol{\varepsilon}^{\mathrm{T}}$, (2.11), (2.13), and (2.14) are the Euler–Lagrange equations for the minimizer of the elastic energy

(2.15)
$$W(U,\boldsymbol{\varepsilon}^{\mathrm{T}}) = \frac{1}{E_c}\min_{\boldsymbol{u}}\frac{1}{2L}\int_{\Pi_L}(\varepsilon_{ij}-\varepsilon_{ij}^{\mathrm{T}})c_{ijkl}(\varepsilon_{kl}-\varepsilon_{kl}^{\mathrm{T}})dxdy, \quad \boldsymbol{u} \text{ subject to (2.13)},$$

where

$$E_c = \frac{4\mathrm{G}(\lambda+\mathrm{G})}{\lambda+2\mathrm{G}}, \tag{2.16}$$

is the two-dimensional Young modulus. Thus it can be verified that the boundary value problem (2.11), (2.13), and (2.14) admits a unique solution.

Overall polycrystalline energy. As a displacement (2.13) is imposed on the polycrystal, each grain may undergo a stress-free transformation into one of the three possible stress-free states. The overall energy $\mathcal{W}_n(U,\boldsymbol{\theta})$ in the polycrystal is determined by global minimization of the *misfit energy* $W(U,\boldsymbol{\varepsilon}^{\mathrm{T}})$ of the polycrystal over all admissible configurations [9]

$$\mathcal{W}_n(U,\boldsymbol{\theta}) = \min_{\boldsymbol{\varepsilon}^{\mathrm{T}}\in\tilde{\Omega}_n|_\theta} W(U,\boldsymbol{\varepsilon}^{\mathrm{T}}). \tag{2.17}$$

The (possible nonunique) array(s) of transformations strains that arise in the polycrystal is (are) the minimizer(s) in (2.17).

Simplifying assumptions of our model. The idea that the energy minimization in composites and polycrystals can explain correlations has been long pursued in material science (see, e.g., [16, 22]). In this paper we use a quadratic form of the polycrystal's energy proposed in [9] and further developed analytically in [20, 6] and numerically in [6, 7, 8].

The probabilistic model introduced in this work captures many of the essential features of the general physical phenomenon of misfit and, at the same time, is amenable to rigorous analytical treatment.

Clearly, however, our model is too simple to reflect the rich phenomena that occur in actual three-dimensional polycrystals. For example, we consider isotropic elasticity, whereas typically, the crystalline lattice of each of the martensite variants has less symmetry than that of the austenite. Further, for sufficiently large grains, the lattices associated with the various martensite variants could be combined, giving rise to *twins* and/or higher-rank laminates of two or more different variants of martensite within each grain [21, 4, 15, 5, 17]—an effect that our model does not allow. We also note that, in general, a stress-free transformation is a time-dependent process that involves energy dissipation. Our study assumes that the final state of a polycrystal is determined by minimizers of a time-independent, dissipation-free misfit energy (see, e.g., [9, 6, 8] and references therein). Importantly, however, we do not assume that the grains in the polycrystal transform without elastic stresses (self-accommodation); see, e.g., [3, 2] and references therein.

Although not explicitly considered in this work, related phenomena, including electrical and magnetic polarizations in electro- and magneto-rheological materials and the combined elastic and magnetic-electric misfits arising from magnetostriction and electrostriction in composite materials, could be treated by similar methods.

3. Elasticity kernel. In this section we give a representation for the elastic energy $W(U,\boldsymbol{\varepsilon}^{\mathrm{T}})$ in terms of a certain integral kernel $K_L((x-t))$, and we then present

asymptotics of this kernel under two regimes that are relevant in our studies of the statistics of transformation strains in sections 5, 6, and 7. Denote spatial averages as

$$\langle g\rangle = \frac{1}{|\Pi_L|}\int_{\Pi_L} g(x,y)dxdy = \frac{1}{2L}\int_{\Pi_L} g(x,y)dxdy. \tag{3.1}$$

It turns out that the most convenient mathematical formulation of the elastic energy is in terms of

$$s(x) = \varepsilon^{\mathrm{T}}_{22}(x), \tag{3.2}$$

and the volume fraction of grains that undergo a phase transition[2]

$$f = \langle I\rangle,\ I = \varepsilon^{\mathrm{T}}_{11} + \varepsilon^{\mathrm{T}}_{22}. \tag{3.3}$$

Then the elastic energy

$$W(U,\boldsymbol{\varepsilon}^{\mathrm{T}}) = \langle (U-f+s)^2\rangle - \int_{-1}^{1}\int_{-1}^{1}(s(x)-\langle s\rangle)K_L((x-t))(s(t)-\langle s\rangle)dxdt, \tag{3.4}$$

where K_L is an even, 2-periodic integral kernel whose cosine Fourier coefficients

$$\hat{K}_L(m) = \int_{-1}^{1} K_L(x)\cos(\pi m x)dx \tag{3.5}$$

are explicitly given in Appendix A.1 by formulas (A.5) and (A.6).

The idea of the proof of (3.4) is to decompose the solution $\boldsymbol{u} = (u_1, u_2)$ of the boundary value problem (2.11), (2.13), and (2.14) in the form $\boldsymbol{u} = \tilde{\boldsymbol{u}} + \bar{\boldsymbol{u}}$, where $\tilde{\boldsymbol{u}}$ solves the elasticity equations (2.15) for infinitely long grains ($L = \infty$) and the remainder $\bar{\boldsymbol{u}}$ is the correction for finite L. It turns out that $\tilde{\boldsymbol{u}}$ is a piecewise linear function of the form

$$\begin{cases} \tilde{u}^i_1 = a_i x + f_i, \\ \tilde{u}^i_2 = c_i x + dy + g_i, \end{cases} \tag{3.6}$$

and $\bar{\boldsymbol{u}}$ satisfies the boundary value problem

$$\begin{cases} \partial_j \bar{\sigma}_{ij} = 0 \text{ for } i = 1,2, \\ \bar{u}_1(0,y) = \bar{u}_1(1,y) = 0, \quad \partial_1\bar{u}_2(0,y) = \partial_1\bar{u}_2(1,y) = 0, \\ \bar{\sigma}_{12}(x,\pm L) = 0, \quad \bar{\sigma}_{22}(x,\pm L) = E_c(\varepsilon^{\mathrm{T}}_{22} - \langle\varepsilon^{\mathrm{T}}_{22}\rangle). \end{cases} \tag{3.7}$$

where

$$\begin{cases} \bar{\sigma}_{ij} = c_{ijkl}\bar{\varepsilon}_{kl}, \\ \bar{\varepsilon}_{11} = \partial_1\bar{u}_1, \quad \bar{\varepsilon}_{12} = \bar{\varepsilon}_{21} = (\partial_2\bar{u}_1 + \partial_1\bar{u}_2)/2, \quad \bar{\varepsilon}_{22} = \partial_2\bar{u}_2. \end{cases}$$

Both functions can be computed explicitly: $\tilde{\boldsymbol{u}}$ is obtained from direct computations and $\bar{\boldsymbol{u}}$ is found by the Airy function method. A detailed proof is provided in Appendix A.1. The asymptotics of the convolution kernel in two important limiting

[2]Since the transformation strains defined in (2.4) satisfy $\varepsilon^1_{11} + \varepsilon^1_{22} = 1$, $\varepsilon^2_{11} + \varepsilon^2_{22} = 1$, and $\varepsilon^0_{11} + \varepsilon^0_{22} = 0$, it follows that the quantity f equals the volume fraction of grains that undergo a phase transition.

cases, in turn, are summarized in the following two lemmas. The proofs of the lemmas are provided in Appendices A.2 and A.3, respectively.

LEMMA 3.1 (kernel asymptotics as $L \to \infty$). *$K_L(x)$ can be represented in the form*

$$K_L(x) = \frac{B}{L} K_\infty(x) + O(\exp(-L)), \text{ as } L \to \infty, \tag{3.8}$$

where

$$K_\infty = -\ln|\sin(\pi x/2)|,$$

and

$$B = \frac{5\lambda + 9G}{4\pi(\lambda + 2G)} > 0.$$

We now consider polycrystals for which the height is commensurate with the grain widths $L = l_0/(2n)$. The parameter l_0 is the height-to-width ratio. In particular, when $L = 1/(2n)$, such polycrystals can be viewed as chains of square grains.

LEMMA 3.2 (kernel asymptotics as $L \to 0$). *Suppose $L = l_0/(2n)$, where $l_0 > 0$, and $n \to \infty$. Then for each fixed height-to-width ratio l_0 there exists a positive-definite function $\mathcal{K}_{l_0}(x)$, $x \in \mathbb{R}$ independent of n, such that*

$$||K_L(x) - n\mathcal{K}_{l_0}(nx)||_{L_\infty([-1,1])} \to 0, \text{ as } n \to \infty, \tag{3.9}$$

where $\mathcal{K}_{l_0}(x)$ is even: $\mathcal{K}_{l_0}(-x) = \mathcal{K}_{l_0}(x)$, and

$$|\mathcal{K}_{l_0}(x)| \le \frac{c}{l_0} \exp(-|x|/l_0). \tag{3.10}$$

The constant c in (3.10) *does not depend on n and l_0.*

We will find it useful, especially for chains of rectangular grains, to identify sequences (vectors) $(f_1, \dots . f_n)$ (of real number, matrices, etc.) with the corresponding piecewise constant (real valued, matrix valued, etc.) functions defined in the interval $[0, 1]$ that take the values $f(x) = f_k$ for $x \in \Pi_L^k$, $k = 1, \dots, n$. For example, the argument $\boldsymbol{\varepsilon}^{\mathrm{T}}$ of W in (2.15), which is a matrix-valued function defined in the interval $[0, 1]$, will often be replaced by a sequence of n matrices $(\boldsymbol{\varepsilon}_1^{\mathrm{T}}, \boldsymbol{\varepsilon}_2^{\mathrm{T}}, \boldsymbol{\varepsilon}_3^{\mathrm{T}}, \dots, \boldsymbol{\varepsilon}_n^{\mathrm{T}})$. As another example, note that the dependence on n of the quantity on the left-hand side of (2.17) arises merely from the fact that ε^{T} on the right-hand side of that formula is a piecewise constant function determined by a sequence of n matrices. For a function f defined by a sequence $(f_1, \dots . f_n)$ the spatial average (3.1) is

$$\langle f \rangle = \frac{1}{n} \sum_{i=1}^{n} f_i.$$

Further, on the space of the piecewise constant functions the integral representation of the misfit energy (3.4) can be viewed as an algebraic nonnegative definite quadratic form:

$$W_n(U, \boldsymbol{\varepsilon}^{\mathrm{T}}) = \langle (U - f + s)^2 \rangle - \langle (s - \langle s \rangle) M(n, L)(s - \langle s \rangle) \rangle, \tag{3.11}$$

where $M = M(n, L)$ is a $n \times n$ symmetric Toeplitz matrix with entries $M_{ij}(n, L) = \lambda_{i-j}(n, L)$ defined by

$$(s - \langle s \rangle) M (s - \langle s \rangle) = \int_{-1}^{1} \int_{-1}^{1} (s(x) - \langle s \rangle) K_L(x - t)(s(t) - \langle s \rangle) dx dt \tag{3.12}$$

for piecewise constant functions $s = (s_1, \dots, s_n)$.

As an immediate consequence of Lemma 3.2 we can estimate the decay of the coefficients $\lambda_{i-j}(n, L)$ for a chain of rectangular grains ($L = l_0/(2n)$): As $n \to \infty$, $\lambda_k(n, L) \to \lambda_k(l_0)$ where

$$|\lambda_k(l_0)| \le \frac{c}{n} \exp(-|k|/l_0). \tag{3.13}$$

By (3.13) for a chain of rectangular grains $\lambda_{i-j}(n, L)$ can be accurately approximated by a (truncated) Toeplitz matrix by setting $\lambda_k(n, L) \equiv 0$, for $|k| > k_0$, and the misfit energy is approximately

$$W_n^{k_0}(U, \boldsymbol{\varepsilon}^{\mathrm{T}}) = \langle (U - f + s)^2 \rangle - \sum_{k=-k_0}^{k_0} \sum_{i=1}^{n} \lambda_k(n)(s_i - \langle s \rangle)(s_{i+k} - \langle s \rangle), \tag{3.14}$$

where we define s_{i+k} for $i + k > n$ by periodicity

$$s_{i+k} = \begin{cases} s_{i+k}, \text{ if } i + k \le n, \\ s_{i+k-n}, \text{ if } i + k > n. \end{cases}$$

The approximation (3.14) provides a justification, in a one-dimensional context, of numerical schemes which are used in practical evaluation of the misfit energy [9, 6, 7, 8]. The approximation (3.14) takes into account only interaction with the nearest neighbors. Hence we call (3.14) k_0-nearest neighbors energy. The next proposition shows that for any finite value of n and finite k_0, this k_0-nearest neighbors approximation has an exponential in k_0 error. Therefore the computational complexity of finding the misfit energy can be significantly reduced if (3.11) is replaced by (3.14). In [6, 7, 8] this truncation was implemented for general two- and three-dimensional polycrystals, and the convergence was verified numerically. The following proposition justifies this convergence analytically in the case of chains of rectangular grains, and provides an explicit exponential error estimate.

PROPOSITION 3.3. *For a given U and a given vector of orientation angles $\boldsymbol{\theta}$, suppose $\boldsymbol{\varepsilon}^T(k_0) \in \tilde{\Omega}_n(\boldsymbol{\theta})$ is a minimizer of the k_0-nearest neighbors energy $W_n^{k_0}(U, \boldsymbol{\varepsilon}^T)$ given by* (3.14). *Then there is a universal constant c, independent of n, such that $W_n(U, \boldsymbol{\theta})$, the minimum of the misfit energy* (3.11), *satisfies*

$$|W_n(U, \boldsymbol{\theta}) - W_n^{k_0}(U, \boldsymbol{\varepsilon}^T(k_0))| \le c \exp(-k_0). \tag{3.15}$$

A proof the theorem is given in Appendix A.4. Finally, applying Proposition 3.3 and Lemma 3.2, the misfit energy (3.4) of a chain of short grains becomes

$$\begin{aligned} W_n(U, \boldsymbol{\varepsilon}^{\mathrm{T}}) = \langle (U - f + s)^2 \rangle - \lambda_0 \langle (s - \langle s \rangle)^2 \rangle - \frac{B e^{-1/l_0}}{n} \sum_{i=1}^{n} \\ (s_i - \langle s \rangle)(s_{i+1} - \langle s \rangle) + O(e^{-2/l_0}), \text{ as } l_0 \to 0, \end{aligned} \tag{3.16}$$

where $\lambda_0 > 0$, $B > 0$.[3] Thus, when $l_0 \ll 1$, the misfit energy is approximated by the *nearest neighbor* energy.

[3]From numerical computations λ_k, $k \ge 2$ are negligible even for large $l_0 = 1$: $\sum_{k=2}^{\infty} |\lambda_k| \le .1\lambda_1$, $\lambda_1 \le .17\lambda_0$.

4. A probabilistic model. Our probabilistic model is set to describe energy minimizers within a random setting. In detail, we consider orientation angles θ_i, $i = 1, 2, \dots$ as a sequence of random variables[4]

$$\theta_i : \Omega \to \Omega^\theta = [0, \pi/4]$$

on a common probability space $(\Omega, \mathcal{F}, P)$. We denote by ρ_n the usual induced probability measure of a sequence of the first n orientation angles on $(\Omega_n^\theta, \sigma_n^\theta)$, where $\Omega_n^\theta = \left(\Omega^\theta\right)^n = \Omega^\theta \times \Omega^\theta \times \cdots \times \Omega^\theta$, and σ_n^θ is the Borel σ-algebra on Ω_n^θ.

We may define and work with the probability space $(\tilde{\Omega}_n, \tilde{\sigma}_n, \tilde{\mu}_n)$ of arrays of transformation strains $(\boldsymbol{\varepsilon}_i^{\mathrm{T}})_{i=1}^n$, where we recall that $\tilde{\Omega}_n$ is the set of all sequences of strains that are admissible for some sequence of orientation angles (see (2.5)).

We will, however, work with a different, but equivalent probability space. Recall that the misfit energy (3.4) depends only on a sequence of pairs $\{(s_i, I_i)\}$, $i = 1, \dots, n$, where

$$I_i = \varepsilon_{11,i}^{\mathrm{T}} + \varepsilon_{22,i}^{\mathrm{T}} \in \{0, 1\}, \text{ and } s_i = \varepsilon_{22,i}^{\mathrm{T}} \in [0, 1].$$

Therefore, we will study the probability space

$$\left(\Omega_n^{\mathrm{T}}, \sigma_n^{\mathrm{T}}, \mu_n\right), \ \Omega_n^{\mathrm{T}} = \left(\Omega^{\mathrm{T}}\right)^n = \Omega^{\mathrm{T}} \times \Omega^{\mathrm{T}} \times \cdots \times \Omega^{\mathrm{T}}, \ (s_i, I_i) \in \Omega^{\mathrm{T}} = [0, 1] \times \{0, 1\},$$

where σ_n^{T} is the Borel σ-algebra on Ω_n^{T} and μ_n is a probability measure, which we will define next.

Probability measure in the space of transformation strains. Suppose the applied deformation U is given. For a fixed sequence of orientation angles $\boldsymbol{\theta}$, there are up to 3^n different admissible arrays of transformation strains $(\boldsymbol{\varepsilon}_i^{\mathrm{T}})_{i=1}^n$, $\mathrm{T} = 0, 1, 2$; see (2.6) with (2.3) and (2.4). They correspond to 3^n different arrays of pairs $\{(s_i, I_i)\}$, $i = 1, \dots, n$ using the rule that the matrices $\boldsymbol{\varepsilon}^1(\theta_i), \boldsymbol{\varepsilon}^2(\theta_i)$, and $\boldsymbol{\varepsilon}^0(\theta_i)$ in (2.4) correspond to the pairs $(s_i, I_i) = (\sin^2\theta_i, 1)$, $(s_i, I_i) = (\cos^2\theta_i, 1)$, and $(s_i, I_i) = (0, 0)$, respectively. Some of these arrays $\{(s_i, I_i)\}$, $i = 1, \dots, n$, say a number k of them, minimize the misfit energy (3.4) amongst all admissible arrays. We assume, as it may indeed be natural from a physics perspective, that each of these energy minimizing arrays occurs with equal probability. In other words, we will define the probability measure μ_n in such a way that the conditional probability measure $\mu_n((s_i, I_i)_{i=1}^n | \boldsymbol{\theta})$ satisfies

$$\mu_n((s_i, I_i)_{i=1}^n | \boldsymbol{\theta}) = \begin{cases} 1/k, & \text{if } \boldsymbol{\varepsilon}^{\mathrm{T}} \text{ is a minimizer} \\ 0, & \text{if } \boldsymbol{\varepsilon}^{\mathrm{T}} \text{ is not a minimizer,} \end{cases} \tag{4.1}$$

where k is the number of minimizers of the misfit energy (3.4) for the fixed $\boldsymbol{\theta}$. The probability measure μ_n on the sequences $(s_i, I_i)_{i=1}^n$ is thus defined by

$$\mu_n(A) = \int_{\Omega_n^\theta} \mu_n(A | \boldsymbol{\theta}) \mathrm{d}\rho_n(\boldsymbol{\theta}) \tag{4.2}$$

for any Borel set $A \in \sigma_n^{\mathrm{T}}$.

[4]In principle $\theta \in [0, \pi/2]$. It follows from (2.4) that we are concerned only with $\sin^2\theta$ and $\cos^2\theta$. Since $\sin(\pi/2 - \theta) = \cos\theta$ and $\cos(\pi/2 - \theta) = \sin\theta$, we may and do assume that $\theta \in [0, \pi/4]$. The orientation of a square crystalline lattice can be described uniquely by a value $\theta \in [0, \pi/4]$.

The measure μ_n describes statistics of the transformation strains; the main objective of this paper is to describe it in detail for $n \gg 1$ and for various polycrystal configurations.

Distribution of angles. In the remainder of this paper we will assume that the angles θ_i are independent and identically distributed with the induced probability measure (distribution) ρ on $\Omega^\theta = [0, \pi/4]$—although other types of angle distributions could be considered within the present context. In other words, the probability measure ρ_n on the sequences of the orientation angles will be taken to be a product measure of the form

$$\rho_n = \Pi_{i=1}^n \rho.$$

In particular, to illustrate our theory we will consider two specific probability distributions ρ: (1) the *uniform distribution*, in which ρ is proportional to the Lebesgue measure, and (2) the *Bernoulli trials model* for which ρ is concentrated in two θ values

$$\theta_i = \begin{cases} \alpha, & \text{with probability } q, \\ \beta, & \text{with probability } 1-q, \end{cases} \tag{4.3}$$

$0 \le \beta \le \alpha \le \pi/4$.

5. Statistics of asymptotic energy minimizers 1: Infinitely long grains ($L = \infty$).

5.1. The main theorem. Suppose the grains are infinitely long ($L = \infty$). Then, by Lemma 3.1 the misfit energy for a given admissible sequence of transformation strains on the array of n grains is given by

$$W_n(U, \boldsymbol{\varepsilon}^{\mathrm{T}}) = \langle \left(U - f + \varepsilon_{22}^{\mathrm{T}}\right)^2 \rangle = \langle (U - f + s)^2 \rangle. \tag{5.1}$$

The sequence of measures (4.2) contains convergent subsequences [18]; each such limit $\mu_{\lim}$ is a measure on the set of infinite sequences of transformation strains; the limits along various subsequences may, in principle, not all coincide. In fact, in all cases we consider, however, all such limits do coincide, and the full sequences (4.2) are convergent. For the sake of simplicity, in the subsequent analysis we assume this is the case and we denote $\mu_{\lim} = \lim_{n\to\infty} \mu_n$.

As we shall show the limits $\mu_{\lim}$ are convex combinations of product measures. This is a consequence of the de Finetti's representation theorem (see [13] for a general version of this theorem). In order to motivate the advantages of this observation in our context, we first consider one such limit $\mu_{\lim}$ and we *assume* (this may or may not be true!) that (1) for each finite n the minimizers are unique, and (2) the measure $\mu_{\lim}$ is given by a product of the form

$$\mu_{\lim} = \tilde{\mu} := \Pi_{i=1}^\infty \mu, \tag{5.2}$$

for a certain measure μ so that, according to $\mu_{\lim}$, the random variables $\{(s_i, I_i)\}_{i=1}^\infty$ are i.i.d.

Since, as we have seen above, for each i we must necessarily have $s_i = \sin^2(\theta_i)$ or $s_i = \cos^2(\theta_i)$ whenever $I_i = 1$, under the uniqueness assumption (1) above the measure μ in (5.2) must satisfy

$$\begin{cases} \mu\left(s = \sin^2(\theta) | \theta, I = 1\right) + \mu\left(s(\theta) = \cos^2(\theta) | \theta, I = 1\right) = 1 \\ \mu\left(s = \sin^2(\theta) | \theta, I = 1\right) = 0 \quad \text{or} \quad 1 \quad \text{and} \quad \mu\left(s(\theta) = \cos^2(\theta) | \theta, I = 1\right) = 0 \quad \text{or} \quad 1. \end{cases}$$

Hence, we can define a function $\kappa(\theta)$, $\kappa(\theta) = 0$, or $\kappa(\theta) = 1$ such that

$$\kappa(\theta) = \begin{cases} 1, & \text{if } \mu\left(s = \sin^2(\theta)|\theta, I = 1\right) = 1, \\ 0, & \text{if } \mu\left(s = \cos^2(\theta)|\theta, I = 1\right) = 1. \end{cases} \tag{5.3}$$

By the law of large numbers, as $n \to \infty$ we have for the misfit energy (5.1)

$$W_n(U, \boldsymbol{\varepsilon}^{\mathrm{T}}) \to \int_0^{\pi/4} g(\theta)\chi(\theta)d\rho(\theta) + (U - f)^2(1 - f), \tag{5.4}$$

where $g(\theta) = (U - f + \sin^2\theta)^2\kappa(\theta) + (U - f + \cos^2\theta)^2(1 - \kappa(\theta))$ and where we have set

$$\chi(\theta) = \mu(I = 1|\theta), \ (0 \le \chi(\theta) \le 1), \tag{5.5}$$

and $f = \int_0^{\pi/4} \chi(\theta)d\rho(\theta)$.

Clearly, in the present context the limiting values of the energy function $W_n(U, \boldsymbol{\varepsilon}^{\mathrm{T}})$ are determined uniquely by the functions $\kappa(\theta)$ and $\chi(\theta)$. Since $\mu_{\lim}$ is the limit of probability measures $\{\mu_n\}$ given by (4.1) with (5.1), it follows that the measure μ in (5.2) must minimize (5.4). In other words, under the assumption (5.2), the overall minimization problem has been reduced to the following minimization problem for the functions $\kappa(\theta)$ and $\chi(\theta)$:

$$\begin{aligned} &\min_{\kappa(\theta),\chi(\theta)} \int_0^{\pi/4} g(\theta)\chi(\theta)d\rho(\theta) + (U - f)^2(1 - f) \\ &g(\theta) = (U - f + \sin^2\theta)^2\kappa(\theta) + (U - f + \cos^2\theta)^2(1 - \kappa(\theta)), \\ &f = \int_0^{\pi/4} \chi(\theta)d\rho(\theta), \ 0 \le \chi(\theta) \le 1, \ 0 \le \kappa(\theta) \le 1. \end{aligned} \tag{5.6}$$

One can anticipate that, generally, the assumption (5.2) does not hold. Indeed, even working under the assumption (5.2), we note that a solution μ (i.e., (κ, χ)) to the minimization problem (5.6) may not be unique. If there are two such solutions μ_1 and μ_2 to this problem, then the corresponding infinite products $\tilde{\mu}_1$ and $\tilde{\mu}_2$ could, conceivably, equal to the limit of a subsequence of μ_n. As shown in Theorem 5.2, however, in general $\mu_{\lim}$ will equal a convex combination of such infinite products. The following definition will be useful in these regards.

DEFINITION 5.1. *Consider the set $\mathbb{S}$ of all measures on the set of pairs $(s, I) \in [0, 1] \times \{0, 1\}$. For each measure $\mu \in \mathbb{S}$ define the associated product measure*

$$\tilde{\mu} = \Pi_{i=1}^{\infty}\mu$$

on $\Pi_{i=1}^{\infty}(s_i, I_i)$. A measure γ is called a convex combination of such product measures, if there exists a positive measure $\nu(\mu)$ on the set $\mathbb{S}$, such that

$$\gamma(A) = \int_{\mathbb{S}} \tilde{\mu}(A)d\nu(\mu), \ \int_{\mathbb{S}} d\nu(\mu) = 1. \tag{5.7}$$

THEOREM 5.2. *Consider infinitely long grains and an arbitrary i.i.d. angle distribution with the probability measure ρ, and assume the limit $\mu_{\lim}$ of the sequence*

μ_n defined by (4.2) *and* (4.1) *exists. Then $\mu_{\lim}$ is given by a convex combination of product measures arising from minimization problems. In detail, we have*

$$\mu_{\lim} = \int_{\mathbb{S}} \tilde{\mu} d\nu(\mu), \ \int_{\mathbb{S}} d\nu(\mu) = 1,$$

where $\mathbb{S}$ is the set of product measures: $\tilde{\mu} = \Pi_{i=1}^{\infty}\mu$, and each μ is defined by

$$\mu((s,I)|\theta) = \begin{cases} \chi(\theta)\kappa(\theta), & \text{if } I = 1, \\ 1 - \chi(\theta), & \text{if } I = 0, \end{cases}$$

where $\kappa(\theta)$ and $\chi(\theta)$ are minimizers of

$$\min_{\kappa(\theta),\chi(\theta)} \int_0^{\pi/4} (U - f + s(\theta))^2 \chi(\theta) d\rho(\theta) + (U-f)^2(1-f), \tag{5.8}$$
$$f = \int_0^{\pi/4} \chi(\theta) d\rho(\theta), \ \chi : [0, \pi/4] \to [0,1],$$

$$s(\theta) = \kappa(\theta)\sin^2(\theta) + (1 - \kappa(\theta))\cos^2(\theta), \quad \text{where } \kappa : [0,\pi/4] \to \{0,1\}. \tag{5.9}$$

If the minimizer of (5.8) *is unique, then transformation strains in different grains are independent identically distributed; that is, $\mu_{\lim}$ is a product measure $\mu_{\lim} = \Pi_{i=1}^{\infty}\mu$.*

Proof. A key property of the energy W_n (5.1) in the case of infinitely long grains is that it is *invariant under permutations*; e.g., for a three-grain polycrystal, if $(s_1, I_1), (s_2, I_2), (s_3, I_3)$ is a minimizing sequence for the angles $(\theta_1, \theta_2, \theta_3)$, then $(s_2, I_2), (s_1, I_1), (s_3, I_3)$ is a minimizing sequence for the angles $(\theta_2, \theta_1, \theta_3)$ with the same probability. More generally, the form of the misfit energy and our assumption (4.1) imply that the probability measure μ_n (defined by (4.1), (4.2)) on the minimizers must be symmetric;[5] that is, for any finite permutation $\tau \in S(n)$

$$\mu_n((s_1,I_1) \in A_1 \ldots, (s_n,I_n) \in A_n) = \mu_n((s_{\tau(1)}, I_{\tau(1)}) \in A_1, \ldots, (s_{\tau(n)}, I_{\tau(n)}) \in A_n).$$

As $n \to \infty$, the probability measure μ_n converges to a certain $\mu_{\lim}$. Clearly, $\mu_{\lim}$ must be symmetric as well: For any n and $\tau \in S(n)$

$$\begin{aligned} &\mu_{\lim}\left((s_1,I_1) \in A_1, \ldots, (s_n,I_n) \in A_n, (s_{n+1}, I_{n+1}) \in A_{n+1}, \ldots\right) \\ &= \mu_{\lim}\left((s_{\tau(1)}, I_{\tau(1)}) \in A_1, \ldots, (s_{\tau(n)}, I_{\tau(n)}) \in A_n, (s_{n+1}, I_{n+1}) \in A_{n+1}, \ldots\right). \end{aligned}$$

Hence, we can apply the de Finetti's representation theorem [13], and $\mu_{\lim}$ must be a convex combination of product measures $\tilde{\mu}$:

$$\mu_{\lim} = \int_{\mathbb{S}} \tilde{\mu} d\nu(\mu), \ \int_{\mathbb{S}} d\nu(\mu) = 1, \ \nu(\mu) \geq 0;$$

see Definition 5.1, as claimed.

Let us now show that

$$\begin{aligned} &\lim_{n\to\infty} \int_{\Omega_n^{\mathrm{T}}} W_n(U, \boldsymbol{\varepsilon}^{\mathrm{T}}) d\mu_n \\ &\qquad = \int_{\mathbb{S}} \left[\int_0^{\pi/4} (U - f + s(\theta))^2 \chi(\theta) d\rho(\theta) + (U-f)^2(1-f) \right] d\nu(\mu). \end{aligned} \tag{5.10}$$

[5] Sometimes the term exchangeable is used instead of symmetric.

The key issue here is classical: Given that $\mu_n \to \mu_{\lim}$ *weakly*, we cannot, in general, conclude convergence of $\int W_n d\mu_n$. In our case, however, we can, because the measures μ_n are symmetric. It follows that for an n-grain sample, the functions $\{(s_i, I_i)\}_{i=1}^n$ satisfy, for any i, j, $1 \le i, j \le n$,

$$\int I_i I_j d\mu_n = \begin{cases} \int I_1 I_2 d\mu_n, & i \neq j, \\ \int I_1^2 d\mu_n, & i = j, \end{cases}$$

and similar equalities hold for $s_i I_j$, and $(U - \langle I \rangle + s_i)^2$. Therefore, W_n can be written as

$$\int W_n(U, \boldsymbol{\varepsilon}^{\mathrm{T}}) d\mu_n = \int F(U, s_1, I_1, I_2) d\mu_n + \frac{1}{n} \int G(s_1, I_1, I_2) d\mu_n,$$

where *both* functions $F(U, s_1, I_1, I_2)$ and $G(s_1, I_1, I_2)$ depend *continuously* (they are quadratic polynomials) on the values of s_i and I_i only in two grains $i = 1, 2$ (and, thus, do not depend on n), and they are explicitly given as

$$F = (U + s_1)^2 - 2UI_1 - 2s_1 I_2 + I_1 I_2, \quad G = -2s_1 I_1 + 2s_1 I_2 + I_1^2 - I_1 I_2.$$

Therefore

$$\begin{aligned} \lim_{n\to\infty} \int_{\Omega_n^{\mathrm{T}}} W_n(U, \boldsymbol{\varepsilon}^{\mathrm{T}}) d\mu_n &= \lim_{n\to\infty} \int_{\Omega_n^{\mathrm{T}}} F(U, s_1, I_1, I_2) d\mu_n + \lim_{n\to\infty} \frac{1}{n} \int_{\Omega_n^{\mathrm{T}}} G(s_1, I_1, I_2) d\mu_n \\ &= \int_{\Omega_\infty^{\mathrm{T}}} F(U, s_1, I_1, I_2) d\mu_{\lim}. \end{aligned}$$

Hence it only remains to show that

$$\begin{aligned} &\int_{\Omega_\infty^{\mathrm{T}}} F(U, s_1, I_1, I_2) d\mu_{\lim} \\ &\quad = \int_{\mathbb{S}} \left[\int_0^{\pi/4} (U - f + s(\theta))^2 \chi(\theta) d\rho(\theta) + (U - f)^2 (1 - f) \right] d\nu(\mu). \end{aligned} \tag{5.11}$$

The last equality is obtained by explicit computations provided in Appendix B.1. The proof of the identity (5.10) is now complete.

Further, up to a set of ν-measure zero, each μ must minimize (5.8). Otherwise, we can choose a $\delta > 0$ such that the set

$$A = \left\{ \mu : \int F(U, s_1, I_1, I_2) d\tilde{\mu} - \delta > \min \int F(U, s_1, I_1, I_2) d\tilde{\mu} \right\}$$

has positive measure: $\nu(A) > 0$. Then, if $\nu(\mathbb{S} \setminus A) \neq 0$ we define a new measure $\tilde{\nu}$ by

$$\tilde{\nu}(B) = \nu(B \setminus A)/\nu(\mathbb{S} \setminus A).$$

Clearly

$$\int_{\mathbb{S}} \left[\int F(U, s_1, I_1, I_2) d\tilde{\mu} \right] d\tilde{\nu}(\mu) < \int_{\mathbb{S}} \left[\int F(U, s_1, I_1, I_2) d\tilde{\mu} \right] d\nu(\mu),$$

which contradicts the assumption that ν yields a limit of minimizers of the misfit energy (5.8) as indicated in (5.11). If $\nu(\mathbb{S} \setminus A) \neq 0$, in turn, we can select a single minimizer and assign $\tilde{\nu}$ measure 1 to it, arriving again to a contradiction.

To establish (5.9), note that for a minimizer μ of (5.6), the μ probabilities conditional to a given angle θ and to $I = 1$, satisfy

$$\mu\left(s(\theta) = \cos^2(\theta)|\theta, I = 1\right) = 0, \text{ or } \mu\left(s(\theta) = \sin^2(\theta)|\theta, I = 1\right) = 0.$$

Hence $\kappa(\theta)$ takes only the values 0 and 1 and (5.9) holds.

Finally, suppose (5.8) admits a unique minimizer μ. Since, as established above, the limit $\mu_{\lim}$ must be a convex combination of product measures that minimize (5.8); in the case of uniqueness of solution to (5.8), $\mu_{\lim}$ must be the product measure $\mu_{\lim} = \Pi_{i=1}^{\infty}\mu$, as claimed. □

A few remarks about Theorem 5.2 are in order. Firstly, we are aware of some examples when the minimization problem (5.8) has more than one solution. One of these examples is to consider a deterministic sequence $\theta_i = 0$ and $U = 1$. Then there are two solutions $\kappa_1 \equiv 0$, $\chi_1 \equiv 1$, and $\kappa_2 \equiv 0$, $\chi_2 \equiv 0$ to the minimization problem (5.8), which give rise to corresponding measures μ_1 and μ_2, and, thus, product measures $\tilde{\mu}_1$ and $\tilde{\mu}_2$. For both $i = 1, 2$ we have

$$\int_0^{\pi/4} (U - f_i + s_i(\theta))^2 \chi_i(\theta) d\rho(\theta) + (U - f_i)^2(1 - f_i) = 1.$$

In view of our symmetrization assumption 4.1, the limit of μ_n exists and it is equal to a convex combination of product measures as implied by Theorem (5.2); the convex combination is given by $\tilde{\mu}_1/2 + \tilde{\mu}_2/2$. We expect that generically the minimization problem (5.8) has a unique solution. We give two explicit examples when this measure is unique: Bernoulli trials, Lemma 5.3, and Uniform distribution, Lemma 5.6, in section 5.2.

Secondly, in principle, the (unique) solution to the minimization problem (5.8) may be such that $\chi(\theta)$ takes only two values 0 or 1, i.e., $\chi : [0, \pi/4] \to \{0, 1\}$. This, indeed, happens for uniform distribution (Lemma 5.6 in section 5.2). If we know that $\chi : [0, \pi/4] \to \{0, 1\}$, then the proof of Theorem 5.2 becomes straightforward. However, there are examples, when $0 < \chi(\theta) < 1$ for some θ, and one of them is the Bernoulli trials (Corollary 5.4 in section 5.2).

Finally, if we do not assume the uniform conditional probability (4.1), then $\mu_{\lim}$ may not be unique even if the minimizer of (5.8) is unique. We discuss this issue for Bernoulli trials after the proof of Corollary 5.4 below.

Motivated by the above remarks, we next investigate in more detail how measure μ, a solution to the minimization problem (5.8), depends on the underlying probability measure ρ for two specific probability measures ρ: the Bernoulli trials model and the uniform distribution of θ.

5.2. Bernoulli trials and uniform distribution. For the Bernoulli trials model (4.3), the minimization problem (5.8) from Theorem 5.2 is

$$\begin{aligned} &\min_{q_\alpha, q_\beta, s(\alpha), s(\beta)} W(q_\alpha, q_\beta, s(\alpha), s(\beta)), \\ &W = (U - f + s(\alpha))^2 q_\alpha + (U - f + s(\beta))^2 q_\beta + (U - f)^2(1 - f), \end{aligned} \tag{5.12}$$

with $f = q_\alpha + q_\beta$,

$$0 \le q_\alpha \le q, \ 0 \le q_\beta \le 1 - q, \tag{5.13}$$

$0 \le \beta \le \alpha \le \pi/4$, and $s(\alpha)$, $s(\beta)$ are defined by $s(\theta) = \sin^2\theta$ or $s(\theta) = \cos^2\theta$. In particular, the minimization with respect to χ is reduced to determining the proportions q_α and q_β of grains with angles α and β that do not undergo a stress-free transformation.

LEMMA 5.3. *For the Bernoulli trials model* (4.3) *with* $0 < q < 1$ *the minimizer* (κ, χ) *of* (5.12) *is unique. For the minimizer* $\kappa(\theta) \equiv 1$, *i.e.,*

$$s(\theta) = \sin^2 \theta, \tag{5.14}$$

and χ *depends on* U *and can be described as follows. For a given* U *the total proportion of grains that undergoes a stress-free transformation* $f = f(U)$ *is a (deterministic) nondecreasing function of* U. *For a given* f *we have several cases*

- *if* $f < 1 - q$, *then*

$$q_\alpha = 0,\ 0 \le q_\beta \le 1 - q, \tag{5.15}$$

  i.e., $\chi(\alpha) = 0$, $\chi(\beta) = q_\beta/(1-q)$,
- *if* $f > 1 - q$, *then*

$$0 \le q_\alpha \le q,\ q_\beta = 1 - q, \tag{5.16}$$

  i.e. $\chi(\beta) = q_\alpha/q$, $\chi(\beta) = 1$.

A proof of the Lemma is in Appendix B.2.

COROLLARY 5.4. *The probability distribution*

$$\theta_i = \begin{cases} \pi/4, & \text{with probability } q,\ q \ge 1/2, \\ 0, & \text{with probability } 1 - q \end{cases}$$

is an example, where the minimizer (κ, χ) *of* (5.12) *is unique, but*

$$\chi : [0, \pi/4] \to K \subset [0,1], K \ne \{0,1\}.$$

Indeed, in this case

$$\begin{aligned} &\chi(0) = U, \chi(\pi/4) = 0, \text{ if } U \le 1 - q, \\ &\chi(0) = 1, \chi(\pi/4) = 0, \text{ if } 1 - q \le U \le 5/4 - q, \\ &\chi(0) = 1, \chi(\pi/4) = U - 5/4 + q, \text{ if } 5/4 - q \le U \le 3/4, \text{ or} \\ &\chi(0) = 1, \chi(\pi/4) = 1, \text{ if } 3/4 \le U. \end{aligned}$$

Let us now discuss our final remark that if we do not assume the uniform conditional probability (4.1), then $\mu_{\lim}$ may not be unique even if the minimizer of (5.12) is unique. It depends on whether χ takes more than two values, that is, on whether $\chi : [0, \pi/4] \to K$, but $K \ne \{0,1\}$. For example, consider the Bernoulli trials with $0 < q_\alpha < q$. Then one can choose the grains with $\theta_i = \alpha$, which do not undergo a stress-free transformation, arbitrarily, provided that their total proportion is q_α. Thus, if we remove our assumption of equal probability (4.1), in the case of infinitely long grains there are many minimizers of the misfit energy in addition to minimizers described in Theorem 5.2. Hence, it is possible to construct the limiting measure $\mu_{\lim}$ on the infinite sequence of pairs $\{(s_i, I_i)\}_{i=1}^\infty$ so that it is not a product measure. Moreover, actual construction of the exact minimizers $\boldsymbol{\varepsilon}^{\mathrm{T}}$ of the energy (5.1) (for a given sequence $\{\theta_i\}_{i=1}^n$) in practice [9, 6, 7, 8] is done numerically. Thus it typically results in finding an *almost* minimizer $\tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}}$, such that

$$|W_n(U, \boldsymbol{\varepsilon}^{\mathrm{T}}) - W_n(U, \tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}})| \le \delta,\ \delta > 0. \tag{5.17}$$

Thus, it is natural to ask which characteristic properties of exact minimizers are approximated by characteristic properties of almost minimizers. The property

that $\mu_{\lim}$ is a product measure is *not* characteristic, but the proportion of grains that undergo a stress-free transformation is such property. For example, for Bernoulli trials, $q_\alpha(\boldsymbol{\theta})$ characterizes the proportion of grains with $\theta_i = \alpha$, $i = 1, \dots, n$, which undergo a stress-free transformation (grains for which $I = 1$), and we have the following immediate result.

LEMMA 5.5. *For every $\delta' > 0$ there exists $\delta > 0$ such that an almost minimizer $\tilde{\boldsymbol{\varepsilon}}^T$ in the sense* (5.17) *satisfies*

$$|q_\alpha(\boldsymbol{\theta}) - \tilde{q}_\alpha(\boldsymbol{\theta})| < \delta', \ |q_\beta(\boldsymbol{\theta}) - \tilde{q}_\beta(\boldsymbol{\theta})| < \delta',$$

where $q_\alpha(\boldsymbol{\theta})$ and $\tilde{q}_\alpha(\boldsymbol{\theta})$ correspond to the exact and almost minimizers, respectively ($q_\beta(\boldsymbol{\theta})$ and $\tilde{q}_\beta(\boldsymbol{\theta})$ are defined analogously).

Analogous to the Bernoulli trials model, direct computations show the following result for the uniform distribution.

LEMMA 5.6. *For the uniform distribution of $\theta \in [0, \pi/4]$, the minimizer (κ, χ) of* (5.12) *is unique. For the minimizer $\kappa(\theta) \equiv 1$, hence*

$$s(\theta) = \sin^2 \theta, \tag{5.18}$$

and χ depends on U and can be described as follows. For a given U the total proportion of grains that undergoes a stress-free transformation $f(U)$ is a (deterministic) nondecreasing function of U given by

$$f(U) = \begin{cases} \frac{4\pi}{8+\pi} U + g(U), & \text{if } U \le \frac{1}{4} + \frac{2}{\pi} \approx .88662, \\ 1 & \text{otherwise}, \end{cases}$$

where the small correction $g(U)$ is concave and it satisfies $g(0) = f(1/4 + 2/\pi) = 0$, $-0.055 < g(U) \le 0$. For a given $f < 1$

$$\chi(\theta) = \begin{cases} 1, & \text{if } \sin^2 \theta \le 2(f - U), \\ 0, & \text{otherwise}. \end{cases}$$

If $f = 1$, then $\chi \equiv 1$.

Lemmas 5.3 and 5.6 together with Theorem 5.2 imply the following.

COROLLARY 5.7. *For Bernoulli trials and uniform distribution, the unique minimizing sequence of transformation strains is i.i.d, and, in particular, it is uncorrelated.*

6. Statistics of asymptotic energy minimizers 2: Thin long grains (finite $L \gg 1$).

6.1. Basic definitions and formulation of the main theorem. In contrast to the case of infinitely long grains, if L is large but finite, then each grain may undergo a stress-free transformation which, as we show in this section, is correlated to stress-free transformations of other grains. In particular, for Bernoulli trials in case 1 ($L = \infty$, $n \to \infty$) the minimizers are shown to be i.i.d. (see Corollary 5.7 in the previous section), whereas in case 2 ($n \to \infty$, followed by $L \to \infty$) the minimizers are no longer i.i.d. (see Theorem 6.4 below).[6]

By Lemma 3.1 the misfit energy for $L \gg 1$ has the following asymptotic representation (up to higher order terms):

$$W_n(U, \boldsymbol{\varepsilon}^{\mathrm{T}}) = \langle (U - f + s)^2 \rangle - \frac{B}{L} \bar{W}_n(U, \boldsymbol{\varepsilon}^{\mathrm{T}}), \text{ as } L \to \infty, \tag{6.1}$$

[6] In this sense, we prove that the limits for large n and large L do not commute.

where

$$\bar{W}_n(U, \boldsymbol{\varepsilon}^{\mathrm{T}}) = \int_{-1}^{1}\int_{-1}^{1} (s(x) - \langle s\rangle)(s(t) - \langle s\rangle) K_\infty(x-t)\,dx\,dt, \tag{6.2}$$

$$K_\infty = -\ln|\sin(\pi x/2)| > 0. \tag{6.3}$$

For $L \gg 1$ the second term on the right-hand side of (6.1) amounts to a small correction to the misfit energy of infinitely long grains (5.1) (the first term in (6.1)), a situation that bears connections with the concept of almost minimizers (5.17) introduced in the previous section. In the present context we have:

LEMMA 6.1. *For any $\delta > 0$, there exists a sufficiently large $L_0 > 0$, so that for any $L \geq L_0$*

$$s(\theta) = \sin^2\theta, \tag{6.4}$$

$$|q_\alpha(\boldsymbol{\theta}) - q_\alpha^L(\boldsymbol{\theta})| < \delta,\ |q_\beta(\boldsymbol{\theta}) - q_\beta^L(\boldsymbol{\theta})| < \delta, \tag{6.5}$$

where $q_\alpha(\boldsymbol{\theta})$ and $q_\alpha^L(\boldsymbol{\theta})$ correspond to minimizers of (5.1) *and* (6.1)*, respectively ($q_\beta(\boldsymbol{\theta})$ and $q_\beta^L(\boldsymbol{\theta})$ are defined analogously).*

A proof of Lemma 6.1 is in Appendix C.1. As we pointed out after Corollary 5.4, when $L = \infty$ and $0 < q_\alpha < q$ or $0 < q_\beta < 1 - q$ there is an ambiguity: The solution to the minimization problem (5.1) (the first term in (6.1)) is not unique. We now show that the second term (6.2) plays a role of regularization—it resolves this ambiguity by reducing the number of minimizers, and it gives rise to correlation of transformations in different grains of a laminated polycrystal. The correlations arise from *maximization* of $\bar{W}$ (note the negative sign in front of the second term in (6.1)). We formalize this idea in the next definition.

DEFINITION 6.2. *For a given deformation U and a sequence of angles $\{\theta_i\}_{i=1}^n$, an asymptotic energy minimizer is a pair of piecewise constant functions*

$$(s(x), I(x)) = (s_i, I_i) \text{ if } x \in \left(\frac{i-1}{n}, \frac{i}{n}\right],\ i = 1, 2, \ldots, n,\ s(x), I(x) \in \mathcal{H}_n,$$

which maximizes (6.2) *(the second term in* (6.1)*) amongst all*[7] *minimizers of the misfit energy of infinitely long grains* (5.1) *(the first term in* (6.1)*).*

When the distribution of angles ρ_n is given, we denote by μ_n^a the corresponding probability measure of the distribution of asymptotic energy minimizers $(s_i, I_i)_{i=1}^n$. The behavior of the asymptotic energy minimizers for the distribution of angles ρ_n given by the Bernoulli trials (4.3) model will be described by the Riesz symmetrically rearranged minimizer which we define as follows.

DEFINITION 6.3. *Consider Bernoulli trials* (4.3) *model (θ is α or β with probabilities q or $1-q$). For a given displacement U, let q_α and q_β be the proportion of grains for which $I = 1$ with $\theta = \alpha$ and $\theta = \beta$, respectively,*

$$0 \leq q_\alpha \leq q, \quad 0 \leq q_\beta \leq 1 - q, \tag{6.6}$$

[7]Note that we consider here *all* possible minimizers of the misfit energy of infinitely long grains (5.1) (the first term in (6.1)). In other words, here minimizers of (5.1) may not satisfy our assumption of equal probability (4.1).

as given by Lemma 5.3. *For a given sequence of angles* $\{\theta_i\}_{i=1}^n$ *a Riesz left-rearranged sequence of transformation strains is a pair of functions* $(s_l(x), I_l(x)) \in \mathcal{H}_n$ *given for each* $x \in ((i-1)/n, i/n]$ *as*

$$(s_l(x), I_l(x)) = \left(s_l\left(\frac{i}{n}\right), I_l\left(\frac{i}{n}\right)\right) = \begin{cases} (\sin^2\alpha, 1), \ \text{if } \theta = \alpha, \ 0 \le i/n \le q_\alpha/q, \\ (0,0), \ \text{if } \theta = \alpha, \ q_\alpha/q \le i/n \le 1, \\ (\sin^2\beta, 1), \ \text{if } \theta = \beta, \ 0 \le i/n \le q_\beta/(1-q), \\ (0,0), \ \text{if } \theta = \beta, \ q_\beta/(1-q) \le i/n \le 1. \end{cases}$$

Similarly, a Riesz right-rearranged sequence $(s_r(x), I_r(x)) \in \mathcal{H}_n$ *is determined by*

$$\begin{aligned}(s_r(x), I_r(x)) &= \left(s_r\left(\frac{i}{n}\right), I_r\left(\frac{i}{n}\right)\right) \\ &= \begin{cases} (\sin^2\alpha, 1), \ \text{if } \theta = \alpha, \ 1 - q_\alpha/q \le i/n \le 1, \\ (0,0), \ \text{if } \theta = \alpha, \ 0 \le i/n \le 1 - q_\alpha/q, \\ (\sin^2\beta, 1), \ \text{if } \theta = \beta, \ 1 - q_\beta/(1-q) \le i/n \le 1, \\ (0,0), \ \text{if } \theta = \beta, \ 0 \le i/n \le 1 - q_\beta/(1-q). \end{cases}\end{aligned}$$

We denote by μ_n^l *and* μ_n^r *the corresponding probability measures. A Riesz symmetrically rearranged probability measure* μ_n^s *is the average of right- and left-rearranged measures* $\mu_n^s = 1/2\mu_n^l + 1/2\mu_n^r$. *Finally, the Riesz symmetrically rearranged probability measure* μ^s, *the weak limit of probability measures:* $\mu_n^s \Rightarrow \mu^s$, *as* $n \to \infty$.

The rearranged minimizers quantitatively describe the rise of correlations for *asymptotic energy minimizers* (Definition 6.2), because, as we prove in Lemma 6.7, the asymptotic and rearranged minimizers coincide in the limit as $n \to \infty$. In other words, Definitions 6.2 and 6.3 characterize the same measure as $n \to \infty$. Moreover, the following theorem shows that the minimizer probability measure of the full misfit energy (6.1) converges to the Riesz symmetrically rearranged probability measure when we let $n \to \infty$, and then let $L \to \infty$.

THEOREM 6.4. *Consider the Bernoulli trials* (4.3). *For a given* U, *let* q_α *and* q_β *be defined as in Lemma* 5.3 *by* (5.15) *or* (5.16). *Then as* $n \to \infty$ *and subsequently* $L \to \infty$ *the probability measure of the energy minimizer of the misfit energy* (6.1) *converges weakly to the Riesz symmetrically rearranged probability measure* μ^s.

6.2. Riesz rearrangement inequalities and proof of Theorem 6.4. The key idea of the proof comes from the classical Riesz rearrangement inequality (see e.g., [12], [14]). In particular, this inequality motivated the name for minimizers in Definition 6.3. The simplest form of this inequality, which is sufficient for our purposes is as follows.

LEMMA 6.5. *Riesz rearrangement inequality on a circle.*

Consider two classes of even, bounded, and positive functions on $[-1, 1]$:

$$\mathcal{A}_i = \{f(x) | f(x) = f(-x), 0 \le f(x) \le 1, \int_0^1 f(x)dx = p_i\}, \ i = 1, 2.$$

Let $\chi_{p_i}^1(x) \in \mathcal{A}_i$ *be the characteristic function of the set* $[-p_i, p_i]$ *and* $\chi_{p_i}^2(x) \in \mathcal{A}_i$ *be the characteristic function of the set* $[-1, -1+p_i] \cup [1-p_i, 1]$. *Suppose* $K(x)$ *is an even positive locally integrable* 2*-periodic function on* $\mathbb{R}$ *that decreases on* $[0, 1]$:

(6.7)

$$K(x) \ge 0, \ K(x) = K(-x), \ \int_0^1 K(x)dx < \infty, \ K(x+2) = K(x), \ \text{and } K(x_1) > K(x_2),$$

if $0 < x_1 < x_2 \le 1$. *Then for any* $f(x) \in \mathcal{A}_1$, $g(x) \in \mathcal{A}_2$

$$\int_{-1}^{1}\int_{-1}^{1}(f(x)-p_1)(g(t)-p_2)K(x-t)dxdt$$
$$\le \int_{-1}^{1}\int_{-1}^{1}(\chi^1_{p_1}(x)-p_1)(\chi^1_{p_2}(t)-p_2)K(x-t)dxdt, \tag{6.8}$$

where the equality holds only in the following two cases

$$(a)\ f(x)=\chi^1_{p_1}(x), g(x)=\chi^1_{p_2}(x),\ (b)\ f(x)=\chi^2_{p_1}(x), g(x)=\chi^2_{p_2}(x). \tag{6.9}$$

Moreover, for any $\delta > 0$ *there exists* $\delta' > 0$ *so that if* $\min(e_1, e_2) \ge \delta'$,

$$e_1 = \int_0^1 (|f-\chi^1_{p_1}| + |g-\chi^1_{p_2}|)dx,\ e_2 = \int_0^1 (|f-\chi^2_{p_1}| + |g-\chi^2_{p_2}|)dx, \tag{6.10}$$

then

$$\int_{-1}^{1}\int_{-1}^{1}(f(x)-p_1)(g(t)-p_2)K(x-t)dxdt$$
$$\le \int_{-1}^{1}\int_{-1}^{1}(\chi^1_{p_1}(x)-p_1)(\chi^1_{p_2}(t)-p_2)K(x-t)dxdt-\delta. \tag{6.11}$$

We assumed in this lemma that the functions are bounded from above by one. This assumption can be replaced by any positive number with obvious modifications of the results. The proof of Lemma 6.5 follows from considerations similar to those found in [1]. It basically says that among all possible functions $0 \le f(x) \le 1$, $0 \le g(x) \le 1$ on a circle $[-1, 1]$ (where the endpoints $x = \pm 1$ are identified) the maximum of the integral

$$\int_{-1}^{1}\int_{-1}^{1}(f(x)-p_1)(g(t)-p_2)dxdt, \text{ with } \int_0^1 f(x)dx = p_1, \int_0^1 g(x)dx = p_2 \tag{6.12}$$

is achieved on characteristic functions of the intervals of length $2p_1$ and $2p_2$. The reason why the intervals centered at $x = 0$ and $x = 1$ is due to our assumption that $f(x)$ and $g(x)$ are even. In order to explain how Lemma 6.5 must be modified and applied for our case, we decompose

$$s(x) - \langle s\rangle = \sin^2\alpha\varepsilon_\alpha + \sin^2\beta\varepsilon_\beta,\ \varepsilon_\alpha = (\chi_\alpha\tilde{\chi}_\alpha - q_\alpha),\ \varepsilon_\beta = (\chi_\beta\tilde{\chi}_\beta - q_\beta),$$

where χ_α and χ_β, $\chi_\alpha + \chi_\beta = 1$ are (random) characteristic functions of the angle distributions $\theta = \alpha$ and $\theta = \beta$, respectively; $\tilde{\chi}_\alpha$ and $\tilde{\chi}_\beta$ are the characteristic function of the grains with $\theta = \alpha$ and $\theta = \beta$, respectively, for which $I = 1$.

The term (6.2) (the second term in (6.1)) equals

$$\bar{W}(U, \boldsymbol{\varepsilon}^{\mathrm{T}}) = \sin^4\alpha\int_{-1}^{1}\int_{-1}^{1}\varepsilon_\alpha(x)\varepsilon_\alpha(t)K_\infty(x-t)dxdt$$
$$+ 2\sin^2\alpha\sin^2\beta\int_{-1}^{1}\int_{-1}^{1}\varepsilon_\alpha(x)\varepsilon_\beta(t)K_\infty(x-t)dxdt \tag{6.13}$$
$$+ \sin^4\beta\int_{-1}^{1}\int_{-1}^{1}\varepsilon_\beta(x)\varepsilon_\beta(t)K_\infty(x-t)dxdt.$$

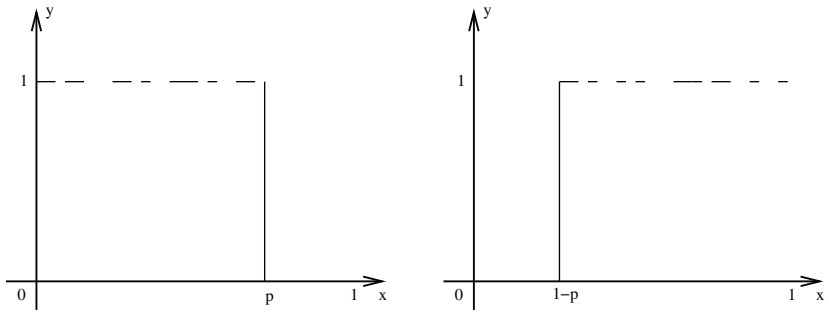


FIG. 6.1. *A sample of two random intervals* $\chi_\alpha(x)\chi^1_{q_\alpha}(x)$ *and* $\chi_\alpha(x)\chi^2_{q_\alpha}(x)$*, where* $p = q_\alpha/q$.

Each of the three integral terms in (6.13) has the form described in the previous Lemma 6.5, because by Lemma 3.1 the integral kernel $K_\infty = -\ln|\sin(\pi x/2)|$, and hence it satisfies all the conditions (6.7). As in Lemma 6.5 we need to maximize the integral (6.13) by varying the characteristic functions $\tilde\chi_\alpha$ and $\tilde\chi_\beta$. The only difference is the additional constraint that χ_α and χ_β are random characteristic functions. This additional constraint, loosely speaking, requires that the maximizers of (6.13) are "random intervals" still centered at $x = 0$ or $x = 1$. More precisely, note that the values of the characteristic functions $\tilde\chi_\alpha$ and $\tilde\chi_\beta$ in (6.13) are important only where $\chi_\alpha = 1$ and $\chi_\beta = 1$, respectively. Hence for a sequence of Bernoulli random variables θ_i, $i = 1, \dots, n$ we can define characteristic functions of random intervals of length $2q_\alpha$ on $[-1, 1]$ centered at $x = 0$ and $x = 1$ as a product of two characteristic functions $\chi_\alpha(x)\chi^1_{q_\alpha}(x)$ and $\chi_\alpha(x)\chi^2_{q_\alpha}(x)$, respectively, where

$$\chi^1_{q_\alpha}(x) = \begin{cases} 1, -\frac{q_\alpha}{q} \le x \le \frac{q_\alpha}{q}, \\ 0, \text{ otherwise}, \end{cases} \qquad \chi^2_{q_\alpha}(x) = \begin{cases} 0, -1 + \frac{q_\alpha}{q} < x < 1 - \frac{q_\alpha}{q}, \\ 1, \text{ otherwise}. \end{cases} \tag{6.14}$$

Similarly, functions $\chi_\alpha(x)\chi^1_{q_\alpha}(x)$ and $\chi_\alpha(x)\chi^2_{q_\alpha}(x)$ are random intervals of length $2q_\beta$ centered at $x = 0$ and $x = 1$ where

$$\chi^1_{q_\beta}(x) = \begin{cases} 1, -\frac{q_\beta}{1-q} \le x \le \frac{q_\beta}{1-q}, \\ 0, \text{ otherwise}, \end{cases} \qquad \chi^2_{q_\beta}(x) = \begin{cases} 0, -1 + \frac{q_\beta}{1-q} < x < 1 - \frac{q_\beta}{1-q}, \\ 1, \text{ otherwise}. \end{cases} \tag{6.15}$$

For an illustration see Figure 6.1. The above discussion is made rigorous by the following.

LEMMA 6.6 (randomized Riesz rearrangement inequality for asymptotic energy minimizers). *Consider Bernoulli trials* (4.3). *Suppose* χ_α *and* χ_β, $\chi_\alpha + \chi_\beta = 1$ *are (random) characteristic functions of the angle distributions* $\theta = \alpha$ *and* $\theta = \beta$, *respectively. Let* $q_\alpha(\boldsymbol{\theta})$ *and* $q_\beta(\boldsymbol{\theta})$ *be random variables of* $\boldsymbol{\theta}$ *with values*

$$0 \le q_\alpha(\boldsymbol{\theta}) \le q, \ 0 \le q_\beta(\boldsymbol{\theta}) \le 1 - q.$$

Suppose $K(x)$ *satisfies* (6.7). *Then for every* $\delta > 0$ *there exists* $\delta' > 0$ *so that if*

$$|q_\alpha(\boldsymbol{\theta}) - q_\alpha| < \delta', \ |q_\beta(\boldsymbol{\theta}) - q_\beta| < \delta' \tag{6.16}$$

for some fixed q_α and q_β, then almost surely[8] *as $n \to \infty$ the maximizers of*

(6.17)
$$\max_{\tilde{\chi}_\alpha, \tilde{\chi}_\beta} \int_{-1}^{1}\int_{-1}^{1} (a\varepsilon_\alpha(x) + b\varepsilon_\beta(x))(a\varepsilon_\alpha(t) + b\varepsilon_\beta(t))K(x-t)dxdt,$$
$$\varepsilon_\alpha = (\chi_\alpha\tilde{\chi}_\alpha - q_\alpha(\boldsymbol{\theta})), \varepsilon_\beta = (\chi_\beta\tilde{\chi}_\beta - q_\beta(\boldsymbol{\theta})), a > 0, b > 0$$
$$\chi^\alpha(x) = \chi^\alpha(-x),\ \chi^\beta(x) = \chi^\beta(-x),\ \int_0^1 \chi^\alpha\tilde{\chi}_\alpha dx = q_\alpha(\boldsymbol{\theta}),\ \int_0^1 \chi^\beta\tilde{\chi}_\beta dx = q_\beta(\boldsymbol{\theta})$$

satisfy

$$\min(e_1, e_2) < \delta, \tag{6.18}$$

where

(6.19)
$$e_1 = \int_0^1 (|\tilde{\chi}_\alpha - \chi^1_{q_\alpha}|\chi_\alpha + |\tilde{\chi}_\beta - \chi^1_{q_\beta}|\chi_\beta)dx, e_2 = \int_0^1 (|\tilde{\chi}_\alpha - \chi^2_{q_\alpha}|\chi_\alpha + |\tilde{\chi}_\beta - \chi^2_{q_\beta}|\chi_\beta)dx,$$

$\chi^i_{q_\alpha}$ and $\chi^i_{q_\beta}$, $i = 1, 2$ are defined in (6.14) *and* (6.15)*, respectively.*

The proof of Lemma 6.6 is by contradiction to the law of large numbers and it is provided in Appendix C.2. The next lemma shows the equivalence of Definitions 6.2 and 6.3 as $n \to \infty$.

LEMMA 6.7. *Consider the Bernoulli trials* (4.3)*. For a given U, let q_α and q_β be defined as in Lemma* 5.3 *by* (5.15) *or* (5.16)*. Then the probability measure μ^a_n of the asymptotic energy minimizer of the misfit energy* (6.1) *(see Definition* 6.2*)) and the probability measure μ^s_n of the Riesz symmetrically rearranged energy minimizer (Definition* 6.3*) have the same weak limit μ^s. Moreover,*

$$\lim_{n\to\infty} e_1 = 0,\ \textit{or}\ \lim_{n\to\infty} e_2 = 0\ \textit{with equal probability}\ 1/2, \tag{6.20}$$

where e_1 and e_2 are defined by (6.19) *in Lemma* 6.17 *above.*

Proof. By Lemmas 5.3 and 5.5, depending on U, the minimizer of the first term in (6.1) satisfies (5.15) or (5.16) for almost every $\boldsymbol{\theta}$ as $n \to \infty$. Since the condition (6.16) of the randomized Riesz rearrangement inequality is satisfied for any $\delta > 0$, by Lemma 6.17 for almost every $\boldsymbol{\theta}$,

$$\lim_{n\to\infty} \min(e_1, e_2) = 0.$$

Hence $\{(s_i, I_i)\}_{i=1}^n$ is either a left-rearranged or right-rearranged sequence almost surely as $n \to \infty$. For every n the measure μ^a_n must be symmetric with respect to the to reflection about the point $x = 1/2$, i.e. with equal probability either $e_1 \to 0$ or $e_2 \to 0$ as $n \to \infty$. This proves (6.20). Clearly (6.20) implies that μ^a_n and the Riesz symmetrically rearranged measure μ^s_n have the same weak limit as $n \to \infty$. ∎

End of proof of Theorem 6.4. Again, by the symmetry of the problem with respect to reflection about the point $x = 1/2$, μ_n also must be similarly symmetric. By Lemma 6.1 the minimizer of (6.1) is an almost minimizer, i.e. for any $\delta > 0$ there is L_0 so that the condition (6.16) of the randomized Riesz rearrangement inequality holds.

[8]Here and in the sequel almost sure convergence is considered in the probability space $(\Omega, \mathcal{F}, P)$ set in the beginning of section 4.

Hence it implies that for sufficiently large L_0 the minimizing sequences $\{(s_i, I_i)\}_{i=1}^n$ for any $L > L_0$ are arbitrarily close to the Riesz symmetrically rearranged minimizing sequences, namely either

(6.21)
$$\int_0^1 (|\tilde{\chi}_\alpha - \chi^1_{q_\alpha}|\chi_\alpha + |\tilde{\chi}_\beta - \chi^1_{q_\beta}|\chi_\beta)dx < \delta, \text{ or } \int_0^1 (|\tilde{\chi}_\alpha - \chi^2_{q_\alpha}|\chi_\alpha + |\tilde{\chi}_\beta - \chi^2_{q_\beta}|\chi_\beta)dx < \delta,$$

with equal probability $1/2$ as $n \to \infty$. If $L_0 \to \infty$, then $\delta \to 0$, and this completes the proof.

It follows from (6.21) that long/short-range correlations in the minimizing sequences $\{(s_i, I_i)\}_{i=1}^n$ are determined by long/short-range correlations of the Riesz symmetrically rearranged minimizing sequences. There is no short-range correlations of the Riesz symmetrically rearranged minimizing sequences. Riesz rearranged measure is, however, correlated on the large-scale: For example, suppose $q_\alpha = q$ and $q_\beta < 1 - q$, then for the right-rearranged measure:

$$(s, I) = \begin{cases} (\sin^2 \alpha, 1), \text{ with probability } q, \\ (\sin^2 \beta, 1), \text{ with probability } 1-q, \text{ if } 1 - q_\beta/(1-q) \le x \le 1, \\ (0, 0), \text{ with probability } 1-q, \text{ if } 0 \le x < 1 - q_\beta/(1-q), \end{cases} \tag{6.22}$$

and for the left-rearranged measure:

$$(s, I) = \begin{cases} (\sin^2 \alpha, 1), \text{ with probability } q, \\ (\sin^2 \beta, 1), \text{ with probability } 1-q, \text{ if } 0 \le x < q_\beta/(1-q), \\ (0, 0), \text{ with probability } 1-q, \text{ if } q_\beta/(1-q) \le x \le 1. \end{cases} \tag{6.23}$$

The long-range correlation of transformation strains for the symmetrically rearranged minimizer probability measure can be read off the formulas (6.22) and (6.23).

7. Statistics of asymptotic energy minimizers 3: Chain of short grains.

7.1. Basic definitions and ideas. In this section we will show how exponentially decaying correlations arise when the scaling of the polycrystal is such that $L = l_0/(2n)$, $l_0 \ll 1$ (short grains). By estimate (3.16) the misfit energy for $l_0 \ll 1$ is given (up to higher order terms) by the nearest neighbor energy

$$W_n(U, \boldsymbol{\varepsilon}^{\mathrm{T}}) = \langle (U - f + s)^2 \rangle - \lambda_0 \langle (s - \langle s \rangle)^2 \rangle - \frac{B l_0}{n} \sum_{i=1}^n (s_i - \langle s \rangle)(s_{i+1} - \langle s \rangle), \tag{7.1}$$

where $B > 0$, $\lambda_0 > 0$. In this case, we show that for Bernoulli trials (4.3) exponentially decaying correlations arise when $n \to \infty$, followed by $l_0 \to 0$.

Qualitatively, the misfit energy $W_n^1(U, \boldsymbol{\varepsilon}^{\mathrm{T}})$ has three terms which are analogous to the case of thin long grains (6.1). The minimization of the first two terms,

$$W_n^0(U, \boldsymbol{\varepsilon}^{\mathrm{T}}) = \langle (U - f + s)^2 \rangle - \lambda_0 \langle (s - \langle s \rangle)^2 \rangle \tag{7.2}$$

determines, as in Lemma 5.3, q_α and q_β, the *total* amount of the grains that undergo a stress-free transformation. The minimizers of (7.2) are, in general, not unique. The third term provides a small correction to (7.2), and, as in section 6, plays a role of regularization, that is it selects the unique minimizer of (7.2) that maximizes

$$\bar{W}_n(U, \boldsymbol{\varepsilon}^{\mathrm{T}}) := \frac{1}{n} \sum_{i=1}^n (s_i - \langle s \rangle)(s_{i+1} - \langle s \rangle). \tag{7.3}$$

Analogous to Definition 6.2, the above considerations motivate the following definition:

DEFINITION 7.1. *For a fixed sequence θ_i, $i = 1, \dots, n$, the asymptotic energy minimizer of the nearest neighbor model* (7.1) *is the sequence (s_i, I_i), $i = 1, \dots, n$ such that it minimizes* (7.2) *and maximizes* (7.3) *among minimizers of* (7.2).

LEMMA 7.2. *For the Bernoulli trials model* (4.3) *the minimizing sequence (s_i, I_i), $i = 1, \dots, n$ of the misfit energy $W_n^0(U, \boldsymbol{\varepsilon}^T)$ given by* (7.2) *satisfies*

$$s(\theta) = \sin^2 \theta, \tag{7.4}$$

and as $n \to \infty$,

$$q_\alpha(\boldsymbol{\theta}) \to q_\alpha, q_\beta(\boldsymbol{\theta}) \to q_\beta$$

almost surely. The values q_α and q_β are determined as follows. For a given U the total proportion of grains that undergoes a stress-free transformation $f = f(U)$ is a (deterministic) nondecreasing function of U. For a given f we have several cases:

- *if $f < 1 - q$, then*

$$q_\alpha = 0,\ 0 \le q_\beta \le 1 - q, \tag{7.5}$$

$\chi(\alpha) = 0$, $\chi(\beta) = q_\beta/(1 - q)$,

- *if $f > 1 - q$, then*

$$0 \le q_\alpha \le q,\ q_\beta = 1 - q, \tag{7.6}$$

$\chi(\beta) = q_\alpha/q$, $\chi(\beta) = 1$.

There are values of α, β, and U for which $0 < q_\beta < 1 - q$ or $0 < q_\alpha < q$.

The proof of this lemma is analogous to Lemma 5.3. Clearly, the function $f(U)$ in Lemma 7.2 is different from the one for the infinitely long grains in Lemma 5.3. However, the characteristic property of the measure that it is determined by q_α and q_β with either (7.5) or (7.6) still holds. One of the consequences of the previous lemma is that there are, again, some values q_α and q_β that determine the proportion of grains that undergo a stress-free transformation and they satisfy $q_\alpha = 0$ or $q_\beta = 1 - q$. This is exactly the characteristic property that we need to be able to prove exponential decay of correlations by applying the isoperimetric inequality (7.8) to the sequences described in Definition 7.1. Following the logic in section 6, we obtain that asymptotic energy minimizer of the nearest neighbor model (7.1) arises in the limit $n \to \infty$, followed by $l_0 \to \infty$. The proof of this statement is similar to the proof of the analogous statement in case 2; see end of the proof of Theorem 6.4 in section 6.2. Hence we only need to find a statistical characterization of maximizers of (7.3) for fixed q_α and q_β found from Lemma 7.2. This is given in Theorems 7.4 and 7.5 below.

7.2. Isoperimetric inequalities and characterization of maximizers of (7.3). Here, it is convenient to characterize any point in the composite $x \in [0, 1]$ as a point that belongs to a (maximal) uninterrupted string of identical values of θ.

DEFINITION 7.3. *For a fixed $\boldsymbol{\theta} = \theta_1, \theta_2, \dots, \theta_n$ we say that a string*

$$\theta_\alpha^m = \theta_{i+1}, \theta_{i+2}, \dots, \theta_{i+m},\ \theta_\alpha^m \subset \boldsymbol{\theta}$$

is a (maximal) uninterrupted string of $\theta = \alpha$ of length m if all $\theta_{i+j} = \alpha$, $i = 1, \dots, m$ and $\theta_i = \theta_{i+1+m} = \beta$. We say that $x \in [0, 1]$ belongs to an uninterrupted string of values α of length m if $x \in \theta_\alpha^m$. The notion $x \in \theta_\beta^m$ is defined analogously.

Recall our notation

$$s(x) - \langle s \rangle = \sin^2 \alpha \varepsilon_\alpha + \sin^2 \beta \varepsilon_\beta,$$

$$\varepsilon_\alpha = (\chi_\alpha \tilde{\chi}_\alpha - q_\alpha), \quad \varepsilon_\beta = (\chi_\beta \tilde{\chi}_\beta - q_\beta),$$

where χ_α and χ_β, $\chi_\alpha + \chi_\beta = 1$ are (random) characteristic functions of the angle distributions $\theta = \alpha$ and $\theta = \beta$, respectively; $\tilde{\chi}_\alpha$ and $\tilde{\chi}_\beta$ are the characteristic function of the grains with $\theta = \alpha$ and $\theta = \beta$, respectively, for which $I = 1$.

By Lemma 7.2, we have two cases: Either $q_\beta = 1 - q$, and then $\tilde{\chi}_\beta \equiv 1$, or $q_\alpha = 0$, and then $\tilde{\chi}_\alpha \equiv 0$. Let us study these two cases separately.

Suppose $q_\alpha = 0$. Let us look at maximization of

$$\frac{1}{n} \sum_{i=1}^{n} s_i s_{i+1}, \ \langle s \rangle = q_\beta \sin^2 \beta \tag{7.7}$$

only. Each of s_i (up to the constant $\sin^2 \beta$) is either 1 or 0; therefore the maximization of the nearest neighbors term (7.7) can be understood as the minimization of the boundary of a set with constant area:

$$\min_{D \in \mathcal{A}} \partial D, \ \mathcal{A} = \left\{ D | D = \{x | s_i(x) \neq 0\}, \int_D dx = q_\beta \right\}. \tag{7.8}$$

Then the usual isoperimetric inequality implies that the maximizer of (7.7) is such that the grains with $\theta_i = \beta$ undergo a stress-free transformation, if they belong to a "long" uninterrupted sequence θ_β^m of the grains with the same $\theta = \beta$. If $\theta_i = \beta$ belongs to a "short" uninterrupted sequence θ_β^m, then it does not undergo a stress-free transformation. Hence there should be short-range correlations. The notion of short and long sequences is relative to the value of the total number of grains that must undergo a stress-free transformation. The above ideas are formulated more precisely in the next theorem.

THEOREM 7.4. *Consider the Bernoulli trials* (4.3). *Denote by* $\tilde{\chi}_\alpha$ *and* $\tilde{\chi}_\beta$ *the characteristic function of the grains with* $\theta = \alpha$ *and* $\theta = \beta$, *respectively, for which* $I = 1$. *Suppose* U *is such that the minimizer of the first term in* (7.1) *satisfies* $q_\alpha = 0$. *Then in the limit* $n \to \infty$ *the sequence* (s_i, I_i), $i = 1, 2, \ldots, n, \ldots$ *is a stationary process with exponentially decaying short-range correlations, long-range correlations are zero, and* $\tilde{\chi}_\alpha \equiv 1$. *Moreover, almost surely as* $n \to \infty$ *the minimizer of the nearest neighbor model* (7.1) *satisfies*

$$\tilde{\chi}_\beta(x) = \begin{cases} 1, \ \textit{if } x \in \theta_\beta^m, \ m > k, \\ 0, \ \textit{if } x \in \theta_\beta^m, \ m < k, \\ 1, \ \textit{with probability } r, \ \textit{if } x \in \theta_\beta^k, \\ 0, \ \textit{with probability } 1 - r, \ \textit{if } x \in \theta_\beta^k, \end{cases}$$

where

$$k = \max(m) \ \textit{such that } q_\beta < (1 - q)^m, \tag{7.9}$$

and r *is found from*

$$q_\beta = rq(1 - q)^k - (1 - q)^{k+1}. \tag{7.10}$$

Proof. By the law of the large numbers, $\tilde{\chi}_\alpha = 0$ and $q_\beta(\boldsymbol{\theta}) \to q_\beta$ almost surely; therefore it is sufficient to study (7.7) or, equivalently, (7.8). By the isoperimetric inequality for every $\boldsymbol{\theta}$ the function $\tilde{\chi}_\beta(x)$ must be such that if $\theta_\beta^{m_1}$, $\theta_\beta^{m_2}$, $\dots$ $\theta_\beta^{m_t}$ are all the (maximal) uninterrupted sequences $\theta_\beta^{m_i} \in \boldsymbol{\theta}$ of values β ordered so that the indices are decreasing $m_1 \geq m_2 \geq \cdots \geq m_t$, then there exists an i: $1 \leq i \leq t$ so that $\tilde{\chi}(x) = 1$, if $x \in \theta_\beta^{m_j}$, $j \leq i$, and $\tilde{\chi}(x) = 0$, if $x \in \theta_\beta^{m_j}$, $j > i$ with the exception of at most one $j \geq i$.

By construction as $n \to \infty$, the process (s_i, I_i), $i = 1, \dots, n, \dots$ is stationary. Since all θ_α^m, θ_β^m are geometrically distributed independent random variables [11], it means explicitly that the probability of a string θ_β^m is given by

$$\rho_\infty(\dots \theta_\beta^m \dots) = q(1-q)^{m-1}.$$

Hence if the total proportion of grains that undergoes a stress-free transformation is q_β, we must have, as $n \to \infty$

$$q_\beta = rq(1-q)^k + \sum_{i=k+1} q(1-q)^i = rq(1-q)^k + (1-q)^{k+1},$$

where, due to our assumption of equal probability (4.1), r is the probability that $\tilde{\chi}(x) = 1$, if $x \in \theta_\beta^k$. Therefore k is found so that $(1-q)^{k+1} \leq q_\beta < (1-q)^k$, i.e., (7.9), and r is found from (7.10).

Since θ_α^m, θ_β^m are independent random variables, the limiting process has exponentially decaying short-range correlations. It implies simultaneously two results: long-range correlations are zero, and short-range correlations decay exponentially with k. These correlations are not zero and can be computed explicitly. □

Suppose $q_\beta = 1 - q$. This case is slightly more technically complicated, but the methods are the same as in the case $q_\alpha = 0$. The main new issue is that s_i may now take *three* values, and, therefore, we have to account for three possible different interfaces. Direct computations show that we have here three different situations, depending on the relative value of α and β. If $\sin^2\alpha > 2\sin^2\beta$, then the maximizer of (7.7) is such that $\tilde{\chi}_\alpha(x) = 1$ if x belongs to the *longest* (maximal) uninterrupted strings θ_α^m. If, however, $\sin^2\alpha < 2\sin^2\beta$, then $\tilde{\chi}_\alpha(x) = 1$ if x belongs to to the *shortest* uninterrupted strings θ_α^m. If $\sin^2\alpha = 2\sin^2\beta$, then there is no difference, and the only statement that is possible to make here is that $\tilde{\chi}_\alpha(x) = \tilde{\chi}_\alpha(y)$, if x and y belong to the same uninterrupted string θ_α^m. Due to our assumption of equal probability (4.1), it is possible to conclude that if $\sin^2\alpha = 2\sin^2\beta$, then there is no correlation at all, therefore we will omit the discussion of this case. Combining these arguments with the arguments in the proof of Theorem 7.4 we have the following result.

THEOREM 7.5. *Consider the Bernoulli trials* (4.3). *Denote by* $\tilde{\chi}_\alpha$ *and* $\tilde{\chi}_\beta$ *the characteristic function of the grains with* $\theta = \alpha$ *and* $\theta = \beta$, *respectively, for which* $I = 1$. *Suppose* U *is such that the minimizer of the first term in* (7.1) *satisfies* $q_\beta = 1 - q$. *Then in the limit* $n \to \infty$ *the sequence* (s_i, I_i), $i = 1, 2, \dots, n, \dots$ *is a stationary process with exponentially decaying short-range correlations; long-range correlations are zero, and* $\tilde{\chi}_\beta \equiv 1$. *Moreover, almost surely as* $n \to \infty$, *the minimizer of the nearest neighbor model* (7.1) *satisfies: If* $\sin^2\alpha > 2\sin^2\beta$, *then*

$$\tilde{\chi}_\alpha(x) = \begin{cases} 1, \text{ if } x \in \theta_\alpha^m, \ m > k, \\ 0, \text{ if } x \in \theta_\alpha^m, \ m < k, \\ 1, \text{ with probability } r, \text{ if } x \in \theta_\alpha^k, \\ 0, \text{ with probability } 1-r, \text{ if } x \in \theta_\alpha^k, \end{cases}$$

where $k = \max(m)$ *such that* $q_\alpha < q^m$ *and* r *solves* $q_\alpha = r(1-q)q^k - q^{k+1}$*; if* $\sin^2\alpha < 2\sin^2\beta$*, then*

$$\tilde{\chi}_\alpha(x) = \begin{cases} 1, \ \textit{if } x \in \theta_\alpha^m, \ m < k, \\ 0, \ \textit{if } x \in \theta_\alpha^m, \ m > k, \\ 1, \ \textit{with probability } r, \ \textit{if } x \in \theta_\alpha^k, \\ 0, \ \textit{with probability } 1-r, \ \textit{if } x \in \theta_\alpha^k, \end{cases}$$

where $k = \max(m)$ *such that* $q - q_\alpha < q^m$ *and* r *solves* $q - q_\alpha = r(1-q)q^k - q^{k+1}$.

Appendix A. Proofs for section 3.

A.1. Proof of formula (3.4). To obtain the representation (3.4) we begin by decomposing $\boldsymbol{u} = (u_1, u_2)$ of the boundary value problem (2.11), (2.13), and (2.14) in the form $\boldsymbol{u} = \tilde{\boldsymbol{u}} + \bar{\boldsymbol{u}}$, where $\tilde{\boldsymbol{u}}$ solves (3.6). The constants a_i, c_i, d, f_i, and g_i are chosen to satisfy the continuity of the displacement $\tilde{\boldsymbol{u}}$ and traction (condition (2.12)) and, denoting

$$\tilde{\sigma}_{ij} = c_{ijkl}(\tilde{\varepsilon}_{kl} - \varepsilon^{\mathrm{T}}_{kl}), \quad \tilde{\varepsilon}_{11} = \partial_1\tilde{u}_1, \quad \tilde{\varepsilon}_{12} = \tilde{\varepsilon}_{21} = (\partial_2\tilde{u}_1 + \partial_1\tilde{u}_2)/2, \quad \tilde{\varepsilon}_{22} = \partial_2\tilde{u}_2,$$

the boundary conditions

$$\tilde{u}_1(0,y) = 0, \ \ \tilde{u}_1(1,y) = U, \ \ \tilde{u}_2(0,0) = 0, \ \ \tilde{\sigma}_{12}(x,y) = 0, \ \text{for } x = 0, 1.$$

The stresses are

$$\begin{aligned} \tilde{\sigma}_{i,11} &= (\lambda + 2\mathrm{G})(a_i - \varepsilon^{\mathrm{T}}_{i,11}) + \lambda(d - \varepsilon^{\mathrm{T}}_{i,22}), \\ \tilde{\sigma}_{i,22} &= (\lambda + 2\mathrm{G})(d - \varepsilon^{\mathrm{T}}_{i,22}) + \lambda(a_i - \varepsilon^{\mathrm{T}}_{i,11}), \\ \tilde{\sigma}_{i,12} &= \mathrm{G}(c_i - 2\varepsilon^{\mathrm{T}}_{i,12}). \end{aligned}$$

Hence $[\tilde{\sigma}_{12}] = 0$, if $c_i = 2\varepsilon^{\mathrm{T}}_{i,12}$. Similarly $[\tilde{\sigma}_{i,11}] = 0$, if

$$a_i = \varepsilon^{\mathrm{T}}_{i,11} + \frac{\lambda}{\lambda + 2\mathrm{G}}\varepsilon^{\mathrm{T}}_{i,22} + \mathrm{Const},$$

where the constant can be found from the condition that the displacement of the right boundary is U:

$$\mathrm{Const} = U - \langle\varepsilon^{\mathrm{T}}_{11}\rangle - \frac{\lambda}{\lambda + 2\mathrm{G}}\langle\varepsilon^{\mathrm{T}}_{22}\rangle.$$

Finally,

$$\begin{aligned} \Big\langle \tilde{\sigma}_{22}\rangle &= \Big\langle (\lambda + 2\mathrm{G})(d - \varepsilon^{\mathrm{T}}_{22}) + \lambda\left(\frac{\lambda}{\lambda + 2\mathrm{G}}\varepsilon^{\mathrm{T}}_{22} + U - \langle\varepsilon^{\mathrm{T}}_{11}\rangle - \frac{\lambda}{\lambda + 2\mathrm{G}}\langle\varepsilon^{\mathrm{T}}_{22}\rangle\right)\Big\rangle \\ &= (\lambda + 2\mathrm{G})(d - \langle\varepsilon^{\mathrm{T}}_{22}\rangle) + \lambda(U - \langle\varepsilon^{\mathrm{T}}_{11}\rangle). \end{aligned}$$

Setting $\langle\tilde{\sigma}_{22}\rangle = 0$ we have

$$d = \langle\varepsilon^{\mathrm{T}}_{22}\rangle - \frac{\lambda}{\lambda + 2\mathrm{G}}(U - \langle\varepsilon^{\mathrm{T}}_{11}\rangle).$$

The values of f_i and g_i in (3.6) are unimportant for our analysis, and we omit them. We have

$$\tilde{\sigma}_{11} = E_c(U - \langle\varepsilon^{\mathrm{T}}_{11}\rangle), \ \tilde{\sigma}_{i,22} = E_c(\langle\varepsilon^{\mathrm{T}}_{22}\rangle - \varepsilon^{\mathrm{T}}_{i,22}), \ \tilde{\sigma}_{12} = 0, \tag{A.1}$$

where the Young's modulus E_c is given by (2.16). The elastic misfit energy associated with $\tilde{\boldsymbol{u}}$ is

$$\frac{1}{E_c}\frac{1}{4L}\int_{-L}^{L}\int_{-1}^{1}\Big[\frac{1}{E_c}((\tilde{\sigma}_{11})^2+(\tilde{\sigma}_{22})^2)-2\frac{\lambda}{4\mathrm{G}(\lambda+\mathrm{G})}\tilde{\sigma}_{11}\tilde{\sigma}_{22}+\frac{1}{2\mathrm{G}}(\tilde{\sigma}_{12})^2\Big]dxdy.$$

Since $\tilde{\sigma}_{12}\equiv 0$, $\tilde{\sigma}_{11}=\text{const}$, and $\tilde{\sigma}_{22}$ is mean-zero, the above equation becomes

$$\frac{1}{E_c^2}\frac{1}{4L}\int_{-L}^{L}\int_{-1}^{1}\left((\tilde{\sigma}_{11})^2+(\tilde{\sigma}_{22})^2\right)dxdy=\langle\left(U-f+\varepsilon_{22}^{\mathrm{T}}\right)^2\rangle=\langle(U-f+s)^2\rangle,$$

which is the first term in (3.4).

Let us now find $\bar{\boldsymbol{u}}$. It solves (3.7) on a bounded domain Π_L. A useful periodic setting for (3.7) is obtained by assuming this equation is posed on an infinite strip $]-\infty,\infty[\times[-L,L]$ with data that is even and periodic in x:

$$\begin{aligned}&(a)\ \theta(-x)=\theta(x),\ \theta(x+2)=\theta(x),\\ &(b)\ \boldsymbol{\varepsilon}^{\mathrm{T}}(x,y)=\boldsymbol{\varepsilon}^{\mathrm{T}}(-x,y),\ \boldsymbol{\varepsilon}^{\mathrm{T}}(x+2,y)=\boldsymbol{\varepsilon}^{\mathrm{T}}(x,y).\end{aligned}\tag{A.2}$$

Thus, $\bar{\boldsymbol{u}}$ equals to the restriction of the solution of (3.7) on an infinite strip $x\in\ ]-\infty,\infty[$, $y\in[-L,L]$ with periodicity conditions defined by (A.2). Since $\bar{\sigma}_{22}(x,\pm L)$ is a periodic, mean-zero, even function, it can be represented as a cosine Fourier series. The solution on the infinite strip with a sinusoidal symmetric stress $\cos(k\pi x)$ at $y=\pm L$ can be computed explicitly for any k by the Airy function method. Namely, since we are given that $\bar{\sigma}_{22}(x,\pm L)$ is a periodic, mean-zero, even function, it can be represented as

$$\bar{\sigma}_{22}(x,\pm L)=E_c\sum_{k=1}^{\infty}c_k\cos(k\pi x),\ E_c=\frac{4\mathrm{G}(\lambda+\mathrm{G})}{\lambda+2\mathrm{G}},$$

where c_k are the corresponding Fourier coefficients of $\varepsilon^{\mathrm{T}}(x)$. The solution for the infinite strip with a sinusoidal symmetric stress $\cos(k\pi x)$ at $y=\pm L$ is given (see [19]) by the Airy function

$$\begin{aligned}&\Phi^k(x,y)\\ &\quad=2\frac{\cos(k\pi x)}{(k\pi)^2}\frac{k\pi y\sinh(k\pi L)\sinh(k\pi y)-[k\pi L\cosh(k\pi L)+\sinh(k\pi L)]\cosh(k\pi y)}{2k\pi L+\sinh(2k\pi L)}.\end{aligned}$$

This Airy function gives rise to the following stresses

$$\begin{aligned}\bar{\sigma}_{11}^k=\frac{\partial^2\Phi^k}{\partial y^2}=d_{11}^k\cos(k\pi x),\ \bar{\sigma}_{12}^k=-\frac{\partial^2\Phi^k}{\partial x\partial y}\\ =d_{12}^k\sin(k\pi x),\ \bar{\sigma}_{22}^k=\frac{\partial^2\Phi^k}{\partial x^2}=d_{22}^k\cos(k\pi x),\end{aligned}$$

where

$$\begin{aligned}d_{11}^k&=2\frac{k\pi y\sinh(k\pi L)\sinh(k\pi y)-[k\pi L\cosh(k\pi L)-\sinh(k\pi L)]\cosh(k\pi y)}{2k\pi L+\sinh(2k\pi L)},\\ d_{12}^k&=2\frac{k\pi y\sinh(k\pi L)\cosh(k\pi y)-k\pi L\cosh(k\pi L)\sinh(k\pi y)}{2k\pi L+\sinh(2k\pi L)},\\ d_{22}^k&=2\frac{-k\pi y\sinh(k\pi L)\sinh(k\pi y)+[k\pi L\cosh(k\pi L)+\sinh(k\pi L)]\cosh(k\pi y)}{2k\pi L+\sinh(2k\pi L)}.\end{aligned}\tag{A.3}$$

Therefore the total stresses are

$$\sigma_{11} = E_c(U - \langle \varepsilon_{11}^{\mathrm{T}} \rangle) + E_c \sum_{k=1}^{\infty} c_k d_{11}^k \cos(k\pi x),$$

$$\sigma_{12} = E_c \sum_{k=1}^{\infty} c_k d_{12}^k \sin(k\pi x), \;\; \sigma_{22} = E_c \sum_{k=1}^{\infty} c_k (d_{22}^k - 1) \cos(k\pi x),$$

where in the last equation we have $d_{22}^k - 1$ instead of d_{22}^k, because (see (A.1))

$$\tilde{\sigma}_{22} = -E_c \sum_{k=1}^{\infty} c_k \cos(k\pi x).$$

By definition

$$W = \frac{1}{E_c} \frac{1}{4L} \int_{-L}^{L} \int_{-1}^{1} \Big[\frac{1}{E_c} ((\sigma_{11})^2 + (\sigma_{22})^2) - 2 \frac{\lambda}{4\mathrm{G}(\lambda + \mathrm{G})} \sigma_{11}\sigma_{22} + \frac{1}{2\mathrm{G}} (\sigma_{12})^2 \Big] dxdy.$$

Since $\int_{-1}^{1} \cos(k\pi x) \cos(m\pi x) dx = \delta_{km}$

$$W(U, \boldsymbol{\varepsilon}^{\mathrm{T}}) = \langle (U - f + \varepsilon_{22}^{\mathrm{T}})^2 \rangle + \sum_{m=1}^{\infty} c_m^2 \hat{K}_L(m),$$

where

$$\text{(A.4)} \quad \hat{K}_L(m) = \frac{1}{4L} \int_{-L}^{L} \Big[(d_{11}^m)^2 + d_{22}^m (d_{22}^m - 2) - \frac{2\lambda d_{11}^m (d_{22}^m - 1)}{\lambda + 2\mathrm{G}} + \frac{2(\lambda + \mathrm{G})(d_{12}^m)^2}{\lambda + 2\mathrm{G}} \Big] dy.$$

Denoting $a = 2\pi mL$ and using Mathematica® we obtain an explicit form of (A.4):

$$\text{(A.5)} \qquad \hat{K}_L(m) = \frac{5\lambda + 9\mathrm{G}}{2(\lambda + 2\mathrm{G})} S_1(a) - \frac{\lambda + \mathrm{G}}{\lambda + 2\mathrm{G}} S_2(a), \;\; a = 2\pi mL, \;\; \text{where}$$

$$\text{(A.6)} \qquad S_1(a) = \frac{(\cosh(a) - 1)}{a(a + \sinh(a))}, \quad S_2(a) = \frac{a^2(2 + \cosh(a))}{6(a + \sinh(a))^2}.$$

A.2. Proof of Lemma 3.1. Using (A.5) and (A.6) from Appendix A.1, direct computations show that the Fourier coefficients of $K_L(x)$ (3.5) are given by

$$\text{(A.7)} \qquad \hat{K}_L(m) = \frac{B}{L} \frac{1}{m} + O(\exp(-L)),$$

so that, defining

$$\text{(A.8)} \qquad K_\infty(x) = \sum_{m=1}^{\infty} \frac{1}{m} \cos(\pi m x),$$

(3.8) is satisfied. A closed form expression for this sum is known: $K_\infty(x) = -\ln|\sin(\pi x/2)|$ (see, e.g., [10]), and the Lemma follows.

A.3. Proof of Lemma 3.2. Set $L = l_0/(2n)$, then for a defined in (A.5) we obtain $a = 2\pi mL = l_0\pi m/n$. Substitute (A.6) into (A.5) and observe that the Fourier coefficients $\hat{K}_L(m)$ depend on the variable $a = l_0\pi m/n$. Thus we can introduce the notation

$$\hat{K}(a) := \hat{K}_L(m),\ a = l_0\pi\frac{m}{n}.$$

In other words the Fourier series of $K_L(x)$ can be written in the form

$$K_L(x) = \sum_{m=-\infty}^{\infty} \hat{K}\left(l_0\pi\frac{m}{n}\right)\cos(\pi m x).$$

For $y = nx$, $y \in [-n, n]$ let

$$K^{(n)}(y) := \frac{1}{n}K_L(y/n) = \frac{1}{n}\sum_{m=-\infty}^{\infty} \hat{K}\left(l_0\pi\frac{m}{n}\right)\cos\left(\pi\frac{m}{n}y\right),$$

and $K^{(n)}(y) = 0$ for $|y| \geq n$. Set

$$\mathcal{K}_{l_0}(y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{K}\left(l_0\zeta\right)\cos(\zeta y)d\zeta,\ y \in \mathbb{R}.$$

Note that

$$\hat{K}\left(l_0\zeta\right) = \int_{-\infty}^{\infty} \mathcal{K}_{l_0}(y)\cos\left(\zeta y\right)dy,$$

provided $\mathcal{K}_{l_0}(y)$ is smooth and it decays sufficiently fast as $y \to \infty$. The function $\mathcal{K}_{l_0}(y)$ is exactly the limiting function mentioned in Lemma 3.2; that is, $K^{(n)}(y) = K_L(y/n)/n$ converges to $\mathcal{K}_{l_0}(y)$ as $n \to \infty$. Let us first verify properties of $\mathcal{K}_{l_0}$ described in Lemma 3.2 and then establish convergence. Direct calculations using (A.5) show that $\hat{K}(a) > 0$ for all physical choices of the Lamé constants, so that, by Bochner's theorem [18, Vol. 1], $\mathcal{K}_{l_0}$ is positive-definite. We verify (3.10) by applying the Paley–Wiener type Theorems [18, Vol. 1]. Indeed, from (A.6) it follows that the Fourier transform of $\mathcal{K}_{l_0}$ can be analytically extended into a finite strip $|\mathrm{Im}(a)| \leq c_2$ around the real axis provided there is no solution of the equation $a + \sinh(a) = 0$, $a \neq 0$, or, equivalently

$$\text{(A.9)}\quad \mathrm{Re}(a) = -\sinh(\mathrm{Re}(a))\cos(\mathrm{Im}(a)),\ \mathrm{Im}(a) = -\cosh(\mathrm{Re}(a))\sin(\mathrm{Im}(a)), a \neq 0.$$

There is no solution of the last equation in (A.9) at least in the strip $|\mathrm{Im}(a)| \leq \pi$. This implies the exponential decay (3.10) of $\mathcal{K}_{l_0}(y)$ by Paley–Wiener theorems. Finally, for every n consider the $2n$-periodization of $\mathcal{K}_{l_0}(y)$:

$$\mathcal{K}_{l_0}^{(n)}(y) = \sum_{k=-\infty}^{+\infty} \mathcal{K}_{l_0}(y + 2nk).$$

Since $\mathcal{K}_{l_0}(y)$ decays exponentially as $y \to \infty$, we have that $\mathcal{K}_{l_0}^{(n)}(y)$ is a smooth $2n$ periodic function and its Fourier coefficients

$$\begin{aligned}\hat{\mathcal{K}}_{l_0}^{(n)}(m) &= \frac{1}{n}\int_{-n}^{n} \mathcal{K}_{l_0}^{(n)}(y)\cos\left(\pi\frac{m}{n}y\right)dy \\ &= \frac{1}{n}\int_{-n}^{n}\left(\sum_{k=-\infty}^{+\infty}\mathcal{K}_{l_0}(y+2nk)\right)\cos\left(\pi\frac{m}{n}y\right)dy \\ = \frac{1}{n}\int_{-\infty}^{\infty}\mathcal{K}_{l_0}(y)\cos\left(\pi\frac{m}{n}y\right)dy &= \frac{1}{n}\hat{K}\left(l_0\pi\frac{m}{n}\right) = \frac{1}{n}\hat{K}_L(m).\end{aligned}$$

Thus $\mathcal{K}_{l_0}^{(n)}(y) = K_L(y/n)/n = K^{(n)}(y)$ on $[-n,n]$. Therefore

$$\begin{aligned}||K_L(x) - n\mathcal{K}_{l_0}(nx)||_{L_\infty([-1,1])} &= n||K^{(n)}(y) - \mathcal{K}_{l_0}(y)||_{L_\infty([-n,n])} \\ &= n||\mathcal{K}_{l_0}^{(n)}(y) - \mathcal{K}_{l_0}(y)||_{L_\infty([-n,n])} \\ &= n\left\|\sum_{k\neq 0}\mathcal{K}_{l_0}(y+2nk)\right\|_{L_\infty([-n,n])} \leq Cn\sum_{k=1}^{\infty}e^{-ckn} \\ &\leq Cne^{-cn} \to 0,\end{aligned}$$

as $n \to \infty$, and $C = C(l_0) > 0$ is independent of n.

A.4. Proof of Proposition 3.3. For any U and $\boldsymbol{\varepsilon}^{\mathrm{T}}$ we can estimate the error of the truncation (3.14) as

$$\begin{aligned}|W_n(U,\boldsymbol{\varepsilon}^{\mathrm{T}}) - W_n^{k_0}(U,\boldsymbol{\varepsilon}^{\mathrm{T}})| &\leq c\sum_{k>k_0}|\lambda_k(n)|\langle(s-\langle s\rangle)^2\rangle \\ &\leq c\exp(-k_0)\langle(s-\langle s\rangle)^2\rangle \leq c\exp(-k_0).\end{aligned}$$

Since

$$W_n^{k_0}(U,\boldsymbol{\varepsilon}_{k_0}^{\mathrm{T}}) = \min_{\tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}}\in\tilde{\Omega}_n(\theta)} W_n^{k_0}(U,\tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}}),$$

we have

$$W_n(U,\boldsymbol{\varepsilon}^{\mathrm{T}}) \leq \min_{\tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}}\in\tilde{\Omega}_n(\theta)} W_n(U,\tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}}) + c\exp(-k_0) = W_n(U,\boldsymbol{\theta}) + c\exp(-k_0).$$

By definition of the minimizer $W_n(U,\boldsymbol{\theta}) \leq W_n(U,\boldsymbol{\varepsilon}^{\mathrm{T}})$. This implies (3.15).

Appendix B. Proofs for section 5.

B.1. Verification of (5.11). For a given product measure $\tilde{\mu}$ we have

$$\int I_1 I_2 d\tilde{\mu} = \left(\int I_1 d\tilde{\mu}\right)^2, \quad \int s_1 I_2 d\tilde{\mu} = \left(\int s_1 d\tilde{\mu}\right)\left(\int I_1 d\tilde{\mu}\right),$$

because the events in the first and the second grains are independent and identically distributed. Hence

$$\begin{aligned}\int I_1 I_2 d\tilde{\mu} &= \left(\int_0^{\pi/4}\chi(\theta)d\rho(\theta)\right)^2 \\ &= f^2, \quad \int (U+s_1)I_2 d\tilde{\mu} = f\left(\int_0^{\pi/4}(U+s(\theta))\chi(\theta)d\rho(\theta) + (1-f)U\right).\end{aligned}$$

It gives

$$\begin{aligned}
&\int F(U,s_1,I_1,I_2)d\tilde{\mu}\\
&\quad=\int\left((U+s_1)^2-2(U+s_1)I_2+I_1I_2\right)d\tilde{\mu}\\
&\quad=\int_0^{\pi/4}(U+s(\theta))^2\chi(\theta)d\rho(\theta)+(1-f)U^2\\
&\qquad-2f\left(\int_0^{\pi/4}(U+s(\theta))\chi(\theta)d\rho(\theta)+(1-f)U\right)+f^2\\
&\quad=\int_0^{\pi/4}\left((U+s(\theta)-f)^2-f^2\right)\chi(\theta)d\rho(\theta)+(1-f)(U^2-2Uf)+f^2\\
&\quad=\int_0^{\pi/4}(U+s(\theta)-f)^2\chi(\theta)d\rho(\theta)+(U-f)^2(1-f),
\end{aligned}$$

where the last equality is obtained by noting the following identity:

$$f^2=f^2(1-f)+\int_0^{\pi/4}f^2\chi(\theta)d\rho(\theta).$$

B.2. Proof of Lemma 5.3. The proof is the direct evaluation and comparison of all possible scenarios in (5.12). Using (5.12), we simply consider four functions of U, q, q_α, and q_β:

$$\begin{aligned}
W_1&=(U-q_\alpha-q_\beta+\cos^2\alpha)^2q_\alpha\\
&+(U-q_\alpha-q_\beta+\cos^2\beta)^2q_\beta+(U-q_\alpha-q_\beta)^2(1-q_\alpha-q_\beta),\\
W_2&=(U-q_\alpha-q_\beta+\cos^2\alpha)^2q_\alpha\\
&+(U-q_\alpha-q_\beta+\sin^2\beta)^2q_\beta+(U-q_\alpha-q_\beta)^2(1-q_\alpha-q_\beta),\\
W_3&=(U-q_\alpha-q_\beta+\sin^2\alpha)^2q_\alpha\\
&+(U-q_\alpha-q_\beta+\cos^2\beta)^2q_\beta+(U-q_\alpha+q_\beta)^2(1-q_\alpha-q_\beta),\\
W_4&=(U-q_\alpha-q_\beta+\sin^2\alpha)^2q_\alpha\\
&+(U-q_\alpha-q_\beta+\sin^2\beta)^2q_\beta+(U-q_\alpha-q_\beta)^2(1-q_\alpha-q_\beta),
\end{aligned}$$

and compare their values for each fixed U and q, where q_α and q_β are in the range (5.13). It is easy to check that the minimum is always achieved for W_4; hence equation (5.14) is satisfied.

Appendix C. Proofs for section 6.

C.1. Proof of Lemma 6.4. Since the Fourier coefficients of a convolution equal the products of the Fourier coefficients, and since from (A.8) we know the Fourier coefficients of K_∞ are ≤ 1 in absolute value, and in view of Plancherel's theorem, we have

$$\int_{-1}^{1}\int_{-1}^{1}(s(x)-\langle s\rangle)(s(t)-\langle s\rangle)K_\infty(x-t)dxdt \le \int_{-1}^{1}(s(x)-\langle s\rangle)^2dx,$$

and $|s(x)|\le 1$, for any $\delta>0$, there exists $L_0>0$, so that for any $L\ge L_0$

$$\left|\frac{B}{L}\bar{W}_n(U,\boldsymbol{\varepsilon}^{\mathrm{T}})\right|\le\delta.$$

Hence for any $s=\varepsilon_{22}^{\mathrm{T}}$

$$\langle(U-f+s)^2\rangle < W_n(U,\boldsymbol{\varepsilon}^{\mathrm{T}}) \le \langle(U-f+s)^2\rangle+\delta.$$

It follows that

$$\min_{s,f}\langle(U-f+s)^2\rangle < \min_{s,f} W_n(U,\boldsymbol{\varepsilon}^{\mathrm{T}}) \le \min_{s,f}\langle(U-f+s)^2\rangle+\delta$$

for any sequence of angles $\boldsymbol{\theta}$. Applying Lemma 5.5 we complete the proof of (6.5). Equality (6.4) follows from direct computations as in Lemma 5.3.

C.2. Proof of Lemma 6.6. Suppose (6.18) and (6.19) do not hold. It means that there exists $\delta>0$ such that for every $\delta'>0$ there is a sequence of sets $A_{n_k}\in\sigma^{\theta}_{n_k}$, $\{n_k\}\to\infty$ with probability $\rho_{n_k}(A_{n_k})>2C>0$ such that for every fixed $\boldsymbol{\theta}\in A_{n_k}$ (or, equivalently, $(\chi_\alpha,\chi_\beta)\in A_{n_k}$) there is a (at least one) maximizer $\tilde{\chi}_\alpha$, $\tilde{\chi}_\beta$ of (6.17) such that

$$\min(e_1,e_2)>\delta, \tag{C.1}$$

where

(C.2)

$$e_1=\int_0^1(|\tilde{\chi}_\alpha-\chi^1_{q_\alpha}|\chi_\alpha+|\tilde{\chi}_\beta-\chi^1_{q_\beta}|\chi_\beta)dx,\ e_2=\int_0^1(|\tilde{\chi}_\alpha-\chi^2_{q_\alpha}|\chi_\alpha+|\tilde{\chi}_\beta-\chi^2_{q_\beta}|\chi_\beta)dx;$$

$\chi^i_{q_\alpha}$ and $\chi^i_{q_\beta}$ are defined in (6.14) and (6.15), respectively. Since $K(x)\in L^1[-1,1]$ is fixed and $\tilde{\chi}_\alpha\chi_\alpha$, $\tilde{\chi}_\beta\chi_\beta$ are uniformly bounded in $L^\infty[-1,1]$, for any $\delta_1>0$ there exists $h>0$ so that

$$\begin{aligned}
&\left|\int_{-1}^{1}\int_{-1}^{1}\varepsilon_\alpha(x)\varepsilon_\alpha(t)K(x-y)dxdt-\int_{-1}^{1}\int_{-1}^{1}\bar{\varepsilon}_\alpha(x)\bar{\varepsilon}_\alpha(t)K(x-y)|dxdt\right|<\delta_1,\\
&\left|\int_{-1}^{1}\int_{-1}^{1}\varepsilon_\alpha(x)\varepsilon_\beta(t)K(x-y)|dxdt-\int_{-1}^{1}\int_{-1}^{1}\bar{\varepsilon}_\alpha(x)\bar{\varepsilon}_\beta(t)K(x-y)|dxdt\right|<\delta_1,\\
&\left|\int_{-1}^{1}\int_{-1}^{1}\varepsilon_\beta(x)\varepsilon_\beta(t)K(x-y)|dxdt-\int_{-1}^{1}\int_{-1}^{1}\bar{\varepsilon}_\beta(x)\bar{\varepsilon}_\beta(t)K(x-y)|dxdt\right|<\delta_1,
\end{aligned} \tag{C.3}$$

where

$$\bar{\varepsilon}_\alpha(x)=\frac{1}{2h}\int_{-h}^{h}\bar{\varepsilon}_\alpha(x+t)dt,\ \bar{\varepsilon}_\beta(x)=\frac{1}{2h}\int_{-h}^{h}\bar{\varepsilon}_\beta(x+t)dt.$$

By the law of the large numbers, for any $\delta_2 > 0$ there is $\bar{A}_{n_k} \in A_{n_k}$, $\rho_{n_k}(A_{n_k}) > C$ so that for every $\boldsymbol{\theta} \in \bar{A}_{n_k}$ (or, equivalently, $(\chi_\alpha^{n_k}, \chi_\beta^{n_k}) \in \bar{A}_{n_k}$) and $n_k > N_0$, we have

$$\bar{\varepsilon}_\alpha(x) = q\chi_\alpha(x, \boldsymbol{\theta}) - q_\alpha(\boldsymbol{\theta}), \bar{\varepsilon}_\beta(x) = (1-q)\chi_\beta(x, \boldsymbol{\theta}) - q_\alpha(\boldsymbol{\theta}),$$

where (random in $\boldsymbol{\theta} \in \bar{A}_{n_k}$) functions $\chi_\alpha(x, \boldsymbol{\theta})$, $\chi_\beta(x, \boldsymbol{\theta})$ satisfy

$$0 \le \chi_\alpha(x, \boldsymbol{\theta}) \le 1 + \delta_2,\ 0 \le \chi_\beta(x, \boldsymbol{\theta}) \le 1 + \delta_2,$$

and $\min(e_1, e_2) > \delta - \delta_2$ where

$$\begin{aligned} e_1 &= \int_0^1 (|\chi_\alpha(x, \boldsymbol{\theta}) - \chi^1_{q_\alpha}| + |\chi_\beta(x, \boldsymbol{\theta}) - \chi^1_{q_\beta}|)dx,\ e_2 \\ &= \int_0^1 (|\chi_\alpha(x, \boldsymbol{\theta}) - \chi^2_{q_\alpha}| + |\chi_\beta(x, \boldsymbol{\theta}) - \chi^2_{q_\beta}|)dx. \end{aligned}$$

The classical Riesz rearrangement inequality (6.11) implies that when δ_1 and δ_2 are sufficiently small, there is $\delta' > 0$ so that for every $\boldsymbol{\theta} \in \bar{A}_{n_k}$

$$\begin{aligned} &\int_{-1}^1 \int_{-1}^1 (a\varepsilon_\alpha(x) + b\varepsilon_\beta(x))(a\varepsilon_\alpha(t) + b\varepsilon_\beta(t))K(x-t)dxdt \\ &\le \int_{-1}^1 \int_{-1}^1 (a\varepsilon^1_\alpha(x) + b\varepsilon^1_\beta(x))(a\varepsilon^1_\alpha(t) + b\varepsilon^1_\beta(t))K(x-t)dxdt - \delta', \end{aligned}$$

where

$$\varepsilon^1_\alpha(x) = \chi^1_{q_\alpha}\chi_\alpha,\ \varepsilon^1_\alpha(x) = \chi^1_{q_\beta}\chi_\beta.$$

Hence we have that $\tilde{\chi}_\alpha$ and $\tilde{\chi}_\beta$ with (C.1) and (C.2) cannot be maximizers of (6.17). This leads to contradiction with our assumption that (6.18) and (6.19) do not hold.

Acknowledgments. We thank Alexei Borodin and Omri Sarig for useful discussions. We are grateful to anonymous referees for their useful suggestions.

REFERENCES

[1] A. Baernstein, II, *Convolution and rearrangement on the circle*, Complex Variables Theory Appl., 12 (1989), pp. 33–37.

[2] K. Bhattacharya, B. Li, and M. Luskin, *The simply laminated microstructure in martensitic crystals that undergo a cubic-to-orthorhombic phase transformation*, Arch. Ration. Mech. Anal., 149 (1999), pp. 123–154.

[3] K. Bhattacharya and R. V. Kohn, *Elastic energy minimization and the recoverable strains of polycrystalline shape-memory materials*, Arch. Ration. Mech. Anal., 139 (1997), pp. 99–180.

[4] J. S. Bowles and J. K. Mackenzie, *The crystallography of martensite transformations* I, Acta Metall., 2 (1954), pp. 129–137.

[5] J. S. Bowles and J. K. Mackenzie, *The crystallography of martensite transformations* III. *Face-centered cubic to body-centered tetragonal transformations*, Acta Metall., 2 (1954), pp. 224–234.

[6] O. P. Bruno and G. H. Goldsztein, *A fast algorithm for the simulation of polycrystalline misfits: Martensitic transformations in two space dimensions*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 455 (1999), pp. 4245–4276.

[7] O. P. Bruno and G. H. Goldsztein, *Numerical simulation of martensitic transformations in two- and three-dimensional polycrystals*, The J. R. Willis 60th anniversary volume, J. Mech. Phys. Solids, 48 (2000), pp. 1175–1201.

[8] O. P. Bruno and G. H. Goldsztein, *A fast algorithm for the simulation of polycrystalline misfits*. II. *Martensitic transformations in three space dimensions*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 460 (2004), pp. 1613–1630.

[9] O. P. Bruno, F. Reitich, and P. Leo, *The overall elastic energy of polycrystalline martensitic solids*, J. Mech. Phys. Solids, 44 (1996), pp. 1051–1101.

[10] B. Demidovich, ed., *Problems in Mathematical Analysis*, Peace Publishers, Moscow, 1965.

[11] W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd ed., Vol. I, John Wiley & Sons, New York-London-Sydney, 1968.

[12] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, 2nd ed., Cambridge University Press, Cambridge, 1952.

[13] E. Hewitt and L. J. Savage, *Symmetric measures on Cartesian products*, Trans. Amer. Math. Soc., 80 (1955), pp. 470–501.

[14] E. H. Lieb and M. Loss, *Analysis*, 2nd ed., Graduate Studies in Mathematics, 14, American Mathematical Society, Providence, RI, 2001.

[15] J. K. Mackenzie and J. S. Bowles, *The crystallography of martensite transformations* II, Acta Metall., 2 (1954), pp. 138–147.

[16] G. W. Milton, *The Theory of Composites*, Cambridge University Press, Cambridge, 2002.

[17] R. L. Patterson and C. M. Wayman, *Internal twinning in ferrous martensites*, Acta Met., 12 (1964), pp. 1306–11.

[18] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, 2nd ed., Academic Press, Inc., New York, 1980.

[19] S. Timoshenko and J. N. Goodier, *Theory of Elasticity*, 2nd ed., McGraw-Hill Book Company, Inc., New York, Toronto, London, 1951.

[20] V. P. Smyshlyaev and J. R. Willis, *A "non-local" variational approach to the elastic energy minimization of martensitic polycrystals*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci. 454 (1998), pp. 1573–1613.

[21] M. S. Wechsler, D. S. Lieberman, and T. A. Read, *On the theory of the formation of martensite*, AIME Trans. J. Metals, 197 (1953), pp. 1503–1515.

[22] J. R. Willis, *Variational and related methods for the overall properties of composites*, Adv. Appl. Mech., 21 (1981), pp. 1–78.

SIAM J. MATH. ANAL.
Vol. 40, No. 4, pp. 1585–1620

THE PERIODIC UNFOLDING METHOD IN HOMOGENIZATION*

D. CIORANESCU†, A. DAMLAMIAN‡, AND G. GRISO§

Abstract. The periodic unfolding method was introduced in 2002 in [Cioranescu, Damlamian, and Griso, *C.R. Acad. Sci. Paris, Ser.* 1, 335 (2002), pp. 99–104] (with the basic proofs in [*Proceedings of the Narvik Conference* 2004, GAKUTO Internat. Ser. Math. Sci. Appl. 24, Gakkōtosho, Tokyo, 2006, pp. 119–136]). In the present paper we go into all the details of the method and include complete proofs, as well as several new extensions and developments. This approach is based on two distinct ideas, each leading to a new ingredient. The first idea is the change of scale, which is embodied in the unfolding operator. At the expense of doubling the dimension, this allows one to use standard weak or strong convergence theorems in L^p spaces instead of more complicated tools (such as two-scale convergence, which is shown to be merely the weak convergence of the unfolding; cf. Remark 2.15). The second idea is the separation of scales, which is implemented as a macro-micro decomposition of functions and is especially suited for the weakly convergent sequences of Sobolev spaces. In the framework of this method, the proofs of most periodic homogenization results are elementary. The unfolding is particularly well-suited for multiscale problems (a simple backward iteration argument suffices) and for precise corrector results without extra regularity on the data. A list of the papers where these ideas appeared, at least in some preliminary form, is given with a discussion of their content. We also give a list of papers published since the publication [Cioranescu, Damlamian, and Griso, *C.R. Acad. Sci. Paris, Ser.* 1, 335 (2002), pp. 99–104], and where the unfolding method has been successfully applied.

Key words. homogenization, periodic unfolding, multiscale problems

AMS subject classifications. 49J45, 35B27, 74Q05

DOI. 10.1137/080713148

1. Introduction.

The notion of two-scale convergence was introduced in 1989 by Nguetseng in [58], further developed by Allaire in [1] and by Lukkassen, Nguetseng, and Wall in [55] with applications to periodic homogenization. It was generalized to some multiscale problems by Ene and Saint Jean Paulin in [38], Allaire and Briane in [2], Lions et al. in [52] and Lukkassen, Nguetseng, and Wall in [55].

In 1990, Arbogast, Douglas, and Hornung defined a "dilation" operator in [5] to study homogenization for a periodic medium with double porosity. This technique was used again in [16], [3], [4], [48], [49], [50], [51], [54], [20], [21], [22], and [23].

In [24], we expanded on this idea and presented a general and quite simple approach for classical or multiscale periodic homogenization, under the name of "unfolding method." Originally restricted to the case of domains consisting of a union of ε-cells, it was extended to general domains (see the survey of Damlamian [34]). In the present work, we give a complete presentation of this method, including all of the proofs, as well as several new extensions and developments. The relationship of the papers listed above with our work is discussed at the end of this introduction.

The periodic unfolding method is essentially based on two ingredients. The first one is the unfolding operator $\mathcal{T}_\varepsilon$ (similar to the dilation operator), defined in section 2,

*Received by the editors January 1, 2008; accepted for publication (in revised form) May 5, 2008; published electronically November 26, 2008.

http://www.siam.org/journals/sima/40-4/71314.html

†Corresponding author. Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, Boîte courrier 187, 4 Place Jussieu, 75252 Paris Cedex 05, France (cioran@ann.jussieu.fr).

‡Université Paris-Est, Laboratoire d'Analyse et de Mathématiques Appliquées, CNRS UMR 8050, Centre Multidisciplinaire de Créteil, 94010, Créteil, Cedex, France (damla@univ-paris12.fr).

§Corresponding author. Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, Boîte courrier 187, 4 Place Jussieu, 75252 Paris Cedex 05, France (griso@ann.jussieu.fr).

where its properties are investigated. Let Ω be a bounded open set, and Y a reference cell in $\mathbb{R}^n$. By definition, the operator $\mathcal{T}_\varepsilon$ associates to any function v in $L^p(\Omega)$, a function $\mathcal{T}_\varepsilon(v)$ in $L^p(\Omega \times Y)$. An immediate (and interesting) property of $\mathcal{T}_\varepsilon$ is that it enables one to transform any integral over Ω in an integral over $\Omega \times Y$. Indeed, by Proposition 2.6 below

$$\int_\Omega w(x)\, dx \sim \frac{1}{|Y|} \int_{\Omega \times Y} \mathcal{T}_\varepsilon(w)(x,y)\, dx\, dy \quad \forall w \in L^1(\Omega). \tag{1.1}$$

Proposition 2.14 shows that the two-scale convergence in the $L^p(\Omega)$-sense of a sequence of functions $\{v_\varepsilon\}$ is equivalent to the weak convergence of the sequence of unfolded functions $\{\mathcal{T}_\varepsilon(v_\varepsilon)\}$ in $L^p(\Omega \times Y)$. Thus, the two-scale convergence in Ω is reduced to a mere weak convergence in $L^p(\Omega \times Y)$, which conceptually simplifies proofs.

In section 2 are also introduced a local average operator $\mathcal{M}_\varepsilon$ and an averaging operator $\mathcal{U}_\varepsilon$, the latter being, in some sense, the inverse of the unfolding operator $\mathcal{T}_\varepsilon$.

The second ingredient of the periodic unfolding method consists of separating the characteristic scales by decomposing every function φ belonging to $W^{1,p}(\Omega)$ in two parts. In section 3 it is achieved by using the local average. In section 4, the original proof of this scale-splitting, inspired by the finite element method (FEM), is given. The confrontation of the two methods of sections 3 and 4 is interesting in itself (Theorem 3.5 and Proposition 4.8). In both approaches, φ is written as $\varphi = \varphi_1^\varepsilon + \varepsilon\varphi_2^\varepsilon$, where φ_1^ε is a macroscopic part designed not to capture the oscillations of order ε (if there are any), while the microscopic part φ_2^ε is designed to do so. The main result states that, from any bounded sequence $\{w^\varepsilon\}$ in $W^{1,p}(\Omega)$, weakly convergent to some w, one can always extract a subsequence (still denoted $\{w^\varepsilon\}$) such that $w^\varepsilon = w_1^\varepsilon + \varepsilon w_2^\varepsilon$, with

$$\begin{aligned}
&\text{(i)} \quad w_1^\varepsilon \rightharpoonup w \quad \textit{weakly in } W^{1,p}(\Omega),\\
&\text{(ii)} \quad \mathcal{T}_\varepsilon(w^\varepsilon) \rightharpoonup w \quad \textit{weakly in } L^p(\Omega; W^{1,p}(Y)),\\
&\text{(iii)} \quad \mathcal{T}_\varepsilon(w_2^\varepsilon) \rightharpoonup \widehat{w} \quad \textit{weakly in } L^p(\Omega; W^{1,p}(Y)),\\
&\text{(iv)} \quad \mathcal{T}_\varepsilon(\nabla w^\varepsilon) \rightharpoonup \nabla w + \nabla_y \widehat{w} \quad \textit{weakly in } L^p(\Omega \times Y),
\end{aligned} \tag{1.2}$$

where $\widehat{w}$ belongs to $L^p(\Omega; W^{1,p}_{per}(Y))$.

In section 5 we apply the periodic unfolding method to a classical periodic homogenization problem. We point out that, in the framework of this method, the proof of the homogenization result is elementary. It relies essentially on formula (1.1), on the properties of $\mathcal{T}_\varepsilon$, and on convergences (1.2). It applies directly for both homogeneous Dirichlet or Neumann boundary conditions without hypothesis on the regularity of $\partial\Omega$. For nonhomogenous boundary conditions (or for Robin-type condition), some regularity of $\partial\Omega$ is required for the problem to make sense, in which case the method applies also directly (see Remark 5.12).

Section 6 is devoted to a corrector result, which holds without any additional regularity on the data (contrary to all previous proofs; see [11], [30], and [59]). This result follows from the use of the averaging operator $\mathcal{U}_\varepsilon$. The idea of using averages to improve corrector results first appeared in Dal Maso and Defranceschi [33]. We also give some error estimates and a new corrector result for the case of domains with

a smooth boundary (obtained by Griso in [42], [43], [44], and [45]). These results are explicitely connected to the unfolding method and improve on known classical ones (see [11] and [59]).

The periodic unfolding method is particularly well-suited for the case of multiscale problems. This is shown in section 7 by a simple backward iteration argument. This problem has a long history; one of the first papers on the subject is due to Bruggeman [19]. Its mathematical treatment by homogenization goes back to the book of Bensoussan, Lions, and Papanicolaou [11], where for this problem, the method of asymptotic expansions is used. For more recent references of multiscale homogenization and its applications, we refer to the books of Braides and Defranceschi [17], Milton [57], and the articles by Damlamian and Donato [35], Lukkassen and Milton [54], Lukkassen [53], Braides and Lukkassen [18], Babadjian and Baía [6], and Barchiesi [8].

The final section gives a list of papers where the method has been successfully applied since the publication of [24].

To conclude, let us turn back to the papers quoted at the beginning of this introduction and point out their relationships with our results. The dilation operation from Arbogast, Douglas, and Hornung [5] was defined in a domain which is an exact union of εY-cells. It consists in a change of variables, similar to that used in Definition 2.1 below. By this operation, any integral on Ω can be written as an integral over $\Omega \times Y$. Some elementary properties of the dilation operator in the space L^2 were also contained in Lemma 2 of [5].

The same dilation operator was used by Bourgeat, Luckhaus, and Mikelic in [16] under the name of "periodic modulation." Proposition 4.6 of [16] showed that if a sequence two-scale converges and its periodic modulation converges weakly, they have the same limit.

In the context of two-scale convergence, Allaire and Conca [3] defined a pair of extension and projection operators (suited to Bloch decompositions) which are adjoint of each other. They are similar to our operators $\mathcal{T}_\varepsilon$ and $\mathcal{U}_\varepsilon$ and the equivalent of property (2.12) and Proposition 2.18(ii) below, are proved in Lemma 4.2 of [3]. These properties were exploited by Allaire, Conca, and Vanninathan in [4] for a general bounded domain by extending all functions by zero on its complement.

In [48], Lenczner used the dilation operator (here called "two-scale transformation") in order to treat the homogenization of discrete electrical networks (by nature, the domain is a union of ε-cells). The convergence of the two-scale transform is called two-scale convergence (this would be confusing except that it was shown to be equivalent to the original two-scale convergence). As an aside, a convergence similar to (1.2)(iv) was also treated. In Lenczner and Mercier [49], Lenczner and Senouci-Bereksi [50], and Lenczner, Kader, and Perrier [51], this theory was applied to periodic electrical networks.

Finally, Casado Díaz and Luna-Laynez [21], Casado Díaz, Luna-Laynez, and Martin [22] and [23] used the dilation operator in the case of reticulated structures. In this framework, they obtained the equivalent of (3.7)(i) of Theorem 3.5 below.

2. Unfolding in L^p-spaces.

2.1. The unfolding operator $\mathcal{T}_\varepsilon$. In $\mathbb{R}^n$, let Ω be an open set and Y a reference cell (e.g., $]0,1[^n$, or more generally, a set having the paving property, with respect to a basis $(b_1, \dots, b_n)$ defining the periods).

By analogy with the notation in the one-dimensional case, for $z \in \mathbb{R}^n$, $[z]_Y$ denotes the unique integer combination $\sum_{j=1}^n k_j b_j$ of the periods such that $z - [z]_Y$ belongs

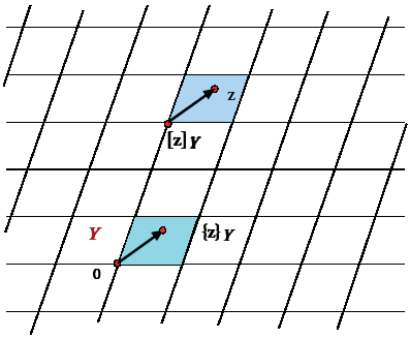


FIG. 1. *Definition of* $[z]_Y$ *and* $\{z\}_Y$.

to Y, and set

$$\{z\}_Y = z - [z]_Y \in Y \quad \text{a.e. for} \quad z \in \mathbb{R}^n.$$

Then for each $x \in \mathbb{R}^n$, one has

$$x = \varepsilon\Big(\Big[\frac{x}{\varepsilon}\Big]_Y + \Big\{\frac{x}{\varepsilon}\Big\}_Y\Big) \quad \text{a.e. for} \quad x \in \mathbb{R}^n \text{ (See Figure 1).}$$

We use the following notations:

$$\begin{cases} \Xi_\varepsilon = \big\{\xi \in \mathbb{Z}^N, \varepsilon(\xi + Y) \subset \Omega\big\}, \\ \widehat{\Omega}_\varepsilon = \text{interior}\Big\{\bigcup_{\xi\in\Xi_\varepsilon} \varepsilon\big(\xi + \overline{Y}\big)\Big\}, \\ \Lambda_\varepsilon = \Omega \setminus \widehat{\Omega}_\varepsilon. \end{cases} \tag{2.1}$$

The set $\widehat{\Omega}_\varepsilon$ is the largest union of $\varepsilon(\xi + \overline{Y})$ cells ($\xi \in \mathbb{Z}^n$) included in Ω, while Λ_ε is the subset of Ω containing the parts from $\varepsilon\big(\xi + \overline{Y}\big)$ cells intersecting the boundary $\partial\Omega$ (see Figure 2).

DEFINITION 2.1. *For ϕ Lebesgue-measurable on Ω, the unfolding operator $\mathcal{T}_\varepsilon$ is defined as follows:*

$$\mathcal{T}_\varepsilon(\phi)(x,y) = \begin{cases} \phi\Big(\varepsilon\Big[\frac{x}{\varepsilon}\Big]_Y + \varepsilon y\Big) & \text{a.e. for} \quad (x,y) \in \widehat{\Omega}_\varepsilon \times Y, \\ 0 & \text{a.e. for} \quad (x,y) \in \Lambda_\varepsilon \times Y. \end{cases}$$

Observe that the function $\mathcal{T}_\varepsilon(\phi)$ is Lebesgue-measurable on $\Omega \times Y$ and vanishes for x outside of the set $\widehat{\Omega}_\varepsilon$.

As in classical periodic homogenization, two different scales appear in the definition of $\mathcal{T}_\varepsilon$: the "macroscopic" scale x gives the position of a point in the domain Ω, while the "microscopic" scale y $(= x/\varepsilon)$ gives the position of a point in the cell Y. The unfolding operator doubles the dimension of the space and puts all of the oscillations in the second variable, in this way separating the two scales (see Figures 3, 4 and Figures 5, 6).

The following property of $\mathcal{T}_\varepsilon$ is a simple consequence of Definition 2.1 for v and w Lebesgue-measurable; it will be used extensively:

$$\mathcal{T}_\varepsilon(vw) = \mathcal{T}_\varepsilon(v)\, \mathcal{T}_\varepsilon(w). \tag{2.2}$$

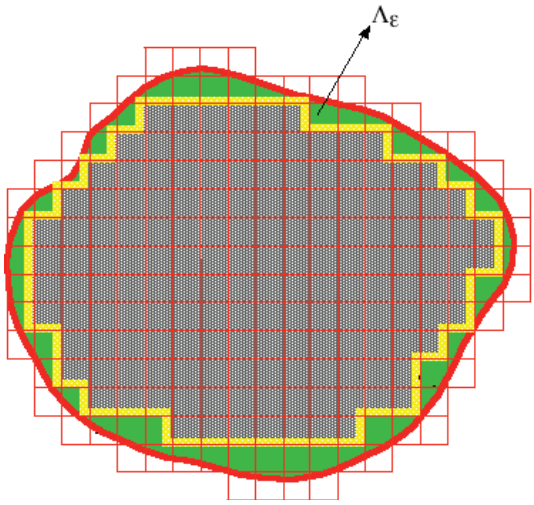


FIG. 2. *The domains* $\widehat{\Omega}_\varepsilon$ *and* Λ_ε.

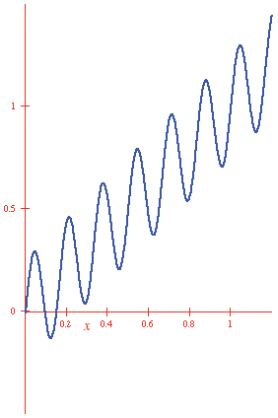


FIG. 3. $f_\varepsilon(x) = \frac{1}{4}\sin(2\pi\frac{x}{\varepsilon}) + x;\ \varepsilon = \frac{1}{6}$.

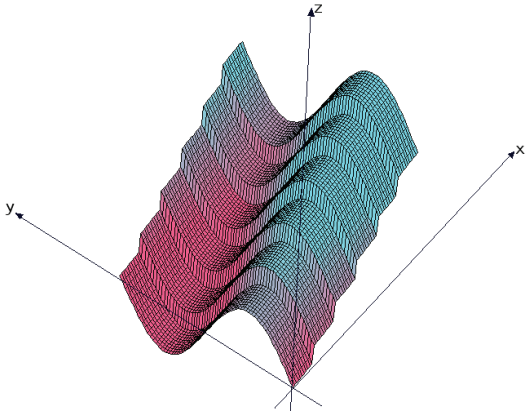


FIG. 4. $\mathcal{T}_\varepsilon(f_\varepsilon)$.

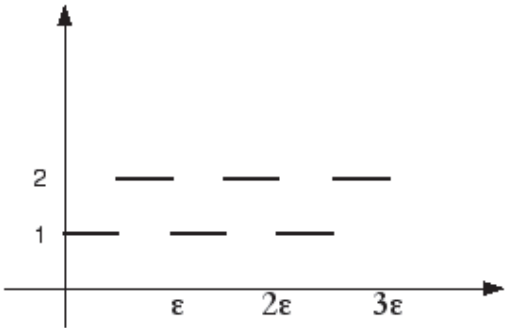


FIG. 5. $f_\varepsilon = f(\{\frac{x}{\varepsilon}\}_Y)$.

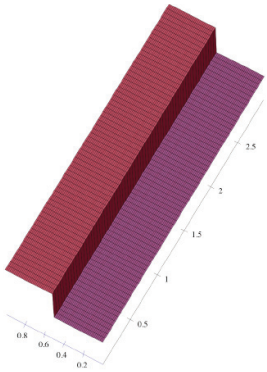


FIG. 6. $\mathcal{T}_\varepsilon(f_\varepsilon)$.

Another simple consequence of Definition 2.1 is the following result concerning highly oscillating functions.

PROPOSITION 2.2. *For f measurable on Y, extended by Y-periodicity to the whole of $\mathbb{R}^n$, define the sequence $\{f_\varepsilon\}$ by*

$$f_\varepsilon(x) = f\left(\frac{x}{\varepsilon}\right) \quad \text{a.e. for } x \in \mathbb{R}^n. \tag{2.3}$$

Then

$$\mathcal{T}_\varepsilon(f_\varepsilon|_\Omega)(x,y) = \begin{cases} f(y) & \text{a.e. for } (x,y) \in \widehat{\Omega}_\varepsilon \times Y, \\ 0 & \text{a.e. for } (x,y) \in \Lambda_\varepsilon \times Y. \end{cases}$$

If f belongs to $L^p(Y)$, $p \in [1, +\infty[$, and if Ω is bounded,

$$\mathcal{T}_\varepsilon(f_\varepsilon|_\Omega) \to f \quad \text{strongly in } L^p(\Omega \times Y). \tag{2.4}$$

Remark 2.3. An equivalent way to define f_ε in (2.3) is to take simply $f_\varepsilon(x) = f(\{\frac{x}{\varepsilon}\}_Y)$. For example, with

$$f(y) = \begin{cases} 1 & \text{for } y \in \left(0, {}^1/_2\right), \\ 2 & \text{for } y \in \left({}^1/_2, 1\right), \end{cases}$$

f_ε is the highly oscillating periodic function, with period ε from Figure 5.

Remark 2.4. Let f in $L^p(Y)$, $p \in [1, +\infty[$, and f_ε be defined by (2.3). It is well-known that $\{f_\varepsilon|_\Omega\}$ converges weakly in $L^p(\Omega)$ to the mean value of f on Y, and not strongly unless f is a constant (see Remark 2.11 below).

The next two results, essential in the study of the properties of the unfolding operator, are also straightforward from Definition 2.1.

PROPOSITION 2.5. *For $p \in [1, +\infty[$, the operator $\mathcal{T}_\varepsilon$ is linear and continuous from $L^p(\Omega)$ to $L^p(\Omega \times Y)$. For every ϕ in $L^1(\Omega)$ and w in $L^p(\Omega)$,*

(i) $\displaystyle \frac{1}{|Y|}\int_{\Omega\times Y} \mathcal{T}_\varepsilon(\phi)(x,y)\,dx\,dy = \int_\Omega \phi(x)\,dx - \int_{\Lambda_\varepsilon} \phi(x)\,dx = \int_{\widehat{\Omega}_\varepsilon} \phi(x)\,dx,$

(ii) $\displaystyle \frac{1}{|Y|}\int_{\Omega\times Y} |\mathcal{T}_\varepsilon(\phi)|\,dxdy \le \int_\Omega |\phi|\,dx,$

(iii) $\displaystyle \left|\int_\Omega \phi\,dx - \frac{1}{|Y|}\int_{\Omega\times Y} \mathcal{T}_\varepsilon(\phi)\,dxdy\right| \le \int_{\Lambda_\varepsilon} |\phi|\,dx,$

(iv) $\|\mathcal{T}_\varepsilon(w)\|_{L^p(\Omega\times Y)} = |\,Y\,|^{\frac{1}{p}}\,\|w\,1_{\widehat{\Omega}_\varepsilon}\|_{L^p(\Omega)} \le |\,Y\,|^{\frac{1}{p}}\,\|w\|_{L^p(\Omega)}.$

Proof. Recalling Definition 2.2 of $\widehat{\Omega}_\varepsilon$, one has

$$\begin{aligned}\frac{1}{|Y|}\int_{\Omega\times Y} \mathcal{T}_\varepsilon(\phi)(x,y)\,dx\,dy &= \frac{1}{|Y|}\int_{\widehat{\Omega}_\varepsilon\times Y} \mathcal{T}_\varepsilon(\phi)(x,y)\,dx\,dy \\ &= \frac{1}{|Y|}\sum_{\xi\in\Xi_\varepsilon}\int_{(\varepsilon\xi+\varepsilon Y)\times Y} \mathcal{T}_\varepsilon(\phi)(x,y)\,dx\,dy.\end{aligned}$$

On each $(\varepsilon\xi+\varepsilon Y)\times Y$, by definition, $\mathcal{T}_\varepsilon(\phi)(x,y) = \phi(\varepsilon\xi+\varepsilon y)$ is constant in x. Hence, each integral in the sum on the right-hand side successively equals

$$\begin{aligned}\int_{(\varepsilon\xi+\varepsilon Y)\times Y} \mathcal{T}_\varepsilon(\phi)(x,y)\,dx\,dy &= |\varepsilon\xi+\varepsilon Y|\int_Y \phi(\varepsilon\xi+\varepsilon y)\,dy \\ &= \varepsilon^n|Y|\int_Y \phi(\varepsilon\xi+\varepsilon y)\,dy = |Y|\int_{(\varepsilon\xi+\varepsilon Y)} \phi(x)\,dx.\end{aligned}$$

By summing over Ξ_ε, the right-hand side becomes $\int_{\widehat{\Omega}_\varepsilon} \phi(x)\,dx$, which gives the result. ∎

Property (iii) in Proposition 2.5 shows that any integral of a function on Ω is "almost equivalent" to the integral of its unfolded on $\Omega \times Y$; the "integration defect" arises only from the cells intersecting the boundary $\partial\Omega$ and is controlled by its integral over Λ_ε.

The next proposition, which we call **unfolding criterion for integrals** (u.c.i.), is a very useful tool when treating homogenization problems.

PROPOSITION 2.6 (u.c.i.). *If $\{\phi_\varepsilon\}$ is a sequence in $L^1(\Omega)$ satisfying*

$$\int_{\Lambda_\varepsilon} |\phi_\varepsilon|\,dx \to 0,$$

then

$$\int_\Omega \phi_\varepsilon\,dx - \frac{1}{|Y|}\int_{\Omega\times Y} \mathcal{T}_\varepsilon(\phi_\varepsilon)\,dxdy \to 0.$$

Based on this result, we introduce the following notation.

Notation. If $\{w_\varepsilon\}$ is a sequence satisfying u.c.i., we write

$$\int_\Omega w_\varepsilon dx \overset{\mathcal{T}_\varepsilon}{\simeq} \frac{1}{|Y|}\int_{\Omega\times Y} \mathcal{T}_\varepsilon(w_\varepsilon)\, dxdy.$$

PROPOSITION 2.7. *Let $\{u_\varepsilon\}$ be a bounded sequence in $L^p(\Omega)$, with $p \in]1,+\infty]$ and $v \in L^{p'}(\Omega)$ $(1/p + 1/p' = 1)$, then*

$$\int_\Omega u_\varepsilon v\, dx \overset{\mathcal{T}_\varepsilon}{\simeq} \frac{1}{|Y|}\int_{\Omega\times Y} \mathcal{T}_\varepsilon(u_\varepsilon)\mathcal{T}_\varepsilon(v)\, dxdy. \tag{2.5}$$

Suppose $\partial\Omega$ is bounded. Let $\{u_\varepsilon\}$ be a bounded sequence in $L^p(\Omega)$ and $\{v_\varepsilon\}$ a bounded sequence in $L^q(\Omega)$, with $1/p + 1/q < 1$, then

$$\int_\Omega u_\varepsilon v_\varepsilon dx \overset{\mathcal{T}_\varepsilon}{\simeq} \frac{1}{|Y|}\int_{\Omega\times Y} \mathcal{T}_\varepsilon(u_\varepsilon)\mathcal{T}_\varepsilon(v_\varepsilon)\, dxdy. \tag{2.6}$$

Proof. Observe that $1_{\Lambda_\varepsilon}(x) \to 0$ for all $x \in \Omega$. Consequently, by the Lebesgue dominated convergence theorem, one gets $\int_{\Lambda_\varepsilon} |v|^{p'}\, dx \to 0$, and then by the Hölder inequality, $\int_{\Lambda_\varepsilon} |u_\varepsilon v|\, dx \to 0$. This proves (2.5). If $\partial\Omega$ is bounded, then one immediately has $|\Lambda_\varepsilon| \to 0$ when $\varepsilon \to 0$, and this implies (2.6). ∎

We now investigate the convergence properties related to the unfolding operator when $\varepsilon \to 0$. For ϕ uniformly continuous on Ω, with modulus of continuity m_ϕ, it is easy to see that

$$\sup_{x\in\widehat{\Omega}_\varepsilon, y\in Y} |\mathcal{T}_\varepsilon(\phi)(x,y) - \phi(x)| \le m_\phi(\varepsilon).$$

So, as ε goes to zero, even though $\mathcal{T}_\varepsilon(\phi)$ is not continuous, it converges to ϕ uniformly on any open set strongly included in Ω. By density, and making use of Proposition 2.5, further convergence properties can be expressed using the mean value of a function defined on $\Omega \times Y$.

DEFINITION 2.8. *The mean value operator $\mathcal{M}_Y : L^p(\Omega \times Y) \mapsto L^p(\Omega)$ for $p \in [1,+\infty]$, is defined as follows:*

$$\mathcal{M}_Y(\Phi)(x) = \frac{1}{|Y|}\int_Y \Phi(x,y)\, dy \quad \text{a.e. for} \quad x \in \Omega. \tag{2.7}$$

Observe that an immediate consequence of this definition is the estimate

$$\|\mathcal{M}_Y(\Phi)\|_{L^p(\Omega)} \le |Y|^{-\frac{1}{p}} \|\Phi\|_{L^p(\Omega\times Y)} \quad \text{for every } \Phi \in L^p(\Omega \times Y).$$

PROPOSITION 2.9. *Let p belong to $[1,+\infty[$.*

(i) *For $w \in L^p(\Omega)$,*

$$\mathcal{T}_\varepsilon(w) \to w \quad \text{strongly in } L^p(\Omega \times Y).$$

(ii) *Let $\{w_\varepsilon\}$ be a sequence in $L^p(\Omega)$ such that*

$$w_\varepsilon \to w \quad \text{strongly in } L^p(\Omega).$$

Then

$$\mathcal{T}_\varepsilon(w_\varepsilon) \to w \quad \text{strongly in } L^p(\Omega \times Y).$$

(iii) *For every relatively weakly compact sequence* $\{w_\varepsilon\}$ *in* $L^p(\Omega)$*, the corresponding* $\{\mathcal{T}_\varepsilon(w_\varepsilon)\}$ *is relatively weakly compact in* $L^p(\Omega \times Y)$*. Furthermore, if*

$$\mathcal{T}_\varepsilon(w_\varepsilon) \rightharpoonup \widehat{w} \quad \text{weakly in } L^p(\Omega \times Y),$$

then

$$w_\varepsilon \rightharpoonup \mathcal{M}_Y(\widehat{w}) \quad \text{weakly in } L^p(\Omega).$$

(iv) *If* $\mathcal{T}_\varepsilon(w_\varepsilon) \rightharpoonup \widehat{w}$ *weakly in* $L^p(\Omega \times Y)$*, then*

$$\|\widehat{w}\|_{L^p(\Omega\times Y)} \le \liminf_{\varepsilon\to 0} |Y|^{\frac{1}{p}} \|w_\varepsilon\|_{L^p(\Omega)}. \tag{2.8}$$

(v) *Suppose* $p > 1$*, and let* $\{w_\varepsilon\}$ *be a bounded sequence in* $L^p(\Omega)$*. Then, the following assertions are equivalent:*

(a) $\mathcal{T}_\varepsilon(w_\varepsilon) \rightharpoonup \widehat{w}$ *weakly in* $L^p(\Omega \times Y)$ *and* $\limsup_{\varepsilon\to 0} |Y|^{\frac{1}{p}} \|w_\varepsilon\|_{L^p(\Omega)} \le \|\widehat{w}\|_{L^p(\Omega\times Y)}$,

(b) $\mathcal{T}_\varepsilon(w_\varepsilon) \to \widehat{w}$ *strongly in* $L^p(\Omega \times Y)$ *and* $\displaystyle\int_{\Lambda_\varepsilon} |w_\varepsilon|^p \, dx \to 0$.

Proof. (i) The result is obvious for any $w \in \mathcal{D}(\Omega)$. If $w \in L^p(\Omega)$, let $\phi \in \mathcal{D}(\Omega)$. Then, by using (iv) from Proposition 2.5,

$$\begin{aligned}\|\mathcal{T}_\varepsilon(w) - w\|_{L^p(\Omega\times Y)} &= \|\mathcal{T}_\varepsilon(w-\phi) + \big(\mathcal{T}_\varepsilon(\phi) - \phi\big) + (\phi - w)\|_{L^p(\Omega\times Y)} \\ &\le 2|Y|^{\frac{1}{p}} \|w-\phi\|_{L^p(\Omega)} + \|\mathcal{T}_\varepsilon(\phi) - \phi\|_{L^p(\Omega\times Y)},\end{aligned}$$

hence,

$$\limsup_{\varepsilon\to 0} \|\mathcal{T}_\varepsilon(w) - w\|_{L^p(\Omega\times Y)} \le 2|Y|^{\frac{1}{p}} \|w-\phi\|_{L^p(\Omega)},$$

from which statement (i) follows by density.

(ii) The following estimate, a consequence of Proposition 2.5(iv), gives the result

$$\|\mathcal{T}_\varepsilon(w_\varepsilon) - \mathcal{T}_\varepsilon(w)\|_{L^p(\Omega\times Y)} \le |\ Y\ |^{\frac{1}{p}} \|w_\varepsilon - w\|_{L^p(\Omega)} \quad \forall w \in L^p(\Omega).$$

(iii) For $p \in]1, +\infty[$, by Proposition 2.5(iv), boundedness is preserved by $\mathcal{T}_\varepsilon$. Suppose that $\mathcal{T}_\varepsilon(w_\varepsilon) \rightharpoonup \widehat{w}$ weakly in $L^p(\Omega \times Y)$, and let $\psi \in L^{p'}(\Omega)$. From Proposition 2.7,

$$\int_\Omega w_\varepsilon(x)\, \psi(x)\, dx \overset{\mathcal{T}_\varepsilon}{\sim} \frac{1}{|Y|} \int_{\Omega\times Y} \mathcal{T}_\varepsilon(w_\varepsilon)(x,y)\, \mathcal{T}_\varepsilon(\psi)(x,y)\, dx\, dy.$$

In view of (i), one can pass to the limit in the right-hand side to obtain

$$\lim_{\varepsilon\to 0} \int_\Omega w_\varepsilon(x)\, \psi(x)\, dx = \int_\Omega \left\{ \frac{1}{|Y|} \int_Y \widehat{w}(x,y)\, dy \right\} \psi(x)\, dx.$$

For $p = 1$, one uses the extra property satisfied by weakly convergent sequences in $L^1(\Omega)$, in the form of the De La Vallée–Poussin criterion (which is equivalent to

relative weak compactness): there exists a continuous convex function $\Phi : \mathbb{R}^+ \mapsto \mathbb{R}^+$ such that

$$\lim_{t\to+\infty} \frac{\Phi(t)}{t} = +\infty, \quad \text{and the set } \left\{ \int_\Omega \big(\Phi \circ |w_\varepsilon|\big)(x)\, dx \right\} \text{ is bounded.}$$

Unfolding the last integral shows that

$$\left\{ \int_{\Omega\times Y} \big(\Phi \circ |\mathcal{T}_\varepsilon(w_\varepsilon)|\big)(x,y)\, dxdy \right\} \text{ is bounded,}$$

which completes the proof of weak compactness of $\{\mathcal{T}_\varepsilon(w_\varepsilon)\}$ in $L^1(\Omega \times Y)$ in the case of Ω with finite measure. For the case where the measure of Ω is not finite, a similar argument shows that the equiintegrability at infinity of the sequence $\{w_\varepsilon\}$ carries over to $\{\mathcal{T}_\varepsilon(w_\varepsilon)\}$.

If $\mathcal{T}_\varepsilon(w_\varepsilon) \rightharpoonup \widehat{w}$ weakly in $L^1(\Omega \times Y)$, let ψ be in $\mathcal{D}(\Omega)$. For ε sufficiently small, one has

$$\int_\Omega w_\varepsilon(x)\, \psi(x)\, dx = \frac{1}{|Y|} \int_{\Omega\times Y} \mathcal{T}_\varepsilon(w_\varepsilon)(x,y)\, \mathcal{T}_\varepsilon(\psi)(x,y)\, dx\, dy.$$

In view of (i), one can pass to the limit in the right-hand side to obtain

$$\lim_{\varepsilon\to 0} \int_\Omega w_\varepsilon(x)\, \psi(x)\, dx = \int_\Omega \left\{ \frac{1}{|Y|} \int_Y \widehat{w}(x,y)\, dy \right\} \psi(x)\, dx.$$

(iv) Inequality (2.8) is a simple consequence of Proposition 2.5(ii).

(v) Proposition 2.5(i) applied to the function $|w_\varepsilon|^p$ gives

$$\frac{1}{|Y|} \|\mathcal{T}_\varepsilon(w_\varepsilon)\|^p_{L^p(\Omega\times Y)} + \int_{\Lambda_\varepsilon} |w_\varepsilon|^p\, dx = \|w_\varepsilon\|^p_{L^p(\Omega)}.$$

This identity implies the required equivalence. □

COROLLARY 2.10. *Let f be in $L^p(Y)$, $p \in [1, +\infty[$, and $\{f_\varepsilon\}$ be the sequence defined by* (2.3). *Then*

$$f_\varepsilon|_\Omega \rightharpoonup \mathcal{M}_Y(f) \quad \textit{weakly in } L^p(\Omega). \tag{2.9}$$

Proof. Proposition 2.2 gives the strong (hence weak) convergenge of $\{\mathcal{T}_\varepsilon(f_\varepsilon|_\Omega)\}$ to f in $L^p(\Omega \times Y)$. Convergence (2.9) follows from Proposition 2.9(iii).[1] □

Remark 2.11. In general, in the case where Λ_ε is not null set (for every ε), the strong (resp. weak) convergence of the sequence $\{\mathcal{T}_\varepsilon(w_\varepsilon)\}$ does not imply the corresponding convergence for the sequence $\{w_\varepsilon\}$, since it gives no control of the sequence $\{w_\varepsilon 1_{\Lambda_\varepsilon}\}$. If $\{w_\varepsilon 1_{\Lambda_\varepsilon}\}$ is bounded in $L^p(\Omega)$ and if $\{\mathcal{T}_\varepsilon(w_\varepsilon)\}$ converges weakly, so does $\{w_\varepsilon\}$ by Proposition 2.9(iii). On the other hand, even if $\{w_\varepsilon 1_{\Lambda_\varepsilon}\}$ converges strongly to 0 in $L^p(\Omega)$, the strong convergence of $\{\mathcal{T}_\varepsilon(w_\varepsilon)\}$ does not imply that of $\{w_\varepsilon\}$ as it is shown by the sequence $\{f_\varepsilon|_\Omega\}$ in Corollary 2.10, unless f is a constant on Y.

COROLLARY 2.12. *Let p belong to $]1, +\infty[$, let $\{u_\varepsilon\}$ be a sequence in $L^p(\Omega)$ such that*

$$\mathcal{T}_\varepsilon(u_\varepsilon) \rightharpoonup u \quad \textit{weakly in } L^p(\Omega \times Y),$$

[1] Note that the proof of convergence (2.9) is really straightforward when using unfolding!

and let $\{v_\varepsilon\}$ *be a sequence in* $L^{p'}(\Omega)$ $(1/p + 1/p' = 1)$*, with*

$$\mathcal{T}_\varepsilon(v_\varepsilon) \to v \quad \textit{strongly in } L^{p'}(\Omega \times Y).$$

Then, for any φ *in* $C_c(\Omega)$*, one has*

$$\int_\Omega u_\varepsilon(x)\, v_\varepsilon(x)\, \varphi(x)dx \to \frac{1}{|Y|} \int_{\Omega\times Y} u(x,y)\, v(x,y)\, \varphi(x)dxdy.$$

Moreover, if

$$\int_{\Lambda_\varepsilon} |v_\varepsilon|^{p'}\, dx \to 0,$$

then, for any φ *in* $C(\overline{\Omega})$*, one has*

$$\int_\Omega u_\varepsilon(x)\, v_\varepsilon(x)\, \varphi(x)dx \to \frac{1}{|Y|} \int_{\Omega\times Y} u(x,y)\, v(x,y)\, \varphi(x)dxdy.$$

Proof. The result follows from the fact that, in both cases, the sequence $\{u_\varepsilon\, v_\varepsilon \phi\}$ satisfies the u.c.i. by the Hölder inequality. □

Remark 2.13. A consequence of (iii) of Proposition 2.9, together with (iv) of Proposition 2.5, is the following. Suppose the sequence $\{w_\varepsilon\}$ converges weakly to w in $L^p(\Omega)$. Then the sequence $\{\mathcal{T}_\varepsilon(w_\varepsilon)\}$ is relatively weakly compact in $L^p(\Omega \times Y)$, and each of its weak-limit points $\widehat{w}$ satisfies $\mathcal{M}_{_Y}(\widehat{w}) = w$.

Now recall the following definition from Nguetseng [58] and Allaire [1].

Two-scale convergence. *Let* $p \in]1, +\infty[$*. A bounded sequence* $\{w_\varepsilon\}$ *in* $L^p(\Omega)$ *two-scale converges to some* w *belonging to* $L^p(\Omega \times Y)$*, whenever, for every smooth function* φ *on* $\Omega \times Y$*, the following convergence holds:*

$$\int_\Omega w_\varepsilon(x) \varphi\Big(x, \frac{x}{\varepsilon}\Big)\, dx \to \frac{1}{|Y|} \int\int_{\Omega\times Y} w(x,y)\, \varphi(x,y)\, dxdy.$$

The next result reduces two-scale convergence of a sequence to a mere weak $L^p(\Omega \times Y)$-convergence of the unfolded sequence.

PROPOSITION 2.14. *Let* $\{w_\varepsilon\}$ *be a bounded sequence in* $L^p(\Omega)$*, with* $p \in]1, +\infty[$*. The following assertions are equivalent:*

(i) $\{\mathcal{T}_\varepsilon(w_\varepsilon)\}$ *converges weakly to* w *in* $L^p(\Omega \times Y)$,

(ii) $\{w_\varepsilon\}$ *two-scale converges to* w*.*

Proof. To prove this equivalence, it is enough to check that, for every φ in a set of admissible test functions for two-scale convergence (for instance, $\mathcal{D}(\Omega, L^q(Y))$), the sequence $\{\mathcal{T}_\varepsilon[\varphi(x, x/\varepsilon)]\}$ converges strongly to φ in $L^q(\Omega \times Y))$. This follows from the definition of $\mathcal{T}_\varepsilon$, indeed

$$\mathcal{T}_\varepsilon\Big[\varphi\Big(x, \frac{x}{\varepsilon}\Big)\Big](x,y) = \varphi\Big(\varepsilon\Big[\frac{x}{\varepsilon}\Big]_{_Y} + \varepsilon y, y\Big). \qquad \square$$

Remark 2.15. Proposition 2.14 shows that the two-scale convergence of a sequence in $L^p(\Omega)$, $p \in]1, +\infty[$, is equivalent to the weak$-L^p(\Omega \times Y)$ convergence of the unfolded sequence. Notice that, by definition, to check the two-scale convergence, one has to use special test functions. To check a weak convergence in the space $L^p(\Omega \times Y)$, one simply makes the use of functions in the dual space $L^{p'}(\Omega \times Y)$. Moreover, due to density properties, it is sufficient to check this convergence only on smooth functions from $\mathcal{D}(\Omega \times Y)$.

2.2. The averaging operator $\mathcal{U}_\varepsilon$. In this section, we consider the adjoint $\mathcal{U}_\varepsilon$ of $\mathcal{T}_\varepsilon$, which we call averaging operator. In order to do so, let v be in $L^p(\Omega\times Y)$, and let u be in $L^{p'}(\Omega)$. We have successively,

$$\begin{aligned}\frac{1}{|Y|}\int_{\Omega\times Y}\mathcal{T}_\varepsilon(u)(x,y)\,v(x,y)\,dxdy &= \frac{1}{|Y|}\int_{\widehat{\Omega}_\varepsilon\times Y}\mathcal{T}_\varepsilon(u)(x,y)\,v(x,y)\,dxdy\\ &=\frac{1}{|Y|}\sum_{\xi\in\Xi_\varepsilon}\int_{\varepsilon(\xi+Y)\times Y}u(\varepsilon\xi+\varepsilon y)\,v(x,y)\,dxdy\\ &=\sum_{\xi\in\Xi_\varepsilon}\frac{1}{|Y|}\int_{Y\times Y}u(\varepsilon\xi+\varepsilon y)\,v(\varepsilon\xi+\varepsilon z,y)\varepsilon^N\,dzdy\\ &=\sum_{\xi\in\Xi_\varepsilon}\frac{1}{|Y|}\int_Y dz\int_{\varepsilon(\xi+Y)}u(x)\,v\Big(\varepsilon\Big[\frac{x}{\varepsilon}\Big]_Y+\varepsilon z,\Big\{\frac{x}{\varepsilon}\Big\}_Y\Big)\,dx\\ &=\int_{\widehat{\Omega}_\varepsilon}u(x)\left(\frac{1}{|Y|}\int_Y v\left(\varepsilon\left[\frac{x}{\varepsilon}\right]_Y+\varepsilon z,\left\{\frac{x}{\varepsilon}\right\}_Y\right)\,dz\right)dx.\end{aligned}$$

This gives the formula for the averaging operator $\mathcal{U}_\varepsilon$.

DEFINITION 2.16. *For p in $[1,+\infty]$, the averaging operator $\mathcal{U}_\varepsilon : L^p(\Omega\times Y)\mapsto L^p(\Omega)$ is defined as*

$$\mathcal{U}_\varepsilon(\Phi)(x)=\begin{cases}\dfrac{1}{|Y|}\displaystyle\int_Y\Phi\Big(\varepsilon\Big[\frac{x}{\varepsilon}\Big]_Y+\varepsilon z,\Big\{\frac{x}{\varepsilon}\Big\}_Y\Big)\,dz & \text{a.e. for } x\in\widehat{\Omega}_\varepsilon,\\ 0 & \text{a.e. for } x\in\Lambda_\varepsilon.\end{cases}$$

Consequently, for $\psi\in L^p(\Omega)$ and $\Phi\in L^{p'}(\Omega\times Y)$, one has

$$\int_\Omega\mathcal{U}_\varepsilon(\Phi)(x)\,\psi(x)\,dx=\frac{1}{|Y|}\int_{\Omega\times Y}\Phi(x,y)\,\mathcal{T}_\varepsilon(\psi)(x,y)\,dxdy.$$

As a consequence of the duality (Hölder's inequality) and of Proposition 2.5(iv), we get the following.

PROPOSITION 2.17. *Let p belong to $[1,+\infty]$. The averaging operator is linear and continuous from $L^p(\Omega\times Y)$ to $L^p(\Omega)$ and*

$$\|\mathcal{U}_\varepsilon(\Phi)\|_{L^p(\Omega)}\le|\,Y\,|^{-\frac{1}{p}}\,\|\Phi\|_{L^p(\Omega\times Y)}. \tag{2.10}$$

The operator $\mathcal{U}_\varepsilon$ maps $L^p(\Omega\times Y)$ into the space $L^p(\Omega)$. It allows one to replace the function $x\mapsto\Phi(x,\{\frac{x}{\varepsilon}\}_Y)$, which is meaningless, in general, by a function which always makes sense. Notice that this implies, in particular, that the largest set of test functions for two-scale convergence is actually the set $\mathcal{U}_\varepsilon(\Phi)$, with Φ in $L^{p'}(\Omega\times Y)$.

It is immediate from its definition that $\mathcal{U}_\varepsilon$ is almost a left-inverse of $\mathcal{T}_\varepsilon$, since

$$\mathcal{U}_\varepsilon\big(\mathcal{T}_\varepsilon(\phi)\big)(x)=\begin{cases}\phi(x) & \text{a.e. for } x\in\widehat{\Omega}_\varepsilon,\\ 0 & \text{a.e. for } x\in\Lambda_\varepsilon,\end{cases} \tag{2.11}$$

for every ϕ in $L^p(\Omega)$, while

$$\mathcal{T}_\varepsilon(\mathcal{U}_\varepsilon(\Phi))(x,y)=\begin{cases}\dfrac{1}{|\,Y\,|}\displaystyle\int_Y\Phi\Big(\varepsilon\Big[\frac{x}{\varepsilon}\Big]_Y+\varepsilon z,y\Big)\,dz & \text{a.e. for } (x,y)\in\widehat{\Omega}_\varepsilon\times Y,\\ 0 & \text{a.e. for } (x,y)\in\Lambda_\varepsilon\times Y,\end{cases} \tag{2.12}$$

for every Φ in $L^p(\Omega\times Y)$.

PROPOSITION 2.18 (properties of $\mathcal{U}_\varepsilon$). *Suppose that p is in $[1,+\infty[$.*

(i) *Let $\{\Phi_\varepsilon\}$ be a bounded sequence in $L^p(\Omega\times Y)$ such that $\Phi_\varepsilon \rightharpoonup \Phi$ weakly in $L^p(\Omega\times Y)$. Then*

$$\mathcal{U}_\varepsilon(\Phi_\varepsilon) \rightharpoonup \mathcal{M}_{_Y}(\Phi) = \frac{1}{|Y|}\int_Y \Phi(\,\cdot\,,y)\,dy \quad \textit{weakly in } L^p(\Omega).$$

In particular, for every $\Phi\in L^p(\Omega\times Y)$,

$$\mathcal{U}_\varepsilon(\Phi) \rightharpoonup \mathcal{M}_{_Y}(\Phi) \quad \textit{weakly in } L^p(\Omega),$$

but not strongly, unless Φ is independent of y.

(ii) *Let $\{\Phi_\varepsilon\}$ be a sequence such that $\Phi_\varepsilon \to \Phi$ strongly in $L^p(\Omega\times Y)$. Then*

$$\mathcal{T}_\varepsilon(\mathcal{U}_\varepsilon(\Phi_\varepsilon)) \to \Phi \quad \textit{strongly in } L^p(\Omega\times Y).$$

(iii) *Suppose that $\{w_\varepsilon\}$ is a sequence in $L^p(\Omega)$. Then, the following assertions are equivalent:*

(a) $\mathcal{T}_\varepsilon(w_\varepsilon) \to \widehat{w}$ *strongly in* $L^p(\Omega\times Y)$,

(b) $w_\varepsilon\, 1_{\widehat{\Omega}_\varepsilon} - \mathcal{U}_\varepsilon(\widehat{w}) \to 0$ *strongly in* $L^p(\Omega)$.

(iv) *Suppose that $\{w_\varepsilon\}$ is a sequence in $L^p(\Omega)$. Then, the following assertions are equivalent:*

(c) $\mathcal{T}_\varepsilon(w_\varepsilon) \to \widehat{w}$ *strongly in* $L^p(\Omega\times Y)$ *and* $\displaystyle\int_{\Lambda_\varepsilon} |w_\varepsilon|^p\,dx \to 0$,

(d) $w_\varepsilon - \mathcal{U}_\varepsilon(\widehat{w}) \to 0$ *strongly in* $L^p(\Omega)$.

Proof. (i) This follows from Proposition 2.9(ii) by duality for $p>1$. It still holds for $p=1$ in the same way as the proof of Proposition 2.9(ii). Indeed, if the De La Vallée–Poussin criterion is satisfied by the sequence $\{\Phi_\varepsilon\}$, it is also satisfied by the sequence $\{\mathcal{U}_\varepsilon(\Phi_\varepsilon)\}$, since for F convex and continuous, Jensen's inequality implies that

$$F(\mathcal{U}_\varepsilon(\Phi_\varepsilon))(x) \le \mathcal{U}_\varepsilon(F(\Phi_\varepsilon))(x).$$

(ii) The proof follows the same lines as that of (i)–(ii) of Proposition 2.9.

(iii) The implication (a)⇒(b) follows from (2.10) applied to $\Phi_\varepsilon = w_\varepsilon\, 1_{\widehat{\Omega}_\varepsilon} - \mathcal{U}_\varepsilon(\widehat{w})$ and from (2.11).

As for the converse (b)⇒(a), Proposition 2.9(ii) implies that

$$\mathcal{T}_\varepsilon(w_\varepsilon\, 1_{\widehat{\Omega}_\varepsilon} - \mathcal{U}_\varepsilon(\widehat{w})) \to 0 \quad \text{strongly in } L^p(\Omega\times Y).$$

Since $\mathcal{T}_\varepsilon(w_\varepsilon) = \mathcal{T}_\varepsilon(w_\varepsilon\, 1_{\widehat{\Omega}_\varepsilon})$, from (ii) above it converges to $\widehat{w}$ strongly in $L^p(\Omega\times Y)$.

(iv) The implication (c)⇒(d) follows from (iii) and the second condition of (c).

Its converse (d)⇒(c) is a consequence of from (iii): since $\mathcal{U}_\varepsilon(\widehat{w})\,1_{\Lambda_\varepsilon} = 0$, (d) implies (b) and $w_\varepsilon\, 1_{\Lambda_\varepsilon} \to 0$ in $L^p(\Omega)$. ∎

Remark 2.19. The statement of Proposition 2.18(iii) does not hold with weak convergences instead of strong ones, contrary to an erroneous statement made in [24]. In view of (2.11) and Proposition 2.18(i), if $\mathcal{T}_\varepsilon(w_\varepsilon) \rightharpoonup \widehat{w}$ weakly in $L^p(\Omega\times Y)$, then $w_\varepsilon\, 1_{\widehat{\Omega}_\varepsilon} - \mathcal{U}_\varepsilon(\widehat{w})$ converges weakly to 0 in $L^p(\Omega)$.

But the converse of this last implication cannot hold. Indeed, choose $\widehat{v}$ with $\mathcal{M}_{_Y}(\widehat{v}) = 0$. By Proposition 2.18(i), $\mathcal{U}_\varepsilon(\widehat{v})$ converges weakly to $\mathcal{M}_{_Y}(\widehat{v}) = 0$. Consequently, the weak limit of $w_\varepsilon\, 1_{\widehat{\Omega}_\varepsilon} - \mathcal{U}_\varepsilon(\widehat{w})$ is also the weak limit of $w_\varepsilon\, 1_{\widehat{\Omega}_\varepsilon} - \mathcal{U}_\varepsilon(\widehat{w}+\widehat{v})$. If the converse were true, it would imply that $\mathcal{T}_\varepsilon(w_\varepsilon)$ converges weakly to both $\widehat{w}$ and $\widehat{w}+\widehat{v}$. So $\widehat{v} = 0$. In other words, $\mathcal{M}_{_Y}(\widehat{v}) = 0$ would imply $\widehat{v} = 0$.

Remark 2.20. Assertions (iii)(b) and (iv)(d) are corrector–type results.

Remark 2.21. The condition (iii)(a) is used by some authors to define the notion of "strong two-scale convergence." From the above considerations, condition (c) of Proposition 2.18(iv) is a better candidate for this definition.

2.3. The local average operator $\mathcal{M}_\varepsilon$. In this section, we consider the classical average operator associated to the partition of Ω by ε-cells Y (setting it to be zero on the cells intersecting the boundary $\partial\Omega$).

DEFINITION 2.22. *The local average operator $\mathcal{M}_\varepsilon : L^p(\Omega) \mapsto L^p(\Omega)$, for $p \in [1,+\infty]$, is defined by*

$$\mathcal{M}_\varepsilon(\phi)(x) = \begin{cases} \dfrac{1}{\varepsilon^N |Y|} \displaystyle\int_\varepsilon \left[\frac{x}{\varepsilon}\right]_y \phi(\zeta)\, d\zeta & \text{if } x \in \widehat{\Omega}_\varepsilon, \\ 0 \qquad \text{if } x \in \Lambda_\varepsilon. \end{cases} \tag{2.13}$$

Remark 2.23. It turns out that the local average $\mathcal{M}_\varepsilon$ is connected to the unfolding operator $\mathcal{T}_\varepsilon$. Indeed, by the usual change of variable cell by cell,

$$\mathcal{M}_\varepsilon(\phi)(x) = \frac{1}{|Y|} \int_Y \mathcal{T}_\varepsilon(\phi)(x,y)\, dy = \mathcal{M}_{_Y}\big(\mathcal{T}_\varepsilon(\phi)\big)(x).$$

Remark 2.24. Note that, for any ϕ in $L^p(\Omega)$, one has $\mathcal{T}_\varepsilon(\mathcal{M}_\varepsilon(\phi)) = \mathcal{M}_\varepsilon(\phi)$ on the set $\Omega \times Y$. Moreover, one also has $\mathcal{U}_\varepsilon(\phi) = \mathcal{M}_\varepsilon(\phi)$.

PROPOSITION 2.25 (properties of $\mathcal{M}_\varepsilon$).

(i) *Suppose that p is in $[1,+\infty]$. For any any ϕ in $L^p(\Omega)$,*

$$\|\mathcal{M}_\varepsilon(\phi)\|_{L^p(\Omega)} \le \|\phi\|_{L^p(\Omega)}.$$

(ii) *Suppose that p is in $[1,+\infty]$. For $\phi \in L^p(\Omega)$ and $\psi \in L^{p'}(\Omega)$,*

$$\int_\Omega \mathcal{M}_\varepsilon(\phi)\, \psi\, dx = \int_\Omega \mathcal{M}_\varepsilon(\phi)\, \mathcal{M}_\varepsilon(\psi)\, dx = \int_\Omega \phi\, \mathcal{M}_\varepsilon(\psi)\, dx. \tag{2.14}$$

(iii) *Suppose that p is in $[1,+\infty[$. Let $\{v_\varepsilon\}$ be a sequence such that $v_\varepsilon \to v$ strongly in $L^p(\Omega)$. Then*

$$\mathcal{M}_\varepsilon(v_\varepsilon) \to v \quad \text{strongly in } L^p(\Omega).$$

In particular, for every $\phi \in L^p(\Omega)$,

$$\mathcal{M}_\varepsilon(\phi) \to \phi \quad \text{strongly in } L^p(\Omega). \tag{2.15}$$

(iv) *Suppose that p is in $[1,+\infty[$. Let $\{v_\varepsilon\}$ be a sequence such that $v_\varepsilon \rightharpoonup v$ weakly in $L^p(\Omega)$. Then*

$$\mathcal{M}_\varepsilon(v_\varepsilon) \rightharpoonup v \quad \text{weakly in } L^p(\Omega).$$

The same holds true for the weak-$*$ topology in $L^\infty(\Omega)$.

Proof. The proofs of (i) and (ii) are straightforward. The proof of (iii) is a simple consequence of (ii) of Proposition 2.9. For the proof of (iv), let ϕ be in $L^{p'}(\Omega)$, with $p' \in [1, +\infty[$ $(p \neq 1)$, and use (2.14) and (2.15) to obtain

$$\int_\Omega \phi\, \mathcal{M}_\varepsilon(v_\varepsilon)\, dx = \int_\Omega \mathcal{M}_\varepsilon(\phi)\, v_\varepsilon\, dx \to \int_\Omega \phi\, v\, dx.$$

For $p = 1$, in the same way as the proof of Proposition 2.9(ii) and Proposition 2.18(i), if the De La Vallée–Poussin criterion is satisfied by the sequence $\{v_\varepsilon\}$, it is also satisfied by the sequence $\{\mathcal{M}_\varepsilon(v_\varepsilon)\}$, since for F convex and continuous, Jensen's inequality implies that

$$F(\mathcal{M}_\varepsilon(v_\varepsilon))(x) \leq \mathcal{M}_\varepsilon(F(v_\varepsilon))(x),$$

which ends the proof. □

COROLLARY 2.26. *Suppose that p is in $[1, +\infty[$. Let w be in $L^p(\Omega)$ and $\{w_\varepsilon\}$ be a sequence in $L^p(\Omega)$ satisfying $\mathcal{T}_\varepsilon(w_\varepsilon) \to w$ strongly in $L^p(\Omega \times Y)$. Then,*

$$w_\varepsilon\, 1_{\widehat{\Omega}_\varepsilon} \to w \quad \text{strongly in } L^p(\Omega).$$

Furthermore, if $\int_{\Lambda_\varepsilon} |w_\varepsilon|^p \to 0$, then, $w_\varepsilon \to w$ strongly in $L^p(\Omega)$.

Proof. Since w does not depend on y, one has $\mathcal{U}_\varepsilon(w) = \mathcal{M}_\varepsilon(w)$ which, by Proposition 2.25(iii), converges strongly to w. The conclusion follows from Proposition 2.18(iii), respectively, (iv). □

3. Unfolding and gradients. This section is devoted to the properties of the restriction of the unfolding operator to the space $W^{1,p}(\Omega)$. Some results require no extra hypotheses, but many others are sensitive to the boundary conditions and the regularity of the boundary itself.

Observe that, for w in $W^{1,p}(\Omega)$, one has

$$\nabla_y(\mathcal{T}_\varepsilon(w)) = \varepsilon \mathcal{T}_\varepsilon(\nabla w), \quad \forall w \in W^{1,p}(\Omega) \quad \text{a.e. for} \quad (x, y) \in \Omega \times Y. \tag{3.1}$$

Then, Proposition 2.5(iv) implies that $\mathcal{T}_\varepsilon$ maps $W^{1,p}(\Omega)$ into $L^p(\Omega; W^{1,p}(Y))$.

For simplicity, we assume that $Y =]0, 1[^n$. Nevertheless, the results we prove here hold true in the case of a general Y, with minor modifications.

PROPOSITION 3.1 (gradient in the direction of a period). *Let k in $[1, \ldots, n]$ and $\{w_\varepsilon\}$ be a bounded sequence in $L^p(\Omega)$, with $p \in]1, +\infty]$, satisfying*

$$\varepsilon \left\| \frac{\partial w_\varepsilon}{\partial x_k} \right\|_{L^p(\Omega)} \leq C. \tag{3.2}$$

Then, there exist a subsequence (still denoted ε) and $\widehat{w}$ in $L^p(\Omega \times Y)$, with $\frac{\partial \widehat{w}}{\partial y_k}$ in $L^p(\Omega \times Y)$ such that

(3.3)

$$\mathcal{T}_\varepsilon(w_\varepsilon) \rightharpoonup \widehat{w} \quad \text{weakly in } L^p(\Omega \times Y),$$

$$\varepsilon \mathcal{T}_\varepsilon\left(\frac{\partial w_\varepsilon}{\partial x_k}\right) = \frac{\partial \mathcal{T}_\varepsilon(w_\varepsilon)}{\partial y_k} \rightharpoonup \frac{\partial \widehat{w}}{\partial y_k} \quad \text{weakly in } L^p(\Omega \times Y), \text{ (weakly-* for } p = +\infty).$$

Moreover, the limit function $\widehat{w}$ is 1-periodic, with respect to the y_k coordinate.

Proof. Convergences (3.3) are a simple consequence of (3.1) and (3.2). It remains to prove the periodicity of $\widehat{w}$. Without loss of generality, assume $k = n$ and write $y = (y', y_n)$, with y' in $Y' \doteq]0,1[^{n-1}$ and $y_n \in]0,1[$.

Let $\psi \in \mathcal{D}(\Omega \times Y')$. Convergences (3.3) imply that the sequence $\{\mathcal{T}_\varepsilon(w_\varepsilon)\}$ is bounded in $L^p(\Omega \times Y'; W^{1,p}(0,1))$ so that $\{\mathcal{T}_\varepsilon(w_\varepsilon)|_{\{y_n=s\}}\}$ is bounded in $L^p(\Omega \times Y')$ for every $s \in [0,1]$. The periodicity with respect to y_n results from the following computation with an obvious change of variable:

$$\begin{aligned}
&\int_{\Omega\times Y'} \big[\mathcal{T}_\varepsilon(w_\varepsilon)(x,(y',1)) - \mathcal{T}_\varepsilon(w_\varepsilon)(x,(y',0)\big]\,\psi(x,y')\,dx\,dy' \\
&\quad= \int_{\Omega\times Y'} \Big\{w_\varepsilon\Big(\varepsilon\Big[\frac{x}{\varepsilon}\Big]_Y + \varepsilon(y',1)\Big) - w_\varepsilon\Big(\varepsilon\Big[\frac{x}{\varepsilon}\Big]_Y + \varepsilon(y',0)\Big)\Big\}\psi(x,y')\,dx\,dy' \\
&\quad= \int_{\Omega\times Y'} w_\varepsilon\Big(\varepsilon\Big[\frac{x}{\varepsilon}\Big]_Y + \varepsilon(y',0)\Big)\big[\psi(x-\varepsilon e_n,y') - \psi(x,y')\big]\,dx\,dy', \\
&\quad= \int_{\Omega\times Y'} \mathcal{T}_\varepsilon(w_\varepsilon)(x,(y',0))\big[\psi(x-\varepsilon e_n,y') - \psi(x,y')\big]\,dx\,dy',
\end{aligned}$$

which goes to zero. ∎

COROLLARY 3.2. *Let $\{w_\varepsilon\}$ be in $W^{1,p}(\Omega)$, with $p \in]1,+\infty[$, and assume that $\{w_\varepsilon\}$ is a bounded sequence in $L^p(\Omega)$ satisfying*

$$\varepsilon\|\nabla w_\varepsilon\|_{L^p(\Omega)} \le C.$$

Then, there exist a subsequence (still denoted ε) and $\widehat{w} \in L^p(\Omega; W^{1,p}(Y))$ such that

$$\begin{aligned}
&\mathcal{T}_\varepsilon(w_\varepsilon) \rightharpoonup \widehat{w} \quad \text{weakly in } L^p(\Omega; W^{1,p}(Y)),\\
&\varepsilon\mathcal{T}_\varepsilon(\nabla w_\varepsilon) \rightharpoonup \nabla_y\widehat{w} \quad \text{weakly in } L^p(\Omega\times Y).
\end{aligned}$$

Moreover, the limit function $\widehat{w}$ is Y-periodic, i.e., belongs to $L^p(\Omega; W^{1,p}_{per}(Y))$, where $W^{1,p}_{per}(Y)$ denotes the Banach space of Y-periodic functions in $W^{1,p}_{loc}(\mathbb{R}^n)$, with the $W^{1,p}(Y)$ norm.

COROLLARY 3.3. *Let p be in $]1,+\infty[$ and $\{w_\varepsilon\}$ be a sequence converging weakly in $W^{1,p}(\Omega)$ to w. Then,*

$$\mathcal{T}_\varepsilon(w_\varepsilon) \rightharpoonup w \quad \text{weakly in } L^p(\Omega; W^{1,p}(Y)).$$

Furthermore, if $\{w_\varepsilon\}$ converges strongly to w in $L^p(\Omega)$, the above convergence is strong (this is the case if, for example, $W^{1,p}(\Omega)$ is compactly embedded in $L^p(\Omega)$).

Proof. Using (3.1), since $\{w_\varepsilon\}$ weakly converges, one has the estimates

$$\begin{aligned}
&\|\mathcal{T}_\varepsilon(w_\varepsilon)\|_{L^p(\Omega\times Y)} \le C,\\
&\|\nabla_y(\mathcal{T}_\varepsilon(w_\varepsilon))\|_{L^p(\Omega\times Y)} \le \varepsilon C,
\end{aligned}$$

so that there exist a subsequence (still denoted ε) and $\widehat{w}$ in $L^p(\Omega; W^{1,p}(Y))$ such that

$$\begin{aligned}
&\mathcal{T}_\varepsilon(w_\varepsilon) \rightharpoonup \widehat{w} \quad \text{weakly in } L^p(\Omega; W^{1,p}(Y)),\\
&\nabla_y(\mathcal{T}_\varepsilon(w_\varepsilon)) \to 0 \quad \text{strongly in } L^p(\Omega\times Y),
\end{aligned} \tag{3.4}$$

and $\nabla_y\widehat{w} = 0$. Consequently, $\widehat{w}$ does not depend on y, and Proposition 2.9(iii) immediately gives $w = \mathcal{M}_Y(\widehat{w}) = \widehat{w}$. Moreover, convergence (3.4) holds for the entire

sequence ε. Finally, if the sequence $\{w_\varepsilon\}$ converges strongly to w in $L^p(\Omega)$, so does the sequence $\{\mathcal{T}_\varepsilon(w_\varepsilon)\}$, thanks to Proposition 2.9(ii). □

PROPOSITION 3.4. *Suppose that p is in $[1,+\infty[$. Let $\{w_\varepsilon\}$ be a sequence which converges strongly to some w in $W^{1,p}(\Omega)$. Then,*

$$\text{(i)}\quad \mathcal{T}_\varepsilon(\nabla w_\varepsilon) \to \nabla w \quad \textit{strongly in } L^p(\Omega\times Y),$$

$$\text{(ii)}\quad \frac{1}{\varepsilon}\Big(\mathcal{T}_\varepsilon(w_\varepsilon) - \mathcal{M}_\varepsilon(w_\varepsilon)\Big) \to y^c\cdot\nabla w \qquad \textit{strongly in } \ L^p(\Omega; W^{1,p}(Y)),$$

where

$$y^c = \left(y_1 - \frac{1}{2},\dots, y_n - \frac{1}{2}\right).$$

Proof. The first asssertion follows from Proposition 2.9(i). To prove (ii), set

$$Z_\varepsilon \doteq \frac{1}{\varepsilon}\Big(\mathcal{T}_\varepsilon(w_\varepsilon) - \mathcal{M}_\varepsilon(w_\varepsilon)\Big),$$

which has mean value zero in Y. Since

$$\nabla_y Z_\varepsilon = \frac{1}{\varepsilon}\nabla_y\big(\mathcal{T}_\varepsilon(w_\varepsilon)\big) = \mathcal{T}_\varepsilon\big(\nabla w_\varepsilon\big),$$

thanks to assertion (i),

$$\nabla_y Z_\varepsilon \to \nabla w \quad \text{ strongly in } L^p(\Omega\times Y).$$

Then recall the Poincaré–Wirtinger inequality in Y:

$$\forall \psi\in W^{1,p}(Y),\quad \big\|\psi - \mathcal{M}_{_Y}(\psi)\big\|_{L^p(Y)} \le C\|\nabla\psi\|_{L^p(Y)}. \tag{3.5}$$

Applying it to the function $Z_\varepsilon - y^c\cdot\nabla w$ (which is of mean value zero) gives

$$\big\|Z_\varepsilon - y^c\cdot\nabla w\big\|_{L^p(\Omega\times Y)} \le C\|\nabla_y Z_\varepsilon - \nabla w\|_{L^p(\Omega\times Y)}, \tag{3.6}$$

and this concludes the proof. □

THEOREM 3.5. *Suppose that p is in $]1,+\infty[$. Let $\{w_\varepsilon\}$ be a sequence converging weakly to some w in $W^{1,p}(\Omega)$. Up to a subsequence, there exists some $\widehat{w}$ in $L^p(\Omega; W^{1,p}_{per}(Y))$ such that*

$$\begin{aligned}&\text{(i)}\quad \mathcal{T}_\varepsilon(\nabla w_\varepsilon) \rightharpoonup \nabla w + \nabla_y\widehat{w} \quad \textit{weakly in } L^p(\Omega\times Y),\\ &\text{(ii)}\quad \frac{1}{\varepsilon}\Big(\mathcal{T}_\varepsilon(w_\varepsilon) - \mathcal{M}_\varepsilon(w_\varepsilon)\Big) \rightharpoonup \widehat{w} + y^c\cdot\nabla w \quad \textit{weakly in } \ L^p(\Omega; W^{1,p}(Y)).\end{aligned} \tag{3.7}$$

Moreover, $\mathcal{M}_{_Y}(\widehat{w}) = 0$.

Proof. Following the same lines as in the previous proof, introduce

$$Z_\varepsilon = \frac{1}{\varepsilon}\Big(\mathcal{T}_\varepsilon(w_\varepsilon) - \mathcal{M}_\varepsilon(w_\varepsilon)\Big),$$

which has mean value zero in Y. Since $\nabla_y Z_\varepsilon = \mathcal{T}_\varepsilon\big(\nabla w_\varepsilon\big)$, (ii) implies (i).

To prove (ii), note that the sequence $\{\nabla_y Z_\varepsilon\}$ is bounded in $L^p(\Omega \times Y)$. By (3.6),

$$\left\| Z_\varepsilon - y^c \cdot \nabla w \right\|_{L^p(\Omega \times Y)} \leq C$$

so that there exists $\widehat{w}$ in $L^p(\Omega; W^{1,p}(Y))$ such that, up to a subsequence,

$$Z_\varepsilon - y^c \cdot \nabla w \rightharpoonup \widehat{w} \quad \text{weakly in} \quad L^p(\Omega; W^{1,p}(Y)).$$

Since, by construction, $\mathcal{M}_Y(y^c)$ vanishes, so does $\mathcal{M}_Y(\widehat{w})$.

It remains to prove the Y-periodicity of $\widehat{w}$. This is obtained in the same way as in the proof of Proposition 3.1 by using a test function $\psi \in \mathcal{D}(\Omega \times Y')$. One has successively,

$$\begin{aligned}
&\int_{\Omega \times Y'} \left[Z_\varepsilon(x,(y',1)) - Z_\varepsilon(x,(y',0)) \right] \psi(x,y')\, dx\, dy' \\
&\quad = \int_{\Omega \times Y'} \frac{1}{\varepsilon} \Big\{ w_\varepsilon\Big(\varepsilon \Big[\frac{x}{\varepsilon}\Big]_Y + \varepsilon(y',1)\Big) - w_\varepsilon\Big(\varepsilon \Big[\frac{x}{\varepsilon}\Big]_Y + \varepsilon(y',0)\Big) \Big\} \psi(x,y')\, dx\, dy' \\
&\quad = \int_{\Omega \times Y'} w_\varepsilon\Big(\varepsilon \Big[\frac{x}{\varepsilon}\Big]_Y + \varepsilon(y',0)\Big) \frac{1}{\varepsilon} \left[\psi(x - \varepsilon e_n, y') - \psi(x,y') \right] dx\, dy', \\
&\quad = \int_{\Omega \times Y'} \mathcal{T}_\varepsilon(w_\varepsilon)(x,(y',0)) \frac{1}{\varepsilon} \left[\psi(x - \varepsilon e_n, y') - \psi(x,y') \right] dx\, dy'.
\end{aligned}$$

By Proposition 2.9(ii), $\{\mathcal{T}_\varepsilon(w_\varepsilon)\}$ converges strongly to w in $L^p(\Omega \times Y)$, and by (3.7) (i), it converges weakly to the same w in $L^p(\Omega; W^{1,p}(Y))$. By the trace theorem in $W^{1,p}(Y)$, the trace of $\mathcal{T}_\varepsilon(w_\varepsilon)$ on $\Omega \times Y'$ converges weakly to w in $L^p(\Omega \times Y')$. Hence, the last integral converges to

$$-\int_{\Omega \times Y'} w(x) \frac{\partial \psi}{\partial x_n}(x,y')\, dx\, dy'. \tag{3.8}$$

Similarly, since $(y^c \cdot \nabla w)(y',1) - (y^c \cdot \nabla w)(y',0) = \frac{\partial w}{\partial x_n}$, we obtain

$$\begin{aligned}
&\int_{\Omega \times Y'} \left[(y^c \cdot \nabla w)(y',1) - (y^c \cdot \nabla w)(y',0) \right] \psi(x,y')\, dx\, dy' \\
&\qquad = \int_{\Omega \times Y'} \frac{\partial w}{\partial x_n}\, \psi(x,y')\, dx\, dy' = -\int_{\Omega \times Y'} w(x) \frac{\partial \psi}{\partial x_n}(x,y')\, dx\, dy'.
\end{aligned}$$

This, together with (3.8) and convergence (3.7)(ii), shows that

$$\int_{\Omega \times Y'} \left[\widehat{w}(x,(y',1)) - \widehat{w}(x,(y',0)) \right] \psi(x,y')\, dx\, dy' = 0,$$

so that $\widehat{w}$ is y_n-periodic. The same holds in the directions of all of the other periods. □

Theorem 3.5 can be generalized to the case of $W^{k,p}(\Omega)$-spaces, with $k \geq 1$ and $p \in]1, +\infty[$. In order to do so, for $r = (r_1, \ldots, r_n) \in \mathbb{N}^n$ with $|r| = r_1 + \cdots + r_n \leq k$, introduce the notation D^r and D^r_y:

$$D^r = \frac{\partial^{|r|}}{\partial x_1^{r_1} \ldots \partial x_n^{r_n}}, \qquad D^r_y = \frac{\partial^{|r|}}{\partial y_1^{r_1} \ldots \partial y_n^{r_n}}.$$

Then the following result holds.

THEOREM 3.6. *Let $\{w_\varepsilon\}$ be a sequence converging weakly in $W^{k,p}(\Omega)$ to w, $k \geq 1$, and $p \in]1, +\infty[$. There exist a subsequence (still denoted ε) and $\widehat{w}$ in the space $L^p(\Omega; W^{k,p}_{per}(Y))$ such that*

$$\begin{cases} \mathcal{T}_\varepsilon(D^l w_\varepsilon) \rightharpoonup D^l w \quad \text{weakly in } L^p(\Omega; W^{k-l,p}(Y)), \quad |l| \leq k-1, \\ \mathcal{T}_\varepsilon(D^l w_\varepsilon) \rightharpoonup D^l w + D^l_y \widehat{w} \quad \text{weakly in } L^p(\Omega \times Y), \quad |l| = k. \end{cases} \tag{3.9}$$

Furthermore, if $\{w_\varepsilon\}$ converges strongly to w in $W^{k-1,p}(\Omega)$, the above convergences are strong in $L^p(\Omega; W^{k-l,p}(Y))$ for $|l| \leq k-1$.

Proof. We give a brief proof for $k = 2$. The same argument generalizes for $k > 2$. If $|l| = 1$, the first convergence in (3.9) follows directly from Corollary 3.3. Set

$$W_\varepsilon = \frac{1}{\varepsilon^2}\Big[\mathcal{T}_\varepsilon(w_\varepsilon) - \mathcal{M}_\varepsilon(w_\varepsilon) - y^c \cdot \mathcal{M}_\varepsilon\big(\nabla w_\varepsilon\big)\Big].$$

The sequence $\{w_\varepsilon\}$ is bounded in $W^{2,p}(\Omega)$, hence proceeding as in the proof of Proposition 2.25(iii), one obtains

$$\big\|W_\varepsilon\big\|_{L^p(\Omega\times Y)} \leq C.$$

Moreover,

$$\nabla_y\big(W_\varepsilon\big) = \frac{1}{\varepsilon^2}\Big(\mathcal{T}_\varepsilon\big(\nabla w_\varepsilon\big) - \mathcal{M}_\varepsilon(\nabla w_\varepsilon)\Big),$$

and

$$D^l_y\big(W_\varepsilon\big) = \mathcal{T}_\varepsilon\big(D^l w_\varepsilon\big), \quad \text{with} \quad |l| = 2.$$

This implies that the sequence $\{W_\varepsilon\}$ is bounded in $L^p(\Omega; W^{2,p}(Y))$. Therefore, there exist a subsequence (still denoted ε) and $\widetilde{w} \in L^p(\Omega; W^{2,p}(Y))$ such that

$$\begin{aligned} &W_\varepsilon \rightharpoonup \widetilde{w} \quad \text{weakly in } L^p(\Omega; W^{2,p}(Y)), \\ &\frac{\partial W_\varepsilon}{\partial y_i} = \frac{1}{\varepsilon^2}\left(\mathcal{T}_\varepsilon\left(\frac{\partial w_\varepsilon}{\partial x_i}\right) - \mathcal{M}_\varepsilon\left(\frac{\partial w_\varepsilon}{\partial x_i}\right)\right) \rightharpoonup \frac{\partial \widetilde{w}}{\partial y_i} \quad \text{weakly in } L^p(\Omega; W^{1,p}(Y)). \end{aligned} \tag{3.10}$$

Consequently,

$$D^l_y(W_\varepsilon) = \mathcal{T}_\varepsilon(D^l w_\varepsilon) \rightharpoonup D^l_y \widetilde{w} \quad \text{weakly in } L^p(\Omega \times Y), \quad |l| = 2. \tag{3.11}$$

Now, apply Theorem 3.5 to each of the derivatives $\frac{\partial w_\varepsilon}{\partial x_i}$, $i \in \{1, \ldots, n\}$. There exist a subsequence (still denoted ε) and $\widehat{w}_i \in L^p(\Omega; W^{1,p}_{per}(Y))$ such that $\mathcal{M}_{_Y}(\widehat{w}_i) \equiv 0$ and

$$\frac{1}{\varepsilon}\left(\mathcal{T}_\varepsilon\left(\frac{\partial w_\varepsilon}{\partial x_i}\right) - \mathcal{M}_\varepsilon\left(\frac{\partial w_\varepsilon}{\partial x_i}\right)\right) \rightharpoonup y^c \cdot \nabla \frac{\partial w}{\partial x_i} + \widehat{w}_i \quad \text{weakly in } L^p(\Omega \times Y).$$

Then (3.10) gives

$$\forall i \in \{1, \ldots, n\}, \qquad \frac{\partial \widetilde{w}}{\partial y_i} = y^c \cdot \nabla \frac{\partial w}{\partial x_i} + \widehat{w}_i. \tag{3.12}$$

Set

$$\widehat{w} = \widetilde{w} - \frac{1}{2}\sum_{i,j=1}^{n} \big(y^c_i\, y^c_j - \mathcal{M}_{_Y}(y^c_i\, y^c_j)\big)\frac{\partial^2 w}{\partial x_i \partial x_j}.$$

By construction, the function $\widehat{w}$ belongs to $L^p(\Omega; W^{2,p}(Y))$. Furthermore,

$$\mathcal{M}_{_Y}(\widehat{w}) = 0, \qquad \frac{\partial \widehat{w}}{\partial y_i} = \frac{\partial \widetilde{w}}{\partial y_i} - y^c \cdot \nabla \left(\frac{\partial w}{\partial x_i} \right) = \widehat{w}_i, \quad \text{and} \quad \mathcal{M}_{_Y}(\nabla_y \widehat{w}) = 0.$$

The last equality implies that $\widehat{w}$ belongs to $L^p(\Omega; W^{2,p}_{per}(Y))$. Finally from (3.12) one gets

$$D^l_y \widetilde{w} = D^l w + D^l_y \widehat{w}, \qquad \text{with} \quad |l| = 2,$$

which together with (3.11), proves the last convergence of (3.9). □

COROLLARY 3.7. *Let $\{w_\varepsilon\}$ be a sequence converging weakly in $W^{2,p}(\Omega)$ to w, and $p \in]1, +\infty[$. Then, there exist a subsequence (still denoted ε) and $\widehat{w}$ in the space $L^p(\Omega; W^{2,p}_{per}(Y))$ such that*

$$\frac{1}{\varepsilon^2}\Big[\mathcal{T}_\varepsilon(w_\varepsilon) - \mathcal{M}_\varepsilon(w_\varepsilon) - y^c \cdot \mathcal{M}_\varepsilon\big(\nabla w_\varepsilon\big)\Big] \rightharpoonup \frac{1}{2} \sum_{i,j=1}^{n} \big(y^c_i\, y^c_j - \mathcal{M}_{_Y}(y^c_i\, y^c_j)\big) \frac{\partial^2 w}{\partial x_i \partial x_j} + \widehat{w}$$

weakly in $L^p(\Omega; W^{2,p}(Y))$, where $\widehat{w}$ is such that $\mathcal{M}_{_Y}(\widehat{w}) = 0$.

Remark 3.8. For the case $Y =]0,1[^n$, y^c was defined in Proposition 3.4. For a general Y, all of the statements of this section hold true, with $y^c = y - \mathcal{M}_{_Y}(y)$.

4. Macro-micro decomposition: The scale-splitting operators $\mathcal{Q}_\varepsilon$ and $\mathcal{R}_\varepsilon$. In this section, we give a different proof of Theorem 3.5, which was the one given originally in [24]. It is based on a scale-separation decomposition which is useful in some specific situations, for example, in the statement of general corrector results (see section 6).

The procedure is based on a splitting of functions ϕ in $W^{1,p}(\Omega)$ (or in $W^{1,p}_0(\Omega)$) for $p \in [1, +\infty]$, in the form

$$\phi = \mathcal{Q}_\varepsilon(\phi) + \mathcal{R}_\varepsilon(\phi),$$

where $\mathcal{Q}_\varepsilon(\phi)$ is an approximation of ϕ having the same behavior as ϕ, while $\mathcal{R}_\varepsilon(\phi)$ is a remainder of order ε. Applied to the sequence $\{w_\varepsilon\}$ converging weakly to w in $W^{1,p}(\Omega)$, it shows that, while $\{\nabla w_\varepsilon\}$, $\{\nabla(\mathcal{Q}_\varepsilon(w_\varepsilon))\}$ and $\{\mathcal{T}_\varepsilon(\nabla \mathcal{Q}_\varepsilon(w_\varepsilon))\}$ have the same weak limit ∇w in $L^p(\Omega)$, respectively, in $L^p(\Omega \times Y)$, the sequence $\mathcal{T}_\varepsilon\big(\nabla(\mathcal{R}_\varepsilon(w_\varepsilon))\big)$ converges weakly in $L^p(\Omega \times Y)$ to $\nabla_y \widehat{w}'$ for some $\widehat{w}'$ in $L^p(\Omega; W^{1,p}_{per}(Y))$.

We will distinguish between the case $W^{1,p}_0(\Omega)$ and the case $W^{1,p}(\Omega)$. For the former, any function ϕ in $W^{1,p}_0(\Omega)$ is extended by zero to the whole of $\mathbb{R}^n$, and this extension is denoted by $\widetilde{\phi}$. In the latter case, we suppose that $\partial\Omega$ is smooth enough so that there exists a continuous extension operator $\mathcal{P} : W^{1,p}(\Omega) \mapsto W^{1,p}(\mathbb{R}^n)$ satisfying

$$\|\mathcal{P}(\phi)\|_{W^{1,p}(\mathbb{R}^n)} \le C \, \|\phi\|_{W^{1,p}(\Omega)}, \quad \forall \phi \in W^{1,p}(\Omega),$$

where C is a constant depending on p and $\partial\Omega$ only.

The construction of $\mathcal{Q}_\varepsilon$ is based on the Q_1 interpolate of some discrete approximation, as is customary in FEM. The idea of using these types of interpolate was already present in Griso [40], [41] for the study of truss-like structures. For the purpose of this paper, it is enough to take the average on $\varepsilon\xi + \varepsilon Y$ to construct the discrete approximations, but the average on $\varepsilon\xi + \varepsilon Y'$, where Y' is any fixed open subset of Y,

or any open subset of a manifold of codimension 1 in Y. The only property which is needed is the Poincaré–Wirtinger inequality, which holds in both of these cases.

DEFINITION 4.1. *The operator $\mathcal{Q}_\varepsilon : L^p(\mathbb{R}^n) \mapsto W^{1,\infty}(\mathbb{R}^n)$, for $p \in [1, +\infty]$, is defined as follows:*

$$\mathcal{Q}_\varepsilon(\phi)(\varepsilon\xi) = \mathcal{M}_\varepsilon(\phi)(\varepsilon\xi) \quad \text{for } \xi \in \varepsilon\mathbb{Z}^n, \tag{4.1}$$

and for any $x \in \mathbb{R}^n$, we set

$$\begin{aligned} &\mathcal{Q}_\varepsilon(\phi)(x) \text{ is the } Q_1 \text{ interpolate of the values of } \mathcal{Q}_\varepsilon(\phi) \text{ at the vertices} \\ &\text{of the cell } \varepsilon\Big[\frac{x}{\varepsilon}\Big]_Y + \varepsilon Y. \end{aligned} \tag{4.2}$$

In the case of the space $W_0^{1,p}(\Omega)$, the operator $\mathcal{Q}_\varepsilon : W_0^{1,p}(\Omega) \mapsto W^{1,\infty}(\Omega)$ is defined by

$$\mathcal{Q}_\varepsilon(\phi) = \mathcal{Q}_\varepsilon(\widetilde{\phi})|_\Omega,$$

where $\mathcal{Q}_\varepsilon(\widetilde{\phi})$ is given by (4.1).

In the case of the space $W^{1,p}(\Omega)$, the operator $\mathcal{Q}_\varepsilon : W^{1,p}(\Omega) \mapsto W^{1,\infty}(\Omega)$ is defined by

$$\mathcal{Q}_\varepsilon(\phi) = \mathcal{Q}_\varepsilon(\mathcal{P}(\phi))|_\Omega,$$

where $\mathcal{Q}_\varepsilon(\mathcal{P}(\phi))$ is given by (4.1).

We start with the following estimates.

PROPOSITION 4.2 (properties of $\mathcal{Q}_\varepsilon$ on $\mathbb{R}^n$). *For ϕ in $L^p(\mathbb{R}^n)$, $p \in [1, +\infty]$, there exists a constant C depending on n and Y only, such that*

$$\text{(i)} \quad \|\mathcal{Q}_\varepsilon(\phi)\|_{L^p(\mathbb{R}^n)} \le C\|\phi\|_{L^p(\mathbb{R}^n)}, \qquad \text{(ii)} \quad \|\nabla\mathcal{Q}_\varepsilon(\phi)\|_{L^p(\mathbb{R}^n)} \le \frac{C}{\varepsilon}\|\phi\|_{L^p(\mathbb{R}^n)},$$

$$\text{(iii)} \quad \|\mathcal{Q}_\varepsilon(\phi)\|_{L^\infty(\mathbb{R}^n)} \le \frac{C}{\varepsilon^{n/p}}\|\phi\|_{L^p(\mathbb{R}^n)}, \qquad \text{(iv)} \quad \|\nabla\mathcal{Q}_\varepsilon(\phi)\|_{L^\infty(\mathbb{R}^n)} \le \frac{C}{\varepsilon^{1+n/p}}\|\phi\|_{L^p(\mathbb{R}^n)}.$$

Furthermore, for any ψ in $L^p(Y)$,

$$\Big\|\mathcal{Q}_\varepsilon(\phi)\psi\Big(\Big\{\frac{\cdot}{\varepsilon}\Big\}_Y\Big)\Big\|_{L^p(\mathbb{R}^n)} \le C\|\phi\|_{L^p(\mathbb{R}^n)}\|\psi\|_{L^p(Y)}; \tag{4.3}$$

if ψ is in $W^{1,p}_{per}(Y)$, then

$$\Big\|\mathcal{Q}_\varepsilon(\phi)\psi\Big(\Big\{\frac{\cdot}{\varepsilon}\Big\}_Y\Big)\Big\|_{W^{1,p}(\mathbb{R}^n)} \le \frac{C}{\varepsilon}\|\phi\|_{L^p(\mathbb{R}^n)}\|\psi\|_{W^{1,p}(Y)}. \tag{4.4}$$

Proof. By definition, the Q_1 interpolate is Lipschitz-continuous and reaches its maximum at some $\varepsilon\xi$. So, to estimate the L^∞ norm of $\mathcal{Q}_\varepsilon(\phi)$, it suffices to estimate the $\mathcal{Q}_\varepsilon(\phi)(\varepsilon\xi)'s$. By (4.1),

$$|\mathcal{Q}_\varepsilon(\phi)(\varepsilon\xi)|^p \le \frac{1}{|Y|}\int_Y |\phi(\varepsilon\xi + \varepsilon z)|^p\, dz = \frac{1}{\varepsilon^n|Y|}\int_{\varepsilon\xi+\varepsilon Y} |\phi(x)|^p\, dx. \tag{4.5}$$

Since

$$\frac{1}{\varepsilon^n|Y|}\int_{\varepsilon\xi+\varepsilon Y} |\phi(x)|^p\, dx \le \frac{1}{\varepsilon^n|Y|}\|\phi\|^p_{L^p(\mathbb{R}^n)},$$

estimate (iii) follows, with $C = \frac{1}{|Y|^{1/p}}$.

The space $Q_1(Y)$ is of dimension 2^n, hence all of the norms are equivalent. So, there are constants c_1, c_2, and c_3 (depending only upon p and Y) such that, for every $\Phi \in Q_1(Y)$,

$$\|\nabla\Phi\|_{L^\infty(Y)} \le c_1 \sum_{\kappa\in\{0,1\}^n} \left|\Phi\left(\sum_{j=1}^n \kappa_j b_j\right)\right|,$$
$$\|\Phi\|_{L^p(Y)} \le c_2 \left(\sum_{\kappa\in\{0,1\}^n} \left|\Phi\left(\sum_{j=1}^n \kappa_j b_j\right)\right|^p\right)^{1/p},$$
$$\|\nabla\Phi\|_{L^p(Y)} \le c_3 \left(\sum_{\kappa\in\{0,1\}^n} \left|\Phi\left(\sum_{j=1}^n \kappa_j b_j\right)\right|^p\right)^{1/p}.$$

Rescaling these inequalities for $\Phi(y) \doteq \mathcal{Q}_\varepsilon(\phi)(\varepsilon\xi + \varepsilon y)$, gives

$$\|\nabla\mathcal{Q}_\varepsilon(\phi)\|_{L^\infty(\varepsilon\xi+\varepsilon Y)} \le \frac{c_1}{\varepsilon} \sum_{\kappa\in\{0,1\}^n} \left|\mathcal{Q}_\varepsilon(\phi)\left(\varepsilon\xi + \varepsilon\sum_{j=1}^n \kappa_j b_j\right)\right|,$$
$$\|\mathcal{Q}_\varepsilon(\phi)\|_{L^p(\varepsilon\xi+\varepsilon Y)} \le c_2\varepsilon^{n/p} \left(\sum_{\kappa\in\{0,1\}^n} \left|\mathcal{Q}_\varepsilon(\phi)\left(\varepsilon\xi + \varepsilon\sum_{j=1}^n \kappa_j b_j\right)\right|^p\right)^{1/p},$$
$$\|\nabla\mathcal{Q}_\varepsilon(\phi)\|_{L^p(\varepsilon\xi+\varepsilon Y)} \le c_3\varepsilon^{n/p-1} \left(\sum_{\kappa\in\{0,1\}^n} \left|\mathcal{Q}_\varepsilon(\phi)\left(\varepsilon\xi + \varepsilon\sum_{j=1}^n \kappa_j b_j\right)\right|^p\right)^{1/p}.$$

Using (4.5), we have

$$\|\nabla\mathcal{Q}_\varepsilon(\phi)\|_{L^\infty(\mathbb{R}^n)} \le \frac{2^n c_1}{\varepsilon^{1+n/p}|Y|^{1/p}} \|\phi\|_{L^p(\mathbb{R}^n)},$$

which gives (iv). Similarly,

$$\|\mathcal{Q}_\varepsilon(\phi)\|^p_{L^p(\varepsilon\xi+\varepsilon Y)} \le \frac{c_2^p}{|Y|} \sum_{\kappa\in\{0,1\}^n} \int_{\varepsilon\xi+\varepsilon\sum_{j=1}^n \kappa_j b_j+\varepsilon Y} |\phi(x)|^p \, dx,$$

which, by summation on $\xi \in \Xi_\varepsilon$, gives (i), with $C = \frac{(2c_2)^{n/p}}{|Y|^{1/p}}$.

Estimate (ii), with $C = \frac{(2c_3)^{n/p}}{|Y|^{1/p}}$, follows by a similar computation.

To prove (4.3), observe first that the function $\mathcal{Q}_\varepsilon(\phi)\psi(\{\frac{\cdot}{\varepsilon}\}_Y)$ belongs to $L^p(\mathbb{R}^n)$, since $\mathcal{Q}_\varepsilon(\phi)$ is in $L^\infty(\mathbb{R}^n)$ and $\psi(\{\frac{\cdot}{\varepsilon}\}_Y)$ is in $L^p(\mathbb{R}^n)$. Moreover,

$$\left\|\psi\left(\left\{\frac{\cdot}{\varepsilon}\right\}_Y\right)\right\|^p_{L^p(\varepsilon\xi+\varepsilon Y)} = \varepsilon^n \|\psi\|^p_{L^p(Y)},$$

while, by (4.5),

$$\|\mathcal{Q}_\varepsilon(\phi)\|^p_{L^\infty(\varepsilon\xi+\varepsilon Y)} \le \sum_{\kappa\in\{0,1\}^n} \frac{1}{\varepsilon^n|Y|} \int_{\varepsilon\xi+\varepsilon Y+\varepsilon\sum_{j=1}^n \kappa_j b_j} |\phi(x)|^p.$$

Using these two estimates and summing on Ξ_ε gives (4.3), with $C = \frac{2^{n/p}}{|Y|^{1/p}}$.

Estimate (4.4) is obtained in a similar fashion, with $C = \frac{(2)^{n/p}+(2c_3)^{n/p}}{|Y|^{1/p}}$. □

COROLLARY 4.3. *For ϕ in $L^p(\mathbb{R}^n)$, $p \in [1, +\infty[$, the following convergences hold:*

$$\mathcal{Q}_\varepsilon(\phi) \to \phi \quad \textit{strongly in} \quad L^p(\mathbb{R}^n),$$
$$\varepsilon \nabla \mathcal{Q}_\varepsilon(\phi) \to 0 \quad \textit{strongly in} \quad (L^p(\mathbb{R}^n))^n.$$

DEFINITION 4.4. *The remainder $\mathcal{R}_\varepsilon(\phi)$ is given by*

$$\mathcal{R}_\varepsilon(\phi) = \phi - \mathcal{Q}_\varepsilon(\phi) \quad \textit{for any } \phi \in W^{1,p}(\Omega).$$

The following proposition is well-known from the FEM.

PROPOSITION 4.5 (properties of $\mathcal{Q}_\varepsilon$ and $\mathcal{R}_\varepsilon$). *For the case $W_0^{1,p}(\Omega)$, one has*

$$\text{(i)} \quad \|\mathcal{Q}_\varepsilon(\phi)\|_{W^{1,p}(\Omega)} \le C\|\phi\|_{W_0^{1,p}(\Omega)},$$
$$\text{(ii)} \quad \|\mathcal{R}_\varepsilon(\phi)\|_{L^p(\Omega)} \le \varepsilon C\|\phi\|_{W_0^{1,p}(\Omega)},$$
$$\text{(iii)} \quad \|\nabla\mathcal{R}_\varepsilon(\phi)\|_{L^p(\Omega)} \le C\|\nabla\phi\|_{L^p(\Omega)}.$$

The constant C depends on Y (via its diameter and its Poincaré–Wirtinger constant) only, and depends neither on Ω nor on ε.

Similarly, for the case $W^{1,p}(\Omega)$, one has

$$\text{(iv)} \quad \|\mathcal{Q}_\varepsilon(\phi)\|_{W^{1,p}(\Omega)} \le C\,\|\mathcal{P}\|\|\phi\|_{W^{1,p}(\Omega)},$$
$$\text{(v)} \quad \|\mathcal{R}_\varepsilon(\phi)\|_{L^p(\Omega)} \le \varepsilon\, C\,\|\mathcal{P}\|\|\phi\|_{W^{1,p}(\Omega)},$$
$$\text{(vi)} \quad \|\nabla\mathcal{R}_\varepsilon(\phi)\|_{L^p(\Omega)} \le C\,\|\mathcal{P}\|\|\nabla\phi\|_{L^p(\Omega)}.$$

Moreover, in both cases,

$$\text{(4.6)} \qquad \left\| \frac{\partial^2 \mathcal{Q}_\varepsilon(\phi)}{\partial x_i \partial x_j} \right\|_{L^p(\Omega)} \le \frac{C'}{\varepsilon}\|\nabla\phi\|_{L^p(\Omega)} \quad \textit{for } i,j \in [1,\dots,n],\ i \ne j,$$

where C' does not depend on ε.

Proof. We start with ϕ in $W^{1,p}(\mathbb{R}^n)$. From Proposition 2.5(i) and inequality (3.5), we get

$$\text{(4.7)} \quad \|\phi - \mathcal{M}_\varepsilon(\phi)\|_{L^p(\mathbb{R}^n)} = |\,Y\,|^{-\frac{1}{p}}\|\mathcal{T}_\varepsilon(\phi) - \mathcal{M}_\varepsilon(\phi)\|_{L^p(\mathbb{R}^n\times Y)} \le \varepsilon C\|\nabla\phi\|_{L^p(\mathbb{R}^n)}.$$

On the other hand, for any $\psi \in W^{1,p}(\text{interior}(\overline{Y \cup (Y+e_i)}))$, $i \in \{1,\dots,n\}$, we have

$$\begin{aligned} |\,\mathcal{M}_{Y+e_i}(\psi) - \mathcal{M}_Y(\psi)\,| &= |\,\mathcal{M}_Y\big(\psi(\cdot+e_i) - \psi(\cdot)\big)\,| \\ &\le \|\psi(\cdot+e_i) - \psi(\cdot)\|_{L^p(Y)} \le C\|\nabla\psi\|_{L^p(Y\cup(Y+e_i))}. \end{aligned}$$

By a scaling argument and using Definition 4.1, this gives

$$\text{(4.8)} \qquad |\mathcal{Q}_\varepsilon(\phi)(\varepsilon\xi) - \mathcal{Q}_\varepsilon(\phi)(\varepsilon\xi + \varepsilon e_i)| \le \varepsilon C\|\nabla\phi\|_{L^p}(\varepsilon(\xi+Y)\cup\varepsilon(\xi+e_i+Y))$$

for all $\xi \in \varepsilon\mathbb{Z}^n$.

Let $x \in \varepsilon\big(\xi + Y\big)$, and set for every $\kappa = (\kappa_1,\dots,\kappa_n) \in \{0,1\}^n$,

$$\overline{x}_l^{(\kappa_l)} = \begin{cases} \dfrac{x_l - \xi_l}{\varepsilon} & \text{if} \quad \kappa_l = 1, \\ 1 - \dfrac{x_l - \xi_l}{\varepsilon} & \text{if} \quad \kappa_l = 0. \end{cases}$$

If $\xi \in \varepsilon\mathbb{Z}^n$, for every $\kappa \in \{0,1\}^n$, by definition we have

$$\mathcal{Q}_\varepsilon\big(\phi\big)(x) = \sum_{\kappa\in\{0,1\}^n} \mathcal{Q}_\varepsilon(\phi)\big(\varepsilon\xi + \varepsilon\kappa\big)\overline{x}_1^{(\kappa_1)} \dots \overline{x}_n^{(\kappa_n)}, \tag{4.9}$$

and so, for example,

$$\frac{\partial \mathcal{Q}_\varepsilon(\phi)}{\partial x_1}(x)$$
$$= \sum_{\kappa_2,\,\dots,\kappa_n} \frac{\mathcal{Q}_\varepsilon(\phi)\big(\varepsilon\xi + \varepsilon(1,\kappa_2,\dots,\kappa_n)\big) - \mathcal{Q}_\varepsilon(\phi)\big(\varepsilon\xi + \varepsilon(0,\kappa_2,\dots,\kappa_n)\big)}{\varepsilon}\overline{x}_2^{(\kappa_2)} \dots \overline{x}_n^{(\kappa_n)},$$

and a same expression for the other derivatives. This last formula and (4.7)–(4.9) imply estimate (i) written in $\mathbb{R}^n$.

Now, from (4.9), we get

$$\phi(x) - \mathcal{Q}_\varepsilon\big(\phi\big)(x) = \sum_{\kappa\in\{0,1\}^n} \Big(\phi(x) - \mathcal{Q}_\varepsilon(\phi)\big(\varepsilon\xi + \varepsilon\kappa\big)\Big)\overline{x}_1^{(\kappa_1)} \dots \overline{x}_n^{(\kappa_n)},$$

and (ii) (in $\mathbb{R}^n$) follows by using estimate (4.7). Estimate (iii) (again in $\mathbb{R}^n$) is straightforward from the previous ones.

If ϕ is in $W_0^{1,p}(\Omega)$, let $\widetilde{\phi}$ be its extension to the whole of $\mathbb{R}^n$. To derive (i)–(iii), it suffices to write down the estimates in $\mathbb{R}^n$ obtained above. Similarly, applying them to $\mathcal{P}(\phi)$ for ϕ in $W^{1,p}(\Omega)$ gives (iv)–(vi).

To finish the proof, it remains to show estimate (4.6) . To do so, it is enough to take the derivative with respect to any x_k, with $k \neq 1$ in the formula of $\frac{\partial \mathcal{Q}_\varepsilon(\phi)}{\partial x_1}$ above, and use estimate (4.8). ∎

Remark 4.6. By construction (see explicit formula (4.9)), the function $\mathcal{Q}_\varepsilon(\phi)$ is separately piecewise linear on each cell. Observe also that, for any $k \in \{1,\dots,n\}$, $\frac{\partial \mathcal{Q}_\varepsilon(\phi)}{\partial x_k}$ is independent of x_k in each cell $\varepsilon\big(\xi + Y\big)$.

PROPOSITION 4.7. *Let $\{w_\varepsilon\}$ be a sequence converging weakly in $W_0^{1,p}(\Omega)$ (resp. $W^{1,p}(\Omega)$) to w. Then, the following convergences hold:*

$$\text{(i)}\quad \mathcal{R}_\varepsilon(w_\varepsilon) \to 0 \quad \textit{strongly in } L^p(\Omega),$$
$$\text{(ii)}\quad \mathcal{Q}_\varepsilon(w_\varepsilon) \rightharpoonup w \quad \textit{weakly in } W^{1,p}(\Omega),$$
$$\text{(iii)}\quad \mathcal{T}_\varepsilon(\nabla \mathcal{Q}_\varepsilon(w_\varepsilon)) \rightharpoonup \nabla w \quad \textit{weakly in } L^p(\Omega\times Y).$$

Proof. Convergence (i) is a direct consequence of estimate (ii) (resp. (v)) of Proposition 4.5, and it implies convergence (ii). Together with (i), Proposition 2.9(ii) implies $\mathcal{T}_\varepsilon(\mathcal{Q}_\varepsilon(w_\varepsilon)) \rightharpoonup w$ weakly in $L^p(\Omega\times Y)$. From (4.5),

$$\left\| \frac{\partial}{\partial x_i}\left(\frac{\partial \mathcal{Q}_\varepsilon(w_\varepsilon)}{\partial x_j}\right)\right\|_{L^p(\Omega)} \leq \frac{C}{\varepsilon} \quad \text{for } i,j \in [1,\dots,n],\ \ i \neq j.$$

Now, by Proposition 3.1, there exist a subsequence (still denoted ε) and $\widehat{w}_j \in L^p(\Omega\times Y)$, with $\frac{\partial \widehat{w}_j}{\partial y_i} \in L^p(\Omega\times Y)$ such that, for $i \neq j$,

$$\mathcal{T}_\varepsilon\left(\frac{\partial \mathcal{Q}_\varepsilon(w_\varepsilon)}{\partial x_j}\right) \rightharpoonup \widehat{w}_j \quad \text{weakly in } L^p(\Omega\times Y),$$
$$\varepsilon\mathcal{T}_\varepsilon\left(\frac{\partial^2 \mathcal{Q}_\varepsilon(w_\varepsilon)}{\partial x_i \partial x_j}\right) \rightharpoonup \frac{\partial \widehat{w}_j}{\partial y_i} \quad \text{weakly in } L^p(\Omega\times Y),$$

where $\widehat{w}_j$ is y_i-periodic for every $i \neq j$. Moreover, from Remark 4.6, the function $\widehat{w}_j$ does not depend on y_j, hence it is Y-periodic. But, by Remark 4.6 again, $\widehat{w}_j$ is also piecewise linear, with respect to any variable y_i. Consequently, $\widehat{w}_j$ is independent of y. On the other hand, from (ii) above we have

$$\frac{\partial \mathcal{Q}_\varepsilon(w_\varepsilon)}{\partial x_j} \rightharpoonup \frac{\partial w}{\partial x_j} \quad \text{weakly in } L^p(\Omega).$$

Now Proposition 2.9(iii) gives $\widehat{w}_j = \frac{\partial w}{\partial x_j}$, and convergence (iii) holds for the whole sequence ε. □

PROPOSITION 4.8 (Theorem 3.5 revisited). *Let $\{w_\varepsilon\}$ be a sequence converging weakly in $W_0^{1,p}(\Omega)$ (resp. in $W^{1,p}(\Omega)$) to w. Then, up to a subsequence there exists some $\widehat{w}'$ in the space $L^p(\Omega; W^{1,p}_{per}(Y))$ such that the following convergences hold:*

$$\frac{1}{\varepsilon}\mathcal{T}_\varepsilon\big(\mathcal{R}_\varepsilon(w_\varepsilon)\big) \rightharpoonup \widehat{w}' \quad \textit{weakly in } L^p(\Omega; W^{1,p}(Y)),$$
$$\mathcal{T}_\varepsilon\big(\nabla \mathcal{R}_\varepsilon(w_\varepsilon)\big) \rightharpoonup \nabla_y \widehat{w}' \quad \textit{weakly in } L^p(\Omega \times Y),$$
$$\mathcal{T}_\varepsilon(\nabla w_\varepsilon) \rightharpoonup \nabla w + \nabla_y \widehat{w}' \quad \textit{weakly in } L^p(\Omega \times Y).$$

Actually, the connection with the $\widehat{w}$ of Theorem 3.5 is given by

$$\widehat{w} = \widehat{w}' - \mathcal{M}_{_Y}(\widehat{w}').$$

Proof. Due to estimates of Proposition 4.5, up to a subsequence, there exists $\widehat{w}'$ in $L^p(\Omega; W^{1,p}_{per}(Y))$ such that

$$\frac{1}{\varepsilon}\mathcal{T}_\varepsilon\big(\mathcal{R}_\varepsilon(w_\varepsilon)\big) \rightharpoonup \widehat{w}' \quad \text{weakly in } L^p(\Omega; W^{1,p}(Y)),$$
$$\mathcal{T}_\varepsilon\big(\nabla \mathcal{R}_\varepsilon(w_\varepsilon)\big) \rightharpoonup \nabla_y \widehat{w}' \quad \text{weakly in } L^p(\Omega \times Y).$$

Combining with convergence (iii) of Proposition 4.7 shows that

$$\mathcal{T}_\varepsilon\big(\nabla w_\varepsilon\big) \rightharpoonup \nabla w + \nabla_y \widehat{w}' \quad \text{weakly in } L^p(\Omega \times Y).$$

So $\nabla_y \widehat{w} \equiv \nabla_y \widehat{w}'$ in $L^p(\Omega \times Y)$. Since $\mathcal{M}_{_Y}(\widehat{w}) = 0$, it follows that $\widehat{w} = \widehat{w}' - \mathcal{M}_{_Y}(\widehat{w}')$. □

Remark 4.9. In the previous proposition, one can actually compute the average of $\widehat{w}'$. One can check that $\mathcal{M}_{_Y}(\widehat{w}') = -\mathcal{M}_{_Y}(y) \cdot \nabla w$, and consequently,

$$\frac{1}{\varepsilon}\Big(\mathcal{T}_\varepsilon(w_\varepsilon) - \mathcal{M}_\varepsilon(w_\varepsilon)\Big) \rightharpoonup y \cdot \nabla w + \widehat{w}' \qquad \text{weakly in } L^p(\Omega; W^{1,p}(Y)).$$

5. Periodic unfolding and the standard homogenization problem.

DEFINITION 5.1. *Let $\alpha, \beta \in \mathbb{R}$, such that $0 < \alpha < \beta$ and $\mathcal{O}$ be an open subset of $\mathbb{R}^n$. Denote by $M(\alpha, \beta, \mathcal{O})$ the set of the $n \times n$ matrices $A = (a_{ij})_{1 \leq i,j \leq n} \in (L^\infty(\mathcal{O}))^{n \times n}$ such that, for any $\lambda \in \mathbb{R}^n$ and a.e. on $\mathcal{O}$,*

$$(A(x)\lambda, \lambda) \geq \alpha|\lambda|^2, \qquad |A(x)\lambda| \leq \beta|\lambda|.$$

Let $A^\varepsilon = (a_{ij}^\varepsilon)_{1\le i,j\le n}$ be a sequence of matrices in $M(\alpha,\beta,\Omega)$. For f given in $H^{-1}(\Omega)$, consider the Dirichlet problem

$$\begin{cases} -\text{div } (A^\varepsilon \nabla u_\varepsilon) = f & \text{in } \Omega, \\ u^\varepsilon = 0 & \text{on } \partial\Omega. \end{cases} \tag{5.1}$$

By the Lax–Milgram theorem, there exists a unique $u^\varepsilon \in H_0^1(\Omega)$ satisfying

$$\int_\Omega A^\varepsilon \nabla u_\varepsilon \, \nabla v \, dx = \langle f, v\rangle_{H^{-1}(\Omega),H_0^1(\Omega)}, \quad \forall v \in H_0^1(\Omega), \tag{5.2}$$

which is the variational formulation of (5.1). Moreover, one has the apriori estimate

$$\|u_\varepsilon\|_{H_0^1(\Omega)} \le \frac{1}{\alpha}\|f\|_{H^{-1}(\Omega)}. \tag{5.3}$$

Consequently, there exist u_0 in $H_0^1(\Omega)$ and a subsequence, still denoted ε, such that

$$u_\varepsilon \rightharpoonup u_0 \quad \text{weakly in } H_0^1(\Omega). \tag{5.4}$$

We are now interested to give a limit problem, the "homogenized" problem, satisfied by u_0. This is called standard homogenization, and the answer, for some classes of A^ε, can be found in many works, starting with the classical book by Bensoussan, Lions, and Papanicolaou [11] (see, for instance, Cioranescu and Donato [30] and the references herein). We now recall it.

THEOREM 5.2 (standard periodic homogenization). *Let $A = (a_{ij})_{1\le i,j\le n}$ belong to $M(\alpha,\beta,Y)$, where $a_{ij} = a_{ij}(y)$ are Y-periodic. Set*

$$A^\varepsilon(x) = \left(a_{ij}\left(\frac{x}{\varepsilon}\right)\right)_{1\le i,j\le n} \quad \text{a.e. on } \Omega. \tag{5.5}$$

Let u_ε be the solution of the corresponding problem (5.1), with f in $H^{-1}(\Omega)$. Then the whole sequence $\{u_\varepsilon\}$ converges to a limit u_0, which is the unique solution of the homogenized problem

$$\begin{cases} -\text{div } (A^0 \nabla u_0) = \displaystyle\sum_{i,j=1}^n a_{ij}^0 \frac{\partial^2 u_0}{\partial x_i \partial x_j} = f & \text{in } \Omega, \\ u_0 = 0 & \text{on } \partial\Omega, \end{cases} \tag{5.6}$$

where the constant matrix $A^0 = (a_{ij}^0)_{1\le i,j\le n}$ is elliptic and given by

$$a_{ij}^0 = \mathcal{M}_Y\left(a_{ij} - \sum_{k=1}^n a_{ik}\frac{\partial \widehat{\chi}_j}{\partial y_k}\right) = \mathcal{M}_Y(a_{ij}) - \mathcal{M}_Y\left(\sum_{k=1}^n a_{ik}\frac{\partial \widehat{\chi}_j}{\partial y_k}\right). \tag{5.7}$$

In (5.7), the functions $\widehat{\chi}_j$ $(j = 1,\dots,n)$, often referred to as correctors, are the solutions of the cell systems

$$\begin{cases} -\displaystyle\sum_{i,k=1}^n \frac{\partial}{\partial y_i}\left(a_{ik}\frac{\partial(\widehat{\chi}_j - y_j)}{\partial y_k}\right) = 0 & \text{in } Y, \\ \mathcal{M}_Y(\widehat{\chi}_j) = 0, \\ \widehat{\chi}_j \quad Y\text{-periodic}. \end{cases} \tag{5.8}$$

As will be seen below, using the periodic unfolding, the proof of this theorem is elementary! Actually, with the same proof, a more general result can be obtained, with matrices A^ε.

THEOREM 5.3 (periodic homogenization via unfolding). *Let u_ε be the solution of problem (5.1), with f in $H^{-1}(\Omega)$, and $A^\varepsilon = (a_{ij}^\varepsilon)_{1\le i,j\le n}$ be a sequence of matrices in $M(\alpha,\beta,\Omega)$. Suppose that there exists a matrix B such that*

$$B^\varepsilon \doteq \mathcal{T}_\varepsilon\big(A^\varepsilon\big) \to B \quad \text{strongly in } [L^1(\Omega\times Y)]^{n\times n}. \tag{5.9}$$

Then there exists $u_0 \in H_0^1(\Omega)$ and $\widehat{u} \in L^2(\Omega; H^1_{per}(Y))$ such that

$$\begin{aligned} &u_\varepsilon \rightharpoonup u_0 \quad \text{weakly in } H_0^1(\Omega),\\ &\mathcal{T}_\varepsilon(u_\varepsilon) \rightharpoonup u_0 \quad \text{weakly in } L^2(\Omega; H^1(Y)),\\ &\mathcal{T}_\varepsilon(\nabla u_\varepsilon) \rightharpoonup \nabla u_0 + \nabla_y \widehat{u} \quad \text{weakly in } L^2(\Omega\times Y), \end{aligned} \tag{5.10}$$

and the pair $(u_0, \widehat{u})$ is the unique solution of the problem

$$\begin{cases} \forall \Psi \in H_0^1(\Omega),\ \forall \Phi \in L^2(\Omega;\, H^1_{per}(Y)),\\ \dfrac{1}{|Y|}\displaystyle\int_{\Omega\times Y} B(x,y)\big[\nabla u_0(x) + \nabla_y \widehat{u}(x,y)\big]\big[\nabla \Psi(x) + \nabla_y \Phi(x,y)\big]\, dxdy \\ \qquad\qquad = \ \langle f, \Psi\rangle_{H^{-1}(\Omega), H_0^1(\Omega)}. \end{cases} \tag{5.11}$$

Remark 5.4. System (5.11) is the unfolded formulation of the homogenized limit problem. It is of standard variational form in the space

$$\mathcal{H} = H_0^1(\Omega) \times L^2(\Omega;\, H^1_{per}(Y)/\mathbb{R}).$$

Remark 5.5. Hypothesis (5.9) implies that $B \in M(\alpha,\beta,\Omega\times Y)$.

Remark 5.6. If A^ε is of the form (5.5), then $B(x,y) = A(y)$. In the case where $A^\varepsilon(x) = A_1(x)A_2(\frac{x}{\varepsilon})$, one has (5.9), with $B(x,y) = A_1(x)A_2(y)$.

Remark 5.7. Let us point out that every matrix $B \in M(\alpha,\beta,\Omega\times Y)$ can be approached by the sequence of matrices A^ε in $M(\alpha,\beta,\Omega)$, with A^ε defined as follows:

$$A^\varepsilon = \begin{cases} \mathcal{U}_\varepsilon(B) & \text{in } \widehat{\Omega}_\varepsilon,\\ \alpha I_n & \text{in } \Lambda_\varepsilon. \end{cases}$$

Proof of Theorem 5.3. Convergences (5.10) follow from estimate (5.3), Corollary 3.3, and Proposition 4.7, respectively.

Let us choose $v = \Psi$, with $\Psi \in H_0^1(\Omega)$ as test function in (5.2). The integration formula (2.5) from Proposition 2.7 gives

$$\frac{1}{|Y|}\int_{\Omega\times Y} B^\varepsilon\ \mathcal{T}_\varepsilon\big(\nabla u_\varepsilon\big)\mathcal{T}_\varepsilon\big(\nabla\Psi\big)\, dxdy \overset{\mathcal{T}_\varepsilon}{\simeq}\ \langle f, \Psi\rangle_{H^{-1}(\Omega), H_0^1(\Omega)}. \tag{5.12}$$

We are allowed to pass to the limit in (5.12), due to (5.9), (5.10), and Proposition 2.9(i), to get

$$\frac{1}{|Y|}\int_{\Omega\times Y} B(x,y)\big[\nabla u_0(x) + \nabla_y \widehat{u}(x,y)\big]\nabla\Psi(x)\, dxdy = \ \langle f, \Psi\rangle_{H^{-1}(\Omega), H_0^1(\Omega)}. \tag{5.13}$$

Now, taking in (5.2), as test function $v^\varepsilon(x) = \varepsilon\Psi(x)\psi(\frac{x}{\varepsilon})$, $\Psi \in \mathcal{D}(\Omega)$, $\psi \in H^1_{per}(Y)$, one has, due to (2.5) and Proposition 2.7,

$$\frac{1}{|Y|}\int_{\Omega\times Y} B^\varepsilon\ \mathcal{T}_\varepsilon\big(\nabla u_\varepsilon\big)\,\varepsilon\psi(y)\mathcal{T}_\varepsilon\big(\nabla\Psi\big)\,dxdy + \frac{1}{|Y|}\int_{\Omega\times Y} B^\varepsilon\ \mathcal{T}_\varepsilon\big(\nabla u_\varepsilon\big)\nabla_y\psi(y)\mathcal{T}_\varepsilon\big(\Psi\big)\,dxdy \overset{\mathcal{T}_\varepsilon}{\simeq}\ \langle f, v_\varepsilon\rangle_{H^{-1}(\Omega),H^1_0(\Omega)}.$$

Since $v^\varepsilon \rightharpoonup 0$ in $H^1_0(\Omega)$, we get at the limit

$$\frac{1}{|Y|}\int_{\Omega\times Y} B(x,y)\big[\nabla u_0(x) + \nabla_y\widehat{u}(x,y)\big]\Psi(x)\nabla_y\psi(y)\,dxdy = 0.$$

By the density of the tensor product $\mathcal{D}(\Omega)\otimes H^1_{per}(Y)$ in $L^2(\Omega; H^1_{per}(Y))$, this holds for all Φ in $L^2(\Omega; H^1_{per}(Y))$. □

Remark 5.8. As in the two-scale method, (5.11) gives $\widehat{u}$ in terms of ∇u_0 and yields the standard form of the homogenized equation, i.e., (5.6). In the simple case where $A(x,y) = A(y) = (a_{ij}(y))_{1\le i,j\le n}$, it is easily seen that

$$\widehat{u} = \sum_{i=1}^n \frac{\partial u_0}{\partial x_i}\widehat{\chi}_i, \tag{5.14}$$

and that the limit B is precisely the matrix A^0 which was defined in Theorem 5.2 by (5.7) and (5.8).

PROPOSITION 5.9 (convergence of the energy). *Under the hypotheses of Theorem* 5.3*,*

$$\begin{cases} \lim\limits_{\varepsilon\to 0}\displaystyle\int_\Omega A^\varepsilon\nabla u_\varepsilon\nabla u_\varepsilon\,dx = \frac{1}{|Y|}\int_{\Omega\times Y} B\big[\nabla u_0 + \nabla_y\widehat{u}\big]\,\big[\nabla u_0 + \nabla_y\widehat{u}\big]\,dx\,dy, \\ \lim\limits_{\varepsilon\to 0}\displaystyle\int_{\Lambda_\varepsilon} |\nabla u_\varepsilon|^2\,dx = 0. \end{cases} \tag{5.15}$$

Proof. By standard weak lower-semicontinuity, one successively obtains

$$\begin{aligned} \frac{1}{|Y|}\int_{\Omega\times Y} B\big[\nabla u_0 + \nabla_y\widehat{u}\big]\,\big[\nabla u_0 + \nabla_y\widehat{u}\big]\,dx\,dy & \\ \le \liminf_{\varepsilon\to 0}\frac{1}{|Y|}\int_{\Omega\times Y} B^\varepsilon\ \mathcal{T}_\varepsilon\big(\nabla u_\varepsilon\big)\mathcal{T}_\varepsilon\big(\nabla u_\varepsilon\big)\,dx\,dy & \\ \le \limsup_{\varepsilon\to 0}\frac{1}{|Y|}\int_{\Omega\times Y} B^\varepsilon\ \mathcal{T}_\varepsilon\big(\nabla u_\varepsilon\big)\mathcal{T}_\varepsilon\big(\nabla u_\varepsilon\big)\,dx\,dy & \\ \le \limsup_{\varepsilon\to 0}\int_\Omega A^\varepsilon\ \nabla u_\varepsilon\nabla u_\varepsilon\,dx = \limsup_{\varepsilon\to 0}\langle f, u_\varepsilon\rangle_{H^{-1}(\Omega),H^1_0(\Omega)} & \\ = \langle f, u_0\rangle_{H^{-1}(\Omega),H^1_0(\Omega)} = \frac{1}{|Y|}\int_{\Omega\times Y} B\big[\nabla u_0 + \nabla_y\widehat{u}\big]\,\big[\nabla u_0 + \nabla_y\widehat{u}\big]\,dx\,dy, & \end{aligned}$$

which gives the first convergence in (5.15), as well as

$$\limsup_{\varepsilon\to 0}\int_{\Lambda_\varepsilon} A^\varepsilon\ \nabla u_\varepsilon\nabla u_\varepsilon\,dx = 0,$$

which implies the second convergence in (5.15). □

Remark 5.10. From the above proof, we also have

$$\lim_{\varepsilon\to 0}\frac{1}{|Y|}\int_{\Omega\times Y} B^\varepsilon\ \mathcal{T}_\varepsilon\big(\nabla u_\varepsilon\big)\mathcal{T}_\varepsilon\big(\nabla u_\varepsilon\big)\ dx\,dy = \frac{1}{|Y|}\int_{\Omega\times Y} B\big[\nabla u_0+\nabla_y\widehat{u}\big]\ \big[\nabla u_0+\nabla_y\widehat{u}\big]\ dx\,dy.$$

COROLLARY 5.11. *The following strong convergence holds:*

$$\mathcal{T}_\varepsilon(\nabla u_\varepsilon)\to\nabla u_0+\nabla_y\widehat{u}\quad \textit{strongly in } L^2(\Omega\times Y). \tag{5.16}$$

Proof. We have successively

$$\begin{aligned}\frac{1}{|Y|}\int_{\Omega\times Y} B^\varepsilon\big[\mathcal{T}_\varepsilon\big(\nabla u_\varepsilon\big)-\nabla u_0-\nabla_y\widehat{u}\big]\big[\mathcal{T}_\varepsilon\big(\nabla u_\varepsilon\big)-\nabla u_0-\nabla_y\widehat{u}\big]\ dx\,dy&\\ = \frac{1}{|Y|}\int_{\Omega\times Y} B^\varepsilon\mathcal{T}_\varepsilon\big(\nabla u_\varepsilon\big)\mathcal{T}_\varepsilon\big(\nabla u_\varepsilon\big)\ dx\,dy&\\ -\frac{1}{|Y|}\int_{\Omega\times Y} B^\varepsilon\big[\nabla u_0+\nabla_y\widehat{u}\big]\mathcal{T}_\varepsilon\big(\nabla u_\varepsilon\big)\ dx\,dy&\\ -\frac{1}{|Y|}\int_{\Omega\times Y} B^\varepsilon\mathcal{T}_\varepsilon\big(\nabla u_\varepsilon\big)\big[\nabla u_0+\nabla_y\widehat{u}\big]\ dx\,dy&\\ +\frac{1}{|Y|}\int_{\Omega\times Y} B^\varepsilon\big[\nabla u_0+\nabla_y\widehat{u}\big]\ \big[\nabla u_0+\nabla_y\widehat{u}\big]\ dx\,dy.&\end{aligned}$$

Each term in the right-hand side converges, the first one due to Remark 5.10, and the others due to (5.10) and hypothesis (5.9). So, the right-hand side term converges to zero. Then convergence (5.16) is a consequence of the ellipticity of B^ε. ∎

Remark 5.12. One can consider problem (5.1) with a homogeneous Neumann boundary condition on $\partial\Omega$ provided a zero order term is added to the operator. This problem is variational on the space $H^1(\Omega)$ without any regularity condition on the boundary. The exact same method applies and gives the corresponding limit problem. In order for a nonhomogeneous Neumann boundary condition (or Robin condition) on $\partial\Omega$ to make sense, a well-behaved trace operator is needed from $H^1(\Omega)$ to $L^2(\Omega)$. In that case, the same method applies.

6. Some corrector results and error estimates. Under additional regularity assumptions on the homogenized solution u_0 and the cell-functions $\widehat{\chi}_j$, the strong convergence for the gradient of u_0 with a corrector is known (cf. [11] Chapter 1, section 5, [30] Chapter 8, section 3 and references therein). More precisely, suppose that $\nabla_y\widehat{\chi}_j\in(L^r(Y))^n,\ j=1,\dots,n$ and $\nabla u^0\in L^s(\Omega)$, with $1\le r,s<+\infty$ and such that $1/r+1/s=1/2$. Then

$$\nabla u_\varepsilon-\nabla u_0-\sum_{j=1}^n\frac{\partial u_0}{\partial x_j}\ \big(\nabla_y\widehat{\chi}_j\big)\Big(\frac{\cdot}{\varepsilon}\Big)\to 0\quad\text{strongly in } L^2(\Omega).$$

Our next result gives a corrector result *without any additional regularity assumption* on $\widehat{\chi}_j$, and its proof reduces to a few lines. We also include a new type of corrector.

THEOREM 6.1. *Under the hypotheses of Theorem* 5.2*, one has*

$$\nabla u_\varepsilon-\nabla u_0-\mathcal{U}_\varepsilon\big(\nabla_y\widehat{u}\big)\to 0\quad \textit{strongly in } L^2(\Omega). \tag{6.1}$$

In the case where $A^\varepsilon(x) = A(\{\frac{x}{\varepsilon}\}_Y)$, *the function* $u_0 + \varepsilon \sum_{i=1}^n \mathcal{Q}_\varepsilon(\frac{\partial u_0}{\partial x_i})\chi_i(\{\frac{\cdot}{\varepsilon}\}_Y)$ *belongs to* $H^1(\Omega)$, *and one has*

$$u_\varepsilon - u_0 - \varepsilon \sum_{i=1}^n \mathcal{Q}_\varepsilon\left(\frac{\partial u_0}{\partial x_i}\right)\chi_i\left(\left\{\frac{\cdot}{\varepsilon}\right\}_Y\right) \to 0 \quad \text{strongly in } H^1(\Omega). \tag{6.2}$$

Proof. From (5.15), (5.16), and Proposition 2.18(iii), one immediately has

$$\nabla u_\varepsilon - \mathcal{U}_\varepsilon\big(\nabla u_0\big) - \mathcal{U}_\varepsilon\big(\nabla_y \widehat{u}\big) \to 0 \quad \text{strongly in } L^2(\Omega).$$

But since ∇u_0 belongs to $L^2(\Omega)$, Corollary 2.26 implies that

$$\mathcal{U}_\varepsilon\big(\nabla u_0\big) \to \nabla u_0 \quad \text{strongly in } L^2(\Omega),$$

whence (6.1). From (4.4) in Proposition 4.2, the function $u_0+\varepsilon \sum_{i=1}^n \mathcal{Q}_\varepsilon(\frac{\partial u_0}{\partial x_i})\chi_i(\{\frac{\cdot}{\varepsilon}\}_Y)$ belongs to $H^1(\Omega)$. From (5.14), we obtain

$$\begin{aligned}
\nabla u_0 + \mathcal{U}_\varepsilon\left(\nabla_y \widehat{u}\right) - \nabla\left[u_0 + \varepsilon \sum_{i=1}^n \mathcal{Q}_\varepsilon\left(\frac{\partial u_0}{\partial x_i}\right)\chi_i\left(\left\{\frac{\cdot}{\varepsilon}\right\}_Y\right)\right] \\
= -\sum_{i=1}^n \left[\mathcal{Q}_\varepsilon\left(\frac{\partial u_0}{\partial x_i}\right) - \mathcal{M}_\varepsilon\left(\frac{\partial u_0}{\partial x_i}\right)\right]\nabla_y \chi_i\left(\left\{\frac{\cdot}{\varepsilon}\right\}_Y\right) \\
- \varepsilon \sum_{i=1}^n \nabla\left[\mathcal{Q}_\varepsilon\left(\frac{\partial u_0}{\partial x_i}\right)\right]\chi_i\left(\left\{\frac{\cdot}{\varepsilon}\right\}_Y\right),
\end{aligned}$$

and using estimate (4.2), Proposition 2.25(iii), and Corollary 4.3, one immediately gets the strong convergence in $L^2(\Omega)$ of the right-hand side in the above equality. Thanks to (6.1), one has (6.2). ∎

We end this section by recalling the error estimates obtained by Griso in [42], [44], and [45] for problem (5.1), with $f \in L^2(\Omega)$.

THEOREM 6.2 (see [42], [44]). *Suppose that* $\partial\Omega$ *is of class* $\mathcal{C}^{1,1}$. *The solution* u_ε *of* (5.1) *satisfies the following estimates:*

$$\left\|\nabla u_\varepsilon - \nabla u_0 - \sum_{i=1}^n \mathcal{Q}_\varepsilon\left(\frac{\partial u_0}{\partial x_i}\right)\nabla_y \widehat{\chi}_i\left(\left\{\frac{\cdot}{\varepsilon}\right\}\right)\right\|_{[L^2(\Omega)]^n} \le C\varepsilon^{1/2}\|f\|_{L^2(\Omega)},$$

$$\|u_\varepsilon - u_0\|_{L^2(\Omega)} + \left\|\rho\left(\nabla u_\varepsilon - \nabla u_0 - \sum_{i=1}^n \mathcal{Q}_\varepsilon\left(\frac{\partial u_0}{\partial x_i}\right)\nabla_y \widehat{\chi}_i\left(\frac{\cdot}{\varepsilon}\right)\right)\right\|_{[L^2(\Omega)]^n} \le C\varepsilon\|f\|_{L^2(\Omega)},$$

where $\widehat{\chi}_i$ *for* $i = 1, \dots, n$ *is defined by* (5.8) *and* $\rho = \rho(x)$ *is the distance between* $x \in \Omega$ *and the boundary* $\partial\Omega$. *The constant* C *depends on* n, A, *and* $\partial\Omega$.

COROLLARY 6.3 (see [44]). *Let* Ω' *be an open set strongly included in* Ω, *then*

$$\left\|u_\varepsilon - u_0 - \varepsilon \sum_{i=1}^n \mathcal{Q}_\varepsilon\left(\frac{\partial u_0}{\partial x_i}\right)\widehat{\chi}_i\left(\frac{\cdot}{\varepsilon}\right)\right\|_{H^1(\Omega')} \le C\varepsilon\|f\|_{L^2(\Omega)}.$$

The constant depends on n, A, Ω', *and* $\partial\Omega$.

In what follows in this paragraph, we suppose that the open set Ω is a bounded domain in $\mathbb{R}^n$, $n = 2$ or 3, of polygonal ($n = 2$) or polyhedral ($n = 3$) boundary. We

assume that Ω is on one side only of its boundary, and that Γ_0 is the union of some edges ($n = 2$) or some faces ($n = 3$) of $\partial\Omega$. Recall that classical regularity results show that the solution u_0 of the homogenized problem (5.6) belongs to $H^{1+s}(\Omega)$ for s in $]1/2, 1[$ ($s = 1$ if the domain is convex) depending only on $\partial\Omega$, on A^0, and satisfies the estimate

$$\|\nabla u_0\|_{H^{1+s}(\Omega)} \leq C\|f\|_{L^2(\Omega)}.$$

The error estimate for this case is given in the following result.

THEOREM 6.4 (see [45]). *The solution u_ε of problem (5.1) satisfies the estimate*

$$\left\|\nabla u_\varepsilon - \nabla u_0 - \sum_{i=1}^{n} \mathcal{Q}_\varepsilon\left(\frac{\partial u_0}{\partial x_i}\right) \nabla_y \widehat{\chi}_i\left(\frac{\cdot}{\varepsilon}\right)\right\|_{[L^2(\Omega)]^n} \leq C\varepsilon^{s/2}\|f\|_{L^2(\Omega)},$$

$$\|u_\varepsilon - u_0\|_{L^2(\Omega)} + \left\|\rho\left(\nabla u_\varepsilon - \nabla u_0 - \sum_{i=1}^{n} \mathcal{Q}_\varepsilon\left(\frac{\partial u_0}{\partial x_i}\right) \nabla_y \widehat{\chi}_i\left(\frac{\cdot}{\varepsilon}\right)\right)\right\|_{[L^2(\Omega)]^n} \leq C\varepsilon^{s}\|f\|_{L^2(\Omega)}.$$

The constants depend on n, A, and $\partial\Omega$.

COROLLARY 6.5 (see [45]). *Let Ω' be an open set strongly included in Ω, then*

$$\left\|u_\varepsilon - u_0 - \varepsilon \sum_{i=1}^{n} \mathcal{Q}_\varepsilon\left(\frac{\partial u_0}{\partial x_i}\right) \widehat{\chi}_i\left(\frac{\cdot}{\varepsilon}\right)\right\|_{H^1(\Omega')} \leq C\varepsilon^{s}\|f\|_{L^2(\Omega)}.$$

The constant depends on n, A, Ω', and $\partial\Omega$.

7. Periodic unfolding and multiscales. As we mentioned in the Introduction, the periodic unfolding method turns out to be particularly well-adapted to multiscales problems. As an example, we treat here a problem with two different small scales.

Consider two periodicity cells Y and Z, both having the properties introduced at the beginning of section 2 (each associated to its set of periods). Suppose that Y is "partitioned" in two nonempty disjoint open subsets Y_1 and Y_2, i.e., such that $Y_1 \cap Y_2 = \emptyset$ and $\overline{Y} = \overline{Y}_1 \cup \overline{Y}_2$.

Let $A^{\varepsilon\delta}$ be a matrix field defined by

$$A^{\varepsilon\delta}(x) = \begin{cases} A_1\left(\left\{\dfrac{x}{\varepsilon}\right\}_Y\right) & \text{for} \quad \left\{\dfrac{x}{\varepsilon}\right\}_Y \in Y_1, \\ A_2\left(\left\{\dfrac{\left\{\frac{x}{\varepsilon}\right\}_Y}{\delta}\right\}_Z\right) & \text{for} \quad \left\{\dfrac{x}{\varepsilon}\right\}_Y \in Y_2, \end{cases}$$

where A_1 is in $M(\alpha, \beta, Y_1)$ and A_2 in $M(\alpha, \beta, Z)$ (cf. Definition 5.1). Here we have two small scales, namely, ε and $\varepsilon\delta$, associated, respectively, to the cells Y and Z (see Figure 7).

Consider the problem

$$\int_\Omega A^{\varepsilon\delta}\nabla u_{\varepsilon\delta}\nabla w \, dx = \int_\Omega f\, w \, dx \qquad \forall w \in H_0^1(\Omega),$$

with f in $L^2(\Omega)$. The Lax–Milgram theorem immediately gives the existence and uniqueness of $u_{\varepsilon\delta}$ in $H_0^1(\Omega)$ satisfying the estimate

$$\|u_{\varepsilon\delta}\|_{H_0^1(\Omega)} \leq \frac{1}{\alpha}\|f\|_{L^2(\Omega)}.$$

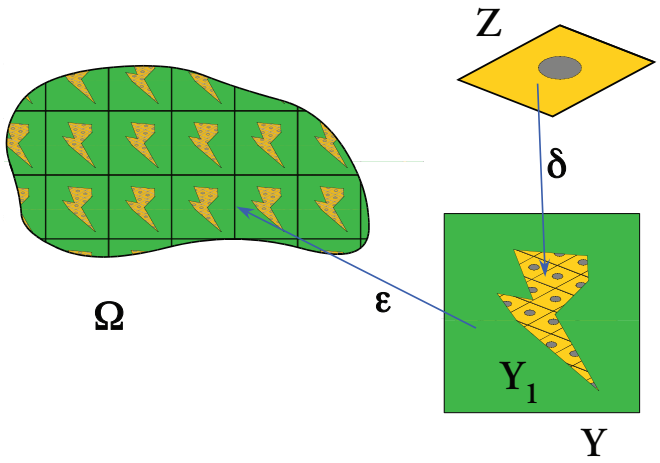


FIG. 7. *A domain with periodic scales ε and $\varepsilon\delta$.*

So, there is some u_0 such that, up to a subsequence,

$$u_{\varepsilon\delta} \rightharpoonup u_0 \quad \text{weakly in } H_0^1(\Omega).$$

Using the unfolding method for scale ε, as before we have

$$\begin{aligned}
&\mathcal{Q}_\varepsilon\big(u_{\varepsilon\delta}\big) \rightharpoonup u_0 \quad \text{weakly in } H_0^1(\Omega),\\
&\mathcal{T}_\varepsilon(u_{\varepsilon\delta}) \rightharpoonup u_0 \quad \text{weakly in } L^2(\Omega;\, H^1(Y)),\\
&\frac{1}{\varepsilon}\mathcal{T}_\varepsilon\big(\mathcal{R}_\varepsilon(u_{\varepsilon\delta})\big) \rightharpoonup \widehat{u} \quad \text{weakly in } L^2(\Omega;\, H^1(Y)),\\
&\mathcal{T}_\varepsilon\big(\nabla u_{\varepsilon\delta}\big) \rightharpoonup \nabla u_0 + \nabla_y \widehat{u} \quad \text{in } L^2(\Omega\times Y).
\end{aligned}$$

These convergences do not see the oscillations at the scale $\varepsilon\delta$. In order to capture them, one considers the restrictions to the set $\Omega\times Y_2$ defined by

$$v_{\varepsilon\delta}(x,y) \doteq \frac{1}{\varepsilon}\mathcal{T}_\varepsilon\big(\mathcal{R}_\varepsilon(u_{\varepsilon\delta})\big)|_{Y_2}.$$

Obviously,

$$v_{\varepsilon\delta} \rightharpoonup \widehat{u}|_{Y_2} \quad \text{weakly in } L^2(\Omega; H^1(Y_2)).$$

Now, we apply to $v_{\varepsilon\delta}$, a similar unfolding operation, denoted $\mathcal{T}_\delta^y$, for the variable y, thus adding a new variable $z\in Z$.

$$\mathcal{T}_\delta^y(v_{\varepsilon\delta})(x,y,z) = v_{\varepsilon\delta}\left(x, \delta\left[\frac{y}{\delta}\right]_Z + \delta z\right) \quad \text{for } x\in\Omega,\ \ y\in Y_2,\ \text{and } z\in Z.$$

It is essential to remark that all of the estimates and weak convergence properties which were shown for the original unfolding $\mathcal{T}_\varepsilon$ still hold for $\mathcal{T}_\delta^y$, with x being a mere parameter. For example, Proposition 4.7 and Theorem 3.5 adapted to this case imply that

$$\begin{aligned}
&\mathcal{T}_\delta^y\big(\nabla_y v_{\varepsilon\delta}\big) \rightharpoonup \nabla_y \widehat{u}|_{\Omega_2} + \nabla_z \widetilde{u} \quad \text{weakly in } L^2(\Omega\times Y_2\times Z),\\
&\mathcal{T}_\delta^y\Big(\mathcal{T}_\varepsilon\big(\nabla u_{\varepsilon\delta}\big)\Big) \rightharpoonup \nabla u_0 + \nabla_y \widehat{u} + \nabla_z \widetilde{u} \quad \text{weakly in } L^2(\Omega\times Y_2\times Z).
\end{aligned}$$

Under these conditions, the limit functions u_0, $\widehat{u}$, and $\widetilde{u}$ are characterized by the following result.

THEOREM 7.1. *The functions*

$$u_0 \in H^1_0(\Omega), \quad \widehat{u} \in L^2(\Omega, H^1_{per}(Y)/\mathbb{R}), \quad \widetilde{u} \in L^2(\Omega \times \Omega_2, H^1_{per}(Z)/\mathbb{R})$$

are the unique solutions of the variational problem

$$\begin{cases} \dfrac{1}{|Y||Z|}\displaystyle\int_\Omega\int_{Y_2}\int_Z A_2(z)\Big\{\nabla u_0 + \nabla_y\widehat{u} + \nabla_z\widetilde{u}\Big\}\Big\{\nabla\Psi + \nabla_y\Phi + \nabla_z\Theta\Big\}\,dx\,dy\,dz \\ \qquad + \dfrac{1}{|Y|}\displaystyle\int_\Omega\int_{Y_1} A_1(y)\Big\{\nabla u_0 + \nabla_y\widehat{u}\Big\}\Big\{\nabla\Psi + \nabla_y\Phi\Big\}\,dx\,dy = \displaystyle\int_\Omega f\,\Psi\,dx, \\ \forall\Psi \in H^1_0(\Omega), \quad \forall\Phi \in L^2(\Omega;\, H^1_{per}(Y)/\mathbb{R}), \forall\Theta \in L^2(\Omega \times \Omega_2, H^1_{per}(Z)/\mathbb{R}). \end{cases}$$

The proof uses test functions of the form

$$\Psi(x) + \varepsilon\Psi_1(x)\Phi_1\left(\frac{x}{\varepsilon}\right) + \varepsilon\delta\Psi_2(x)\Phi_2\left(\left\{\frac{x}{\varepsilon}\right\}_Y\right)\Theta_2\left(\frac{1}{\delta}\left\{\frac{x}{\varepsilon}\right\}_Y\right),$$

where Ψ, Ψ_1, Ψ_2 are in $\mathcal{D}(\Omega)$, Φ_1 in $H^1_{per}(Y)$, $\Phi_2 \in \mathcal{D}(Y_2)$, and $\Theta_2 \in H^1_{per}(Z)$.

Remark 7.2. The same theorem holds true for a general $A^{\varepsilon\delta}$ under the hypotheses

$$\mathcal{T}_\varepsilon(A^{\varepsilon\delta})\,1_{Y_1} \to A_1 \quad \text{strongly in } [L^1(\Omega \times Y_1)]^{n\times n},$$
$$\mathcal{T}^y_\delta\big(\mathcal{T}_\varepsilon(A^{\varepsilon\delta})\,1_{Y_2}\big) \to A_2 \quad \text{strongly in } [L^1(\Omega \times Y_2 \times Z)]^{n\times n}.$$

Proposition 5.9 (convergence of the energy) and Corollary 5.11 extend without any difficulty to the multiscale case.

PROPOSITION 7.3. *The convergence for the energy holds true:*

$$\begin{aligned} \lim_{\varepsilon,\delta\to 0}\int_\Omega A^{\varepsilon\delta}\nabla u_{\varepsilon\delta}\nabla u_{\varepsilon\delta}\,dx &= \frac{1}{|Y||Z|}\int_\Omega\int_{Y_2}\int_Z A_2(z)\Big\{\nabla u_0 + \nabla_y\widehat{u} + \nabla_z\widetilde{u}\Big\}\Big\{\nabla u_0 + \nabla_y\widehat{u} + \nabla_z\widetilde{u}\Big\}\,dx\,dy\,dz \\ &\quad + \frac{1}{|Y|}\int_\Omega\int_{Y_1} A_1(y)\Big\{\nabla u_0 + \nabla_y\widehat{u}\Big\}\Big\{\nabla u_0 + \nabla_y\widehat{u}\Big\}\,dx\,dy. \end{aligned}$$

COROLLARY 7.4. *The following strong convergences hold true:*

$$\mathcal{T}^y_\delta\big(\nabla_y v_{\varepsilon\delta}\big) \rightharpoonup \nabla_y\widehat{u}|_{\Omega_2} + \nabla_z\widetilde{u} \quad \textit{strongly in } L^2(\Omega \times Y_2 \times Z),$$
$$\mathcal{T}^y_\delta\Big(\mathcal{T}_\varepsilon\big(\nabla u_{\varepsilon\delta}\big)\Big) \rightharpoonup \nabla u_0 + \nabla_y\widehat{u} + \nabla_z\widetilde{u} \quad \textit{strongly in } L^2(\Omega \times Y_2 \times Z).$$

Remark 7.5. A corrector result, similar to that of Theorem 6.1, can be obtained.

Remark 7.6. Theorem 7.1 can be extended to the case of any finite number of distinct scales by a simple reiteration.

8. Further developments. The unfolding method has some interesting properties which make it suitable for more general situations than that presented here. In problems which are set on a domain Ω_ε which depends on the parameter ε, it may be difficult to have a good notion of convergence for the sequence of solutions u_ε. The traditional way is to extend the solution by 0 outside Ω_ε; however, this precludes the strong convergence of these extended functions in general. For the case of holes of the

size of order ε distributed ε-periodically, the unfolded sequence lives on a fixed domain. Similarly, for domains with ε-oscillating boundaries, a partial unfolding yields a function which is defined on a fixed domain.

We conclude by giving a list of publications making use of the unfolding method in several of these directions (both for linear and nonlinear problems).

– Reiterated homogenization: Meunier and Van Schaftingen [56].

– Electro-magnetism: Banks et al. [7], Bossavit, Griso, and Miara [15].

– Homogenization of thin piezoelectric shells: Ghergu et al. [39].

– Homogenization of diffusion deformation media: Griso and Rohan [46].

– Homogenization of the Stokes problem in porous media: Cioranescu, Damlamian, and Griso [25].

– Homogenization in perforated domains with Robin boundary conditions: Cioranescu, Donato and Zaki [31], [32].

– Homogenization in domains with oscillating boundaries: Damlamian and Pettersson [36].

– Homogenization of nonlinear integrals of the calculus of variations: Cioranescu, Damlamian, and De Arcangelis [27], [28], and [29].

– Homogenization of multivalued monotone operators of Leray–Lions type: Damlamian, Meunier, and Van Schaftingen [37].

– Thin junctions in linear elasticity: Blanchard, Gaudiello, and Griso [12], [13], Blanchard and Griso [14].

– Thin domains and free boundary problems arising in lubrication theory: Bayada, Martin, and Vazquez [9], [10].

– Elasticity problems in perforated domains: Griso and Sanchez-Rua [47].

– Neumann sieve and Dirichlet shield problems: Onofrei [60], Cioranescu et al. [26]. This last paper treats the case of domains with ε-periodically distributed "very small" holes (their size being a power of ε) on the boundary of which a homogeneous Dirichlet condition is prescribed. This requires the introduction of a rescaled unfolding operator (which originally appeared in the framework of the two-scale convergence in Casado-Díaz [20]).

Aknowledgments. We thank Petru Mironescu and Riccardo De Arcangelis for helpful comments and corrections.

REFERENCES

[1] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[2] G. Allaire and M. Briane, *Multiscale convergence and reiterated homogenization*, Proc. Roy. Soc. Edinburgh Sect. A, 126 (1996), pp. 297–342.

[3] G. Allaire and C. Conca, *Bloch wave homogenization and spectral asymptotic analysis*, J. Math. Pures Appl., 2 (1998), pp. 153–208.

[4] G. Allaire, C. Conca, and M. Vanninathan, *Spectral asymptotics of the Helmholtz model in fluid-solid structures*, Internat. J. Numer. Methods Engrg., 9 (1999), pp. 1463–1504.

[5] T. Arbogast, J. Douglas, and U. Hornung, *Derivation of the double porosity model of single phase flow via homogenization theory*, SIAM J. Math. Anal., 21 (1990), pp. 823–836.

[6] J. F. Babadjian and M. Baía, *Multiscale nonconvex relaxation and application to thin films*, Asymptot. Anal., 48 (2006), pp. 173–218.

[7] H. T. Banks, V. A. Bokil, D. Cioranescu, N. L. Gibson, G. Griso, and B. Miara, *Homogenization of periodically varying coefficients in electromagnetic materials*, J. Sci. Comput., 28 (2006), pp. 191–221.

[8] M. Barchiesi, *Multiscale homogenization of convex functionals with discontinuous integrand*, J. Convex Anal., 14 (2007), pp. 205–226.

[9] G. Bayada, S. Martin, and C. Vazquez, *Two-scale homogenization of a hydrodynamic Elrod-Adams model*, Asymptot. Anal., 44 (2005), pp. 75–110.
[10] G. Bayada, S. Martin, and C. Vazquez, *Homogenization of a nonlocal elastohydrodynamic lubrication problem: A new free boundary model*, Math. Models Methods Appl. Sci., 15 (2005), pp. 1923–1956.
[11] A. Bensoussan, J. L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, Stud. Math. Appl. 5, North-Holland, Amsterdam, 1978.
[12] D. Blanchard, A. Gaudiello, and G. Griso, *Junction of a periodic family of elastic rods with a* 3*d plate, Part* I, J. Math. Pures Appl., 88 (2007), pp. 1–33.
[13] D. Blanchard, A. Gaudiello, and G. Griso, *Junction of a periodic family of elastic rods with a thin plate, Part* II, J. Math. Pures Appl., 88 (2007), pp. 149–190.
[14] D. Blanchard and G. Griso, *Microscopic effects in the homogenization of the junction of rods and a thin plate*, Asymptot. Anal., 56 (2008), pp. 1–36.
[15] A. Bossavit, G. Griso, and B. Miara, *Modelling of periodic electromagnetic structures: Bianisotropic materials with memory effects*, J. Math. Pures Appl., 84 (2005), pp. 819–850.
[16] A. Bourgeat, S. Luckhaus, and A. Mikelic, *Convergence of the homogenization process for a double-porosity model of immiscible two-phase flow*, SIAM. J. Math. Anal., 27 (1996), pp. 1520–1543.
[17] A. Braides and A. Defranceschi, *Homogenization of Multiple Integrals*, Oxford Lect. Ser. Math. Appl. 12, Oxford University Press, Oxford, 1998.
[18] A. Braides and D. Lukkassen, *Reiterated homogenization of integral functionals*, Math. Models Methods Appl. Sci., 10 (2000), pp. 1–25.
[19] D. A. G. Bruggeman, *Berechnung verschiedener physikalisher konstanten von heterogenen substanzen*, Ann. Physik., 24 (1935), p. 634.
[20] J. Casado-Díaz, *Two-scale convergence for nonlinear Dirichlet problems in perforated domains*, Proc. Roy. Soc. Edinburgh Sect. A, 130 (2000), pp. 249–276.
[21] J. Casado-Díaz and M. Luna-Laynez, *A multiscale method to the homogenization of elastic thin reticulated structures*, in Homogenization 2001, GAKUTO Internat. Ser. Math. Sci. Appl. 18, Gakkōtosho, Tokyo, 2003, pp. 155–168.
[22] J. Casado-Díaz, M. Luna-Laynez, and J. D. Martín, *An adaptation of the multi-scale methods for the analysis of very thin reticulated structures*, C.R. Acad. Sci. Paris, Ser. 1, 332 (2001), pp. 223–228.
[23] J. Casado-Díaz, M. Luna-Laynez, and J. D. Martín, *A new approach to the analysis of thin reticulated structures*, in Homogenization 2001, GAKUTO Internat. Ser. Math. Sci. Appl. 18, Gakkōtosho, Tokyo, 2003, pp. 257–262.
[24] D. Cioranescu, A. Damlamian, and G. Griso, *Periodic unfolding and homogenization*, C.R. Acad. Sci. Paris, Ser. 1, 335 (2002), pp. 99–104.
[25] D. Cioranescu, A. Damlamian, and G. Griso, *The Stokes problem in perforated domains by the periodic unfolding method*, in Proceedings of the Conference on New Trends in Continuum Mechanics, M. Suliciu, ed., Theta, Bucarest, 2005, pp. 67–80.
[26] D. Cioranescu, A. Damlamian, G. Griso, and D. Onofrei, *The periodic unfolding method for perforated domains and Neumann sieve models*, J. Math. Pures Appl., 89 (2008), pp. 248–277.
[27] D. Cioranescu, A. Damlamian, and R. De Arcangelis, *Homogenization of nonlinear integrals via the periodic unfolding method*, C.R. Acad. Sci. Paris, Ser. 1, 339 (2005), pp. 77–82.
[28] D. Cioranescu, A. Damlamian, and R. De Arcangelis, *Homogenization of integrals with pointwise gradient constraints via the periodic unfolding method*, Ricerche Mat., 55 (2006), pp. 31–54.
[29] D. Cioranescu, A. Damlamian, and R. De Arcangelis, *Homogenization of quasiconvex integrals via the periodic unfolding method*, SIAM J. Math. Anal., 37 (2006), pp. 1435–1453.
[30] D. Cioranescu and P. Donato, *An Introduction to Homogenization*, Oxford Lecture Ser. in Math. Appl. 17, Oxford University Press, Oxford, 1999.
[31] D. Cioranescu, P. Donato, and R. Zaki, *The periodic unfolding method in perforated domains*, Port. Math., 63 (2006), pp. 467–496.
[32] D. Cioranescu, P. Donato, and R. Zaki, *Asymptotic behavior of elliptic problems in perforated domains with nonlinear boundary conditions*, Asymptot. Anal., 53 (2007), pp. 209–235.
[33] G. Dal Maso and A. Defranceschi, *Correctors for the homogenization of monotone operators*, Differential Integral Equations, 3 (1990), pp. 1151–1166.

[34] A. Damlamian, *An elementary introduction to periodic unfolding*, in Proceedings of the Narvik Conference 2004, GAKUTO Internat. Ser. Math. Sci. Appl. 24, A. Damlamian, D. Lukkassen, A. Meidell, and A. Piatnitski, eds., Gakkōtosho, Tokyo, 2006, pp. 119–136.
[35] A. Damlamian and P. Donato, *H^0-convergence and iterated homogenization*, Asymptot. Anal., 39 (2004), pp. 45–60.
[36] A. Damlamian and K. Pettersson, *Homogenization of oscillating boundaries*, Discrete Contin. Dyn. Syst., 23 (2009), pp. 197–219.
[37] A. Damlamian, N. Meunier, and J. Van Schaftingen, *Periodic homogenization of monotone multivalued operators*, Nonlinear Anal., 67 (2007), pp. 3217–3239.
[38] A. Ene and J. Saint Jean Paulin, *On a model of fractured porous media*, in Mathematical Modelling of Flow Through Porous Media, A. Bourgeat, C. Carasso, S. Luckhaus, and A. Mikelic, eds., World Scientific, River Edge, NJ, 1995, pp. 402–409.
[39] M. Ghergu, G. Griso, H. Mechkour, and B. Miara, *Homogenization of thin piezoelectric shells*, ESAIM: M2AN Math. Modeling Numer. Anal., 41 (2007), pp. 875–895.
[40] G. Griso, *Analyse asymptotique de structures réticulées*, Thèse Université Pierre et Marie Curie (Paris VI), Paris, 1996.
[41] G. Griso, *Thin reticulated structures*, in Progress in Partial Differential Equations, The Metz Surveys 3, M. Chipot, J. Saint Jean Paulin, and I. Shafrir, eds., Pitman, London, 1994, pp. 161–182.
[42] G. Griso, *Error estimate and unfolding for periodic homogenization*, Asymptot. Anal., 40 (2004), pp. 269–286.
[43] G. Griso, *Les méthodes d'éclatement en homogénéisation périodique et en élasticité linéarisée*, Thèse d'Habilitation, Université Pierre et Marie Curie, Paris, 2005.
[44] G. Griso, *Interior error estimates for periodic homogenization*, C.R. Acad. Sci. Paris, Ser. 1, 340 (2005), pp. 251–254.
[45] G. Griso, *Interior error estimates for periodic homogenization*, Anal. Appl., 4 (2006), pp. 61–79.
[46] G. Griso and E. Rohan, *On homogenization of diffusion-deformation problem in strongly heterogeneous media*, Ricerche Mat., 56 (2007), pp. 161–188.
[47] G. Griso and T. Sanchez-Rua, *Homogenization of an elasticity problem for a catalyst support by using the unfolding method*, submitted.
[48] M. Lenczner, *Homogénéisation d'un circuit électrique*, C.R. Acad. Sci. Paris, Ser. 2, 324 (1997), pp. 537–542.
[49] M. Lenczner and D. Mercier, *Homogenization of periodic electrical networks including voltage to current amplifiers*, SIAM Multiscale Model. Simul., 2 (2004), pp. 359–397.
[50] M. Lenczner and G. Senouci-Bereksi, *Homogenization of electrical networks including voltage-to-voltage amplifiers*, Math. Models Methods Appl. Sci., 9 (1999), pp. 899–932.
[51] M. Lenczner, M. Kader, and P. Perrier, *Modèle à deux échelles de l'équation des ondes à coefficients oscillants*, C.R. Acad. Sci. Paris, Ser. 1, 328 (2000), pp. 335–340.
[52] J. L. Lions, D. Lukkassen, L. E. Persson, and P. Wall, *Reiterated homogenization of monotone operators*, Chinese Ann. Math. Ser. B, 22 (2001), pp. 1–12.
[53] D. Lukkassen, *Reiterated homogenization of non-standard Lagrangians*, C.R. Acad. Sci. Paris, Ser. 1, 332 (2001), pp. 999–1004.
[54] D. Lukkassen and G. W. Milton, *On hierarchical structures and reiterated homogenization*, in Proceedings of the Conference on Function Spaces, Interpolation Theory and Related Topics in Honor of Jack Peetre on his 65th Birthday, M. Cwikwl, M. Engliš, A. Kufner, L.-E. Persson, and G. Sparr, eds, Walter de Gruyter, Berlin, 2002, pp. 311–324.
[55] D. Lukkassen, G. Nguetseng, and P. Wall, *Two-scale convergence*, Int. J. Pure Appl. Math., 2 (2002), pp. 35–86.
[56] N. Meunier and J. Van Schaftingen, *Periodic reiterated homogenization for elliptic functions*, J. Math. Pures Appl., 9 (2005), pp. 1716–1743.
[57] G. Milton, *The Theory of Composites*, Cambridge University Press, Cambridge, 2002.
[58] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–629.
[59] O. A. Oleinik, A. S. Shamaev, and G. A. Yosifian, *Mathematical Problems in Elasticity and Homogenization*, North-Holland, Amsterdam, 1992.
[60] D. Onofrei, *The unfolding operator near a hyperplane and its applications to the Neumann sieve model*, Adv. Math. Sci. Appl., 16 (2006), pp. 239–258.

SIAM J. MATH. ANAL.
Vol. 40, No. 4, pp. 1621–1670

ON THE ASYMPTOTIC STABILITY OF SMALL NONLINEAR DIRAC STANDING WAVES IN A RESONANT CASE*

NABILE BOUSSAID†

Abstract. We study the behavior of perturbations of small nonlinear Dirac standing waves. We assume that the linear Dirac operator of reference $H = D_m + V$ has only two double eigenvalues and that degeneracies are due to a symmetry of H (theorem of Kramers). In this case, we can build a small four-dimensional manifold of stationary solutions tangent to the first eigenspace of H. Then we assume that a resonance condition holds, and we build a center manifold of real codimension 8 around each stationary solution. Inside this center manifold any H^s perturbation of stationary solutions, with $s > 2$, stabilizes towards a standing wave. We also build center-stable and center-unstable manifolds, each one of real codimension 4. Inside each of these manifolds, we obtain stabilization towards the center manifold in one direction of time, while in the other, we have instability. Eventually, outside all of these manifolds, we have instability in the two directions of time. For localized perturbations inside the center manifold, we obtain a nonlinear scattering result.

Key words. Dirac equation, nonlinear PDE, asymptotic stability, Strichartz estimates, smoothness estimates, center manifold

AMS subject classifications. 35Q40, 35Q55, 35Q75

DOI. 10.1137/070684641

Introduction. We study the asymptotic stability of stationary solutions of a time-dependent nonlinear Dirac equation.

A localized stationary solution of a given time-dependent equation represents a bound state of a particle. Like Rañada [39], we call it a *particle-like solution* (PLS). Many works have been devoted to the proof of the existence of such solutions for a wide variety of equations. Although their stability is a crucial problem (in particular, in a numerical computation or experiment), less attention has been devoted to this issue.

In this paper, we deal with the problem of stability of small PLSs of the following nonlinear Dirac equation:

$$\mathrm{i}\partial_t\psi = (D_m + V)\psi + \nabla F(\psi), \tag{0.1}$$

where ∇F is the gradient of $F : \mathbb{C}^4 \mapsto \mathbb{R}$ for the standard scalar product of $\mathbb{R}^8$. Here, D_m is the usual Dirac operator (see Thaller [48]) acting on $L^2(\mathbb{R}^3, \mathbb{C}^4)$:

$$D_m = \alpha \cdot (-\mathrm{i}\nabla) + m\beta = -\mathrm{i}\sum_{k=1}^{3} \alpha_k \partial_k + m\beta,$$

where $m \in \mathbb{R}_+^*$, $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, and β are $\mathbb{C}^4$ Hermitian matrices satisfying

$$\begin{cases} \alpha_i\alpha_k + \alpha_k\alpha_i = 2\delta_{ik}\mathbf{1}_{\mathbb{C}^4}, & i,k \in \{1,2,3\}, \\ \alpha_i\beta + \beta\alpha_i = \mathbf{0}_{\mathbb{C}^4}, & i \in \{1,2,3\}, \\ \beta^2 = \mathbf{1}_{\mathbb{C}^4}. \end{cases}$$

*Received by the editors March 8, 2007; accepted for publication (in revised form) June 4, 2008; published electronically November 26, 2008. This work was partially supported by EPSRC grant EP/D054621.

http://www.siam.org/journals/sima/40-4/68464.html

†Laboratoire de Mathématiques de Besançon, Université de Franche-Comté, 16 Route de Gray, F-25030 Besançon Cedex, France (nabile.boussaid@univ-fcomte.fr).

Here, we choose

$$\alpha_i = \begin{pmatrix} 0 & \sigma_i \\ \sigma_i & 0 \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} I_{\mathbb{C}^2} & 0 \\ 0 & -I_{\mathbb{C}^2} \end{pmatrix},$$

$$\text{where } \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}, \quad \text{and} \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

In (0.1), V is the external potential field, and $F : \mathbb{C}^4 \mapsto \mathbb{R}$ is a nonlinearity with the following gauge invariance:

$$\forall (\theta, z) \in \mathbb{R} \times \mathbb{C}^4, \quad F(e^{i\theta} z) = F(z). \tag{0.2}$$

Some additional assumptions on F and V will be made in what follows. Nonlinearity with no potential arises in some Dirac models introduced by physicists to model either extended particles with self-interaction or particles in space-time with geometrical structure. In the latter case, physicists have shown that a relativistic theory sometimes imposes a fourth order nonlinear potential (i.e., a cubic nonlinearity) such as the square of a quadratic form on $\mathbb{C}^4$; see Rañada [39] and the references therein. We added a potential with special features to ensure the existence of small stationary solutions, whose existence and stability are easier to study.

Stationary solutions (PLSs) of (0.1) take the form $\psi(t, x) = e^{-\mathrm{i}Et}\phi(x)$, where ϕ satisfies

$$E\phi = (D_m + V)\phi + \nabla F(\phi). \tag{0.3}$$

We show that there exists a manifold of small solutions to (0.3) tangent to the first eigenspace of $D_m + V$ (see Proposition 1.1 below).

Concerning the asymptotic stability in the Schrödinger equation, the question has been solved in several cases. For small stationary solutions in the simple eigenvalue case it has been studied by Soffer and Weinstein [43, 44], Pillet and Wayne [38], and Gustafson, Nakanishi, and Tsai [22]. For the two eigenvalue case under a resonance condition for an excited state, the problem has been studied by Tsai and Yau [51, 53, 54, 52, 50] and Soffer and Weinstein [45, 46]. Another problem has been studied by Cuccagna [16, 17, 18], who considered the case of a big PLS, when the linearized operator has only one eigenvalue, and obtained the asymptotic stability of the manifold of ground states. Schlag [41] proved that any ground state of the cubic nonlinear Schrödinger equation in dimension 3 is orbitally unstable but possesses a stable manifold of codimension 9.

We also would like to mention the works of Buslaev and Perel'mann [12, 10, 11, 9], Buslaev and Sulem [14, 13], Weder [55], and Krieger and Schlag [31] in the one-dimensional Schrödinger case. Krieger and Schlag [31] proved a result similar to [41] in the one-dimensional case. Concerning two-dimensional nonlinear Schrödinger equations, Mizumachi [35, 36] studied the properties of unstable solutions, while Kirr and Zarnescu [30] proved the existence of small asymptotically stable states in a semilinear case.

In [7], we prove that there are stable directions for the PLS manifold under a nonresonance assumption on the spectrum of $H := D_m + V$. This gives a stable manifold containing the PLS manifold. But we were not able to say anything about solutions starting outside the stable manifold.

The results we present here state the existence of a stable manifold and describe the behavior of solutions starting outside of it. In fact, we prove the instability of the perturbations starting outside the stable manifold. We also prove stabilization towards stationary solutions inside the stable manifold for H^s perturbation with $s > 2$. We have been able to obtain it since we impose a resonance condition (see Assumption 1.5 below), while in [7], we assumed that there is no resonance phenomena.

When the perturbations are localized, we are able to push the analysis further and obtain a nonlinear scattering.

This paper is organized as follows.

In section 1, we present our main results and the assumptions we need. Subsection 1.1 is devoted to the statement of the time decay estimates of the propagator associated with $H = D_m + V$ on the continuous subspace. One estimate is a kind of smoothness result, in the sense of Kato (see, e.g., [27]), and the other is a Strichartz-type result. We prove these estimates with the propagation and dispersive estimates proved in [7]. In subsection 1.2, we state the existence of small stationary states forming a manifold tangent to an eigenspace of H. The study of the dynamics around such states leads us to our main results; see subsection 1.3 and 1.4. In subsection 1.3, we split a neighborhood of a stationary state into different parts, each giving rise to stabilization or instability. In subsection 1.4, we state our scattering result.

To prove our theorems, we consider our nonlinear system as a small perturbation of a linear equation. More precisely, in subsection 2.2, we show that the spectral properties of the linearized operator around a stationary state, presented in section 2, permit us to obtain, as in the linear case, some properties of the dynamics around a stationary state. We obtain center, center-stable, and center-unstable manifolds. In section 3, we obtain, with our time decay estimates, a stabilization towards the PLS manifold for H^s perturbation with $s > 2$ in the center manifold. Section 4 deals with the dynamics outside the center manifold. Eventually in section 5, we conclude our study.

Our results are the analogue, in the Dirac case, of some results of Tsai and Yau [54], Soffer and Weinstein [43], Pillet and Wayne [38], and Gustafson, Nakanishi, and Tsai [22] about the semilinear Schrödinger equation. The point which we did not investigate is the long time behavior of perturbation in the unstable direction. In [51, 53, 54, 52, 50, 45, 46], the authors show that such perturbations relax towards the ground states. Their analyses are based on the asymptotic stability of ground states under a resonance condition. In our case this question is still open and should be investigated. But we believe that the techniques used in the previously cited works will allow us to obtain a similar result. Indeed the phenomenon that leads to this relaxation, namely nonlinear resonance, is not specific to the Shrödinger operator or nonnegative operators. Such a phenomenon comes from the interaction between discrete and continuous modes allowed by the nonlinearity. So a Fermi golden rule assumption should have the same consequences in both Shrödinger and Dirac equations.

We should also mention that four manifolds appear in this paper: the PLS manifold, the center manifold, the center-stable manifold, and the center-unstable manifold. The PLS manifold used to be called center manifold by previous papers on the same subject. The terminology we choose to adopt here comes from dynamical system theory and especially from the center manifold theorem. All these manifolds are invariant manifolds of our equation, at least locally in time. The PLS manifold is a set of stationary solutions, while the center manifold is a set of solutions converging

to the PLS manifold and containing the PLS manifold. The center manifold in turn is the intersection of the center-stable and center-unstable manifolds. The roles of the center-stable and center-unstable manifolds are exchanged when time is reversed.

This paper is devoted to the study of the asymptotic stability of some Dirac equations. So we are investigating convergence of orbits. As far as we know, the question of the orbital stability which requires the orbits to stay close and not necessarily to converge is still open for Dirac equations. In the Schrödinger case, orbital stability results (see, e.g., [15], [56, 57], or [42, 21]) give that any solution stays near the PLS manifold. Unfortunately, orbital stability criteria applied to Schrödinger equations use the fact that Schrödinger operators are bounded from below. Hence the question of orbital stability for Dirac standing waves cannot be solved by a straightforward application of the methods used in the Schrödinger case.

1. Assumptions and statements.

1.1. Time decay estimates. We generalize, to small nonlinear perturbations, stability results for linear systems. These results, as in [7], follow from linear decay estimates. Here we use smoothness-type and Strichartz-type estimates deduced from propagation and dispersive estimates of [7]. Hence, we work within the same assumptions for V and $D_m + V$.

Assumption 1.1. The potential $V : \mathbb{R}^3 \mapsto S_4(\mathbb{C})$ *(self-adjoint* 4×4 *matrices)* is a smooth function such that there exists $\rho > 5$ with

$$\forall \alpha \in \mathbb{N}^3,\ \exists C > 0, \qquad \forall x \in \mathbb{R}^3,\ |\partial^\alpha V|(x) \leq \frac{C}{\langle x\rangle^{\rho+|\alpha|}}.$$

We notice that by the Kato–Rellich theorem, the operator

$$H := D_m + V$$

is essentially self-adjoint on $\mathcal{C}_0^\infty(\mathbb{R}^3, \mathbb{C}^4)$ and self-adjoint on $H^1(\mathbb{R}^3, \mathbb{C}^4)$.

We also mention that Weyl's theorem gives that the essential spectrum of H is $(-\infty, -m] \cup [+m, +\infty)$, and the work of Berthier and Georgescu [5, Theorems 6 and A] gives that there is no embedded eigenvalue. Hence the thresholds $\pm m$ are the only points of the continuous spectrum which can be associated with a wave of zero velocity. These waves perturb the spectral density and diminish the decay rate in the propagation and the dispersive estimates. We will work (as in [7]) within the following assumption.

Assumption 1.2. The operator H presents no resonance at thresholds and no eigenvalue at thresholds.

A resonance is a stationary solution in $H^{1/2}_{-\sigma} \setminus H^{1/2}$ for any $\sigma \in (1/2, \rho - 2)$, where H^t_σ is given by the following definition.

Definition 1.1 (weighted Sobolev space). *The weighted Sobolev space is defined by*

$$H^t_\sigma(\mathbb{R}^3, \mathbb{C}^4) = \left\{ f \in \mathcal{S}'(\mathbb{R}^3),\ \|\langle Q\rangle^\sigma \langle P\rangle^t f\|_2 < \infty \right\}$$

for $\sigma, t \in \mathbb{R}$. *We endow it with the norm*

$$\|f\|_{H^t_\sigma} = \|\langle Q\rangle^\sigma \langle P\rangle^t f\|_2.$$

If $t = 0$, *we write* L^2_σ *instead of* H^0_σ.

We have used the usual notation: $\langle u\rangle = \sqrt{1+u^2}$, $P = -i\nabla$, and Q is the operator of multiplication by x in $\mathbb{R}^3$.

Remark 1.1. Assumption 1.2 is generic in the following sense.

If V does not fulfill Assumption 1.2, then $(1 \pm \varepsilon)V$ does for nonzero small ε. As in [26, section 3] one can prove that resonances and eigenvalues at theresholds at $\pm m$ are in the kernel of $1 + (D_m \mp m)^{-1}V$ in $H^{1/2}_\sigma$ for any $\sigma \in (1/2, \rho - 2)$. Since $(D_m \mp m)^{-1}V$ is a compact operator in $H^{1/2}_\sigma$ (see proof of Lemma 1.1 below) for any $\sigma \in (1/2, \rho - 2)$, its spectrum is discrete, and hence if -1 is in its spectrum, then -1 is not in the spectrum of $\lambda(D_m \mp m)^{-1}V$ for $\lambda \neq 1$ but close to 1. This remark can be extended to any analytic operator valued functions $\lambda \to V(\lambda)$ satisfying Assumption 1.1.

Under the previous assumption one can prove the following lemma.

LEMMA 1.1. *The discrete spectrum of H is finite.*

Proof. To prove this lemma we show that there is no eigenvalue in a neighborhood of the thresholds $\pm m$. Since there is no other possible accumulation point for the discrete spectrum, this will be sufficient.

We will do the proof for m, since it is similar for $-m$. To do so, we show that $(H - z)^{-1}$ exists from $H^{-1/2}_\sigma$ to $H^{1/2}_{-\sigma}$ for any $\sigma \in (1/2, \rho - 2)$ and z close to m and $\Re z < m$. Since isolated eigenvectors belong to $H^{-1/2}_\sigma$ (see [23]), this will give the desired result.

We use the formula

$$H - z = (D_m - z)\left(1 + (D_m - z)^{-1}V\right)$$

for z close to m and $\Re z < m$. From [7, Proposition 2.4] (or [26, Lemma 2.1] and [48, section 1.E]), $z \mapsto (D_m - z)^{-1}$ is a continuous map from $\{z \in \mathbb{C}, \Im z > 0, \Re z > 0\}$ to $\mathcal{B}(H^{-1/2}_\sigma, H^{1/2}_{-\sigma})$ for any $\sigma \in (1/2, \rho - 2)$. The point is to prove the invertibility of $\left(1 + (D_m - z)^{-1}V\right)$ in $H^{-1/2}_\sigma$. Using continuity and Assumption 1.1, we see that it is enough to prove that $\left(1 + (D_m - m)^{-1}V\right)$ is invertible in $H^{-1/2}_\sigma$. This follows from Assumption 1.2 (see [7, Lemma 2.2 and Proof of Proposition 2.3]). ∎

Now let

$$\mathbf{P}_c(H) = \mathbf{1}_{(-\infty, -m|\cup[+m, +\infty)}(H)$$

be the projector associated with the continuous spectrum of H and let $\mathcal{H}_c$ be its range. Using [7, Theorem 1.1], we obtain a limiting absorption principle which gives the H-smoothness of $\langle Q\rangle^{-1}$ in the sense of Kato.

THEOREM 1.1 (Kato smoothness estimates). *If Assumptions* 1.1 *and* 1.2 *hold, then for any $\sigma \geq 1$ and $s \in \mathbb{R}$ one has*

$$\left\| \langle Q\rangle^{-\sigma} e^{-\mathrm{i}tH} \mathbf{P}_c(H)\, \psi \right\|_{L^2_t(\mathbb{R}, H^s)} \leq C \|\psi\|_{H^s}, \tag{1.1}$$

$$\left\| \int_{\mathbb{R}} e^{\mathrm{i}tH} \mathbf{P}_c(H) \langle Q\rangle^{-\sigma} F(t)\, dt \right\|_{H^s} \leq C \|F\|_{L^2_t(\mathbb{R}, H^s)}, \tag{1.2}$$

$$\left\| \int_{s<t} \langle Q\rangle^{-\sigma} e^{-\mathrm{i}(t-s)H} \mathbf{P}_c(H) \langle Q\rangle^{-\sigma} F(s)\, ds \right\|_{L^2_t(\mathbb{R}, H^s)} \leq C \|F\|_{L^2_t(\mathbb{R}, H^s)}. \tag{1.3}$$

Proof. We first prove (1.1). For $s = 0$, it is (see, e.g., [1, Proposition 7.11] or [40, Theorem XIII.25]) a consequence of the limiting absorption principle:

$$\sup_{\Im z\in(0,1)} \left\{ \left\| \langle Q\rangle^{-\sigma} (H-z)^{-1} P_c(H) \langle Q\rangle^{-\sigma} \right\|_2 \right\} < \infty, \tag{1.4}$$

which follows from [7, Theorem 1.1] (or Theorem 3.1 below) for $\sigma > 5/2$ using the fact that the Fourier transform in time of the propagator is the resolvent. Actually, the Fourier transform of

$$\langle Q\rangle^{-\sigma} e^{-\mathrm{i}t(H-\mathrm{i}\varepsilon)} \mathbf{P}_c(H) \mathbf{1}_{\mathbb{R}_+^*}(t) \langle Q\rangle^{-\sigma} f$$

in time is

$$\langle Q\rangle^{-\sigma} (H-\lambda-\mathrm{i}\varepsilon)^{-1} \mathbf{P}_c(H) \langle Q\rangle^{-\sigma} f$$

for $f \in L^2(\mathbb{R}^3, \mathbb{C}^4)$. Then we use the Born expansion

$$(H-z)^{-1} = (D_m-z)^{-1} - (D_m-z)^{-1}V(D_m-z)^{-1} + (D_m-z)^{-1}V(H-z)^{-1}V(D_m-z)^{-1},$$

the limiting absorption in [25, Theorem 2.1(i)] (the authors prove the identity (1.4) for $H = D_m$ when $\sigma = 1$), [7, Proposition 2.3], and the fact that $\|(1-P_c(H))\|_{\mathcal{B}(H^{1/2}_\sigma)}$ is bounded (since the discrete spectrum is finite and eigenvectors are exponentially decaying; see [23]) to obtain (1.4) for $\sigma = 1$. Hence we have concluded the proof for $s = 0$ and $\sigma \geq 1$. For $s \in 2\mathbb{Z}$ and $\sigma \geq 1$ it follows from the previous cases using the boundedness of $\langle H\rangle^s \langle D_m\rangle^{-s}$ and $\langle H\rangle^{-s}\langle D_m\rangle^s$ (which follow from the boundedness of V and its derivatives) and the boundedness of $\langle Q\rangle^{\mp\sigma}[\langle Q\rangle^{\pm\sigma}, H\rangle^s]\langle H\rangle^{-s}$ (which follow from multicommutator estimates; see [24, Appendix B]). The rest of the claim (1.1) follows by interpolation.

Estimates (1.1) and (1.2) are equivalent by duality.

To prove estimate (1.3) when $s = 0$ (the general case will follow in the same way as above), we notice that we have to prove that there exists $C > 0$ such that, for all $F, G \in L^2_t(\mathbb{R}, L^2(\mathbb{R}^3, \mathbb{C}^4))$, we have

$$\left| \iint_{\mathbb{R}^2} \left\langle G(t), \langle Q\rangle^{-\sigma} e^{-\mathrm{i}(t-s)H} \mathbf{P}_c(H) \mathbf{1}_{\mathbb{R}_+^*}(t-s) \langle Q\rangle^{-\sigma} F(s) \right\rangle \, dsdt \right| \leq C \|G\|_{L^2_t(\mathbb{R}, L^2(\mathbb{R}^3,\mathbb{C}^4))} \|F\|_{L^2_t(\mathbb{R}, L^2(\mathbb{R}^3,\mathbb{C}^4))}.$$

We can suppose that F and G are smooth functions with compact support from $\mathbb{R}\times\mathbb{R}^3$ to $\mathbb{C}^4$ and we just need to prove that there exists $C > 0$ such that, for all $\varepsilon > 0$ and for all $F, G \in \mathcal{C}^\infty_0(\mathbb{R}, L^2(\mathbb{R}^3, \mathbb{C}^4))$, we have

$$\left| \iint_{\mathbb{R}^2} \left\langle G(t), \langle Q\rangle^{-\sigma} e^{-\mathrm{i}(t-s)(H-\mathrm{i}\varepsilon)} \mathbf{P}_c(H) \mathbf{1}_{\mathbb{R}_+^*}(t-s) \langle Q\rangle^{-\sigma} F(s) \right\rangle \, dsdt \right| \leq C \|G\|_{L^2_t(\mathbb{R}, L^2(\mathbb{R}^3,\mathbb{C}^4))} \|F\|_{L^2_t(\mathbb{R}, L^2(\mathbb{R}^3,\mathbb{C}^4))}.$$

Then we take the limit as $\varepsilon \to 0$ and conclude using density arguments. Let us write $A_\varepsilon(t)$ for $\langle Q\rangle^{-\sigma} e^{-\mathrm{i}t(H-\mathrm{i}\varepsilon)} \mathbf{P}_c(H) \mathbf{1}_{\mathbb{R}_+^*}(t) \langle Q\rangle^{-\sigma}$; then we have to prove

$$\left| \int_{\mathbb{R}} \langle G(t), (A_\varepsilon * F)(t)\rangle \, dt \right| \leq C \|G\|_{L^2_t(\mathbb{R}, L^2(\mathbb{R}^3,\mathbb{C}^4))} \|F\|_{L^2_t(\mathbb{R}, L^2(\mathbb{R}^3,\mathbb{C}^4))}.$$

Using Plancherel's identity in $L^2_t(\mathbb{R}, L^2(\mathbb{R}^3, \mathbb{C}^4))$ and $A_\varepsilon * F \in L^2_t(\mathbb{R}, L^2(\mathbb{R}^3, \mathbb{C}^4))$, we just need to prove

$$\left| \int_{\mathbb{R}} \left\langle \widehat{G}(\lambda), \widehat{A_\varepsilon * F}(\lambda) \right\rangle \, d\lambda \right| \leq C \left\| \widehat{G} \right\|_{L^2_\lambda(\mathbb{R}, L^2(\mathbb{R}^3, \mathbb{C}^4))} \left\| \widehat{F} \right\|_{L^2_\lambda(\mathbb{R}, L^2(\mathbb{R}^3, \mathbb{C}^4))}.$$

Since the Fourier transform in time of the propagator is the resolvent, F is smooth with compact support, and $\varepsilon > 0$, we obtain

$$\widehat{A_\varepsilon * F}(\lambda) = \langle Q \rangle^{-\sigma} (H - \lambda - \mathrm{i}\varepsilon)^{-1} \mathbf{P}_c(H) \langle Q \rangle^{-\sigma} \widehat{F}(\lambda).$$

Hence we just have to prove

$$\left| \int_{\mathbb{R}} \left\langle \widehat{G}(\lambda), \langle Q \rangle^{-\sigma} (H - \lambda - \mathrm{i}\varepsilon)^{-1} \mathbf{P}_c(H) \langle Q \rangle^{-\sigma} \widehat{F}(\lambda) \right\rangle \, d\lambda \right|$$
$$\leq C \left\| \widehat{G} \right\|_{L^2_\lambda(\mathbb{R}, L^2(\mathbb{R}^3, \mathbb{C}^4))} \left\| \widehat{F} \right\|_{L^2_\lambda(\mathbb{R}, L^2(\mathbb{R}^3, \mathbb{C}^4))}.$$

This in turn follows from the limiting absorption principle (1.4) just proved. ∎

To state the next result, we need the following definition.

DEFINITION 1.2 (Besov space). *For $s \in \mathbb{R}$ and $1 \leq p, q \leq \infty$, the Besov space $B^s_{p,q}(\mathbb{R}^3, \mathbb{C}^4)$ is the space of all $f \in \mathcal{S}'(\mathbb{R}^3, \mathbb{C}^4)$ (dual of the Schwartz space) such that*

$$\|f\|_{B^s_{p,q}} = \left(\sum_{j \in \mathbb{N}} 2^{jsq} \|\varphi_j * f\|_p^q \right)^{\frac{1}{q}} < +\infty$$

with $\widehat{\varphi} \in \mathcal{C}^\infty_0(\mathbb{R}^n \setminus \{0\})$ such that $\sum_{j \in \mathbb{Z}} \widehat{\varphi}(2^{-j}\xi) = 1$ for all $\xi \in \mathbb{R}^3 \setminus \{0\}$, $\widehat{\varphi_j}(\xi) = \widehat{\varphi}(2^{-j}\xi)$ for all $j \in \mathbb{N}^*$ and for all $\xi \in \mathbb{R}^3$, and $\widehat{\varphi_0} = 1 - \sum_{j \in \mathbb{N}^*} \widehat{\varphi_j}$. It is endowed with the natural norm $f \in B^s_{p,q}(\mathbb{R}^3, \mathbb{C}^4) \mapsto \|f\|_{B^s_{p,q}}$.

Using the dispersive estimates of [7, Theorem 1.2] and [29, Theorem 10.1], we obtain the following theorem.

THEOREM 1.2 (Strichartz estimates). *If Assumptions 1.1 and 1.2 hold, then for any $2 \leq p, q \leq \infty$, $\theta \in [0, 1]$, with $(1 - \frac{2}{q})(1 \pm \frac{\theta}{2}) = \frac{2}{p}$ and $(p, \theta) \neq (2, 0)$, and for any reals s, s' with $s' - s \geq \alpha(q)$, where $\alpha(q) = (1 + \frac{\theta}{2})(1 - \frac{2}{q})$, there exists a positive constant C such that*

$$\left\| e^{-\mathrm{i}tH} P_c(H) \psi \right\|_{L^p_t(\mathbb{R}, B^s_{q,2}(\mathbb{R}^3, \mathbb{C}^4))} \leq C \|\psi\|_{H^{s'}(\mathbb{R}^3, \mathbb{C}^4)}, \tag{1.5}$$

$$\left\| \int e^{\mathrm{i}tH} P_c(H) F(t) \, dt \right\|_{H^s} \leq C \|F\|_{L^{p'}_t(\mathbb{R}, B^{s'}_{q',2}(\mathbb{R}^3, \mathbb{C}^4))}, \tag{1.6}$$

$$\left\| \int_{s<t} e^{-\mathrm{i}(t-s)H} P_c(H) F(s) \, ds \right\|_{L^p_t(\mathbb{R}, B^{-s}_{q,2}(\mathbb{R}^3, \mathbb{C}^4))} \leq C \|F\|_{L^{\tilde{p}'}_t(\mathbb{R}, B^{\tilde{s}}_{\tilde{q}',2}(\mathbb{R}^3, \mathbb{C}^4))}, \tag{1.7}$$

for any $r \in [1, \infty]$, $(\tilde{q}, \tilde{p})$ chosen like (q, p), and $s + \tilde{s} \geq \alpha(q) + \alpha(\tilde{q})$.

Proof. This is a consequence of [29, Theorem 10.1] applied to $U(t) = e^{-\mathrm{i}tH} P_c(H)$, using [7, Theorem 1.2] (or Theorem 3.2 below) and

$$B^{(1+\frac{\theta}{2})(1-\frac{2}{q})+s}_{q,2} \hookrightarrow (H^s, B^{1+\theta/2+s}_{1,2})_{2/((1\pm\theta/2)p),2}$$

continuously for $p \geq 2$ ($p \neq 2$ if $\theta = 0$) and $1/q = 1 - 1/((1 \pm \theta/2)p)$. For these embeddings, we refer to the proof of [4, Theorem 6.4.5] as well as the properties of the real interpolation (see [4] or [49]). More precisely, for $\theta = 0$ or 1 it is obvious. In the other cases, we work as in the proof of [4, Theorem 6.4.5(3)].

We use [4, Theorem 6.4.3] ($B^s_{p,2}$ is a retract of $l^s_2(L^p)$ for $s \in \mathbb{R}$ and $p, q \in [1, \infty]$) and [4, Theorem 5.6.2] (about the interpolation of $l^s_2(L^p)$ spaces) with [4, Theorem 5.2.1] (about the interpolation of L^p spaces). Then we conclude using the injection of L^p spaces into some Lorentz spaces [4, section 1.3 and Exercice 1.6.8].

In the case $\theta \neq 0$, the proof is actually simpler. We can prove it using the usual TT^* method and the Hölder inequality instead of the Hardy–Littlewood–Sobolev inequality. □

1.2. The manifold of PLS. We study the nonlinear Dirac equation

$$\begin{cases} \mathrm{i}\partial_t \psi = H\psi + \nabla F(\psi), \\ \psi(0, \cdot) = \psi_0 \end{cases} \tag{1.8}$$

with $\psi \in \mathcal{C}^1(I, H^1(\mathbb{R}^3, \mathbb{C}^4))$ for some real open interval I which contains 0 and $H = D_m + V$. The nonlinearity $F : \mathbb{C}^4 \mapsto \mathbb{R}$ is a differentiable map for the real structure of $\mathbb{C}^4$, and hence the ∇ symbol has to be understood for the real structure of $\mathbb{C}^4$: for the usual Hermitian product of $\mathbb{C}^4$, one has

$$DF(v)h = \Re\langle \nabla F(v), h\rangle.$$

If F has a gauge invariance (see (0.2) or Assumption 1.4), this equation may have stationary solutions, i.e., solutions of the form $e^{-\mathrm{i}Et}\phi_0$, where ϕ_0 satisfies the nonlinear stationary equation

$$E\phi_0 = H\phi_0 + \nabla F(\phi_0).$$

We notice that the Dirac operator D_m has an interesting invariance property due to its matrix structure. This invariance can be shared by some perturbed Dirac operators and gives a consequence of a theorem of Kramers; see [2, 37]. Indeed if we introduce K, the antilinear operator defined by

$$K \begin{pmatrix} \psi \\ \chi \end{pmatrix} = \begin{pmatrix} \sigma_2 \overline{\psi} \\ \sigma_2 \overline{\chi} \end{pmatrix} \quad \text{with } \sigma_2 = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}, \tag{1.9}$$

the operator D_m commutes with K. So if V also commutes with K, we obtain that the eigenspaces of H are always of even dimension. Here we work with the following assumption.

Assumption 1.3. The potential V commutes to K. The operator $H := D_m + V$ has only two double eigenvalues $\lambda_0 < \lambda_1$, with $\{\phi_0, K\phi_0\}$ and $\{\phi_1, K\phi_1\}$ as associated orthonormalized bases.

Remark 1.2. This assumption is fulfilled by $V = \lambda\phi\beta$, where ϕ is a smooth version of the characteristic function of the unit ball and λ is big enough to obtain a first set of double eigenvalues $\pm E$ with E close to m for H but not too much to avoid the appearance of a second set of eigenvalues. Computations have been sketched in [6], and similar ones have been done in detail for $V = \lambda\phi Id_{\mathbb{C}^4}$ in [47].

We also need the next assumption.

Assumption 1.4. The function $F : \mathbb{C}^4 \mapsto \mathbb{R}$ is in $\mathcal{C}^\infty(\mathbb{R}^8, \mathbb{R})$ and satisfies $F(z) = O(|z|^4)$ as $z \to 0$. Moreover, it has the following invariance property:

$$\forall z \in \mathbb{C}^4,\ \forall u_1, u_2 \in \mathbb{C}^2 \text{ with } |u_1|^2 + |u_2|^2 = 1,\ F(u_1 z + u_2 K z) = F(z).$$

Remark 1.3. The invariance property is actually equivalent to both the gauge invariance with respect to the semigroups generated by i and $e^{\mathrm{i}\theta}K$ for all real θ.

We note that this assumption includes the cubic nonlinearity mentionned in the introduction. We were not able to go beyond the cubic order in the present work. This comes from the fact that here orbital stability does not hold or, equivalently, we have no a priori control in the L^2 norm of the perturbation. We bound it with $L^\infty_t H^s$ and $L^2_t B^\beta_{\infty,2}$ ($s > \beta + 2 > 2$) using Duhamel's formula and Theorem 1.2. Our method asks the nonlinearity to be at least cubic; see the proof of Lemma 3.8.

We also note that both the nonlinearity and the potential are smooth. This is not optimal. Concerning the nonlinearity, what we have in mind is a smooth cubic nonlinearity. For instance, a nonlinearity of class $\mathcal{C}^4$ should be enough. It can be improved, but we are not looking for an optimal result in this direction. The advantage of choosing smooth potentials and nonlinearities here is that it allows stationary states to be smooth. It also avoids having to introduce a new parameter giving a bound for the regularity of our solutions.

We obtain the following proposition.

PROPOSITION 1.1 (PLS manifold). *If Assumptions* 1.1–1.4 *hold, then for any* $\sigma \in \mathbb{R}^+$*, there exist* Ω*, a neighborhood of* 0 *in* $\mathbb{C}^2$*; a smooth map*

$$h : \Omega \mapsto \{\phi_0, K\phi_0\}^\perp \cap H^2(\mathbb{R}^3, \mathbb{C}^4) \cap L^2_\sigma(\mathbb{R}^3, \mathbb{C}^4);$$

and a smooth map $E : \Omega \mapsto \mathbb{R}$ *such that* $S((u_1, u_2)) = u_1\phi_0 + u_2 K\phi_0 + h((u_1, u_2))$ *satisfies, for all* $U \in \Omega$*,*

$$HS(U) + \nabla F(S(U)) = E(U)S(U), \tag{1.10}$$

with the following properties:

$$\begin{cases} h((u_1,u_2)) = \left(\frac{u_1}{|(u_1,u_2)|} Id_{\mathbb{C}^4} + \frac{u_2}{|(u_1,u_2)|} K\right) h\left((|(u_1,u_2)|, 0)\right) \quad \forall U = (u_1,u_2) \in \Omega, \\ h(U) = O(|U|^2), \\ E(U) = E(|U|), \\ E(U) = \lambda_0 + O(|U|^2). \end{cases}$$

Proof. This result is adapted from [38, Proposition 2.2] after the reduction due to the invariance of the problem with respect to K. ∎

Moreover, we have the following lemma.

LEMMA 1.2 (exponential decay). *For any* $\beta \in \mathbb{N}^4$*,* $s \in \mathbb{R}^+$*, and* $p, q \in [1, \infty]$*, there exist* $\gamma > 0$*,* $\varepsilon > 0$*, and* $C > 0$ *such that for all* $U \in B_{\mathbb{C}^2}(0, \varepsilon)$ *one has*

$$\|e^{\gamma\langle Q\rangle} \partial^\beta_U S(U)\|_{B^s_{p,q}} \le C \|S(U)\|_2,$$

where

$$\partial^\beta_{(u_1,u_2)} = \frac{\partial^{|\beta|}}{\partial^{\beta_1}\Re u_1 \partial^{\beta_2}\Im u_1 \partial^{\beta_3}\Re u_2 \partial^{\beta_4}\Im u_2}.$$

Proof. This is proved as in [7, Lemma 4.1], where we used ideas from [23]. ∎

1.3. The unstable manifold and the stabilization. Each stationary solution previously introduced has, as in [7], a stable manifold. Under the following assumption, we can prove that a small perturbation of a stationary solution starting outside of this manifold leaves any neighborhood of this stationary solution. We work with the following assumption.

Assumption 1.5. The resonant condition

$$|\lambda_1 - \lambda_0| > \min\{|\lambda_0 + m|, |\lambda_0 - m|\}$$

holds. Moreover, we have the Fermi golden rule

$$\Gamma(\phi) > 0, \tag{1.11}$$

where, for any nonzero eigenvector ϕ associated with λ_0, $\Gamma(\phi)$ is given by

$$\Gamma(\phi) = \lim_{\varepsilon \to 0, \varepsilon > 0} \left\langle d^2F(\phi)\phi_1, \Im\left((H - \lambda_0) + (\lambda_1 - \lambda_0) - \mathrm{i}\varepsilon\right)^{-1} P_c(H) d^2F(\phi)\phi_1 \right\rangle.$$

In this assumption, the notation d^2F denotes the differential of ∇F with respect to the real structure of $\mathbb{C}^4$. Let us introduce the linearized operator $JH(U)$ around a stationary state $S(U)$:

$$H(U) = H + d^2F(S(U)) - E(U).$$

We note that the operator $H(U)$ is not $\mathbb{C}$-linear but only $\mathbb{R}$-linear. Hence we work with the space $L^2(\mathbb{R}^3, \mathbb{R}^4 \times \mathbb{R}^4)$ instead of $L^2(\mathbb{R}^3, \mathbb{C}^4)$ by writing

$$\begin{pmatrix} \Re\phi \\ \Im\phi \end{pmatrix}$$

instead of ϕ. The multiplication by $-\mathrm{i}$ becomes the operator

$$J = \begin{pmatrix} 0 & I_{\mathbb{R}^4} \\ -I_{\mathbb{R}^4} & 0 \end{pmatrix}.$$

Now we mention some spectral properties of the *real operator* $JH(U)$ in $L^2(\mathbb{R}^3, \mathbb{C}^8)$ (the complexification of $L^2(\mathbb{R}^3, \mathbb{R}^4 \times \mathbb{R}^4)$) which are needed to state and to understand our main theorem. These properties will be proved in subsection 2.1.

PROPOSITION 1.2 (spectrum of $JH(U)$). *The operator $JH(U)$ in $L^2(\mathbb{R}^3, \mathbb{C}^4 \times \mathbb{C}^4)$ has a four-dimensional algebraic kernel, and its spectrum is symmetric with respect to the imaginary and real axes.*

Outside the imaginary axis the spectrum is discrete. The sum of the algebraic eigenspaces associated with the right of the imaginary axis (i.e., associated with $\{z \in \mathbb{C}, \Re z > 0\}$) is stable under complex conjugation, and its dimension is 4. The same is true for the left of the imainary axis.

The rest of the spectrum is the essential (or continuous) spectrum. We write $\mathcal{H}_c(U)$ for the space associated with the continuous spectrum. The space $J\mathcal{H}_c(U)$ is the orthogonal of the previous eigenspaces and the geometric kernel of $JH(U)$ and is invariant by the complex conjugation.

Proof. See subsection 2.1 below. □

We will work on the real part of the previous spaces. Let $X_u(U) \subset L^2(\mathbb{R}^3, \mathbb{R}^4 \times \mathbb{R}^4)$ be the real part of the sum of the spectral subspaces associated with $\{z \in \mathbb{C}, \Re z > 0\}$. We introduce a real basis $(\xi_i(U))_{i=1,\dots,4}$ of $X_u(U)$.

We will also work in the real part of the sum of the spectral subspaces associated with $\{z \in \mathbb{C}, \Re z < 0\}$: $X_s(U) \subset L^2(\mathbb{R}^3, \mathbb{R}^4 \times \mathbb{R}^4)$. We introduce a real basis $(\xi_i(U))_{i=5,\dots,8}$ of $X_s(U)$.

We define $\xi(U) = (\xi_i(U))_{i=1,\dots,8}$. We can state our main theorems which will be proved in sections 2–5.

THEOREM 1.3 (central manifold and asymptotic stability). *If Assumptions* 1.1–1.5 *hold, then, for $s > \beta + 2 > 2$ and $\sigma > 3/2$, there exist $\varepsilon > 0$, a continuous map $r : B_{\mathbb{C}^2}(0,\varepsilon) \mapsto \mathbb{R}$ with $r(U) = O(|U|^2)$, $C > 0$, $\mathcal{V}$ a neighborhood of $(0,0)$ in*

$$\mathcal{S} = \{(U,z)\,;\, U \in B_{\mathbb{C}^2}(0,\varepsilon),\ z \in \mathcal{H}_c(U) \cap B_{H^s}(0,r(U))\}$$

endowed with the metric of $\mathbb{C}^2 \times H^s$, and a map $\Psi : \mathcal{V} \mapsto \mathbb{R}^8$, smooth on $\mathcal{V} \setminus (0,0)$, satisfying for any nonzero $U \in B_{\mathbb{C}^2}(0,\varepsilon)$

$$\|\Psi(U,z)\| = O(\|z\|_{H^s}^2)$$

for all $z \in \mathcal{H}_c(U) \cap B_{H^s}(0,r(U))$ with $(U,z) \in \mathcal{V}$ such that the following is true.

For any initial condition of the form

$$\psi_0 = S(U_0) + z_0 + A \cdot \xi(U_0)$$

with $(U_0,z_0) \in \mathcal{V}$ and $A = \Psi(U_0,z_0)$, there exists a solution $\psi \in \cap_{k=0}^2 \mathcal{C}^k\left(\mathbb{R}, H^{s-k}\right)$ of (1.8) *with initial condition ψ_0, and this solution is unique in $L^\infty((-T,T), H^s(\mathbb{R}^3,\mathbb{C}^4))$ for any $T > 0$.*

Moreover, we have for all $t \in \mathbb{R}$

$$\psi(t) = e^{-\mathrm{i}\int_0^t E(U(v))\,dv} S(U(t)) + \varepsilon(t) \tag{1.12}$$

with $\|\dot{U}\|_{L^q(\mathbb{R})} \le C\|z_0\|_{H^s}^2$ for all $q \in [1,\infty]$, $\lim_{t\to\pm\infty} U(t) = U_{\pm\infty}$, and

$$\max\left\{ \|\varepsilon\|_{L^\infty(\mathbb{R}^\pm, H^s)}\,,\ \|\varepsilon\|_{L^2(\mathbb{R}^\pm, H^s_{\pm\sigma})}\,, \|\varepsilon\|_{L^2(\mathbb{R}^\pm, B^\beta_{\infty,2})} \right\} \le C\|z_0\|_{H^s}.$$

Remark 1.4. We mention that to our knowledge the local well-posedness of the Cauchy problem associated with (1.8) has not been obtained for initial data in the energy space H^1. Some results are available for slightly bigger spaces; see, for instance, the work of Escobedo and Vega [19] and Machihara, Nakamura, Nakanishi, and Ozawa [34, 33, 32].

In our case, we work with data in H^s with $s > 2$; this comes from the fact that we use inhomogeneous Strichartz estimates (1.7) with $q = \tilde{q} = \infty$ (see the proof of Lemma 3.8).

THEOREM 1.4 (center-stable and center-unstable manifolds). *With the same assumptions and notation as in Theorem* 1.3, *let $\mathcal{CM}$ be the graph of $(U,z) \in \mathcal{V} \mapsto S(U) + z + \Psi(U,z)\cdot\xi(U)$. Then for the set*

$$\widetilde{\mathcal{S}} = \{(U,z,p)\,;\, U \in B_{\mathbb{C}^2}(0,\varepsilon),\ z \in \mathcal{H}_c(U) \cap B_{H^s}(0,r(U)),\ p \in B_{\mathbb{R}^4}(0,r(U))\}$$

endowed with the metric of $\mathbb{C}^2 \times H^s \times \mathbb{R}^4$, there exist $C > 0$, $\gamma > 0$, neighborhoods $\mathcal{W}_\pm$ of $(0,0,0)$ in $\widetilde{\mathcal{S}}$, and maps $\Phi_\pm : \mathcal{W}_\pm \mapsto \mathbb{R}^4$, smooth on $\mathcal{W}_\pm \setminus \{(0,0,0)\}$ with

$$\|\Phi_\pm(U,z,p)\| = O(\|z\|_{H^s}^2 + \|p\|^2)$$

for all $(U,z,p)\in\mathcal{W}_\pm$*, such that for any initial condition of the form*

$$\psi_0 = S(U_0)+z_0+(p_+,p_-).\xi(U_0)$$

not in $\mathcal{CM}$*, i.e.,*

$$(p_+,p_-)\neq\Psi(U_0,z_0),$$

the following is true.

1. *If* $(U_0,z_0,p_+)\in\mathcal{W}_+$ *and* $p_-=\Phi_+(U_0,z_0,p_+)$ *(resp., if* $(U_0,z_0,p_-)\in\mathcal{W}_-$ *and* $p_+=\Phi_-(U_0,z_0,p_-)$*), then for any small neighborhood* $\mathcal{O}$ *of* $S(U_0)$ *containing* ψ_0 *there exist* $t_+>0$ *(resp.,* $t_->0$*) and a solution*

$$\psi_+\in\cap_{k=0}^2\mathcal{C}^k([-t_+;+\infty),H^{s-k}),$$

respectively,

$$\psi_-\in\cap_{k=0}^2\mathcal{C}^k((-\infty;t_-],H^{s-k}),$$

of (1.8) *with initial condition* ψ_0*, and in* $L^\infty((-T',T),H^s(\mathbb{R}^3,\mathbb{C}^4))$ *this solution is unique for any* $T>0$ *(resp.,* $T\in(0,t_-)$*) and any* $T'\in(0,t_+)$ *(resp.,* $T'<0$*).*
Moreover, there exist $C>0$*,* $\phi_\pm(t)\in\mathcal{CM}$*, and* $\rho_+(t)\in X_s(U_0)$ *(resp.,* $\rho_-(t)\in X_u(U_0)$*) for all* $t>-t_+$ *(resp., for all* $t<t_-$*) such that* $\psi_\pm(t)=\phi_\pm(t)+\rho_\pm(t)$ *with*

$$\|\rho_\pm(t)\|_{H^s}\le C\,\|\rho_\pm(0)\|_{H^s}\,e^{\mp\gamma t}\ \text{ as } t\to\pm\infty\quad\text{ and }\quad\psi_\pm(\mp t_\pm)\notin\mathcal{O}.$$

We also have

$$\phi_\pm(t)=e^{-\mathrm{i}\int_0^t E(U_\pm(v))\,dv}S(U_\pm(t))+\varepsilon_\pm(t)\quad\forall t>t_-\ (\text{resp., }\forall t<-t_+)$$

with

$$\left\|\dot U_+\right\|_{L^q((-t_+,+\infty))}\le C\left(\|z_0\|_{H^s}+\|\rho_\pm(0)\|_{H^s}\right)^2,$$

respectively,

$$\left\|\dot U_-\right\|_{L^q((-\infty,t_-))}\le C\left(\|z_0\|_{H^s}+\|\rho_\pm(0)\|_{H^s}\right)^2,$$

for all $q\in[1,\infty]$*,* $\lim\limits_{t\to\pm\infty}U_\pm(t)=U_{\pm\infty}$*, and*

$$\max\left\{\|\varepsilon_+\|_{L^\infty((-t_+,+\infty),H^s)},\ \|\varepsilon_+\|_{L^2((-t_+,+\infty),H^s_{-\sigma})},\|\varepsilon_+\|_{L^2((-t_+,+\infty),B^\beta_{\infty,2})}\right\}\le C\left(\|z_0\|_{H^s}+\|\rho_\pm(0)\|_{H^s}\right),$$

respectively,

$$\max\left\{\|\varepsilon_-\|_{L^\infty((-\infty,t_-),H^s)},\ \|\varepsilon_-\|_{L^2((-\infty,t_-),H^s_{-\sigma})},\|\varepsilon_-\|_{L^2((-\infty,t_-),B^\beta_{\infty,2})}\right\}\le C\left(\|z_0\|_{H^s}+\|\rho_\pm(0)\|_{H^s}\right).$$

2. *If* $(U_0, z_0, p_+) \in \mathcal{W}_+$ *and* $(U_0, z_0, p_-) \in \mathcal{W}_-$ *with* $p_- \neq \Phi_+(U_0, z_0, p_+)$ *and* $p_+ \neq \Phi_-(U_0, z_0, p_-)$*, then there exist* $t_+(\psi_0) > 0$*,* $t_-(\psi_0) < 0$*, and a unique solution* ψ *of* (1.8) *with initial condition* ψ_0 *such that, for any small neighborhood* $\mathcal{O}$ *of* $S(U_0)$ *containing* ψ_0*,* $\phi \in \cap_{k=0}^{2} \mathcal{C}^k([t_-; t_+], H^{s-k})$ *with* $\psi(t_+) \notin \mathcal{O}$ *and* $\psi(t_-) \notin \mathcal{O}$*. This solution is unique in* $L^\infty((T', T), H^s(\mathbb{R}^3, \mathbb{C}^4))$ *for any* $T \in (0, t_+)$ *and any* $T' \in (t_-, 0)$*.*

Remark 1.5. In Theorem 1.4, we note that when $U_0 = 0$, then $z_0 = 0$ and $p = 0$, and so the theorem does not say anything for this case. In fact, the charge conservation gives the orbital stability of 0. But we cannot extend the previous results to 0 since we can build a manifold of stationary states tangent to the eigenspace associated with λ_1, similarly to Proposition 1.1.

One can also note that the previous result (and the following ones) are still valid when switching the roles of λ_0 and λ_1 as long as the resonance condition, Assumption 1.5, is fulfilled with λ_0 and λ_1 switched. If the resonance condition does not hold, the result in [7] can be applied if the strict nonresonance condition holds:

$$|\lambda_0 - \lambda_1| < \min\{|\lambda_1 + m|, |\lambda_1 - m|\},$$

and we obtain a stable manifold associated with λ_1. As far as we know, the cases

$$|\lambda_0 - \lambda_1| = \min\{|\lambda_1 + m|, |\lambda_1 - m|\} \quad \text{and} \quad |\lambda_1 - \lambda_0| = \min\{|\lambda_0 + m|, |\lambda_0 - m|\}$$

are open.

1.4. The nonlinear scattering. If we choose a localized z_0, we are able to further expand (1.12) as stated by the following theorems that are also proved in sections 2–5.

THEOREM 1.5. *With the assumptions and the notation of Theorem* 1.3*, for the set*

$$\mathcal{S}_\sigma = \left\{(U, z);\ U \in B_{\mathbb{C}^2}(0, \varepsilon),\ z \in \mathcal{H}_c(U) \cap B_{H^s_\sigma}(0, r(U))\right\}$$

endowed with the metric of $\mathbb{C}^2 \times H^s_\sigma$*, there exists a neighborhood* $\mathcal{V}_\sigma$ *of* $(0, 0)$ *in* $\mathcal{S}_\sigma$ *such that the following is true. If* $A = \Psi(U_0, z_0)$ *with* $(U_0, z_0) \in \mathcal{V}_\sigma$*, there exist* $\mathcal{V}^\pm_\sigma$ *open neighborhoods of* $(0, 0)$ *in* $\mathcal{S}_\sigma$ *and* $(V_{\pm\infty}; z_{\pm\infty}) \in \mathcal{V}^\pm_\sigma$*, such that*

$$|V_{\pm\infty} - U_0| \leq C\|z_0\|^2_{H^s_\sigma}, \quad \|z_{\pm\infty} - z_0\|_{H^s} \leq C\|z_0\|^2_{H^s_\sigma},$$

and for all $t \in \mathbb{R}$

$$\psi(t) = e^{-\mathrm{i}tE(V_{\pm\infty})} S(V_\pm(t)) + e^{JtE(V_{\pm\infty})} e^{JtH(V_{\pm\infty})} z_{\pm\infty} + \varepsilon_\pm(t)$$

with

$$\begin{aligned}
\left|\dot V_\pm(t) + \mathrm{i}\,(E(V_\pm(t)) - E(V_{\pm\infty}))\right| &\leq \frac{C}{\langle t\rangle^2}\|z_0\|^2_{H^s_\sigma},\\
|V_\pm(t) - V_{\pm\infty}| &\leq \frac{C}{\langle t\rangle}\|z_0\|_{H^s_\sigma},\\
\max\left\{\|\varepsilon_\pm(t)\|_{H^s},\ \|\varepsilon_\pm(t)\|_{H^{s}_{-\sigma}}, \|\varepsilon_\pm(t)\|_{B^\beta_{\infty,2}}\right\} &\leq \frac{C}{\langle t\rangle^2}\|z_0\|^2_{H^s_\sigma},\\
\text{and} \quad \left\|e^{-JtH(V_{\pm\infty})} e^{J\int_0^t (E(V_\pm(s)) - E(V_{\pm\infty})\,ds} \varepsilon_\pm(t)\right\|_{H^s_{\frac{3}{2}}} &\leq \frac{C}{\langle t\rangle^{\frac{1}{2}}}\|z_0\|^2_{H^s_\sigma}
\end{aligned}$$

for all $t \in \mathbb{R}$.

Moreover, the maps

$$(U_0; z_0) \in \mathcal{V}_\sigma \mapsto (V_{\pm\infty}; z_{\pm\infty}) \in \mathcal{V}_\sigma^\pm$$

are bijective.

Remark 1.6. The fact that z_0 is localized gives us the convergence of

$$\int_0^t E(U(v))\, dv - tE(U_{\pm\infty})$$

as $t \to \pm\infty$ and allows us to obtain an asymptotic profile for the dispersive part of the perturbed solution ϕ. Indeed, to obtain an asymptotic profile for z, we need an asymptotic profile for the phase and hence the existence of

$$\lim_{t\to\pm\infty} \int_0^t E(U(v))\, dv - tE(U_{\pm\infty}) = \int_0^{\pm\infty} \left(E(U(v)) - E(U_{\pm\infty})\right)\, dv. \tag{1.13}$$

But in the present case, we control $\dot{U}$ by the third power of some spatial norm of z. If z belongs just to some Lebesgue space in time, we are able to obtain just that

$$\lim_{t\to\pm\infty} \left(E(U(v)) - E(U_{\pm\infty})\right) = 0,$$

which gives nothing about the existence of (1.13). To obtain this existence, we need pointwise estimates in time for z (see subsection 3.2.4). This requires z_0 to be localized in space. Moreover, we believe that a result similar to [22, Theorem 1.9] showing slow pointwise decay for the perturbation in the nonlocalized case can be adapted here. Such phenomena prevent (1.13) from existing.

What we call the nonlinear scattering result is essentially the fact that the maps

$$(U_0; z_0) \in \mathcal{V}_\sigma \mapsto (U_{\pm\infty}; z_{\pm\infty}) \in \mathcal{V}_\sigma^\pm$$

are well defined and bijective.

Using wave operators for the pair $(JH(U), JD_m)$, we can obtain an expansion of the form $\psi(t) = e^{-\mathrm{i}\int_0^t E(U(v))\, dv} S(U_{\pm\infty}) + e^{-\mathrm{i}tD_m} z_\pm + \varepsilon_\pm(t)$ but will only have

$$\|z_{\pm\infty} - z_0\|_{H^s} \leq C\|z_0\|_{H^s}$$

and

$$\max\left\{ \|\varepsilon_\pm\|_{L^\infty(\mathbb{R}^\pm, H^s)},\ \|\varepsilon_\pm\|_{L^2(\mathbb{R}^\pm, H^s_{-\sigma})}, \|\varepsilon_\pm\|_{L^2(\mathbb{R}^\pm, B^\beta_{\infty,2})} \right\} \leq C\|z_0\|_{H^s}.$$

Or, using wave operators for the pair $(JH(U), JH)$, we can obtain an expansion of the form $\psi(t) = e^{-\mathrm{i}\int_0^t E(U(v))\, dv} S(U_{\pm\infty}) + e^{-\mathrm{i}tH} z_\pm + \varepsilon_\pm(t)$ with

$$\|z_{\pm\infty} - z_0\|_{H^s} \leq C\left(|U_0| + \|z_0\|_{H^s}\right)\|z_0\|_{H^s}$$

and

$$\max\left\{ \|\varepsilon_\pm\|_{L^\infty(\mathbb{R}^\pm, H^s)},\ \|\varepsilon_\pm\|_{L^2(\mathbb{R}^\pm, H^s_{-\sigma})}, \|\varepsilon_\pm\|_{L^2(\mathbb{R}^\pm, B^\beta_{\infty,2})} \right\} \leq C\left(|U_0| + \|z_0\|_{H^s}\right)\|z_0\|_{H^s}.$$

But in these cases, we cannot obtain a nice asymptotic completness statement.

We have (in the previous expansions)

$$\max\left\{\sup_{t\in\mathbb{R}}\left(\left\|e^{JtH(U_{\pm\infty})}z_{\pm\infty}\right\|_{H^s}\right),\ \sup_{t\in\mathbb{R}}\left(\langle t\rangle^{3/2}\left\|e^{JtH(U_{\pm\infty})}z_{\pm\infty}\right\|_{H^s_{-\sigma}}\right),\right.$$
$$\left.\sup_{t\in\mathbb{R}}\left(\langle t\rangle^{3/2}\left\|e^{JtH(U_{\pm\infty})}z_{\pm\infty}\right\|_{B^\beta_{\infty,2}}\right)\right\}\le C\|z_{\pm\infty}\|_{H^s}.$$

This follows from Lemmas 3.13 and 3.14.

Outside the center manifold, we can also have an expansion of the same type. But due to the presence of exponentially stable and unstable directions, one cannot expect a scattering result of the same type. Actually we cannot obtain the injectivity of the corresponding mappings. We have the following theorem.

THEOREM 1.6. *With the assumptions and notation of Theorem* 1.4, *for the sets*

$$\widetilde{\mathcal{S}}_\sigma=\left\{(U,z,p);\ U\in B_{\mathbb{C}^2}(0,\varepsilon),\ z\in\mathcal{H}_c(U)\cap B_{H^s_\sigma}(0,r(U)),\ p\in B_{\mathbb{R}^4}(0,r(U))\right\}$$

endowed with the metric of $\mathbb{C}^2\times H^s_\sigma\times\mathbb{R}^4$, *there exist* $C>0$, $\gamma>0$, *and neighborhoods* $\mathcal{W}^\pm_\sigma$ *of* $(0,0,0)$ *in* $\mathcal{S}_\sigma$ *such that the following is true.*

If $\psi_0\notin\mathcal{CM}$, $(U_0,z_0,p_+)\in\mathcal{W}^+_\sigma$ *and* $p_-=\Phi_+(U_0,z_0,p_+)$ *(resp.,* $(U_0,z_0,p_-)\in\mathcal{W}^-_\sigma$ *and* $p_+=\Phi_-(U_0,z_0,p_-)$*), then there exist* $C>0$, $\phi_\pm(t)\in\mathcal{CM}$, *and* $\rho_\pm(t)\in X_s(U_0)$ *for all* $t>-t_+$ *(resp., for all* $t<t_-$*) such that* $\psi_\pm(t)=\phi_\pm(t)+\rho_\pm(t)$ *with*

$$\|\rho_\pm(t)\|_{H^s}\le C\,\|\rho_\pm(0)\|_{H^s}\,e^{\mp\gamma t}\ \text{ as } t\to\pm\infty\quad\text{ and }\quad\psi(\mp t_\pm)\notin\mathcal{O}.$$

Further, there exist $(V_{\pm\infty};z_{\pm\infty})\in\mathcal{S}$ *such that*

$$|V_{\pm\infty}-U_0|\le C\left(\|z_0\|_{H^s_\sigma}+\|\rho_\pm(0)\|_{H^s}\right)^2,$$
$$\|z_{\pm\infty}-z_0\|_{H^s}\le C\left(\|z_0\|_{H^s_\sigma}+\|\rho_\pm(0)\|_{H^s}\right)^2,$$

and for all $t>-t_+$ *(resp., for all* $t<t_-$*)*

$$\phi_\pm(t)=e^{-\mathrm{i}tE(V_{\pm\infty})}S(V_\pm(t))+e^{JtE(V_{\pm\infty})}e^{JtH(V_{\pm\infty})}z_{\pm\infty}+\varepsilon_\pm(t)$$

with

$$\left|\dot V_\pm(t)+\mathrm{i}\left(E(V_\pm(t))-E(V_{\pm\infty})\right)\right|\le\frac{C}{\langle t\rangle^2}\left(\|z_0\|_{H^s_\sigma}+\|\rho_\pm(0)\|_{H^s}\right)^2,$$
$$|V_\pm(t)-V_{\pm\infty}|\le\frac{C}{\langle t\rangle}\left(\|z_0\|_{H^s_\sigma}+\|\rho_\pm(0)\|_{H^s}\right)^2,$$
$$\max\left\{\|\varepsilon_\pm(t)\|_{H^s},\ \|\varepsilon_\pm(t)\|_{H^s_{-\sigma}},\|\varepsilon_\pm(t)\|_{B^\beta_{\infty,2}}\right\}\le\frac{C}{\langle t\rangle^2}\left(\|z_0\|_{H^s_\sigma}+\|\rho_\pm(0)\|_{H^s}\right)^2,$$
$$\left\|e^{-JtH(V_{\pm\infty})}e^{J\int_0^t(E(V_\pm(s))-E(V_{\pm\infty}))\,ds}\varepsilon_\pm(t)\right\|_{H^s_{\frac{3}{2}}}\le\frac{C}{\langle t\rangle^{\frac12}}\left(\|z_0\|_{H^s_\sigma}+\|\rho_\pm(0)\|_{H^s}\right)^2$$

for all $t>-t_+$ *(resp., for all* $t<t_-$*).*

2. Linearized operator and exponentially stable and unstable manifolds. We study the dynamics associated with (1.8) around a stationary state. We will use spectral properties of the linearized operator around a stationary state.

2.1. The spectrum of the linearized operator. Here we study the spectrum of the linearized operator associated with (1.8) around a stationary state $S(U)$. Let us recall

$$H(U) = H + d^2F(S(U)) - E(U),$$

where d^2F is the differential of ∇F. The operator $H(U)$ is $\mathbb{R}$-linear but not $\mathbb{C}$-linear. Replacing $L^2(\mathbb{R}^3, \mathbb{C}^4)$ by $L^2(\mathbb{R}^3, \mathbb{R}^4 \times \mathbb{R}^4)$ with the inner product obtained by taking the real part of the inner product of $L^2(\mathbb{R}^3, \mathbb{C}^4)$, we obtain a symmetric operator. We then complexify this real Hilbert space and obtain $L^2(\mathbb{R}^3, \mathbb{C}^4 \times \mathbb{C}^4)$ with its natural Hermitian product. This process transforms the operator $-\mathrm{i}$ into

$$J = \begin{pmatrix} 0 & Id_{\mathbb{C}^4} \\ -Id_{\mathbb{C}^4} & 0 \end{pmatrix}.$$

For $\phi \in L^2(\mathbb{R}^3, \mathbb{R}^4 \times \mathbb{R}^4) \subset L^2(\mathbb{R}^3, \mathbb{C}^4 \times \mathbb{C}^4)$, we still write ϕ instead of

$$\begin{pmatrix} \Re\phi \\ \Im\phi \end{pmatrix}.$$

The extension of $H(U)$ to $L^2(\mathbb{R}^3, \mathbb{C}^4 \times \mathbb{C}^4)$ is also written $H(U)$ and is now a real operator. The extension of K (see (1.9)) is also written K.

The linearized operator associated with (1.8) around the stationary state $S(U)$ is given by $JH(U)$. We shall now study its spectrum.

Differentiating (1.10), we have that for $U = (u_1, u_2) \in \Omega$

$$\mathcal{H}_0(u_1, u_2) = \mathrm{span}\left\{ \frac{\partial}{\partial \Re u_1} S(u_1, u_2), \frac{\partial}{\partial \Im u_1} S(u_1, u_2), \frac{\partial}{\partial \Re u_2} S(u_1, u_2), \frac{\partial}{\partial \Im u_2} S(u_1, u_2) \right\}$$

is invariant under the action of $JH(U)$. Differentiating the gauge invariance property for S, we notice that $JS(U) \in \mathcal{H}_0(U)$; differentiating the gauge invariance property for F, we also obtain

$$JH(U)JS(U) = 0;$$

and differentiating (1.10), we obtain for any $\beta \in \mathbb{N}^4$ with $|\beta| = 1$

$$JH(U)\partial_U^\beta S(U) = (\partial_U^\beta E)(U)JS(U).$$

The space $\mathcal{H}_0(U)$ is contained in the algebraic null space of $JH(U)$; in fact, it is exactly the algebraic null space as proved in what follows.

Now we state our results on the spectrum of $JH(U)$. The first deals with the excited states part. Using the function Γ introduced in Assumption 1.5, we have the following proposition.

PROPOSITION 2.1. *If Assumptions* 1.1–1.5 *hold, let*

$$\widetilde{\Gamma}(|U|) = \inf_{V \in \mathbb{C}^2, |V| = |U|} \Gamma(S(V)).$$

For any sufficiently small U, we have

$$\widetilde{\Gamma}(|U|) > 0.$$

There exist continuous maps $E_1^i : B_{\mathbb{C}^2}(0,\varepsilon) \mapsto \mathbb{R}$ *smooth outside* $U = 0$ *with*

$$\begin{cases} \Im E_1^i(U) = (\lambda_1 - E(U)) + O(|U|^4), \\ \Re E_1^i(U) = \Gamma^i(U) + O(|U|^6), \end{cases}$$

where

$$\Gamma^i(U) \geq \frac{1}{2}\widetilde{\Gamma}(|U|)$$

for $i \in \{1,2\}$ *such that* $E_1^i(U)$, $\overline{E_1^i}(U)$, $-E_1^i(U)$, *and* $-\overline{E_1^i}(U)$ *are eigenvalues of* $JH(U)$. *These eigenvalues are simple if* $E_1^1 \neq E_1^2$; *otherwise the associated algebraic eigenspace is of dimension* 2. *We can define maps* $\Phi_\pm^i(U) : B_{\mathbb{C}^2}(0,\varepsilon) \mapsto H^s$ *such that*

- $\{\Phi_+^1(U), \Phi_+^2(U)\}$ *is an orthonormal basis of the sum of the eigenspaces associated with* $E_1^1(U)$ *and* $E_1^2(U)$,
- $\{\overline{\Phi_+^1(U)}, \overline{\Phi_+^2(U)}\}$ *is an orthonormal basis of the eigenspaces associated with* $\overline{E_1^1(U)}$ *and* $\overline{E_1^2(U)}$,
- $\{\Phi_-^1(U), \Phi_-^2(U)\}$ *is an orthonormal basis of the sum of the eigenspaces associated with* $-E_1^1(U)$ *and* $-E_1^2(U)$,
- $\{\overline{\Phi_-^1(U)}, \overline{\Phi_-^2(U)}\}$ *is an orthonormal basis of the sum of the eigenspaces associated with* $-\overline{E_1^1(U)}$ *and* $-\overline{E_1^2(U)}$.

In the case of double eigenvalues $E_1^1(U) = E_1^2(U)$ *one should consider algebraic eigenspaces instead of the sum of the eigenspaces. In any case, the associated projectors are continuous around* $U = 0$ *and smooth outside* $U = 0$.

Moreover, for any $\beta \in \mathbb{N}^4$, $s \in \mathbb{R}^+$ *and* $p, q \in [1,\infty]$. *There exist* $\gamma > 0$, $\varepsilon > 0$, *and* $C > 0$ *such that, for all* $U \in B_{\mathbb{C}^2}(0,\varepsilon) \setminus \{0\}$ *and for any* $i \in \{1,2\}$, *one has*

$$\|e^{\gamma\langle Q\rangle}\partial_U^\beta \Phi_\pm^i(U)\|_{B^s_{p,q}} \leq C\|\Phi_\pm^i(U)\|_2, \tag{2.1}$$

where

$$\partial^\beta_{(u_1,u_2)} = \frac{\partial^{|\beta|}}{\partial^{\beta_1}\Re u_1 \partial^{\beta_2}\Im u_1 \partial^{\beta_1}\Re u_2 \partial^{\beta_2}\Im u_2}.$$

Proof. Using Weyl's sequences, we prove that the essential spectrum of $JH(U)$, for small U, is the essential spectrum of $J(H - E(U))$, i.e., $\mathrm{i}(\mathbb{R} \setminus (-c,c))$ with $c = \min\{m - E(U), m + E(U)\}$. So z with nonzero real part is in the spectrum of $JH(U)$ if and only if it is an isolated eigenvalue with finite multiplicity. The multiplicity of z here is the dimension of the algebraic kernel of $JH(U) - z$.

We now investigate properties of the eigenvalues. To do so, we use ideas from the proof of [54, Theorem 2.2]. The equation to solve for excited states is

$$(JH(U) - z)\,\phi = 0. \tag{2.2}$$

We consider solutions of the form $\phi = (v_1 S_1 + v_2 K S_1) + \eta$, where S_1 is the normalized eigenvector of JH,

$$S_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 \\ -i\phi_1 \end{pmatrix},$$

with $(v_1, v_2) \in \mathbb{C}^2$ such that $|v_1|^2 + |v_2|^2 = 1$ and $\eta \in \{S_1, KS_1\}^\perp$. The orthogonal relation is taken, in fact, with respect to J (but since $JS_1 = \mathrm{i}S_1$ and $JKS_1 = \mathrm{i}KS_1$, we can take it in the usual way). For $z \in \mathbb{C} \setminus \mathrm{i}\mathbb{R}$, we obtain the equation

$$\eta = (J(H - E(U)) - z)^{-1} P_1^\perp W(U) \{(v_1 S_1 + v_2 K S_1) + \eta\} \tag{2.3}$$

with $P_1^\perp$ the orthogonal projector into $\{JS_1\}^\perp = \{S_1\}^\perp$ and $W(U) = JH(U) - J(H - E(U))$. We notice that $\{S_1\}^\perp$ is invariant under the action of $J(H - E(U))$. To solve this equation in η for a fixed U and z, we note that if

$$\Re z > 0 \quad \text{and} \quad |\Im z| \geq c,$$

where $c = \min\{m + E(U), m - E(U)\}$, the series

$$k(U,z)(v_1 S_1 + v_2 K S_1) = (J(H - E(U)) - z)^{-1}$$

$$\times P_1^\perp \sum_{k \geq 0} \left(-W(U) (J(H - E(U)) - z)^{-1} P_1^\perp \right)^k W(U)(v_1 S_1 + v_2 K S_1)$$

is convergent in L^2 for sufficiently small $|U|$ and $|\Re z| = O(|U|^2)$ using the limiting absorption principle (1.4) and the bound of the resolvent $\left\| (H - z')^{-1} \right\| \leq |\Im z'|^{-1}$ in L^2. Hence, we have a solution of (2.3).

Then we solve the equation in z. We obtain from (2.2) the equations

$$\langle (JH(U) - z)\, \phi, S_1 \rangle = 0 \quad \text{and} \quad \langle (JH(U) - z)\, \phi, KS_1 \rangle = 0$$

with $\phi = (v_1 S_1 + v_2 K S_1) + k(U,z)(v_1 S_1 + v_2 K S_1)$. Hence, z is an eigenvalue of the matrix $A(U,z)$ defined by

$$\begin{pmatrix} \langle JH(U)(S_1 + k(U,z)S_1), S_1 \rangle & \langle JH(U)(S_1 + k(U,z)S_1), KS_1 \rangle \\ \langle JH(U)(KS_1 + k(U,z)KS_1), S_1 \rangle & \langle JH(U)(KS_1 + k(U,z)KS_1), KS_1 \rangle \end{pmatrix}$$

in $\mathbb{C}^2$. So we study the zeros of $\det(A(U,z) - z)$. We have that $d_\varepsilon : z \mapsto \det(A(U, z + \varepsilon) - z)$ is an analytic function for z in $\{z \in \mathbb{C}, \Re z > -\varepsilon\}$ for any $\varepsilon \geq 0$, $\det(A(0, z + \varepsilon) - z) = (z - \mathrm{i}(\lambda_1 - \lambda_0))^2$ and $\det(A(U, z + \varepsilon) - z) - \det(A(0, z + \varepsilon) - z) \to 0$ as $U \to 0$ uniformly in z and ε in a bounded set. Since at $(U, \varepsilon) = (0, \varepsilon)$ $(\varepsilon > 0)$ we have a double zero, for $(U, 0)$ close to $(0, 0)$, we have two zeros counted with multiplicity.

Hence if z is a zero of $d_0(U, \cdot)$ and hence an eigenvalue of $A(U,z)$, we have with an associated normalized eigenvector $(v_1, v_2) \in \mathbb{C}^2$ and $\psi = (v_1 S_1 + v_2 K S_1)$

$$z = \mathrm{i}\,(\lambda_1 - \lambda_0) + \langle W(U)\psi, \psi \rangle + \sum_{k \geq 0} \Big\langle JH(U) (J(H - E(U)) - z)^{-1}$$

$$\times P_1^\perp \left(-W(U) (J(H - E(U)) - z)^{-1} P_1^\perp \right)^k W(U)\psi, \psi \Big\rangle$$

$$= \mathrm{i}\,(\lambda_1 - \lambda_0) + \langle W(U)\psi, \psi \rangle$$

$$+ \sum_{k \geq 0} \left\langle P_1^\perp \left(-W(U) (J(H - E(U)) - z)^{-1} P_1^\perp \right)^k W(U)\psi, \psi \right\rangle$$

$$+\sum_{k\geq 0}\Big\langle (W(U)+z)\,(J(H-E(U))-z)^{-1}$$

$$\times\, P_1^{\perp}\Big(-W(U)\,(J(H-E(U))-z)^{-1}P_1^{\perp}\Big)^{k}W(U)\psi,\psi\Big\rangle.$$

Since $P_1^{\perp}(v_1S_1+v_2KS_1)=0$, we introduce the function

$$f(z)=\mathrm{i}\,(\lambda_1-\lambda_0)+\Big\langle W(U)\psi,\psi\Big\rangle+\sum_{k\geq 0}\big\langle W(U)\,(J(H-E(U))-z)^{-1}$$

$$\times\, P_1^{\perp}\left(-W(U)\,(J(H-E(U))-z)^{-1}\,P_1^{\perp}\right)^{k}W(U)\psi,\psi\Big\rangle.$$

Since $J(v_1S_1+v_2KS_1)=-\mathrm{i}(v_1S_1+v_2KS_1)$, we obtain that

$$\Re\,\langle W(U)(v_1S_1+v_2KS_1),(v_1S_1+v_2KS_1)\rangle=0.$$

Thus for $z\in\mathbb{C}\setminus\mathrm{i}\mathbb{R}$, we have

$$\begin{aligned}
&\Re f(z)\\
&=\Re\Big\langle W(U)\,(J(H-E(U))-z)^{-1}\,P_1^{\perp}W(U)(v_1S_1+v_2KS_1),(v_1S_1+v_2KS_1)\Big\rangle\\
&\quad+O(|U|^6)\\
&=\Im\big\langle d^2F(S(U))\,((H-E(U))+zJ)^{-1}\\
&\quad\times P_1^{\perp}d^2F(S(U))(v_1S_1+v_2KS_1),(v_1S_1+v_2KS_1)\big\rangle+O(|U|^6).
\end{aligned}$$

Then using (1.11) and

$$\begin{aligned}
&((H-E(U))+zJ)^{-1}\\
&\qquad=\frac{1}{2}\left(((H-E(U))-\mathrm{i}z)^{-1}\,(I_{C^2}+\mathrm{i}J)+((H-E(U))+\mathrm{i}z)^{-1}\,(I_{C^2}-\mathrm{i}J)\right),
\end{aligned}$$

we obtain

$$\begin{aligned}
&\Im\big\langle d^2F(S(U))((H-E(U))+zJ)^{-1}P_1^{\perp}d^2F(S(U))(v_1S_1+v_2KS_1),(v_1S_1+v_2KS_1)\big\rangle\\
&=\frac{1}{2}\big(\Im\big\langle d^2F(S(U))((H-E(U))-\mathrm{i}z)^{-1}d^2P_1^{\perp}F(S(U))(v_1S_1+v_2KS_1),(v_1S_1+v_2KS_1)\big\rangle\\
&+\Im\big\langle d^2F(S(U))((H-E(U))+\mathrm{i}z)^{-1}P_1^{\perp}d^2F(S(U))(v_1S_1+v_2KS_1),(v_1S_1+v_2KS_1)\big\rangle\big)\\
&-\Im\big\langle d^2F(S(U))\big((H-E(U))^2+z^2\big)^{-1}zJP_1^{\perp}d^2F(S(U))(v_1S_1+v_2KS_1),(v_1S_1+v_2KS_1)\big\rangle,
\end{aligned}$$

and so, using regularity results of the resolvent of [20, Theorem 1.7], we obtain

$$\begin{aligned}
&\Im\big\langle d^2F(S(U))\,((H-E(U))+(\mathrm{i}(\lambda_1-\lambda_0)+0)\,J)^{-1}\\
&\qquad\times P_1^{\perp}d^2F(S(U))(v_1S_1+v_2KS_1),(v_1S_1+v_2KS_1)\big\rangle\\
&=\frac{1}{2}\Im\big\langle d^2F(S(U))\,((H-E(U))+(\lambda_1-\lambda_0)-\mathrm{i}0)^{-1}\\
&\qquad\times P_c(H)d^2F(S(U))(v_1S_1+v_2KS_1),(v_1S_1+v_2KS_1)\big\rangle.
\end{aligned}$$

Using Assumption 1.5, the limiting absorption principle (1.4), and regularity results of [20, Theorem 1.7], we obtain

$$\Re f(z) = \frac{1}{2}\Im\big\langle d^2F(S(U))\left((H-E(U))+(\lambda_1-\lambda_0)-\mathrm{i}0\right)^{-1} \\ \times P_c(H)d^2F(S(U))(v_1S_1+v_2KS_1),(v_1S_1+v_2KS_1)\big\rangle + O(|U|^6)$$

for z in a ball of radius of order $|U|^2$ around $\mathrm{i}\,(\lambda_1-\lambda_0)$ and for small U. We also prove in the same way that

$$\Im f(z) = (\lambda_1-\lambda_0) + O(|U|^4)$$

for z in a ball of radius of order $|U|^2$ around $\mathrm{i}\,(\lambda_1-\lambda_0)$ and for small U.

This proves that the zeros of $A(U,z)$ at the right of the imaginary axis are close to $\mathrm{i}\,(\lambda_1-\lambda_0)$. Hence we have obtained that, for small U, $JH(U)$ has at most two eigenvalues. Counted with algebraic multiplicity, we have that it is exactly two at the right side of the imaginary axis. Indeed, a complex number z is in the resolvent set of $JH(U)$ if and only if $JH(U)-z$ is invertible or if and only if $H(U)+Jz$ is invertible.

Let us consider for $\varepsilon>0$

$$J_\varepsilon = \begin{pmatrix} \frac{\varepsilon}{2} & (1+\mathrm{i}\frac{\varepsilon}{2})Id_{\mathbb{C}^4} \\ (-1-\mathrm{i}\frac{\varepsilon}{2})Id_{\mathbb{C}^4} & \frac{\varepsilon}{2} \end{pmatrix}$$

and the set of z for which the operator

$$H(U) - J_\varepsilon z$$

is invertible. When $U=0$, we have after multiplication by

$$P = \frac{1}{\sqrt{2}}\begin{pmatrix} \mathrm{i} & 1 \\ 1 & \mathrm{i} \end{pmatrix}$$

on the left and by P^{-1} on the right (that is to say, diagonalization of J_ε)

$$\begin{pmatrix} H-\lambda_0+\mathrm{i}z+\mathrm{i}\varepsilon & 0 \\ 0 & H-\lambda_0-\mathrm{i}z \end{pmatrix},$$

which is invertible if

$$z \notin \mathrm{i}\left\{\mathbb{R}\setminus(-m-\lambda_0,m-\lambda_0)\cup\{0,\lambda_1-\lambda_0\}\right\}$$

and

$$z \notin -\mathrm{i}\left\{\mathbb{R}\setminus(-m-\lambda_0,m-\lambda_0)\cup\{0,\lambda_1-\lambda_0\}\right\}-\varepsilon.$$

Since we have closed operators, applying Theorems 3.16 and 5.33 of [28, Chapter IV] to $J_\varepsilon^{-1}H(U)$, we obtain that close to $\mathrm{i}(\lambda_1-\lambda_0)$ the operator $J_\varepsilon^{-1}H(U)$ has two simple eigenvalues or one double (algebraic) eigenvalue when U is small. Now consider ε as a parameter and U fixed, and applying again Theorems 3.16 and 5.33 of [28, Chapter IV], we obtain that, counted with multiplicity, $JH(U)$ has two eigenvalues since the discrete spectrum of $JH(U)=J_0^{-1}H(U)$ close to $\mathrm{i}(\lambda_1-\lambda_0)$ at the right

of the imaginary axis is outside the imaginary axis. Theorems 3.16 and 5.33 of [28, Chapter IV] give as well the continuity with respect to U of the associated projector.

The continuity of eigenvalues follows from the continuity of the zeros of d_ε introduced earlier.

Letting $P_1(U)$ be the spectral projector associated with $\left\{E_1^i(U), i \in \{1,2\}\right\}$, the eigenvalue of $JH(U)$, we have the formula

$$P_1(U) = -\frac{1}{2\mathrm{i}\pi}\int_\Gamma (JH(U) - z)^{-1} dz,$$

where Γ is a positively oriented circle at the right of the imaginary axis around $E_1^1(U')$ and $E_1^2(U')$ for some fixed U' and U sufficiently close to U'; see [28, Formula I.1.16]. Since the resolvent is smooth outside $U = 0$, we have that $P_1(U)$ is smooth outside $U = 0$.

Using complex conjugation, we obtain the corresponding result for the spectrum at the right of the imaginary axis in a neighborhood of $\mathrm{i}\,(\lambda_1 - \lambda_0)$.

Using Weyl's sequences, we prove that the essential spectrum of $(JH(U))^* = -H(U)J$, for small U, is the essential spectrum of $-(H - E(U))J = -J(H - E(U))$. So z with nonzero real part is in the spectrum of $(JH(U))^*$ if and only if it is an isolated eigenvalue; see [28, Theorem IV.5.33]. Then to obtain $-E_1^i(U)$ and $-\overline{E_1^i(U)}$, we notice that $E_1^i(U)$ and $\overline{E_1^i(U)}$ are eigenvalues of $(JH(U))^*$. Using the symmetry $J(JH(U)) = -(JH(U))^*J$, we show that any algebraic eigenvector ϕ of $(JH(U))^*$ associated with λ, $J\phi$ is an algebraic eigenvector of $JH(U)$ associated with $-\lambda$. Hence, repeating the previous proof for $(JH(U))^*$, we also obtain the multiplicity two. Except for the symmetry of the eigenvalues, this can also be obtained by adapting the previous proof to the left part of the complex plane.

The exponential decay works as in Lemma 1.2. □

Remark 2.1. If $F(z)$ is homogeneous of order p, then there exist $\varepsilon, \Gamma_1, \Gamma_2 > 0$ such that for all $U \in B_{\mathbb{C}^2}(0, \varepsilon)$

$$|U|^{p-2}\,\Gamma_1 \le \widetilde{\Gamma}(|U|) \le |U|^{p-2}\,\Gamma_2.$$

We just write $S((u_1, u_2)) = u_1\phi_0 + u_2 K\phi_0 + h((u_1, u_2))$, expand $\Gamma(U)$, and use Assumption 1.5 with the regularity results of the resolvent from [20, Theorem 1.7].

The following proposition deals with the essential spectrum of our linearized operator.

PROPOSITION 2.2. *If Assumptions* 1.1–1.5 *hold, for any sufficiently small nonzero* $U \in \mathbb{C}^2$, *let*

$$\mathcal{H}_1(U) = \operatorname{span}\left\{\Phi_+^1(U), \Phi_+^2(U), \overline{\Phi_+^1(U)}, \overline{\Phi_+^2(U)}, \Phi_-^1(U), \Phi_-^2(U), \overline{\Phi_-^1(U)}, \overline{\Phi_-^2(U)}\right\}.$$

The orthogonal space of $\mathcal{H}_0(U) \oplus \mathcal{H}_1(U)$,

$$\mathcal{H}_c(U) = \{J\mathcal{H}_0(U) \oplus \mathcal{H}_1(U)\}^\perp,$$

is invariant under the action of $JH(U)$.

We also have that, for $\mathbf{P}_c(U)$, *the orthogonal projector onto* $J\mathcal{H}_c(U)$ *is a bounded operator from* $H^s_\sigma(\mathbb{R}^3, \mathbb{C}^8)$ *or* $B^s_{p,q}(\mathbb{R}^3, \mathbb{C}^8)$ *to itself for any reals* $s, \sigma \in \mathbb{R}$ *and* $p, q \in [1, \infty]$, *and for* $U' \in B_{\mathbb{C}^2}(U, \varepsilon)$, *with sufficiently small* $\varepsilon > 0$, *that*

$$\mathbf{P}_c(U)|_{\mathcal{H}_c(U')} : \mathcal{H}_c(U') \mapsto \mathcal{H}_c(U)$$

is an isomorphism and its inverse $R(U',U)$ can be extended to a bounded operator from $H^s_\sigma(\mathbb{R}^3,\mathbb{C}^8)$ or $B^s_{p,q}(\mathbb{R}^3,\mathbb{C}^8)$ to itself for any reals $s,\sigma\in\mathbb{R}$ and $p,q\in[1,\infty]$.

Moreover, there exists $C>0$ such that we have

$$\|\psi\|_X \le C\,\|P_c\psi\|_X \tag{2.4}$$
$$\forall\psi\in\mathcal{H}_c(U)\cap X \text{ with } X=H^s_\sigma(\mathbb{R}^3,\mathbb{C}^8) \text{ or } B^s_{p,q}(\mathbb{R}^3,\mathbb{C}^8),$$
$$\forall s,\sigma\in\mathbb{R},\quad \forall p,q\in[1,\infty];$$

$$\int_{\mathbb{R}} \|\langle Q\rangle^{-\sigma} e^{sJH(U)}\mathbf{P}_c(U)\psi\|_2^2\,ds \le C\|\psi\|_2^2 \tag{2.5}$$
$$\forall\psi\in L^2,\quad \forall\sigma\ge 1;$$

$$\left\|e^{tJH(U)}\mathbf{P}_c(U)\psi\right\|_2 \le C\|\psi\|_2 \tag{2.6}$$
$$\forall t\in\mathbb{R},\quad \forall\psi\in L^2,$$

and $\mathcal{H}_c(U)$ contains no eigenvector.

Remark 2.2. We use the same notation for $\mathcal{H}_c(U)$ and its real part (see definition below) that appears in our main theorems. We just note that $\mathcal{H}_c(U)$ appears when we discuss spectral properties in our proof. Then when we talk about dynamical properties, we deal with its real part. We remind the reader that the real part of $\mathcal{H}_c(U)$ is left invariant by $JH(U)$.

Proof. We prove that there is no other eigenvector by proving that smoothness estimate (2.5) takes place over

$$\mathcal{H}_c(U)=\{J\mathcal{H}_0(U)\oplus\mathcal{H}_1(U)\}^\perp.$$

First we prove that

$$\mathbf{P}_c((U))|_{\mathcal{H}_c(U')}:\mathcal{H}_c(U')\mapsto\mathcal{H}_c(U)$$

is an isomorphism. To prove it, we exhibit an inverse $R(U',U)$ which is the projector onto $\mathcal{H}_c(U')$ associated with the decomposition $\mathcal{H}_0(U)\oplus\mathcal{H}_1(U)\oplus\mathcal{H}_c(U')$ of $L^2(\mathbb{R}^3,\mathbb{C}^8)$. Indeed, we have $\{\mathcal{H}_0(U)\oplus\mathcal{H}_1(U)\}\cap\mathcal{H}_c(U')=\{0\}$ when U' and U are close to one another and $\mathrm{codim}\mathcal{H}_c(U')=\dim\{\mathcal{H}_0(U)\oplus\mathcal{H}_1(U)\}$. We have a decomposition of $L^2(\mathbb{R}^3,\mathbb{C}^8)$ into closed subspaces; hence the associated projectors are continuous. So $R(U',U)$ should be of the form

$$R(U',U)=Id+\sum_i |J\xi_i(U)\rangle\,\langle\alpha_i(U',U)|\,,$$

where $\xi_i(U)$ is a basis of the eigenspaces of $JH(U)$ and $(\alpha_i(U',U))_i$ solve the equations

$$J\xi_j(U')+\sum_i \langle J\xi_i(U),J\xi_j(U')\rangle\,\alpha_i(U',U)=0.$$

Such an α exists because the matrix $(\langle J\xi_i(U),J\xi_j(U')\rangle)_{i,j}$ is invertible. Indeed, if it is not invertible, there exists $\phi\in\mathcal{H}_1(U')=\mathrm{span}\{\xi_j(U')\}$ orthogonal to $\mathcal{H}_1(U)=$

span$\{\xi_j(U)\}$, and hence $\phi \in \mathcal{H}_c(U)$ with $P_c(U')\phi = 0$. But for all $\psi \in \mathcal{H}_c(U)$, we have

$$\|\psi\| \le \|P_c(U')\psi\| + \|(1 - P_c(U'))\psi\| \le \|P_c(U')\psi\| + C|U - U'|\,\|\psi\|,$$

and hence for ϕ this gives a contradiction.

The boundedness of R in $\mathcal{B}(H^s_\sigma(\mathbb{R}^3, \mathbb{C}^8))$ or $\mathcal{B}(B^s_{p,q}(\mathbb{R}^3, \mathbb{C}^8))$ follows from the exponential decay of eigenvectors and their derivatives; see Proposition 1.1, Lemma 1.2, and Proposition 2.1.

Let us now consider the orthogonal projector P_c associated with the continuous subspace of JH. Since the eigenvectors of JH are exponentially decaying, we can extend P_c to obtain an operator of $L^2_{\pm\sigma}$ into itself. The same is true for $P_c(U)$, and hence we can consider the extension of $\mathcal{H}_c(U)$ to $L^2_{\pm\sigma}$. We still call it $\mathcal{H}_c(U)$.

For all $\psi \in \mathcal{H}_c(U)$,

$$\|\psi\|_{L^2_{-\sigma}} \le \|P_c\psi\|_{L^2_{-\sigma}} + \|(1 - P_c)\psi\|_{L^2_{-\sigma}}.$$

Since $1 - P_c$ is the projector into the eigenspaces of H and ψ is orthogonal to the eigenvectors of $JH(U)$, we obtain that

$$\|(1 - P_c)\psi\|_{L^2_{-\sigma}} = \|(1 - P_c(U))\psi\|_{L^2_{-\sigma}} + O(|U|)\|\psi\|_{L^2_{-\sigma}}.$$

Since the components of eigenvectors with respect of $\mathcal{H}_c(U)$ are small with respect to U (see the proof of Lemma 2.1), the projection of ψ in the eigenspaces of H is small. Moreover, for a sufficiently small nonzero U, we obtain estimate (2.4) for $X = L^2_{-\sigma}$ with $\sigma > 0$.

The rest of estimate (2.4) follows in the same way using the exponential decay of eigenvectors (estimate (2.1) and Lemma 1.2).

We infer that

$$\begin{aligned}
&\|\langle Q\rangle^{-\sigma} e^{tJH(U)} \mathbf{P}_c(U)\psi\| \\
&\le C\|\langle Q\rangle^{-\sigma} \mathbf{P}_c e^{tJH(U)} \mathbf{P}_c(U)\psi\| \\
&\le C\|\langle Q\rangle^{-\sigma} \mathbf{P}_c e^{-\mathrm{i}t(H - E(U))} \mathbf{P}_c(U)\psi\| \\
&\quad + C\left\|\langle Q\rangle^{-\sigma} \int_0^t \mathbf{P}_c e^{-\mathrm{i}(t-s)(H-E(U))} D\nabla F(S(U)) e^{sJH(U)} \mathbf{P}_c(U)\psi\, ds\right\|.
\end{aligned}$$

Using estimates (1.1) and (1.3) of Theorem 1.1, we obtain estimate (2.5) for sufficiently small U:

$$\int_{\mathbb{R}} \|\langle Q\rangle^{-\sigma} e^{-sJH(U)} \mathbf{P}_c(U)\psi\|^2\, ds \le C\|\psi\|^2.$$

Hence there is no eigenvector in the range of $P_c(U)$. Using the inequalities (2.5); the conservation law for H; and Duhamel's formula,

$$e^{JtH(U)} = e^{-\mathrm{i}t(H-E(U))} + \int_0^t e^{-\mathrm{i}(t-s)(H-E(U))} Jd^2\nabla F(S(U)) e^{JsH(U)}\, ds,$$

we prove estimate (2.6). ∎

Since $\mathcal{H}_c(U)$ is closed and $\operatorname{codim}\mathcal{H}_c(U) = \dim\{\mathcal{H}_0(U) \oplus \mathcal{H}_1(U)\}$ and $\mathcal{H}_c(U) \cap \{\mathcal{H}_0(U) \oplus \mathcal{H}_1(U)\} = \{0\}$, we obtain $\mathcal{H}_0(U) \oplus \mathcal{H}_1(U) \oplus \mathcal{H}_c(U) = L^2(\mathbb{R}^3, \mathbb{C}^8)$ and the following proposition.

PROPOSITION 2.3. *Suppose that Assumptions* 1.1–1.5 *hold. Then the space* $\mathcal{H}_0(U)$ *is the geometric null space of* $JH(U)$.

2.2. Stable, unstable, and center manifolds. We can now obtain results similar to those of Bates and Jones [3]. We note that we will not prove that the Cauchy problem (1.8) is locally well-posed for an initial condition outside some manifolds (built below). However, this can be proved with the methods we present here or by generalizing to our case the results of Escobedo and Vega [19].

We have that $JH(U)$ as an operator in $L^2(\mathbb{R}^3,\mathbb{R}^8)$ is a closed densely defined operator that generates a continuous semigroup on $L^2(\mathbb{R}^3,\mathbb{R}^8)$. The spectrum of $JH(U)$ in $L^2(\mathbb{R}^3,\mathbb{R}^8)$ is the same as that of $JH(U)$ in $L^2(\mathbb{R}^3,\mathbb{C}^8)$, and so it splits into three parts,

$$\sigma_s(U)=\{\lambda\in\sigma(JH(u)),\ \Re\lambda<0\}=\left\{-E_1^1(U),-E_1^2(U),-\overline{E_1^1(U)},-\overline{E_1^2(U)}\right\},$$
$$\sigma_c(U)=\{\lambda\in\sigma(JH(u)),\ \Re\lambda=0\}=\{0\}\cup \mathrm{i}\left\{\mathbb{R}\setminus(-c,c)\right\},$$
$$\text{where } c=\min\{m+E(U),m-E(U)\},$$
$$\sigma_u(U)=\{\lambda\in\sigma(JH(u)),\ \Re\lambda>0\}=\left\{E_1^1(U),E_1^2(U),\overline{E_1^1(U)},\overline{E_1^2(U)}\right\},$$

each of which is associated with a spectral real subspace, respectively,

$$X_s(U)=\operatorname{span}_{\mathbb{R}}\left\{\Re\Phi_-^1(U),\Im\Phi_-^1(U),\Re\Phi_-^2(U),\Im\Phi_-^2(U)\right\},$$
$$X_u(U)=\operatorname{span}_{\mathbb{R}}\left\{\Re\Phi_+^1(U),\Im\Phi_+^1(U),\Re\Phi_+^2(U),\Im\Phi_+^2(U)\right\},$$
$$X_c(U)=\Re\mathcal{H}_0(U)\oplus\Re\mathcal{H}_c(U),$$

where we used the notation $\Re\Psi=(1/2)\left(\Psi+\overline{\Psi}\right)$, $\Im\Psi=-(\mathrm{i}/2)\left(\Psi-\overline{\Psi}\right)$, and $\Re X=\{\Re\Psi,\Psi\in X\}$, the real part of the space X. The spaces $X_s(U)$ and $X_u(U)$ are finite dimensional. Let us write $\pi^c(U)$, $\pi^s(U)$, and $\pi^u(U)$ for the projectors associated with the decomposition $X_c(U)\oplus X_s(U)\oplus X_u(U)$. Since the eigenvectors also belong to L^2_σ for any $\sigma\in\mathbb{R}$, the projectors $P_c(U)$ and $\pi^c(U)$ can be defined in L^2_σ for any real σ. We can extend, in this way, the spaces $\mathcal{H}_c(U)$ and $X_c(U)$ to L^2_σ for any $\sigma\in\mathbb{R}$. We have the following lemma.

LEMMA 2.1. *If Assumptions* 1.1–1.5 *hold, then, for any* $\sigma\in\mathbb{R}$*, there exist positive reals* r,C_1,C_2 *such that, for all* $t\in\mathbb{R}$*, we have*

$$C_1\frac{1}{f_+(t)}\le\left\|e^{tJH(U)}\pi^s(U)\right\|_{\mathcal{B}(L^2_\sigma)}\le C_2\frac{1}{f_-(t)},\tag{2.7}$$

$$C_1f_-(t)\le\left\|e^{tJH(U)}\pi^u(U)\right\|_{\mathcal{B}(L^2_\sigma)}\le C_2f_+(t),\tag{2.8}$$

$$\left\|e^{tJH(U)}\pi^c(U)\right\|_{\mathcal{B}(L^2_\sigma)}\le C_2\langle t\rangle^r,\tag{2.9}$$

where

$$f_\pm(t)=\begin{cases}e^{\gamma_\pm t} & \text{if } t\ge0,\\ e^{\gamma_\mp t} & \text{if } t\le0,\end{cases}$$

with

$$\gamma_-<\Gamma_-(U)=\min\{\Re E_1^1(U),\Re E_1^2(U)\},$$
$$\gamma_+>\Gamma_+(U)=\max\{\Re E_1^1(U),\Re E_1^2(U)\}.$$

Proof. The statements for the spaces $X^s(U)$ and $X^u(U)$ follow from (2.1).

The statement about $X^c(U)$ is a little more complicated. We note that we are not looking for an optimal r.

First, the result for $e^{-\mathrm{i}t(D_m+V)}$ in L^2_σ with $\sigma \in 2\mathbb{N}$ follows from [48, Theorem 8.5] (see also Proposition 3.1 below), which is based on the charge conservation. The case $\sigma \in \mathbb{R}$ follows by duality and interpolation.

Then for $e^{tJH(U)}\pi^c(U)$, we use Duhamel's formula,

$$e^{JtH(U)}\pi^c(U) = e^{-\mathrm{i}t(H-E(U))}\pi^c(U) + \int_0^t e^{-\mathrm{i}(t-s)(H-E(U))} Jd^2\nabla F(S(U)) e^{JsH(U)}\pi^c(U)\, ds;$$

then the assertion for $e^{tJH(U)}\pi^c(U)$ follows from the assertion for $e^{-\mathrm{i}t(D_m+V)}$, the charge conservation of $e^{tJH(U)}P_c(U)$ (see (2.6)), the fact that $e^{tJH(U)}S(U) = S(U)$, $e^{tJH(U)}\partial_U^\beta S(U) = \partial_U^\beta S(U) + t\partial_U^\beta E(U)S(U)$, and Lemma 1.2. ∎

By now we do not restrict our study to the space $L^2(\mathbb{R}^3,\mathbb{R}^8)$; we extend it to $L^2_\sigma(\mathbb{R}^3,\mathbb{R}^8)$ for any $\sigma \in \mathbb{R}$, but we still write $\mathcal{H}_c(U)$ and $X_c(U)$ for the extensions of these spaces to $L^2_\sigma(\mathbb{R}^3,\mathbb{R}^8)$ for any $\sigma \in \mathbb{R}$.

We now study the behavior of the solutions in L^2_σ of (1.8) centered around $S(U)$:

$$\partial_t\phi = JH(U)\phi + JN(U,\phi), \tag{2.10}$$

where $H(U) = H + d^2F(S(U)) - E(U)$, $N(U,\phi) = \nabla F(S(U)+\phi) - \nabla F(S(U)) - d^2F(S(U))\phi$, and d^2F is the differential of ∇F.

In this subsection, we study a modified equation which coincides with (2.10) as long as the solution stays in a neighborhood of a small $S(U)$:

$$\partial_t\phi = JH(U)\phi + JN_\varepsilon(U,\phi), \tag{2.11}$$

where $N_\varepsilon(U,\eta) = \rho(\varepsilon^{-1}\eta)N(U,\eta)$ and ρ is a smoothed version of the characteristic function of the unit ball of $\mathbb{R}^8$.

We state the following proposition.

PROPOSITION 2.4 (center-stable manifold). *If Assumptions* 1.1–1.5 *hold, then for any sufficiently small nonzero U, there exists around $S(U)$ a unique invariant smooth center-stable manifold $W^{cs}(U)$ for* (2.11) *built as a graph with value in $X_u(U)$ and tangent to $S(U) + X_c(U) \oplus X_s(U)$ at $S(U)$.*

Any solution $\phi \in L^2_\sigma$ of (2.11) *initially in the neighborhood of $S(U)$ tends as $t \to -\infty$ to $W^{cs}(U)$ with*

$$dist_{L^2_\sigma}(\phi(t), W^{cs}(U)) = O(e^{\gamma t}) \text{ as } t \to -\infty$$

for any $\gamma \in (0, \Gamma_-(U))$ and any $s,\sigma \in \mathbb{R}$, and for any sufficiently small neighborhood V of $S(U)$ any solution in V not in $W^{cs}(U)$ leaves V in finite positive time.

Remark 2.3. For any $s \in \mathbb{R}^+$, due to the exponential decay of eigenvectors, if $\phi \notin H^s_\sigma$, there exists $\psi \in W^{cs}(U)$ such that $\phi - \psi \in H^s_\sigma$, and we have

$$dist_{H^s_\sigma}(\phi(t), W^{cs}(U)) = O(e^{\gamma t}) \text{ as } t \to -\infty,$$

as shown in the following proof.

If we consider only small solutions, we obtain a locally invariant manifold for (2.10); that is to say, for any initial condition in the manifold there exists a corresponding solution of (2.10) which stays in this manifold in a small interval of time

around 0. We notice that in the following proofs the size of this invariant manifold, which is given by ε, is a function of U and this function is $O(\Gamma_-(U))$. From now on, we call this function r.

Proof. Our proof is an adaptation of that of Bressan [8], and we refer to it for more details. We make the proof only for the case $\sigma = 0$; the proof in the general case is similar.

First we prove that there is a global solution of (2.11) which does not grow much as $t \to +\infty$. We look for the solution as a fixed point:

$$y(t) = \mathcal{G}_\varepsilon(y_0, y)(t)$$

for any $y_0 \in X_s(U) \oplus X_c(U)$, where for small positive ε

$$\mathcal{G}_\varepsilon(y_0,\eta)(t) = e^{tJH(U)}y_0 + \int_0^t e^{(t-s)JH(U)}\pi^c(U)JN_\varepsilon(U,\eta(s))\,ds$$

$$+\int_0^t e^{(t-s)JH(U)}\pi^s(U)JN_\varepsilon(U,\eta(s))\,ds - \int_t^{+\infty} e^{(t-s)JH(U)}\pi^u(U)JN_\varepsilon(U,\eta(s))\,ds,$$

with $\pi^*(U)$ the projector into $X^*(U)$ with respect to the decomposition $\oplus_{*\in\{c,s,u\}}X^*(U)$.

Let us introduce, for any Γ smaller than $\Gamma_-(U)$, the space

$$Y_\Gamma = \left\{ y : \mathbb{R} \mapsto L^2(\mathbb{R}^3,\mathbb{C}^4),\ \exists C > 0,\ \|y(t)\|_2 \le Ce^{\Gamma|t|}\ \forall t \in \mathbb{R} \right\}.$$

For sufficiently small $\varepsilon > 0$, the map $\mathcal{G}_\varepsilon(y_0,\cdot)$ leaves Y_Γ invariant and is continuous for the norm

$$N_\Gamma : y \mapsto \sup_{t\in\mathbb{R}} \left\{ \|y(t)\|_2 e^{-\Gamma|t|} \right\}.$$

Moreover, it is a strict contraction for sufficiently small U and $\varepsilon > 0$. Actually we choose ε as a function of Γ which is $O(\Gamma)$. In fact, since $Y_\Gamma \subset Y_{\Gamma'}$ for $\Gamma < \Gamma'$, we obtain that ε as a function of U is $O(\Gamma_-(U))$. This proves the existence of the fixed point y.

Then we fix $h^{cs}_U(y_0) = y(0) - y_0$. The invariance of the graph of h^{cs}_U by the flow of (2.11) is immediate.

Now we prove the smoothness property. We have that $N_\varepsilon(U,\eta)$ is l times differentiable in η from $Y_{\Gamma'}$ to Y_Γ if $(l+1)\Gamma' \le \Gamma$ and that $\mathcal{G}_\varepsilon$ is l times differentiable from $X_c(U) \oplus X_s(U) \times Y_{\Gamma''}$ to Y_Γ if $2l\Gamma'' \le \Gamma$ (see [8]). We introduce the family $(\eta_n)_{n\in\mathbb{N}}$ satisfying

$$\eta_0 = 0 \quad \text{and} \quad \eta_{n+1} = \mathcal{G}_\varepsilon(y_0,\eta_n).$$

This sequence converges to y (the fixed point) in Y_Γ. Moreover, as functions of y_0, the convergence is uniform in Y_Γ (endowed with the norm N_Γ) on bounded sets of $X_s(U) \oplus X_c(U)$.

We want to prove that the sequence of their derivatives of order k with respect to η also converges in Y_Γ on bounded sets for any $\Gamma < \gamma(U)$. We prove it by induction in k. So suppose that $(\partial^j\eta_n)_{n\in\mathbb{N}}$ is converging in Y_Γ for all $j < k$ and any $\Gamma < \gamma(U)$. Then we have that for any $k \ge 2$ (see [8])

$$\partial\eta_n = \partial\mathcal{G}_\varepsilon(y_0,\eta_{n-1}) = L + M\left(\partial_\eta N_\varepsilon(U,\eta_{n-1})\partial\eta_{n_1}\right),$$
$$\partial^k\eta_n = \partial^k\mathcal{G}_\varepsilon(y_0,\eta_{n-1}) = M\left(\partial_\eta N_\varepsilon(U,\eta_{n-1})\partial^k\eta_{n-1} + \Psi_k(\eta_{n-1},\dots,\partial^{k-1}\eta_{n-1})\right),$$

with $L = e^{tJH(U)}$,

$$(M\eta)(t) = -\int_0^t e^{(t-s)JH(U)}\pi^{cs}(U)J\eta(s)\,ds + \int_t^{+\infty} e^{(t-s)JH(U)}\pi^{u}(U)J\eta(s)\,ds,$$

and Ψ_k a smooth function of k parameters. Hence since $M \circ \partial_\eta N_\varepsilon(U, y_{n-1})$ is a strict contraction in Y_Γ for sufficiently small ε and U (once more ε is $O(\Gamma_-(U))$), this proves the convergence of the sequence of kth derivatives in Y_Γ on bounded sets for any $\Gamma < \Gamma_-(U)$. Hence the sequences of derivatives of $(\eta_n)_{n\in\mathbb{N}}$ converge in Y_Γ on bounded sets for any $\Gamma < \Gamma_-(U)$. This gives the differentiability at any order of $y(0) = h(y_0)$. This also gives, since $N(U,\eta) = O(|\eta|^2)$ around zero, that $h(y_0) = O(|y_0|^2)$ around zero.

Now we want to prove that $W^{cs}(U)$ is attractive in negative time. In fact, $W^{cs}(U)$ is the graph of a smooth function $h : X^{cs} \mapsto X^u(U)$. Letting η be such that $S(U) + \eta$ is a solution of (1.8), we have

$$\partial_t \eta = JH(U)\eta + JN_\varepsilon(U,\eta),$$

where

$$\eta = y + r = y + h(y) + z$$

with $y = \pi^{cs}(U)\eta$, and we have the following equation for $z \in X^u(U)$:

$$\partial_t z = JH(U)z + M(U,y,z),$$

where

$$\begin{aligned} M(U,y,z) &= \pi^u(U)\left\{JN_\varepsilon(U,\eta) - JN_\varepsilon(U,y+h(y))\right\} \\ &\quad -Dh(y)\pi^{cs}(U)\left\{JN_\varepsilon(U,\eta) - JN_\varepsilon(U,y+h(y))\right\}. \end{aligned}$$

Using Duhamel's formula, we obtain

$$z(t) = e^{tJH(U)}z(0) + \int_0^t e^{(t-s)JH(U)}M(U,y(s),z(s))\,ds.$$

We obtain, since $z \in X^u(U)$,

$$\|z(t)\| \le e^{\gamma(U)t}\|z(0)\| + C\int_0^t e^{(t-s)\gamma(U)}\|M(U,y(s),z(s))\|\,ds,$$

and so for $\gamma \in (0, \Gamma_-(U))$

$$e^{-\gamma t}\|z(t)\| \le \|z(0)\| + C|U|\sup_{s\in[0,t]}\left\{e^{-\gamma s}\|z(s)\|\right\} + o\left(\sup_{s\in[0,t]}\left\{e^{-\gamma s}\|z(s)\|\right\}\right),$$

where C does not depend on U and z. Hence if $z(0)$ and U are small, we have that there exists $c > 0$ such that $\|z(t)\| \le ce^{\gamma t}$ for all $t \le 0$. We notice that, since $X^u(U) \subset H^s_\sigma$ for any $s \in \mathbb{R}^+$ and is finite dimensional (see Lemma 2.1), the time decay in L^2_σ also gives a time decay in H^s_σ gives for any $s \in \mathbb{R}^+$.

Now choose V a sufficiently small neighborhood of 0 and ϕ a solution of (2.11) initially in V but not in $W^{cs}(U)$. Suppose ϕ stays in V in positive time. We obtain that $\phi \in Y_\Gamma$. We have

$$\begin{aligned}\phi(t) &= e^{tJH(U)}\left(\pi^s(U)+\pi^c(U)\right)\phi(0)\\ &+\int_0^t e^{(t-s)JH(U)}\pi^s(U)JN_\varepsilon(U,\phi(s))\,ds+\int_0^t e^{(t-s)JH(U)}\pi^c(U)JN_\varepsilon(U,\phi(s))\,ds\\ &+e^{tJH(U)}\pi^u(U)\left(\pi^u(U)\phi(0)+\int_0^\infty e^{-sJH(U)}\pi^u(U)JN_\varepsilon(U,\phi(s))\,ds\right)\\ &-\int_t^\infty e^{(t-s)JH(U)}\pi^u(U)JN_\varepsilon(U,\phi(s))\,ds\end{aligned}$$

with

$$\pi^u(U)\phi(0)+\int_0^\infty e^{-sJH(U)}\pi^u(U)JN_\varepsilon(U,\phi(s))\,ds\neq 0.$$

Hence we obtain with (2.8) that $\phi(t)$ exponentially tends to infinity in norm. This is a contradiction, so ϕ leaves V in finite time. ∎

Then reversing the time direction, that is to say, replacing H by $-H$ and F by $-F$, we obtain with this theorem a locally invariant center-unstable manifold with the following corresponding properties.

PROPOSITION 2.5 (center-unstable manifold). *If Assumptions* 1.1–1.5 *hold, then for any sufficiently small nonzero U, there exists around $S(U)$ a unique smooth invariant center-unstable manifold $W^{cu}(U)$ for* (2.11)*, built as a graph with value in $X_S(U)$ and tangent to $S(U)+X_c(U)\oplus X_u(U)$ at $S(U)$.*

Any solution $\phi \in L^2_\sigma$ of (2.11) *initially in the neighborhood of $S(U)$ tends as $t\to+\infty$ to $W^{cu}(U)$ with, for any $s\in\mathbb{R}^+$,*

$$dist_{H^s_\sigma}(\phi(t),W^{cu}(U))=O(e^{-\gamma t})\ \text{as}\ t\to+\infty,$$

and for $\gamma\in(0,\Gamma_+(U))$, any $s,\sigma\in\mathbb{R}$, and for any V a sufficiently small neighborhood of $S(U)$, any solution in V not in $W^{cu}(U)$ leaves V in finite negative time.

We can build in the same way a center manifold which is the intersection of the previous two manifolds.

PROPOSITION 2.6 (center manifold). *If Assumptions* 1.1–1.5 *hold, then for any sufficiently small nonzero U, there exists around $S(U)$ a unique smooth invariant center manifold $W^c(U)$ for* (2.11)*, built as a graph with value in $X_s(U)\oplus X_u(U)$ and tangent to $S(U)+X_c(U)$ at $S(U)$.*

Moreover, we have that $W^c(U)=W^{cs}(U)\cap W^{cu}(U)$ and $W^c(U)$ contains the part of the PLS manifold which is in a small neighborhood of $S(U)$.

Proof. We build the center manifold with the same method as in the previous cases. We can also build a center-unstable manifold inside a center-stable manifold. More precisely, let $h^s_U : X_c(U)\oplus X_s(U)\mapsto X_u(U)$ be the map defining the center-stable manifold and let $h^u_U : X_c(U)\oplus X_u(U)\mapsto X_s(U)$ be the map defining the center-unstable manifold. A solution $y=S(U)+y_c+y_s+y_u$ with $y_*\in X_*(U)$ for $*\in\{c,s,u\}$ is in the center-stable manifold if $y_u=h^s_U(y_c,y_s)$. Hence to obtain a center-unstable manifold inside a center-stable manifold one has to solve, for each y_c, the equation

$$y_s=h^u_U(y_c,h^s_U(y_c,y_s)),$$

which can be solved inside a small ball for small y_c and small U by means of the fixed point theorem, since $h^*_U(y_c, z)$ is $O(|y_c|^2 + |z|^2)$ around zero for $* \in \{s, u\}$.

In the same way, we can also build a center-stable manifold inside the center-unstable manifold.

Using the uniqueness of the center manifold, we obtain that these two manifolds are equal to the center manifold and $W^c(U) = W^{cs}(U) \cap W^{cu}(U)$.

Then any stationary states in a small neighborhood of $S(U)$ converge to $W^{cs}(U)$ and $W^{cu}(U)$ using the stabilization results of Propositions 2.4 and 2.5. Hence, we have that it belongs to $W^{cs}(U) \cap W^{cu}(U) = W^c(U)$. ∎

In the following two sections, we study the dynamics inside and outside the center manifold, respectively.

3. The dynamics inside the center manifold. In this section, we prove that the dynamics inside the center manifold around $S(V_0)$, for small nonzero V_0, relaxes towards the PLS manifold. To this end, we use Theorems 1.1 and 1.2 regarding the time decay of the propagator associated with H.

3.1. Decomposition of the system. As in [7], we decompose a solution $\phi \in W^c(V_0)$ of (1.8) with respect to the spectrum of $JH(U)$, with U specified in what follows, and we study the equations for these different parts of the decomposition. We introduce

$$\mathcal{H}_0^{\perp_J}(u_1, u_2) = \Bigg\{ \eta \in L^2(\mathbb{R}^3, \mathbb{C}^8), \Big\langle J\eta, \frac{\partial}{\partial \Re u_1} S(u_1, u_2) \Big\rangle = \Big\langle J\eta, \frac{\partial}{\partial \Im u_1} S(u_1, u_2) \Big\rangle$$
$$= \Big\langle J\eta, \frac{\partial}{\partial \Re u_2} S(u_1, u_2) \Big\rangle = \Big\langle J\eta, \frac{\partial}{\partial \Im u_2} S(u_1, u_2) \Big\rangle = 0 \Bigg\}.$$

In fact, we have

$$\mathcal{H}_0^{\perp_J}(U) = \mathcal{H}_1(U) \oplus \mathcal{H}_c(U),$$

which is invariant under the action of $JH(U)$. We recall that $\mathcal{H}_1(U)$ is defined in Proposition 2.1 and $\mathcal{H}_c(U)$ in Proposition 2.2. We have the following lemma.

LEMMA 3.1. *If Assumptions* 1.1–1.4 *hold, let* $s, \sigma \in \mathbb{R}$. *Then there exist* $\varepsilon, \varepsilon' > 0$ *such that, for the manifold*

$$\Sigma = \Big\{ (U, \eta),\ U \in B_{\mathbb{C}^2}(0, \varepsilon'),\ \eta \in \mathcal{H}_0^{\perp_J}(U) \Big\}$$

endowed with the metric of $\mathbb{C}^2 \times H^s_\sigma$ *and any function* $\phi \in B_{H^s_\sigma}(0, \varepsilon)$, *there exists a unique* $(U, \eta) \in \Sigma$ *with*

$$\phi = S(U) + \eta.$$

Moreover, there exists a neighborhood $\mathcal{O}$ *of* $(0, 0) \in \Sigma$ *such that the mapping* $\phi \mapsto (U, \eta) \in \mathcal{O}$ *is smooth.*

Proof. To prove that Σ is a manifold, we use Proposition 2.2, which gives that it is locally isomorphic to some open subset of $\mathbb{C}^2 \times \mathcal{H}_c$ endowed with the metric of $\mathbb{C}^2 \times H^s_\sigma$. Then we work as in [22, Lemma 2.3]. Indeed we are looking, for all $\phi \in B_{H^s_\sigma}(0, \varepsilon)$, for a solution U in $\mathbb{C}^2 = \mathbb{R}^4$ of the equation

$$F(\phi, U) = (0, 0, 0, 0),$$

where

$$F(\phi,U)=\left(\left\langle J\left(\phi-S(U)\right),\frac{\partial}{\partial\Re u_1}S(u_1,u_2)\right\rangle,\left\langle J\left(\phi-S(U)\right),\frac{\partial}{\partial\Im u_1}S(u_1,u_2)\right\rangle,\right.$$
$$\left.\left\langle J\left(\phi-S(U)\right),\frac{\partial}{\partial\Re u_2}S(u_1,u_2)\right\rangle,\left\langle J\left(\phi-S(U)\right),\frac{\partial}{\partial\Im u_2}S(u_1,u_2)\right\rangle\right)$$

is smooth and satisfies $F(0,0)=0$ and

$$D_UF(0,0)=\begin{pmatrix}0&1&1&1\\1&0&1&1\\1&1&0&1\\1&1&1&0\end{pmatrix},$$

which is invertible in $\mathbb{R}^4$. Hence we apply the implicit function theorem. □

For any solution ϕ of (1.8) on an interval of time I containing 0, we write for $t\in I$

$$\phi(t)=e^{-\mathrm{i}\int_0^t E(U(s))\,ds}\left(S(U(t))+\eta(t)\right),$$

where $\eta(t)\in\mathcal{H}_0^{\perp_J}(U(t))$, and we want to solve the equation

$$\begin{aligned}\mathrm{i}\partial_t\eta&=\left\{H-E(U)\right\}\eta+\left\{\nabla F(S(U)+\eta)-\nabla F(S(U))\right\}-\mathrm{i}dS(U)\dot U\\&=\left\{H+d^2F(S(U))-E(U)\right\}\eta+N(U,\eta)-\mathrm{i}dS(U)\dot U\end{aligned}\tag{3.1}$$

for $\eta(t)\in\mathcal{H}_0^{\perp_J}(U(t))$. Here d^2F is the differential of ∇F and dS is the differential of S in $\mathbb{R}^2$. To close the system, we need the equation for U. This follows from the condition

$$\langle\eta(t),JdS(U(t))\rangle=0.$$

After a time derivation (as in [7]), we obtain the equation

$$\dot U(t)=-A(U(t),\eta(t))\langle N(U(t),\eta(t)),dS(U(t))\rangle,$$

where

$$A(U,\eta)=[\langle JdS(U),dS(U)\rangle-\langle J\eta,d^2S(U)\rangle]^{-1};$$

indeed the matrix $[\langle JdS(U(t)),dS(U(t))\rangle-\langle J\eta(t),d^2S(U(t))\rangle]$ is invertible for small $|U(t)|$ and $\|\eta(t)\|_2$ since we have

$$[\langle JdS(U(t)),dS(U(t))\rangle-\langle J\eta(t),d^2S(U(t))\rangle]=\begin{pmatrix}J&0_2\\0_2&J\end{pmatrix}+O(|U(t)|+\|\eta(t)\|_2).$$

We will need the following lemma.

LEMMA 3.2. *For any $s,s',\sigma\in\mathbb{R}$, any $p,q\in[1,\infty]$, and any $V_0\in\mathbb{C}^2\setminus\{0\}$ sufficiently small, there exist $\varepsilon,\varepsilon'>0$ such that, for the manifold*

$$\mathcal{S}(V_0,\varepsilon)=\left\{(U,z);\ U\in B_{\mathbb{C}^2}(V_0,\varepsilon),z\in\mathcal{H}_c(U)\cap B_{H^{s'}_\sigma}(0,\varepsilon')\right\}$$

endowed with the metric of $\mathbb{C}^2 \times H^{s'}_\sigma$*, there exists a unique map* $g : S(V_0,\varepsilon) \mapsto B^s_{p,q}(\mathbb{R}^3,\mathbb{C}^4)$ *which is smooth and satisfies* $g(U,z) \in \mathcal{H}_1(U)$*,* $z + g(U,z) \in \mathcal{H}_0^{\perp_J}(U)$*, and* $S(U) + z + g(U,z) \in W^c(V_0)$ *for all* $(U,z) \in S(V_0,\varepsilon)$*. Moreover, we have* $\|g(U,z)\|_{B^s_{p,q}} = O(\|z\|^2_{H^{s'}})$*.*

Proof. The fact that $\mathcal{S}(V_0,\varepsilon)$ is a manifold here is proved as in Lemma 3.1.

Let h^c be the function for which $W^c(V_0)$ is the graph. Any $\phi \in L^2(\mathbb{R}^3,\mathbb{R}^8)$ can be written in the form $S(V_0)+\tilde{U}\cdot DS(V_0)+\xi+\rho$ with $\rho \in \mathcal{H}_1(V_0)$ and $\xi \in \mathcal{H}_c(V_0)$. It can also be written in the form $S(U)+z+r$ with $r \in \mathcal{H}_1(U)$ and $z \in \mathcal{H}_c(U)$. These two decompositions in fact define two bijective smooth maps in sufficiently small sets (for the first we have a linear decomposition; for the second see Lemma 3.1). We write Ψ for the first and Φ for the second. Then $f = \Psi \circ \Phi^{-1}$ has three components following the decomposition $\mathcal{H}_0(V_0) \oplus \mathcal{H}_1(V_0) \oplus \mathcal{H}_c(V_0)$; we write them as (f_1, f_2, f_3). Then g is the solution of the implicit equation in r,

$$F(U,z,r) = f_2(U,z,r) - h^c(f_1(U,z,r), f_3(U,z,r)) = 0,$$

which can be solved by the implicit function theorem in $H^{s'}_\sigma$ since $\partial_r F(V_0,0,0)$ is invertible from $\mathcal{H}_1(V_0)$ to itself because $\partial_r f_2(V_0,0,0)$ ($f_2(V_0,r,0) = r$) is invertible from $\mathcal{H}_1(V_0)$ to itself and $Dh_c(0,0) = 0$.

The smoothness of g in the Besov spaces follows from the fact that $g(U,z) \in \mathcal{H}_1(U)$, and the exponential decay for excited states and their derivatives is given by (2.1).

Then we note that, for any U close to V_0, the previous proof can be applied to $W^c(U)$. It shows that $W^c(U)$, $W^c(V_0)$ are, in a neighborhood of $S(V_0)$, graphs of two functions on $\mathcal{S}(V_0,\varepsilon)$, which are equal up to a translation in $\mathbb{C}^2$ of the first parameter. Hence their graphs are equal, so locally $W^c(U) = W^c(V_0)$. The last assertion then follows from the fact that at $S(U)$, $W^c(U)$ is tangent to $S(U)+X^c(U)$ and $X^c(U)\cap\mathcal{H}_1(U) = \{0\}$. ∎

Hence decomposing η with respect to the spectrum of $JH(U)$, we write

$$\eta(t) = g(U(t),z(t)) + z(t)$$

with $z \in \mathcal{H}_c(U) \cap L^2(\mathbb{R}^3,\mathbb{R}^8)$. We obtain the system

$$\begin{cases} \dot{U} = -A(U,\eta)\langle N(U,\eta), dS(U)\rangle, \\ \partial_t z = JH(U)z + \mathbf{P}_c(U)JN(U,\eta) \\ \quad + \mathbf{P}_c(U(v))dS(U(v))A(U(v),\eta(v))\langle N(U(v),\eta(v)), dS(U(v))\rangle \\ \quad + (dP_c(U))A(U,\eta)\langle N(U,\eta), dS(U)\rangle\eta \end{cases}$$

with

$$\eta(t) = z(t) + g\left(U(t),z(t)\right).$$

We note that this equation is defined only *for* z *small with real values and* U *small*. We now study this system.

3.2. The stabilization towards the PLS manifold. We now show that any solution of (1.8) which belongs to the center manifold $W^c(V_0)$, for a small nonzero V_0, stabilizes as $t \to \pm\infty$ towards the manifold of the stationary states inside $W^c(V_0)$. To this end, we will use Theorems 1.1 and 1.2 to prove that z tends to zero in some sense.

Let us define for any ε, $\delta > 0$

$$\mathcal{U}(\varepsilon,\delta) = \left\{ U \in \mathcal{C}^1(\mathbb{R}, B_{\mathbb{C}^2}(V_0,\varepsilon)),\ \|\dot{U}\|_{L^1(\mathbb{R})\cap L^\infty(\mathbb{R})} \leq \delta^2 \right\},$$

and for any $U \in \mathcal{U}(\varepsilon,\delta)$, let s, β *be such that* $s > \beta + 2 > 2$ *and* $\sigma > 3/2$,

$$\mathcal{Z}(U,\delta) = \Big\{ z \in \mathcal{C}(\mathbb{R}, L^2(\mathbb{R}^3,\mathbb{R}^8)),\ z(t) \in \mathcal{H}_c(U(t)),$$

$$\max\left[\|z\|_{L^\infty(\mathbb{R},H^s)}, \|z\|_{L^2(\mathbb{R},H^s_{-\sigma})}, \|z\|_{L^2(\mathbb{R},B^\beta_{\infty,2})} \right] \leq \delta \Big\},$$

and ε, δ are small enough to ensure that for $U \in \mathcal{U}(\varepsilon,\delta)$ and $z \in \mathcal{Z}(U,\delta)$

$$S(U) + z + g(U,z) \in W^c(V_0) \cap B_{H^s}(S(V_0), r(V_0)),$$

where g is defined in Lemma 3.2 and r is defined in Remark 2.3. It will appear later that δ is of the same order as $\|z_0\|_{H^s}$ (see Lemma 3.8).

3.2.1. Some useful lemmas. In the rest of our study, we will need some technical lemmas, which we collect here.

LEMMA 3.3. *If Assumptions* 1.1–1.4 *hold, let* $\sigma, \sigma' \in \mathbb{R}$, $s > 1$, *and* $p, \widetilde{p_1}, p_1, p_2, q$ *in* $[1,\infty]$ *be such that*

$$\frac{1}{p} + \frac{s}{3} \geq \frac{1}{p_1} + \frac{1}{p_2} \geq \frac{1}{p}$$

and

$$\frac{1}{p} + \frac{s}{3} \geq \frac{1}{\widetilde{p_1}}.$$

Then there exist $\varepsilon > 0$ *and* $C > 0$ *such that, for all* $U \in B_{\mathbb{C}}(0,\varepsilon)$ *and* $\eta \in B^s_{p_2,q}(\mathbb{R}^3,\mathbb{R}^8) \cap L^\infty(\mathbb{R}^3,\mathbb{R}^8)$ *with* $\langle Q\rangle^\sigma \eta \in B^s_{p_1,q}(\mathbb{R}^3,\mathbb{R}^8)$ *and* $\langle Q\rangle^{\sigma'}\eta \in B^s_{\widetilde{p_1},q}(\mathbb{R}^3,\mathbb{R}^8)$, *we have*

$$\|\langle Q\rangle^\sigma N(U,\eta)\|_{B^s_{p,q}} \leq C\left(s,F,|U| + \|\eta\|_{L^\infty}\right)|U|\,\|\eta\|_{L^\infty}\left\|\langle Q\rangle^{\sigma'}\eta\right\|_{B^s_{\widetilde{p_1},q}} + C\left(s,F,|U| + \|\eta\|_{L^\infty\cap B^s_{p_2,q}}\right)\|\eta\|^2_{L^\infty}\,\|\langle Q\rangle^\sigma\eta\|_{B^s_{p_1,q}}. \tag{3.2}$$

Proof. We recall the definition

$$N(U,\eta) = \nabla F(S(U)+\eta) - \nabla F(S(U)) - d^2F(S(U))\eta.$$

We have

$$N(U,\eta) = \int_0^1\int_0^1 d^3F(S(U)+\theta'\theta\eta)\cdot\eta\cdot\theta\eta\,d\theta'd\theta$$

or

$$N(U,\eta) = \frac{1}{2}d^3F(S(U))\cdot\eta\cdot\eta + \int_0^1\int_0^1 d^4F(S(U)+\theta''\theta'\theta\eta)\cdot\theta'\theta\eta\cdot\eta\cdot\theta\eta\,d\theta''d\theta'd\theta.$$

Then we use, for $s \in \mathbb{R}^*_+$ and p, p_1, p_2, $q \in [1,\infty]$ such that $\frac{1}{p} + \frac{s}{3} \geq \frac{1}{p_1} + \frac{1}{p_2} \geq \frac{1}{p}$,

$$\|uv\|_{B^s_{p,q}} \leq C\|u\|_{B^s_{p_1,q}}\|v\|_{B^s_{p_2,q}},$$

and for $s > 1$, we use [19, Proposition 2.1]

$$\|d^k F(\psi)\|_{B^s_{p2,q}} \leq C\left(s, F, \|\psi\|_{L^\infty}\right) \|\psi\|_{B^s_{p2,q}}$$

for $k = 3$ or $k = 4$ and $d^4F(z) = O(|z|)$; otherwise we decompose $d^4F(z) = A + O(|z|)$, where A is a constant operator.

Eventually using Lemma 1.2 and

$$\left\| \langle Q\rangle^\sigma |\eta|^l \right\|_{B^s_{p_1,q}} \leq C \|\eta\|^{l-1}_{L^\infty} \|\langle Q\rangle^\sigma \eta\|_{B^s_{p_1,q}}$$

for $l \in \mathbb{N}$, we conclude the proof. □

LEMMA 3.4. *If Assumptions* 1.1–1.4 *hold, let* $\sigma \in \mathbb{R}$, $s > 1$, $p, p_1, p_2, q \in [1, \infty]$, *and* $\sigma_1, \sigma_2 \in \mathbb{R}$ *such that*

$$\frac{1}{p} + \frac{s}{3} \geq \frac{1}{p_1} + \frac{1}{p_2} \geq \frac{1}{p}.$$

Then there exist $\varepsilon > 0$ *and* $C > 0$ *such that, for all* $U \in B_{\mathbb{C}}(0, \varepsilon)$ *and* $\eta \in B^s_{p,q}(\mathbb{R}^3, \mathbb{R}^8) \cap L^\infty(\mathbb{R}^3, \mathbb{R}^8)$ *with* $\langle Q\rangle^{\sigma_1}\eta \in B^s_{p_1,q}(\mathbb{R}^3, \mathbb{R}^8)$ *and* $\langle Q\rangle^{\sigma_2}\eta \in B^s_{p_2,q}(\mathbb{R}^3, \mathbb{R}^8)$, *we have*

$$\begin{aligned}
&\|\langle Q\rangle^\sigma \left(\nabla F(S(U) + \eta) - \nabla F(S(U)) - \nabla F(\eta)\right)\|_{B^s_{p,q}} \\
&\leq C(s, F, |U| + \|\eta\|_{L^\infty}) \left(|U| + \|\langle Q\rangle^{\sigma_1}\eta\|_{B^s_{p_1,q}}\right) |U| \, \|\langle Q\rangle^{\sigma_2}\eta\|_{B^s_{p_2,q}}.
\end{aligned}$$

Proof. The proof is similar to that of Lemma 3.3. □

LEMMA 3.5. *If Assumptions* 1.1–1.4 *hold, let* $\sigma \in \mathbb{R}$, $s > 1$, *and* $p, q \in [1, \infty]$ *such that* $sp \geq 3$. *Then there exist* $\varepsilon > 0$ *and* $C > 0$ *such that, for all* U, $U' \in B_{\mathbb{C}^2}(0, \varepsilon)$ *and* η, $\eta' \in B^s_{p,q}(\mathbb{R}^3, \mathbb{R}^8)$, *we have*

$$\begin{aligned}
\|\langle Q\rangle^\sigma \{N(U, \eta) - N(U', \eta')\}\|_{B^s_{p,q}} &\leq C\left(s, F, |U| + |U'| + \|\eta\|_{B^s_{p,q}} + \|\eta'\|_{B^s_{p,q}}\right) \\
\times \Big\{ &\left(\|\langle Q\rangle^{\sigma_1}\eta\|_{B^s_{p,q}} + \|\langle Q\rangle^{\sigma_1}\eta'\|_{B^s_{p,q}}\right)^2 \left(|U - U'| + \|\langle Q\rangle^{\sigma_2}(\eta - \eta')\|_{B^s_{p,q}}\right) \\
&+ \left(|U| + |U'| + \left\|\langle Q\rangle^{\sigma_1'}\eta\right\|_{B^s_{p,q}} + \left\|\langle Q\rangle^{\sigma_1'}\eta'\right\|_{B^s_{p,q}}\right) \\
&\times \left(\|\langle Q\rangle^{\sigma_2'}\eta\|_{B^s_{p,q}} + \|\langle Q\rangle^{\sigma_2'}\eta'\|_{B^s_{p,q}}\right) \|\langle Q\rangle^{\sigma_3'}(\eta - \eta')\|_{B^s_{p,q}} \Big\},
\end{aligned}$$

with $2\sigma_1 + \sigma_2 = \sigma_1' + \sigma_2' + \sigma_3' = \sigma$ *if* $\langle Q\rangle^w \eta$, $\langle Q\rangle^w \eta' \in B^s_{p,q}(\mathbb{R}^3, \mathbb{R}^8)$ *for* $w \in \{\sigma_1, \sigma_2, \sigma_1', \sigma_2', \sigma_3'\}$.

Proof. Using the identity

$$N(u, \eta) = \int_0^1 \int_0^1 d^3F(S(u) + \theta'\theta\eta) \cdot \eta \cdot \theta\eta \, d\theta' d\theta,$$

we can restrict our study to $d^3F(\phi) - d^3F(\phi')$. If $F = O(|z|^5)$, we have

$$\|\langle Q\rangle^\sigma \left(d^3F(\phi) - d^3F(\phi')\right)\|_{B^s_{p,q}} \leq \int_0^1 \|d^4F(\phi + t(\phi - \phi'))\|_{B^s_{p,q}} \|\langle Q\rangle^\sigma(\phi - \phi')\|_{B^s_{p,q}} \, dt.$$

Then since $s > 1$ and $sp \geq 3$, we use

$$\|d^4F(\psi)\|_{B^s_{p,q}} \leq C(s, F, \|\psi\|_{B^s_{p,q}}).$$

Using Lemma 1.2, we conclude the proof when $F = O(|z|^5)$.

Otherwise, if F is a homogeneous polynomial of order 4, the proof is easily adaptable since d^4F is a constant tensor.

The case $F = O(|z|^4)$ follows by summing the two previous cases since, as a function of $u \in \mathbb{R}^8$, $F(u) = Au^{\otimes 4} + O(|u|^5)$. ∎

LEMMA 3.6. *If Assumptions* 1.1–1.4 *hold, let* $\sigma \in \mathbb{R}$, $s \in \mathbb{R}$, *and* $p, q \in [1, \infty]$. *Then there exist* $\varepsilon > 0$, $M > 0$, *and* $C > 0$ *such that, for all* $U, U' \in B_{\mathbb{C}^2}(0, \varepsilon)$ *and* $\eta, \eta' \in B_{L^2(\mathbb{R}^3, \mathbb{R}^8)}(0, M)$ *with* $\langle Q \rangle^\sigma \{\eta - \eta'\} \in B^s_{p,q}(\mathbb{R}^3, \mathbb{R}^8)$, *one has*

$$|A(U, \eta) - A(U', \eta')| \leq C \left\{ |U - U'| + \|\langle Q \rangle^\sigma \{\eta - \eta'\}\|_{B^s_{p,q}} \right\}. \tag{3.3}$$

Proof. We recall that

$$A(U, \eta) = [\langle JdS(U), dS(U) \rangle - \langle J\eta, d^2 S(U) \rangle]^{-1}.$$

We have

$$\begin{aligned} A(U, \eta) - A(u', \eta') &= -[\langle JdS(U), dS(U) \rangle - \langle J\eta, d^2 S(U) \rangle]^{-1} \\ &\times \left\{ \langle JdS(U), dS(U) \rangle - \langle J\eta, d^2 S(U) \rangle - \langle JdS(U'), dS(U') \rangle + \langle J\eta', d^2 S(U') \rangle \right\} \\ &\times [\langle JdS(U'), dS(U') \rangle - \langle J\eta', d^2 S(U') \rangle]^{-1}. \end{aligned}$$

The lemma then follows from Lemma 1.2. ∎

3.2.2. Global well-posedness for z. Let $U \in \mathcal{U}(\varepsilon, \delta)$ and $z_0 \in \mathcal{H}_c(U(0)) \cap H^s$. Let us write $U_\infty = \lim_{t \to +\infty} U(t)$; then we define $\mathcal{T}_{U, z_0}(z)$ by

$$\begin{aligned} \mathcal{T}_{U,z_0}(z)(t) &= e^{-\mathrm{i}tH + \mathrm{i}\int_0^t E(U(r))\, dr} z_0 \\ &+ \int_0^t e^{-\mathrm{i}(t-v)H + \mathrm{i}\int_v^t E(U(r))\, dr} \mathbf{P}_c(U(v)) J \nabla F(\eta(v))\, dv \\ &+ \int_0^t e^{-\mathrm{i}(t-v)H + \mathrm{i}\int_v^t E(U(r))\, dr} \\ &\quad \times \mathbf{P}_c(U(v)) J \{\nabla F(S(U(v)) + \eta(v)) - \nabla F(S(U(v)) - \nabla F(\eta(v))\}\, dv \\ &+ \int_0^t e^{-\mathrm{i}(t-v)H + \mathrm{i}\int_v^t E(U(r))\, dr} \\ &\quad \times \mathbf{P}_c(U(v)) dS(U(v)) A(U(v), \eta(v)) \langle N(U(v), \eta(v)), dS(U(v)) \rangle\, dv \\ &- \int_0^t e^{-\mathrm{i}(t-v)H + \mathrm{i}\int_v^t E(U(r))\, dr} d\mathbf{P}_c(U(v)) \dot{U}(v) \eta(v)\, dv \end{aligned}$$

with

$$\eta(t) = z(t) + g\left(U(t), z(t)\right).$$

First, we have a local well-posedness result for z with the following lemma.

LEMMA 3.7. *If Assumptions* 1.1–1.5 *hold, then there exist* $\varepsilon_0 > 0$ *and* $\delta_0 > 0$ *such that for any* $\varepsilon \in (0, \varepsilon_0)$, $\delta \in (0, \delta_0)$, $U \in \mathcal{U}(\delta, \varepsilon)$, *and* $z_0 \in B_{H^s(0,\delta)} \cap \mathcal{H}_c(U(0))$ *there are* $T^\pm(z_0, U) > 0$ *and a solution*

$$z \in \cap_{k=0}^2 \mathcal{C}^k((-T^-(z_0, U); +T^+(z_0, U)), H^{s-k}(0, \delta))$$

of the equation

$$\begin{cases} \partial_t z = JH(U)z + \mathbf{P}_c(U)JN(u,\eta) - (d\mathbf{P}_c(U))\dot{U}\eta, \\ z(0) = z_0, \end{cases} \tag{3.4}$$

where $\eta(t) = z(t) + g\,(U(t), z(t))$.

Moreover, z *is unique in* $L^\infty((-T', T), H^s)$ *for any* $T \in (0, T^+(z_0, U))$ *and* $T' \in (0, T^-(z_0, U))$, *and we have that if* $T^+(z_0, U) < +\infty$, *then*

$$\lim_{t \to T^+(z_0,U)} \|z(t)\|_{H^s} \geq \delta,$$

and if $T^-(z_0, U) = +\infty$, *then*

$$\lim_{t \to -T^-(z_0,U)} \|z(t)\|_{H^s} \geq \ \delta.$$

Proof. The proof is a consequence of the fixed point theorem applied to $\mathcal{T}_{U,z_0}$.

Using Lemmas 3.3, 3.5, and 3.6 with the estimate (2.7)–(2.9) and the properties of g given by Lemma 3.2, we obtain that $\mathcal{T}_{U,z_0}$ leaves a small ball in H^s invariant and is a contraction inside this ball.

Hence there exists a unique solution defined on the interval $[-T, T]$. Classical arguments permit us to extend the solution over a maximal interval $(-T^-(z_0, U), T^+(z_0, U))$ such that if $T^+(z_0, U) < \infty$, then necessarily the solution should leave a small ball in H^s at time $T^+(z_0, U)$. ∎

We now have a global well-posedness result as stated in the following lemma.

LEMMA 3.8. *If Assumptions* 1.1–1.5 *hold, there exist* $\varepsilon_0 > 0$, $\delta_0 > 0$, *and* $C > 0$ *such that for any* $\varepsilon \in (0, \varepsilon_0)$, $\delta \in (0, \delta_0)$, $U \in \mathcal{U}(\varepsilon, \delta)$, *and* $z_0 \in B_{H^s}(0, \delta) \cap \mathcal{H}_c(U(0))$ *we obtain, for the Cauchy problem* (3.4), $T^+(U, z_0) = +\infty$, $T^-(U, z_0) = +\infty$, $z \in \mathcal{Z}(U, \delta)$, *and*

$$\max\left[\|z\|_{L^\infty(\mathbb{R},H^s)}, \|z\|_{L^2(\mathbb{R},H^s_{-\sigma})}, \|z\|_{L^2(\mathbb{R},B^\beta_{\infty,2})}\right] \leq C\,\|z_0\|_{H^s}\,.$$

Proof. We have $(1 - P_c(U))z \equiv 0$ because $(1 - P_c(U(0)))z(0) = 0$ and its time derivative is zero.

Let us introduce for any $0 < T < T^+(U, z_0)$ the function

$$m(T) = \sup_{t \in (-T,T)} \left\{ \|z\|_{L^\infty((-T,T),H^s)}\,,\ \|z\|_{L^2((-T,T),H^s_{-\sigma})}\,,\ \|z\|_{L^2((-T,T),B^\beta_{\infty,2})} \right\}.$$

First, we study the estimation of $L^2((-T, T), H^s_{-\sigma})$. We use estimate (2.4) and the estimates of Theorem 1.1:

$$\begin{aligned} &\|z\|_{L^2((-T,T),H^s_{-\sigma})} \\ &\leq C_0\,\|z_0\|_{H^s} + C\left\|\mathbf{P}_c\int_0^t e^{-\mathrm{i}(t-v)H + \mathrm{i}\int_v^t E(U(r))\,dr}\mathbf{P}_c(U(v))J\nabla F(\eta(v))\,dv\right\|_{L^2((-T,T),H^s_{-\sigma})} \\ &\quad + C\,\|\nabla F(S(U)+\eta) - \nabla F(S(U) - \nabla F(\eta)\|_{L^2((-T,T),H^s_\sigma)} \\ &\quad + C\,\|dS(U)A(U,\eta)\langle N(U,\eta), dS(U)\rangle\|_{L^2((-T,T),H^s_\sigma)} \\ &\quad + C\left\|(d\mathbf{P}_c(U)\dot{U}\eta\right\|_{L^2((-T,T),H^s_\sigma)}. \end{aligned}$$

We now study the estimation of the third term of the right-hand side:

$$\left\| \int_0^t e^{-\mathrm{i}(t-v)H+\mathrm{i}\int_v^t E(U(r))\,dr} \mathbf{P}_c \mathbf{P}_c(U(v)) J\nabla F(\eta(v))\,dv \right\|_{L^2_t((-T,T),H^s_{-\sigma})}$$

$$\leq \int_{-T}^T \left\| e^{-\mathrm{i}(t-v)H+\mathrm{i}\int_v^t E(U(r))\,dr} \mathbf{P}_c \mathbf{P}_c(U(v)) J\nabla F(\eta(v)) \right\|_{L^2_t((-T,T),H^s_{-\sigma})} dv$$

$$\leq C(U) \left\| \nabla F(\eta) \right\|_{L^1((-T,T),H^s)}$$

$$\leq C(U) \left\| \eta \right\|^2_{L^2((-T,T),L^\infty)} \left\| \eta \right\|_{L^\infty((-T,T),H^s)},$$

where we used Theorem 1.1, estimate (1.2). Hence for the $L^2H^s_{-\sigma}$ estimate, we obtain

$$\begin{aligned} \left\| z \right\|_{L^2((-T,T),H^s_{-\sigma})} \leq{}& C_0 \left\| z_0 \right\|_{H^s} + C \left\| \eta \right\|^2_{L^2((-T,T),L^\infty)} \left\| \eta \right\|_{L^\infty((-T,T),H^s)} \\ &+ C \left(\left\| U \right\|_\infty + \left\| \eta \right\|_{L^\infty((-T,T),H^s_{-\sigma})} \right) \left\| U \right\|_\infty \left\| \eta \right\|_{L^2((-T,T),H^s_{-\sigma})} \\ &+ C \left\| \eta \right\|^2_{L^2((-T,T),L^\infty)} + C \left\| \dot{U} \right\|_{L^2} \left\| \eta \right\|_{L^\infty((-T,T),H^s)}. \end{aligned}$$

Using Lemma 3.2, we obtain

$$\left\| z \right\|_{L^2((-T,T),H^s_{-\sigma})} \leq C_0 \left\| z_0 \right\|_{H^s} + Cm(T)^3 + Cm(T)^2 + C\varepsilon m(T) + C\delta^2 m(T),$$

where C depends on $\left\| U \right\|_\infty$ and $\left\| \eta \right\|_{L^\infty((-T,T),H^s)}$.

Then, we estimate the H^s norm. Using estimate (2.4), we have

$$\begin{aligned} \left\| z(t) \right\|_{H^s} \leq{}& \left\| z_0 \right\|_{H^s} + \int_{-T}^T \left\| \nabla F(\eta(v)) \right\|_{H^s} dv \\ &+ \Bigg\| \int_0^t e^{-\mathrm{i}(t-v)H+\mathrm{i}\int_v^t E(U(r))\,dr} \mathbf{P}_c(U(v)) \\ &\quad \times J\{\nabla F(S(U(v))+\eta(v)) - \nabla F(S(U(v))) - \nabla F(\eta(v))\}\,dv \Bigg\|_{H^s} \\ &+ \int_{-T}^T \left\| dS(U(v)) A(U(v),\eta(v)) \langle N(U(v),\eta(v)), dS(U(v)) \rangle \right\|_{H^s} dv \\ &+ \int_{-T}^T \left\| d\mathbf{P}_c(U(v)) \dot{U}(v) \eta(v) \right\|_{H^s} dv. \end{aligned}$$

To estimate the third term of the right-hand side, we use the H-smoothness estimates,

more precisely, Theorem 1.1, estimate (1.1), and then we use Lemma 3.4:

$$\begin{aligned}
&\Big\| \int_0^t e^{-\mathrm{i}(t-v)H+\mathrm{i}\int_v^t E(U(r))\,dr} \\
&\qquad \times \mathbf{P}_c(U(v))J\{\nabla F(S(U(v))+\eta(v)) - \nabla F(S(U(v))) - \nabla F(\eta(v))\}\,dv \Big\|_{H^s} \\
&\le \Big\| \int_0^t e^{\mathrm{i}vH-\mathrm{i}\int_0^v E(U(r))\,dr} \\
&\qquad \times \mathbf{P}_c(U(v))J\{\nabla F(S(U(v))+\eta(v)) - \nabla F(S(U(v))) - \nabla F(\eta(v))\}\,dv \Big\|_{H^s} \\
&\le C\left\|\nabla F(S(U)+\eta) - \nabla F(S(U)) - \nabla F(\eta)\right\|_{L^2((-T,T),H^s_\sigma)} \\
&\le C\left(\|U\|_\infty + \|\eta(v)\|_{L^\infty((-T,T),H^s)}\right)\|U\|_\infty \|\eta\|_{L^2((-T,T),H^s_{-\sigma})}.
\end{aligned}$$

Hence for the $L^\infty H^s$ estimate, we obtain

$$\begin{aligned}
&\|z(t)\|_{H^s} \le \|z_0\|_{H^s} + C\|\eta\|_{L^\infty((-T,T),H^s)}\|\eta\|^2_{L^2((-T,T),L^\infty)} \\
&+ C\left(\|U\|_{L^\infty((-T,T))} + \|\eta(v)\|_{L^\infty((-T,T),H^s)}\right)\|U\|_{L^\infty((-T,T))}\|\eta\|_{L^2((-T,T),H^s_{-\sigma})} \\
&+ C\|\eta\|^2_{L^2((-T,T),L^\infty)} + \left\|\dot U\right\|_{L^1((-T,T))}\|\eta\|_{L^\infty((-T,T),H^s)}.
\end{aligned}$$

Using Lemma 3.2, we obtain

$$\|z(t)\|_{H^s} \le \|z_0\|_{H^s} + Cm(T)^3 + Cm(T)^2 + C\varepsilon m(T) + C\delta^2 m(T),$$

where C depends on $\|U\|_\infty$ and $\|\eta\|_{L^\infty((-T,T),H^s)}$.

For the $L^2B^\beta_{\infty,2}$ estimate, by Proposition 2.2 and Theorem 1.2, we have for any $\delta>0$, any $p_\delta > 3/\delta$, and $\theta_\delta = \frac{4}{p_\delta - 2}$

$$\begin{aligned}
\|z\|_{L^2((-T,T),B^\beta_{\infty,2})} &\le \|z\|_{L^2((-T,T),B^{\beta+\delta}_{p_\delta,2})} \\
&\le C_0\|z_0\|_{H^{\beta+1+\delta+\theta_\delta/2}} + C\left\|d^2F(S(U))\cdot\eta\right\|_{L^2((-T,T),B^{\beta+2+\delta+\theta_\delta}_{p'_\delta,2})} \\
&\quad + C\|N(U,\eta)\|_{L^1((-T,T),H^{\beta+1+\delta+\theta_\delta/2})} \\
&\quad + C\|dS(U)A(U,\eta)\langle N(U,\eta), dS(U)\rangle\|_{L^1((-T,T),H^{\beta+1+\delta+\theta_\delta/2})} \\
&\quad + C\left\|(d\mathbf{P}_c(U))\dot U\eta\right\|_{L^1((-T,T),H^{\beta+1+\delta+\theta_\delta/2})}\,dv.
\end{aligned}$$

With Lemmas 3.3 and 3.4, we infer that

$$\begin{aligned}
&\|z\|_{L^2(\mathbb{R},B^\beta_{\infty,2})} \le C_0\|z_0\|_{H^{\beta+1+\delta+\theta_\delta/2}} + C|U|_\infty\|\eta\|_{L^2((-T,T),H^{\beta+2+\delta+\theta_\delta}_{-\sigma})} \\
&+ C|U|_\infty\|z\|_{L^2((-T,T),L^\infty)}\|z\|_{L^2((-T,T),H^{\beta+1+\delta+\theta_\delta/2})} \\
&+ C(|U|_\infty + \|\eta\|_{L^\infty((-T,T),H^{\beta+1+\delta+\theta_\delta/2})})\|\eta\|^2_{L^2((-T,T),L^\infty)}\|z\|_{L^\infty((-T,T),H^{\beta+1+\delta+\theta_\delta/2})}
\end{aligned}$$

$$+ C(|U|_\infty + \|\eta\|_{L^\infty((-T,T),H^{\beta+1+\delta+\theta_\delta/2})}) \|\eta\|_{L^2((-T,T),H^{\beta+1+\delta+\theta_\delta/2}_{-\sigma})}$$

$$\times \|\eta\|_{L^\infty((-T,T),H^{\beta+1+\delta+\theta_\delta/2})}$$

$$+ C\|\dot{U}\|_{L^1} \|\eta\|_{L^\infty((-T,T),H^{\beta+1+\delta+\theta_\delta/2})}.$$

Since for small $\delta > 0$, $s \geq \beta + 2 + \delta + \theta_\delta$ and using Lemma 3.2, we infer that

$$\|z\|_{L^2((-T,T),B^\beta_{\infty,2})} \leq C_0\|z_0\|_{H^{\beta+1+\delta+\theta_\delta/2}} + Cm(T)^3 + Cm(T)^2 + C\varepsilon m(T) + C\delta^2 m(T).$$

Hence we obtain

$$m(T) \leq C_0\|z_0\|_{H^{\beta+1+\delta+\theta_\delta/2}} + C\varepsilon m(T) + C\delta^2 m(T) + Cm(T)^3 + Cm(T)^2,$$

where C_0 does not depend on m and C is a nondecreasing function of $\|z\|_{L^\infty((-T,T),H^s)}$ and $\|U\|_\infty$ and hence can be bounded by a nondecreasing function of m.

If $\|z_0\|_{H^s}$ is small, then $m(0)$ is small and $m(T)$ stays small. Therefore we have that $z \in \mathcal{Z}(U,\delta)$ if $\|z_0\|_{H^s}, \delta$, and ε are small enough; moreover,

$$\max\left[\|z\|_{L^\infty(\mathbb{R},H^s)}, \|z\|_{L^2(\mathbb{R},H^s_{-\sigma})}, \|z\|_{L^2(\mathbb{R},B^\beta_{\infty,2})}\right] \leq f(\|z_0\|_{H^s}),$$

where f is such that there exists $C > 0$ with

$$f(\|z_0\|_{H^s}) \leq C\|z_0\|_{H^s}. \qquad \square$$

The solution z just found is a function of z_0 and U; writing it $z[z_0, U]$, we have the following important property given by the following lemma.

LEMMA 3.9. *If Assumptions* 1.1–1.5 *hold, then for any $T > 0$, there exist $\varepsilon_0 > 0$, $\delta_0 > 0$, $C > 0$, and $\kappa \in (0,1)$ such that for any $\varepsilon \in (0,\varepsilon_0)$, $\delta \in (0,\delta_0)$, U, $U' \in \mathcal{U}(\varepsilon,\delta)$, $z_0 \in \mathcal{H}_c(U(0))$, $z_0' \in \mathcal{H}_c(U'(0))$, $z \in \mathcal{Z}(U,\delta)$, and $z' \in \mathcal{Z}(U',\delta)$, one has*

$$\|z[z_0',U'] - z[z_0,U]\|_{L^\infty((-T;T),H^s)\cap L^2((-T;T),L^\infty)\cap L^2((-T;T),H^s_{-\sigma})} \leq C\|z_0 - z_0'\|_{H^s} + \kappa\left\{\|U - U'\|_{L^\infty((-T;T))} + \left\|\dot{U} - \dot{U}'\right\|_{L^\infty((-T;T))}\right\}.$$

Proof. We use the techniques of the previous lemma. $\square$

3.2.3. Global well-posedness for U and its stabilization. Here we want to solve the equation for U. We note that z has been built in the previous section and is a function of U and $z_0 \in \mathcal{H}_c(U(0))$. Let us introduce for any $U_0 \in B_{\mathbb{C}}(0,\varepsilon)$ the function on $\mathcal{U}(\varepsilon,\delta)$:

$$f_{U_0}(U)(t) = U_0 - \int_0^t A(U(v),\eta(v))\langle N(U(v),\eta(v)), dS(U(v))\rangle\, dv,$$

where $\eta = z(t) + g[U(t),z(t)]$. We have the following lemma.

LEMMA 3.10. *If Assumptions* 1.1–1.5 *hold, there exist $\varepsilon_0 > 0$ and $\delta_0 > 0$ such that, for any $\varepsilon \in (0,\varepsilon_0)$, $\delta \in (0,\delta_0)$, the function f_{U_0} maps $\mathcal{U}(\varepsilon,\delta)$ into itself if U_0 and $z_0 \in H^s \cap \mathcal{H}_c(U_0)$ are small enough.*

Proof. By means of Lemma 3.3, we obtain

$$\|\partial_t f_{U_0}(U)\|_{L^1(\mathbb{R})\cap L^\infty(\mathbb{R})} \leq C \|N(U(v),\eta(v))\|_{L^1(\mathbb{R},H^s_{-\sigma})\cap L^\infty(\mathbb{R},H^s)} \leq \delta^2$$

and

$$\|f_{U_0}(U)\|_{L^\infty(\mathbb{R})} \leq |U_0| + C \|N(U(v),\eta(v))\|_{L^1(\mathbb{R},H^s)} \leq |U_0| + \delta^2;$$

hence for sufficiently small U_0 and δ, we obtain the lemma. ∎

The function f_{U_0} also has a local Lipschitz property, as stated by the following lemma.

LEMMA 3.11. *If Assumptions* 1.1–1.5 *hold, for any* $T > 0$, *there exist* $\varepsilon_0 > 0$, $\delta_0 > 0$, *and* $\kappa \in (0,1)$ *such that for any* $\varepsilon \in (0,\varepsilon_0)$, $\delta \in (0,\delta_0)$, U, $U' \in \mathcal{U}(\varepsilon,\delta)$, *for any* $z_0 \in \mathcal{H}_c(U(0)) \cap H^s$, *for any* $z_0' \in \mathcal{H}_c(U'(0)) \cap H^s$ *small enough, and for* U_0, U_0' *small enough, one has*

$$\begin{aligned}&\left|f_{U_0}(U) - f_{U_0'}(U')\right|_{L^\infty((-T;T))} + \left|\partial_t f_{U_0}(U) - \partial_t f_{U_0'}(U')\right|_{L^1((-T;T))}\\ &\leq |U_0 - U_0'| + \kappa\left(\|U - U'\|_{L^\infty((-T;T))} + \|\dot U - \dot U'\|_{L^1((-T;T))} + \|z_0 - z_0'\|_{H^s}\right).\end{aligned}$$

Proof. The proof is a straightforward consequence of Lemmas 3.5, 3.6, and 3.9. ∎

We now obtain the following lemma.

LEMMA 3.12. *If Assumptions* 1.1–1.5 *hold, there exist* $\varepsilon > 0$ *and* $\delta > 0$ *such that, for any* $U_0 \in \mathbb{C}$ *small and* $z_0 \in \mathcal{H}_c(U_0) \cap H^s_\sigma$ *small, the equation*

$$\left\{\begin{array}{lcl}\dot U &=& -A(U,\eta)\langle N(U,\eta), dS(U)\rangle,\\ U(0) &=& U_0,\end{array}\right. \tag{3.5}$$

where $\eta(t) = z(t) + g\,[U(t),z(t)]$, *has a unique solution in* $\mathcal{U}(\delta,\varepsilon)$. *Moreover, there exists* $C > 0$ *such that*

$$|U_{\pm\infty} - U_0| \leq C \|z_0\|^2_{H^s}.$$

Proof. This is also a fixed point result for f_{U_0}. Let us fix $T > 0$ and consider, for any $V \in \mathcal{U}(\delta,\varepsilon)$ with sufficiently small $\delta > 0$ and $\varepsilon > 0$, the sequence

$$\left\{\begin{array}{l}V_{n+1} = f_{U_0}(V_n) \quad \forall n \in \mathbb{N},\\ V_0 = V\end{array}\right.$$

for any $n \in \mathbb{N}$, $V_n \in \mathcal{U}(\delta,\varepsilon)$. With Lemma 3.11, the fixed point theorem gives us the convergence for the norms of $L^\infty((-T,T))$ and $\dot W^{1,1}((-T,T))$ of $(V_n)_{n\in\mathbb{N}}$.

Then we notice that, for any $T' \in \mathbb{R}$, we have

$$V_{n+1}(t) = f_{f_{U_0}(V_n)(T')}(V_n)(t - T').$$

Since for $T' \in (-T;T)$, $(f_{U_0}(V_n)(T'))$ is a Cauchy sequence, Lemma 3.11 gives us the convergence of (V_n) for the norms of $L^\infty((T'-T;T'+T))$ and $\dot W^{1,1}((T'-T;T'+T))$.

Iterating this process, we obtain that the sequence converges uniformly locally in time and we prove the lemma since the other statements are classical. We note just that the last statement follows from the fact that there exists $C > 0$ such that

$$\int_{\mathbb{R}^\pm} \left|\dot U(v)\right| \, dv \leq \int_{\mathbb{R}^\pm} |A(U(v),\eta(v))\langle N(U(v),\eta(v)), dS(U(v))\rangle| \, dv \leq C \|z_0\|^2_{H^s}. \qquad ∎$$

3.2.4. The asymptotic profile of z. In this section, our aim is to specify the asymptotic profile of z when z_0 is localized. First we state the following proposition.

PROPOSITION 3.1. *There exists $\varepsilon > 0$ such that for all $U \in B_{\mathbb{C}^2}(0,\varepsilon)$ and $\alpha \in \mathbb{R}^+$ there exists $C > 0$ such that*

$$\left\| \langle Q \rangle^{\alpha} e^{JtH(U)} P_c(U) \psi \right\| \leq C_\alpha \sum_{\beta=0}^{\alpha} \langle t \rangle^{\beta} \left\| \langle Q \rangle^{\alpha-\beta} \psi \right\|$$

for any $\psi \in L^2(\mathbb{R}^3, \mathbb{C}^8)$.

Proof. From Proposition 2.2, we obtain the lemma for $\alpha = 0$; then we just need the estimate

$$\left\| Q^{\alpha} e^{JtH(U)} \psi \right\|^2 \leq C_\alpha^2 \sum_{0 \leq \beta \leq \alpha} |t|^{2|\beta|} \left\| Q^{\alpha-\beta} \psi \right\|^2$$

for any $\psi \in L^2(\mathbb{R}^3, \mathbb{C}^8)$, $\alpha \in \mathbb{N}^3$, and some $C > 0$ independent of ψ. The rest of the proposition will follow by interpolation.

For $U = 0$, this follows by an iterated proof from the identity

$$\frac{d}{dt} e^{\mathrm{i}tH} Q e^{-\mathrm{i}tH} = e^{\mathrm{i}tH} \alpha e^{-\mathrm{i}tH},$$

where α is the 3-vector of Dirac–Pauli matrices defined in the introduction. For $U \neq 0$, we use the same proof with the exponential decay of Proposition 1.1. ∎

We can improve Lemma 3.13 if we use [7, Theorem 1.1] and [7, Theorem 1.2], which we repeat in what follows.

THEOREM 3.1 (Theorem 1.1 of [7]: propagation for perturbed Dirac dynamics). *Assume that Assumptions* 1.1 *and* 1.2 *hold, and let $\sigma > \frac{5}{2}$. Then one has*

$$\| e^{-\mathrm{i}tH} \mathbf{P}_c(H) \|_{B(L^2_\sigma, L^2_{-\sigma})} \leq C \langle t \rangle^{-\frac{3}{2}}.$$

We also have the following proposition.

PROPOSITION 3.2 (Proposition 2.2 of [7]: propagation far from thresholds). *Suppose that Assumption* 1.1 *holds. Then, for any $\chi \in \mathcal{C}^\infty(\mathbb{R}^3, \mathbb{C}^4)$ bounded with support in $\mathbb{R} \setminus (-m; m)$ and for any $\sigma \geq 0$, there is $C > 0$ such that*

$$\| e^{-\mathrm{i}tH} \chi(H) \|_{B(L^2_\sigma, L^2_{-\sigma})} \leq C \langle t \rangle^{-\sigma}.$$

Using Duhamel's formula as in Proposition 2.2 and interpolating with estimate (2.6), we obtain the following corollary.

COROLLARY 3.1. *Assume that Assumptions* 1.1 *and* 1.2 *hold, and let $\theta \geq 0$ and $\sigma > \frac{5}{2}\theta$. Then there exists $\varepsilon > 0$ such that for all $U \in B_{\mathbb{C}^2}(0,\varepsilon)$ one has*

$$\| e^{JtH(U)} \mathbf{P}_c(U) \|_{B(L^2_\sigma, L^2_{-\sigma})} \leq C \langle t \rangle^{-\frac{3\theta}{2}}.$$

THEOREM 3.2 (Theorem 1.2 of [7]: dispersion for perturbed Dirac dynamics). *Assume that Assumptions* 1.1 *and* 1.2 *hold. Then for $p \in [1,2]$, $\theta \in [0,1]$, $s - s' \geq (2+\theta)(\frac{2}{p} - 1)$, and $q \in [1,\infty]$ there exists a constant $C > 0$ such that*

$$\| e^{-\mathrm{i}tH} \mathbf{P}_c(H) \|_{B^s_{p,q}, B^{s'}_{p',q}} \leq C (K(t))^{\frac{2}{p}-1}$$

with $\frac{1}{p}+\frac{1}{p'}=1$, *and*

$$K(t)=\begin{cases} |t|^{-1+\theta/2} & \text{if } |t|\in(0,1],\\ |t|^{-1-\theta/2} & \text{if } |t|\in[1,\infty). \end{cases}$$

Using Duhamel's formula once more, Theorem 3.1, and Corollary 3.1, we obtain the following corollary.

COROLLARY 3.2. *Assume that Assumptions* 1.1 *and* 1.2 *hold, and let* $p\in[1,2]$, $\theta\in[0,1]$, $s-s'\geq(2+\theta)(\frac{2}{p}-1)$, $q\in[1,\infty]$, *and* $\sigma>\max\{\frac{3}{2},(\frac{2}{p}-1)(1+\frac{\theta}{2})\}$. *Then there exists* $\varepsilon>0$ *such that for all* $U\in B_{\mathbb{C}^2}(0,\varepsilon)$ *one has*

$$\|e^{JtH(U)}\mathbf{P}_c(U)\|_{H^s_\sigma,B^{s'}_{p',q}}\leq C\,(K(t))^{\frac{2}{p}-1}$$

with $\frac{1}{p}+\frac{1}{p'}=1$, *and*

$$K(t)=\begin{cases} |t|^{-1+\theta/2} & \text{if } |t|\in(0,1],\\ |t|^{-1-\theta/2} & \text{if } |t|\in[1,\infty). \end{cases}$$

Proof. We first prove this for $U=0$. We have to study the high and low energy parts in a different manner. For the low energy part, we iterate twice Duhamel's formula with respect to D_m in order to use Theorems 3.1 and 3.2 for the free case.

In the high energy part, we use also Duhamel's formula but with Theorem 3.2 for the free case and Proposition 3.2.

Then for $U\neq 0$, we work as for estimate (2.6). ∎

We obtain the following lemma.

LEMMA 3.13. *If Assumptions* 1.1–1.5 *hold, there exist* $\varepsilon_0>0$, $\delta_0>0$, *and* $C>0$ *such that for any* $\varepsilon\in(0,\varepsilon_0)$, $\delta\in(0,\delta_0)$, $U_0\in B_{\mathbb{C}^2}(0,\varepsilon)$, *and* $z_0\in B_{H^s_\sigma}(0,\delta)\cap\mathcal{H}_c(U_0)$ *we obtain for the Cauchy problem* (3.4) *(with* U *the solution of* (3.5)) *a global solution* z *such that*

$$\max\Big[\sup_{t\in\mathbb{R}}(\|z(t)\|_{H^s}),\ \sup_{t\in\mathbb{R}}(\langle t\rangle^{3/2}\|z(t)\|_{H^s_{-\sigma}}),$$

$$\sup_{t\in\mathbb{R}}(\langle t\rangle^{3/2}\|z(t)\|_{B^\beta_{\infty,2}}),\ \sup_{t\in\mathbb{R}}(\langle t\rangle^{-3/2}\|z(t)\|_{H^s_{3/2}})\Big]\leq C\,\|z_0\|_{H^s_\sigma}.$$

Proof. The proof is similar to that of Lemma 3.8 with some adaptations involving the norm H^s_σ; we also refer to the proof of [7, Lemma 5.5]. Indeed, we work in the spaces

$$\widetilde{\mathcal{U}}(\varepsilon,\delta)=\Big\{U\in\mathcal{C}^1(\mathbb{R},B_{\mathbb{C}^2}(V_0,\varepsilon)),\ \|\dot V(t)\|\leq\frac{\delta^2}{\langle t\rangle^3}\Big\}$$

and

$$\mathcal{Z}(U,\delta)=\Big\{z\in\mathcal{C}(\mathbb{R},L^2(\mathbb{R}^3,\mathbb{R}^8)),\ z(t)\in\mathcal{H}_c(U(t)),$$

$$\max\Big[\sup_{t\in\mathbb{R}}(\|z(t)\|_{H^s}),\ \sup_{t\in\mathbb{R}}(\langle t\rangle^{3/2}\|z(t)\|_{H^s_{-\sigma}}),$$

$$\sup_{t\in\mathbb{R}}(\langle t\rangle^{3/2}\|z(t)\|_{B^\beta_{\infty,2}}),\ \sup_{t\in\mathbb{R}}(\langle t\rangle^{-3/2}\|z(t)\|_{H^s_{3/2}})\Big]\leq\delta\Big\}.$$

Let

$$t \mapsto \xi_\pm(t) = e^{J\int_0^t (E(U(v))-E(U_{\pm\infty}))\,dv} z(t)$$

and

$$t \mapsto V_\pm(t) = e^{-\mathrm{i}\int_0^t (E(U(v))-E(U_{\pm\infty}))\,dv} U(t).$$

We have that $V_{\pm\infty} \lim_{\pm\infty} V_\pm(t)$ exist and we use exactly the same method as that of Lemma 3.8, applied to

$$\begin{aligned}
\xi_\pm(t) &= e^{JtH(V_{\pm\infty})} z_0 \\
&+ \int_0^t e^{J(t-s)H(V_{\pm\infty})} \mathbf{P}_c(V_\pm(v)) J \left(d^2F(S(V_\pm(v))) - d^2F(S(V_{\pm\infty})) \right) \xi_\pm(v)\,dv \\
&+ \int_0^t e^{J(t-s)H(V_{\pm\infty})} \mathbf{P}_c(V_\pm(v)) J N(V_\pm(v), \tilde\eta_\pm(v))\,dv \\
&+ \int_0^t e^{J(t-s)H(V_{\pm\infty})} \mathbf{P}_c(V_\pm(v)) dS(V(v)) A(V_\pm(v), \tilde\eta_\pm(v)) \\
&\qquad \times \langle N(V_\pm(v), \tilde\eta_\pm(v)), dS(V_\pm(v)) \rangle\,dv \\
&- \int_0^t e^{J(t-s)H(V_{\pm\infty})} (d\mathbf{P}_c(V_\pm(v))) A(V_\pm(v), \tilde\eta_\pm(v)) \\
&\qquad \times \langle N(V_\pm(v), \tilde\eta_\pm(v)), dS(V_\pm(v)) \rangle \tilde\eta_\pm(v)\,dv,
\end{aligned}$$

with $\tilde\eta_\pm(t) = e^{J\int_0^t (E(U(v))-E(U_{\pm\infty}))\,dv} (z(t) + g(U(t), z(t)))$, but using the previous time decay estimates.

There are two differences from the proof of Lemma 3.8:

One is in the estimate of the $H^s_{-\sigma}$ norm. In fact, before using the time decay estimates for $e^{-\mathrm{i}tH} P_c(H)$, we split the space associated with the continuous spectrum into two parts: one associated with energy close to the thresholds and one associated with the rest of the spectrum. In the first part, we use the fact that $\sigma > 3/2$ to estimate the $H^s_{-\sigma}$ by the $B^\beta_{\infty,2}$ norm since we work with bounded energies. In the second part, since we work far from thresholds, we use Proposition 3.2 after estimating the $H^s_{-\sigma}$ by the $H^s_{-3/2}$ norm.

The other difference is in the estimation of the $B^\beta_{\infty,2}$ norm. We use Corollary 3.2 for $e^{JtH(V_{\pm\infty})} z_0$ and Theorem 3.2 for the integrals. □

We have that $\lim_{\pm\infty} U = U_{\pm\infty}$ exist. If $z_0 \in H^s_\sigma$, then the associated solution U satisfies

$$\left|\dot U\right| \le \frac{C}{\langle t\rangle^3} \left\| z_0 \right\|_{H^s_\sigma},$$

and we have

$$\int_0^t \left(E(U(v)) - E(U_{\pm\infty})\right)\,dv \to E_{\pm\infty} \quad \text{as } t \to \pm\infty$$

for some real $E_{\pm\infty}$. We introduce

$$V_\pm(t) = e^{-\mathrm{i}\int_0^t (E(U(v))-E(U_{\pm\infty}))\,dv} U(t),$$

which have a limit as $t \to \pm\infty$, respectively, as being

$$V_{\pm\infty} = e^{-\mathrm{i}E_{\pm\infty}} U_{\pm\infty}.$$

Then we note that we can also obtain an asymptotic profile for $e^{\mathrm{i}tH+\mathrm{i}tE(U_\infty)}z(t)$ if z_0 is localized. But we prefer to obtain a scattering result with respect to $e^{JtH(V_\infty))}$. We have the following lemma.

LEMMA 3.14. *If Assumptions* 1.1–1.5 *hold, then there exist* $\varepsilon_0 > 0$, $\delta_0 > 0$ *such that for any* $\varepsilon \in (0, \varepsilon_0)$, $\delta \in (0, \delta_0)$, $U_0 \in B_{\mathbb{C}^2}(0, \varepsilon)$, *and* $z_0 \in B_{H^s_\sigma}(0, \delta) \cap \mathcal{H}_c(U_0)$ *and for the solution* z *of* (3.4) *(with* U *the solution of* (3.5)*) given in Lemma* 3.7, *the limit*

$$z_{\pm\infty} = \lim_{t\to\pm\infty} e^{-JtH(V_{\pm\infty})} e^{J\int_0^t (E(U(v))-E(U_{\pm\infty}))\,dv} z(t)$$

exists in H^s. *Moreover, we have* $z_{\pm\infty} \in \mathcal{H}_c(V_{\pm\infty}) \cap H^s_\sigma$, *and there exists* $C > 0$ *such that*

$$\max\Big\{ \|e^{-J\int_0^t (E(U(v))-E(U_{\pm\infty}))\,dv} e^{JtH(V_{\pm\infty})\,dr} z_{\pm\infty} - z(t)\|_{H^s},$$
$$\|e^{-J\int_0^t (E(U(v))-E(U_{\pm\infty}))\,dv} e^{JtH(V_{\pm\infty})\,dr} z_{\pm\infty} - z(t)\|_{H^s_{-\sigma}},$$
$$\|e^{-J\int_0^t (E(U(v))-E(U_{\pm\infty}))\,dv} e^{JtH(V_{\pm\infty})\,dr} z_{\pm\infty} - z(t)\|_{B^\beta_{\infty,2}} \Big\} \le \tfrac{C}{\langle t\rangle^2} \|z_0\|^2_{H^s_\sigma}$$

and

$$\|z_{\pm\infty} - e^{-JtH(V_{\pm\infty})} e^{J\int_0^t (E(U(v))-E(U_{\pm\infty}))\,dv} z(t)\|_{H^s_{3/2}} \le \frac{C}{\langle t\rangle^{\frac12}} \|z_0\|^2_{H^s_\sigma}.$$

Proof. Let

$$t \mapsto \xi_\pm(t) = e^{J\int_0^t (E(U(v))-E(U_{\pm\infty}))\,dv} z(t)$$

and

$$t \mapsto V_\pm(t) = e^{-\mathrm{i}\int_0^t (E(U(v))-E(U_{\pm\infty}))\,dv} U(t).$$

Using exactly the same method as that of Lemma 3.8, applied to

$$\begin{aligned}
e^{-JtH(V_{\pm\infty})}\xi_\pm(t) &= z_0 \\
&+ \int_0^t e^{\quad -JsH(V_{\pm\infty})} \mathbf{P}_c(V_\pm(v)) J \left(d^2F(S(V_\pm(v))) - d^2F(S(V_{\pm\infty})) \right) \xi_\pm(v)\,dv \\
&+ \int_0^t e^{\quad -JsH(V_{\pm\infty})} \mathbf{P}_c(V_\pm(v)) J N(V_\pm(v), \tilde\eta_\pm(v))\,dv \\
&+ \int_0^t e^{-JsH(V_{\pm\infty})} \mathbf{P}_c(V_\pm(v)) dS(V(v)) A(V_\pm(v), \tilde\eta_\pm(v)) \\
&\qquad \times \langle N(V_\pm(v), \tilde\eta_\pm(v)), dS(V_\pm(v))\rangle\,dv \\
&- \int_0^t e^{-JsH(V_{\pm\infty})} (d\mathbf{P}_c(V_\pm(v))) A(V_\pm(v), \tilde\eta_\pm(v)) \\
&\qquad \times \langle N(V_\pm(v), \tilde\eta_\pm(v)), dS(V_\pm(v))\rangle \tilde\eta_\pm(v)\,dv,
\end{aligned}$$

with $\tilde{\eta}_\pm(t) = e^{J\int_0^t (E(U(v))-E(U_{\pm\infty}))\,dv}\,(z(t)+g(U(t),z(t)))$, we prove that the limits

$$\lim_{t\to\pm\infty} e^{-JtH(V_{\pm\infty})}\xi_\pm(t) = z_{\pm\infty}$$

exist. If we use the method of Lemma 3.13, we obtain the estimates on the convergence of $e^{JtH(V_{\pm\infty})}z_{\pm\infty} - \xi_\pm(t)$. Then multiplying by $e^{-J\int_0^t (E(U(v))-E(U_{\pm\infty}))\,dv}$, we obtain the estimates and the convergence of

$$e^{-J\int_0^t (E(U(v))-E(U_{\pm\infty}))\,dv} e^{-JtH(V_{\pm\infty})}z_{\pm\infty} - z(t).$$

Then since $(1-P_c(U(t)))\,z(t) = 0$, we have $(1-P_c(V_{\pm\infty}))\,z_{\pm\infty} = 0$, and hence $z_{\pm\infty}$ belongs to $\mathcal{H}_c(V_{\pm\infty})$. □

4. The dynamics outside the center manifold. We can make the same study in the center-stable manifold and the center-unstable manifold but only in one direction of time. Let us explain it for the center-stable manifold in positive time since it is similar for the center-unstable manifold. Actually it is equivalent if we revert the time direction.

We give just a sketch of the proof since it is similar to the previous study. Using the idea of the proof of exponential stabilization for Proposition 2.4, we write any solution ψ in the form $\phi+\rho+f(\phi,\rho)$ with ϕ in the center manifold, $\rho\in X_s(V_0)$, and f a function to be specified and which ensures that ϕ is in the center stable manifold.

Indeed, $W^c(V_0)$ is the graph of a smooth function $h^c : X_c(V_0)\mapsto X_s(V_0)\oplus X_u(V_0)$, and $W^{cs}(V_0)$ is the graph of a smooth function $h^u : X_c(V_0)\oplus X_s(V_0)\mapsto X_u(V_0)$. Let ν be such that $\psi = S(V_0)+\nu$ satisfies (1.8); then we have

$$\begin{aligned}
\partial_t\nu &= JH(V_0)\nu + JN(V_0,\nu),\\
\nu &= y + h_c(y) + \rho + h^u(y, h^c(y)+\rho)\\
&= \phi(y) - S(V_0) + (\rho - \pi^s(V_0)h^c(y)) + (h^u(y,h^c(y)+\rho) - \pi^u(V_0)h^c(y))\\
&= \phi(y) - S(V_0) + \rho + f(y,\rho)
\end{aligned}$$

with $y = \pi^c(V_0)\nu = \pi^c(V_0)\,(\psi - S(V_0))$, $\phi(y) = S(V_0) + y + h^c(y)$ in the center manifold, and $\rho\in X_s(V_0)$. We have the following equation for ρ:

$$\partial_t\rho = JH(V_0)\rho + M(V_0,y,\rho), \tag{4.1}$$

where

$$\begin{aligned}
M(V_0,y,\rho) &= \pi^s(V_0)\left\{JN(V_0,y+h^c(y)+\rho+f(y,\rho)) - JN(V_0,y+h^c(y))\right\}\\
&- \pi^s(V_0)Dh^c(y)\pi^c(V_0)\left\{JN(V_0,y+h^c(y)+\rho+f(y,\rho)) - JN(V_0,y+h^c(y))\right\}.
\end{aligned}$$

Then we obtain for ϕ the equation

$$\begin{aligned}
\partial_t\phi &= JH\phi + J\nabla F(\phi) + R(\phi,\rho),\\
R(\phi,\rho) &= J\nabla F(\phi+\rho+f(y,\rho)) - J\nabla F(\phi)\\
&\quad - Jd^2F(S(V_0))\rho - M(V_0,\pi^c(V_0)(\phi - S(V_0)),\rho),
\end{aligned}$$

noting that $|R(\phi,\rho)|\le C(\|\phi\|_{H^s},\|\rho\|_{L^\infty})|\rho|$.

Working as in section 3, we write $\phi = S(U) + \eta$ with $\eta = z + g(U, z)$ and we have the following equations for U and z:

$$\begin{cases} \dot{U} = -A(U,\eta)\langle N(U,\eta) - JR(U,\eta,\rho), dS(U)\rangle, \\ \partial_t z = JH(U)z + \mathbf{P}_c(U)JN(U,\eta) \\ + \mathbf{P}_c(U(v))dS(U(v))A(U(v),\eta(v))\langle N(U(v),\eta(v)) - JR(U,\eta,\rho), dS(U(v))\rangle \\ + (dP_c(U))A(U,\eta)\langle N(U,\eta) - JR(U,\eta,\rho), dS(U)\rangle\eta + \mathbf{P}_c(U)R(U,\eta,\rho) \end{cases}$$

with

$$\eta(t) = z(t) + g\left(U(t), z(t)\right),$$

where g is defined by Lemma 3.2 and

$$R(U,\eta,\rho) = R(S(U) + \eta, \rho).$$

These equations are similar to those we have studied but with an extra term coming from R which is exponentially decaying in positive time. Indeed, let us introduce for any $T_0 < 0$, $\gamma \in (0, \gamma(V_0))$, and $\delta > 0$ the set

$$\mathcal{R}_{T_0,\gamma}(\delta) = \left\{\rho \in \mathcal{C}((T_0, +\infty), X_s(V_0)),\ |\rho(t)|_{H^s} \leq \delta e^{-\gamma t}\ \forall t > T_0\right\}.$$

We study (4.1) in $\mathcal{R}_{T_0,\gamma}(\delta)$ with small initial condition ρ_0. We also define for any $\varepsilon > 0$

$$\mathcal{U}_{T_0}(\varepsilon,\delta) = \left\{U \in \mathcal{C}^1((T_0,+\infty), B_{\mathbb{C}^2}(V_0,\varepsilon)),\ \|\dot{U}\|_{L^1((T_0,+\infty))\cap L^\infty((T_0,+\infty))} \leq \delta^2\right\},$$

and for any $U \in \mathcal{U}_{T_0}(\varepsilon)$, let s, β be such that $s > \beta + 2 > 2$ and $\sigma > 3/2$,

$$\mathcal{Z}_{T_0}(U,\delta) = \big\{z \in \mathcal{C}((T_0,+\infty), L^2(\mathbb{R}^3,\mathbb{R}^8)),\ z(t) \in \mathcal{H}_c(U(t)),$$
$$\max\left[\|z\|_{L^\infty((T_0,+\infty),H^s)}, \|z\|_{L^2((T_0,+\infty),H^s_{-\sigma})}, \|z\|_{L^2((T_0,+\infty),B^\beta_{\infty,2})}\right] \leq \delta\big\},$$

and ε, δ are small enough to ensure that for $U \in \mathcal{U}(\varepsilon,\delta)$ and $z \in \mathcal{Z}(U,\delta)$

$$S(U) + z + g(U,z) \in W^c(V_0) \cap B_{H^s}(S(V_0), r(V_0)).$$

For a sufficiently small T_0, we solve the equation for z first, then the one for ρ, and eventually the one for U using the method of section 3. This gives us the desired exponential decay for ρ as well as similar results for U and z.

We note that instead of Lemma 3.9, we obtain the following lemma.

LEMMA 4.1. *If Assumptions* 1.1–1.5 *hold, then for any $T > 0$, there exist $T_0 > 0$, $\varepsilon_0 > 0$, $\delta_0 > 0$, $C > 0$, and $\kappa \in (0,1)$ such that for any $\varepsilon \in (0,\varepsilon_0)$, $\delta \in (0,\delta_0)$, $U, U' \in \mathcal{U}_{T_0}(\varepsilon,\delta)$, $\rho, \rho' \in \mathcal{R}_{T_0,\gamma}$, $z_0 \in \mathcal{H}_c(U(0))$, $z_0' \in \mathcal{H}_c(U'(0))$, $z \in \mathcal{Z}_{T_0}(U,\delta)$, and $z' \in \mathcal{Z}_{T_0}(U',\delta)$, one has*

$$\begin{aligned} &\|z[z_0', U', \rho'] - z[z_0, U, \rho]\|_{L^\infty((T_0,T),H^s)\cap L^2((T_0,T),L^\infty)\cap L^2((T_0,T),H^s_{-\sigma})} \\ &\quad \leq C\left\|z_0 - z_0'\right\|_{H^s} + \kappa\big\{\left\|U - U'\right\|_{L^\infty((T_0,T),\mathbb{C}^2)} + \left\|\dot{U} - \dot{U}'\right\|_{L^\infty((T_0,T),\mathbb{C}^2)} \\ &\qquad + \|e^{\gamma t}(\rho - \rho')(t)\|_{L^\infty_t((T_0,T),X^s(V_0))}\big\}. \end{aligned}$$

Then for ρ as a function of U, z_0, and ρ_0 (the initial condition for ρ), we obtain the following lemma.

LEMMA 4.2. *If Assumptions* 1.1–1.5 *hold, then for any* $T > 0$ *there exist* $T_0 > 0$, $\varepsilon_0 > 0$, $\delta_0 > 0$, $C > 0$, *and* $\kappa \in (0,1)$ *such that for any* $\varepsilon \in (0, \varepsilon_0)$, $\delta \in (0, \delta_0)$, $U, U' \in \mathcal{U}_{T_0}(\varepsilon, \delta)$, $r_0, r_0' \in X_s(V_0)$, $z_0 \in \mathcal{H}_c(U(0))$, $z_0' \in \mathcal{H}_c(U'(0))$, $z \in \mathcal{Z}_{T_0}(U, \delta)$, *and* $z' \in \mathcal{Z}_{T_0}(U', \delta)$ *one has*

$$\begin{aligned}&\|e^{\gamma t}\left(\rho[z_0', U', \rho_0'] - \rho[z_0, U, \rho_0]\right)\|_{L_t^\infty((T_0,T),X^s(V_0))}\\ &\leq C\|z_0 - z_0'\|_{H^s} + \kappa\big\{\,\|U - U'\|_{L^\infty((T_0,T),\mathbb{C}^2)}) + \big\|\dot U - \dot U'\big\|_{L^\infty((T_0,T),\mathbb{C}^2)} + \|\rho_0 - \rho_0'\|_{L^2}\,\big\}.\end{aligned}$$

We also note that the proof gives the well-posedness of (4.1) in $\mathcal{R}_{T_0,\gamma}(\delta)$ with small initial condition ρ_0 and that there exists $C > 0$ such that the solution ρ satisfies

$$\|\rho(t)\|_{H^s} \leq C\,\|\rho_\pm(0)\|\,e^{-\gamma t} \quad \forall t > T_0.$$

The asymptotic behaviors of U and z are obtained as in the previous section when z_0 is localized.

5. End of the proof of main theorems. We note that the small *locally invariant* center manifold built in section 2.2 for (2.10) is now a small *invariant* (globally in time) center manifold. Indeed, we have just proved the stabilization towards the PLS manifold; this ensures that a solution in the center manifold will stay inside this manifold in the two directions of time.

Now let us consider $\mathcal{C}M$ as being the union of all these small globally invariant center manifolds and 0. Using the uniqueness of the center manifold and Lemma 3.2, we prove that $\mathcal{C}M \setminus \{0\}$ is a manifold. Now we generalize Lemma 3.2 by the following lemma.

LEMMA 5.1. *For any* $s, s', \sigma \in \mathbb{R}$ *and* $p, q \in [1, \infty]$, *there exist* $\varepsilon > 0$, *a continuous map* $r : B^2_{\mathbb{C}}(0,\varepsilon) \mapsto \mathbb{R}^+$ *with* $r(U) = O(\Gamma(U))$, *and a continuous map* $\Psi : S \mapsto \mathcal{C}M$, *where*

$$\mathcal{S}_\sigma = \big\{(U,z);\ U \in B_{\mathbb{C}^2}(0,\varepsilon),\ z \in \mathcal{H}_c(U) \cap B_{H^{s'}_\sigma}(0, r(U))\big\}$$

is endowed with the metric of $\mathbb{C}^2 \times H^{s'}_\sigma$.

Moreover, Ψ *is bijective from* $\mathcal{S}$ *to an open neigborhood of* $(0,0)$ *in* $\mathcal{C}M$ *and smooth on* $\mathcal{S} \setminus \{(0,0)\}$. *For all* $U \in B_{\mathbb{C}^2}(0,\varepsilon)$, *there exists* $C > 0$ *such that, for all* $z \in \mathcal{H}_c(U) \cap B_{H^{s'}}(0, r(U))$, $\Psi(U,z) \in \mathcal{H}_1(U)$, $z + \Psi(U,z) \in \mathcal{H}_0(U)^\perp$, *and* $S(U) + z + \Psi(U,z) \in \mathcal{C}M$. *For sufficiently small nonzero* U, *we have* $\|\Psi(U,z)\|_{B^s_{p,q}} = O(\|z\|^2_{H^{s'}})$ *for* $z \in H^{s'}$ *such that* $(U,z) \in \mathcal{S}$.

Proof. The proof works like that of Lemma 3.2. The statements for r follow from Remark 2.3. ∎

The scattering result follows from a one-to-one correspondence of the initial profile with the asymptotic profile as stated in the following proposition.

PROPOSITION 5.1. *If Assumptions* 1.1–1.5 *hold, there exist* $\varepsilon > 0$ *and a continuous map* $r : B^2_{\mathbb{C}}(0,\varepsilon) \mapsto \mathbb{R}^+$ *with* $r(U) = O(\Gamma(U))$ *and* $\mathcal{V}_\sigma$, $\mathcal{V}_\pm$ *neighborhoods of* $(0,0)$ *in*

$$S_\sigma = \big\{(U,z);\ U \in \mathbb{C}^2,\ z \in \mathcal{H}_c(U) \cap B_{H^s_\sigma}(0, r(U))\big\}$$

endowed with the norm of $\mathbb{C}^2 \times H^s_\sigma$ *such that the maps*

$$\mathcal{P}_\pm : \begin{pmatrix} U_0 \\ z_0 \end{pmatrix} \in \mathcal{V}_\sigma \mapsto \begin{pmatrix} V_{\pm\infty} \\ z_{\pm\infty} \end{pmatrix} \in \mathcal{V}_\pm$$

are bijections and are smooth on $\mathcal{V}_0 \setminus \{(0,0)\}$.

Proof. We choose, for example,

$$\mathcal{V}_\sigma = \left\{(U,z)\,;\, U \in B_{\mathbb{C}^2}(0,\varepsilon),\ z \in \mathcal{H}_c(U) \cap B_{H^s_\sigma}(0,r(U))\right\}$$

for some positive ε, and we work on the manifold $\mathcal{V}_\sigma \setminus \{(0,0)\}$ which is locally isomorphic to an open set of $\mathbb{C}^2 \times \mathcal{H}_c(U) \cap H^s_\sigma$. We write

$$\mathcal{P}^{U_0}_\pm (U,z) = (U,z) + \mathcal{R}^{U_0}_\pm (U,z)\,.$$

Since

$$\|(U_\infty, z_\infty) - (U_0, z_0)\|_{H^s_\sigma} = O\left(|U_0|^2 + \|z_0\|^2_{H^s_\sigma}\right),$$

we need only prove the statement locally. Hence we prove that, in a neighborhood of $(U_0,0)$, the maps $\mathcal{P}^{U_0}_\pm (U,z) \mapsto (Id_{\mathbb{C}^2}, P_c(U_0))\mathcal{P}_\pm (U, R(U,U_0)z)$ are bijective (P_c and R are defined in Proposition 2.2).

To prove that $\mathcal{P}^{U_0}_\pm$ is bijective (i.e., the scattering exists), let us prove it for $\mathcal{P}^{U_0}_+$ (it is similar for $\mathcal{P}^{U_0}_-$). It is enough to prove that the following system has a unique solution in an open neighborhood of $(0,0)$ in $\mathcal{S}_\sigma$:

$$\begin{aligned}
&V_\pm(t) = V_{\pm\infty} \\
&+\int_t^\infty A(V_\pm(v), e^{JsH(V_{\pm\infty})}\tilde\eta_\pm(v))\langle N(U(v), e^{JsH(V_{\pm\infty})}\tilde\eta_\pm(v)), dS(V_\pm(v))\rangle\, dv
\end{aligned}$$

and

$$\begin{aligned}
\tilde\xi_+(t) &= z_\infty - \int_t^\infty e^{-JsH(V_{+\infty})} \\
&\quad\times \mathbf{P}_c(V_+(v)) J\left(d^2F(S(V_+(v))) - d^2F(S(V_{+\infty}))\right) e^{JsH(V_{+\infty})}\tilde\xi_+(v)\, dv \\
&-\int_t^\infty e^{-JsH(V_{+\infty})} \\
&\quad\times \mathbf{P}_c(V_+(v)) JN(V_+(v), e^{JsH(V_{+\infty})}\tilde\eta_+(v))\, dv \\
&-\int_t^\infty e^{-JsH(V_{+\infty})}\mathbf{P}_c(V_+(v)) dS(V(v)) A(V_+(v), e^{JsH(V_{+\infty})}\tilde\eta_+(v)) \\
&\quad\times\langle N(V_+(v), e^{JsH(V_{+\infty})}\tilde\eta_+(v)), dS(V_+(v))\rangle\, dv \\
&+\int_t^\infty e^{-JsH(V_{+\infty})}(d\mathbf{P}_c(V_+(v))) A(V_+(v), e^{JsH(V_{+\infty})}\tilde\eta_+(v)) \\
&\quad\times\langle N(V_+(v), e^{JsH(V_{+\infty})}\tilde\eta_+(v)), dS(V_+(v))\rangle e^{JsH(V_{+\infty})}\tilde\eta_+(v)\, dv
\end{aligned}$$

with $\tilde\eta_+(t) = \tilde\xi_+(t) + e^{-JsH(V_{+\infty})} g\big(V_+(t), e^{JsH(V_{+\infty})}\tilde\xi_+(t)\big)$.

This system can be solved by a fixed point argument in the set of functions such that

$$\begin{aligned}
&\max\Bigg[\sup_{t\in\mathbb{R}}(\|z_{+\infty} - \tilde\xi_+(t)\|_{H^s}),\ \sup_{t\in\mathbb{R}}(\langle t\rangle^{3/2}\|z_{+\infty} - \tilde\xi_+(t)\|_{H^s_{-\sigma}}), \\
&\sup_{t\in\mathbb{R}}(\langle t\rangle^{3/2}\|z_{+\infty} - \tilde\xi_+(t)\|_{B^\beta_{\infty,2}}),\ \sup_{t\in\mathbb{R}}(\langle t\rangle^{-3/2}\|z_{+\infty} - \tilde\xi_+(t)\|_{H^s_{3/2}})\Bigg]
\end{aligned}$$

and

$$\langle t\rangle^2 \left|V_+(t) - V_{+\infty}\right|$$

are small with the method we used in Lemma 3.14. □

For the same reasons, the small locally invariant center-stable manifold built in section 2.2 is invariant in positive time. We can also consider the union of these manifolds, and we can obtain a map Φ_+ similar to the map Ψ built in Lemma 5.1.

The instability in negative time is, in fact, a consequence of Proposition 2.5.

The corresponding conclusion holds for the center-unstable manifold.

The statements on the instability outside these manifolds follow from Propositions 2.4 and 2.5.

Acknowledgments. I would like to thank Éric Séré for fruitful discussions and advice during the preparation of this work. I am indebted to Galina Perel'man for her careful reading and her comments and suggestions, to Michael Levitin for his advice, to Sylvain Golénia for having checked some proofs, to Matthieu Brassart for his answers to my questions, and to Mathieu Lewin for all his help. I am grateful to the referees for their careful reading, corrections, and remarks.

REFERENCES

[1] W. O. Amrein, A. Boutet de Monvel, and V. Georgescu, *C_0-Groups, Commutator Methods and Spectral Theory of N-Body Hamiltonians*, Progr. Math. 135, Birkhäuser Verlag, Basel, 1996.

[2] E. Balslev and B. Helffer, *Limiting absorption principle and resonances for the Dirac operator*, Adv. in Appl. Math., 13 (1992), pp. 186–215.

[3] P. W. Bates and C. K. R. T. Jones, *Invariant manifolds for semilinear partial differential equations*, in Dynamics Reported, Vol. 2, Dynam. Report. Ser. Dynam. Systems Appl. 2, Wiley, Chichester, UK, 1989, pp. 1–38.

[4] J. Bergh and J. Löfström, *Interpolation Spaces. An Introduction*, Grundlehren Math. Wiss. 223, Springer-Verlag, Berlin, 1976.

[5] A. Berthier and V. Georgescu, *On the point spectrum of Dirac operators*, J. Funct. Anal., 71 (1987), pp. 309–338.

[6] N. Boussaid, *Étude de la stabilité des petites solution stationnaires pour une classe d'équations de Dirac non linéaires*, Ph.D. thesis, Université Paris-Dauphine, Paris, 2006.

[7] N. Boussaid, *Stable directions for small nonlinear Dirac standing waves*, Comm. Math. Phys., 268 (2006), pp. 757–817.

[8] A. Bressan, *A tutorial on the center manifold theorem*, in Proceedings of the C.I.M.E. Course, Cetraro, Italy, 2003, Lecture Notes in Math. 1911, P. Marcati, ed., Springer-Verlag, New York, 2007, pp. 327–344.

[9] V. S. Buslaev and G. S. Perel'man, *Nonlinear scattering: States that are close to a soliton*, Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI), 200 (1992), pp. 38–50, 70, 187.

[10] V. S. Buslaev and G. S. Perel'man, *On nonlinear scattering of states which are close to a soliton*, Astérisque, no. 210 (1992), pp. 6, 49–63.

[11] V. S. Buslaev and G. S. Perel'man, *Scattering for the nonlinear Schrödinger equation: States that are close to a soliton*, Algebra i Analiz, 4 (1992), pp. 63–102.

[12] V. S. Buslaev and G. S. Perel'man, *On the stability of solitary waves for nonlinear Schrödinger equations*, in Nonlinear Evolution Equations, Amer. Math. Soc. Transl. Ser. 2 164, AMS, Providence, RI, 1995, pp. 75–98.

[13] V. S. Buslaev and C. Sulem, *Asymptotic stability of solitary waves for nonlinear Schrödinger equations*, in The Legacy of the Inverse Scattering Transform in Applied Mathematics (South Hadley, MA, 2001), Contemp. Math. 301, AMS, Providence, RI, 2002, pp. 163–181.

[14] V. S. Buslaev and C. Sulem, *On asymptotic stability of solitary waves for nonlinear Schrödinger equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 20 (2003), pp. 419–475.

[15] T. Cazenave and P.-L. Lions, *Orbital stability of standing waves for some nonlinear Schrödinger equations*, Comm. Math. Phys., 85 (1982), pp. 549–561.

[16] S. Cuccagna, *Stabilization of solutions to nonlinear Schrödinger equations*, Comm. Pure Appl. Math., 54 (2001), pp. 1110–1145.
[17] S. Cuccagna, *On asymptotic stability of ground states of NLS*, Rev. Math. Phys., 15 (2003), pp. 877–903.
[18] S. Cuccagna, *Erratum: "Stabilization of solutions to nonlinear Schrödinger equations"* [Comm. Pure Appl. Math., 54 (2001), pp. 1110–1145, MR1835384], Comm. Pure Appl. Math., 58 (2005), p. 147.
[19] M. Escobedo and L. Vega, *A semilinear Dirac equation in $H^s(\mathbf{R}^3)$ for $s > 1$*, SIAM J. Math. Anal., 28 (1997), pp. 338–362.
[20] V. Georgescu and M. Măntoiu, *On the spectral theory of singular Dirac type Hamiltonians*, J. Operator Theory, 46 (2001), pp. 289–321.
[21] M. Grillakis, J. Shatah, and W. Strauss, *Stability theory of solitary waves in the presence of symmetry*. I, J. Funct. Anal., 74 (1987), pp. 160–197.
[22] S. Gustafson, K. Nakanishi, and T.-P. Tsai, *Asymptotic stability and completeness in the energy space for nonlinear Schrödinger equations with small solitary waves*, Int. Math. Res. Not., no. 66 (2004), pp. 3559–3584.
[23] P. D. Hislop, *Exponential decay of two-body eigenfunctions: A review*, in Proceedings of the Symposium on Mathematical Physics and Quantum Field Theory (Berkeley, CA, 1999), Electron. J. Differ. Equ. Conf. 4, Southwest Texas State University, San Marcos, TX, 2000, pp. 265–288.
[24] W. Hunziker and I. M. Sigal, *Time-dependent scattering theory of N-body quantum systems*, Rev. Math. Phys., 12 (2000), pp. 1033–1084.
[25] A. Iftimovici and M. Măntoiu, *Limiting absorption principle at critical values for the Dirac operator*, Lett. Math. Phys., 49 (1999), pp. 235–243.
[26] A. Jensen and T. Kato, *Spectral properties of Schrödinger operators and time-decay of the wave functions*, Duke Math. J., 46 (1979), pp. 583–611.
[27] T. Kato, *Wave operators and similarity for some non-selfadjoint operators*, Math. Ann., 162 (1965/1966), pp. 258–279.
[28] T. Kato, *Perturbation Theory for Linear Operators*, Classics Math., Springer-Verlag, Berlin, 1995 (reprint of the 1980 edition).
[29] M. Keel and T. Tao, *Endpoint Strichartz estimates*, Amer. J. Math., 120 (1998), pp. 955–980.
[30] E. Kirr and A. Zarnescu, *On the asymptotic stability of bound states in 2D cubic Schrödinger equation*, Comm. Math. Phys., 272 (2007), pp. 443–468.
[31] J. Krieger and W. Schlag, *On the focusing critical semi-linear wave equation*, Amer. J. Math., 129 (2007), pp. 843–913.
[32] S. Machihara, M. Nakamura, K. Nakanishi, and T. Ozawa, *Endpoint Strichartz estimates and global solutions for the nonlinear Dirac equation*, J. Funct. Anal., 219 (2005), pp. 1–20.
[33] S. Machihara, M. Nakamura, and T. Ozawa, *Small global solutions for nonlinear Dirac equations*, Differential Integral Equations, 17 (2004), pp. 623–636.
[34] S. Machihara, K. Nakanishi, and T. Ozawa, *Small global solutions and the nonrelativistic limit for the nonlinear Dirac equation*, Rev. Mat. Iberoamericana, 19 (2003), pp. 179–194.
[35] T. Mizumachi, *Instability of bound states for 2D nonlinear Schrödinger equations*, Discrete Contin. Dyn. Syst., 13 (2005), pp. 413–428.
[36] T. Mizumachi, *A remark on linearly unstable standing wave solutions to NLS*, Nonlinear Anal., 64 (2006), pp. 657–676.
[37] B. Parisse, *Résonances paires pour l'opérateur de Dirac*, C. R. Acad. Sci. Paris Sér. I Math., 310 (1990), pp. 265–268.
[38] C.-A. Pillet and C. E. Wayne, *Invariant manifolds for a class of dispersive, Hamiltonian, partial differential equations*, J. Differential Equations, 141 (1997), pp. 310–326.
[39] A. F. Rañada, *Classical nonlinear Dirac field models of extended particles*, in Quantum Theory, Groups, Fields and Particles, Vol. 198, A. O. Barut, ed., Reidel, Amsterdam, pp. 271–291.
[40] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*. IV. *Analysis of Operators*, Academic Press [Harcourt Brace Jovanovich], New York, 1978.
[41] W. Schlag, *Stable Manifolds for an Orbitally Unstable NLS*, preprint, University of Chicago, Chicago, IL, 2004.
[42] J. Shatah and W. Strauss, *Instability of nonlinear bound states*, Comm. Math. Phys., 100 (1985), pp. 173–190.
[43] A. Soffer and M. I. Weinstein, *Multichannel nonlinear scattering for nonintegrable equations*, Comm. Math. Phys., 133 (1990), pp. 119–146.
[44] A. Soffer and M. I. Weinstein, *Multichannel nonlinear scattering for nonintegrable equations*. II. *The case of anisotropic potentials and data*, J. Differential Equations, 98 (1992), pp. 376–390.

[45] A. Soffer and M. I. Weinstein, *Selection of the ground state for nonlinear Schrödinger equations*, Rev. Math. Phys., 16 (2004), pp. 977–1071.
[46] A. Soffer and M. I. Weinstein, *Theory of nonlinear dispersive waves and selection of the ground state*, Phys. Rev. Lett., 95 (2005), 213905.
[47] N. Szpak, *Spontaneous Particle Creation in Time-Dependent Overcritical Fields of QED*, Ph.D. thesis, J. W. Goethe University, Frankfurt am Main, Germany, 2005.
[48] B. Thaller, *The Dirac Equation*, Texts Monogr. Phys., Springer-Verlag, Berlin, 1992.
[49] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, North–Holland Math. Library 18, North–Holland, Amsterdam, 1978.
[50] T.-P. Tsai, *Asymptotic dynamics of nonlinear Schrödinger equations with many bound states*, J. Differential Equations, 192 (2003), pp. 225–282.
[51] T.-P. Tsai and H.-T. Yau, *Asymptotic dynamics of nonlinear Schrödinger equations: Resonance-dominated and dispersion-dominated solutions*, Comm. Pure Appl. Math., 55 (2002), pp. 153–216.
[52] T.-P. Tsai and H.-T. Yau, *Classification of asymptotic profiles for nonlinear Schrödinger equations with small initial data*, Adv. Theor. Math. Phys., 6 (2002), pp. 107–139.
[53] T.-P. Tsai and H.-T. Yau, *Relaxation of excited states in nonlinear Schrödinger equations*, Int. Math. Res. Not., no. 31 (2002), pp. 1629–1673.
[54] T.-P. Tsai and H.-T. Yau, *Stable directions for excited states of nonlinear Schrödinger equations*, Comm. Partial Differential Equations, 27 (2002), pp. 2363–2402.
[55] R. Weder, *Center manifold for nonintegrable nonlinear Schrödinger equations on the line*, Comm. Math. Phys., 215 (2000), pp. 343–356.
[56] M. I. Weinstein, *Modulational stability of ground states of nonlinear Schrödinger equations*, SIAM J. Math. Anal., 16 (1985), pp. 472–491.
[57] M. I. Weinstein, *Lyapunov stability of ground states of nonlinear dispersive evolution equations*, Comm. Pure Appl. Math., 39 (1986), pp. 51–67.

SIAM J. MATH. ANAL.
Vol. 40, No. 4, pp. 1671–1691

ANALYSIS OF MODEL EQUATIONS FOR STRESS-ENHANCED DIFFUSION IN COAL LAYERS. PART I: EXISTENCE OF A WEAK SOLUTION*

ANDRO MIKELIĆ† AND HANS BRUINING‡

Abstract. This paper is motivated by the study of the sorption processes in the coal. They are modeled by a nonlinear degenerate pseudoparabolic equation for stress-enhanced diffusion of carbon dioxide (CO_2) in coal, $\partial_t \phi = \partial_x \left\{ D(\phi)\, \partial_x \phi + \frac{D(\phi)\phi}{B} \partial_x \left(e^{-m\phi} \partial_t \phi \right) \right\}$, where B, m are positive constants and the diffusion coefficient $D(\phi)$ has a small value when the CO_2 volume fraction ϕ is $0 \leq \phi < \phi_c$, representative of coal in the glass state and orders of magnitude higher value for $\phi > \phi_c$, when coal is in the rubber-like state. These types of equations arise in a number of cases when nonequilibrium thermodynamics or extended nonequilibrium thermodynamics is used to compute the flux. For this equation, existence of the travelling wave–type solutions was extensively studied. Nevertheless, the existence seems to be known only for a sufficiently short time. We use the corresponding entropy functional in order to get existence, for any time interval, of an appropriate weak solution with square integrable first derivatives and satisfying uniform L^∞-bounds. Due to the degeneracy, we obtain square integrability of the mixed second order derivative only in the region where the concentration ϕ is strictly positive. In obtaining the existence result it was crucial to have the regularized entropy as unknown for the approximate problem and not the original unknown (the concentration).

Key words. degenerate pseudoparabolic equation, entropy methods, stress-enhanced diffusion

AMS subject classifications. 35K70, 35K65, 76R50, 80A17

DOI. 10.1137/070710172

1. Introduction.

One of the promising methods for reducing the discharge of the "greenhouse gas" carbon dioxide (CO_2) into the atmosphere is its sequestration in unminable coal seams. A typical procedure is the injection of CO_2 via deviated wells drilled inside the coal seams. CO_2 displaces the methane adsorbed on the internal surface of the coal. A production well gathers the methane as free gas. This process, known as CO_2-enhanced coal bed methane production (CO_2-ECBM), is a producer of energy and at the same time reduces greenhouse concentrations as about two CO_2 molecules displace one molecule of methane. Worldwide application of ECBM can reduce greenhouse gas emissions by a few percent. Coal has an extensive fracturing system called the cleat system. In fact, it is possible to discern a number of cleat

*Received by the editors December 5, 2007; accepted for publication (in revised form) July 8, 2008; published electronically November 26, 2008.
http://www.siam.org/journals/sima/40-4/71017.html

†Université de Lyon, Lyon, F-69003, France, and Université Lyon 1, Institut Camille Jordan, UFR Mathématiques, Site de Gerland, Bât. A, 50, avenue Tony Garnier, 69367 Lyon Cedex 07, France (mikelic@univ-lyon1.fr). This author's research was partially supported by the Groupement MOMAS (Modélisation mathématique et simulations numériques liées aux problèmes de gestion des déchets nucléaires) (PACEN/CNRS, ANDRA, BRGM, CEA, EDF, IRSN) as a part of the project "*Modèles de dispersion efficace pour des problèmes de Chimie-Transport: Changement d'échelle dans la modélisation du transport réactif en milieux poreux, en présence des nombres caractéristiques dominants.*" It was initiated during this author's sabbatical visit to the TU Eindhoven in 2006, supported by Visitors grant B-61-602 of the Netherlands Organisation for Scientific Research (NWO).

‡Dietz Laboratory, Geo-Environmental Engineering, Stevinweg 1, 2628 CN Delft, The Netherlands (j.bruining@tudelft.nl). This author's research was supported by Shell International Exploration and Production B.V., NWO-Novem, and the Greenhouse-gas Removal Apprenticeship and Student Program (GRASP) funded by the EC. The idea of stress-enhanced diffusion [29], [30] for coal, worked out in the doctoral thesis of Mazumder [23], was the inspiration for this work.

systems at different scales. In the end, the matrix blocks between the smallest cleat systems typically have diameters of a few tens of microns [13].

The matrix blocks have a polymeric structure (dehydrated cellulose [32]), which provides the adsorption sites for the gases. At low temperatures or low sorption concentration the coal structure behaves like a rigid glassy polymer, in which movement is difficult. At high temperatures or high sorption concentrations, the glassy structure is converted to the less rigid and open rubber-like (swollen) structure [29], [30]. As coal is less dense in the rubber-like state, a conversion from the glassy state to the rubber-like state exhibits swelling. Therefore modeling of diffusion is not only relevant for modeling transport into the matrix blocks, but also for the modeling of swelling, which affects the permeability of the coal seam.

Ritger and Peppas [29], [30] distinguish between transport by Fickean diffusion and a process that occurs on the interface between the glass state and the rubber-like state. Ritger and Peppas state that the conversion process from the glass to the rubber state is controlled by a rate-limiting relaxation phenomenon (see also [2]). Thomas and Windle [31] (see also [16], [17], [19]), however, suggested in their classic paper that the diffusion transport was enhanced by stress gradients that resulted from the accommodation of large molecules in the small cavities providing the adsorption sites. For this, Alfrey, Gurnee, and Lloyd [1] coined the term superdiffusion or case II diffusion. At a critical concentration of the penetrants the glassy polymer is transformed to a rubber state, where the diffusion coefficient is of the order of a factor 1000 larger than in the glassy state.

This paper is the first of a series in which the model equations for case II diffusion [31], [16], [17], [19] will be analyzed. Our longtime interest is to investigate the one-dimensional sorption rate behavior, i.e., whether the equations indeed lead to a rate faster than the square root of time. In this paper, we establish existence of a weak solution for all times.

Nonlinear diffusion equations with a pseudoparabolic regularizing term being the Laplacian of the time derivative are considered in [25] and [26]. Global existence of a strong solution is proved by writing the problem as a linear elliptic operator, acting on the time derivative, equal to the nonlinear diffusion term. Then the linear elliptic operator, acting on the time derivative, is inverted, and the standard geometric theory of nonlinear parabolic equations (see, e.g., [15]) is applicable.

In our situation the physical model leads to a degenerate nonlinear second order elliptic operator, acting on the time derivative, in place of the Laplacian. The invertibility of this nonlinear elliptic operator is not clear anymore and depends on the solution itself. The same type of equation can occur in models that use classical irreversible thermodynamics (CIT) or extended irreversible thermodynamics (EIT). An important example is the model of the two-phase flow through porous media introduced in [14], where the capillary pressure relation is extended with a dynamic term, which contains the time derivative of the saturation. We also refer the reader to [5] for the modeling. This application to multiphase and unsaturated flows through porous media motivated a number of recent papers. In [18] one finds a detailed study of possible travelling wave solutions and in particular of the behavior of such travelling waves near fronts where the concentration is zero. Further studies of the travelling waves are in [8] and [7]. The small- and waiting-time behavior of the equations is studied in [20]. Study of the viscosity limit for the linear relaxation model of the dynamic term is in [10]. Nevertheless, the study of existence of a solution to the nonlinear model from [14] is undertaken only in [4] and [5], where the nondegeneracy is supposed and existence is local in time. Another existence result, also local in

time, is in the paper [9] by Düll, where a related pseudoparabolic equation modeling solvent uptake in polymeric solids is studied. Düll proved the short-time existence of a solution for the problem in $\mathbb{R}$, supposing nonnegative compactly supported initial datum. Contrary to our approach, the problem was written as a system containing a linear elliptic equation and an evolution equation. With such an approach, we did not manage to get estimates as good as those with the entropy approach undertaken in this paper. For studies of travelling waves and sharp fronts in case II diffusion models, we refer the reader to [17] and [33].

We consider the evolution problem

$$\partial_t \phi = \partial_x \left\{ D(\phi)\, \partial_x \phi + \frac{D(\phi)\, \phi}{B} \partial_x \left(e^{-m\phi} \partial_t \phi \right) \right\} \quad \text{in } (0, L) \times (0, T), \tag{1}$$

$$D(\phi)\, \partial_x \phi + \frac{D(\phi)\, \phi}{B} \partial_x \left(e^{-m\phi} \partial_t \phi \right) = 0 \quad \text{on } \{x = L\} \times (0, T), \tag{2}$$

$$\phi(0, t) = \phi_g(t) \quad \text{on } (0, T), \quad \phi = 0 \quad \text{on } (0, L) \times \{0\}. \tag{3}$$

Our goal is to obtain a global existence of a weak solution, for any time interval, as was obtained in [3] for a degenerate pseudoparabolic regularization of a nonlinear forward-backward heat equation. Our PDE allows a natural generalization of the classic Kullback entropy, and its integrand is given by

$$\mathcal{E}(\varphi) = \int_0^{\varphi} \frac{\varphi - \xi}{\xi} \left(e^{-m\xi} \frac{1}{D(\xi)} - \frac{1}{D(0)} \right) d\xi + \frac{1}{D(0)} (\varphi \log \varphi - \varphi). \tag{4}$$

As in [24], we will use $\mathcal{E}'(\varphi)$ as a test function, with the hope of obtaining a convenient a priori estimate. Formal calculation gives the equality

$$\begin{aligned} &\partial_t \int_0^L \left\{ \mathcal{E}(\phi) - \varphi \mathcal{E}'(\varphi_g) + \frac{1}{2B} (e^{-m\phi} \partial_x \phi)^2 \right\} dx \\ &\quad + \int_0^L \left(\frac{1}{\phi} e^{-m\phi} (\partial_x \phi)^2 + \phi \partial_t \mathcal{E}'(\varphi_g) \right) dx = 0. \end{aligned} \tag{5}$$

The presence of the initial and boundary conditions leads to unbounded nonintegrable $\mathcal{E}'$. The equality (5) cannot be used directly, and we do not get the entropy estimates as in [12]. We had to obtain an additional estimate for the time derivative, and our calculations are more complicated than in the literature.

Existence is proved by showing that the "energy" of the system remains bounded during the time evolution of the system. The "energy" equation is derived from the differential equation by multiplying with an appropriate test function and integrating over the domain. The choice of the test function depends strongly on the choice of the nonlinearities. With an appropriate approximation, this can also be the basis of a numerical scheme that leads to an implicit first order nonlinear system of ODEs. The implicit dependence on the time derivative makes its solvability nontrivial. Solvability of our system of ODEs depends strongly on the initial conditions. The fact that the "energy" is bounded means that the numerical scheme is stable. If convergence can be proved, it shows that at least one solution exists.

As already stated, in this case an appropriate test function is $\Phi(\phi)$, where

$$\Phi'(\xi) = \frac{e^{-m\xi}}{\xi D(\xi)},$$

is, however, singular for $\xi = 0$. Another problem of the test function is that for large values of ξ, Φ' is exponentially small. In order to prove existence we need Φ that is bijective from $\mathbb{R}$ to $\mathbb{R}$. Concerning the diffusion coefficient D, we extend it by setting $D(\xi) = D(-\xi)$ for $\xi < 0$.

We introduce Φ_δ by

$$\Phi'_\delta := \frac{e^{-m\min\{|\xi|,1/\delta\}}}{(|\xi|+\delta)\,D(\xi)}, \quad \delta > 0,\ \xi \in \mathbb{R}, \tag{6}$$

and

$$\Phi_\delta(\phi) := \begin{cases} \displaystyle\int_0^\phi \frac{e^{-m\min\{\xi,1/\delta\}}}{(\xi+\delta)D(\xi)}\,d\xi, & \phi > 0, \\ \displaystyle -\int_\phi^0 \frac{e^{-m\min\{-\xi,1/\delta\}}}{(-\xi+\delta)D(-\xi)}\,d\xi, & \phi < 0. \end{cases} \tag{7}$$

Obviously, Φ_δ is odd and strictly increasing on $\mathbb{R}_+$.

In order to obtain an existence result for problem (1)–(3) we study the following regularized problem in $Q_T = (0,L)\times(0,T)$:

$$\partial_t\phi = \partial_x\left\{D(\phi)\,\partial_x\phi + \frac{D(\phi)(|\phi|+\delta)}{B}\partial_x\left(e^{-m\min\{|\phi|,1/\delta\}}\partial_t\phi\right)\right\} \tag{8}$$

with boundary condition at $x = L$,

$$D(\phi)\,\partial_x\phi + \frac{D(\phi)(|\phi|+\delta)}{B}\partial_x\left(e^{-m\min\{|\phi|,1/\delta\}}\partial_t\phi\right)\Big|_{x=L} = 0, \tag{9}$$

and boundary and initial conditions (3).

We start by introducing a variational solution for the problem (8), (9), and (3).

Definition 1. *Let*

$$\mathcal{V} := \{z \in C^\infty[0,L],\ z|_{x=0} = 0\} \quad \textit{and} \quad \mathcal{H} := \{C^\infty[0,T],\ h(T) = 0\}. \tag{10}$$

Then the variational formulation corresponding to the problem (3), (8), *and* (9) *is*

$$\begin{aligned} &-\int_0^T\int_0^L \phi(x,t)\,g(x)\,\partial_t h(t)\,dx\,dt + \int_0^T\int_0^L D(\phi)\,\partial_x\phi(x,t)\,\partial_x g(x)h(t)\,dx\,dt \\ &+\int_0^T\int_0^L \frac{D(\phi)(|\phi|+\delta)}{B}\partial_x g(x)\,h(t)\,\partial_x\left(e^{-m\min\{|\phi|,1/\delta\}}\partial_t\phi\right)dx\,dt = 0 \end{aligned} \tag{11}$$

for all $g \in \mathcal{V}$ and for all $h \in \mathcal{H}$, and at the boundary $x = 0$ we have

$$\phi - \phi_g = 0. \tag{12}$$

Our goal is to prove existence for (11)–(12). In order to have the entropy estimate, we should formulate the approximate problem in terms of it. Otherwise it *would not be possible to use it as a test function* for the approximate problem, which is finite dimensional. Getting a priori estimates without this approach is not clear.

Let $z := \Phi_\delta(\phi)$, $\phi = \Phi_\delta^{-1}(z)$, $z|_{x=0} = \Phi_\delta(\phi_g(t))$. We reformulate the problem (3), (8), and (9) in terms of z:

$$\frac{1}{\Phi_\delta'\left(\Phi_\delta^{-1}(z)\right)}\partial_t z = \partial_x\left\{\frac{D\left(\Phi_\delta^{-1}(z)\right)}{\Phi_\delta'\left(\Phi_\delta^{-1}(z)\right)}\partial_x z\right.$$

$$\left. + \frac{D\left(\Phi_\delta^{-1}(z)\right)\left(\left|\Phi_\delta^{-1}(z)\right|+\delta\right)}{B}\partial_x\left(D\left(\Phi_\delta^{-1}(z)\right)\left(\left|\Phi_\delta^{-1}(z)\right|+\delta\right)\partial_t z\right)\right\} \quad \text{in } Q_T. \tag{13}$$

Moreover we can express the boundary and initial conditions in z as

$$z(0,t) = \Phi_\delta(\phi_g(t)) \quad \text{on } (0,T), \quad z(x,t=0) = \Phi_\delta(0) = 0 \quad \text{on } (0,L), \tag{14}$$

$$\frac{1}{\Phi_\delta'\left(\Phi_\delta^{-1}(z)\right)}\partial_x z + \frac{\left(\left|\Phi_\delta^{-1}(z)\right|+\delta\right)}{B}\partial_x\left(D\left(\Phi_\delta^{-1}(z)\right)\left(\left|\Phi_\delta^{-1}(z)\right|+\delta\right)\partial_t z\right) = 0 \quad \text{at } x = L. \tag{15}$$

Our paper is organized as follows: section 2 describes the physical model, first proposed in [31]. We repeat the derivations from [16], [17], [19] for reasons of easy reference and unified notation.

In section 3 we introduce the discretization of the problem (13)–(15). We get the Cauchy problem for an implicit first order system of ODEs. Next the solvability of the discretized problem is proved, and the uniform L^2 a priori estimates for the first derivatives and the mixed second derivative are obtained for a small time interval $(0, T_0)$. They imply the short-time existence for the regularized problem.

We continue with section 4 where we use the entropy to establish the existence of a solution for the regularized problem for all times. Next, we establish L^∞-bounds independent of the regularization parameter.

The last section 5 concerns the existence for the original problem. Using the entropy, the estimates for the time derivative, and the L^∞-bounds again, we are able to pass to the limit when the regularization parameter tends to zero and prove the existence of at least one solution for the original problem.

2. Model equation for stress-induced diffusion. Consider a coal particle between the fractured cleat systems in coal. The matrix block can be considered as a small (30 μm diameter) cubical particle consisting of glassy coal. The coal face is exposed to the penetrant, in our case CO_2. The coal face of the particle and the mechanism of the sorption process is shown schematically in Figure 1. The coal originates from a cellulose-like polymeric structure [32], with the chemical formula $C_n(H_2O)_m$, from which part of the hydrogen and oxygen have disappeared during the coalification process, which took millions of years. The remaining structure behaves like a glassy polymer, which contains holes (sites) that can accommodate CO_2, CH_4, etc. In other words, sorption of gases by coal is more a dissolution process than is adsorption of gases at a coal surface. The holes receiving the CO_2 are originally too small to accommodate the molecule and need to expand. Consequently, the expanded hole exerts a stress on the neighboring molecules constituting the polymeric coal. Therefore the penetration of CO_2 will lead to both a stress gradient and a concentration gradient. The concentration will be expressed as a volume fraction ϕ, i.e., $\phi = c/\Omega$, where c is the molecular concentration and Ω is the molecular volume. As the CO_2 likes to move toward a region of smaller stress, the transport of the molecule will be caused by both a concentration gradient and a stress gradient. When the stresses become too high, a deformation occurs, in which the glassy polymeric structure is converted to a rubber-like (swollen) structure, which is much more open. Consequently the diffusion coefficient in the rubber-like structure is much higher (more

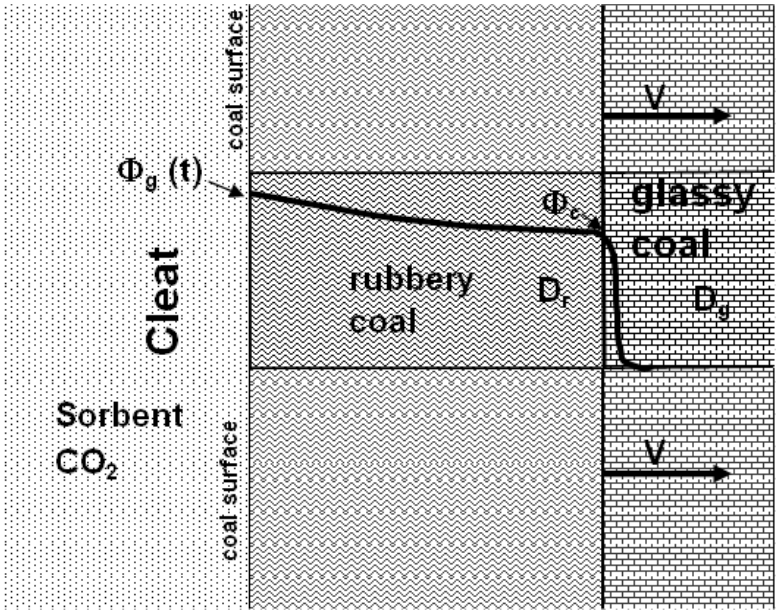


FIG. 1. *A coal face exposed to a sorbent (CO_2). To the far right is the virgin coal, which behaves as a glassy polymer. As the sorbent penetrates in the coal, a reorientation of the polymeric coal structure occurs, and the coal becomes rubber-like. The diffusion coefficient in the rubber-like structure is much higher ($> 1000\times$) than in the glassy structure. The rubber-like structure has also a lower density leading to swelling.*

than 1000 times) than the diffusion coefficient in the glassy structure. The stresses are considered to depend on the CO_2 concentration in the coal, and conversion to the rubber-like structure occurs instantaneously when a certain critical concentration is exceeded.

These ideas were formulated for the first time by Thomas and Windle [31], and the derivation of the model equations will be explained below.

2.1. Derivation of model equations. The salient features of the Thomas and Windle model [31] are well summarized by Hui et al. [16], [17]. We summarize the derivation here with the help of the article by Hui et al. and the book of Landau and Lifshitz [21]; i.e., the molar (diffusive) flux J is not only driven by the volume fraction (ϕ) (concentration) gradient, but also by the stress (P_{xx}) gradient, i.e.,

$$J = -D\left(\frac{\partial \phi}{\partial x} + \frac{\Omega\phi}{kT}\frac{\partial P_{xx}}{\partial x}\right), \tag{16}$$

where k is the Boltzmann constant. As opposed to the equation in [21], which contains a scalar pressure gradient, the idea here is extended in [19] with the use of the stress gradient $\partial_x P_{xx}$. Hui (see [16], [17]) interprets P as the osmotic pressure. Note that J is the flux of a volume fraction and behaves as a velocity. The diffusion coefficient depends on the concentration. Below a critical volume fraction ϕ_c, a diffusion coefficient $D_g > 0$ characteristic of a glassy state is used, and above ϕ_c the diffusion coefficient $D_r > 0$ characteristic of the rubber (swollen) state is used. It can be expected that $D_r/D_g \gg 1$. In the model an abrupt change of the diffusion coefficients at ϕ_c is used, but D_r and D_g are considered constant for $\phi > \phi_c$ and $\phi < \phi_c$, respectively:

$$D(\xi) := \begin{cases} D_g, & 0 \le \xi < \phi_c - \kappa, \\ D_g + (D_r - D_g)(\xi - \phi_c + \kappa)/(2\kappa), & \phi_c - \kappa \le \xi \le \phi_c + \kappa, \\ D_r, & \phi_c + \kappa < \xi < +\infty, \end{cases} \tag{17}$$

where $\kappa > 0$ is a small parameter. Extended nonequilibrium thermodynamics [19] suggests that vice versa also the stress (P_{xx}) is related to the volumetric flux gradient as

$$P_{xx} = -\eta_l \frac{\partial J}{\partial x} = \eta_l \frac{\partial \phi}{\partial t}, \tag{18}$$

where the second equality follows from a mass conservation law that assumes incompressible flow,

$$\frac{\partial \phi}{\partial t} + \frac{\partial J}{\partial x} = 0. \tag{19}$$

With η_l we denote the elongational velocity [6], i.e., the resistance of movement due to a velocity gradient $\frac{\partial J}{\partial x}$ in the direction of flow. Elongational viscosity is caused by a resistance force of a fluid to accelerate. Hence, the force is proportional to the component of the gradient of the velocity in the flow direction. Elongational viscosity η_l is always larger than the shear viscosity η_s, e.g., in Newtonian fluids $\eta_l = 3\eta_s$. In this case the "fluid CO_2" is moving in the coal medium. The resistance to flow is largely determined by the coal-CO_2 interaction and not, as in the usual definition of viscosity, as CO_2-CO_2 interaction. Hence here we deal with an apparent or pseudoviscosity. With increasing CO_2 concentration the coal becomes more rubberlike, i.e., it acquires a more open structure, and the apparent viscosity decreases with increasing concentration (see (20)). The elongational viscosity η_l is supposed to depend on the volume fraction of the penetrant as

$$\eta_l = \eta_o \exp(-m\phi), \tag{20}$$

where m is a material constant and η_0 is the volumetric viscosity of the unswollen coal sample.

Substituting expression (16) for the flux into the mass balance equation (19), where we also use (18), we obtain

$$\partial_t \phi = \partial_x \left\{ D(\phi)\, \partial_x \phi + \frac{D(\phi)\,\phi}{B} \partial_x \left(e^{-m\phi} \partial_t \phi \right) \right\}, \tag{21}$$

where the constant $B = k_B T/(\eta_o \Omega)$. This equation is defined in $Q_T = (0, L) \times (0, T)$.

As initial condition we have that the concentration is

$$\phi(x, t = 0) = 0 \quad \text{on } (0, T). \tag{22}$$

The boundary condition at $x = 0$ must be derived from thermodynamic arguments. The final equilibrium concentration is reached when the coal has swollen to make the stress $P_{xx} = P^0_{xx}$ equal to zero. In this case the volume fraction of CO_2 in the coal is in equilibrium with the CO_2 in the fluid phase outside the coal. Also the CO_2 in the stressed coal is in equilibrium with the CO_2 in the fluid phase. The change in chemical potential is $d\mu = \Omega\, dP_{xx} + k_B T\, d\ln\phi$. Equating the chemical potential in the unstressed and stressed state leads to

$$\Omega P_{xx} + k_B T \ln\phi = \Omega P^0_{xx} + k_B T \ln\phi_o, \tag{23}$$

where ϕ_o is the volume fraction at the coal boundary that would be in equilibrium with the CO_2 in the gas phase if the coal has relaxed to the rubber state with $P^0_{xx} = 0$.

Substituting (18) and (20) into (23) leads to

$$t = -\phi_o \frac{\eta_0 \Omega}{k_B T} \int_0^{\phi/\phi_o} \frac{\exp(-m\phi_o y)}{\ln y}\, dy, \tag{24}$$

where we use the initial condition that $\phi = 0$ at $t = 0$. Singularity of the integrand at $y = 1$ guarantees that ϕ remains bounded by ϕ_o for all times.

At $x = L$ we have the boundary condition on $(0, T)$,

$$D(\phi)\left(\partial_x \phi + \frac{1}{B}\phi \partial_x \left(\exp(-m\phi)\, \partial_t \phi\right)\right)_{x=L} = 0. \tag{25}$$

In summary, we have one initial condition equation (22), one boundary condition at $x = L$, viz. (25), and the implicit boundary condition equation (24), which specifies $\phi(x = 0, t)$ as

$$\phi(0, t) = \phi_g(t). \tag{26}$$

ϕ_g satisfies the conditions

$$0 \le \phi_g \le A_0, \quad \phi_g(0) = 0. \tag{27}$$

Remark 2. Equations like (21) can occur in many transport problems in which the flux is calculated using CIT or EIT. A well-known example for CIT in porous media flow is that the deviation of the capillary pressure P_c from its equilibrium value at a given oil saturation S_o, i.e., $P_c^o = P_c^o(S_o)$, is a driving force leading to a rate of change of the saturation (scalar flux). This leads [14], [27], [28], [22] to $\partial_t S_o = L(P_c - P_c^o)$ and to the transport equation for counter current imbibition:

$$\begin{aligned}\varphi \partial_t S_o &= \partial_x \left(\Lambda(S_o)\, \partial_x P_c\right) \\ &= \partial_x \left(\Lambda(S_o)\, \partial_x P_c^o(S_o)\right) + \partial_x \left(\Lambda(S_o)\, \partial_x \frac{1}{L(S_o)} \partial_t S_o\right).\end{aligned}$$

EIT [19] differs from CIT as it characterizes a system not only by its local thermodynamic variables (pressure, temperature, and concentration) but also by its gradients. The explanation in [19] is difficult to follow by nonspecialists as many thermodynamic relations are considered to be known by the reader. In isothermal systems and in the absence of other applied fields, e.g., electric fields, the volumetric flux J is, according to EIT, given by the following system of equations:

$$\tau_1 \partial_t J + J = -D\left(\frac{\partial \phi}{\partial x} + \frac{\Omega \phi}{kT} \frac{\partial P_{xx}}{\partial x}\right), \tag{28}$$

$$\tau_2 \partial_t P_{xx} + P_{xx} = -\eta_l \frac{\partial J}{\partial x}. \tag{29}$$

Reference [19] uses a mass flux instead of a volumetric flux and therefore uses a factor υ_1, being the partial volume per unit mass. Here τ_1, τ_2 are time constants, which are small with respect to L^2/D. The first terms on the left-hand sides of (28) and (29) appear only in EIT and not in CIT. The first terms on the left are of interest for

short-time behavior and are omitted from the model discussed here. Another example from EIT is the Taylor dispersion problem (see equation 10.34 in [19]), where there is an "xxt" derivative in the concentration, apart from many other terms. Hence, EIT or CIT can lead to transport equations of the form of (21).

3. Short-time existence for the regularized problem. In this section, we first introduce an approximate problem corresponding to (13)–(15). It is a first order system of ODEs for expansion coefficients, with implicit dependence on the time derivatives. First we prove the solvability on some interval $(0, T_N)$, where N is the parameter describing discretization in space. Then, we use the entropy to prove the solvability on interval $(0, T_0)$, where T_0 does not depend on N. Finally, we pass to the limit $N \to +\infty$ and prove that the problem (13)–(15) itself has a solution on $(0, T_0)$.

Let $V := \left\{ g \in H^1(0,L) \mid g(0) = 0 \right\}$ be the closure of $\mathcal{V}$ in $H^1(0,L)$, and let $\{\alpha_j\}_{j\in\mathbb{N}}$ be a C^∞-basis for V. We set $V_N := \operatorname{span}\{\alpha_1, \dots, \alpha_N\}$ and introduce the following coefficients:

$$d_1(z) := \frac{1}{\left(\Phi_\delta'\left(\Phi_\delta^{-1}(z)\right)\right)}, \quad d_2(z) := \frac{D\left(\Phi_\delta^{-1}(z)\right)}{\left(\Phi_\delta'\left(\Phi_\delta^{-1}(z)\right)\right)}, \quad \text{and}$$

$$d(z) := D\left(\Phi_\delta^{-1}(z)\right)\left(\left|\Phi_\delta^{-1}(z)\right| + \delta\right). \tag{30}$$

The coefficients d_1 and d_2 are continuous, nonnegative, and bounded functions of z. d is a continuous function of z, bounded away from zero.

We start study of the initial boundary problem (13)–(15) by constructing an approximate solution for every N. It is defined as follows.

Approximate problem 3. *For $q \in (2, +\infty)$, find $z_N = \sum_{j=1}^N c_j(t)\,\alpha_j(x) + \Phi_\delta(\phi_g(t)) \in W^{1,q}([0,T]; V_N)$ such that*

$$\int_0^L \partial_t z_N d_1(z_N)\,\alpha_k\,dx + \int_0^L d_2(z_N)\,\partial_x z_N \partial_x \alpha_k\,dx$$

$$+ \int_0^L \frac{1}{B} d(z_N)\,\partial_x\left(d(z_N)\,\partial_t z_N\right)\partial_x\alpha_k\,dx = 0 \text{ for } k = 1, \dots, N, \quad \text{and} \tag{31}$$

$$z_N|_{t=0} = P_N\left(z|_{t=0} - \Phi_\delta(\phi_g(0))\right) = 0, \tag{32}$$

where $P : V \to V_N$ is the projector $P_N(f)(x) := \sum_{j=1}^N \alpha_j(x)\,(f, \alpha_j)_V$.

Let the vector valued function $\mathbf{F}$ be given by $F_\kappa(t, \mathbf{c}, \partial_t \mathbf{c}) =$ left-hand side of (31) and let $\mathbf{c}$ be the column vector consisting of elements $(c_1(t) \dots c_N(t))$; then (31), (32) are equivalent to the following Cauchy problem in $\mathbb{R}^N$:

$$\begin{cases} \mathbf{F}(t, \mathbf{c}, \partial_t \mathbf{c}) = 0, \\ \mathbf{c}|_{t=0} = 0. \end{cases} \tag{33}$$

The Cauchy problem (33) is difficult to solve, since the dependence of $\mathbf{F}$ on $\partial_t \mathbf{c}$ is implicit. It is crucial to reduce it to an ordinary Cauchy problem of the form $\partial_t \mathbf{c} = \varrho(t, \mathbf{c})$.

We note that

$$F_k := \sum_{j=1}^N \left\{ \int_0^L d_1(z_N)\,\alpha_k\alpha_j\,dx + \int_0^L \frac{1}{B} d(z_N)\,\partial_x\left(d(z_N)\,\alpha_j\right)\partial_x\alpha_k\,dx \right\} \frac{dc_j}{dt}$$

$$+ \sum_{j=1}^N \left\{ \int_0^L d_2(z_N)\,\partial_x\alpha_j\partial_x\alpha_k\,dx \right\} c_j + \int_0^L d_1(z_N)\,\alpha_k \partial_t \Phi_\delta(\phi_g(t))\,dx. \tag{34}$$

Then, after introducing the matrices $\mathcal{A}(\mathbf{c})$ and $\mathcal{B}(\mathbf{c})$ and the vector $\mathbf{f}(\mathbf{c})$ by

$$\mathcal{A}_{kj}(\mathbf{c}) := \int_0^L d_1(z_N)\,\alpha_k\alpha_j\,dx + \int_0^L \frac{1}{B} d(z_N)\,\partial_x\left(d(z_N)\,\alpha_j\right)\partial_x\alpha_k\,dx, \tag{35}$$

$$\mathcal{B}_{kj}(\mathbf{c}) := \int_0^L d_2(z_N)\,\partial_x\alpha_j\partial_x\alpha_k\,dx, \quad \text{and} \quad f_k(\mathbf{c}) = \int_0^L d_1(z_N)\,\alpha_k\partial_t\Phi_\delta(\phi_g(t))\,dx, \tag{36}$$

$1 \le k, j \le N$, we see that the problem (31)–(32) is equivalent to the following Cauchy problem:

$$\text{Find} \quad \mathbf{c} \in W^{1,q}(0,T)^N \quad \text{such that}$$
$$\mathcal{A}(\mathbf{c})\frac{d\mathbf{c}}{dt} = -\mathcal{B}(\mathbf{c})\mathbf{c} - \mathbf{f}(\mathbf{c}) \ \text{ a.e. in } \ (0,T); \quad \mathbf{c}|_{t=0} = 0. \tag{37}$$

PROPOSITION 4. *There is a $T_N > 0$ such that the problem* (31)–(32) *has a unique solution $z_N \in W^{1,q}(0,T_N;V_N)$ for all $q < +\infty$.*

Proof. It is enough to prove that the Cauchy problem (37) has a solution.

Obviously, $\mathcal{A}$, $\mathcal{B}$, and $\mathbf{f}$ are smooth functions of $\mathbf{c}$. Because of the singularity of $\partial_t\varphi_g$ at $t = 0$, $\mathbf{f}(\mathbf{c}) \in L^q(0,T)$ for all $q < +\infty$, but it is not bounded. Hence, the only property to check is the invertibility of the matrix $\mathcal{A}$. Let $\mathbf{b}$ be an arbitrary vector from $\mathbb{R}^N$ and let $b_\alpha(x) = \mathbf{b}\cdot\alpha(x) = \sum_{j=1}^N b_j\alpha_j(x)$. Then we have

$$\begin{aligned}(\mathcal{A}\mathbf{b})\cdot\mathbf{b} = \sum_{k,j=1}^N \mathcal{A}_{k,j}b_kb_j &= \int_0^L d_1(z_N)(b_\alpha)^2\,dx + \frac{1}{B}\int_0^L d(z_N)\partial_x b_\alpha\partial_x(d(z_N)b_\alpha)\,dx \\ &= \int_0^L d_1(z_N)(b_\alpha)^2\,dx + \frac{1}{B}\int_0^L (d(z_N)\partial_x b_\alpha)^2\,dx \\ &\quad + \frac{1}{B}\int_0^L d(z_N)\partial_x b_\alpha b_\alpha d'(z_N)\partial_x z_N\,dx \\ &\ge \int_0^L \left\{d_1(z_N) - \frac{1}{4B}(d'(z_N))^2(\partial_x z_N)^2\right\}(b_\alpha)^2\,dx. \end{aligned} \tag{38}$$

Since $\partial_x z_N(x,0) = 0$ and functions $\{\alpha_j\}_{j\in\mathbb{N}}$ are linearly independent, the matrix $\mathcal{A}$ is by (38) invertible in a neighborhood of $t = 0$. Then by the classical theory, the problem (37) has a unique solution on some interval $(0,T_N)$. ∎

Next, we want to prove that the existence interval does not depend on N.

PROPOSITION 5. *There is a constant C, independent of N, such that*

$$\|\partial_x z_N\|_{L^\infty(0,T_N;L^2(0,L))} \le C. \tag{39}$$

Consequently, the vector valued function $\mathbf{c}$ *remains bounded at $t = T_N$.*

Proof. In (31) we can replace α_k by $z_N - \Phi_\delta(\phi_g)$. Then after using that $\partial_x\left(d(z_N)\,\partial_t z_N\right) = \partial_t\left(d(z_N)\,\partial_x z_N\right)$, we get

$$\begin{aligned}&\int_0^L d_1(z_N)\,z_N\partial_t z_N\,dx + \int_0^L d_2(z_N)\,(\partial_x z_N)^2\,dx \\ &\quad + \int_0^L \frac{1}{B}\partial_t\left(d(z_N)\,\partial_x z_N\right)d(z_N)\,\partial_x z_N\,dx\end{aligned} \tag{40}$$

$$= \int_0^L d_1(z_N)\,\Phi_\delta(\phi_g)\,\partial_t z_N\,dx = \partial_t \int_0^L \Phi_\delta(\phi_g)(t) \int_0^{z_N} d_1(\xi)\,d\xi\,dx$$
$$- \partial_t \Phi_\delta(\phi_g)(t) \int_0^L \int_0^{z_N} d_1(\xi)\,d\xi\,dx.$$

Integrating over t leads to

$$\int_0^L \left(\int_0^{z_N(x,t)} d_1(\xi)\,\xi\,d\xi \right) dx + \int_0^t \int_0^L d_2(z_N)(\partial_x z_N)^2\,dx\,d\tau + \frac{1}{2B}\int_0^L d(z_N)^2(\partial_x z_N)^2\,dx \tag{41}$$
$$= \int_0^L \left(\int_0^{z_N(x,t)} d_1(\xi)\,d\xi \right) dx\,\Phi_\delta(\phi_g)(t) - \int_0^t \partial_\tau \Phi_\delta(\phi_g)(\tau) \left(\int_0^L \int_0^{z_N} d_1(\xi)\,d\xi\,dx \right) d\tau.$$

We easily find out that

$$\int_0^z d_1(\xi)\,\xi\,d\xi = \int_0^{\Phi_\delta^{-1}(z)} \Phi_\delta(\eta)\,d\eta \quad \text{and} \quad \int_0^z d_1(\xi)\,d\xi = \Phi_\delta^{-1}(z). \tag{42}$$

The growth of the terms in (42) indicates that it will be possible to control the two terms on the right-hand side of (41) by the first term on the left-hand side of (41).

Let $M_\phi := \max_{0\le t\le T} |\Phi_\delta(\phi_g(t))|$. Then by the definition of $\Phi_\delta(\varphi)$, we have $C_0(\delta)\log(1+\varphi/\delta) \le \Phi_\delta(\varphi)$ for all $\varphi \ge 0$. Hence $\int_0^z d_1(\xi)\,\xi\,d\xi \ge C_0(\delta)((|\Phi_\delta^{-1}(z)| + \delta)\log(1+|\Phi_\delta^{-1}(z)|/\delta) - |\Phi_\delta^{-1}(z)|)$, and there is a large enough constant $C_\varphi = C_\varphi(M_\phi,\delta)$ such that $g(z) = C_0(\delta)((|\Phi_\delta^{-1}(z)|+\delta)\log(1+|\Phi_\delta^{-1}(z)|/\delta) - |\Phi_\delta^{-1}(z)|) - M_\phi|\Phi_\delta^{-1}(z)| + C_\varphi > |\Phi_\delta^{-1}(z)|$ for all z. The equality (41) now implies

$$\int_0^L g(z_N(x,t))\,dx + \int_0^t \int_0^L d_2(z_N)(\partial_x z_N)^2\,dx\,d\tau + \frac{1}{2B}\int_0^L d(z_N)^2(\partial_x z_N(t))^2\,dx$$
$$\le C_\varphi(M_\phi,\delta)\,L + \int_0^t |\partial_\tau \Phi_\delta(\phi_g)(\tau)| \left(\int_0^L g(z_N(x,\tau))\,dx \right) d\tau. \tag{43}$$

Since $\partial_\tau \Phi_\delta(\phi_g) \in L^1(0,T)$, we apply Gronwall's inequality, and estimate (39) follows. Hence $\mathbf{c}$ remains bounded at $t = T_N$. ∎

Nevertheless, since the matrix $\mathcal{A}$ could degenerate, some components of $\frac{\partial \mathbf{c}}{\partial t}$ could blow up at $t = T_N$. In order to exclude this possibility and to prove that the maximal solution for (33) exists on $[0,T]$, we need an estimate for the time derivatives. Furthermore, if we want to pass to the limit $N \to +\infty$ in (31), estimate (39) is not sufficient. Our strategy is to obtain an estimate, uniform with respect to N, for $\partial_{xt} z_N$ in $L^2(Q_T)$.

THEOREM 6. *There exists $T_0 > 0$, independent of N, such that*

$$\|\partial_x z_N\|_{L^\infty(0,T_0;L^2(0,L))} \le C, \tag{44}$$
$$\|\partial_t z_N\|_{L^2(0,T_0;L^2(0,L))} \le C, \tag{45}$$
$$\|\partial_{xt} z_N\|_{L^2(0,T_0;L^2(0,L))} \le C, \tag{46}$$
$$\left\| \partial_{xt} \int_0^{z_N} d(\xi)\,d\xi \right\|_{L^2((0,T_0)\times(0,L))} \le C, \tag{47}$$

with constants independent of N. Consequently, the maximal solution for (33) *exists on* $[0,T_0]$.

Proof. We replace α_k in (31) by $\partial_t z_N - \partial_t \Phi_\delta(\phi_g)$. This yields

$$\int_0^L d_1(z_N)(\partial_t z_N)^2\,dx + \int_0^L d_2(z_N)\,\partial_x z_N \partial_{xt} z_N\,dx$$
$$+\frac{1}{B}\int_0^L d(z_N)\,\partial_t\left(d(z_N)\,\partial_x z_N\right)\partial_{xt} z_N\,dx = \int_0^L d_1(z_N)\,\partial_t z_N \partial_t \Phi_\delta(\phi_g)\,dx. \tag{48}$$

In the estimates which follow we will use the fact that integrability of higher order derivatives implies continuity and boundedness in x or in t. We recall that for one-dimensional Sobolev embeddings Morrey's theorem applies and $H^1(0,t)$ (respectively, $H^1(0,L)$) is continuously embedded into the Hölder space $C^{0,1/2}[0,t]$ (respectively, into $C^{0,1/2}[0,L]$). See, e.g., [11] for more details. In our particular situation, we use the explicit dependence of the embedding constant on the length of the time interval and we prefer to derive the estimates directly.

First, as $\partial_x z_N \in L^2\left(0,L;H^1(0,t)\right)$ and $\partial_x z_N|_{\tau=0}=0$, we have for a.e. $x\in(0,L)$ and for every $\tau\in(0,t)$

$$|\partial_x z_N(x,\tau)| = \left|\int_0^\tau \partial_\xi \partial_x z_N(x,\xi)\,d\xi\right| \le \sqrt{\tau}\sqrt{\int_0^\tau |\partial_\xi \partial_x z_N(x,\xi)|^2\,d\xi}. \tag{49}$$

Next, as $\partial_\tau z_N \in L^2\left(0,t;H^1(0,L)\right)$ and $\partial_\tau z_N|_{\tau=0} = \partial_\tau \Phi(\phi_g)$, we have for a.e. $\tau\in(0,t)$ and for every $x\in(0,L)$

$$|\partial_\tau z_N(x,\tau)| \le |\partial_\tau \Phi(\phi_g(\tau))| + \left|\int_0^x \partial_\xi \partial_\tau z_N(\xi,\tau)\,d\xi\right|$$
$$\le |\partial_\tau \Phi(\phi_g(\tau))| + \sqrt{L}\sqrt{\int_0^L |\partial_\xi \partial_\tau z_N(\xi,\tau)|^2\,d\xi}. \tag{50}$$

Estimates (49)–(50) imply

$$\int_0^L\int_0^t |\partial_\tau z_N(x,\tau)|^2\,|\partial_x z_N(x,\tau)|^2\,dx\,d\tau$$
$$\le 2\int_0^L\int_0^t \tau\left(\int_0^t |\partial_\xi \partial_x z_N(x,\xi)|^2\,d\xi\right)\left(|\partial_\tau\Phi(\phi_g(\tau))|^2 + L\int_0^L |\partial_\xi\partial_\tau z_N(\xi,\tau)|^2\,d\xi\right)d\tau\,dx$$
$$\le 2Lt\,\|\partial_{x\tau} z_N\|^4_{L^2((0,t)\times(0,L))} + 2\,\|\partial_{x\tau} z_N\|^2_{L^2((0,t)\times(0,L))}\int_0^t \tau|\Phi'_\delta(\phi_g)|^2|\partial_\tau\phi_g|^2\,d\tau$$
$$\le \left(2\sqrt{L}\sqrt{t}\,\|\partial_{x\tau} z_N\|^2_{L^2((0,t)\times(0,L))} + \frac{1}{\sqrt{2L}}\int_0^t \tau^{1/2}|\Phi'_\delta(\phi_g)|^2|\partial_\tau\phi_g|^2\,d\tau\right)^2. \tag{51}$$

Now we integrate (48) with respect to time, over $(0,t)$, and estimate the obtained terms. The second term is estimated as follows:

$$\left|\int_0^t\int_0^L d_2(z_N)\,\partial_x z_N \partial_{x\tau} z_N\,dx\,d\tau\right| \le C\int_0^t \|\partial_{x\tau} z_N(\tau)\|_{L^2(0,L)}\,\|\partial_x z_N(\tau)\|_{L^2(0,L)}\,d\tau$$
$$\le C\sqrt{\int_0^t \|\partial_{x\tau} z_N(\tau)\|^2_{L^2(0,L)}\,d\tau}\sqrt{\int_0^t\int_0^L (\partial_x z_N)^2\,dx\,d\tau} \le C\,\|\partial_{x\tau} z_N\|_{L^2((0,t)\times(0,L))}, \tag{52}$$

where we have used the estimate (39). We rewrite the third term of (48), omitting the $1/B$ factor, as

$$\int_0^t\int_0^L d(z_N)\partial_\tau\left(d(z_N)\partial_x z_N\right)\partial_{x\tau}z_N\,dx\,d\tau = \int_0^t\int_0^L d(z_N)\left(\partial_{x\tau}z_N\right)^2 dx\,d\tau$$

$$+\int_0^t\int_0^L d(z_N)d'(z_N)\partial_\tau z_N\partial_x z_N\partial_{x\tau}z_N\,dx\,d\tau. \tag{53}$$

The last term in (53) is cubic in derivatives of z_N. Our idea is to use the estimate (51), showing that for small times it enters with a small coefficient and then controlling it using other terms. Using the estimate (51), we find out that it satisfies the following inequality:

$$\left|\int_0^t\int_0^L d(z_N)\partial_{x\tau}z_N\,d'(z_N)\partial_\tau z_N\partial_x z_N\,dx\,d\tau\right| \le C\left\|\partial_{x\tau}z_N\right\|_{L^2(Q_t)}\left\|\partial_\tau z_N\partial_x z_N\right\|_{L^2(Q_t)}$$

$$\le C\sqrt{t}\left(\left\|\partial_{x\tau}z_N\right\|^3_{L^2(Q_t)}+\left\|\tau^{1/6}\Phi'_\delta(\phi_g)\partial_\tau\phi_g\right\|^3_{L^3(0,t)}\right), \tag{54}$$

where $Q_t=(0,t)\times(0,L)$. Let $X^2(t):=\int_0^t\int_0^L\left|\partial_{x\tau}z_N\right|^2dx\,d\tau$. Since $\partial_\tau\Phi_g(\phi_g)\in L^2(0,t)$, estimates (39), (52), (53), and (54) imply

$$\left\|\sqrt{d_1(z_N)}\partial_\tau z_N\right\|^2_{L^2((0,t)\times(0,L))}+X^2(t)-C_1\sqrt{t}X^3(t)\le C_o, \tag{55}$$

where C_o depends on $\left\|\partial_\tau\Phi_\delta(\phi_g)\right\|_{L^2(0,t)}$ and on the constant from estimate (39). We note that the last term on the left-hand side corresponds to the lower bound for the cubic term, corresponding to the stress gradient part of the diffusive flux. Inequality (55) is satisfied for $t=0$. The function $\varrho(X)=X^2-C_1\sqrt{t}X^3$ has its maximum on $(0,+\infty)$ in the point $X_o=3/\left(2C_1\sqrt{t}\right)$. If $C_o<\varrho(X_o)$, then inequality (55) gives an estimate for $X(t)$. We note that $C_o<\varrho(X_o)$ if $t<\frac{4}{27C_1^2C_o}$. Hence for $T\le\frac{4}{27C_1^2C_o}=T_0$ we have estimates (44)–(46).

From (44)–(46) it follows that $\partial_x z_N\partial_t z_N\in L^2\left(0,T_0;L^2(0,L)\right)\le C$, and we have (47) as well. □

The estimates (44)–(47) allow us to pass to the limit $N\to+\infty$. Using classical compactness and weak compactness arguments and due to the a priori estimates (44)–(47), we can extract a subsequence of z_N, denoted by the same subscripts, which converges to an element $z\in H^1((0,T_0)\times(0,L))$, $\partial_{xt}z\in L^2((0,T_0)\times(0,L))$, in the following sense:

$$z_N\to z \text{ strongly in } L^2((0,T_0)\times(0,L)) \text{ and a.e. on } (0,T_0)\times(0,L), \tag{56}$$

$$\partial_x z_N\rightharpoonup\partial_x z \text{ weakly in } L^2((0,T_0)\times(0,L)), \tag{57}$$

$$\partial_t z_N\rightharpoonup\partial_t z \text{ weakly in } L^2((0,T_0)\times(0,L)), \tag{58}$$

$$\partial_{xt} z_N\rightharpoonup\partial_{xt} z \text{ weakly in } L^2((0,T_0)\times(0,L)), \tag{59}$$

$$\partial_{xt}\int_0^{z_N} d(\xi)\,d\xi\rightharpoonup\partial_{xt}\int_0^{z} d(\xi)\,d\xi \text{ weakly in } L^2((0,T_0)\times(0,L)). \tag{60}$$

Now passing to the limit $N\to\infty$ in (31) does not pose problems, and we conclude that z satisfies (13)–(15).

We summarize the results in the following theorem.

THEOREM 7. *Let $\phi_g \in H^1(0,T)$. Then there exists $T_0 > 0$ such that problem* (13)–(15) *has at least one variational solution $z \in H^1((0,T_0)\times(0,L))$, $\partial_{xt} z \in L^2((0,T_0)\times(0,L))$.*

COROLLARY 8. *Let $\phi_g \in H^1(0,T)$. Then there exists $T_0 > 0$ such that the variational formulation* (11)–(12) *of the problem* (8), (9), *and* (3) *has at least one solution $\phi = \Phi_\delta^{-1}(z) \in H^1((0,T_0)\times(0,L))$, $\partial_{xt}\phi \in L^2((0,T_0)\times(0,L))$.*

4. Existence of the regularized problem. In this section, we first use the regularized entropy function to prove the global existence for the problem (8), (9), and (3) (i.e., for the regularized problem). Then we establish the L^∞-bounds for the solution, independent of the regularization parameter.

Let us prove that any solution ϕ for the problem (8), (9), and (3), constructed in Corollary 8, could be extended from $(0,T_0)$ to arbitrary time interval $(0,T)$. First we test (11) by $\Phi_\delta(\phi) - \Phi_\delta(\phi_g(t))$. We have

$$\int_0^t\int_0^L \partial_\tau\phi\Phi_\delta(\phi)\,dx\,d\tau + \int_0^t\int_0^L D(\phi)\,\partial_x\phi\partial_x\Phi_\delta(\phi)\,dx\,d\tau + \int_0^t\int_0^L \frac{D(\phi)(|\phi|+\delta)}{B}\partial_x\left(e^{-m\min\{|\phi|,1/\delta\}}\partial_\tau\phi\right)\partial_x\Phi_\delta(\phi)\,dx\,d\tau = \int_0^t\int_0^L \partial_\tau\phi\Phi_\delta(\phi_g)\,dx\,d\tau,$$

and it follows that

$$\int_0^L\left(\int_0^{\phi(t)}\Phi_\delta(\xi)\,d\xi\right)dx + \int_0^t\int_0^L \frac{1}{2B}\partial_\tau\left(e^{-m\min\{|\phi|,1/\delta\}}\partial_x\phi\right)^2 dx\,d\tau + \int_0^t\int_0^L D(\phi)\Phi_\delta'(\phi)(\partial_x\phi)^2\,dx\,d\tau = \int_0^L \phi(t)\Phi_\delta(\phi_g(t))\,dx - \int_0^t\int_0^L \phi\partial_\tau\Phi_\delta(\phi_g)\,dx\,d\tau$$

and we get as in the proof of Proposition 5

$$\|\partial_x\phi\|_{L^\infty(0,t;L^2(0,L))} \le C. \tag{61}$$

This estimate implies the boundedness of ϕ. We note that C does not depend on the smoothing of D at $\phi = \phi_c$.

Next we test (11) by

$$e^{-m\min\{|\phi|,1/\delta\}}\partial_t\phi - e^{-m\min\{|\phi_g|,1/\delta\}}\partial_t\phi_g$$

and get

$$\int_0^t\int_0^L (\partial_\tau\phi)^2 e^{-m\min\{|\phi|,1/\delta\}}\,dx\,d\tau + \int_0^t\int_0^L D(\phi)\,\partial_x\phi\partial_x\left(e^{-m\min\{|\phi|,1/\delta\}}\partial_\tau\phi\right)dx\,d\tau + \int_0^t\int_0^L \frac{D(\phi)(|\phi|+\delta)}{2B}\left(\partial_{x\tau}\int_0^\phi e^{-m\min\{|\xi|,1/\delta\}}\,d\xi\right)^2 dx\,d\tau$$

$$= \int_0^t \int_0^L \partial_\tau \phi \, e^{-m \min\{|\phi_g|, 1/\delta\}} \partial_\tau \phi_g \, dx \, d\tau.$$

Then, as in the proof of Proposition 5, by estimating the second and the fourth terms and after using (61), we conclude that

$$\|\partial_\tau \phi\|_{L^2((0,t)\times(0,L))} \le C, \tag{62}$$

$$\left\| \partial_{x\tau} \int_0^\phi e^{-m \min\{|\xi|, 1/\delta\}} \, d\xi \right\|_{L^2((0,t)\times(0,L))} \le C, \tag{63}$$

and from this it follows that

$$\|\partial_{x\tau} \phi\|_{L^2((0,t)\times(0,L))} \le C. \tag{64}$$

Therefore, we arrive at the following theorem.

THEOREM 9. *Let $\phi_g \in H^1(0,T)$. Then for all $T > 0$ there exists a weak solution $\phi \in H^1((0,T)\times(0,L))$, $\partial_{xt}\phi \in L^2((0,T)\times(0,L))$ for the variational formulation* (11)–(12) *of the problem* (8), (9), *and* (3).

We conclude this section by establishing uniform L^∞-bounds for ϕ. We have the following proposition.

PROPOSITION 10. *Let $\phi_g \in H^1(0,T)$ and $\phi_g \ge 0$. Then any weak solution ϕ of the problem* (8), (9), *and* (3), *obtained in Theorem* 9, *satisfies $\phi(x,t) \ge 0$ a.e. on Q_T.*

Proof. Let $a_- = -\inf\{a, 0\}$ and $a_+ = \sup\{a, 0\}$. Then $a = a_+ - a_-$ and $\Phi_\delta((\phi_g)_-) = \Phi_\delta(0) = 0$. We test (11) by $\Phi_\delta(\phi_-)$. Note that $\Phi_\delta(\phi_-)|_{x=0} = 0$ and $\Phi_\delta(\phi_-) \ge 0$. Then we have

$$\int_0^t \int_0^L (\partial_\tau \phi) \, \Phi_\delta(\phi_-) \, dx \, d\tau + \int_0^t \int_0^L D(\phi) \, \partial_x \phi \partial_x \Phi_\delta(\phi_-) \, dx \, d\tau$$
$$+ \int_0^t \int_0^L \frac{D(\phi)(|\phi| + \delta)}{B} \partial_x \left(e^{-m \min\{|\phi|, 1/\delta\}} \partial_\tau \phi \right) \Phi_\delta'(\phi_-) \, \partial_x \phi_- \, dx \, d\tau = 0.$$

Since $\phi_-|_{t=0} = 0$, $\phi_+ \phi_- = 0$, and $|\phi| \, \phi_- = \phi_-^2$, we get

$$\int_0^L \left(\int_0^{\phi_-(x,t)} \Phi_\delta(\xi) \, d\xi \right) dx + \int_0^t \int_0^L D(\phi_-) \, \Phi_\delta'(\phi_-) \, (\partial_x \phi_-)^2 \, dx \, d\tau$$
$$+ \int_0^L \frac{D(\phi)(|\phi| + \delta)}{2B} \left(e^{-m \min\{|\phi|, 1/\delta\}} \partial_x \phi_- \right)^2 (t) \, dx = 0.$$

It follows that $\partial_x \phi_- = 0$ and $\phi_-|_{x=0} = 0$. Therefore $\phi_- = 0$, and consequently $\phi = \phi_+ \ge 0$. ∎

In the uniform bounds which follow, we use, for given positive constants m and δ, the function

$$G(z) := \int_0^z \exp\{-m \min\{\xi, 1/\delta\}\} \, d\xi, \quad z \ge 0. \tag{65}$$

Then we have the following bounds.

PROPOSITION 11. *Let $\phi_g \in H^1(0,T)$, $\phi_g \ge 0$ and $\partial_t \phi_g \ge 0$ a.e. on $(0,T)$. Then any weak solution ϕ of the problem* (8), (9), *and* (3), *obtained in Theorem* 9, *satisfies $\phi_g(t) \ge \phi(x,t)$ a.e. on Q_T.*

Proof. Let G be given by (65). We test (11) by $(G(\phi)-G(\phi_g))_+$. Note that $(G(\phi)-G(\phi_g))_+\big|_{x=0}=0$. Then we have

$$\int_0^t\int_0^L \partial_\tau\phi\,(G(\phi)-G(\phi_g))_+\,dx\,d\tau+\int_0^t\int_0^L D(\phi)\,\partial_x\phi\partial_x\,(G(\phi)-G(\phi_g))_+\,dx\,d\tau$$

(66)

$$+\int_0^t\int_0^L \frac{D(\phi)(|\phi|+\delta)}{B}\partial_x\left(e^{-m\min\{|\phi|,1/\delta\}}\partial_\tau\phi\right)\partial_x\,(G(\phi)-G(\phi_g))_+\,dx\,d\tau=0.$$

Note that

(67)

$$\partial_\tau\phi\,(G(\phi)-G(\phi_g))_+=\partial_\tau\left(\int_0^\phi (G(\xi)-G(\phi_g))_+\,d\xi\right)+G'(\phi_g)\partial_\tau\phi_g\,(\phi-\phi_g)_+$$

and

$$\frac{D(\phi)(|\phi|+\delta)}{B}\partial_\tau\partial_x G(\phi)\partial_x\,(G(\phi)-G(\phi_g))_+=\partial_\tau\left(\frac{D(\phi)(|\phi|+\delta)}{2B}(\partial_x\,(G(\phi)\right.$$

(68)

$$\left.-G(\phi_g))_+)^2\right)-(\partial_x\,(G(\phi)-G(\phi_g))_+)^2\partial_\tau\left(\frac{D(\phi)(|\phi|+\delta)}{2B}\right).$$

Then using the monotonicity of ϕ_g and G we obtain from (66), (67), and (68) the following inequality:

$$\int_0^L\left(\int_0^{\phi(x,t)}(G(\xi)-G(\phi_g))_+\,d\xi\right)dx+\int_0^t\int_0^L\frac{D(\phi)}{G'(\phi)}\left(\partial_x\,(G(\phi)-G(\phi_g))_+\right)^2dx\,d\tau$$

$$+\int_0^L\frac{D(\phi)(|\phi|+\delta)}{2B}(\partial_x\,(G(\phi)-G(\phi_g))_+)^2\,dx$$

$$\le\int_0^t\int_0^L(\partial_x\,(G(\phi)-G(\phi_g))_+)^2\partial_\tau\left(\frac{D(\phi)(|\phi|+\delta)}{2B}\right)dx\,d\tau.$$

Since $\partial_\tau\left(\frac{D(\phi)(|\phi|+\delta)}{2B}\right)\in L^2(0,T;L^\infty(0,L))$, we apply Gronwall's lemma and conclude that $(G(\phi)-G(\phi_g))_+=0$, from which it follows that $G(\phi)\le G(\phi_g)$. Inversion of this equation leads to $\phi(x,t)\le\phi_g(t)$ a.e. on Q_T. ∎

PROPOSITION 12. *Let $\phi_g\in H^1(0,T)$, and let us suppose in addition that there are constants $A_0>0$, $\alpha>0$, and $C_0>0$ such that*

$$A_0\ge\phi_g(t)\ge C_0t^\alpha\quad\forall t\in[0,T]. \tag{69}$$

Then any weak solution ϕ of the problem (8), (9), *and* (3), *obtained in Theorem* 9, *satisfies $A_0\ge\phi(x,t)\ge C_0t^\alpha$ a.e. on Q_T.*

Proof. The proof follows the lines of Proposition 11. It is enough to prove the lower bound. We test (11) by $(G(C_0t^\alpha)-G(\phi))_-$. Note that $(G(C_0t^\alpha)-G(\phi))_-\big|_{x=0}=0$. Then as in the proof of Proposition 11 we have

$$\int_0^t\int_0^L\partial_\tau\phi\,(G(C_0\tau^\alpha)-G(\phi))_-\,dx\,d\tau \tag{70}$$

$$+\int_0^t\int_0^L D(\phi)\,\partial_x\phi\partial_x\,(G(C_0\tau^\alpha)-G(\phi))_-\,dx\,d\tau$$

$$+\int_0^t\int_0^L\frac{D(\phi)(|\phi|+\delta)}{B}\partial_x\left(e^{-m\min\{|\phi|,1/\delta\}}\partial_\tau\phi\right)\partial_x\,(G(C_0\tau^\alpha)-G(\phi))_-\,dx\,d\tau=0.$$

Note that

$$G(\phi) = G(C_0 t^\alpha) - \left(G\left(C_0 t^\alpha\right) - G\left(\phi\right)\right)_+ + \left(G\left(C_0 t^\alpha\right) - G\left(\phi\right)\right)_-, \tag{71}$$

$$\partial_\tau \phi \left(G\left(C_0 \tau^\alpha\right) - G\left(\phi\right)\right)_- = \frac{\partial_\tau G(C_0 \tau^\alpha)}{G'(\phi)} \left(G\left(C_0 \tau^\alpha\right) - G\left(\phi\right)\right)_-$$

$$+ \frac{1}{2G'(\phi)} \partial_\tau \left(G\left(C_0 \tau^\alpha\right) - G\left(\phi\right)\right)_-^2 \geq \frac{1}{2G'(\phi)} \partial_\tau \left(G\left(C_0 \tau^\alpha\right) - G\left(\phi\right)\right)_-^2, \tag{72}$$

and

$$\frac{D\left(\phi\right)\left(|\phi| + \delta\right)}{B} \partial_\tau \partial_x G(\phi)\, \partial_x \left(G\left(C_0 \tau^\alpha\right) - G\left(\phi\right)\right)_-$$

$$= \partial_\tau \left(\frac{D\left(\phi\right)\left(|\phi| + \delta\right)}{2B} \left(\partial_x \left(G\left(C_0 \tau^\alpha\right) - G\left(\phi\right)\right)_-\right)^2 \right)$$

$$- \left(\partial_x \left(G\left(C_0 \tau^\alpha\right) - G\left(\phi\right)\right)_-\right)^2 \partial_\tau \left(\frac{D\left(\phi\right)\left(|\phi| + \delta\right)}{2B} \right). \tag{73}$$

Then using the monotonicity of G we obtain from (70), (72), and (73) the following inequality:

$$\int_0^L \frac{\left(G\left(C_0 t^\alpha\right) - G\left(\phi\right)\right)_-^2}{2G'(\phi)}\, dx + \int_0^t \int_0^L \frac{D\left(\phi\right)}{G'\left(\phi\right)} \left(\partial_x \left(G\left(C_0 \tau^\alpha\right) - G\left(\phi\right)\right)_-\right)^2 dx\, d\tau$$

$$+ \int_0^L \frac{D\left(\phi\right)\left(|\phi| + \delta\right)}{2B} \left(\partial_x \left(G\left(C_0 t^\alpha\right) - G\left(\phi\right)\right)_-\right)^2 dx$$

$$\leq \int_0^t \int_0^L \left(\partial_x \left(G\left(C_0 \tau^\alpha\right) - G\left(\phi\right)\right)_-\right)^2 \partial_\tau \frac{1}{2G'(\phi)}\, dx\, d\tau$$

$$+ \int_0^t \int_0^L \left(G\left(C_0 \tau^\alpha\right) - G\left(\phi\right)\right)_-^2 \partial_\tau \left(\frac{D\left(\phi\right)\left(|\phi| + \delta\right)}{2B} \right) dx\, d\tau.$$

Since $\partial_t \left(\frac{D(\phi)(|\phi|+\delta)}{2B}\right)$ and $\partial_t \frac{1}{2G'(\phi)}$ are elements of $L^2(0,T;L^\infty(0,L))$, we apply Gronwall's lemma to the function

$$\int_0^t \left\|\partial_\tau \phi(\tau)\right\|_{L^\infty(0,L)} \left\|\left(G\left(C_0 \tau^\alpha\right) - G\left(\phi(\tau)\right)\right)_-\right\|_{H^1(0,L)}^2 d\tau$$

and conclude that $\left(G\left(C_0 t^\alpha\right) - G\left(\phi\right)\right)_- = 0$, from which it follows that $G\left(\phi\right) \geq G\left(C_0 t^\alpha\right)$. Inversion of G leads to $\phi(x,t) \geq C_0 t^\alpha$ a.e. on Q_T. ∎

THEOREM 13. *Let $\phi_g \in H^1(0,T)$, $A_0 = \max_{0\leq t\leq T} \phi_g(t)$, $A_0 \geq \phi_g \geq C_0 t^\alpha$, and $\alpha > 1$. Then there exists a weak solution ϕ, $C_0 t^\alpha \leq \phi\left(x,t\right) \leq A_0$, $\partial_{xt}\phi \in L^2\left((0,T)\times(0,L)\right)$, $\phi \in H^1\left((0,T)\times(0,L)\right)$, for the problem* (8), (9), *and* (3).

Remark 14.

- By choosing $\delta < 1/A_0$, we can replace $e^{-m\min\{|\phi|,1/\delta\}}$ by $e^{-m\phi}$ and $|\phi| + \delta$ by $\phi + \delta$.
- In addition to the assumptions of Theorem 13 let us suppose that $\partial_t \phi_g \geq 0$. Then there exists a weak solution ϕ, $C_0 t^\alpha \leq \phi\left(x,t\right) \leq \phi_g(t)$, $\partial_{xt}\phi \in L^2\left((0,T)\times(0,L)\right)$, $\phi \in H^1\left((0,T)\times(0,L)\right)$, for the problem (8), (9), and (3).

5. Existence for the original problem. It remains to pass to the limit $\delta \to 0$. This limit will give us the solvability of the starting problem (1)–(3).

After Theorem 13, we are free to replace the nonlinearity $\exp\{-m\phi\}$ by $h(\xi) = e^{-m}\min\{\xi, A_0\}$, $\xi \geq 0$. We have existence for the system (11)–(12); i.e., for every $g \in L^2(0,T;V)$, $V = \{g \in H^1(0,L) \mid g(0) = 0\}$, we have

(74)
$$\int_0^T\int_0^L \partial_t\phi_\delta\, g\, dx\, dt + \int_0^T\int_0^L D(\phi_\delta)\left\{\partial_x\phi_\delta + \frac{(\phi_\delta+\delta)}{B}\partial_x\left(h(\phi_\delta)\,\partial_t\phi_\delta\right)\right\}\partial_x g\, dx\, dt = 0,$$

(75)
$$\phi_\delta|_{x=0} = \phi_g(t) \qquad \text{and} \qquad \phi_\delta|_{t=0} = 0,$$

and we want to pass to the limit $\delta \to 0$.

Let

$$\Psi_\delta'(\xi) := \frac{h(\xi)}{D(\xi)(\xi+\delta)}, \quad \xi \geq 0, \tag{76}$$

and

$$\Psi_\delta(\phi) := \int_0^\phi \frac{1}{\xi+\delta}\left(\frac{h(\xi)}{D(\xi)} - \frac{h(0)}{D(0)}\right)d\xi + \frac{h(0)}{D(0)}\log(\phi+\delta) \quad \text{for} \quad \phi \geq 0. \tag{77}$$

It should be noted that $\Psi_\delta(0) = \frac{h(0)}{D(0)}\log\delta < 0$, which would cause some complications.

THEOREM 15. *Let $\alpha > 0$, C_0, and A_0 be positive constants and let*

$$\phi_g \in H^1(0,T), \quad C_0 t^\alpha \leq \phi_g \leq A_0 \quad \textit{and} \quad \log\phi_g \in L^2(0,T). \tag{78}$$

Then problem (1)–(3) *has at least one weak solution $\phi \in H^1((0,T)\times(0,L))$ such that $\sqrt{\phi}\,\partial_x\left(e^{-m\phi}\partial_t\phi\right) \in L^2((0,T)\times(0,L))$ and $C_0 t^\alpha \leq \phi \leq A_0$.*

Proof.

Step 1 (a priori estimates uniform in δ). We test (74) by $\Psi_\delta(\phi_\delta) - \Psi_\delta(\phi_g)$ and get

$$\begin{aligned}\int_0^t\int_0^L \partial_t\phi_\delta\,\Psi_\delta(\phi_\delta)\,dx\,d\tau + \int_0^t\int_0^L \frac{h(\phi_\delta)}{\phi_\delta+\delta}\left(\partial_x\phi_\delta\right)^2 dx\,d\tau \\ + \frac{1}{B}\int_0^t\int_0^L D(\phi_\delta)(\phi_\delta+\delta)\,\partial_t\left(h(\phi_\delta)\,\partial_x\phi_\delta\right)\frac{h(\phi_\delta)\,\partial_x\phi_\delta}{D(\phi_\delta)(\phi_\delta+\delta)}\,dx\,d\tau \\ = \int_0^t\int_0^L \partial_t\phi_\delta\,\Psi_\delta(\phi_g)\,dx\,d\tau.\end{aligned}$$

This yields

$$\begin{aligned}\int_0^L\left(\int_0^{\phi_\delta(t)}\Psi_\delta(\xi)\,d\xi + \frac{1}{2B}\left(h(\phi_\delta)\,\partial_x\phi_\delta\right)^2\right)dx + \int_0^t\int_0^L \frac{h(\phi_\delta)}{\phi_\delta+\delta}\left(\partial_x\phi_\delta\right)^2 dx\,d\tau \\ = \int_0^t\int_0^L \partial_t\phi_\delta\,\Psi_\delta(\phi_g)\,dx\,d\tau.\end{aligned} \tag{79}$$

In order to get a useful estimate we should find a bound for the first term on the left-hand side of (79). First we note that $\int_0^{\phi_\delta}\int_0^\xi \frac{1}{\eta+\delta}\left(\frac{h(\eta)}{D(\eta)} - \frac{h(0)}{D(0)}\right)d\eta\,d\xi$ defines a continuous function of ϕ_δ. Since ϕ_δ takes values between 0 and A_0, it is bounded independently of δ. Hence

(80)
$$\left|\int_0^L \int_0^{\phi_\delta(t)} \Psi_\delta(\xi)\, d\xi\, dx\right| \le \int_0^L \frac{h(0)}{D(0)} \left|\{\phi_\delta+\delta\}\log\{\phi_\delta+\delta\} - \phi_\delta - \delta\log\delta\right| dx + C.$$

Next $(\phi_\delta(t)+\delta)\log(\phi_\delta(t)+\delta) - \phi(t) - \delta\log\delta$ takes value zero at $t=0$. It is a continuous function of ϕ_δ. Obviously $|(\phi(t)+\delta)\log(\phi(t)+\delta) - \phi(t) - \delta\log\delta| \le \max\{1-\delta+\delta\log\delta,\ (A_0+\delta)\log(A_0+\delta) - A_0 - \delta\log\delta\}$, and it is uniformly bounded with respect to δ.

With (80), (79) leads to

$$\int_0^t \int_0^L \frac{h(\phi_\delta)}{\phi_\delta+\delta} \left(\partial_x\phi_\delta\right)^2 dx\, d\tau \le C + \left|\int_0^t \int_0^L \partial_t\phi_\delta\, \Psi_\delta(\phi_g)\, dx\, d\tau\right|. \tag{81}$$

Next we test (74) by $h(\phi_\delta)\partial_t\phi_\delta - h(\phi_g)\partial_t\phi_g$ and get

$$\int_0^t \int_0^L h(\phi_\delta)\left(\partial_\tau\phi_\delta\right)^2 dx\, d\tau + \int_0^t \int_0^L D(\phi_\delta)\partial_x\phi_\delta\partial_x\left(h(\phi_\delta)\partial_\tau\phi_\delta\right) dx\, d\tau$$
$$+ \frac{1}{B}\int_0^t \int_0^L D(\phi_\delta)(\phi_\delta+\delta)\left(\partial_x\left(h(\phi_\delta)\partial_\tau\phi_\delta\right)\right)^2 dx\, d\tau = \int_0^t \int_0^L \partial_t\phi_\delta\, h(\phi_g)\partial_\tau\phi_g\, dx\, d\tau,$$

and from this

$$\int_0^t \int_0^L h(\phi_\delta)\left(\partial_\tau\phi_\delta\right)^2 dx\, d\tau + \frac{1}{B}\int_0^t \int_0^L D(\phi_\delta)(\phi_\delta+\delta)\left(\partial_x\left(h(\phi_\delta)\partial_\tau\phi_\delta\right)\right)^2 dx\, d\tau$$
$$\le B\int_0^t \int_0^L \frac{D(\phi_\delta)}{\phi_\delta+\delta}\left(\partial_x\phi_\delta\right)^2 dx\, d\tau + \int_0^t \int_0^L \frac{h^2(\phi_g)}{h(\phi_\delta)}\left(\partial_\tau\phi_g\right)^2 dx\, d\tau. \tag{82}$$

Let $h_{\min} = e^{-mA_0}$. Then inserting (81) into (82) yields

$$\int_0^t \int_0^L h(\phi_\delta)\left(\partial_\tau\phi_\delta\right)^2 dx\, d\tau + \frac{1}{B}\int_0^t \int_0^L D(\phi_\delta)(\phi_\delta+\delta)\left(\partial_x\left(h(\phi_\delta)\partial_\tau\phi_\delta\right)\right)^2 dx\, d\tau$$
$$\le C + \frac{BD_r}{h_{\min}}\left|\int_0^t \int_0^L \partial_\tau\phi_\delta\Psi_\delta(\phi_g)\, dx\, d\tau\right| + \int_0^t \int_0^L \frac{h^2(\phi_g)}{h(\phi_\delta)}\left(\partial_\tau\phi_g\right)^2 dx\, d\tau \le C$$
$$+ \frac{1}{2}\int_0^t \int_0^L h(\phi_\delta)\left(\partial_\tau\phi_\delta\right)^2 dx\, d\tau + \frac{B^2\left(D_r\right)^2}{2h_{\min}^3}\left\|\Psi_\delta(\phi_g)\right\|^2_{L^2((0,t)\times(0,L))}$$
$$+ \frac{1}{h_{\min}}\int_0^t \int_0^L \left(\partial_\tau\phi_g\right)^2 dx\, d\tau.$$

Step 2 (weak and strong compactness). From the above a priori estimate and assumptions (78) on ϕ_g, we conclude that

$$\left\|\partial_t\phi_\delta\right\|_{L^2((0,T)\times(0,L))} + \left\|\frac{1}{\sqrt{\phi_\delta+\delta}}\,\partial_x\phi_\delta\right\|_{L^2((0,T)\times(0,L))} \le C, \tag{83}$$

$$\left\|\sqrt{\phi_\delta+\delta}\,\partial_x\left(h(\phi_\delta)\partial_t\phi_\delta\right)\right\|_{L^2((0,T)\times(0,L))} \le C. \tag{84}$$

Hence there are a $\phi \in H^1((0,T)\times(0,L))$ and a subsequence $\{\phi_\delta\}$, denoted by the same subscripts, such that

$$\phi_\delta \to \phi \quad \text{strongly in } L^2((0,T)\times(0,L)) \text{ and a.e. on } \quad (0,T)\times(0,L), \tag{85}$$

$$\partial_t \phi_\delta \rightharpoonup \partial_t \phi \quad \text{weakly in } \quad L^2((0,T)\times(0,L)), \tag{86}$$

$$\partial_x \phi_\delta \rightharpoonup \partial_x \phi \quad \text{weakly in } \quad L^2((0,T)\times(0,L)). \tag{87}$$

With the part of the flux containing the second order operator, the situation is more complicated. Obviously, there is $F \in L^2((0,T)\times(0,L))$ such that

$$\sqrt{\phi_\delta+\delta}\,\partial_{xt}\int_0^{\phi_\delta} h(\xi)\,d\xi \rightharpoonup F \quad \text{weakly in } \quad L^2((0,T)\times(0,L)). \tag{88}$$

Using the lower bound $\phi_\delta \geq C_0 t^\alpha$, we get from the estimate (84) and convergence (85) that

$$\partial_{xt}\int_0^{\phi_\delta} h(\xi)\,d\xi \rightharpoonup \partial_{xt}\int_0^{\phi} h(\xi)\,d\xi \quad \text{weakly in } \quad L^2((0,T)\times(0,L)). \tag{89}$$

The convergences (85) and (89) imply that F in (88) is given by $F = \sqrt{\phi}\,\partial_{xt}\int_0^{\phi} h(\xi)\,d\xi$.

Step 3 (passing to the limit). Consequently for every $g \in L^2(0,T;V)$ we have

$$\int_0^T\int_0^L \partial_t\phi_\delta\, g\,dx\,dt \to \int_0^T\int_0^L \partial_t\phi\, g\,dx\,dt \quad \text{for } \delta\to 0, \tag{90}$$

$$\int_0^T\int_0^L D(\phi_\delta)\,\partial_x\phi_\delta\,\partial_x g\,dx\,dt \to \int_0^T\int_0^L D(\phi)\,\partial_x\phi\,\partial_x g\,dx\,dt \quad \text{for } \delta\to 0, \tag{91}$$

$$\begin{aligned}&\int_0^T\int_0^L \frac{D(\phi_\delta)}{B}(\phi_\delta+\delta)\,\partial_x\left(h(\phi_\delta)\,\partial_t\phi_\delta\right)\partial_x g\,dx\,dt\\ &\qquad\to \int_0^T\int_0^L \frac{D(\phi)}{B}\phi\,\partial_x\left(h(\phi)\,\partial_t\phi\right)\partial_x g\,dx\,dt \quad \text{for } \delta\to 0.\end{aligned} \tag{92}$$

Furthermore, for every $\zeta \in C^\infty([0,L]\times[0,T])$, such that $\zeta(L,t)=0$ on $[0,T]$, we have

$$\int_0^T \phi|_{x=0}\zeta|_{x=0}\,dt = \int_0^T \phi_g(t)\zeta|_{x=0}\,dt - \int_0^T\int_0^L \partial_x\left((\phi-\phi_\delta)\zeta\right)dx\,dt,$$

and, using the convergences (85) and (87), we obtain that the trace of ϕ at $x=0$ satisfies the boundary condition (26). Hence, we conclude that ϕ satisfies the system (1)–(3). □

Acknowledgment. The authors are grateful to the (anonymous) referees for their careful reading of the manuscript and helpful remarks.

REFERENCES

[1] T. Alfrey, E. F. Gurnee, and W. G. Lloyd, *Diffusion in glassy polymers*, J. Polym. Sci. Part C, 12 (1966), pp. 249–261.

[2] A. S. Argon, R. E. Cohen, and A. C. Patel, *A mechanistic model of case-*II *diffusion of a diluent into a glassy polymer*, Polymer, 40 (1999), pp. 6991–7012.

[3] G. I. Barenblatt, M. Bertsch, R. Dal Passo, and M. Ughi, *A degenerate pseudoparabolic regularization of a nonlinear forward-backward heat equation arising in the theory of heat and mass exchange in stably stratified turbulent shear flow*, SIAM J. Math. Anal., 24 (1993), pp. 1414–1439.

[4] A. Beliaev, *Homogenization of two-phase flows in porous media with hysteresis in the capillary relation*, European J. Appl. Math., 14 (2003), pp. 61–84.
[5] A. Yu. Beliaev and S. M. Hassanizadeh, *A theoretical model of hysteresis and dynamic effects in the capillary relation for two-phase flow in porous media*, Transp. Porous Media, 43 (2001), pp. 487–510.
[6] R. B. Bird, R. C. Amstrong, and O. Hassager, *Dynamics of Polymeric Liquids, Volume* 1: *Fluid Mechanics*, 2nd ed., John Wiley, New York, 1987.
[7] C. Cuesta and J. Hulshof, *A model problem for groundwater flow with dynamic capillary pressure: Stability of travelling waves*, Nonlinear Anal., 52 (2003), pp. 1199–1218.
[8] C. Cuesta, C. J. van Duijn, and J. Hulshof, *Infiltration in porous media with dynamic capillary pressure: Travelling waves*, European J. Appl. Math., 11 (2000), pp. 381–397.
[9] W. P. Düll, *Some qualitative properties of solutions to a pseudoparabolic equation modeling solvent uptake in polymeric solids*, Comm. Partial Differential Equations, 31 (2006), pp. 1117–1138.
[10] C. J. van Duijn, L. A. Peletier, and I. S. Pop, *A new class of entropy solutions of the Buckley–Leverett equation*, SIAM J. Math. Anal., 39 (2007), pp. 507–536.
[11] L. C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, AMS, Providence, RI, 1998.
[12] L. C. Evans, *A survey of entropy methods for partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 41 (2004), pp. 409–438.
[13] P. D. Gamson, B. B. Beamish, and D. P. Johnson, *Coal microstructure and micropermeability and their effects on natural-gas recovery*, Fuel, 72 (1993), pp. 87–99.
[14] S. M. Hassanizadeh and W. G. Gray, *Thermodynamic basis of capillary pressure in porous media*, Water Resources Res., 29 (1993), pp. 3389–3405.
[15] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, 1993.
[16] C. Y. Hui, K. C. Wu, R. C. Lasky, and E. J. Kramer, *Case* II *diffusion in polymers* I: *Transient swelling*, J. Appl. Phys., 61 (1987), pp. 5129–5136.
[17] C. Y. Hui, K. C. Wu, R. C. Lasky, and E. J. Kramer, *Case* II *diffusion in polymers* II: *Steady-state front motion*, J. Appl. Phys., 61 (1987), pp. 5137–5149.
[18] J. Hulshof and J. R. King, *Analysis of a Darcy flow model with a dynamic pressure saturation relation*, SIAM J. Appl. Math., 59 (1998), pp. 318–346.
[19] D. Jou, J. Casas-Vasquez, and G. Lebon, *Extended Irreversible Thermodynamics*, Springer-Verlag, Heidelberg, 2001.
[20] J. R. King and C. M. Cuesta, *Small- and waiting-time behavior of a Darcy flow model with a dynamic pressure saturation relation*, SIAM J. Appl. Math., 66 (2006), pp. 1482–1511.
[21] L. D. Landau and E. M. Lifshitz, *Fluid Mechanics*, Pergamon Press, Oxford, London, Paris, Frankfurt, 1959.
[22] C. M. Marle, *On macroscopic equations governing multiphase flow with diffusion and chemical reactions in porous media*, Internat. J. Engrg. Sci., 20 (1982), pp. 643–662.
[23] S. Mazumder, *Dynamics of CO_2 in Coal as a Reservoir*, Doctoral thesis, TU-Delft Civil Engineering and Geosciences, Delft, The Netherlands, 2007.
[24] A. Mikelić and R. Robert, *On the equations describing a relaxation toward a statistical equilibrium state in the two-dimensional perfect fluid dynamics*, SIAM J. Math. Anal., 29 (1998), pp. 1238–1255.
[25] A. Novick-Cohen and R. L. Pego, *Stable patterns in a viscous diffusion equation*, Trans. Amer. Math. Soc., 324 (1991), pp. 331–351.
[26] V. Padron, *Sobolev regularization of a nonlinear ill-posed parabolic problem as a model for aggregating populations*, Comm. Partial Differential Equations, 23 (1998), pp. 457–486.
[27] D. R. Pavone, *Macroscopic equations derived from space averaging for immiscible two-phase flow in porous media*, Rev. Inst. Français Pétrole, 44 (1989), pp. 29–41.
[28] D. R. Pavone, *A Darcy's law extension and a new capillary pressure equation for two-phase flow in porous media*, in Proceedings of the 65th Annual Technical Conference of SPE, paper SPE 20474, 1990.
[29] L. P. Ritger and N. A. Peppas, *Transport of penetrants in the macromolecular structure of coals* IV: *Models for analysis of dynamic penetrant transport*, Fuel, 66 (1987), pp. 815–826.
[30] L. P. Ritger and N. A. Peppas, *Transport of penetrants in the macromolecular structure of coals* VII: *Transport in thin coal sections*, Fuel, 66 (1987), pp. 1379–1388.
[31] N. L. Thomas and A. H. Windle, *A theory of case* II *diffusion*, Polymer, 23 (1982), pp. 529–542.
[32] D. W. Van Krevelen, *Coal: Typology-Physics-Chemistry-Constitution*, 3rd ed., Elsevier, Amsterdam, 1993.
[33] T. P. Witelski, *Traveling wave solutions for case* II *diffusion in polymers*, J. Polym. Sci. B: Polymer Physics, 34 (1996), pp. 141–150.

SIAM J. MATH. ANAL.
Vol. 40, No. 4, pp. 1692–1724

ASYMPTOTIC STABILITY OF THE STATIONARY SOLUTION FOR A HYPERBOLIC FREE BOUNDARY PROBLEM MODELING TUMOR GROWTH*

SHANGBIN CUI†

Abstract. In this paper we study the asymptotic behavior of solutions for a free boundary problem modeling the growth of tumors containing two species of cells: proliferating cells and quiescent cells. This tumor model was proposed by Pettet, Please, and McElwain [*Bull. Math. Biol.*, 63 (2001), pp. 231–257]. By using a functional approach and the C_0 semigroup theory, we prove that the unique stationary solution of this model ensured by the work of Cui and Friedman [*Trans. Amer. Math. Soc.*, 355 (2003), pp. 3537–3590] is locally asymptotically stable in certain function spaces. Key techniques used in the proof include an improvement of the linear estimate obtained by the work of Chen, Cui, and Friedman [*Trans. Amer. Math. Soc.*, 357 (2005), pp. 4771–4804] and a similarity transformation.

Key words. free boundary problem, hyperbolic equations, tumor growth, stationary solution, asymptotic stability

AMS subject classifications. 35C10, 35Q80, 92C15

DOI. 10.1137/080717778

1. Introduction.

During the past thirty years, an increasing number of free boundary problems of partial differential equations have been proposed by groups of researchers to model the growth of various in vivo and in vitro tumors; cf. the reviewing articles [1], [15], [16], and [22] and the references cited therein. Such free boundary problems usually contain one or more reaction-diffusion equations describing the distribution of nutrient and inhibitory materials, and several first-order nonlinear partial differential equations or nonlinear conservation laws with source terms describing the evolution and movement of various tumor cells (proliferating cells, quiescent cells, and dead cells). Rigorous analysis of such tumor models is evidently a significant topic of research and has drawn great attention during the past few years. The main concern of this topic is the dynamics or the long-term behavior of the solutions of such free boundary problems.

Based on the applications of the well-established theories of elliptic and parabolic partial differential equations, parabolic differential equations in Banach spaces (i.e., differential equations in Banach spaces that are treatable with the analytic semigroup theory), and the bifurcation theory, rigorous analysis of models for the growth of tumors containing only one species of tumor cells has achieved great success; cf. [2], [3], [6], [7], [9], [10], [11], [17], [18], [19], [23], [24], and the references cited therein. As far as models for tumors containing more than one species of tumor cells are concerned, however, the results are much less. This is caused by the fact that such tumor models are much more difficult to analyze because they contain nonlinear conservation laws whose dynamical behavior is very hard to grasp.

In this paper we study the following free boundary problem modeling the growth of an in vitro tumor containing two species of cells—proliferating cells and quiescent

*Received by the editors March 5, 2008; accepted for publication (in revised form) July 23, 2008; published electronically November 26, 2008. This work is supported by the China National Natural Science Foundation under grant number 10471157 and funds in Sun Yat-Sen University.
http://www.siam.org/journals/sima/40-4/71777.html
†Institute of Mathematics, Sun Yat-Sen University, Guangzhou, Guangdong 510275, People's Republic of China (cuisb3@yahoo.com.cn).

cells:

$$\nabla^2 C = F(C) \quad \text{for } x \in \Omega(t), \ t \geq 0, \tag{1.1}$$

$$C = C_0 \quad \text{for } x \in \partial\Omega(t), \ t \geq 0, \tag{1.2}$$

$$\frac{\partial P}{\partial t} + \nabla \cdot (\vec{u}P) = \big[K_B(C) - K_Q(C)\big]P + K_P(C)Q \quad \text{for } x \in \Omega(t), \ t \geq 0, \tag{1.3}$$

$$\frac{\partial Q}{\partial t} + \nabla \cdot (\vec{u}Q) = K_Q(C)P - \big[K_D(C) + K_P(C)\big]Q \quad \text{for } x \in \Omega(t), \ t \geq 0, \tag{1.4}$$

$$P + Q = N \quad \text{for } x \in \Omega(t), \ t \geq 0, \tag{1.5}$$

$$\frac{dR}{dt} = \vec{u} \cdot \vec{\nu} \quad \text{for } x \in \partial\Omega(t), \ t \geq 0. \tag{1.6}$$

Here C denotes the concentration of nutrient (with all nutrient materials regarded as one species), P and Q denote the densities of proliferating cells and quiescent cells, respectively, whose mixture makes up the tumor tissue and has a constant density N, $\vec{u}$ denotes the velocity of the cell movement, R denotes the radius of the tumor, $\Omega(t) = \{x \in \mathbb{R}^3 : r = |x| < R(t)\}$ is the domain occupied by the tumor at time t, and $\vec{\nu}$ is the unit outward normal of $\partial\Omega(t)$. Besides, C_0 is a positive constant reflecting the constant nutrient supply that the tumor receives from its surface, $F(C)$ is the nutrient consumption rate function, and $K_B(C)$, $K_D(C)$, $K_P(C)$, and $K_Q(C)$ are the birth rate of proliferating cells, the death rate of quiescent cells, the transferring rate of proliferating cells to quiescent cells, and the transferring rate of quiescent cells to proliferating cells, respectively. We shall only consider radially symmetric solutions of the above problem, so that C, P, and Q are functions of the radial space variable $r = |x|$ and the time variable t, and $\vec{u} = u(r,t)r^{-1}x$, where u is a scalar function. Note that in the above problem, the unknowns are C, P, Q, u, and R, and all of the rest of the constants and functions are given.

The above tumor model was proposed by Pettet et al. in the literature [21]. Its global well-posedness has been established by Cui and Friedman in [12]. A challenging task concerning this free boundary problem is the study of the asymptotic behavior of its solutions as time goes to infinity. For the corresponding model of the growth of tumors with one species of cells, it is known that there exists a unique stationary solution and that all time-dependent solutions converge to it as time goes to infinity, or, in other words, this unique stationary solution is globally asymptotically stable; cf. [7] and [18]. Since the above problem is a natural extension of such one species tumor model to the two species case, we are naturally led to the conjecture that a similar result holds for it. The advancement of the study toward this goal is as follows. In [13], Cui and Friedman proved that the problem (1.1)–(1.6) has a unique stationary solution. In [5], Chen, Cui, and Friedman further proved that this stationary solution is linearly asymptotically stable, namely, the trivial solution of the linearization of (1.1)–(1.6) at the stationary solution is asymptotically stable. However, this last-mentioned result does not imply that the stationary solution of (1.1)–(1.6) is asymptotically stable. In fact, the asymptotic stability of this stationary solution of (1.1)–(1.6), which is one of a number of interesting open problems raised in [15] and [16], is a very difficult problem due to a number of reasons. Firstly, since the problem (1.1)–(1.6) is of the hyperbolic type, not of the parabolic type, there is not a well-developed geometric theory like that for parabolic problems used in [6], [7], [9], [10], [11], [23], [24] to study this problem. Secondly, the stationary solution of this problem does not have an explicit expression, and, in particular, it possibly possesses a singularity at the point $r = 0$ (see Lemma 3.1 below for this point). It follows that

the method of linearized stability as in [3] and [17] (tempted by [5]) does not work. Thirdly, the commonly used characteristic method for hyperbolic conservation laws also does not apply to this problem, because all of its characteristic curves approach to the same point $r = 0$ as $t \to \infty$, and, consequently, analysis along characteristic curves does not give us much information of the asymptotic behavior of the solution. Finally, it is our experience that the other commonly used methods for investigating the asymptotic behavior of the solutions of partial differential equations, such as upper and lower solutions and so on (as either explicitly or implicitly discussed in [5]), also do not work. In order to solve this problem, we have to use some new ideas and develop some new techniques.

In this paper we shall prove that the unique stationary solution of (1.1)–(1.6) ensured by [13] is locally asymptotically stable. Recall that conditions given in [13], which ensure that (1.1)–(1.6) has a unique stationary solution are as follows:

$$\text{(1.7)}\quad F(C),\ K_B(C),\ K_D(C),\ K_P(C) \text{ and } K_Q(C) \text{ are analytic for } 0 \le C \le C_0;$$

$$\text{(1.8)}\quad F(0) = 0, \quad F'(C) > 0 \quad \text{for } 0 \le C \le C_0;$$

$$\text{(1.9)}\quad \begin{cases} K_B'(C) > 0 \text{ and } K_D'(C) < 0 \text{ for } 0 \le C \le C_0,\ K_B(0) = 0 \text{ and } K_D(C_0) = 0; \\ K_P(C) \text{ and } K_Q(C) \text{ satisfy the same conditions as } K_B(C) \text{ and } K_D(C), \\ \qquad \text{respectively}; \\ K_B'(C) + K_D'(C) > 0 \text{ for } 0 \le C \le C_0. \end{cases}$$

The main result of this paper is the following.

THEOREM 1.1. *Assume that the conditions* (1.7)–(1.9) *are satisfied. Let* $(C_*, P_*, Q_*, \vec{u}_*, R_*)$ *be the unique stationary solution of the problem* (1.1)–(1.6), *and let* $(C, P, Q, \vec{u}, R)$ *be a time-dependent solution of it such that* $P|_{t=0} = P_0$, $Q|_{t=0} = Q_0$, *and* $R|_{t=0} = R_0$, *where* P_0, Q_0, *and* R_0 *are given initial data satisfying* $0 \le P_0 \le N$, $0 \le Q_0 \le N$, *and* $P_0 + Q_0 = N$. *Then there exist positive constants* μ, ε, *and* K *such that if* P_0, Q_0, *and* R_0 *satisfy*

$$\max_{0\le r\le 1} \{|P_0(rR_0) - P_*(rR_*)| + |Q_0(rR_0) - Q_*(rR_*)|\} < \varepsilon,$$

$$\sup_{0<r<1} r(1-r) \left\{ \left| \frac{d}{dr} \left(P_0(rR_0) - P_*(rR_*)\right) \right| + \left| \frac{d}{dr} \left(Q_0(rR_0) - Q_*(rR_*)\right) \right| \right\} < \varepsilon,$$

and $|R_0 - R_*| < \varepsilon$; *then, for all* $t \ge 0$, *we have*

$$\max_{0\le r\le 1} \{|P(rR(t),t) - P_*(rR_*)| + |Q(rR(t),t) - Q_*(rR_*)|\} < K\varepsilon e^{-\mu t},$$

$$\sup_{0<r<1} r(1-r) \left\{ \left| \frac{\partial}{\partial r} \left(P(rR(t),t) - P_*(rR_*)\right) \right| + \left| \frac{\partial}{\partial r} \left(Q(rR(t),t) - Q_*(rR_*)\right) \right| \right\} < K\varepsilon e^{-\mu t},$$

and $|R(t) - R_*| < K\varepsilon e^{-\mu t}$.

We shall use a functional approach to prove the above theorem. More precisely, we shall first reduce the problem (1.1)–(1.6) into a differential equation for the unknown $U = (p, z)$ in the Banach space $X = C[0,1] \times \mathbb{R}$, where $p = p(r,t) = P(rR(t),t)$ and $z = z(t) = \log R(t)$. The reduced equation is of the hyperbolic type in the sense of Pazy [20] and is quasi-linear. We next use the Banach fixed point theorem to prove that, for any $U_0 = (p_0, z_0)$ sufficiently closed to the stationary point $U_* = (p_*, z_*)$, where $p_* = p_*(r) = P_*(rR_*)$ and $z_* = \log R_*$, this differential equation imposed, with the initial condition $U|_{t=0} = U_0$ and the decay estimate $\sup_{t\ge 0} e^{\mu t} \|U(t) - U_*\|_{X_0} < \infty$, where X_0 is a subspace of X, has a unique solution in the space $C([0,\infty), X_0)$

(endowed with the norm $|||U||| = \sup_{t\geq 0} e^{\mu t}\|U(t)\|_{X_0}$). To attain this goal we shall use some abstract results for hyperbolic differential equations in Banach spaces presented in [20]. In particular, a family of evolution systems for the linear equations related to the semilinearization of the reduced equation are obtained and applied to convert the semilinearized equations into integral equations. The main difficult and key step in the proof of Theorem 1.1 is the establishment of a uniform decay estimate for the family of evolution systems. In order to obtain such an estimate, we first use a localization technique to get an improvement of the linear estimate established in [5], removing the singularities at $r = 0$ contained in that estimate, and next use a special technique—similarity transformation—to extend this improved linear estimate to the family of evolution systems mentioned above. The similarity transformation technique is the core of this paper. It enables us to transform a transport equation of the form $\partial u/\partial t + w(r,t)\partial u/\partial r = f(r,t)$, where $0 < r < 1$ and $w(0,t) = w(1,t) = 0$, into a similar equation of the form $\partial\widetilde{u}/\partial t + w_*(r)\partial\widetilde{u}/\partial r = \widetilde{f}(r,t)$, where $w_*(0) = w_*(1) = 0$. See section 5 for details of this transformation.

An interesting question raised by one of the referees is: can the above result be extended to the radially nonsymmetric version of the problem (1.1)–(1.6)? Up to now we do not have an answer to this question. A basic difficulty encountered in answering this question is that $P_*, Q_*, \vec{u}_*$ are possibly singular at the center point $r = 0$. More precisely, we know only that P_* is continuous at $r = 0$, but we do not know if it is smooth at $r = 0$ (in fact, we even do not know if it is differentiable at $r = 0$). This possibly existing singularity of the stationary solution induces many obstacles in the analysis of its asymptotic stability for nonradial perturbations. Our conjecture is that $P_*, Q_*, \vec{u}_*$ are smooth for all $r \leq R_*$, and a similar result as that in [11] proved for the one-species model also holds for the radially nonsymmetric version of the problem (1.1)–(1.6).

The structure of the rest part is as follows. In the following section we reduce the problem (1.1)–(1.6) into a differential equation in the Banach space $X = C[0,1]\times\mathbb{R}$. In section 3 we summarize some basic properties of the stationary solution. In section 4 we study a linear equation obtained from semilinearizing the reduced equation and prove that its solution operator is an evolution system so that the semilinearized equation can be converted into an integral equation. Section 5 aims at developing the similarity transformation technique mentioned above. In section 6 we first derive an improvement of the linear estimate established in [5] and next use the similarity transformation technique to extend this estimate to the evolution systems obtained in section 4. After these preparations, in the last section we use the Banach fixed point theorem to prove Theorem 1.1.

Throughout this paper the notation "$'$" denotes both the ordinary derivatives of functions in $\mathbb{R}$ and the Fréchet derivatives of mappings between Banach spaces.

Finally, to end this introduction, we would like to refer the reader to see [4], [8], and [14] for other work related with the problem (1.1)–(1.6).

2. Reduction of the problem. In this section we reduce the system of equations (1.1)–(1.6) into a differential equation in the Banach space $X = C[0,1]\times\mathbb{R}$.

We first note that, by summing up (1.3), (1.4), and using (1.5), we get the following equation:

$$\nabla\cdot\vec{u} = \frac{1}{N}\big[K_B(C)P - K_D(C)Q\big]. \tag{2.1}$$

Conversely, from (1.3), (1.5), and (2.1) we immediately obtain (1.4). Hence, the two groups of equations (1.3), (1.4), (1.5), and (1.3), (1.5), (2.1) are equivalent. We recall that $\vec{u} = u(r,t)r^{-1}x$ for some scalar function u.

By rescaling the space and time variables, setting

$$\bar{c}(\bar{r},t)=\frac{C\left(\bar{r}e^{z(t)},t\right)}{C_0},\quad \bar{p}(\bar{r},t)=\frac{P\left(\bar{r}e^{z(t)},t\right)}{N},$$
$$\bar{u}(\bar{r},t)=u\left(\bar{r}e^{z(t)},t\right)e^{-z(t)},\quad R(t)=e^{z(t)},$$

and using the equivalence of (1.3), (1.4), and (1.5) with (1.3), (1.5), and (2.1), we see that the problem (1.1)–(1.6) can be reformulated into the following problem (for the simplicity of the notation we omit all bars):

$$\frac{\partial^2 c}{\partial r^2}+\frac{2}{r}\frac{\partial c}{\partial r}=e^{2z}F(c)\quad \text{for } 0<r\le 1,\ t\ge 0, \tag{2.2}$$

$$\frac{\partial c}{\partial r}\Big|_{r=0}=0,\quad c|_{r=1}=1\quad \text{for } t\ge 0, \tag{2.3}$$

$$\frac{\partial p}{\partial t}+[u(r,t)-ru(1,t)]\frac{\partial p}{\partial r}=K_P(c)+\big[K_M(c)-K_N(c)\big]p-K_M(c)p^2 \quad \text{for } 0\le r\le 1,\ t\ge 0, \tag{2.4}$$

$$\frac{\partial u}{\partial r}+\frac{2}{r}u=-K_D(c)+K_M(c)p\quad \text{for } 0<r\le 1 \text{ and } u|_{r=0}=0,\ t\ge 0, \tag{2.5}$$

$$\frac{dz}{dt}=u(1,t)\quad \text{for } t\ge 0, \tag{2.6}$$

where $K_M(c)=K_B(c)+K_D(c)$, $K_N(c)=K_P(c)+K_Q(c)$, and $F(c)$, $K_B(c)$, $K_D(c)$, $K_P(c)$, and $K_Q(c)$ are the rescaled forms of the corresponding functions appearing in (1.1)–(1.6).

Clearly, (2.2) and (2.3) can be solved to get c as a function of z. Thus, instead of $c(r,t)$, later on we use the notation $c(r,z(t))$, or simply $c(r,z)$, to denote the solution of (2.2) and (2.3). Similarly, (2.5) can be solved to get u as a functional of p and z. Thus, later on we use the notation $u_{p,z}$ to redenote u. By a simple computation, we have

$$u_{p,z}(r,t)=\frac{1}{r^2}\int_0^r[-K_D(c(\rho,z(t)))+K_M(c(\rho,z(t)))p(\rho,t)]\rho^2 d\rho \tag{2.7}$$

for $0<r\le 1$, $t\ge 0$, and $u_{p,z}(0,t)=0$ for $t\ge 0$. We also denote

$$w_{p,z}(r,t)=u_{p,z}(r,t)-ru_{p,z}(1,t). \tag{2.8}$$

Then (2.2)–(2.6) reduces into the following system of equations:

$$\begin{cases}\dfrac{\partial p}{\partial t}+w_{p,z}(r,t)\dfrac{\partial p}{\partial r}=f(r,p,z)\quad \text{for } 0\le r\le 1,\ t>0,\\[2mm] \dfrac{dz}{dt}=u_{p,z}(1,t)\quad \text{for } t>0,\end{cases} \tag{2.9}$$

where

$$f(r,p,z)=K_P(c(r,z))+\big[K_M(c(r,z))-K_N(c(r,z))\big]p-K_M(c(r,z))p^2.$$

In what follows we rewrite (2.9) as a differential equation in the Banach space $X=C[0,1]\times\mathbb{R}$. To this end we denote $X_0=C_V^1[0,1]\times\mathbb{R}$, where $C_V^1[0,1]$ is the function space

$$C_V^1[0,1]=\big\{p\in C[0,1]\cap C^1(0,1):\ r(1-r)p'(r)\in C[0,1]\big\}$$

endowed with the norm

$$\|p\|_{C_V^1[0,1]} = \max_{0\le r\le 1} |p(r)| + \sup_{0<r<1} |r(1-r)p'(r)| \quad \text{for } p \in C_V^1[0,1].$$

Clearly, endowed with the product norm, X_0 is a Banach space densely and continuously embedded into X. We introduce a mapping $\mathbb{F} : X_0 \to X$ as follows. First, for given $p \in C[0,1]$ and $z \in \mathbb{R}$, let $\mathcal{A}_0(p,z) : C_V^1[0,1] \to C(0,1)$ be the following linear operator: For any $q \in C_V^1[0,1]$,

$$\mathcal{A}_0(p,z)q(r) = -w_{p,z}(r)q'(r) \quad \text{for } 0 < r < 1.$$

Here and hereafter, $w_{p,z}(r)$ represents the function defined by similar formulations as in (2.7) and (2.8), with $p(r,t)$ and $z(t)$ there replaced by $p(r)$ and z, respectively. Later on we shall use the convention that, for a function $f \in C(0,1)$, if both limits $\lim_{r\to 0^+} f(r)$ and $\lim_{r\to 1^-} f(r)$ exist and are finite, then we write $f \in C[0,1]$. Furthermore, when we are concerned with the values of f at $r = 0$ and $r = 1$, we mean that $f(0) = \lim_{r\to 0^+} f(r)$ and $f(1) = \lim_{r\to 1^-} f(r)$. Using this convention, we see easily that, for any $p \in C[0,1]$ and $z \in \mathbb{R}$, we have $w_{p,z}(r)/r(1-r) \in C[0,1]$. It follows that, for any $q \in C_V^1[0,1]$, both limits $\lim_{r\to 0^+} w_{p,z}(r)q'(r)$ and $\lim_{r\to 1^-} w_{p,z}(r)q'(r)$ exist so that $\mathcal{A}_0(p,z)q \in C[0,1]$. It can also be easily seen that $\mathcal{A}_0(p,z) \in L(C_V^1[0,1], C[0,1])$, and

$$\|\mathcal{A}_0(p,z)\|_{L(C_V^1[0,1],C[0,1])} \le \sup_{0<r<1} \left| \frac{w_{p,z}(r)}{r(1-r)} \right|.$$

Next we denote by $\mathcal{F} : C[0,1] \times \mathbb{R} \to C[0,1]$ and $\mathcal{G} : C[0,1] \times \mathbb{R} \to \mathbb{R}$, respectively, the following nonlinear operators:

$$\mathcal{F}(p,z)(r) = f(r, p(r), c(r,z)),$$
$$\mathcal{G}(p,z) = \int_0^1 [-K_D(c(r,z)) + K_M(c(r,z))p(r)]r^2\,dr.$$

We now define $\mathbb{F} : X_0 \to X$ to be the following nonlinear operator:

$$\mathbb{F}(U) = \big(\mathcal{A}_0(p,z)p + \mathcal{F}(p,z), \mathcal{G}(p,z)\big) \quad \text{for } U = (p,z) \in X_0.$$

It is obvious that $\mathbb{F} \in C^\infty(X_0, X)$. Later on we shall also regard $\mathbb{F}$ as an unbounded nonlinear operator in X with domain X_0.

With the above notations, we see that (2.9) can be rewritten as the following differential equation in the Banach space X:

$$\frac{dU}{dt} = \mathbb{F}(U). \tag{2.10}$$

Here $U = U(t)$ represents an X_0-valued unknown function for $t \ge 0$, and the left-hand side denotes the Fréchet derivative of $U = U(t)$ regarded as a mapping from $[0,\infty)$ to the X space.

The equation (2.10) has a quasi-linear structure. To see this we define $\mathbb{A}_0 : X \to L(X_0, X)$ and $\mathbb{F}_0 : X \to X$ as follows: For $U = (p,z) \in X$ and $V = (q,y) \in X_0$, we let

$$\mathbb{A}_0(U)V = \big(\mathcal{A}_0(p,z)q, 0\big) \quad \text{and} \quad \mathbb{F}_0(U) = \big(\mathcal{F}(p,z), \mathcal{G}(p,z)\big).$$

Then we have

$$\mathbb{F}(U) = \mathbb{A}_0(U)U + \mathbb{F}_0(U) \quad \text{for } U \in X_0.$$

Clearly, $\mathbb{A}_0 \in C^\infty(X, L(X_0, X))$ and $\mathbb{F}_0 \in C^\infty(X, X) \cap C^\infty(X_0, X_0)$. Hence, the desired assertion follows. Later on, for given $U \in X$, we shall also regard $\mathbb{A}_0(U)$ as an unbounded linear operator in X with domain X_0.

From [13] we know that under the assumptions (1.7)–(1.9), the problem (2.2)–(2.6) has a unique stationary solution which we denote as (c_*, p_*, u_*, z_*). By definition, $(c_*, p_*, u_*, z_*) = (c_*(r), p_*(r), u_*(r), z_*)$ $(0 \le r \le 1)$ is the solution of the following problem:

$$c_*'' + \frac{2}{r}c_*' = e^{2z_*}F(c_*) \quad \text{for } 0 < r \le 1, \tag{2.11}$$

$$c_*'(0) = 0, \quad c_*(1) = 1, \tag{2.12}$$

$$u_*p_*' = f(r, p_*, z_*) \quad \text{for } 0 \le r \le 1, \tag{2.13}$$

$$u_*' + \frac{2}{r}u_* = -K_D(c_*) + K_M(c_*)p_* \quad \text{for } 0 < r \le 1, \tag{2.14}$$

$$u_*(0) = 0, \quad u_*(1) = 0. \tag{2.15}$$

Let $U_* = (p_*, z_*)$. Then $U_* \in X_0$ (see Lemma 2.1 below), and, by the equivalence of the system (2.2)–(2.6) with the equation (2.10), it follows that U_* is the unique equilibrium of (2.10), i.e., $\mathbb{F}(U_*) = 0$ or $\mathbb{A}_0(U_*)U_* + \mathbb{F}_0(U_*) = 0$. Hence, to study the asymptotic stability of the stationary solution (c_*, p_*, u_*, z_*) of the problem (2.2)–(2.6), we need only to study the asymptotic stability of the stationary solution U_* of (2.10). For this purpose, we rewrite (2.10) into an equivalent equation for the difference $V = U - U_*$. Let $\mathbb{A} : X \to L(X_0, X)$ and $\mathbb{G} : X \to X$ be the following operators:

$$\begin{aligned}\mathbb{A}(V)W &= \mathbb{A}_0(U_* + V)W + [\mathbb{A}_0'(U_*)W]U_* + \mathbb{F}_0'(U_*)W \quad \text{for } V \in X,\ W \in X_0,\\ \mathbb{G}(V) &= [\mathbb{A}_0(U_* + V) - \mathbb{A}_0(U_*) - \mathbb{A}_0'(U_*)V]U_*\\ &\quad + [\mathbb{F}_0(U_* + V) - \mathbb{F}_0(U_*) - \mathbb{F}_0'(U_*)V] \quad \text{for } V \in X.\end{aligned}$$

Then clearly (2.10) is equivalent to the following equation for $V = U - U_*$:

$$\frac{dV}{dt} = \mathbb{A}(V)V + \mathbb{G}(V). \tag{2.16}$$

Later on we shall concentrate our attention on (2.16).

The above deduction leads to the following preliminary result.

LEMMA 2.1. *The system* (2.2)–(2.6) *is equivalent to* (2.16), *and the asymptotic stability of the stationary solution* (c_*, p_*, u_*, z_*) *of* (2.2)–(2.6) *is equivalent to the asymptotic stability of the trivial solution of* (2.16). *Moreover, the assertion of Theorem* 1.1 *is equivalent to the following assertion for* (2.16): *There exist positive constants* K, μ, *and* ε *such that if the solution* $V = V(t)$ *of* (2.16) *satisfies* $\|V(0)\|_{X_0} \le \varepsilon$, *then it also satisfies*

$$\|V(t)\|_{X_0} \le K\varepsilon e^{-\mu t} \quad \textit{for all } t \ge 0. \tag{2.17}$$ □

Since it has been proved (see [12]) that the initial value problem of the system (2.2)–(2.6) is globally well-posed for any C^1 initial data, by the equivalence of (2.2)–(2.6) with (2.16), it follows that, for any $V_0 \in X_0$, (2.16) has a unique solution

$V = V(t)$ for all $t \geq 0$ such that $V(0) = V_0$. Our analysis later on is to prove that (2.17) holds provided $\|V_0\|_{X_0} \leq \varepsilon$.

We note that $\mathbb{A} \in C^\infty(X, L(X_0, X))$, $\mathbb{G} \in C^\infty(X, X)$, and, by using the Taylor expansions up to second-order for Fréchet derivatives of $\mathbb{A}_0$ and $\mathbb{F}_0$, we have

$$\|\mathbb{G}(V)\|_X = O\left(\|V\|_X^2\right) \quad \text{as } \|V\|_X \to 0. \tag{2.18}$$

We also note that, by introducing an operator $\mathbb{B} : X \to X$ by

$$\mathbb{B}W = [\mathbb{A}_0'(U_*)W]U_* + \mathbb{F}_0'(U_*)W \quad \text{for } W \in X,$$

we have

$$\mathbb{A}(0) = \mathbb{F}'(U_*) = \mathbb{A}_0(U_*) + \mathbb{B} \quad \text{and} \quad \mathbb{A}(V) = \mathbb{A}_0(U_* + V) + \mathbb{B}.$$

Note that as an immediate consequence of the facts that $\mathbb{A}_0 \in C^\infty(X, L(X_0, X))$ and $\mathbb{F}_0 \in C^\infty(X, X)$, we have $\mathbb{B} \in L(X)$. Finally, we note that $[V \to \mathbb{A}_0(U_* + V)] \in C^\infty(X, L(X_0, X))$. These facts will play an important role in later analysis.

A simple computation shows that if we denote

$$\begin{aligned} a(r) &= K_M(c_*(r)) - K_N(c_*(r)) - 2K_M(c_*(r))p_*(r), \\ b(r) &= \left\{K_P'(c_*(r)) + [K_M'(c_*(r)) - K_N'(c_*(r))]p_*(r) - K_M'(c_*(r))p_*^2(r)\right\} c_z(r) \end{aligned} \tag{2.19}$$

$$+ rp_*'(r)\left[\int_0^1 g_c(\rho)c_z(\rho)\rho^2 d\rho - \frac{1}{r^3}\int_0^r g_c(\rho)c_z(\rho)\rho^2 d\rho\right], \tag{2.20}$$

$$\mathcal{B}q(r) = rp_*'(r)\left[\int_0^1 g_p(\rho)q(\rho)\rho^2 d\rho - \frac{1}{r^3}\int_0^r g_p(\rho)q(\rho)\rho^2 d\rho\right], \tag{2.21}$$

$$\mathcal{F}(q) = \int_0^1 g_p(\rho)q(\rho)\rho^2 d\rho, \tag{2.22}$$

$$\kappa = \int_0^1 g_c(\rho)c_z(\rho)\rho^2 d\rho, \tag{2.23}$$

where

$$g_p(r) = K_M(c_*(r)), \quad g_c(r) = -K_D'(c_*(r)) + K_M'(c_*(r))p_*(r), \quad c_z(r) = \frac{\partial c}{\partial z}(r, z_*),$$

then we have

$$\mathbb{B} = \begin{pmatrix} a(r) + \mathcal{B} & b(r) \\ \mathcal{F} & \kappa \end{pmatrix}. \tag{2.24}$$

Here and hereafter, when we write $\mathbb{M} = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}$ for bounded linear operators $M_{11} \in L(X_1, Y_1)$, $M_{12} \in L(X_2, Y_1)$, $M_{21} \in L(X_1, Y_2)$, and $M_{22} \in L(X_2, Y_2)$, where X_1, X_2, Y_1, and Y_2 are Banach spaces, we mean that $\mathbb{M}$ is the bounded linear operator from $X_1 \times X_2$ to $Y_1 \times Y_2$ defined by $\mathbb{M}(x_1, x_2) = (M_{11}x_1 + M_{12}x_2, M_{21}x_1 + M_{22}x_2)$ for $(x_1, x_2) \in X_1 \times X_2$. Using this notation we see that

$$\mathbb{A}_0(U_*) = \begin{pmatrix} \mathcal{L}_0 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbb{A}_0(U_* + V) = \begin{pmatrix} \mathcal{L}_V & 0 \\ 0 & 0 \end{pmatrix}, \tag{2.25}$$

where $\mathcal{L}_0 = \mathcal{A}_0(p_*, z_*)$ and $\mathcal{L}_V = \mathcal{A}_0(p_* + \varphi, z_* + \zeta)$ for $V = (\varphi, \zeta) \in X$. Finally, we recall that $a(r) < 0$ for all $0 \leq r \leq 1$; see (2.7) in section 2 of [5].

3. Some basic facts. In this section we summarize some basic properties of the functions $c_*(r) = c(r, z_*)$, $c_z(r) = \frac{\partial c}{\partial z}(r, z_*)$, $p_*(r)$, and $u_*(r)$. These properties will play an important role in later discussions.

LEMMA 3.1. *We have the following assertions:*

(1) $c_*,\ c_z \in C^\infty[0,1]$, *and*

(3.1)
$$0 < c_*(0) \le c_*(r) \le 1 \quad \textit{for } 0 \le r \le 1, \quad c_*'(r) > 0 \quad \textit{for } 0 < r \le 1, \quad c_*'(0) = 0.$$

(2) $p_* \in C[0,1] \cap C^\infty(0,1]$,

$$0 < p_*(0) \le p_*(r) \le 1 \quad \textit{for } 0 \le r \le 1, \quad p_*'(r) > 0 \quad \textit{for } 0 < r \le 1, \tag{3.2}$$

and either $p_* \in C^1[0,1]$ *or there exists* $0 < \gamma < 1$ *such that* $\lim_{r\to 0^+} r^\gamma p_*'(r)$ *exists and is finite, so that* $r^\gamma p_*'(r) \in C[0,1]$. *Moreover,*

$$\lim_{r\to 0^+} r p_*'(r) = 0, \quad \lim_{r\to 0^+} r^2 p_*''(r) = 0, \quad \lim_{r\to 0^+} r^3 p_*'''(r) = 0. \tag{3.3}$$

(3) $u_* \in C^1[0,1] \cap C^\infty(0,1]$, *and there exist positive constants* C_1, C_2 *such that*

$$-C_1 r(1-r) \le u_*(r) \le -C_2 r(1-r) \quad \textit{for } \ 0 \le r \le 1. \tag{3.4}$$

Besides, either $u_* \in C^2[0,1]$ *or there exists* $0 < \gamma < 1$ *such that* $\lim_{r\to 0^+} r^\gamma u_*''(r)$ *exists and is finite, so that* $r^\gamma u_*''(r) \in C[0,1]$. *Moreover,*

$$\lim_{r\to 0^+} r u_*''(r) = 0, \quad \lim_{r\to 0^+} r^2 u_*'''(r) = 0. \tag{3.5}$$

Proof. The assertions that $c_*, c_z \in C^\infty[0,1]$ and relations in (3.1) are immediate. The assertions that $p_* \in C[0,1] \cap C^\infty(0,1]$, $u_* \in C^1[0,1] \cap C^\infty(0,1]$, and relations in (3.2) follow from Theorem 2.1 of [13], by which we also know that $u_*(r) < 0$ for $0 < r < 1$. The last assertion combined with the facts that $u_*'(0) < 0$ and $u_*'(1) > 0$ (see Theorem 7.1 of [13]) immediately yields (3.4). To prove (3.3) we compute

$$\lim_{r\to 0} u_*(r) p_*'(r) = K_P(c_*(0)) + [K_M(c_*(0)) - K_N(c_*(0))] p_*(0) - K_M(c_*(0)) p_*^2(0) = 0$$

(see (8.4) in section 8 of [13]), so that

$$\lim_{r\to 0} r p_*'(r) = \lim_{r\to 0} \frac{r}{u_*(r)} \cdot \lim_{r\to 0} u_*(r) p_*'(r) = 0,$$

and

$$\begin{aligned}\lim_{r\to 0} u_*(r) r p_*''(r) = \lim_{r\to 0} r\big\{ & K_P'(c_*(r)) + [K_M'(c_*(r)) - K_N'(c_*(r))] p_*(r) \\ & - K_M'(c_*(r)) p_*^2(r)\big\} c_*'(r) + \lim_{r\to 0}\{[K_M(c_*(r)) - K_N(c_*(r))] \\ & - 2K_M(c_*(r)) p_*(r)\} r p_*'(r) - \lim_{r\to 0} u_*'(r) r p_*'(r) = 0,\end{aligned}$$

so that

$$\lim_{r\to 0} r^2 p_*''(r) = \lim_{r\to 0} \frac{r}{u_*(r)} \cdot \lim_{r\to 0} u_*(r) r p_*''(r) = 0.$$

This proves the first two relations in (3.3). The proof of the third relation is similar and is omitted. Next, from (2.14) we can easily deduce that

$$\begin{aligned} u_*''(r) &= [-K_D'(c_*(r)) + K_M'(c_*(r))p_*(r)]c_*'(r) \\ &\quad + K_M(c_*(r))p_*'(r) + \frac{2}{r}[K_D(c_*(r)) - K_M(c_*(r))p_*(r)] \\ &\quad + \frac{6}{r^4}\int_0^r [-K_D(c_*(\rho)) + K_M(c_*(\rho))p_*(\rho)]\rho^2 d\rho \\ &= [-K_D'(c_*(r)) + K_M'(c_*(r))p_*(r)]c_*'(r) + \frac{6}{r^4}\int_0^r [K_D(c_*(r)) - K_D(c_*(\rho))]\rho^2 d\rho \\ &\quad + K_M(c_*(r))p_*'(r) - \frac{6}{r^4}\int_0^r [K_M(c_*(r))p_*(r) - K_M(c_*(\rho))p_*(\rho)]\rho^2 d\rho. \end{aligned} \tag{3.6}$$

From this expression and the first relation in (3.3) we readily obtain the first relation in (3.5). The proof of the second relation in (3.5) is similar and is omitted. Finally, by Theorems 5.3 and 5.4 of [13], we know that either $p_* \in C^1[0,1]$ or there exist constants $-1 < \alpha < 0$ and C such that[1]

$$p_*'(r) = Cr^\alpha + O(1) \quad \text{for } r \to 0. \tag{3.7}$$

Suppose that it is the second case. Then, by letting $\gamma = |\alpha|$, we see that $0 < \gamma < 1$ and $r^\gamma p_*'(r) \in C[0,1]$. Finally, from (3.7) we see that

$$p_*(r) = p_*(0) + C(1+\alpha)^{-1} r^{1+\alpha} + O(r) \quad \text{for } r \to 0. \tag{3.8}$$

Substituting (3.7) and (3.8) into (3.6), we get $r^\gamma u_*''(r) \in C[0,1]$. This completes the proof of Lemma 3.1. ∎

COROLLARY 3.2. *Let $a, b, \mathcal{B}, \mathcal{F}$, and $\mathbb{B}$ be as in* (2.19)–(2.22) *and* (2.24). *Then we have $a, b \in C_V^1[0,1], \mathcal{B} \in L(C[0,1], C_V^1[0,1]) \subseteq L(C[0,1]) \cap L(C_V^1[0,1]), \mathcal{F} \in L(C[0,1], \mathbb{R})$, and $\mathbb{B} \in L(X) \cap L(X_0)$. Moreover, we also have $r^2(1-r)^2 a''(r)$, $r^2(1-r)^2 b''(r) \in C[0,1]$.* ∎

COROLLARY 3.3. *$\mathbb{G} \in C^\infty(X,X) \cap C^\infty(X_0, X_0)$ and in addition to* (2.18), *we also have*

$$\|\mathbb{G}(V)\|_{X_0} = O\left(\|V\|_{X_0}^2\right) \quad \text{as } \|V\|_{X_0} \to 0. \tag{3.9}$$

Proof. We have $\mathbb{G}(V) = \mathbb{G}_1(V) + \mathbb{G}_2(V)$, where $\mathbb{G}_1(V) = [\mathbb{A}_0(U_* + V) - \mathbb{A}_0(U_*) - \mathbb{A}_0'(U_*)V]U_*$ and $\mathbb{G}_2(V) = \mathbb{F}_0(U_* + V) - \mathbb{F}_0(U_*) - \mathbb{F}_0'(U_*)V$. Since $\mathbb{F}_0 \in C^\infty(X_0, X_0)$, it is evident that $\mathbb{G}_2 \in C^\infty(X_0, X_0)$ and $\|\mathbb{G}_2(V)\|_{X_0} = O(\|V\|_{X_0}^2)$ as $\|V\|_{X_0} \to 0$. Next, let $V = (\varphi, \zeta)$ and $(p, z) = (p_* + \varphi, z_* + \zeta)$. Then, by (2.25), we have $\mathbb{A}_0(U_* + V)U_* = (-w_{p,z}(r)p_*'(r), 0)$. Using this expression and the first two relations in (3.3), we can easily show that, for every $V \in X$, we have $\mathbb{A}_0(U_* + V)U_* \in X_0$, and the mapping $V \to \mathbb{A}_0(U_* + V)U_*$ belongs to $C^\infty(X, X_0)$. Hence we have $\mathbb{G}_1 \in C^\infty(X, X_0) \subseteq C^\infty(X_0, X_0)$ and $\|\mathbb{G}_1(V)\|_{X_0} = O(\|V\|_X^2) = O(\|V\|_{X_0}^2)$ as $\|V\|_{X_0} \to 0$. Combining these assertions together, we see that the desired assertion follows. ∎

4. Evolution system. In this section we study the following initial value problem:

$$\frac{dU}{dt} = \mathbb{A}(V(t))U + F(t) \quad \text{for } t > 0, \quad U(0) = U_0, \tag{4.1}$$

[1] In the notation of Theorem 5.4 of [13], we have $\alpha = \alpha(\lambda)$ and $C = (1 + \alpha(\lambda))\omega$.

where $V \in C([0,\infty), X)$, $F \in C([0,\infty), X_0)$, and $U_0 \in X_0$ are given. For a small $\varepsilon > 0$, we denote

$$S_\varepsilon = \{V = (\varphi, \zeta) \in X = C[0,1] \times \mathbb{R} : \ \|\varphi\|_\infty \le \varepsilon, \ |\zeta| \le \varepsilon\}.$$

We want to prove that given $V \in C([0,\infty), X)$ such that $V(t) \in S_\varepsilon$ for all $t \ge 0$, the problem (4.1) has a unique solution $U \in C([0,\infty), X_0) \cap C^1([0,\infty), X)$, and there exists a family of bounded linear operators $\{\mathbb{U}(t,s,V) : t \ge s \ge 0\}$ on X_0, the so-called *evolution system* determined by the family of operators $\{\mathbb{A}(V(t)) : t \ge 0\}$, such that the solution of this problem is given by

$$U(t) = \mathbb{U}(t,0,V)U_0 + \int_0^t \mathbb{U}(t,s,V)F(s)ds \quad \text{for } \ t \ge 0. \tag{4.2}$$

For this purpose we shall prove that the family of operators $\{\mathbb{A}(V) : V \in S_\varepsilon\}$ is a stable family of the infinitesimal generators of C_0 semigroups on X, and its part in X_0 is a stable family of the infinitesimal generators of C_0 semigroups on X_0. When these assertions are proved, the desired assertion then follows from the results of sections 5.2–5.5 and 6.4 of [20].

In what follows, for $V \in S_\varepsilon$, we denote by $\tilde{\mathbb{A}}(V)$ the part of $\mathbb{A}(V)$ in X_0. Recall that $\mathrm{Dom}(\tilde{\mathbb{A}}(V)) = \{U \in X_0 : \mathbb{A}(V)U \in X_0\}$, and $\tilde{\mathbb{A}}(V)U = \mathbb{A}(V)U$ for $U \in \mathrm{Dom}(\tilde{\mathbb{A}}(V))$.

Let $w \in C^1[0,1]$ and assume that it satisfies the following condition: There exist positive constants C_1 and C_2 such that

$$-C_1 r(1-r) \le w(r) \le -C_2 r(1-r) \quad \text{for } \ 0 \le r \le 1. \tag{4.3}$$

Note that this assumption particularly implies that $w(0) = w(1) = 0$, $w'(0) < 0$ and $w'(1) > 0$. For a such w, we denote by $\mathcal{L}_0$ the bounded linear operator from $C^1_V[0,1]$ to $C[0,1]$ defined by

$$\mathcal{L}_0 q(r) = -w(r)q'(r) \quad \text{for } \ 0 < r < 1, \quad \text{for } \ q \in C^1_V[0,1].$$

Later on we shall also regard $\mathcal{L}_0$ as an unbounded linear operator in $C[0,1]$ with domain $C^1_V[0,1]$. Note that if $w = u_*$, then $\mathcal{L}_0 = \mathcal{A}_0(p_*, z_*)$.

LEMMA 4.1. *Let the notation and assumption be as above. Then $\mathcal{L}_0$ generates a C_0 semigroup of contractions $e^{t\mathcal{L}_0}$ on $C[0,1]$, i.e.,*

$$\|e^{t\mathcal{L}_0}\|_{L(C[0,1])} \le 1 \quad \textit{for } \ t \ge 0. \tag{4.4}$$

Moreover, $C^1_V[0,1]$ is $\mathcal{L}_0$-admissible,[2] and the restriction of $e^{t\mathcal{L}_0}$ on $C^1_V[0,1]$ is a uniformly bounded C_0 semigroup on $C^1_V[0,1]$, i.e., there exists constant $C > 0$ depending only on the constants C_1 and C_2 in (4.3) such that

$$\|e^{t\mathcal{L}_0}\|_{L(C^1_V[0,1])} \le C \quad \textit{for } \ t \ge 0. \tag{4.5}$$

[2]Recall that for a C_0 semigroup $T(t)$ $(t \ge 0)$ on a Banach space X generated by an unbounded linear operator A in X, a linear subspace Y of X is called *A-admissible* if it is an invariant subspace of $T(t)$ for all $t \ge 0$ and the restriction of $T(t)$ $(t \ge 0)$ to Y is a C_0 semigroup in Y. A necessary and sufficient condition for Y to be A-admissible is that (1) Y is an invariant subspace of $R(\lambda, A)$ for all $\lambda > \omega$, and (2) the part $\bar{A}$ of A in Y is an infinitesimal generator of a C_0 semigroup on Y. In this case we have $e^{t\bar{A}} = e^{tA}|_Y$. See Theorem 5.5 in Chapter 4 of [20].

Proof. Let $\lambda \in \mathbb{C}$ be such that $\mathrm{Re}\lambda > 0$, and let $f \in C[0,1]$. Consider the equation

$$-w(r)q'(r) - \lambda q(r) = f(r) \quad \text{for } 0 < r < 1. \tag{4.6}$$

By using some similar arguments as in section 2 of [8], we can easily show that this equation has a unique solution defined and bounded for $0 < r < 1$, given by

$$q(r) = e^{-\lambda \int_{r_0}^{r} \frac{d\rho}{w(\rho)}} \int_r^1 \frac{f(\eta)}{w(\eta)} e^{\lambda \int_{r_0}^{\eta} \frac{d\rho}{w(\rho)}} d\eta, \quad 0 < r < 1, \tag{4.7}$$

where r_0 is an arbitrarily chosen number in $(0,1)$. Clearly $q \in C^1(0,1)$. By using the L'Hospital's rule, we can easily verify that $q(r)$ has finite limits as $r \to 0^+$ and $r \to 1^-$, so that $q \in C[0,1] \cap C^1(0,1)$. Furthermore, since $r(1-r)/w(r) \in C[0,1]$, from (4.6) we see that also $r(1-r)q'(r) \in C[0,1]$, so that $q \in C_V^1[0,1]$, and, moreover,

$$\begin{aligned} |q(r)| &\le \|f\|_\infty e^{(\mathrm{Re}\lambda) \int_{r_0}^{r} \frac{d\rho}{|w(\rho)|}} \int_r^1 \frac{1}{|w(\eta)|} e^{-(\mathrm{Re}\lambda) \int_{r_0}^{\eta} \frac{d\rho}{|w(\rho)|}} d\eta \\ &= \frac{\|f\|_\infty}{\mathrm{Re}\lambda} \Big[1 - e^{-(\mathrm{Re}\lambda) \int_r^1 \frac{d\rho}{|w(\rho)|}} \Big] \le \frac{\|f\|_\infty}{\mathrm{Re}\lambda} \quad \text{for } 0 < r < 1. \end{aligned} \tag{4.8}$$

This proves that, for any $\lambda \in \mathbb{C}$ with $\mathrm{Re}\lambda > 0$, we have $\lambda \in \rho(\mathcal{L}_0)$ and $\|R(\lambda, \mathcal{L}_0)\|_{L(C[0,1])} \le 1/\mathrm{Re}\lambda$. Thus, by the Hille–Yosida theorem, we see that $\mathcal{L}_0$ generates a strongly continuous semigroup $e^{t\mathcal{L}_0}$ on $C[0,1]$, which satisfies the estimate (4.4).

Next we prove that $C_V^1[0,1]$ is $\mathcal{L}_0$-admissible. Clearly, $C_V^1[0,1]$ is an invariant subspace of $R(\lambda, \mathcal{L}_0)$ for all $\lambda \in \mathbb{C}$, with $\mathrm{Re}\lambda > 0$. Let $\tilde{\mathcal{L}}_0$ be the part of $\mathcal{L}_0$ in $C_V^1[0,1]$. Since $r(1-r)/w(r)$, $w(r)/r(1-r) \in C[0,1]$, we see that, for $q \in C[0,1] \cap C^1(0,1)$, $r(1-r)q'(r) \in C[0,1]$ if and only if $w(r)q'(r) \in C[0,1]$, and there exist positive constants C_1 and C_2 such that

$$C_1 \sup_{0<r<1} |w(r)q'(r)| \le \sup_{0<r<1} |r(1-r)q'(r)| \le C_2 \sup_{0<r<1} |w(r)q'(r)|.$$

By the above assertion, it follows easily that

$$\mathrm{Dom}\left(\tilde{\mathcal{L}}_0\right) = \left\{ q \in C[0,1] \cap C^2(0,1) :\ w(r)q'(r) \in C[0,1],\ w^2(r)q''(r) \in C[0,1] \right\}, \tag{4.9}$$

and $\|q\|'_{C_V^1[0,1]} = \|q\|_\infty + \|wq'\|_\infty$ is an equivalent norm in $C_V^1[0,1]$. For $q \in \mathrm{Dom}(\tilde{\mathcal{L}}_0)$, we have, by definition, $\tilde{\mathcal{L}}_0 q = \mathcal{L}_0 q$. Now let $f \in C_V^1[0,1]$ and $\mathrm{Re}\lambda > 0$. Let q be the solution of (4.6). Using (4.9) we can easily verify that $q \in \mathrm{Dom}(\tilde{\mathcal{L}}_0)$, so that it is the solution of the equation $\tilde{\mathcal{L}}_0 q - \lambda q = f$. Moreover, a simple computation shows that $wq' = R(\mathcal{L}_0, \lambda)(wf')$. Thus, by (4.8), we have

$$\begin{aligned} \|q\|'_{C_V^1[0,1]} &= \|q\|_\infty + \|wq'\|_\infty = \|R(\mathcal{L}, \lambda)f\|_\infty + \|R(\mathcal{L}, \lambda)(wf')\|_\infty \\ &\le \frac{1}{\mathrm{Re}\lambda}\|f\|_\infty + \frac{1}{\mathrm{Re}\lambda}\|wf'\|_\infty = \frac{1}{\mathrm{Re}\lambda}\|f\|'_{C_V^1[0,1]}. \end{aligned}$$

Hence, $\lambda \in \rho(\tilde{\mathcal{L}}_0)$ and $\|R(\lambda, \tilde{\mathcal{L}}_0)\|'_{C_V^1[0,1]} \le (\mathrm{Re}\lambda)^{-1}$. The desired assertion then follows from the Hille–Yosida theorem and footnote 2. ∎

Given $V = (\varphi, \zeta) \in S_\varepsilon$, we set $p(r) = p_*(r) + \varphi(r)$, $z = z_* + \zeta$, and, as before, we denote

$$u_{p,z}(r) = \frac{1}{r^2} \int_0^r [-K_D(c(\rho, z)) + K_M(c(\rho, z))p(\rho)]\rho^2 d\rho, \quad w_{p,z}(r) = u_{p,z}(r) - r u_{p,z}(1).$$

Since $\|\varphi\|_\infty \le \varepsilon$ and $|\zeta| \le \varepsilon$, by a simple computation, we see that

$$-C\varepsilon r(1-r) \le w_{p,z}(r) - u_*(r) \le C\varepsilon r(1-r) \quad \text{for } 0 \le r \le 1. \tag{4.10}$$

Since $-C_1 r(1-r) \le u_*(r) \le -C_2 r(1-r)$, it follows that, for ε sufficiently small, we have

$$-C_1 r(1-r) \le w_{p,z}(r) \le -C_2 r(1-r) \quad \text{for } 0 \le r \le 1. \tag{4.11}$$

Later on we shall also use the notation $w_V(r)$ to redenote $w_{p,z}(r)$. We note that all constants C, C_1, and C_2 that appear in (4.10)–(4.11) are independent of V and ε.

LEMMA 4.2. *$\{\mathbb{A}(V) : V \in S_\varepsilon\}$ is a stable family of the infinitesimal generators of C_0 semigroups on $X = C[0,1] \times \mathbb{R}$, and $\{\tilde{\mathbb{A}}(V) : V \in S_\varepsilon\}$ is a stable family of the infinitesimal generators of C_0 semigroups on X_0.*

Proof. Let $\mathcal{L}_V q(r) = -w_V(r) q'(r)$. Then, by Lemma 4.1, we see that, for any $V \in S_\varepsilon$, $\mathcal{L}_V$ is an infinitesimal generator of a C_0 semigroup of contractions on $C[0,1]$. From this assertion and the second expression in (2.25), we easily infer that, for any $V \in S_\varepsilon$, $\mathbb{A}_0(U_* + V)$ is an infinitesimal generator of a C_0 semigroup of contractions on $X = C[0,1] \times \mathbb{R}$. Hence, $\{\mathbb{A}_0(U_* + V) : V \in S_\varepsilon\}$ is a stable family of the infinitesimal generators of C_0 semigroups on X, with stability constants $(M, \omega) = (1, 0)$. Since $\mathbb{A}(V) = \mathbb{A}_0(U_* + V) + \mathbb{B}$ and $\mathbb{B}$ is a bounded linear operator on X independent of V, by a standard perturbation result (see, e.g., Theorem 2.3 in section 5.2 of [20]), we see immediately that $\{\mathbb{A}(V) : V \in S_\varepsilon\}$ is a stable family of the infinitesimal generators of C_0 semigroups on $X = C[0,1] \times \mathbb{R}$, with stability constants $(M, \omega) = (1, \|\mathbb{B}\|)$.

In order to prove that $\{\tilde{\mathbb{A}}(V) : V \in S_\varepsilon\}$ is a stable family of the infinitesimal generators of a C_0 semigroup on $X_0 = C_V^1[0,1] \times \mathbb{R}$, we first establish an estimate for the semigroup $e^{t\tilde{\mathcal{L}}_V}$ on $C_V^1[0,1]$ different from (4.5), where $\tilde{\mathcal{L}}_V$ represents the part of $\mathcal{L}_V$ in $C_V^1[0,1]$. Let $q_0 \in C_V^1[0,1]$ and $q = e^{t\tilde{\mathcal{L}}_V} q_0 = e^{t\mathcal{L}_V} q_0$. Then q is the solution of the problem

$$\frac{\partial q}{\partial t} + w_V(r)\frac{\partial q}{\partial r} = 0 \quad \text{for } 0 \le r \le 1 \text{ and } t > 0, \quad q|_{t=0} = q_0.$$

Let $l(r,t) = r(1-r)\frac{\partial q(r,t)}{\partial r}$ and $l_0(r) = r(1-r)q_0'(r)$. Formally differentiating the above equation in r and multiplying it with $r(1-r)$, we get

$$\frac{\partial l}{\partial t} + w_V(r)\frac{\partial l}{\partial r} = a_V(r) l \quad \text{for } 0 \le r \le 1 \text{ and } t > 0, \quad l|_{t=0} = l_0,$$

where $a_V(r) = (1-2r)\frac{w_V(r)}{r(1-r)} - w_V'(r)$. Since $w_V \in C^1[0,1], w_V(0) = w_V(1) = 0$, and $a_V \in C[0,1]$, by a standard characteristics argument, we can easily show that this problem has a unique solution $l \in C([0,1] \times [0,\infty))$ such that $[\frac{\partial}{\partial t} + w_V(r)\frac{\partial}{\partial r}]l \in C([0,1] \times [0,\infty))$. Thus, the above formal computation makes sense. Clearly, there exists a nonnegative constant c_0 independent of V such that $a_V(r) \le c_0$ for $0 < r < 1$, for all $V \in S_\varepsilon$. Using this fact and the characteristics argument, we can easily obtain $\|l(\cdot,t)\|_\infty \le \|l_0\|_\infty e^{c_0 t}$ for $t \ge 0$. Combining this estimate with $\|q(\cdot,t)\|_\infty \le \|q_0\|_\infty$ ensured by (4.4), we get $\|q(\cdot,t)\|_{C_V^1[0,1]} \le \|q_0\|_{C_V^1[0,1]} e^{c_0 t}$ for $t \ge 0$, or

$$\|e^{t\tilde{\mathcal{L}}_V}\|_{L(C_V^1[0,1])} \le e^{c_0 t} \quad \text{for } t \ge 0, \quad \text{for all } V \in S_\varepsilon.$$

Hence, $\{\tilde{\mathcal{L}}_V : V \in S_\varepsilon\}$ is a stable family of the infinitesimal generators of C_0 semigroups on $C_V^1[0,1]$, with stability constants $(M, \omega) = (1, c_0)$. Using this assertion and

the second expression in (2.25), we see easily that $\{\tilde{\mathbb{A}}_0(U_* + V) : V \in S_\varepsilon\}$, the part of $\{\mathbb{A}_0(U_* + V) : V \in S_\varepsilon\}$ on $X_0 = C[0,1] \times \mathbb{R}$, is a stable family of the infinitesimal generators of C_0 semigroups on X_0, with stability constants $(M, \omega) = (1, c_0)$. Since $\tilde{\mathbb{A}}(V) = \tilde{\mathbb{A}}_0(U_* + V) + \mathbb{B}$, and, by Corollary 3.2, $\mathbb{B}$ is a bounded linear operator on X_0 independent of V, we conclude, as before, that $\{\mathbb{A}(V) : V \in S_\varepsilon\}$ is a stable family of the infinitesimal generators of C_0 semigroups on $X_0 = C^1_V[0,1] \times \mathbb{R}$, with stability constants $(M, \omega) = (1, c_0 + \|\mathbb{B}\|_{L(C^1_V[0,1])})$. This completes the proof of Lemma 4.2. ∎

Since $\mathbb{A} \in C^\infty(X, L(X_0, X))$, by Lemma 4.2, we see that, for any $V \in C([0,\infty), X)$ such that $V(t) \in S_\varepsilon$ for all $t \geq 0$, $\{\mathbb{A}(V(t)) : t \geq 0\}$ satisfies the conditions (H_1)–(H_3) in section 5.3 of [20]. It follows by Theorem 3.1 in section 5.3 of [20] that given such a function $V = V(t)$, there exists an evolution system determined by $\{\mathbb{A}(V(t)) : t \geq 0\}$, which we denote as $\mathbb{U}(t, s, V)$. By definition, this means that

(1) for any $t \geq s \geq 0$, $\mathbb{U}(t, s, V)$ is a bounded linear operator on X;
(2) $\mathbb{U}(s, s, V) = \mathrm{id}$ for all $s \geq 0$, $\mathbb{U}(t, s, V)\mathbb{U}(s, r, V) = \mathbb{U}(t, r, V)$ for all $t \geq s \geq r$; and
(3) the mapping $(t, s) \to \mathbb{U}(t, s, V)$ is strongly continuous for $t \geq s \geq 0$.

However, the theory developed in [20] does not ensure that $U = \mathbb{U}(t, s, V)U_0$ is a solution of the problem

$$\frac{dU}{dt} = \mathbb{A}(V(t))U \quad \text{for } t > s, \quad U|_{t=s} = U_0, \tag{4.12}$$

even if $U_0 \in X_0$, unless some other conditions are satisfied by $\mathbb{U}(t, s, V)$. These conditions are as follows (see conditions (E_4) and (E_5) in Theorem 4.3 in section 5.4 of [20]):

(4) $\mathbb{U}(t, s, V)X_0 \subseteq X_0$ for any $t \geq s \geq 0$; and
(5) For any $U_0 \in X_0$, the mapping $(t, s) \to \mathbb{U}(t, s, V)U_0$ is continuous in X_0 for $t \geq s \geq 0$.

In the following lemma we shall directly prove that, for any $U_0 \in X_0$, the problem (4.12) has a unique solution $U = U_s(t) \in C([s,\infty), X_0) \cap C^1([s,\infty), X)$. By Theorem 4.2 in section 5.4 of [20], it then follows that $U_s(t) = \mathbb{U}(t, s, V)U_0$, and consequently, the conditions (4) and (5) above are satisfied.

LEMMA 4.3. *Given $V \in C([0,\infty), X)$ such that $V(t) \in S_\varepsilon$ for all $t \geq 0$, for any $s \geq 0$ and any $U_0 \in X_0$, the problem* (4.12) *has a unique solution $U = U_s(t) \in C([s,\infty), X_0) \cap C^1([s,\infty), X)$.*

Proof. Let $U = (q, y)$ and $U_0 = (q_0, y_0)$. Then (4.12) can be rewritten as follows:

$$\begin{cases} \dfrac{\partial q}{\partial t} + w_V(r,t)\dfrac{\partial q}{\partial r} = a(r)q + \mathcal{B}(q) + b(r)y \quad \text{for } 0 \leq r \leq 1, \ t > s, \\ \dfrac{dy}{dt} = \mathcal{F}(q) + \kappa y \quad \text{for } t > s, \\ q|_{t=s} = q_0(r) \quad \text{for } 0 \leq r \leq 1, \quad \text{and} \quad y|_{t=s} = y_0. \end{cases} \tag{4.13}$$

Using the characteristic method and the Banach fixed point theorem, we can easily show that this problem has a unique local solution (q, y), with $q \in C([0,1] \times [0,\delta])$ and $y \in C^1[0,\delta]$ for some $\delta > 0$. Since $w_V(0,t) = w_V(1,t) = 0$ for all $t \geq 0$, we see that the two lines $r = 0$ and $r = 1$ are characteristic curves. It follows that all characteristic curves starting from the open interval $(0,1)$ always lie in it, so that the solution of the above problem exists for all $t \geq s$. It remains to prove that $q \in C([0,\infty), C^1_V[0,1])$. To this end we formally differentiate the first equation in (4.13) in r and multiply it

with $r(1-r)$, which gives, by letting $l(r,t) = r(1-r)\frac{\partial q(r,t)}{\partial r}$, that

$$\frac{\partial l}{\partial t} + w_V(r,t)\frac{\partial l}{\partial r} = a_1(r,t)u + f_1(r,t) \quad \text{for } 0 \le r \le 1, \ \ t > 0, \tag{4.14}$$

where $a_1(r,t) = a(r) + (1-2r)\frac{w_V(r,t)}{r(1-r)} - \frac{\partial w_V(r,t)}{\partial r}$, and

$$f_1(r,t) = r(1-r)a'(r)q(r,t) + r(1-r)\frac{\partial \mathcal{B}q(r,t)}{\partial r} + r(1-r)b'(r)y(t).$$

Clearly, $a_1 \in C([0,1]\times[0,\infty))$. By Corollary 3.2, we see that also $f_1 \in C([0,1]\times[0,\infty)$. Thus, by using the characteristic method, we can easily prove that (4.14) imposed with the initial condition $l(r,0) = r(1-r)q_0'(r)$ has a unique solution $l \in C([0,1]\times[0,\infty))$ such that $[\frac{\partial}{\partial t} + w_V(r,t)\frac{\partial}{\partial r}]l \in C([0,1]\times[0,\infty))$. Thus, the above formal computation makes sense, and consequently, $q \in C([0,\infty), C_V^1[0,1]) \cap C^1([0,\infty), C[0,1])$. The desired assertion now becomes immediate. □

By the above results and Theorems 4.2 and 5.2 in sections 5.4 and 5.5 of [20], we get the following.

COROLLARY 4.4. *Let $V \in C([0,\infty), X)$ be as in Lemma* 4.3, *and let $F \in C([0,\infty), X_0)$. Then, for any $U_0 \in X_0$, the initial value problem* (4.1) *has a unique solution $U \in C([0,\infty), X_0) \cap C^1([0,\infty), X)$, and it is given by* (4.2). □

5. Similarity transformation. In the previous section we proved that given $V \in C([0,\infty), X)$ such that $V(t) \in S_\varepsilon$ for all $t \ge 0$, the family of operators $\{\mathbb{A}(V(t)) : t \ge 0\}$ determines an evolution system $\mathbb{U}(t,s,V)$ $(t \ge s \ge 0)$, so that the solution of (4.1) can be expressed as (4.2). However, the deduction in the previous section does not provide us with an estimate of the form $\|\mathbb{U}(t,s,V)\| \le Ce^{-\mu(t-s)}$, which, however, is essential in order to establish the estimate (2.17). To establish such an estimate, we shall follow a different approach as follows: For $V \equiv 0$, the desired estimate will be obtained by improving the linear estimate established in [5]. For a general $V = V(t)$, we shall use a special transformation, which we call *similarity transformation,* to transform the problem into an equivalent problem which can be estimated by using the estimate for the case $V \equiv 0$. The similarity transformation is a family of C^1 diffeomorphisms $\bar{r} = T(r,t,s)$ of the unit interval $0 \le r \le 1$ to itself, where $t \ge s \ge 0$ are parameters. The aim of this section is to establish this transformation.

Let $w \in C([0,\infty), C^1[0,1])$. We assume that w satisfies the following condition: For some small parameter $\varepsilon > 0$,

$$-C\varepsilon r(1-r)e^{-\mu t} \le w(r,t) - u_*(r) \le C\varepsilon r(1-r)e^{-\mu t} \quad \text{for } 0 \le r \le 1, \ \ t \ge 0, \tag{5.1}$$

where C is a positive constant independent of ε and w. Since $-C_1 r(1-r) \le u_*(r) \le -C_2 r(1-r)$, we see that

$$\sup_{0<r<1}\Big|\frac{w(r,t)}{u_*(r)} - 1\Big| \le C\varepsilon e^{-\mu t} \quad \text{for } 0 \le r \le 1, \ \ t \ge 0, \tag{5.2}$$

and, for ε sufficiently small, $w(r,t)$ satisfies (4.3) for all $t \ge 0$.

Let $0 \le \xi \le 1$ and $s \ge 0$. Consider the following initial value problem:

$$\frac{dr}{dt} = u_*(r) \quad \text{for } t > s, \quad r|_{t=s} = \xi. \tag{5.3}$$

Since $u_* \in C^1[0,1]$, $u_*(r) < 0$ for $0 < r < 1$, and, in particular, $u_*(0) = u_*(1) = 0$, it can be easily shown that this problem has a unique solution $r = \Phi_*(\xi, t, s)$ for all $t \geq s \geq 0$, satisfying the following properties (A):

$$\Phi_*(\xi,t,s) \text{ is continuously differentiable in } (\xi,t,s),$$
$$\Phi_*(0,t,s) = 0, \quad \Phi_*(1,t,s) = 1 \quad \text{for } t \geq s \geq 0,$$
$$0 < \Phi_*(\xi,t,s) < 1 \quad \text{for } 0 < \xi < 1, \ t \geq s \geq 0,$$
$$\frac{\partial \Phi_*(\xi,t,s)}{\partial \xi} > 0, \quad \frac{\partial \Phi_*(\xi,t,s)}{\partial t} < 0 \quad \text{for } 0 < \xi < 1, \ t \geq s \geq 0,$$
$$\Phi_*(\xi,s,s) = \xi \quad \text{for } 0 \leq \xi \leq 1, \ s \geq 0.$$

From these properties we see that, for any $s \geq 0$ and $t \geq s$, the mapping $\xi \to r = \Phi_*(\xi,t,s)$ is a C^1 diffeomorphism of $[0,1]$ to itself. Let $\xi = \Psi_*(r,t,s)$ be the inverse of this mapping. Clearly, Ψ_* also satisfies the properties (A) (with ξ replaced by r). Furthermore, from the relation $\Psi_*(\Phi_*(\xi,t,s),t,s) = \xi$ (for $0 \leq \xi \leq 1$ and $t \geq s \geq 0$), it can be easily shown that $\xi = \Psi_*(r,t,s)$ is the unique solution of the following initial value problem:

$$\frac{\partial \xi}{\partial t} + u_*(r)\frac{\partial \xi}{\partial r} = 0 \quad \text{for } t > s, \quad \xi|_{t=s} = r. \tag{5.4}$$

Next, let $r = \Phi(\xi,t,s)$ $(0 \leq \xi \leq 1,\ t \geq s \geq 0)$ be the solution of the following problem:

$$\frac{dr}{dt} = w(r,t) \quad \text{for } t > s, \quad r|_{t=s} = \xi. \tag{5.5}$$

Similarly as before, $\Phi(\xi,t,s)$ is well-defined for all $0 \leq \xi \leq 1$ and $t \geq s \geq 0$, and it also satisfies the properties (A). It follows that, for any $s \geq 0$ and $t \geq s$, the mapping $\xi \to r = \Phi(\xi,t,s)$ is a C^1 diffeomorphism of $[0,1]$ to itself. Let $\xi = \Psi(r,t,s)$ be the inverse of this mapping. Clearly, Ψ also satisfies the properties (A) (with ξ replaced by r). Furthermore, from the relation $\Psi(\Phi(\xi,t,s),t,s) = \xi$ (for $0 \leq \xi \leq 1$ and $t \geq s \geq 0$), it can be easily shown that $\xi = \Psi(r,t,s)$ is the unique solution of the following initial value problem:

$$\frac{\partial \xi}{\partial t} + w(r,t)\frac{\partial \xi}{\partial r} = 0 \quad \text{for } t > s, \quad \xi|_{t=s} = r. \tag{5.6}$$

In what follows we consider the following initial value problem:

$$\begin{cases} \dfrac{\partial \bar{r}}{\partial t} + w(r,t)\dfrac{\partial \bar{r}}{\partial r} = u_*(\bar{r}) \quad \text{for } 0 \leq r \leq 1, \ t > s, \\ \bar{r}|_{t=s} = r \quad \text{for } 0 \leq r \leq 1. \end{cases} \tag{5.7}$$

LEMMA 5.1. *For any $0 \leq r \leq 1$ and $s \geq 0$, the problem* (5.7) *has a unique solution $\bar{r} = T(r,t,s)$ for all $t \geq s$, and the following relation holds:*

$$T(r,t,s) = \Phi_*(\Psi(r,t,s),t,s) \quad \textit{for } 0 \leq r \leq 1, \ t \geq s \geq 0. \tag{5.8}$$

Proof. Using (5.3) and (5.6) we can easily verify that $\bar{r} = \Phi_*(\Psi(r,t,s),t,s)$ is a solution of the problem (5.7). Thus, (5.8) follows by the uniqueness of the solution. ∎

By (5.8), it is evident that, for any $s \geq 0$ and $t \geq s$, the mapping $r \to \bar{r} = T(r,t,s)$ is a C^1 diffeomorphism of $[0,1]$ to itself, satisfying $T(0,t,s) = 0$, $T(1,t,s) = 1$ for $t \geq s \geq 0$, and $\partial T(r,t,s)/\partial r > 0$ for $0 < r < 1$, $t \geq s$. We denote by $r = S(\bar{r},t,s)$ the inverse of this mapping. By (5.8), it is clear that

$$S(\bar{r},t,s) = \Phi(\Psi_*(\bar{r},t,s),t,s) \quad \text{for } 0 \leq \bar{r} \leq 1, \ \ t \geq s \geq 0. \tag{5.9}$$

It is also clear that $S(\bar{r},t,s)$ satisfies $S(0,t,s) = 0$, $S(1,t,s) = 1$ for $t \geq s \geq 0$ and that $\partial S(\bar{r},t,s)/\partial \bar{r} > 0$ for $0 < \bar{r} < 1$, $t \geq s \geq 0$.

T and S can be expressed in more explicit formulations. To show this we introduce a function F_* as follows:

$$F_*(r) = -\int_{\frac{1}{2}}^{r} \frac{d\eta}{u_*(\eta)} = \int_{\frac{1}{2}}^{r} \frac{d\eta}{|u_*(\eta)|} \quad \text{for } 0 < r < 1. \tag{5.10}$$

Clearly, $F_* \in C^1(0,1)$, $F_*'(r) > 0$ for all $0 < r < 1$, and $\lim_{r\to 0^+} F_*(r) = -\infty$, $\lim_{r\to 1^-} F_*(r) = \infty$. Hence, $\bar{r} = F_*(r)$ is a C^1 diffeomorphism of the open unit interval $(0,1)$ onto the real line $(-\infty,\infty)$. From (5.3) we easily obtain $F_*(\Phi_*(\xi,t,s)) - F_*(\xi) = -t + s$. Thus, we have

$$\Phi_*(\xi,t,s) = F_*^{-1}(F_*(\xi) - t + s), \tag{5.11}$$

and consequently,

$$\Psi_*(r,t,s) = F_*^{-1}(F_*(r) + t - s). \tag{5.12}$$

Next, let

$$g(\xi,t,s) = G(\Phi(\xi,t,s),t), \quad \text{where} \quad G(r,t) = \frac{w(r,t)}{u_*(r)} - 1.$$

Since $w(r,t) = [1 + G(r,t)]u_*(r)$, from (5.5) we see that $r = \Phi(\xi,t,s)$ is a solution of the following problem:

$$\frac{dr}{dt} = [1 + g(\xi,t,s)]u_*(r) \quad \text{for } t > s, \quad r|_{t=s} = \xi. \tag{5.13}$$

Thus, similarly, as before, we have $F_*(\Phi(\xi,t,s)) - F_*(\xi) = -t + s - \int_s^t g(\xi,\tau,s)d\tau$, so that

$$\Phi(\xi,t,s) = F_*^{-1}\left(F_*(\xi) - t + s - \int_s^t g(\xi,\tau,s)d\tau\right), \tag{5.14}$$

$$\Psi(r,t,s) = F_*^{-1}\left(F_*(r) + t - s + \int_s^t g(\Psi(r,t,s),\tau,s)d\tau\right). \tag{5.15}$$

Combining (5.8), (5.9), (5.11), (5.12), (5.14), and (5.15), we see that

$$T(r,t,s) = F_*^{-1}\left(F_*(r) + \int_s^t g(\Psi(r,t,s),\tau,s)d\tau\right), \tag{5.16}$$

$$S(\bar{r},t,s) = F_*^{-1}\left(F_*(\bar{r}) - \int_s^t g(\Psi_*(\bar{r},t,s),\tau,s)d\tau\right). \tag{5.17}$$

LEMMA 5.2. *Assume that $|\zeta| \le C$. Then there exist positive constants C_1 and C_2 depending only on C such that, for any $0 < r < 1$, we have*

$$C_1 r(1-r) \le F_*^{-1}(F_*(r)+\zeta)\left[1 - F_*^{-1}(F_*(r)+\zeta)\right] \le C_2 r(1-r). \tag{5.18}$$

Proof. Since $-C \le \zeta \le C$, by the monotonicity of F_*, we have

$$F_*^{-1}(F_*(r)-C) \le F_*^{-1}(F_*(r)+\zeta) \le F_*^{-1}(F_*(r)+C).$$

Thus,

$$\frac{F_*^{-1}(F_*(r)+\zeta)}{r} \le \frac{F_*^{-1}(F_*(r)+C)}{r} \quad \text{and}$$
$$\frac{1-F_*^{-1}(F_*(r)+\zeta)}{1-r} \le \frac{1-F_*^{-1}(F_*(r)-C)}{1-r}.$$

Moreover, using the facts that $u_*(r) = u_*'(0)r[1+O(r^\beta)]$ as $r \to 0^+$ for some $0 < \beta \le 1$ (see Assertion (3) of Lemma 3.1) and that $u_*(r) = -u_*'(1)(1-r)[1+O(1-r)]$ as $r \to 1^-$, we can easily show that

$$\lim_{r\to 0^+} \frac{F_*^{-1}(F_*(r)+C)}{r} = e^{C|u_*'(0)|} \quad \text{and} \quad \lim_{r\to 1^-} \frac{1-F_*^{-1}(F_*(r)-C)}{1-r} = e^{Cu_*'(1)}.$$

From these relations we immediately obtain the second inequality in (5.18). The proof for the first inequality in (5.18) is similar. ∎

COROLLARY 5.3. *For ε sufficiently small, we have*

$$C_1 r(1-r) \le T(r,t,s)[1-T(r,t,s)] \le C_2 r(1-r), \tag{5.19}$$
$$C_1 \bar{r}(1-\bar{r}) \le S(\bar{r},t,s)[1-S(\bar{r},t,s)] \le C_2 \bar{r}(1-\bar{r}). \tag{5.20}$$

Proof. Let $\zeta = \int_s^t g(\Psi(r,t,s),\tau,s)d\tau$. By (5.2), we have

$$|\zeta| \le \int_s^t |g(\Psi(r,t,s),\tau,s)|d\tau \le C\varepsilon \int_s^t e^{-\mu\tau}d\tau \le C\varepsilon \int_0^\infty e^{-\mu\tau}d\tau \le C\varepsilon \le C.$$

Hence, (5.19) follows from (5.16) and (5.18). Similarly, (5.20) follows from (5.17) and (5.18). ∎

As an immediate consequence of Corollary 5.3, we see that there exists a constant $C > 1$ such that, for ε sufficiently small, we have $C^{-1} \le T(r,t,s)/r \le C$ and $C^{-1} \le S(\bar{r},t,s)/\bar{r} \le C$.

COROLLARY 5.4. *For ε sufficiently small, we have the following inequalities:*

$$C_1\Psi_*(r,t,s)[1-\Psi_*(r,t,s)] \le \Psi(r,t,s)[1-\Psi(r,t,s)] \le C_2\Psi_*(r,t,s)[1-\Psi_*(r,t,s)], \tag{5.21}$$
$$C_1\Phi_*(\bar{r},t,s)[1-\Phi_*(\bar{r},t,s)] \le \Phi(\bar{r},t,s)[1-\Phi(\bar{r},t,s)] \le C_2\Phi_*(\bar{r},t,s)[1-\Phi_*(\bar{r},t,s)]. \tag{5.22}$$

Proof. Let $\bar{r} = \Psi_*(r,t,s)$ and $\zeta = \int_s^t g(\Psi(r,t,s),\tau,s)d\tau$. Then, by (5.12) and (5.15), we have $\Psi(r,t,s) = F_*^{-1}(F_*(\bar{r})+\zeta)$. By this expression and (5.18), we immediately obtain (5.21). The proof of (5.22) is similar. ∎

As an immediate consequence of Corollary 5.4, we see that there exists a constant $C > 1$ such that, for ε sufficiently small, we have $C^{-1} \le \Psi(r,t,s)/\Psi_*(r,t,s) \le C$ and $C^{-1} \le \Phi(\bar{r},t,s)/\Phi_*(\bar{r},t,s) \le C$.

LEMMA 5.5. *We have the following inequalities:*

$$|T(r,t,s) - r| \le C\varepsilon\big(e^{-\mu s} - e^{-\mu t}\big)r(1-r), \tag{5.23}$$

$$|S(\bar{r},t,s) - \bar{r}| \le C\varepsilon\big(e^{-\mu s} - e^{-\mu t}\big)\bar{r}(1-\bar{r}). \tag{5.24}$$

Proof. Similarly as in the proof of Corollary 5.3, we have

$$\int_s^t |g(\Psi(r,t,s),\tau,s)|d\tau \le C\varepsilon\int_s^t e^{-\mu\tau} \le C\varepsilon\Big(e^{-\mu s} - e^{-\mu t}\Big).$$

Thus, by noticing that $\frac{dF_*^{-1}(\eta)}{d\eta} = \frac{1}{F_*'(F_*^{-1}(\eta))} = |u_*(F_*^{-1}(\eta))|$, we see that

$$\begin{aligned}
|T(r,t,s) - r| &= \left|F_*^{-1}\left(F_*(r) + \int_s^t g(\Psi(r,t,s),\tau,s)d\tau\right) - F_*^{-1}(F_*(r))\right| \\
&\le \int_0^1 \left|u_*\left(F_*^{-1}(F_*(r)+\zeta_\theta)\right)\right| d\theta \cdot \int_s^t |g(\Psi(r,t,s),\tau,s)|d\tau \\
&\le C\varepsilon\left(e^{-\mu s} - e^{-\mu t}\right)\cdot\int_0^1 \left|u_*\left(F_*^{-1}(F_*(r)+\zeta_\theta)\right)\right| d\theta,
\end{aligned}$$

where $\zeta_\theta = \theta\int_s^t g(\Psi(r,t,s),\tau,s)d\tau$. Since $|\zeta_\theta| \le C$ and $|u_*(\eta)| \le C\eta(1-\eta)$, by Lemma 5.2, we have

$$\left|u_*\left(F_*^{-1}(F_*(r)+\zeta_\theta)\right)\right| \le CF_*^{-1}(F_*(r)+\zeta_\theta)\left[1 - F_*^{-1}(F_*(r)+\zeta_\theta)\right] \le Cr(1-r).$$

Substituting this estimate into the above inequality, we see that (5.23) follows. The proof of (5.24) is similar. ∎

COROLLARY 5.6. *We have the following inequalities:*

$$|\Phi(\xi,t,s) - \Phi_*(\xi,t,s)| \le C\varepsilon\left(e^{-\mu s} - e^{-\mu t}\right)\Phi_*(\xi,t,s)[1-\Phi_*(\xi,t,s)], \tag{5.25}$$

$$|\Psi(r,t,s) - \Psi_*(r,t,s)| \le C\varepsilon\left(e^{-\mu s} - e^{-\mu t}\right)\Psi_*(r,t,s)[1-\Psi_*(r,t,s)]. \tag{5.26}$$

Proof. Let $r = \Phi(\xi,t,s)$. Then $\xi = \Psi(r,t,s)$ and $\Phi_*(\xi,t,s) = \Phi_*(\Psi(r,t,s),t,s) = T(r,t,s)$. Thus, by (5.23), we have

$$|\Phi(\xi,t,s) - \Phi_*(\xi,t,s)| = |r - T(r,t,s)| \le C\varepsilon\left(e^{-\mu s} - e^{-\mu t}\right)r(1-r).$$

Substituting $r = \Phi(\xi,t,s)$ into the right-hand side of the last inequality and using (5.22), we see that (5.25) follows. The proof of (5.26) is similar. ∎

LEMMA 5.7. *Assume that, in addition to* (5.2)*, there also holds*

$$\max_{0\le r\le 1}\left|\frac{\partial w(r,t)}{\partial r} - u_*'(r)\right| \le C\varepsilon e^{-\mu t}. \tag{5.27}$$

Then we have the following estimates:

$$e^{-C\varepsilon\left(e^{-\mu s}-e^{-\mu t}\right)} \le \frac{\partial T(r,t,s)}{\partial r} \le e^{C\varepsilon\left(e^{-\mu s}-e^{-\mu t}\right)}, \tag{5.28}$$

$$e^{-C\varepsilon\left(e^{-\mu s}-e^{-\mu t}\right)} \le \frac{\partial S(\bar{r},t,s)}{\partial \bar{r}} \le e^{C\varepsilon\left(e^{-\mu s}-e^{-\mu t}\right)}. \tag{5.29}$$

Proof. Recalling that $T(r,t,s) = \Phi_*(\Psi(r,t,s),t,s)$, we have

$$\begin{aligned}\frac{\partial T(r,t,s)}{\partial r} &= \frac{\partial \Phi_*}{\partial \xi}(\Psi(r,t,s),t,s)\frac{\partial \Psi(r,t,s)}{\partial r}\\ &= \frac{\partial \Phi_*}{\partial \xi}(\Psi(r,t,s),t,s)\left[\frac{\partial \Phi}{\partial \xi}(\Psi(r,t,s),t,s)\right]^{-1}\\ &= \exp\left(\int_s^t \left[u_*'(\Phi_*(\xi,\tau,s)) - \frac{\partial w}{\partial r}(\Phi(\xi,\tau,s),\tau)\right]d\tau\right)\Bigg|_{\xi=\Psi(r,t,s)}. \qquad (5.30)\end{aligned}$$

We write

$$\begin{aligned}u_*'(\Phi_*(\xi,\tau,s)) - \frac{\partial w}{\partial r}(\Phi(\xi,\tau,s),\tau) &= [u_*'(\Phi_*(\xi,\tau,s)) - u_*'(\Phi(\xi,\tau,s))]\\ &\quad + \left[u_*'(\Phi(\xi,\tau,s)) - \frac{\partial w}{\partial r}(\Phi(\xi,\tau,s),\tau)\right].\end{aligned}$$

From assumption (5.27) we see that

$$\sup_{\xi\in\mathbb{R}}\left|u_*'(\Phi(\xi,\tau,s)) - \frac{\partial w}{\partial r}(\Phi(\xi,\tau,s),\tau)\right| \le \sup_{0<r<1}\left|u_*'(r) - \frac{\partial w}{\partial r}(r,\tau)\right| \le C\varepsilon e^{-\mu\tau}.$$

Next, by assertion (3) of Lemma 3.1, we know that there exists $0 \le \gamma < 1$ such that $r^\gamma u_*''(r) \in C[0,1]$. With this fact in mind, we use the mean value theorem to compute

$$\begin{aligned}&|u_*'(\Phi_*(\xi,\tau,s)) - u_*'(\Phi(\xi,\tau,s))|\,\big|_{\xi=\Psi(r,t,s)}\\ &= |u_*''(\zeta)|\,|\Phi_*(\xi,\tau,s) - \Phi(\xi,\tau,s)|\,\big|_{\xi=\Psi(r,t,s)}\\ &= |\zeta^\gamma u_*''(\zeta)|\cdot\left[\left(\frac{\Phi(\xi,\tau,s)}{\zeta}\right)^\gamma\cdot(\Phi(\xi,\tau,s))^{-\gamma}\,|\Phi_*(\xi,\tau,s) - \Phi(\xi,\tau,s)|\right]\Bigg|_{\xi=\Psi(r,t,s)},\end{aligned}$$

where $\zeta = \theta\Phi_*(\xi,\tau,s) + (1-\theta)\Phi(\xi,\tau,s)$ for some $0 < \theta < 1$ (depending on ξ, τ, s). Since there exists a constant $0 < c < 1$ such that $\frac{\Phi_*(\xi,\tau,s)}{\Phi(\xi,\tau,s)} \ge c$ for ε sufficiently small, we have $\zeta \ge c\Phi(\xi,\tau,s)$. Thus,

$$\begin{aligned}&\Big|u_*'(\Phi_*(\xi,\tau,s)) - u_*'(\Phi(\xi,\tau,s))\Big|\Big|_{\xi=\Psi(r,t,s)}\\ &\le C\big(\Phi(\xi,\tau,s)\big)^{-\gamma}\big|\Phi_*(\xi,\tau,s) - \Phi(\xi,\tau,s)\big|\big|_{\xi=\Psi(r,t,s)}\\ &\le C\varepsilon\big(e^{-\mu s} - e^{-\mu\tau}\big)\big(\Phi(\xi,\tau,s)\big)^{-\gamma}\Phi_*(\xi,\tau,s)[1-\Phi_*(\xi,\tau,s)]\big|_{\xi=\Psi(r,t,s)}\\ &\le C\varepsilon\big(e^{-\mu s} - e^{-\mu\tau}\big)\big(\Phi(\xi,\tau,s)\big)^{1-\gamma}[1-\Phi(\xi,\tau,s)]\big|_{\xi=\Psi(r,t,s)}\\ &= C\varepsilon\big(e^{-\mu s} - e^{-\mu\tau}\big)\big(\Psi(r,t,\tau)\big)^{1-\gamma}[1-\Psi(r,t,\tau)].\end{aligned}$$

In getting the last equality, we used the following relation:

$$\Phi(\Psi(r,t,s),\tau,s) = \Psi(r,t,\tau) \quad \text{for } 0 \le r \le 1,\ \ s \le \tau \le t. \qquad (5.31)$$

The proof of this relation is as follows: From (5.6) we know that $\rho = \Psi(r,t,\tau)$ is a solution of the following problem:

$$\frac{\partial \rho}{\partial t} + w(r,t)\frac{\partial \rho}{\partial r} = 0 \quad \text{for } 0 \le r \le 1,\ \ t > \tau, \quad \rho|_{t=\tau} = r.$$

But it is easy to verify that $\rho = \Phi(\Psi(r,t,s),\tau,s)$ is also a solution of this problem. Hence, by uniqueness, we have (5.31). Hence, using (5.21) and (5.12), we get

$$\begin{aligned}
&\left|u_*'(\Phi_*(\xi,\tau,s)) - \frac{\partial w}{\partial r}(\Phi(\xi,\tau,s),\tau)\right|\Big|_{\xi=\Psi(r,t,s)} \\
&\le C\varepsilon\left(e^{-\mu s} - e^{-\mu\tau}\right)\left(\Psi_*(r,t,\tau)\right)^{1-\gamma}\left[1-\Psi_*(r,t,\tau)\right] + C\varepsilon e^{-\mu\tau} \\
&= C\varepsilon\left(e^{-\mu s} - e^{-\mu\tau}\right)\left[F_*^{-1}(F_*(r)+t-\tau)\right]^{1-\gamma}\left[1-F_*^{-1}(F_*(r)+t-\tau)\right] + C\varepsilon e^{-\mu\tau}.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\int_s^t \left|u_*'(\Phi_*(\xi,\tau,s)) - \frac{\partial w}{\partial r}(\Phi(\xi,\tau,s),\tau)\right|\Big|_{\xi=\Psi(r,t,s)} d\tau \\
&\le C\varepsilon\left(e^{-\mu s} - e^{-\mu t}\right)\int_s^t \left[F_*^{-1}(F_*(r)+t-\tau)\right]^{1-\gamma}\left[1-F_*^{-1}(F_*(r)+t-\tau)\right] d\tau \\
&\quad + C\varepsilon\int_s^t e^{-\mu\tau}d\tau \\
&= C\varepsilon\left(e^{-\mu s} - e^{-\mu t}\right)\int_{F_*(r)}^{F_*(r)+t-s}\left(F_*^{-1}(\xi)\right)^{1-\gamma}[1-F_*^{-1}(\xi)]d\xi + C\varepsilon\left(e^{-\mu s} - e^{-\mu t}\right) \\
&\quad \left(\xi = F_*(\eta), \quad d\xi = F_*'(\eta)d\eta = \frac{d\eta}{|u_*(\eta)|}\right) \\
&= C\varepsilon\left(e^{-\mu s} - e^{-\mu t}\right)\int_r^{F_*^{-1}(F_*(r)+t-s)} \frac{\eta^{1-\gamma}(1-\eta)}{|u_*(\eta)|}d\eta + C\varepsilon\left(e^{-\mu s} - e^{-\mu t}\right) \\
&\le C\varepsilon\left(e^{-\mu s} - e^{-\mu t}\right)\int_0^1 \eta^{-\gamma}d\eta + C\varepsilon\left(e^{-\mu s} - e^{-\mu t}\right) = C\varepsilon\left(e^{-\mu s} - e^{-\mu t}\right).
\end{aligned}$$

Combining this result with (5.30), we see that (5.28) follows. Finally, (5.29) is an immediate consequence of (5.28). □

COROLLARY 5.8. *Under the assumption of Lemma* 5.7, *for* ε *sufficiently small, we have*

$$\left|\frac{\partial T(r,t,s)}{\partial r} - 1\right| \le C\varepsilon\left(e^{-\mu s} - e^{-\mu t}\right), \quad \left|\frac{\partial S(\bar r,t,s)}{\partial \bar r} - 1\right| \le C\varepsilon\left(e^{-\mu s} - e^{-\mu t}\right),$$

and

$$C^{-1} \le \frac{\partial T(r,t,s)}{\partial r} \le C, \quad C^{-1} \le \frac{\partial S(\bar r,t,s)}{\partial \bar r} \le C. \qquad \square$$

LEMMA 5.9. *Assume that* $a \in C_V^1[0,1]$, *and the assumption of Lemma* 5.7 *holds. Then*

$$\|a(S(\cdot,t,s)) - a\|_\infty \le C\|a\|_1\varepsilon\left(e^{-\mu s} - e^{-\mu t}\right), \tag{5.32}$$

where $\|a\|_1 = \max_{0\le r\le 1} r(1-r)|a'(r)|$. *If further* $r^2(1-r)^2a''(r) \in C[0,1]$, *then we also have*

$$\|a(S(\cdot,t,s)) - a\|_{C_V^1[0,1]} \le C\|a\|_2\varepsilon\left(e^{-\mu s} - e^{-\mu t}\right), \tag{5.33}$$

where $\|a\|_2 = \|a\|_1 + \max_{0\le r\le 1} r^2(1-r)^2|a''(r)|$.

Proof. We have

$$\begin{aligned}|a(S(r,t,s)) - a(r)| &= |a'(\eta)||S(r,t,s) - r| \\ &\le C\varepsilon\eta(1-\eta)|a'(\eta)| \cdot \frac{r(1-r)}{\eta(1-\eta)}\left(e^{-\mu s} - e^{-\mu t}\right) \\ &\le C\|a\|_1\varepsilon\left(e^{-\mu s} - e^{-\mu t}\right),\end{aligned}$$

where $\eta = (1-\theta)r + \theta S(r,t,s)$ for some $0 < \theta < 1$ (depending on r, t, and s). In getting the last inequality, we used the inequality $\eta(1-\eta) \ge Cr(1-r)$ for $0 \le r \le 1$, which follows from (5.20) and the following identity:

$$\eta(1-\eta) = (1-\theta)r(1-r) + \theta S(r,t,s)[1 - S(r,t,s)] + \theta(1-\theta)[r - S(r,t,s)]^2.$$

Hence, (5.32) is proved. Next, we compute

$$\begin{aligned}r(1-r)\left|\frac{\partial a(S(r,t,s))}{\partial r} - a'(r)\right| &= r(1-r)\left|a'(S(r,t,s))\frac{\partial S(r,t,s)}{\partial r} - a'(r)\right| \\ &\le r(1-r)|a'(S(r,t,s))|\left|\frac{\partial S(r,t,s)}{\partial r} - 1\right| \\ &\quad + r(1-r)|a'(S(r,t,s)) - a'(r)|.\end{aligned}$$

By Corollaries 5.3 and 5.8, we see that the first term on the right-hand side is bounded by $C\|a\|_1 \cdot C\varepsilon(e^{-\mu s} - e^{-\mu t})$, and, by a similar argument as before, we can easily show that the second term on the right-hand side is bounded by $C(\max_{0\le r\le 1} r^2(1-r)^2|a''(r)|)\varepsilon(e^{-\mu s} - e^{-\mu t})$. Hence, (5.33) follows. ∎

LEMMA 5.10. *Given $a \in C[0,1]$, we define a bounded linear operator L in $C[0,1]$ by*

$$L(q)(r) = \frac{1}{r^3}\int_0^r a(\rho)q(\rho)\rho^2 d\rho \quad \text{for } q \in C[0,1], \quad 0 < r \le 1,$$

and $L(q)(0) = \lim_{r\to 0^+} L(q)(r) = \frac{1}{3}a(0)q(0)$. Let $\bar r = T(r,t,s)$ and $r = S(\bar r,t,s)$ be as before, and let $\widetilde L$ be the following bounded linear operator in $C[0,1]$:

$$\widetilde L(q)(\bar r) = \frac{1}{r^3}\int_0^r a(\rho)q(T(\rho,t,s))\rho^2 d\rho\Big|_{r=S(\bar r,t,s)} \quad \text{for } q \in C[0,1], \quad 0 < \bar r \le 1,$$

and $\widetilde L(q)(0) = \lim_{\bar r\to 0^+}\widetilde L(q)(\bar r) = \frac{1}{3}a(0)q(0)$. Assume that $a \in C_V^1[0,1]$, and the assumption of Lemma 5.7 holds. Then both L and $\widetilde L$ are bounded linear operators from $C[0,1]$ to $C_V^1[0,1]$, and we have

$$\left\|\widetilde L - L\right\|_{L(C[0,1],C_V^1[0,1])} \le C\|a\|_{C_V^1[0,1]}\varepsilon\left(e^{-\mu s} - e^{-\mu t}\right). \tag{5.34}$$

Proof. We give only the proof of (5.34), because the proof of the assertion that both L and $\widetilde L$ are bounded linear operators from $C[0,1]$ to $C_V^1[0,1]$ follows by a similar argument.

We first note that, for $q \in C[0,1]$ and $0 < \bar r \le 1$, $\widetilde L(q)(\bar r)$ can be rewritten as follows:

$$\widetilde L(q)(\bar r) = \frac{1}{[S(\bar r,t,s)]^3}\int_0^{\bar r} a(S(\rho,t,s))q(\rho)\left[\frac{S(\rho,t,s)}{\rho}\right]^2\frac{\partial S(\rho,t,s)}{\partial\rho}\rho^2 d\rho.$$

Thus,

$$\begin{aligned}\widetilde{L}(q)(\bar{r})-L(q)(\bar{r})=&\left[\frac{\bar{r}}{S(\bar{r},t,s)}\right]^3\cdot\frac{1}{\bar{r}^3}\int_0^{\bar{r}}a(S(\rho,t,s))q(\rho)\\&\left[\frac{S(\rho,t,s)}{\rho}\right]^2\left[\frac{\partial S(\rho,t,s)}{\partial\rho}-1\right]\rho^2d\rho\\&+\left[\frac{\bar{r}}{S(\bar{r},t,s)}\right]^3\cdot\frac{1}{\bar{r}^3}\int_0^{\bar{r}}a(S(\rho,t,s))q(\rho)\left\{\left[\frac{S(\rho,t,s)}{\rho}\right]^2-1\right\}\rho^2d\rho\\&+\left[\frac{\bar{r}}{S(\bar{r},t,s)}\right]^3\cdot\frac{1}{\bar{r}^3}\int_0^{\bar{r}}[a(S(\rho,t,s))-a(\rho)]q(\rho)\rho^2d\rho\\&+\left\{\left[\frac{\bar{r}}{S(\bar{r},t,s)}\right]^3-1\right\}\cdot\frac{1}{\bar{r}^3}\int_0^{\bar{r}}a(\rho)q(\rho)\rho^2d\rho.\end{aligned}$$

From Corollary 5.3, Lemma 5.5, Corollary 5.8, and Lemma 5.9, we know that

$$\begin{aligned}\left|\frac{\bar{r}}{S(\bar{r},t,s)}\right|\le C,\quad&\left|\frac{\bar{r}}{S(\bar{r},t,s)}-1\right|\le C\varepsilon\left(e^{-\mu s}-e^{-\mu t}\right),\\\left|\frac{S(\rho,t,s)}{\rho}\right|\le C,\quad&\left|\frac{\partial S(\rho,t,s)}{\partial\rho}-1\right|\le C\varepsilon\left(e^{-\mu s}-e^{-\mu t}\right),\\|a(S(\rho,t,s))-a(\rho)|\le C\|a\|_1\varepsilon&\left(e^{-\mu s}-e^{-\mu t}\right).\end{aligned}$$

Using the above estimates, we see easily that

$$\max_{0\le\bar{r}\le1}\left|\widetilde{L}(q)(\bar{r})-L(q)(\bar{r})\right|\le C\varepsilon\left(e^{-\mu s}-e^{-\mu t}\right)\|a\|_{C_V^1[0,1]}\|q\|_\infty. \tag{5.35}$$

Next, by a simple computation, we have

$$\begin{aligned}\bar{r}(1-\bar{r})L(q)'(\bar{r})=&(1-\bar{r})a(\bar{r})q(\bar{r})-\frac{3(1-\bar{r})}{\bar{r}^3}\int_0^{\bar{r}}a(\rho)q(\rho)\rho^2d\rho,\\\bar{r}(1-\bar{r})\widetilde{L}(q)'(\bar{r})=&\frac{\bar{r}(1-\bar{r})}{S(\bar{r},t,s)}a(S(\bar{r},t,s))q(\bar{r})\frac{\partial S(\bar{r},t,s)}{\partial\bar{r}}\\&-\frac{3\bar{r}(1-\bar{r})}{[S(\bar{r},t,s)]^4}\frac{\partial S(\bar{r},t,s)}{\partial\bar{r}}\int_0^{\bar{r}}a(S(\rho,t,s))q(\rho)\\&\left[\frac{S(\rho,t,s)}{\rho}\right]^2\frac{\partial S(\rho,t,s)}{\partial\rho}\rho^2d\rho.\end{aligned}$$

Using these expressions and a similar argument as before, we have

$$\sup_{0<\bar{r}<1}\bar{r}(1-\bar{r})\left|\widetilde{L}(q)'(\bar{r})-L(q)'(\bar{r})\right|\le C\varepsilon\left(e^{-\mu s}-e^{-\mu t}\right)\|a\|_{C_V^1[0,1]}\|q\|_\infty. \tag{5.36}$$

To save space, we omit the details here. By (5.35) and (5.36), we see that (5.34) follows. □

6. Decay estimates. In this section we establish a decay estimate for the evolution system $\{\mathbb{U}(t,s,V):t\ge s\ge0\}$ obtained in section 4, where $V=V(t)\in C([0,\infty),S_\varepsilon)$, under an additional assumption that $V(t)$ is exponentially decaying as $t\to\infty$.

We first consider the special case where $V = 0$. In this case we have $\mathbb{U}(t, s, V) = e^{(t-s)\mathbb{A}(0)}$. The main result, Theorem 5.1 of [5], gives a decay estimate for $e^{t\mathbb{A}(0)}$ (see (6.8) below). But that estimate contains some singularity at $r = 0$, so that it does not meet our requirement. In what follows we use the fact that $\lim_{r\to 0^+} rp'_*(r) = 0$ ensured by Lemma 3.1 to establish an improved estimate. To this end we need a preliminary lemma, which gives an estimate for the semigroup generated by the following operator $\mathcal{L} = \mathcal{L}_0 + a$:

$$\mathcal{L}q(r) = -w(r)q'(r) + a(r)q(r) \quad \text{for } 0 \le r \le 1,$$

where w and a are given functions.

LEMMA 6.1. *Assume that $w \in C^1[0,1]$ and satisfies (4.3), and $a \in C^1_V[0,1]$. Then $\mathcal{L}$ generates a C_0 semigroup $e^{t\mathcal{L}}$ on $C[0,1]$ satisfying the following estimate:*

$$\left\|e^{t\mathcal{L}}\right\|_{L(C[0,1])} \le e^{\omega_0 t} \quad \textit{for } t \ge 0, \tag{6.1}$$

where $\omega_0 = \max_{0\le r\le 1} a(r)$. Moreover, $C^1_V[0,1]$ is $\mathcal{L}$-admissible, and, for any $\omega > \omega_0$, we have

$$\left\|e^{t\mathcal{L}}\right\|_{L(C^1_V[0,1])} \le C_\omega e^{\omega t} \quad \textit{for } t \ge 0. \tag{6.2}$$

Here, C_ω is independent of the function w.

The proof is similar to that of Lemma 4.1, so that is omitted. □

LEMMA 6.2. *There exists a constant $\mu^* > 0$ such that, for any $0 < \mu < \mu^*$, the semigroup $e^{t\mathbb{A}(0)}$ $(t \ge 0)$ generated by $\mathbb{A}(0)$ satisfies the following estimate:*

$$\left\|e^{t\mathbb{A}(0)}\right\|_{L(C[0,1])} \le Ce^{-\mu t} \quad \textit{for } t \ge 0. \tag{6.3}$$

Proof. Given $U_0 = (\phi_0, \zeta_0) \in X$, let $U(t) = e^{t\mathbb{A}(0)}U_0 = (\phi(r,t), \zeta(t))$. Then (ϕ, ζ) is the unique solution of the following initial value problem:

$$\partial_t\phi + u_*(r)\partial_r\phi = a(r)\phi + \mathcal{B}(\phi) + b(r)\zeta \quad \text{for } 0 \le r \le 1, \ t > 0, \tag{6.4}$$

$$\frac{d\zeta}{dt} = \mathcal{F}(\phi) + \kappa\zeta \quad \text{for } t > 0, \tag{6.5}$$

$$\phi(r,0) = \phi_0(r) \quad \text{for } 0 \le r \le 1, \quad \zeta(0) = \zeta_0, \tag{6.6}$$

where $a(r), b(r), \mathcal{B}(\phi), \mathcal{F}(\phi)$, and κ are given in (2.19)–(2.23). By Theorem 5.1 of [5] and the remark in the end of section 8 of [5], we know that there exists a constant $\sigma^* > 0$ and a function $\hat\phi \in C^1(0,1]$ satisfying

$$\hat\phi(r) > 0 \quad \text{for } 0 < r \le 1, \quad \hat\phi(r) \sim Cr^{-\theta} \quad \text{for } r \to 0 \tag{6.7}$$

for some constants $1 \le \theta < 3$ and $C > 0$, such that the solution of the above problem satisfies the following estimate:

$$|\zeta(t)| + \sup_{0<r\le 1}\left|\frac{\phi(r,t)}{\hat\phi(r)}\right| \le C\left(|\zeta_0| + \sup_{0<r\le 1}\left|\frac{\phi_0(r)}{\hat\phi(r)}\right|\right)(1+t)^2e^{-\sigma^* t} \quad \text{for } t \ge 0. \tag{6.8}$$

This particularly implies that, for any $0 < \sigma < \sigma^*$ and $\delta \in (0,1)$, we have

$$|\zeta(t)| + \sup_{\delta\le r\le 1}|\phi(r,t)| \le C\left(|\zeta_0| + \sup_{0\le r\le 1}|\phi_0(r)|\right)e^{-\sigma t} \quad \text{for } t \ge 0, \tag{6.9}$$

because $1/\hat{\phi}(r)$ has a positive lower bound for $\delta \le r \le 1$ and a finite upper bound for $0 \le r \le 1$. In what follows we prove that, for δ sufficiently small, there also holds

$$\sup_{0\le r\le\delta} |\phi(r,t)| \le Ce^{-\mu t} \quad \text{for } t \ge 0 \tag{6.10}$$

for some $\mu > 0$.

Take a nonnegative cut-off function $\varphi \in C[0,1]$ such that

$$\varphi(r) \le 1 \quad \text{for } 0 \le r \le 1, \quad \varphi(r) = 1 \quad \text{for } 0 \le r \le \delta, \quad \varphi(r) = 0 \quad \text{for } 2\delta \le r \le 1.$$

We split $\mathcal{B}$ into a sum of two operators as follows:

$$\mathcal{B}(q) = \mathcal{B}_1(q) + \mathcal{B}_2(q) \quad \text{for } q \in C[0,1], \tag{6.11}$$

where

$$\mathcal{B}_1(q) = -rp'_*(r)\varphi(r)\cdot\frac{1}{r^3}\int_0^{\min\{r,\delta\}} g_p(\rho)q(\rho)\rho^2 d\rho,$$

$$\mathcal{B}_2(q) = rp'_*(r)\int_0^1 g_p(\rho)q(\rho)\rho^2 d\rho - r^{-2}p'_*(r)[1-\varphi(r)]\int_0^{\min\{r,\delta\}} g_p(\rho)q(\rho)\rho^2 d\rho$$
$$- r^{-2}p'_*(r)\int_{\min\{r,\delta\}}^r g_p(\rho)q(\rho)\rho^2 d\rho$$

and introduce $f(r,t) = \mathcal{B}_2(\phi(\cdot,t))(r) + b(r)\zeta(t)$. By (6.4) and (6.11), we see that ϕ is the solution of the equation

$$\partial_t\phi + u_*(r)\partial_r\phi = a(r)\phi + \mathcal{B}_1(\phi) + f(r,t) \quad \text{for } 0 \le r \le 1, \ \ t > 0 \tag{6.12}$$

subject to the initial condition $\phi(r,0) = \phi_0(r)$. Introducing operators $\mathcal{L}(q) = -u_*q' + aq$, we see that (6.12) can be rewritten as the following differential equation in $C[0,1]$:

$$\frac{d\phi(\cdot,t)}{dt} = (\mathcal{L}+\mathcal{B}_1)(\phi(\cdot,t)) + f(\cdot,t). \tag{6.13}$$

Using Lemma 6.1 we see that the operator $\mathcal{L}$ generates a strongly continuous semigroup $e^{t\mathcal{L}}$ on $C[0,1]$, and $\|e^{t\mathcal{L}}\| \le e^{-\omega t}$ for all $t \ge 0$ where $\omega = \min_{0\le r\le 1}|a(r)| > 0$ and using the facts that $\lim_{r\to 0} rp'_*(r) = 0$ and $\varphi(r) = 0$ for $r \ge 2\delta$, we can easily deduce that, for any given $\varepsilon > 0$, there exists corresponding $\delta > 0$ such that $\|\mathcal{B}_1(q)\|_\infty \le \varepsilon\|q\|_\infty$, or, in other words, $\mathcal{B}_1$ is a bounded linear operator on $C[0,1]$, with norm dorminated by ε. Thus, the operator $\mathcal{L}+\mathcal{B}_1$ also generates a strongly continuous semigroup $e^{t(\mathcal{L}+\mathcal{B}_1)}$ on $C[0,1]$, and it satisfies

$$\|e^{t(\mathcal{L}+\mathcal{B}_1)}\| \le e^{-(\omega-\varepsilon)t} \quad \text{for all } \ t \ge 0. \tag{6.14}$$

In what follows we assume that ε is sufficiently small such that $\omega - \varepsilon > 0$. By (6.13), we have

$$\phi(\cdot,t) = e^{t(\mathcal{L}+\mathcal{B}_1)}\phi_0 + \int_0^t e^{(t-\tau)(\mathcal{L}+\mathcal{B}_1)} f(\cdot,\tau)d\tau. \tag{6.15}$$

Using (6.7), (6.8), and (6.9), we can easily show that $\|f(\cdot,t)\|_\infty \le C_\delta\|U_0\|e^{-\sigma t}$ for $t \ge 0$. Hence, from (6.15) and (6.14) we see that, for any $0 < \mu < \min\{\sigma, \omega-\varepsilon\}$, there holds

$$\|\phi(\cdot,t)\|_\infty \le \|\phi_0\|_\infty e^{-(\omega-\varepsilon)t} + C\|U_0\|e^{-\mu t} \quad \text{for } \ t \ge 0,$$

from which (6.10) immediately follows.

Combining (6.9) and (6.10), we get (6.3). This completes the proof. □

LEMMA 6.3. *Let μ^* be as in Lemma 6.2. Then, for any $0 < \mu < \mu^*$, in addition to (6.3), we also have the following estimate:*

$$\left\|e^{t\mathbb{A}(0)}\right\|_{L(C_V^1[0,1])} \le Ce^{-\mu t} \quad \text{for } t \ge 0. \tag{6.16}$$

Proof. We first show that $\mathcal{B}$ and $\mathcal{F}$ satisfy the following properties: For any $q \in C_V^1[0,1]$,

$$\|\mathcal{B}(u_*q')\|_\infty + \|u_*\mathcal{B}(q)'\|_\infty \le C\|q\|_\infty, \tag{6.17}$$

$$|\mathcal{F}(u_*q')| \le C\|q\|_\infty. \tag{6.18}$$

Using the facts that $\lim_{r\to 0} rp_*'(r) = 0$ and $\lim_{r\to 0} r^2p_*''(r) = 0$ (see (3.3)), we can easily prove that $\|u_*\mathcal{B}(q)'\|_\infty \le C\|q\|_\infty$ for $q \in C[0,1]$. To estimate $\|\mathcal{B}(u_*q')\|_\infty$, we compute

$$\frac{1}{r^3}\int_0^r g_p(\rho)u_*(\rho)q'(\rho)\rho^2 d\rho = \frac{1}{r}u_*(r)g_p(r)q(r) - \frac{1}{r^3}\int_0^r m(\rho)q(\rho)\rho^2 d\rho,$$

where $m(\rho) = g_p'(\rho)u_*(\rho) + g_p(\rho)u_*'(\rho) + \frac{2}{\rho}u_*(\rho)g_p(\rho)$. Taking $r = 1$ we particularly obtain

$$\int_0^1 g_p(\rho)u_*(\rho)q'(\rho)\rho^2 d\rho = u_*(1)g_p(1)q(1) - \int_0^1 m(\rho)q(\rho)\rho^2 d\rho.$$

Since $g_p \in C^1[0,1]$, $u_* \in C^1[0,1]$, and $u_*(0) = 0$, we see that $\frac{1}{r}u_*g_p$ and m both belong to $C[0,1]$. Hence, from the above expressions we see immediately that

$$\begin{aligned}\|\mathcal{B}(u_*q')\|_\infty &= \sup_{0<r<1}\left|rp_*'(r)\left[\int_0^1 g_p(\rho)u_*(\rho)q'(\rho)\rho^2 d\rho - \frac{1}{r^3}\int_0^r g_p(\rho)u_*(\rho)q'(\rho)\rho^2 d\rho\right]\right| \\ &\le C\|q\|_\infty.\end{aligned}$$

Similarly we have $|\mathcal{F}(u_*q')| = |\int_0^1 g_p(\rho)u_*(\rho)q'(\rho)\rho^2 d\rho| \le C\|q\|_\infty$. This proves (6.17) and (6.18).

We now proceed to prove (6.16). Let $U_0 \in X_0$ and $U = e^{t\mathbb{A}(0)}U_0$. From the proof of Lemma 4.2 we know that $U \in C([0,\infty), X_0) \cap C^1([0,\infty), X)$. Let $U_0 = (q_0, y_0)$ and $U = (q, y)$. Then (q, y) is the solution of the following problem:

$$\begin{cases} \dfrac{\partial q}{\partial t} + u_*(r)\dfrac{\partial q}{\partial r} = a(r)q + \mathcal{B}(q) + b(r)y \quad \text{for } 0 \le r \le 1, \ t > 0, \\ \dfrac{dy}{dt} = \mathcal{F}(q) + \kappa y \quad \text{for } t > 0, \\ q|_{t=0} = q_0(r) \quad \text{for } 0 \le r \le 1, \quad \text{and} \quad y|_{t=0} = y_0. \end{cases}$$

Let $l(r,t) = u_*(r)\frac{\partial q(r,t)}{\partial r}$. By formally differentiating the first equation above in r and multiplying it with $u_*(r)$, we see that (l, y) is a formal solution of the following problem:

$$\begin{cases} \dfrac{\partial l}{\partial t} + u_*(r)\dfrac{\partial l}{\partial r} = a(r)l + \mathcal{B}(l) + b(r)y + f_1(r,t) \quad \text{for } 0 \le r \le 1, \ t > 0, \\ \dfrac{dy}{dt} = \mathcal{F}(l) + \kappa y + c_1(t) \quad \text{for } t > 0, \\ l|_{t=0} = l_0(r) \quad \text{for } 0 \le r \le 1, \quad \text{and} \quad y|_{t=0} = y_0, \end{cases}$$

where $l_0(r) = u_*(r)q_0'(r)$, $c_1(t) = \mathcal{F}(q) - \mathcal{F}(u_*\frac{\partial q}{\partial r})$, and

$$f_1(r,t) = u_*(r)a'(r)q(r,t) - \mathcal{B}\left(u_*\frac{\partial q}{\partial r}\right) + u_*(r)\frac{\partial \mathcal{B}(q)}{\partial r} + [u_*(r)b'(r) - b(r)]y(t).$$

We denote $W(t) = (l(\cdot,t), y(t))$, $W_0 = (l_0, y_0)$, and $F_1(t) = (f_1(\cdot,t), c_1(t))$. Then the above problem can be rewritten as follows:

$$\frac{dW}{dt} = \mathbb{A}(0)W + F_1(t) \quad \text{for } t > 0, \quad W(0) = W_0.$$

Using the fact that $U \in C^1([0,\infty), X)$, Corollary 3.2, (6.17), and (6.18), we can easily prove that $F_1 \in C^1([0,\infty), X)$. Thus, by a standard result, we see that the above problem has a unique mild solution, and consequently, the above formal computation makes sense. Moreover, we have

$$W(t) = e^{t\mathbb{A}(0)}W_0 + \int_0^t e^{(t-s)\mathbb{A}(0)}F_1(s)ds \quad \text{for } t \geq 0.$$

From this expression and Lemma 6.2 we see that, for any given $0 < \mu < \mu_*$, we have

$$\|W(t)\|_X \leq Ce^{-\mu t}\|W_0\|_X + C\int_0^t e^{-\mu(t-s)}\|F_1(s)\|_X ds \quad \text{for } t \geq 0.$$

Using (6.17), (6.18), and the fact that $\|U(t)\|_X \leq Ce^{-\mu t}\|U_0\|_X$ ensured by (6.3), we see that

$$\|F_1(t)\|_X \leq C\|U(t)\|_X \leq Ce^{-\mu t}\|U_0\|_X \quad \text{for } t \geq 0.$$

Substituting this estimate into the above inequality and noticing that $W(t) = (u_*\frac{\partial q(\cdot,t)}{\partial r}, y(t))$ and $W_0 = (u_*q_0', y_0)$, we obtain

$$\|U(t)\|_{X_0} \leq C(1+t)e^{-\mu t}\|U_0\|_{X_0} \quad \text{for } t \geq 0.$$

Now, for any given $0 < \mu < \mu_*$, we arbitrarily take a $\bar{\mu} \in (\mu, \mu_*)$ and first use the above estimate to $\bar{\mu}$ and next use the elementary inequality $(1+t)e^{-\bar{\mu}t} \leq Ce^{-\mu t}$, we see that (6.16) follows. This completes the proof. □

In what follows we consider the evolution system $\mathbb{U}(t,s,V)$ for a general $V = V(t) \in C([0,\infty), X)$ satisfying the following condition: For some positive constants $\bar{\mu}$, ε and C_0,

$$\|V(t)\|_X \leq C_0\varepsilon e^{-\bar{\mu}t} \quad \text{for } t \geq 0. \tag{6.19}$$

LEMMA 6.4. *Assume that $V = V(t) \in C([0,\infty), X)$, and it satisfies* (6.19). *Let μ^* be as in Lemma* 6.2. *Then, for any $0 < \mu < \mu^*$, there exists corresponding $\varepsilon_0 > 0$ (depending on μ, $\bar{\mu}$, and C_0) such that if $0 < \varepsilon \leq \varepsilon_0$, then the following estimates hold:*

$$\|\mathbb{U}(t,s,V)\|_{L(X)} \leq C_1e^{-\mu t} \quad \textit{for } t \geq 0, \tag{6.20}$$

$$\|\mathbb{U}(t,s,V)\|_{L(X_0)} \leq C_2e^{-\mu t} \quad \textit{for } t \geq 0, \tag{6.21}$$

where C_1 and C_2 are positive constants depending only on μ and independent of $\bar{\mu}$ and C_0.

Proof. Given $0 < \mu < \mu^*$ we take a $\mu_1 \in (\mu, \mu^*)$ and fix it. By Lemmas 6.2 and 6.3, we have the following estimates:

$$\left\|e^{t\mathbb{A}(0)}\right\|_{L(C[0,1])} \le C_1 e^{-\mu_1 t} \quad \text{for } t \ge 0, \tag{6.22}$$

$$\left\|e^{t\mathbb{A}(0)}\right\|_{L(C^1_V[0,1])} \le C_2 e^{-\mu_1 t} \quad \text{for } t \ge 0. \tag{6.23}$$

Let $U_0 = (q_0, s_0)$ be an arbitrary point in X, and let $U = \mathbb{U}(t, s, V)U_0$. By definition, U is the solution of the problem (4.12). Let $U = (q, y)$. Then (4.12) can be rewritten as follows:

$$\begin{cases} \dfrac{\partial q}{\partial t} + w_V(r,t)\dfrac{\partial q}{\partial r} = a(r)q + \mathcal{B}q + b(r)y \quad \text{for } 0 \le r \le 1, \ \ t > s, \\ \dfrac{dy}{dt} = \mathcal{F}(q) + \kappa y \quad \text{for } t > s, \\ q|_{t=s} = q_0(r) \quad \text{for } 0 \le r \le 1, \quad \text{and} \quad y|_{t=s} = y_0. \end{cases} \tag{6.24}$$

Let $\bar r = T(r,t,s)$ and $r = S(\bar r, t, s)$ be as defined in section 5, and let $\widetilde q(\bar r, t, s) = q(S(\bar r, t, s), t)$, or, equivalently, $q(r,t) = \widetilde q(T(r,t,s), t, s)$. Then by using (5.7), we see that (6.24) is transformed into the following problem:

$$\begin{cases} \dfrac{\partial \widetilde q}{\partial t} + u_*(\bar r)\dfrac{\partial \widetilde q}{\partial \bar r} = \widetilde a(\bar r, t, s)\widetilde q + \widetilde{\mathcal{B}}\widetilde q + \widetilde b(\bar r, t, s)s \quad \text{for } 0 \le \bar r \le 1, \ \ t > s, \\ \dfrac{ds}{dt} = \widetilde{\mathcal{F}}(\widetilde q)(t,s) + \kappa s \quad \text{for } t > s, \\ \widetilde q|_{t=s} = q_0(\bar r) \quad \text{for } 0 \le \bar r \le 1, \quad \text{and} \quad s|_{t=s} = s_0, \end{cases} \tag{6.25}$$

where $\widetilde a(\bar r, t, s) = a(S(\bar r, t, s))$, $\widetilde b(\bar r, t, s) = b(S(\bar r, t, s))$,

$$\widetilde{\mathcal{B}}\widetilde q = rp'_*(r)\left[\int_0^1 g_p(\rho)\widetilde q(T(\rho,t,s),t,s)\rho^2 d\rho - \frac{1}{r^3}\int_0^r g_p(\rho)\widetilde q(T(\rho,t,s),t,s)\rho^2 d\rho\right]\Bigg|_{r=S(\bar r,t,s)},$$

and $\widetilde{\mathcal{F}}(\widetilde q)(t,s) = \int_0^1 g_p(\rho)\widetilde q(T(\rho,t,s),t,s)\rho^2 d\rho$. We define a family of bounded linear operators $\widetilde{\mathbb{B}}(t,s,V) : X \to X$ $(t \ge s \ge 0)$ as follows:

$$\widetilde{\mathbb{B}}(t,s,V) = \begin{pmatrix} \widetilde a(\cdot,t,s) + \widetilde{\mathcal{B}} & \widetilde b(\cdot,t,s) \\ \widetilde{\mathcal{F}} & \kappa \end{pmatrix}.$$

We also denote $\widetilde U = (\widetilde q, y)$. Then (6.25) can be rewritten as follows:

$$\frac{d\widetilde U}{dt} = \mathbb{A}_0(U_*)\widetilde U + \widetilde{\mathbb{B}}(t,s,V)\widetilde U \quad \text{for } t > s, \quad \widetilde U|_{t=s} = U_0. \tag{6.26}$$

Recalling that $\mathbb{A}(0) = \mathbb{A}_0(U_*) + \mathbb{B}$ and denoting

$$\widetilde{\mathbb{E}}(t,s,V) = \widetilde{\mathbb{B}}(t,s,V) - \mathbb{B} = \begin{pmatrix} \widetilde a(\cdot,t,s) - a + \widetilde{\mathcal{B}} - \mathcal{B} & \widetilde b(\cdot,t,s) - b \\ \widetilde{\mathcal{F}} - \mathcal{F} & 0 \end{pmatrix},$$

we see that

$$\mathbb{A}_0(U_*) + \widetilde{\mathbb{B}}(t,s,V) = \mathbb{A}_0(U_*) + \mathbb{B} + \widetilde{\mathbb{E}}(t,s,V) = \mathbb{A}(0) + \widetilde{\mathbb{E}}(t,s,V).$$

Hence, (6.26) can be further rewritten as follows:

$$\frac{d\widetilde{U}}{dt} = \mathbb{A}(0)\widetilde{U} + \widetilde{\mathbb{E}}(t,s,V)\widetilde{U} \quad \text{for } t > s, \quad \widetilde{U}|_{t=s} = U_0. \tag{6.27}$$

We know that (6.27) is equivalent to the following integral equation:

$$\widetilde{U}(t,s) = e^{(t-s)\mathbb{A}(0)}U_0 + \int_s^t e^{(t-\tau)\mathbb{A}(0)}\widetilde{\mathbb{E}}(\tau,s,V)\widetilde{U}(\tau,s) \quad \text{for } t \geq s. \tag{6.28}$$

It can be easily shown that, under the assumption (6.19), w_V satisfies the estimates (5.1) and (5.27) (with μ replaced by $\bar{\mu}$). Hence, by Corollary 3.2 and Lemma 5.9, we have

$$\|\widetilde{a}(\cdot,t,s) - a\|_{C_V^1[0,1]} \leq C\varepsilon, \qquad \big\|\widetilde{b}(\cdot,t,s) - b\big\|_{C_V^1[0,1]} \leq C\varepsilon,$$

and by Corollary 3.2, Lemma 5.9, and Lemma 5.10, we have

$$\big\|\widetilde{\mathcal{B}} - \mathcal{B}\big\|_{L(C[0,1])} \leq C\varepsilon, \quad \big\|\widetilde{\mathcal{B}} - \mathcal{B}\big\|_{L(C_V^1[0,1])} \leq C\varepsilon, \quad \big\|\widetilde{\mathcal{F}} - \mathcal{F}\big\|_{L(C[0,1],\mathbb{R})} \leq C\varepsilon.$$

It follows that

$$\big\|\widetilde{\mathbb{E}}(\tau,s,V)\big\|_{L(X)} \leq C\varepsilon, \quad \big\|\widetilde{\mathbb{E}}(\tau,s,V)\big\|_{L(X_0)} \leq C\varepsilon. \tag{6.29}$$

From (6.22), (6.23), (6.28), and (6.29) we obtain

$$\big\|\widetilde{U}(t,s)\big\|_X \leq C_1 e^{-\mu_1(t-s)}\|U_0\|_X + C\varepsilon\int_s^t e^{-\mu_1(t-\tau)}\big\|\widetilde{U}(\tau,s)\big\|_X,$$

$$\big\|\widetilde{U}(t,s)\big\|_{X_0} \leq C_2 e^{-\mu_1(t-s)}\|U_0\|_{X_0} + C\varepsilon\int_s^t e^{-\mu_1(t-\tau)}\big\|\widetilde{U}(\tau,s)\big\|_{X_0}.$$

By the Gronwall lemma, these inequalities yield

$$\big\|\widetilde{U}(t,s)\big\|_X \leq C_1 e^{-(\mu_1 - C\varepsilon)t}\|U_0\|_X, \quad \big\|\widetilde{U}(t,s)\big\|_{X_0} \leq C_2 e^{-(\mu_1 - C\varepsilon)t}\|U_0\|_{X_0}.$$

Hence, by taking ε sufficiently small such that $\mu_1 - C\varepsilon \geq \mu$, we obtain (6.20) and (6.21). This completes the proof. □

7. The proof of Theorem 1.1. In order to prove Theorem 1.1, we let μ^* be as in Lemma 6.2 and arbitrarily fix a number $0 < \mu < \mu^*$. Let ε be a positive number to be specified later. For any fixed $U_0 \in X_0$ satisfying $\|U_0\|_{X_0} \leq \varepsilon$, we denote by $\mathbf{M}$ the set of all functions $V = V(t) \in C([0,\infty), X)$ satisfying the following conditions:

$$V(0) = U_0, \quad \|V(t)\|_X \leq 2C_1\varepsilon e^{-\mu t} \quad \text{for } t \geq 0, \tag{7.1}$$

where C_1 is the constant appearing in (6.20). We introduce a metric d on $\mathbf{M}$ by defining $d(V_1,V_2) = \sup_{t\geq 0} e^{\mu t}\|V_1(t) - V_2(t)\|_X$ for $V_1, V_2 \in \mathbf{M}$. It is evident that $(\mathbf{M}, d)$ is a complete metric space. Given $V \in \mathbf{M}$, we consider the following initial value problem:

$$\frac{dU(t)}{dt} = \mathbb{A}(V(t))U(t) + \mathbb{G}(U(t)) \quad \text{for } t > 0, \quad U(0) = U_0. \tag{7.2}$$

LEMMA 7.1. *If ε is sufficiently small, then, for any $V \in \mathbf{M}$, problem (7.2) has a unique solution $U \in C([0,\infty), X_0) \cap C^1([0,\infty), X)$, which satisfies the following estimates:*

(7.3)
$$\|U(t)\|_X \le 2C_1\varepsilon e^{-\mu t}, \quad \|U(t)\|_{X_0} \le C\varepsilon e^{-\mu t}, \quad \|U'(t)\|_X \le C\varepsilon e^{-\mu t} \quad \text{for } t \ge 0,$$

where C_1 is as before, and C is another constant independent of V.

Proof. We denote

$$\widetilde{\mathbf{M}} = \{U \in C([0,\infty), X_0) : \|U(t)\|_X \le 2C_1\varepsilon e^{-\mu t} \text{ and } \|U(t)\|_{X_0} \le 2C_2\varepsilon e^{-\mu t} \text{ for } t \ge 0\}$$

and introduce a metric d on it by defining $d(U_1, U_2) = \sup_{t\ge 0} e^{\mu t}\|U_1(t) - U_2(t)\|_{X_0}$ for $U_1, U_2 \in \widetilde{\mathbf{M}}$. Here, C_1 and C_2 are positive constants appearing in (6.20) and (6.21), respectively. $(\widetilde{\mathbf{M}}, d)$ is clearly a complete metric space. Given $U \in \widetilde{\mathbf{M}}$, we consider the following initial value problem:

$$\frac{d\widetilde{U}(t)}{dt} = \mathbb{A}(V(t))\widetilde{U}(t) + \mathbb{G}(U(t)) \quad \text{for } t > 0, \quad \widetilde{U}(0) = U_0. \tag{7.4}$$

Since $U(t) \in C([0,\infty), X_0)$, by Corollary 3.3, we have $\mathbb{G}(U(t)) \in C([0,\infty), X_0)$. It follows by Corollary 4.4 that the above problem has a unique solution $\widetilde{U} \in C([0,\infty), X_0) \cap C^1([0,\infty), X)$ and is given by

$$\widetilde{U}(t) = \mathbb{U}(t,0,V)U_0 + \int_0^t \mathbb{U}(t,s,V)\mathbb{G}(U(s))ds. \tag{7.5}$$

Using this expression and Lemma 6.4 and (2.18) we have

$$\begin{aligned}\left\|\widetilde{U}(t)\right\|_X &\le C_1 e^{-\mu t}\|U_0\|_X + C_1\int_0^t e^{-\mu(t-s)}\|\mathbb{G}(U(s))\|_X ds \\ &\le C_1\varepsilon e^{-\mu t} + C\int_0^t e^{-\mu(t-s)}\|U(s)\|_X^2 ds \le 2C_1\varepsilon e^{-\mu t},\end{aligned}$$

when ε is sufficiently small. Similarly, by using Lemma 6.4 and (3.9), we also have $\|\widetilde{U}(t)\|_{X_0} \le 2C_2\varepsilon e^{-\mu t}$, when ε is sufficiently small. Hence, $\widetilde{U} \in \widetilde{\mathbf{M}}$. We now define a mapping $\widetilde{\mathbf{S}} : \widetilde{\mathbf{M}} \to \widetilde{\mathbf{M}}$ by setting $\widetilde{\mathbf{S}}(U) = \widetilde{U}$ for every $U \in \widetilde{\mathbf{M}}$. We claim that $\widetilde{\mathbf{S}}$ is a contraction mapping. Indeed, for any $U_1, U_2 \in \widetilde{\mathbf{M}}$, let $\widetilde{U}_1 = \mathbf{S}(U_1)$, $\widetilde{U}_2 = \mathbf{S}(U_2)$, and $W = \widetilde{U}_1 - \widetilde{U}_2$. Then W satisfies

$$\frac{dW(t)}{dt} = \mathbb{A}(V(t))W(t) + [\mathbb{G}(U_1(t)) - \mathbb{G}(U_2(t))] \quad \text{for } t > 0, \quad W(0) = 0,$$

so that $W(t) = \int_0^t \mathbb{U}(t,s,V)[\mathbb{G}(U_1(s)) - \mathbb{G}(U_2(s))]ds$. It follows that

$$\begin{aligned}\|W(t)\|_{X_0} &\le C_2\int_0^t e^{-\mu(t-s)}\|\mathbb{G}(U_1(s) - \mathbb{G}(U_2(s)))\|_{X_0} ds \\ &\le C_2\int_0^t e^{-\mu(t-s)}\|U_1(s) - U_2(s)\|_{X_0} \\ &\quad \left(\int_0^1 \|\mathbb{G}'(\theta U_1(s) + (1-\theta)U_2(s))\|_{L(X_0)} d\theta\right) ds \\ &\le C\int_0^t e^{-\mu(t-s)}\|U_1(s) - U_2(s)\|_{X_0}\left(\int_0^1 \|\theta U_1(s) + (1-\theta)U_2(s)\|_{X_0} d\theta\right) ds,\end{aligned}$$

which yields $\|W(t)\|_{X_0} \le C\varepsilon d(U_1,U_2)e^{-\mu t}$. Thus, for ε sufficiently small, we have $d(\widetilde{U}_1,\widetilde{U}_2) = \sup_{t\ge 0} e^{\mu t}\|W(t)\|_{X_0} \le \frac{1}{2}d(U_1,U_2)$, showing that $\widetilde{\mathbf{S}}$ is a contraction mapping, as we claimed. Thus, by the Banach fixed point theorem, we see that $\widetilde{\mathbf{S}}$ has a unique fixed point in $\widetilde{\mathbf{M}}$, which is clearly a solution of the problem (7.2) in $C([0,\infty),X_0)$. The uniqueness of the solution follows from a standard argument.

From the above argument we see that the solution U of (7.2) satisfies the first two inequalities in (7.3), and $U \in C^1([0,\infty),X)$. It remains to prove that U also satisfies the last inequality in (7.3). The argument is as follows. First, it is straightforward to deduce from condition (7.1) that, for sufficiently small $\varepsilon > 0$, we have $w_V(r,t)/r(1-r) \in C[0,1]$, and there exist positive constants C_1 and C_2 independent of V such that

$$-C_1 r(1-r) \le w_V(r,t) \le -C_2 r(1-r) \quad \text{for } 0 \le r \le 1 \text{ and } t \ge 0. \tag{7.6}$$

It follows that, for the solution $U = (q,s)$ of (7.2), we have

$$\sup_{0\le r\le 1}\left|w_V(r,t)\frac{\partial q(r,t)}{\partial r}\right| \le C \sup_{0\le r\le 1}\left|r(1-r)\frac{\partial q(r,t)}{\partial r}\right| \le C\|q(\cdot,t)\|_{C^1_V[0,1]}.$$

Using this result and (7.2), we see that

$$\|U'(t)\|_X \le \|\mathbb{A}(V(t))U(t)\|_X + \|\mathbb{G}(U(t)\|_X \le C\|U(t)\|_{X_0} + C\|U(t)\|_X^2 \le C\varepsilon e^{-\mu t}$$

for all $t \ge 0$. This completes the proof of Lemma 7.1. ∎

Lemma 7.1, in particular, implies that, for every V in $\mathbf{M}$, the solution U of (7.2) also belongs to $\mathbf{M}$. Thus we can define a mapping $\mathbf{S} : \mathbf{M} \to \mathbf{M}$ as follows: For any $V \in \mathbf{M}$,

$$\mathbf{S}(V) = U = \text{the solution of (7.2)}.$$

LEMMA 7.2. *For ε sufficiently small, $\mathbf{S}$ is a contraction mapping.*

Proof. Let $V_1, V_2 \in \mathbf{M}$ and denote $U_1 = \mathbf{S}(V_1)$, $U_2 = \mathbf{S}(V_2)$, and $W = U_1 - U_2$. Then W satisfies

$$\frac{dW(t)}{dt} = \mathbb{A}(V_1(t))W(t) + [\mathbb{A}(V_1(t)) - \mathbb{A}(V_2(t))]U_2(t) + [\mathbb{G}(U_1(t)) - \mathbb{G}(U_2(t))]$$

for $t > 0$, and $W(0) = 0$. Thus

$$\begin{aligned} W(t) &= \int_0^t \mathbb{U}(t,s,V_1)[\mathbb{A}(V_1(s)) - \mathbb{A}(V_2(s))]U_2(s)ds \\ &\quad + \int_0^t \mathbb{U}(t,s,V_1)[\mathbb{G}(U_1(s)) - \mathbb{G}(U_2(s))]ds. \end{aligned} \tag{7.7}$$

Since the first component of $[\mathbb{A}(V_1(s)) - \mathbb{A}(V_2(s))]U_2(s)$ is $[w_{V_2}(r,s) - w_{V_1}(r,s)]q_2'(r,s)$ and the second component is zero, we have

$$\begin{aligned} \|[\mathbb{A}(V_1(s)) - \mathbb{A}(V_2(s))]U_2(s)\|_X &= \max_{0\le r\le 1}|[w_{V_1}(r,s) - w_{V_2}(r,s)]q_2'(r,s)| \\ &\le \sup_{0\le r\le 1}\left|\frac{w_{V_1}(r,s) - w_{V_2}(r,s)}{r(1-r)}\right| \max_{0\le r\le 1}|r(1-r)q_2'(r,s)| \\ &\le C\|V_1(s) - V_2(s)\|_X \|U_2(s)\|_{X_0}. \end{aligned} \tag{7.8}$$

Besides, from (2.18) we have

$$\|\mathbb{G}(U_1(s)) - \mathbb{G}(U_2(s))\|_X = \left\| \int_0^1 \mathbb{G}'(\theta U_1(s) + (1-\theta)U_2(s))[U_1(s) - U_2(s)]d\theta \right\|_X$$
$$\leq C\left(\|U_1(s)\|_X + \|U_2(s)\|_X\right)\|U_1(s) - U_2(s)\|_X. \tag{7.9}$$

From (7.7)–(7.9) and Lemma 6.4 we get

$$\begin{aligned}\|U_1(t) - U_2(t)\|_X &\leq C\int_0^t e^{-\mu(t-s)}\|V_1(s) - V_2(s)\|_X\|U_2(s)\|_{X_0}ds \\ &\quad + C\int_0^t e^{-\mu(t-s)}\left(\|U_1(s)\|_X + \|U_2(s)\|_X\right)\|U_1(s) - U_2(s)\|_X ds \\ &\leq C\varepsilon e^{-\mu t}d(V_1, V_2) + C\varepsilon e^{-\mu t}d(U_1, U_2),\end{aligned}$$

which yields $d(U_1, U_2) \leq C\varepsilon d(V_1, V_2) + C\varepsilon d(U_1, U_2)$. From this inequality the desired assertion easily follows. □

By Lemma 7.2, if ε is sufficiently small, then the mapping $\mathbf{S}$ has a unique fixed point U in $\mathbf{M}$. Clearly, U is a global solution of (2.16) subject to the initial condition $U(0) = U_0$. Moreover, by Lemma 7.1, we know that the image of $\mathbf{S}$ is contained in $\widetilde{\mathbf{M}}$ so that U satisfies (7.3). From this result and Lemma 2.1 we see that the assertion of Theorem 1.1 follows. This completes the proof of Theorem 1.1.

Acknowledgments. Part of this work was prepared when the author was visiting the Ecole Normale Superieur (ENS) during June 1–August 28, 2007. He wishes to acknowledge his gratefulness to all staff and faculty of the Department of Mathematics and Applications of ENS, particularly Prof. Perthame and Prof. Rosso, for their hospitality, and the French Ministry of Foreign Affairs, particularly the Section of Science and Technology of the French Consulate at Guangzhou, for financial support for his visit. He is also glad to express thanks to the anonymous referees for helpful suggestions.

REFERENCES

[1] R. P. Araujo and D. L. McElwain, *A history of the study of solid tumor growth: The contribution of mathematical modeling*, Bull. Math. Biol., 66 (2004), pp. 1039–1091.

[2] B. Bazaliy and A. Friedman, *A free boundary problem for an elliptic-parabolic system: Application to a model of tumor growth*, Comm. Partial Differential Equations, 28 (2003), pp. 517–560.

[3] B. Bazaliy and A. Friedman, *Global existence and asymptotic stability for an elliptic-parabolic free boundary problem: An application to a model of tumor growth*, Indiana Univ. Math. J., 52 (2003), pp. 1265–1304.

[4] X. Chen and A. Friedman, *A free boundary problem for an elliptic-hyperbolic system: An application to tumor growth*, SIAM J. Math. Anal., 35 (2003), pp. 974–986.

[5] X. Chen, S. Cui, and A. Friedman, *A hyperbolic free boundary problem modeling tumor growth: Asymptotic behavior*, Trans. Amer. Math. Soc., 357 (2005), pp. 4771–4804.

[6] S. Cui, *Analysis of a mathematical model for the growth of tumors under the action of external inhibitors*, J. Math. Biol., 44 (2002), pp. 395–426.

[7] S. Cui, *Analysis of a free boundary problem modelling tumor growth*, Acta Math. Appl. Sin. Engl. Ser., 21 (2005), pp. 1071–1082.

[8] S. Cui, *Existence of a stationary solution for the modified Ward-King tumor growth model*, Adv. Appl. Math., 36 (2006), pp. 421–445.

[9] S. Cui, *Well-posedness of a multidimensional free boundary problem modeling the growth of nonnecrotic tumors*, J. Func. Anal., 245 (2007), pp. 1–18.

[10] S. Cui and J. Escher, *Bifurcation analysis of an elliptic free boundary problem modelling the growth of avascular tumors*, SIAM J. Math. Anal., 39 (2007), pp. 210–235.

[11] S. Cui and J. Escher, *Asymptotic behavior of solutions of a multidimensional moving boundary problem modeling tumor growth*, Comm. Partial Differential Equations, 33 (2008), pp. 636–655.
[12] S. Cui and A. Friedman, *A hyperbolic free boundary problem modeling tumor growth*, Interfaces Free Bound., 5 (2003), pp. 159–181.
[13] S. Cui and A. Friedman, *A free boundary problem for a singular system of differential equations: An application to a model of tumor growth*, Trans. Amer. Math. Soc., 355 (2003), pp. 3537–3590.
[14] S. Cui and X. Wei, *Existence of solutions for a parabolic-hyperbolic free boundary problem*, Acta Math. Appl. Sin. Engl. Ser., 21 (2005), pp. 597–614.
[15] A. Friedman, *A hierarchy of cancer models and their mathematical challenges*, Disc. Cont. Dyna. Syst. B, 4 (2004), pp. 147–159.
[16] A. Friedman, *Mathematical analysis and challenges arising from models of tumor growth*, Math. Models Methods Appl. Sci., 17 (2007), supplement, pp. 1751–1772.
[17] A. Friedman and B. Hu, *Asymptotic stability for a free boundary problem arising in a tumor model*, J. Differential Equations, 227 (2006), pp. 598–639.
[18] A. Friedman and F. Reitich, *Analysis of a mathematical model for the growth of tumors*, J. Math. Biol., 38 (1999), pp. 262–284.
[19] A. Friedman and F. Reitich, *Symmetry-breaking bifurcation of analytic solutions to free boundary problems*, Trans. Amer. Math. Soc., 353 (2000), pp. 1587–1634.
[20] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, New York, 1983.
[21] G. Pettet, C. Please, and M. McElwain, *The migration of cells in multicell tumor spheroids*, Bull. Math. Biol., 63 (2001), pp. 231–257.
[22] T. Roose, S. J. Chapman, and P. K. Maini, *Mathematical models of avascular tumor growth*, SIAM Rev., 49 (2007), pp. 179–208.
[23] J. Wu and S. Cui, *Asymptotic behavior of solutions of a free boundary problem modeling the growth of tumors in the presence of inhibitors*, Nonlinearity, 20 (2007), pp. 2389–2408.
[24] F. Zhou and S. Cui, *Well-posedness and stability of a multidimensional moving boundary problem modeling the growth of tumor cord*, Discrete Contin. Dyn. Syst. Ser. A, 21 (2008), pp. 929–943.

SIAM J. MATH. ANAL.
Vol. 40, No. 5, pp. 1725–1749

WELL-POSEDNESS AND CONVERGENCE TO THE STEADY STATE FOR A MODEL OF MORPHOGEN TRANSPORT*

PIOTR KRZYŻANOWSKI†, PHILIPPE LAURENÇOT‡, AND DARIUSZ WRZOSEK†

Abstract. Well-posedness and large time convergence to the unique steady state are shown for a model which describes the spreading of morphogens by a nonlinear transport mechanism (transcytosis) and couples a quasilinear parabolic partial differential equation with an ordinary differential equation. A simpler model which assumes linear transport is also investigated for comparison. The analysis of both models requires the construction of specific Liapunov functionals. The study is supplemented by numerical simulations of the sensitivity of the models to the variation of their parameters.

Key words. degenerate parabolic system, Liapunov functional, steady state

AMS subject classifications. 35K55, 35B40, 37L45, 35Q80

DOI. 10.1137/070711608

1. Introduction.

Morphogens are signaling molecules that play an important role in the process of cell differentiation, as different morphogen concentrations induce distinct cell fates. In fact, during cell development, spatial gradients of morphogen concentration form while the morphogens move away from a spatially localized source. It is, however, yet unclear whether the spreading of morphogens occurs by passive diffusion in the extracellular medium or by more complex mechanisms such as planar transcytosis (see, e.g., [5, 11] and the references therein). Planar transcytosis explains the transport of morphogens across the tissue as the consequence of repeated cycles of internalization of morphogens inside the cell (endocytosis) and release of morphogens outside the cell (exocytosis) together with the binding and unbinding of morphogens to receptors located at the surface of the cell. A related model accounting only for the latter mechanism was introduced in [6].

A model for the transport of morphogens by passive diffusion was recently derived in [8, Model B] and describes the space and time evolution of the concentrations of free morphogens ℓ and bound receptors s: in dimensionless form, it reads

$$\partial_t \ell = D\, \partial_x^2 \ell + \delta\, s - \ell\,(1-s), \quad (t,x) \in (0,\infty)\times(0,1), \tag{1.1}$$

$$\partial_t s = \ell\,(1-s) - (\delta+\varepsilon)\, s, \quad (t,x) \in (0,\infty)\times(0,1), \tag{1.2}$$

supplemented by suitable boundary and initial conditions. In this model, the total number of receptors (free + bound) per cell is assumed to be constant and normalized to one. The reaction terms account for the depletion of free morphogens resulting from the binding with free receptors ($-\ell\,(1-s)$), the release of free morphogens after unbinding from the receptors ($\delta\, s$), and the degradation of bound receptors ($-\varepsilon\, s$), respectively. An attempt to take into account planar transcytosis is made in

*Received by the editors December 21, 2007; accepted for publication (in revised form) July 11, 2008; published electronically December 17, 2008. This work was partially supported by Polish MEiSW grant 1P03A01730 and by "Projet Hubert Curien" POLONIUM 11605VC (2006/2007).
http://www.siam.org/journals/sima/40-5/71160.html

†Institute of Applied Mathematics, Warsaw University, Banacha 2, 02-097 Warszawa, Poland (p.krzyzanowski@mimuw.edu.pl, d.wrzosek@mimuw.edu.pl).

‡Institut de Mathématiques de Toulouse, CNRS (UMR 5219), and Université de Toulouse, F-31062 Toulouse Cedex 9, France (laurenco@math.univ-toulouse.fr).

[3, 7]: it gives a system of two equations similar to (1.1)–(1.2), with the noticeable difference that it includes a nonconstant diffusion coefficient in the equation for the free morphogens. It describes the fact that free receptors have to be available for the transport of morphogens to take place. Depending on the assumptions made during the derivation, the equation for the free morphogens might also contain an additional drift term of the form $\partial_x (\chi(\ell, s)\ \partial_x s)$. Let us mention here that such a drift term also shows up in the related model proposed in [9]. We will, however, focus in this paper on a model which does not include an additional drift term of the above-mentioned form but accounts for transcytosis by a nonconstant diffusion coefficient which depends on the concentration of bound receptors and vanishes if no free receptor is available ($s = 1$).

More precisely, the model derived in [7] reads, in dimensionless form,

$$\partial_t \ell = D\ \partial_x \left((1-s)\ \partial_x \ell\right) + \delta\ s - \ell\ (1-s), \quad (t,x) \in (0,\infty) \times (0,1), \tag{1.3}$$

$$\partial_t s = \ell\ (1-s) - (\delta + \varepsilon)\ s, \quad (t,x) \in (0,\infty) \times (0,1), \tag{1.4}$$

$$D\ (1 - s(t,0))\ \partial_x \ell(t,0) + \nu = \ell(t,1) = 0, \quad t \in (0,\infty), \tag{1.5}$$

$$(\ell, s)(0,x) = (\ell_0, s_0)(x), \quad x \in (0,1), \tag{1.6}$$

with $D > 0$, $\delta > 0$, $\varepsilon \ge 0$, and $\nu > 0$. The strength of the diffusion is clearly related to the concentration of bound receptors s and no diffusion takes place in the absence of free receptors. As for the reaction terms, they are the same as for (1.1)–(1.2) though their biological meaning is different: Indeed, the term $(-\ell\ (1-s))$ accounts for the endocytosis process, which combines the binding of a free morphogen with a receptor followed by its internalization, and the term (δs) for the exocytosis process, which includes an externalization step and a dissociation step. Finally, the boundary condition at $x = 0$ accounts for a source of morphogens localized on one side of the tissue. We refer to [7] for a complete derivation of the model, and some additional information as well.

The aim of this paper is to supplement the qualitative study of (1.3)–(1.6) performed in [7] with that of mathematical properties, namely, well-posedness and large time behavior. We first point out that, as (1.3)–(1.6) couples a quasilinear parabolic partial differential equation with an ordinary differential equation, it cannot be studied with the usual techniques for systems of parabolic partial differential equations. Nevertheless, a general theory has been developed in [1] to study this kind of system: applying this theory in a suitable functional framework allows us to establish the local well-posedness of (1.3)–(1.6). A Liapunov functional is then constructed from which we deduce the global well-posedness.

More precisely, we assume that

$$(\ell_0, s_0) \in V \times V \quad \text{with} \quad \ell_0 \ge 0 \quad \text{and} \quad 0 \le s_0 < 1, \tag{1.7}$$

where $V := \{v \in H^1(0,1) \ \text{ such that } \ v(1) = 0\}$ and V' denotes the topological dual of V.

THEOREM 1. *Given initial data (ℓ_0, s_0) satisfying (1.7) the initial-boundary value problem (1.3)–(1.6) possesses a unique weak solution*

$$(\ell, s) \in \mathcal{C}([0,\infty); V \times V) \cap \mathcal{C}^1([0,\infty); V' \times V)$$

such that $\partial_t \ell \in L^2((0,T) \times (0,1))$ for every $T > 0$,

$$0 \le \ell(t,x) \quad \textit{and} \quad 0 \le s(t,x) < 1 \quad \textit{for} \quad (t,x) \in [0,\infty) \times [0,1],$$

and

$$\langle \partial_t \ell(t), \psi \rangle_{V',V} + D \int_0^1 (1 - s(t))\ \partial_x \ell(t)\ \partial_x \psi\ dx = \nu\ \psi(0) + \int_0^1 (\delta s(t) - \ell(t)(1 - s(t)))\ \psi\ dx,$$

$$\partial_t s(t) = -(\delta + \varepsilon) s(t) + \ell(t)(1 - s(t))$$

for $\psi \in V$ *and* $t \in [0, \infty)$. *Furthermore, for each* $t > 0$,

$$\mathcal{L}(s(t), \partial_t \sigma(t)) + \int_0^t \mathcal{D}(s(\tau), \partial_t \sigma(\tau))\ d\tau = \mathcal{L}(s_0, \partial_t \sigma(0)), \tag{1.8}$$

where $\sigma := -\ln(1 - s)$ *and*

$$\mathcal{L}(v, w) := \mathcal{L}_0(v) + \frac{1}{2} \int_0^1 w(x)^2\ dx, \tag{1.9}$$

$$\mathcal{L}_0(v) := \int_0^1 \left\{ \frac{D(\delta + \varepsilon)}{2}\ |\partial_x \Sigma_I(v)(x)|^2 + \varepsilon\ (\Sigma_I(v) - v)(x) - \nu\ \Sigma_I(v)(0) \right\} dx, \tag{1.10}$$

$$\mathcal{D}(v, w) := \int_0^1 \left(1 - v(x) + \frac{\delta + \varepsilon}{1 - v(x)}\right)\ w(x)^2\ dx + D \int_0^1 (1 - v(x))\ |\partial_x w(x)|^2\ dx, \tag{1.11}$$

the function Σ_I *being defined by* $\Sigma_I(r) := -\ln(1 - r)$ *for* $r \in [0, 1)$.

We now turn to the behavior for large times of weak solutions to (1.3)–(1.6) and first recall from [7] that (1.3)–(1.5) has a single stationary solution.

PROPOSITION 2. *There is a unique stationary solution* $(\ell_\infty, s_\infty) \in \mathcal{C}([0, 1]; \mathbb{R}^2)$ *to* (1.3)–(1.5) *satisfying the natural constraints*

$$\ell_\infty(x) \geq 0 \quad \textit{and} \quad 0 \leq s_\infty(x) < 1 \quad \textit{for} \quad x \in [0, 1]. \tag{1.12}$$

The function s_∞ *is the unique solution to the boundary value problem*

$$\begin{aligned} -D\ \partial_x^2 \Sigma_I(s_\infty) + \frac{\varepsilon}{\delta + \varepsilon}\ s_\infty = 0 \quad \textit{in} \quad (0, 1), \\ D\ \partial_x \Sigma_I(s_\infty)(0) + \frac{\nu}{\delta + \varepsilon} = s_\infty(1) = 0, \end{aligned} \tag{1.13}$$

while ℓ_∞ *is given by* $\ell_\infty(x) = (\delta + \varepsilon) s_\infty(x)/(1 - s_\infty(x))$ *for* $x \in [0, 1]$.

With this notation, our second result reads as follows.

THEOREM 3. *Consider* (ℓ_0, s_0) *satisfying* (1.7) *and let* (ℓ, s) *be the corresponding weak solution to the initial-boundary value problem* (1.3)–(1.6). *Then* $(\ell(t), s(t))$ *converges (strongly) towards* (ℓ_∞, s_∞) *in* $L^2(0, 1) \times H^1(0, 1)$ *as* $t \to \infty$.

Roughly speaking, the only piece of information on the convergence that can be retrieved from (1.8) is that $\partial_t \Sigma_I(s)$ vanishes as $t \to \infty$ and so does $\partial_t s$. Therefore, by (1.4), $\ell(1 - s) - (\delta + \varepsilon) s$ vanishes for large times, which somehow means that s gets closer and closer to $\ell/(\ell + \delta + \varepsilon)$ as time goes by. However, no information is available

on ℓ at this stage. Nevertheless, it seems reasonable to guess that ℓ behaves as the solution λ to

$$\partial_t \lambda = \partial_x \left(\frac{\delta+\varepsilon}{\lambda+\delta+\varepsilon}\, \partial_x \lambda \right) - \frac{\varepsilon}{\delta+\varepsilon}\, \frac{\lambda}{\lambda+\delta+\varepsilon}, \quad (t,x) \in (0,\infty)\times(0,1), \tag{1.14}$$

$$D\, \frac{\partial_x \lambda(t,0)}{\lambda(t,0)+\delta+\varepsilon} + \frac{\nu}{\delta+\varepsilon} = \lambda(t,1) = 0, \quad t \in (0,\infty), \tag{1.15}$$

obtained from (1.3)–(1.5) by replacing s by $\ell/(\ell+\delta+\varepsilon)$. Noting that ℓ_∞ is also a stationary solution to (1.14)–(1.15) allows us to construct another Liapunov functional for (1.3)–(1.6) resembling that available for (1.14)–(1.15) and providing the needed information on ℓ.

The large time convergence towards the steady state is proved in section 3 after establishing the well-posedness in $V \times V$ in section 2. In section 4, we return to the model (1.1)–(1.2) of morphogen transport by passive diffusion and show that it is also well-posed in $V \times V$ once supplemented by the boundary and initial conditions (1.5)–(1.6). To this end we construct a Liapunov function for this problem which turns out to provide much more information than $\mathcal{L}$. In particular, exponential convergence to the steady state can be proved for (1.1)–(1.2). Finally, we devote section 5 to numerical simulations, where we show that the solutions of both models converge to the steady state at an exponential rate and that this rate cannot be arbitrarily large regardless of the choice of the parameters.

Throughout the paper, C_i, $i \ge 1$, denotes any positive constant depending only on ℓ_0, s_0, D, δ, ε, and ν. For a real number $r \in \mathbb{R}$, $r_+ := \max\{r,0\}$ denotes the positive part of r and we set $\mathrm{sign}_+(r) := 1$ for $r > 0$ and $\mathrm{sign}_+(r) := 0$ for $r \le 0$.

2. Well-posedness. We first show that the Liapunov functional $\mathcal{L}$ is bounded from below.

LEMMA 4. *For $v \in V$, $0 \le v < 1$, we have*

$$\mathcal{L}_0(v) \ge \mathcal{L}_0(s_\infty) + \frac{D(\delta+\varepsilon)}{2} \int_0^1 \left| \partial_x \left(\Sigma_I(v) - \Sigma_I(s_\infty) \right)(x) \right|^2 dx. \tag{2.1}$$

In particular,

$$\mathcal{L}_0(s_\infty) = \min_{v \in V} \{ \mathcal{L}_0(v) \}.$$

Proof. Let $v \in V$, $0 \le v < 1$. It follows from (1.13) by a straightforward computation that

$$\begin{aligned} \mathcal{L}_0(v) - \mathcal{L}_0(s_\infty) = {} & \frac{D(\delta+\varepsilon)}{2} \int_0^1 \left| \partial_x \left(\Sigma_I(v) - \Sigma_I(s_\infty) \right)(x) \right|^2 dx \\ & + \varepsilon \int_0^1 \left\{ (1-s_\infty)\, \left(\Sigma_I(v) - \Sigma_I(s_\infty) \right) - v + s_\infty \right\}(x)\, dx. \end{aligned}$$

Since $s_\infty < 1$ by Proposition 2 we observe that

$$(1-s_\infty)\left(\Sigma_I(v)-\Sigma_I(s_\infty)\right) - v + s_\infty = (1-s_\infty)\left[\Sigma_I(v)-\Sigma_I(s_\infty)-\Sigma_I'(s_\infty)\,(v-s_\infty)\right] \ge 0$$

by the convexity of Σ_I, and (2.1) follows. ∎

Proof of Theorem 1. By [1, Theorem 6.4] (with $G = \mathbb{R} \times (-\infty, 1)$) the initial-boundary value problem (1.3)–(1.6) possesses a unique maximal weak solution

$$(\ell, s) \in \mathcal{C}([0,T_m); V \times H^1(0,1)) \cap \mathcal{C}^1([0,T_m); V' \times H^1(0,1)) \tag{2.2}$$

such that

$$s(t,x) < 1, \quad (t,x) \in [0,T_m) \times [0,1], \tag{2.3}$$

and

$$\begin{aligned} &\langle \partial_t \ell(t), \psi \rangle_{V',V} + D \int_0^1 (1-s(t))\ \partial_x \ell(t)\ \partial_x \psi\ dx \\ &\qquad = \nu\ \psi(0) + \int_0^1 (\delta s(t) - \ell(t)(1-s(t)))\ \psi\ dx, \end{aligned} \tag{2.4}$$

$$\partial_t s(t) = -(\delta+\varepsilon) s(t) + \ell(t)(1-s(t)) \tag{2.5}$$

for $\psi \in V$ and $t \in [0,T_m)$. In addition, $T_m = \infty$ if, for each $T > 0$,

$$(\ell, s) \in \mathcal{BUC}([0,T] \cap [0,T_m); V \times H^1(0,1)) \tag{2.6}$$

and there is $\vartheta_T \in (0,1)$ such that

$$\max_{x\in[0,1]} \{s(t,x)\} \le \vartheta_T \quad \text{for every} \quad t \in [0,T] \cap [0,T_m). \tag{2.7}$$

We first establish the nonnegativity of (ℓ, s). For that purpose we notice that since $1 - s \ge 0$, the right-hand side of (1.3)–(1.4) satisfies

$$\begin{aligned} &[\delta s - \ell(1-s)]\ \operatorname{sign}_+(-\ell) + [\ell(1-s) - (\delta+\varepsilon)s]\ \operatorname{sign}_+(-s) \\ &\ge -\delta(-s)_+ + (1-s)(-\ell)_+ - (1-s)(-\ell)_+ + (\delta+\varepsilon)(-s)_+ \\ &\ge 0. \end{aligned}$$

We may then use classical approximations of $r \mapsto r_+$ to deduce from (2.4) and (2.5) that

$$\int_0^1 [(-\ell)_+(t) + (-s)_+(t)]\ dx \le \int_0^1 [(-\ell_0)_+ + (-s_0)_+]\ dx = 0$$

for $t \in [0,T_m)$. Consequently

$$\ell(t,x) \ge 0 \quad \text{and} \quad s(t,x) \ge 0, \quad (t,x) \in [0,T_m) \times [0,1]. \tag{2.8}$$

It also follows from (1.7), (2.2), and (2.5) that $\partial_t s(t,1) = -(\delta+\varepsilon) s(t,1)$ with $s(0,1) = 0$, whence

$$s(t,1) = 0 \quad \text{and} \quad s(t) \in V \quad \text{for} \quad t \in [0,T_m). \tag{2.9}$$

We next turn to the identity (1.8): owing to (2.2), (2.3), and (2.9), the function $\sigma := \Sigma_I(s) = -\ln(1-s)$ is well-defined in $[0,T_m) \times [0,1]$ and belongs to $\mathcal{C}^1([0,T_m); V)$ with

$$\partial_t \sigma = \frac{\partial_t s}{1-s} \quad \text{and} \quad \partial_x \sigma = \frac{\partial_x s}{1-s}.$$

Furthermore (2.5) also reads

$$\partial_t \sigma = \ell - (\delta+\varepsilon)\ \frac{s}{1-s}, \tag{2.10}$$

which, together with (2.2), ensures that $\partial_t\sigma$ belongs to $\mathcal{C}([0,T_m);V)$ and $\mathcal{C}^1([0,T_m);V')$. For $t\in[0,T_m)$, $\partial_t\sigma(t)$ is then an admissible test function in (2.4) and it follows from (2.4) and (2.5) that

$$\begin{aligned}&\langle\partial_t\ell(t),\partial_t\sigma(t)\rangle_{V',V}+D\int_0^1(1-s(t))\,\partial_x\ell(t)\,\partial_x\partial_t\sigma(t)\,dx\\&=-\int_0^1\left(\partial_t s(t)+\varepsilon\,s(t)\right)\,\partial_t\sigma(t)\,dx+\nu\,\partial_t\sigma(t,0).\end{aligned}$$

On the one hand, expressing ℓ in terms of σ and s with the help of (2.10) yields

$$\langle\partial_t\ell(t),\partial_t\sigma(t)\rangle_{V',V}=\frac{1}{2}\,\frac{d}{dt}\int_0^1|\partial_t\sigma(t)|^2\,dx+(\delta+\varepsilon)\int_0^1\frac{|\partial_t\sigma(t)|^2}{1-s(t)}\,dx.$$

On the other hand, using once more (2.10), we obtain

$$\begin{aligned}D\int_0^1(1-s(t))\,\partial_x\ell(t)\,\partial_x\partial_t\sigma(t)\,dx&=D\int_0^1(1-s(t))\,|\partial_x\partial_t\sigma(t)|^2\,dx\\&\quad+\frac{D(\delta+\varepsilon)}{2}\,\frac{d}{dt}\int_0^1|\partial_x\sigma(t)|^2\,dx.\end{aligned}$$

Combining the above three identities we end up with

$$\frac{d}{dt}\mathcal{L}(s(t),\partial_t\sigma(t))+\mathcal{D}(s(t),\partial_t\sigma(t))=0,\quad t\in[0,T_m),$$

whence

$$\mathcal{L}(s(t),\partial_t\sigma(t))+\int_0^t\mathcal{D}(s(\tau),\partial_t\sigma(\tau))\,d\tau=\mathcal{L}(s_0,\partial_t\sigma(0)),\quad t\in[0,T_m).\tag{2.11}$$

We then deduce from (1.9), (2.11), (2.1), and the nonnegativity of $\mathcal{D}(s,\partial_t\sigma)$ that

$$\frac{D(\delta+\varepsilon)}{2}\int_0^1\left|\partial_x\left(\sigma(t)-\Sigma_I(s_\infty)\right)\right|^2\,dx+\frac{1}{2}\int_0^1|\partial_t\sigma(t)|^2\,dx\le C_0,\tag{2.12}$$

with $C_0:=-\mathcal{L}_0(s_\infty)+\mathcal{L}(s_0,\partial_t\sigma(0))$. Since $\sigma(t)\in V$ for $t\in[0,T_m)$ it follows from (2.12) and the Poincaré inequality

$$\|v\|_2\le\|v\|_\infty\le\|\partial_x v\|_2,\quad v\in V,\tag{2.13}$$

that there is a constant $C_1>0$ such that

$$\|\sigma(t)\|_{H^1}\le C_1,\quad t\in[0,T_m).\tag{2.14}$$

In particular, $\|\sigma(t)\|_\infty\le C_1$, from which we readily deduce that

$$s(t,x)\le 1-e^{-C_1},\quad (t,x)\in[0,T_m)\times[0,1].\tag{2.15}$$

Since $\partial_x s=e^{-\sigma}\,\partial_x\sigma$ and $\sigma\ge 0$ we further obtain from (2.14) and (2.15) that

$$\|s(t)\|_{H^1}\le 1+C_1,\quad t\in[0,T_m).\tag{2.16}$$

We next establish a similar estimate for ℓ and first proceed in a formal way. A rigorous proof based on an approximation argument will be sketched at the end of the proof of Theorem 1. We take $\psi = \partial_t \ell$ in (2.4) and use (2.3), (2.5), (2.12), (2.15), and the nonnegativity of ℓ to obtain

$$\begin{aligned}
&\int_0^1 |\partial_t \ell|^2 \, dx + \frac{D}{2} \, \frac{d}{dt} \int_0^1 (1-s) \, |\partial_x \ell|^2 \, dx - \nu \, \partial_t \ell(0) \\
&= -\frac{D}{2} \int_0^1 \partial_t s \, |\partial_x \ell|^2 \, dx - \int_0^1 (\partial_t s + \varepsilon \, s) \, \partial_t \ell \, dx \\
&\le \frac{D}{2} \, (\delta + \varepsilon) \int_0^1 s \, |\partial_x \ell|^2 \, dx + \frac{1}{2} \, \|\partial_t \ell\|_2^2 + \frac{1}{2} \, \|(1-s) \, \partial_t \sigma + \varepsilon \, s\|_2^2 \\
&\le \frac{D(\delta+\varepsilon) e^{C_1}}{2} \int_0^1 (1-s) \, |\partial_x \ell|^2 \, dx + \frac{1}{2} \, \|\partial_t \ell\|_2^2 + \|\partial_t \sigma\|_2^2 + \varepsilon \\
&\le \frac{D(\delta+\varepsilon) e^{C_1}}{2} \int_0^1 (1-s) \, |\partial_x \ell|^2 \, dx + \frac{1}{2} \, \|\partial_t \ell\|_2^2 + 2C_0 + \varepsilon.
\end{aligned}$$

Consequently, there is a constant $C_2 > 0$ such that

$$\int_0^1 |\partial_t \ell|^2 \, dx + \frac{d}{dt} \left\{ \int_0^1 D \, (1-s) \, |\partial_x \ell|^2 \, dx - \nu \, \ell(0) \right\} \le C_2 \, \left(1 + \int_0^1 (1-s) \, |\partial_x \ell|^2 \, dx \right),$$

whence

$$\begin{aligned}
&D \int_0^1 (1 - s(t)) \, |\partial_x \ell(t)|^2 \, dx - \nu \, \ell(t,0) + \int_0^t \int_0^1 |\partial_t \ell(\tau)|^2 \, dx d\tau \\
&\qquad \le (1 + D \, \|\ell_0\|_{H^1}^2) \, e^{C_2 t}
\end{aligned} \tag{2.17}$$

for $t \in [0, T_m)$. Recalling (2.15) we readily deduce from (2.17) and the Poincaré inequality (2.13) that

$$\|\ell(t)\|_{H^1}^2 + \int_0^t \|\partial_t \ell(\tau)\|_2^2 \, d\tau \le C_3 \, e^{C_3 t} \tag{2.18}$$

for $t \in [0, T_m)$. According to (2.16) and (2.18), we have shown so far that, given $T > 0$, there is a positive constant M_T depending only on ℓ_0, s_0, D, δ, ε, ν, and T such that

$$\|\ell(t)\|_{H^1} + \|s(t)\|_{H^1} \le M_T \quad \text{for} \quad t \in [0,T] \cap [0, T_m).$$

It remains to improve this estimate to a uniform bound on the modulus of continuity with respect to time for ℓ and s in $H^1(0,1)$ as required by (2.6). To this end, we first notice that (2.5) and (2.18) entail that

$$\|\partial_t s(t)\|_{H^1} \le 2 \, \|\ell(t)\|_{H^1} \, \|(1-s)(t)\|_{H^1} + (\delta + \varepsilon) \, \|s(t)\|_{H^1} \le C_4 \, e^{C_4 t} \quad \text{for } t \in [0, T_m). \tag{2.19}$$

It next follows from (2.4) that $\xi := \partial_t \ell$ formally solves

$$\begin{aligned}
\partial_t \xi &= D \, \partial_x \left((1-s) \, \partial_x \xi - \partial_t s \, \partial_x \ell \right) \\
&\qquad + (\ell + \delta) \, \partial_t s - \xi \, (1-s), \quad (t,x) \in (0, T_m) \times (0,1),
\end{aligned} \tag{2.20}$$

$$(1 - s(t,0)) \, \partial_x \xi(t,0) - \partial_t s(t,0) \, \partial_x \ell(t,0) = \xi(t,1) = 0, \quad t \in (0, T_m). \tag{2.21}$$

We multiply (2.20) by ξ and integrate over $(0,1)$ to obtain, thanks to (2.21),

$$\begin{aligned}\frac{1}{2}\,\frac{d}{dt}\|\xi(t)\|_2^2 &= -D\,\int_0^1 \partial_x\xi\,\left((1-s)\,\partial_x\xi-\partial_t s\,\partial_x\ell\right)\,dx\\ &\quad+\int_0^1(\ell+\delta)\,\partial_t s\,\xi\,dx-\int_0^1(1-s)\,\xi^2\,dx\\ &\le -D\,\int_0^1(1-s)\,|\partial_x\xi|^2\,dx-\int_0^1(1-s)\,\xi^2\,dx\\ &\quad+D\,\|\partial_t s\|_\infty\,\|\partial_x\ell\|_2\,\|\partial_x\xi\|_2+(\|\ell\|_\infty+\delta)\,\|\partial_t s\|_2\,\|\xi\|_2.\end{aligned}$$

Recalling (2.15), (2.18), (2.19), and the continuous embedding of $H^1(0,1)$ in $L^\infty(0,1)$, we deduce from the Young inequality that

$$\begin{aligned}\frac{1}{2}\,\frac{d}{dt}\|\xi(t)\|_2^2 &\le -D\,e^{-C_1}\,\|\partial_x\xi\|_2^2-e^{-C_1}\,\|\xi\|_2^2+D\,e^{-C_1}\,\|\partial_x\xi\|_2^2\\ &\quad+e^{-C_1}\,\|\xi\|_2^2+C_5\,\|\partial_t s\|_{H^1}^2\,\left(\|\ell\|_{H^1}+\delta\right)^2\\ &\le C_6\,e^{C_6 t},\end{aligned}$$

whence

$$\|\xi(t)\|_2^2\le\|\xi(\tau)\|_2^2+2\,e^{C_6 t},\quad 0\le\tau\le t<T_m.$$

Integrating with respect to τ over $(0,t)$ and using (2.18), we end up with

$$\begin{aligned}t\,\|\xi(t)\|_2^2&\le\int_0^t\|\xi(\tau)\|_2^2\,d\tau+2\,t\,e^{C_6 t},\\ \|\xi(t)\|_2^2&\le C_7\,\left(1+\frac{1}{t}\right)\,e^{C_7 t},\quad t\in(0,T_m).\end{aligned}\tag{2.22}$$

We next infer from (1.3) and (2.16) that

$$\partial_x^2\ell=\frac{\partial_t\ell}{D\,(1-s)}+\frac{\ell}{D}-\frac{\delta}{D}\,\frac{s}{1-s}+\partial_x\sigma\,\partial_x\ell.$$

Therefore, by (2.14), (2.15), (2.18), (2.22), and the interpolation inequality

$$\|\partial_x w\|_\infty\le C_8\,\|w\|_{H^2}^{3/4}\,\|w\|_2^{1/4},\quad w\in H^2(0,1),$$

we have

$$\begin{aligned}\|\partial_x^2\ell(t)\|_2&\le\frac{e^{C_1}}{D}\,\|\partial_t\ell(t)\|_2+\frac{\|\ell(t)\|_2}{D}+\frac{\delta\,e^{C_1}}{D}+\|\partial_x\sigma(t)\|_2\,\|\partial_x\ell(t)\|_\infty\\ &\le\frac{\sqrt{C_7}\,e^{C_1}}{D}\,\left(1+\frac{1}{t}\right)^{1/2}\,e^{C_7 t/2}+\frac{\sqrt{C_3}\,e^{C_3 t/2}}{D}\\ &\quad+\frac{\delta\,e^{C_1}}{D}+C_1\,C_8\,\|\ell(t)\|_{H^2}^{3/4}\,\|\ell(t)\|_2^{1/4}\\ &\le C_9\,\left(1+\frac{1}{\sqrt{t}}\right)\,e^{C_9 t}+\frac{1}{2}\,\|\ell(t)\|_{H^2}\\ &\le C_9\,\left(1+\frac{1}{\sqrt{t}}\right)\,e^{C_9 t}+\frac{1}{2}\,\sqrt{C_3}\,e^{C_3 t/2}+\frac{1}{2}\,\|\partial_x^2\ell(t)\|_2,\end{aligned}$$

whence

$$\|\ell(t)\|_{H^2} \le C_{10}\ \left(1+\frac{1}{\sqrt{t}}\right)\ e^{C_{10}t}, \quad t\in(0,T_m). \tag{2.23}$$

Consider next $t_1\in(0,T_m)$ and $t_2\in(t_1,T_m)$. It follows from (2.18), (2.23), and the interpolation inequality

$$\|\partial_x w\|_2 \le C_{11}\ \|w\|_{H^2}^{1/2}\ \|w\|_2^{1/2}, \quad w\in H^2(0,1),$$

that

$$\begin{aligned}
\|\ell(t_2)-\ell(t_1)\|_{H^1}^2 &\le C_{11}^2\ \|\ell(t_2)-\ell(t_1)\|_{H^2}\ \|\ell(t_2)-\ell(t_1)\|_2\\
&\le C_{11}^2\ \left(\|\ell(t_2)\|_{H^2}+\|\ell(t_1)\|_{H^2}\right)\ \int_{t_1}^{t_2}\|\partial_t\ell(\tau)\|_2\ d\tau\\
&\le 2\ C_{10}\ C_{11}^2\ \left(1+\frac{1}{\sqrt{t_1}}\right)\ e^{C_{10}t_2}\ (t_2-t_1)^{1/2}\ \left(\int_{t_1}^{t_2}\|\partial_t\ell(\tau)\|_2^2\ d\tau\right)^{1/2}\\
&\le C_{12}\ \left(1+\frac{1}{\sqrt{t_1}}\right)\ e^{C_{12}t_2}\ (t_2-t_1)^{1/2},
\end{aligned}$$

$$\|\ell(t_2)-\ell(t_1)\|_{H^1} \le C_{13}\ \left(1+t_1^{-1/4}\right)\ e^{C_{13}t_2}\ (t_2-t_1)^{1/4}. \tag{2.24}$$

Since ℓ belongs to $\mathcal{BC}([0,T]\cap[0,T_m);H^1(0,1))$ for any $T>0$ by (2.2) and (2.18), it readily follows from (2.24) that

$$\ell\in\mathcal{BUC}([0,T]\cap[0,T_m);H^1(0,1)) \quad \text{for any} \quad T>0. \tag{2.25}$$

Recalling (2.16) and (2.19), we also have

$$s\in\mathcal{BUC}([0,T]\cap[0,T_m);H^1(0,1)) \quad \text{for any} \quad T>0,$$

which, together with (2.15) and (2.25), allows us to conclude that $T_m=\infty$ by (2.6) and (2.7). The proof of Theorem 1 is then complete, provided we can justify the above computations.

To this end we note that s belongs to $\mathcal{C}^1([0,T_m);H^1(0,1))$ by (2.2). Setting $\alpha_1:=e^{-C_1}/2$, classical approximation arguments ensure that, for each $\alpha\in(0,\alpha_1)$ and $T\in(0,T_m)$, there is a function $s_{\alpha,T}\in\mathcal{C}^\infty([0,T]\times[0,1])$ such that

$$\|s_{\alpha,T}-s\|_{\mathcal{C}^1([0,T];H^1(0,1))}+\|s_{\alpha,T}-s\|_{\mathcal{C}([0,T]\times[0,1])}\le\alpha. \tag{2.26}$$

Owing to (2.15) and (2.26) we have $1-s_{\alpha,T}\ge\alpha_1>0$ and the initial-boundary value problem

$$\begin{aligned}
&\partial_t\ell_{\alpha,T}-D\ \partial_x\left((1-s_{\alpha,T})\ \partial_x\ell_{\alpha,T}\right)=\delta\ s_{\alpha,T}-(1-s_{\alpha,T})\ \ell_{\alpha,T}, \quad (t,x)\in(0,T)\times(0,1),\\
&D\ (1-s_{\alpha,T}(t,0))\ \partial_x\ell_{\alpha,T}(t,0)+\nu=\ell_{\alpha,T}(t,1)=0, \quad t\in(0,T),\\
&\ell_{\alpha,T}(0,x)=\ell_0(x), \quad x\in(0,1),
\end{aligned}$$

has a unique nonnegative classical solution $\ell_{\alpha,T}\in\mathcal{C}([0,T]\times[0,1])\cap\mathcal{C}^\infty((0,T]\times[0,1])$. On the one hand it is rather straightforward to check that

$$\sup_{t\in[0,T]}\left\{\|(\ell_{\alpha,T}-\ell)(t)\|_2^2\right\}\le\alpha\ \int_0^T\left(D^2\ \|\partial_x\ell(t)\|_2^2+\delta^2+\|\ell(t)\|_2^2\right)\ dt. \tag{2.27}$$

On the other hand, the smoothness of $\ell_{\alpha,T}$ allows us to proceed as in the derivation of (2.17) and obtain that

$$\|\partial_x \ell_{\alpha,T}(t)\|_2^2 + \int_0^t \|\partial_t \ell_{\alpha,T}(\tau)\|_2^2 \, d\tau \le C_{14} \, e^{C_{14} t}, \quad t \in [0,T]. \tag{2.28}$$

Owing to (2.27) we may let $\alpha \to 0$ in (2.28) and deduce by weak convergence arguments that $\ell \in L^\infty(0,T;V)$ and $\partial_t \ell \in L^2((0,T)\times(0,1))$ and satisfy an inequality similar to (2.18). Still proceeding as above, we next show that $\ell_{\alpha,T}$ enjoys the properties (2.22) and (2.23), and we use once more weak compactness arguments to provide a rigorous proof of the $\mathcal{BUC}$-estimate (2.25). ∎

3. Convergence to the steady state. As a first step towards the proof of Theorem 3, we derive the following information on s and $\sigma := \Sigma_I(s)$ from (1.8).

LEMMA 5. *There is a real number $\vartheta \in (0,1)$ such that*

$$0 \le s(t,x) \le 1 - \vartheta \quad \textit{for all} \quad (t,x) \in [0,\infty) \times [0,1] \tag{3.1}$$

and

$$\sigma \in L^\infty(0,\infty;H^1(0,1)), \qquad \partial_t \sigma \in L^2(0,\infty;H^1(0,1)). \tag{3.2}$$

Proof. By (1.8) and (2.1) we have

$$\mathcal{L}_0(s_\infty) + \frac{D(\delta+\varepsilon)}{2} \, \|\partial_x \sigma(t) - \partial_x \Sigma_I(s_\infty)\|_2^2 \le \mathcal{L}(s(t), \partial_t \sigma(t)) \le \mathcal{L}(s_0, \partial_t \sigma(0))$$

for all $t \ge 0$, from which we readily conclude that $\partial_x \sigma$ belongs to $L^\infty(0,\infty;L^2(0,1))$. Since $\sigma(t,1) = 0$ for every $t \ge 0$, the Poincaré inequality (2.13) entails that σ belongs to $L^\infty(0,\infty;H^1(0,1))$, which gives the first assertion in (3.2). The space $H^1(0,1)$ being continuously embedded in $L^\infty(0,1)$, we further deduce that σ belongs to $L^\infty((0,\infty)\times(0,1))$, whence (3.1).

Next, since $r + (\delta+\varepsilon)/r \ge 2(\delta+\varepsilon)^{1/2} \ge 2\delta^{1/2}$ for $r \ge 0$ and $1 - s \ge \vartheta > 0$ by (3.1), we have

$$\mathcal{D}(s, \partial_t \sigma) \ge 2\delta^{1/2} \, \|\partial_t \sigma\|_2^2 + \vartheta D \, \|\partial_x \partial_t \sigma\|_2^2,$$

and we infer from (1.8) and (2.1) that

$$\mathcal{L}_0(s_\infty) + \int_0^t \left(2\delta^{1/2} \, \|\partial_t \sigma\|_2^2 + \vartheta D \, \|\partial_x \partial_t \sigma\|_2^2 \right) \, d\tau \le \mathcal{L}(s_0, \partial_t \sigma(0))$$

for all $t \ge 0$, which gives the second assertion in (3.2). ∎

We next turn to ℓ and establish the following inequality.

LEMMA 6. *For $t \ge 0$ we have*

$$\mathcal{L}_1(\ell(t)) + \int_0^t \mathcal{D}_1(\ell, s) \, d\tau \le \mathcal{L}_1(\ell_0) + C_{15} \, \int_0^t \left\| \partial_x \ln\left(\frac{\ell + \delta + \varepsilon}{\ell_\infty + \delta + \varepsilon} \right) \right\|_2 \, \|\partial_t \sigma\|_2 \, d\tau \tag{3.3}$$

with

$$\mathcal{L}_1(u) := \int_0^1 \left[(u + \delta + \varepsilon) \, \left(\ln\left(\frac{u + \delta + \varepsilon}{\ell_\infty + \delta + \varepsilon} \right) - 1 \right) + \ell_\infty + \delta + \varepsilon \right] \, dx \ge 0$$

and

$$\mathcal{D}_1(u,v) := D\ \int_0^1 (1-v)\ (u+\delta+\varepsilon)\ \left|\partial_x \ln\left(\frac{u+\delta+\varepsilon}{\ell_\infty+\delta+\varepsilon}\right)\right|^2\ dx$$
$$+\ \varepsilon(\delta+\varepsilon)\ \int_0^1 \frac{u-\ell_\infty}{(u+\delta+\varepsilon)(\ell_\infty+\delta+\varepsilon)}\ \ln\left(\frac{u+\delta+\varepsilon}{\ell_\infty+\delta+\varepsilon}\right)\ dx \geq 0.$$

Proof. We put $\Lambda := \ln(\ell+\delta+\varepsilon)$ and $\Lambda_\infty := \ln(\ell_\infty+\delta+\varepsilon)$. By (1.4) we have

$$1-s = (\delta+\varepsilon+\partial_t s)\ e^{-\Lambda} \quad \text{and} \quad s = (\ell-\partial_t s)\ e^{-\Lambda}. \tag{3.4}$$

Let $\psi \in V$. Using (1.4) and (3.4) the weak formulation of (1.3) reads

$$\begin{aligned}
&\langle \partial_t \ell, \psi\rangle_{V',V} + D\ \int_0^1 (\delta+\varepsilon+\partial_t s)\ e^{-\Lambda}\ \partial_x \ell\ \partial_x \psi\ dx \\
&\qquad = \nu\ \psi(0) - \int_0^1 (\partial_t s + \varepsilon\ s)\ \psi\ dx \\
&\qquad = \nu\ \psi(0) - \int_0^1 \left(\partial_t s + \varepsilon\ (\ell-\partial_t s)\ e^{-\Lambda}\right)\ \psi\ dx \\
&\qquad = \nu\ \psi(0) - \int_0^1 \left(1-\varepsilon\ e^{-\Lambda}\right)\ \partial_t s\ \psi\ dx - \varepsilon\ \int_0^1 \left(1-(\delta+\varepsilon)\ e^{-\Lambda}\right)\ \psi\ dx.
\end{aligned}$$

Similarly, as (ℓ_∞, s_∞) is also a weak solution to (1.3)–(1.5), we have

$$\begin{aligned}
&D\ \int_0^1 (\delta+\varepsilon)\ e^{-\Lambda_\infty}\ \partial_x \ell_\infty\ \partial_x \psi\ dx \\
&\qquad = \nu\ \psi(0) - \varepsilon\ \int_0^1 \left(1-(\delta+\varepsilon)\ e^{-\Lambda_\infty}\right)\ \psi\ dx.
\end{aligned}$$

Subtracting the above two identities gives

$$\begin{aligned}
&\langle \partial_t \ell, \psi\rangle_{V',V} + D\ \int_0^1 (\delta+\varepsilon+\partial_t s)\ \partial_x(\Lambda-\Lambda_\infty)\ \partial_x \psi\ dx \\
&\qquad + \varepsilon(\delta+\varepsilon)\ \int_0^1 \left(e^{-\Lambda_\infty} - e^{-\Lambda}\right)\ \psi\ dx \\
&\qquad = -D\ \int_0^1 \partial_t s\ \partial_x \Lambda_\infty\ \partial_x \psi\ dx - \int_0^1 \left(1-\varepsilon\ e^{-\Lambda}\right)\ \partial_t s\ \psi\ dx.
\end{aligned} \tag{3.5}$$

Clearly, $\Lambda-\Lambda_\infty$ belongs to V and we may thus take $\psi = \Lambda-\Lambda_\infty$ in (3.5) to obtain

$$\begin{aligned}
&\langle \partial_t \ell, \Lambda-\Lambda_\infty\rangle_{V',V} + D\ \int_0^1 (\delta+\varepsilon+\partial_t s)\ |\partial_x(\Lambda-\Lambda_\infty)|^2\ dx \\
&\qquad + \varepsilon(\delta+\varepsilon)\ \int_0^1 \left(e^{-\Lambda_\infty} - e^{-\Lambda}\right)\ (\Lambda-\Lambda_\infty)\ dx \\
&\qquad = -D\ \int_0^1 \partial_t s\ \partial_x \Lambda_\infty\ \partial_x(\Lambda-\Lambda_\infty)\ dx - \int_0^1 \left(1-\varepsilon\ e^{-\Lambda}\right)\ \partial_t s\ (\Lambda-\Lambda_\infty)\ dx.
\end{aligned}$$

On the one hand, we have

$$\langle \partial_t \ell, \Lambda-\Lambda_\infty\rangle_{V',V} = \frac{d}{dt}\int_0^1 (\ell+\delta+\varepsilon)\ (\ln(\ell+\delta+\varepsilon) - 1 - \Lambda_\infty)\ dx = \frac{d}{dt}\mathcal{L}_1(\ell).$$

On the other hand, $\partial_x\Lambda_\infty$ clearly belongs to $L^\infty(0,1)$ and

$$\begin{aligned}
&-D\int_0^1 \partial_t s\ \partial_x\Lambda_\infty\ \partial_x(\Lambda-\Lambda_\infty)\ dx - \int_0^1 \left(1-\varepsilon\ e^{-\Lambda}\right)\ \partial_t s\ (\Lambda-\Lambda_\infty)\ dx \\
&\le D\ \|\partial_x\Lambda_\infty\|_\infty\ \|\partial_t s\|_2\ \|\partial_x(\Lambda-\Lambda_\infty)\|_2 + \|\partial_t s\|_2\ \|\Lambda-\Lambda_\infty\|_2 \\
&\le C_{16}\ \|\partial_t s\|_2\ \|\partial_x(\Lambda-\Lambda_\infty)\|_2,
\end{aligned}$$

the last inequality being a consequence of the Poincaré inequality (2.13) which can be applied here since $\Lambda(t)-\Lambda_\infty\in V$ for all $t\ge 0$. Therefore,

$$\frac{d}{dt}\mathcal{L}_1(\ell)+\mathcal{D}_1(\ell,s)\le C_{16}\ \|\partial_t s\|_2\ \|\partial_x(\Lambda-\Lambda_\infty)\|_2.$$

Integrating the above inequality with respect to time and using the fact that $|\partial_t s| = |(1-s)\ \partial_t\sigma|\le|\partial_t\sigma|$ complete the proof of (3.3). □

Several bounds on ℓ may be deduced from (3.3) and are listed now.

LEMMA 7. *Introducing the space $W:=H^2(0,1)\cap V$ and denoting its topological dual by W', we have*

$$\ell\ \ln(1+\ell)\in L^\infty(0,\infty;L^1(0,1)), \tag{3.6}$$

$$(\ell+\delta+\varepsilon)^{1/2}\ \partial_x\ln\left(\frac{\ell+\delta+\varepsilon}{\ell_\infty+\delta+\varepsilon}\right)\in L^2((0,\infty)\times(0,1)), \tag{3.7}$$

$$\frac{1}{\ell+\delta+\varepsilon}-\frac{1}{\ell_\infty+\delta+\varepsilon}\in L^2((0,\infty)\times(0,1)), \tag{3.8}$$

$$\partial_t\ell\in L^2(0,\infty;W'). \tag{3.9}$$

Proof. Keeping the notation $\Lambda:=\ln(\ell+\delta+\varepsilon)$ and $\Lambda_\infty:=\ln(\ell_\infty+\delta+\varepsilon)$, we infer from (3.1), the nonnegativity of ℓ, and the inequalities $\left|e^{-\Lambda_\infty}-e^{-\Lambda}\right|\le|\Lambda-\Lambda_\infty|$ and $\ell+\delta+\varepsilon\ge(\ell+2\delta+\varepsilon)/2$ that

$$\mathcal{D}_1(\ell,s)\ge\frac{\vartheta D}{2}\ \int_0^1(\ell+2\delta+\varepsilon)\ |\partial_x(\Lambda-\Lambda_\infty)|_2^2\ dx+\varepsilon(\delta+\varepsilon)\ \left\|e^{-\Lambda_\infty}-e^{-\Lambda}\right\|_2^2.$$

It then follows from (3.3) and the Young inequality that

$$\begin{aligned}
&\mathcal{L}_1(\ell(t))+\frac{\vartheta D}{2}\ \int_0^t\int_0^1(\ell+2\delta+\varepsilon)\ |\partial_x(\Lambda-\Lambda_\infty)|_2^2\ dxd\tau \\
&+\varepsilon(\delta+\varepsilon)\ \int_0^t\left\|e^{-\Lambda_\infty}-e^{-\Lambda}\right\|_2^2\ d\tau \\
&\le\mathcal{L}_1(\ell(t))+\int_0^t\mathcal{D}_1(\ell,s)\ d\tau \\
&\le\mathcal{L}_1(\ell_0)+\frac{\vartheta D\delta}{2}\ \int_0^t\|\partial_x(\Lambda-\Lambda_\infty)\|_2^2\ d\tau+C_{17}\ \int_0^t\|\partial_t\sigma\|_2^2\ d\tau,
\end{aligned}$$

hence

$$\begin{aligned}
&\mathcal{L}_1(\ell(t))+\frac{\vartheta D}{2}\ \int_0^t\int_0^1(\ell+\delta+\varepsilon)\ |\partial_x(\Lambda-\Lambda_\infty)|_2^2\ dxd\tau \\
&+\varepsilon(\delta+\varepsilon)\ \int_0^t\left\|e^{-\Lambda_\infty}-e^{-\Lambda}\right\|_2^2\ d\tau \\
&\le C_{18}\ \left(1+\int_0^t\|\partial_t\sigma\|_2^2\ d\tau\right).
\end{aligned}$$

Recalling that $\partial_t\sigma \in L^2((0,\infty)\times(0,1))$ by Lemma 5, we readily deduce from the previous inequality, the Poincaré inequality (2.13), and the nonnegativity of $\mathcal{L}_1$ that $\mathcal{L}_1(\ell)$ belongs to $L^\infty(0,\infty)$ and that (3.7) and (3.8) hold true. In turn, the L^∞-bound on $\mathcal{L}_1(\ell)$ gives (3.6).

It remains to check (3.9). For that purpose, consider $\psi \in W$. Since $\delta+\varepsilon+\partial_t s = (1-s)(\ell+\delta+\varepsilon)$ by (1.4), it follows from (3.5) and the Hölder inequality that

$$\begin{aligned}
|\langle \partial_t \ell, \psi\rangle_{V',V}| &\le D \int_0^1 (1-s)(\ell+\delta+\varepsilon)\ |\partial_x(\Lambda-\Lambda_\infty)|\ |\partial_x\psi|\ dx \\
&\quad + \varepsilon(\delta+\varepsilon)\ \left\|e^{-\Lambda_\infty}-e^{-\Lambda}\right\|_2\ \|\psi\|_2 + \|\partial_t s\|_2\ \left(D\ \|\partial_x\Lambda_\infty\|_\infty + 1\right)\|\psi\|_{H^1} \\
&\le D\left(\int_0^1 (\ell+\delta+\varepsilon)|\partial_x(\Lambda-\Lambda_\infty)|^2\ dx\right)^{1/2}\left(\int_0^1 (\ell+\delta+\varepsilon)\ dx\right)^{1/2}\|\partial_x\psi\|_\infty \\
&\quad + \varepsilon(\delta+\varepsilon)\ \left\|e^{-\Lambda_\infty}-e^{-\Lambda}\right\|_2\ \|\psi\|_2 + C_{19}\ \|\partial_t\sigma\|_2\ \|\psi\|_{H^1}.
\end{aligned}$$

Thanks to (3.2), (3.6), (3.7), (3.8), and the continuous embedding of W in $W^{1,\infty}(0,1)$, we deduce from the above inequality that

$$\int_0^t |\langle \partial_t\ell,\psi\rangle_{V',V}|^2\ d\tau \le C_{20}\ \|\psi\|_{H^2},$$

which implies the claim (3.9). □

Proof of Theorem 3. Let $(t_n)_{n\ge1}$ be a sequence of positive times such that $t_n\to\infty$ as $n\to\infty$. It follows from (3.1), (3.2), (3.6), and the Dunford–Pettis theorem that $(\ell(t_n))_{n\ge1}$ is weakly compact in $L^1(0,1)$ and $(s(t_n))_{n\ge1}$ is weakly compact in $H^1(0,1)$. There are thus $\ell_*\in L^1(0,1)$, $s_*\in H^1(0,1)$, and a subsequence of $(\ell(t_n),s(t_n))_{n\ge1}$ (not relabeled) such that

$$\ell(t_n) \rightharpoonup \ell_* \quad\text{in}\quad L^1(0,1), \tag{3.10}$$

$$s(t_n) \rightharpoonup s_* \quad\text{in}\quad H^1(0,1). \tag{3.11}$$

Owing to the compactness of the embedding of $H^1(0,1)$ in $\mathcal{C}([0,1])$, we further obtain

$$s(t_n) \longrightarrow s_* \quad\text{in}\quad \mathcal{C}([0,1]). \tag{3.12}$$

We now aim at showing that $(\ell_*,s_*)=(\ell_\infty,s_\infty)$. For that purpose, we introduce as usual the functions $(\ell_n,s_n)\in\mathcal{C}([0,1];V\times V)$ defined by

$$(\ell_n(t,x),s_n(t,x)) := (\ell(t_n+t,x),s(t_n+t,x)) \quad\text{for}\quad (t,x)\in[0,1]^2 \quad\text{and}\quad n\ge1,$$

and claim that

$$\ell_n \longrightarrow \ell_* \quad\text{in}\quad \mathcal{C}([0,1];L^1_w(0,1)), \tag{3.13}$$

$$s_n \longrightarrow s_* \quad\text{in}\quad \mathcal{C}([0,1]\times[0,1]), \tag{3.14}$$

$L^1_w(0,1)$ denoting the usual space $L^1(0,1)$ endowed with its weak topology. Indeed, on the one hand, the continuous embedding of $H^1(0,1)$ in $L^\infty(0,1)$ and the Hölder inequality entail that

$$\begin{aligned}
\|s_n(t)-s_n(0)\|_\infty &\le \int_{t_n}^{t_n+t}\|\partial_t s\|_\infty\ d\tau \le \int_{t_n}^{t_n+t}\|\partial_t\sigma\|_\infty\ d\tau \\
&\le \left(\int_{t_n}^{t_n+t}\|\partial_t\sigma\|_{H^1}^2\ d\tau\right)^{1/2}
\end{aligned}$$

for $t \in [0,1]$, so that

$$\sup_{t\in[0,1]} \|s_n(t) - s_n(0)\|_\infty \le \left(\int_{t_n}^\infty \|\partial_t \sigma\|_{H^1}^2 \, d\tau \right)^{1/2}.$$

By (3.2), the right-hand side of the above inequality converges to zero as $n \to \infty$, which, together with (3.12), gives the claim (3.14). On the other hand, if $\psi \in W = H^2(0,1) \cap V$ (defined in Lemma 7) and $t \in [0,1]$,

$$\begin{aligned}
\left| \int_0^1 (\ell_n(t) - \ell_n(0)) \, \psi \, dx \right| &= \left| \int_{t_n}^{t_n+1} \langle \partial_t \ell(\tau), \psi \rangle_{V',V} \, d\tau \right| \\
&\le \int_{t_n}^{t_n+1} \|\partial_t \ell(\tau)\|_{W'} \, \|\psi\|_W \, d\tau \\
&\le \|\psi\|_W \, \left(\int_{t_n}^{t_n+1} \|\partial_t \ell(\tau)\|_{W'}^2 \, d\tau \right)^{1/2} \\
&\le \|\psi\|_W \, \left(\int_{t_n}^{\infty} \|\partial_t \ell(\tau)\|_{W'}^2 \, d\tau \right)^{1/2}.
\end{aligned}$$

Owing to (3.9) and (3.10), we pass to the limit as $n \to \infty$ in the above inequality to conclude that

$$\lim_{n\to\infty} \sup_{t\in[0,1]} \left| \int_0^1 (\ell_n(t) - \ell_*) \, \psi \, dx \right| = 0 \quad \text{for all} \quad \psi \in W.$$

A classical approximation combined with (3.6) allows us to extend the above convergence to every $\psi \in L^\infty(0,1)$ and completes the proof of (3.13).

We next infer from (3.8) that

$$\begin{aligned}
&\lim_{n\to\infty} \int_0^1 \left\| \frac{1}{\ell_n(t) + \delta + \varepsilon} - \frac{1}{\ell_\infty + \delta + \varepsilon} \right\|_2^2 dt \\
&= \lim_{n\to\infty} \int_{t_n}^{t_n+1} \left\| \frac{1}{\ell(t) + \delta + \varepsilon} - \frac{1}{\ell_\infty + \delta + \varepsilon} \right\|_2^2 dt \\
&= 0.
\end{aligned}$$

Consequently, after extracting a further subsequence if necessary, we may assume that

$$\ell_n \longrightarrow \ell_\infty \quad \text{a.e. in} \quad (0,1) \times (0,1). \tag{3.15}$$

Recalling (3.13) we readily conclude from (3.15) that $\ell_* = \ell_\infty$.

Next, (3.1) and (3.14) imply that $0 \le s_*(x) \le 1 - \vartheta < 1$ for $x \in [0,1]$, and we infer from (1.4), (3.2), (3.13), and (3.14) that, if $\psi \in L^\infty(0,1)$,

$$\begin{aligned}
\left| \int_0^1 \left(\ell_* - (\delta+\varepsilon) \, \frac{s_*}{1-s_*} \right) \psi \, dx \right| &= \left| \lim_{n\to\infty} \int_0^1 \int_0^1 \left(\ell_n(t) - (\delta+\varepsilon) \, \frac{s_n(t)}{1-s_n(t)} \right) \psi(x) \, dxdt \right| \\
&= \left| \lim_{n\to\infty} \int_0^1 \partial_t \sigma(t_n + t) \, \psi \, dxdt \right| \\
&\le \|\psi\|_\infty \, \left(\lim_{n\to\infty} \int_{t_n}^{t_n+1} \|\partial_t \sigma(t)\|_2^2 \, dt \right)^{1/2} \\
&= 0.
\end{aligned}$$

Therefore, $(\delta+\varepsilon) s_*/(1-s_*) = \ell_* = \ell_\infty$ a.e. in $(0,1)$ and thus $s_* = s_\infty$.

We have actually shown that (ℓ_∞, s_∞) is the unique cluster point of $(\ell(t), s(t))$ as $t \to \infty$ for the weak topology of $L^1(0,1)\times H^1(0,1)$. The trajectory $\{(\ell(t), s(t));\ t \ge 0\}$ being relatively compact for that topology by (3.2), (3.6), and the Dunford–Pettis theorem, we conclude that

$$(\ell(t), s(t)) \rightharpoonup (\ell_\infty, s_\infty) \quad \text{in} \quad L^1(0,1) \times H^1(0,1)$$

as $t \to \infty$. The embedding of $H^1(0,1)$ in $L^2(0,1)$ being compact, we further obtain that

$$s(t) \longrightarrow s_\infty \quad \text{in} \quad L^2(0,1). \tag{3.16}$$

Let us finally improve the topology in which the convergence to the steady state takes place. For $\psi \in V$ it follows from the weak formulation of (1.3) and Proposition 2 that

$$\begin{aligned} &\langle \partial_t \ell, \psi \rangle_{V',V} + D \int_0^1 (1-s)\ \partial_x(\ell - \ell_\infty)\ \partial_x \psi\ dx \\ &= D \int_0^1 (s - s_\infty)\ \partial_x \ell_\infty\ \partial_x \psi\ dx + \int_0^1 \left[(\ell_\infty + \delta)\ (s - s_\infty) - (1-s)\ (\ell - \ell_\infty) \right]\ \psi\ dx. \end{aligned}$$

Since $\ell - \ell_\infty$ belongs to V, we may take $\psi = \ell - \ell_\infty$ and use the fact that ℓ_∞ does not depend on time to obtain

$$\begin{aligned} &\frac{1}{2}\ \frac{d}{dt} \|\ell - \ell_\infty\|_2^2 + \int_0^1 (1-s)\ \left(D\ \left| \partial_x (\ell - \ell_\infty) \right|^2 + |\ell - \ell_\infty|^2 \right)\ dx \\ &\le \|s - s_\infty\|_2\ \left(\|\partial_x \ell_\infty\|_\infty\ \left\| \partial_x (\ell - \ell_\infty) \right\|_2 + (\delta + \|\ell_\infty\|_\infty)\ \|\ell - \ell_\infty\|_2 \right) \\ &\le \frac{D\vartheta}{2}\ \left\| \partial_x (\ell - \ell_\infty) \right\|_2^2 + \frac{\vartheta}{2}\ \|\ell - \ell_\infty\|_2^2 + C_{21}\ \|s - s_\infty\|_2^2, \end{aligned}$$

the last inequality following from the boundedness of ℓ_∞ and $\partial_x \ell_\infty$ and the parameter ϑ being defined in Lemma 5. Using (3.1) to bound from below $(1-s)$ in the second term on the left-hand side of the above inequality we get

$$\frac{d}{dt} \|\ell - \ell_\infty\|_2^2 + D\vartheta\ \left\| \partial_x (\ell - \ell_\infty) \right\|_2^2 + \vartheta\ \|\ell - \ell_\infty\|_2^2 \le 2C_{21}\ \|s - s_\infty\|_2^2. \tag{3.17}$$

We next infer from (1.4) and Proposition 2 that

$$\partial_t s = (1-s)\ (\ell - \ell_\infty) - (\ell_\infty + \delta + \varepsilon)\ (s - s_\infty).$$

We differentiate the above identity with respect to x and take the scalar product in $L^2(0,1)$ of the result with $\partial_x(s - s_\infty)$. Recalling the Poincaré inequality (2.13) and the bound (3.2) we obtain

$$\begin{aligned} &\frac{1}{2}\ \frac{d}{dt}\ \left\| \partial_x (s - s_\infty) \right\|_2^2 + \int_0^1 (\ell_\infty + \delta + \varepsilon)\ \left| \partial_x (s - s_\infty) \right|_2^2\ dx \\ &\le \|\partial_x s\|_2\ \|\ell - \ell_\infty\|_\infty\ \left\| \partial_x (s - s_\infty) \right\|_2 + \left\| \partial_x (\ell - \ell_\infty) \right\|_2\ \left\| \partial_x (s - s_\infty) \right\|_2 \\ &+ \|\partial_x \ell_\infty\|_\infty\ \|s - s_\infty\|_2\ \left\| \partial_x (s - s_\infty) \right\|_2 \\ &\le \|(1-s)\ \partial_x \sigma\|_2\ \left\| \partial_x (\ell - \ell_\infty) \right\|_2\ \left\| \partial_x (s - s_\infty) \right\|_2 + \left\| \partial_x (\ell - \ell_\infty) \right\|_2\ \left\| \partial_x (s - s_\infty) \right\|_2 \\ &+ \|\partial_x \ell_\infty\|_\infty\ \|s - s_\infty\|_2\ \left\| \partial_x (s - s_\infty) \right\|_2 \\ &\le C_{22}\ \left(\left\| \partial_x (\ell - \ell_\infty) \right\|_2 + \|s - s_\infty\|_2 \right)\ \left\| \partial_x (s - s_\infty) \right\|_2 \\ &\le \frac{\delta}{2}\ \left\| \partial_x (s - s_\infty) \right\|_2^2 + \frac{C_{23}}{2}\ \left(\left\| \partial_x (\ell - \ell_\infty) \right\|_2^2 + \|s - s_\infty\|_2^2 \right). \end{aligned}$$

Since ℓ_∞ and ε are nonnegative we end up with

$$\frac{d}{dt}\,\|\partial_x\,(s-s_\infty)\|_2^2+\delta\;\|\partial_x\,(s-s_\infty)\|_2^2\le C_{23}\;\left(\|\partial_x\,(\ell-\ell_\infty)\|_2^2+\|s-s_\infty\|_2^2\right). \tag{3.18}$$

We now multiply (3.18) by $D\vartheta/C_{23}$ and add the result to (3.17) to obtain

$$\begin{aligned}\frac{d}{dt}\left(\|\ell-\ell_\infty\|_2^2+\frac{D\vartheta}{C_{23}}\,\|\partial_x\,(s-s_\infty)\|_2^2\right)+\vartheta\;\|\ell-\ell_\infty\|_2^2\\ +\;\frac{D\vartheta\delta}{C_{23}}\,\|\partial_x\,(s-s_\infty)\|_2^2\le C_{24}\;\|s-s_\infty\|_2^2.\end{aligned}$$

Introducing $\omega:=\min\{\vartheta,\delta\}>0$ and integrating the above differential inequality give

$$\begin{aligned}\left(\|\ell(t)-\ell_\infty\|_2^2+\frac{D\vartheta}{C_{23}}\,\|\partial_x\,(s(t)-s_\infty)\|_2^2\right)\le\left(\|\ell_0-\ell_\infty\|_2^2+\frac{D\vartheta}{C_{23}}\,\|\partial_x\,(s_0-s_\infty)\|_2^2\right)\;e^{-\omega t}\\ +\,C_{24}\;\int_0^t\|s(\tau)-s_\infty\|_2^2\;e^{\omega(\tau-t)}\;d\tau\end{aligned}$$

for $t\ge0$. Since $\|s(\tau)-s_\infty\|_2\longrightarrow0$ as $\tau\to\infty$ by (3.16), we deduce from the above inequality that both $\|\ell(t)-\ell_\infty\|_2$ and $\|\partial_x\,(s(t)-s_\infty)\|_2$ converge to zero as $t\to\infty$, and the proof of Theorem 3 is complete. □

4. Well-posedness of (1.1)–(1.2). In this section, we consider the initial-boundary value problem

$$\partial_t\ell=D\;\partial_x^2\ell+\delta\;s-\ell\;(1-s),\quad(t,x)\in(0,\infty)\times(0,1), \tag{4.1}$$

$$\partial_t s=\ell\;(1-s)-(\delta+\varepsilon)\;s,\quad(t,x)\in(0,\infty)\times(0,1), \tag{4.2}$$

$$D\;\partial_x\ell(t,0)+\nu=\ell(t,1)=0,\quad t\in(0,\infty), \tag{4.3}$$

$$(\ell,s)(0,x)=(\ell_0,s_0)(x),\quad x\in(0,1), \tag{4.4}$$

with $D>0$, $\delta>0$, $\varepsilon\ge0$, and $\nu>0$.

THEOREM 8. *Given initial data (ℓ_0,s_0) satisfying (1.7) the initial-boundary value problem (4.1)–(4.4) possesses a unique weak solution*

$$(\ell,s)\in\mathcal{C}([0,\infty);V\times V)\cap\mathcal{C}^1([0,\infty);V'\times V)$$

such that $\partial_t\ell\in L^2((0,T)\times(0,1))$ for every $T>0$,

$$0\le\ell(t,x)\quad\textit{and}\quad0\le s(t,x)<1\quad\textit{for}\quad(t,x)\in[0,\infty)\times[0,1], \tag{4.5}$$

and

$$\begin{aligned}\langle\partial_t\ell(t),\psi\rangle_{V',V}+D\;\int_0^1\partial_x\ell(t)\;\partial_x\psi\;dx\\ =\nu\;\psi(0)+\int_0^1\left(\delta s(t)-\ell(t)(1-s(t))\right)\;\psi\;dx,\end{aligned}$$

$$\partial_t s(t)=-(\delta+\varepsilon)s(t)+\ell(t)(1-s(t))$$

for $\psi\in V$ and $t\in[0,\infty)$. Furthermore, for each $t>0$,

$$\Lambda(\ell(t),s(t))+\int_0^t\mathcal{D}_\Lambda(\ell(\tau),s(\tau))\;d\tau=\Lambda(\ell_0,s_0), \tag{4.6}$$

where

$$\Lambda(u,v) := \frac{1}{2}\ \|u-\ell_\infty\|_2^2 + \Lambda_0(u,v),$$

$$\Lambda_0(u,v) := \int_0^1 (1-s_\infty)\ (\ell_\infty+\delta+2\varepsilon)\ \left[\Sigma_I(v)-\Sigma_I(s_\infty)-\frac{v-s_\infty}{1-s_\infty}\right]\ dx \ge 0,$$

$$\mathcal{D}_\Lambda(u,v) := D\ \|\partial_x(u-\ell_\infty)\|_2^2 + \int_0^1 \frac{|\partial_t v|^2+\varepsilon\ (\ell_\infty+\delta+\varepsilon)\ (v-s_\infty)^2}{1-v}\ dx,$$

$s_\infty := \ell_\infty/(\ell_\infty+\delta+\varepsilon)$, *and* ℓ_∞ *denotes the unique solution to the boundary-value problem*

$$-D\ \partial_x^2\ell_\infty + \frac{\varepsilon\ \ell_\infty}{\ell_\infty+\delta+\varepsilon} = 0 \quad in \quad (0,1), \quad D\ \partial_x\ell_\infty(0)+\nu = \ell_\infty(1) = 0. \tag{4.7}$$

We first point out that the nonnegativity of Λ_0 is a consequence of the convexity of Σ_I. We next observe that Λ differs strongly from $\mathcal{L}$ and provide valuable information on the large time behavior. In fact, exponential convergence to the steady state can be shown; see Proposition 9 below.

Proof. The local well-posedness of (4.1)–(4.4) in $V\times V$ and the bounds (4.5) for (ℓ,s) on the maximal existence time interval $[0,T_m)$ are established as in the proof of Theorem 1 by employing the abstract theory developed in [1]. In addition, since $u\longmapsto \varepsilon u/(u+\delta+\varepsilon)$ is a nondecreasing function, the existence and uniqueness of the solution ℓ_∞ to (4.7) are proved by classical arguments [2, 4].

We next turn to the proof of the identity (4.6) for $t\in[0,T_m)$. For that purpose, we take $\psi=\ell-\ell_\infty$ in the weak formulation of (4.1) and (4.7) for ℓ and ℓ_∞ and use (4.2) for s and s_∞ to obtain

$$\frac{1}{2}\ \frac{d}{dt}\|\ell-\ell_\infty\|_2^2 + D\ \|\partial_x(\ell-\ell_\infty)\|_2^2 = -\int_0^1 (\partial_t s+\varepsilon\ (s-s_\infty))\ (\ell-\ell_\infty)\ dx. \tag{4.8}$$

Owing to (4.2) we have

$$\ell-\ell_\infty = \frac{\partial_t s}{1-s} + (\delta+\varepsilon)\ \frac{s-s_\infty}{(1-s)\ (1-s_\infty)}.$$

Consequently,

$$\begin{aligned}
&-\int_0^1 (\partial_t s+\varepsilon\ (s-s_\infty))\ (\ell-\ell_\infty)\ dx\\
&= -\int_0^1 \frac{|\partial_t s|^2}{1-s}\ dx - \int_0^1 \left[\varepsilon+\frac{\delta+\varepsilon}{1-s_\infty}\right]\ (s-s_\infty)\ \frac{\partial_t s}{1-s}\ dx\\
&\quad -\varepsilon(\delta+\varepsilon)\ \int_0^1 \frac{(s-s_\infty)^2}{(1-s)\ (1-s_\infty)}\ dx\\
&= -\int_0^1 \frac{|\partial_t s|^2}{1-s}\ dx - \int_0^1 (\ell_\infty+\delta+2\varepsilon)\ \partial_t\left((1-s_\infty)\ \Sigma_I(s)-s\right)\ dx\\
&\quad -\varepsilon\ \int_0^1 (\ell_\infty+\delta+\varepsilon)\ \frac{(s-s_\infty)^2}{1-s}\ dx,
\end{aligned}$$

where we have used $(\delta+\varepsilon)/(1-s_\infty)=\ell_\infty+\delta+\varepsilon$ to deduce the last equality. Inserting the previous equality in (4.8) gives (4.6) for $t\in[0,T_m)$ after integration with respect to time.

A useful consequence of (4.6) is that

$$\|\ell(t)\|_2 \le \|\ell_\infty\|_2 + \sqrt{2\ \Lambda(\ell_0,s_0)}, \quad t\in[0,T_m). \tag{4.9}$$

We next proceed as in the proof of (2.18) and take $\psi=\partial_t\ell$ in the weak formulation of (4.1) (recall that a regularization procedure has to be used to justify this step as $\partial_t\ell\notin V$). This gives

$$\|\partial_t\ell\|_2^2 + \frac{D}{2}\ \frac{d}{dt}\|\partial_x\ell\|_2^2 = \int_0^1 \partial_t\ell\ (\delta\ s-\ell\ (1-s))\ dx \le \frac{1}{2}\ \|\partial_t\ell\|_2^2 + \left(\delta+\|\ell\|_2^2\right).$$

We then infer from (4.9) that

$$\int_0^t \|\partial_t\ell(\tau)\|_2^2\ d\tau + D\ \|\partial_x\ell(t)\|_2^2 \le C_{25}\ (1+t), \quad t\in[0,T_m). \tag{4.10}$$

We next differentiate (4.2) to obtain the equation solved by $\partial_x s$ and deduce from (4.10) that

$$\|\partial_x s(t)\|_2 \le C_{26}\ (1+t), \quad t\in[0,T_m). \tag{4.11}$$

Finally, we infer from (4.2), (4.5), (4.9), (4.10), (4.11), and the continuous embedding of $H^1(0,1)$ in $L^\infty(0,1)$ that

$$s\in W^{1,\infty}([0,T]\cap[0,T_m);H^1(0,1)) \quad\text{and}\quad \delta\ s-\ell\ (1-s)\in L^2((0,T)\cap(0,T_m);H^1(0,1))$$

for all $T>0$. Combining this last property with (4.1) and [10, Theorem 4.3.1] ensures that

$$\ell\in\mathcal{C}^{1/2}((0,T]\times(0,T_m);H^1(0,1)) \quad\text{for all}\quad T>0.$$

Collecting the above information allows us to conclude that $T_m=\infty$ and complete the proof. □

We next turn to the large time behavior and prove the following result.

PROPOSITION 9. *Consider initial data (ℓ_0,s_0) satisfying (1.7) and denote by (ℓ,s) the corresponding weak solution to (4.1)–(4.4) given by Theorem 8. Then, for $t\ge 0$,*

$$\|\ell(t)-\ell_\infty\|_2^2 + (\delta+\varepsilon)\ \|s(t)-s_\infty\|_2^2 \le 2\ \Lambda(\ell_0,s_0)\ e^{-\chi t} \tag{4.12}$$

with

$$\chi := \min\left\{D, \frac{D\ (\delta+\varepsilon)}{2(D+2)} + \frac{\varepsilon}{2}\right\}. \tag{4.13}$$

Proposition 9 is a simple consequence of (4.6) and the following two functional inequalities.

LEMMA 10. *Under the assumptions and notation of Proposition 9 we have for all $t\ge 0$*

$$\mathcal{D}_\Lambda(\ell(t),s(t)) \ge \chi\ \Lambda(\ell(t),s(t)), \tag{4.14}$$

$$\Lambda_0(s(t)) \ge \frac{\delta+\varepsilon}{2}\ \|s(t)-s_\infty\|_2^2. \tag{4.15}$$

Proof. By the elementary inequality $a^2 \le ((D+2)/D)\ (a-b)^2 + ((D+2)/2)\ b^2$ we have

$$\begin{aligned}
&\int_0^1 \left(\frac{\delta+\varepsilon}{1-s_\infty}\right)^2 \ \frac{(s-s_\infty)^2}{1-s}\ dx \\
&\le \frac{D+2}{D}\ \int_0^1 (1-s)\ \left[\frac{\delta+\varepsilon}{1-s_\infty}\ \frac{s-s_\infty}{1-s} - (\ell-\ell_\infty)\right]^2\ dx \\
&+ \frac{D+2}{2}\ \int_0^1 (1-s)\ (\ell-\ell_\infty)^2\ dx \\
&\le \frac{D+2}{D}\ \left\{\int_0^1 (1-s)\ \left(\frac{\partial_t s}{1-s}\right)^2\ dx + \frac{D}{2}\ \|\ell-\ell_\infty\|_2^2\right\}, \qquad (4.16)
\end{aligned}$$

the last inequality following from the identities $\ell+\delta+\varepsilon = (\partial_t s+\delta+\varepsilon)/(1-s)$ and

$$\frac{\delta+\varepsilon}{1-s_\infty} = \ell_\infty+\delta+\varepsilon \ge \delta+\varepsilon. \qquad (4.17)$$

We next infer from the Poincaré inequality (2.13), (4.16), and (4.17) that

$$\begin{aligned}
\mathcal{D}_\Lambda(\ell,s) &\ge D\ \|\ell-\ell_\infty\|_2^2 + \int_0^1 \frac{|\partial_t s|^2}{1-s}\ dx + \varepsilon\ (\delta+\varepsilon)\ \int_0^1 \frac{(s-s_\infty)^2}{(1-s)(1-s_\infty)}\ dx \\
&\ge \frac{D}{2}\ \|\ell-\ell_\infty\|_2^2 + \frac{D}{D+2}\ \int_0^1 \left(\frac{\delta+\varepsilon}{1-s_\infty}\right)^2\ \frac{(s-s_\infty)^2}{1-s}\ dx \\
&\quad + \varepsilon\ (\delta+\varepsilon)\ \int_0^1 \frac{(s-s_\infty)^2}{(1-s)(1-s_\infty)}\ dx, \\
\mathcal{D}_\Lambda(\ell,s) &\ge \frac{D}{2}\ \|\ell-\ell_\infty\|_2^2 + (\delta+\varepsilon)\ \left(\frac{D(\delta+\varepsilon)}{D+2}+\varepsilon\right)\ \int_0^1 \frac{(s-s_\infty)^2}{(1-s)(1-s_\infty)}\ dx. \qquad (4.18)
\end{aligned}$$

We next claim that

$$\frac{(z-a)^2}{2} \le \Sigma_I(z)-\Sigma_I(a)-\frac{z-a}{1-a} \le \frac{1}{2}\ \frac{1+a}{1-a}\ \frac{(z-a)^2}{1-z}, \quad (a,z)\in[0,1)\times[0,1). \qquad (4.19)$$

Indeed, the first inequality in (4.19) follows from the convexity of Σ_I, while the second can be shown by studying the variation of the function

$$z \longmapsto (1-z)\ \left[\Sigma_I(z)-\Sigma_I(a)-\frac{z-a}{1-a}\right] - \frac{1}{2}\ \frac{1+a}{1-a}\ (z-a)^2,$$

the real number a being fixed in $[0,1)$. Using the second inequality in (4.19) with $z=s(t,x)$ and $a=s_\infty(x)$ gives

$$\int_0^1 \frac{(s-s_\infty)^2}{(1-s)(1-s_\infty)}\ dx \ge \int_0^1 \frac{2(\delta+\varepsilon)}{1+s_\infty}\left(\Sigma_I(s)-\Sigma_I(s_\infty)-\frac{s-s_\infty}{1-s_\infty}\right)\ dx.$$

Using once more (4.17) we realize that

$$\begin{aligned}
\frac{2(\delta+\varepsilon)}{1+s_\infty} &= \frac{2(\delta+\varepsilon)}{1+s_\infty}\ \frac{(1-s_\infty)\ (\ell_\infty+\delta+2\varepsilon)}{\delta+\varepsilon+\varepsilon(1-s_\infty)} \\
&\ge (\delta+\varepsilon)\ \frac{(1-s_\infty)\ (\ell_\infty+\delta+2\varepsilon)}{\delta+2\varepsilon} \\
&\ge \frac{(1-s_\infty)\ (\ell_\infty+\delta+2\varepsilon)}{2},
\end{aligned}$$

so that

$$\int_0^1 \frac{(s-s_\infty)^2}{(1-s)(1-s_\infty)}\,dx \geq \frac{1}{2}\,\Lambda_0(s).$$

Inserting the previous inequality in (4.18) gives

$$\mathcal{D}_\Lambda(\ell,s) \geq \frac{D}{2}\,\|\ell-\ell_\infty\|_2^2 + \frac{1}{2}\left(\frac{D(\delta+\varepsilon)}{D+2}+\varepsilon\right)\Lambda_0(s) \geq \chi\,\Lambda(\ell,s),$$

and hence (4.14).

Inequality (4.15) readily follows from (4.17) and the first inequality in (4.19). ∎

5. Numerical simulations. Above, we have developed a theory which provides insight into the stationary states and large time behavior of two models of morphogen transport. It is therefore interesting to verify the dependence on the parameters of both the properties of these stationary solutions and the convergence rate to the equilibrium. For this, we have performed numerical simulations on an x86-64 machine under MATLAB 7.2. All source files are available upon request.

We discuss certain qualitative properties of (1.3)–(1.6) and (4.1)–(4.4) for a range of their parameters with default values $D=\nu=\delta=1$ and $\varepsilon=0$. For an evaluation of the two models for biologically relevant parameter values, see [7]. In all experiments, we supply zero initial data, $\ell_0(x)=s_0(x)=0$. For clarity, in what follows we shall refer to (1.3)–(1.6) as the nonlinear diffusion (ND) model and to (4.1)–(4.4) as the linear diffusion (LD) model.

Discretization. To discretize the system (1.3)–(1.6) we employ a finite difference scheme in the spatial variable on a regular mesh with $N=240$ nodes. Introducing $\ell_i(t)\approx\ell(t,ih)$, $s_i(t)\approx s(t,ih)$, where h is the mesh size and using the first order flux approximation,

$$j_i = \left(1-\frac{s_{i+1}+s_i}{2}\right)\frac{\ell_{i+1}-\ell_i}{h},$$

we arrive at a system of $2N$ ordinary differential equations (ODEs)

$$\frac{d\ell_i}{dt} = \frac{j_i-j_{i-1}}{h} + \delta s_i - \ell_i\,(1-s_i)\,, \tag{5.1a}$$

$$\frac{ds_i}{dt} = -(\delta+\varepsilon)s_i + \ell_i\,(1-s_i) \tag{5.1b}$$

for $i=1,\ldots,N$. This discretization has the truncation error of second order with respect to h and is supplemented by a first order approximation of the flux boundary condition. In order to solve (5.1) we use a stiff ODE solver `ode15s` from MATLAB. Since we trace the behavior of the solution up to very small time derivatives, we adjust the absolute and relative error tolerance parameters in `ode15s` to the value of 10^{-13}. Model LD is discretized and solved analogously.

Steady state computation. As observed in [7], without degradation ($\varepsilon=0$) one has explicit formulas for the stationary states, since from (1.13) it follows that for the ND model there holds

$$\ell_\infty(x) = \delta\,\left(\exp\left(\frac{\nu}{\delta D}(1-x)\right)-1\right), \tag{5.2}$$

$$s_\infty(x) = 1-\exp\left(-\frac{\nu}{\delta\,D}\,(1-x)\right), \tag{5.3}$$

while for the LD model, according to Theorem 8,

$$\ell_\infty(x) = \frac{\nu}{D}(1-x), \tag{5.4}$$

$$s_\infty(x) = 1 - \frac{1}{1 + (\nu/(\delta\, D))\,(1-x)}. \tag{5.5}$$

But when $\varepsilon > 0$ there is no closed formula and we have to approximate the steady states numerically. To this end, we either run the simulation until the L^2 norm of the right-hand side of the discrete ODE system drops (unless otherwise stated) below 10^{-6} or we solve the nonlinear system (5.1), where we replace time derivatives with zeros. Note, however, that although in practice the first procedure turns out to be more robust, none of these procedures guarantees that the obtained solution is sufficiently close to the stationary state, particularly when s_∞ is too close to 1; see [7] for a more detailed discussion.

5.1. Dependence of the convergence rate on the parameters. According to Theorem 3, the solutions of the ND model converge to the stationary state. From Proposition 9 we know that for the LD model the distance between the solution and the equilibrium decays at least at an exponential rate. It may therefore be interesting to identify the actual convergence rate of the two models under consideration for several values of the parameters. In order to do so, we investigate the L^2 norm of the relative difference between the stationary solution (ℓ_∞, s_∞) and the solution at time t,

$$d_{rel}(t) = \frac{||(\ell, s)(t) - (\ell_\infty, s_\infty)||_2}{||(\ell_\infty, s_\infty)||_2}.$$

We stop the analysis when d_{rel} drops below 10^{-5}; otherwise the results would have been polluted with the discretization error.

We consider four cases, when only one of the parameters ε, ν, D, or δ is allowed to vary, while the others are kept at their default values.

The convergence to the stationary state seems to occur indeed at an exponential rate; see the top plots in Figures 1–4. Therefore, we can next compare the dependence of the rates on various parameters of the models. To this end, we introduce an approximate measure of the convergence rate ζ as the mean value of the ratios

$$\zeta_i = -\frac{\ln(d_{rel}(t_{i+1})) - \ln(d_{rel}(t_i))}{t_{i+1} - t_i}$$

for several subsequent t_i at which d_{rel} was measured.

As could be expected by comparing the diffusive parts, the LD model always converges faster than the ND model for the same parameter set. From the modeling point of view, as the constants D and δ have different interpretation in each of the two models [7], it is worth mentioning that for moderate values of D, an approximately tenfold increase of D in the nonlinear diffusion model is needed to recover the convergence speed of the linear diffusion model (see Figure 3). Although in the absence of degradation the stationary states of the two models depend only on the ratio $\nu/(\delta\, D)$, the convergence rate may be different for the same ratios of $\nu/(\delta\, D)$, as indicated in Figures 2–4.

Both models show similar dependence of the convergence speed ζ on the degradation rate ε, the unbinding parameter δ, and the diffusion coefficient D. However,

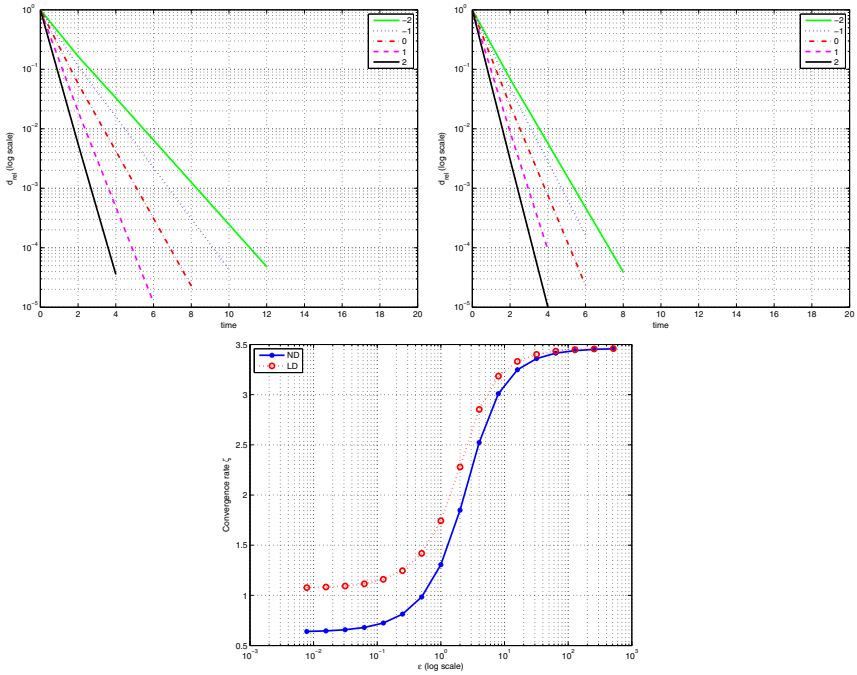


FIG. 1. *Convergence to the equilibrium for different values of ε. The ND model (top left) vs. the LD model (top right) for $\varepsilon = 2^k$, $k = -2, \dots, 2$. The comparison of the convergence rates ζ for a broader range of ε (bottom).*

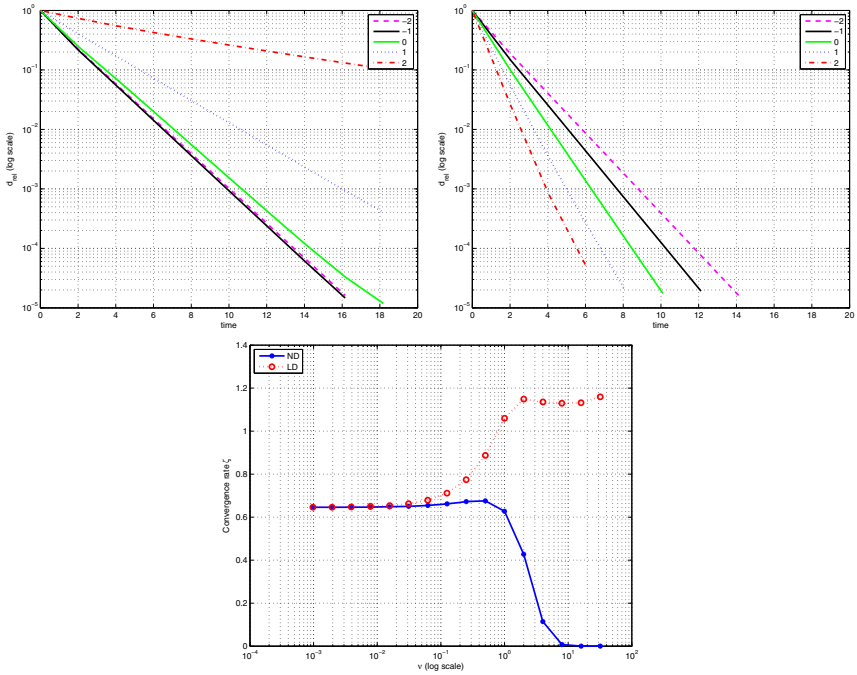


FIG. 2. *Convergence to the equilibrium for different values of ν. The ND model (top left) vs. the LD model (top right) for $\nu = 2^k$, $k = -2, \dots, 2$. The comparison of the convergence rates ζ for a broader range of ν (bottom).*

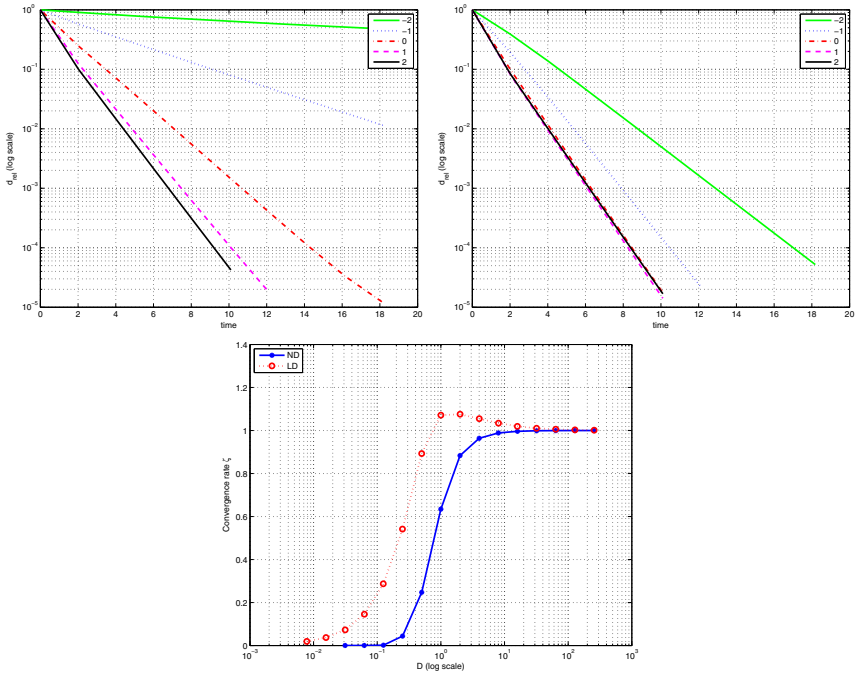


FIG. 3. *Convergence to the equilibrium for different values of D. The ND model (top left) vs. the LD model (top right) for $D = 2^k$, $k = -2, \ldots, 2$. The comparison of the convergence rates ζ for a broader range of D (bottom).*

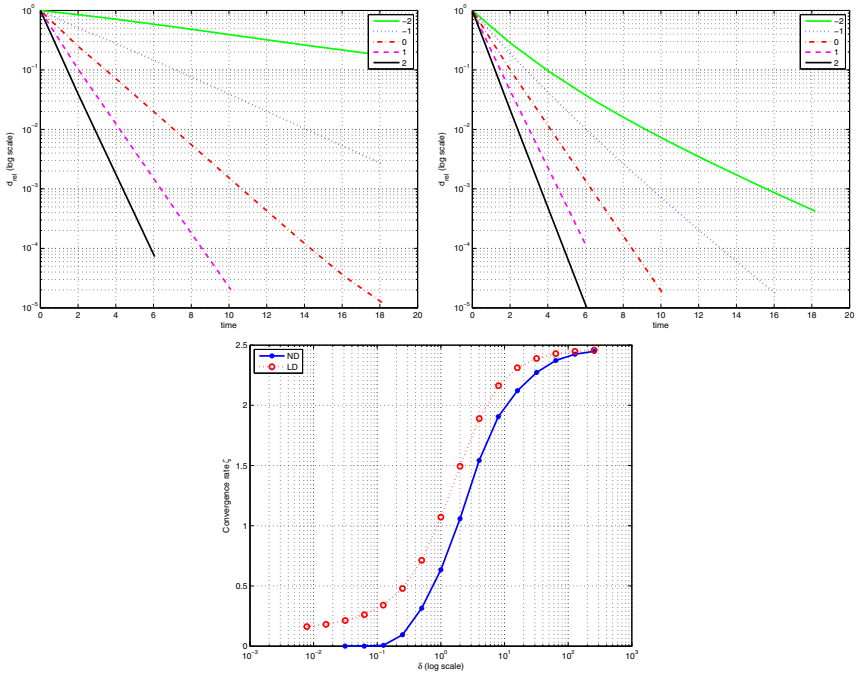


FIG. 4. *Convergence to the equilibrium for different values of δ. The ND model (top left) vs. the LD model (top right) for $\delta = 2^k$, $k = -2, \ldots, 2$. The comparison of the convergence rates ζ for a broader range of δ (bottom).*

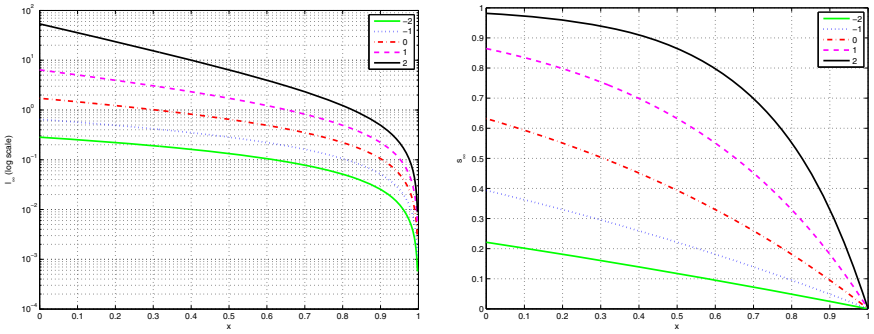


FIG. 5. *Stationary states of the ND model,* $\nu = 2^k$, $k = -2, -1, 0, 1, 2$, *without degradation.*

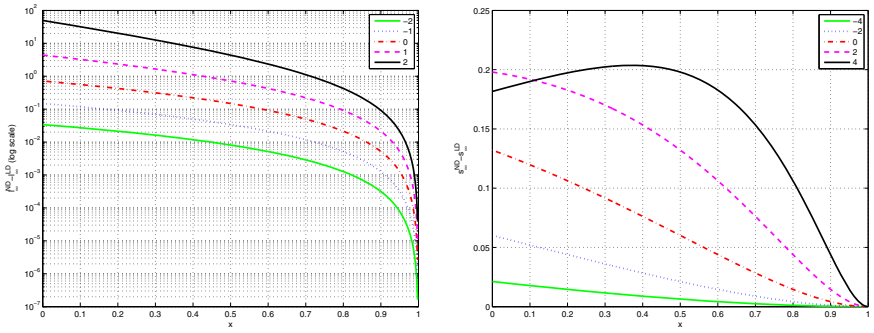


FIG. 6. *Difference between the stationary states of the ND and LD models,* $\nu = 2^k$, $k = -2, -1, 0, 1, 2$, *without degradation.*

the two models differ significantly with respect to their sensitivity to the injection parameter ν: the convergence speed increases with ν for the LD model, while the effect of ν on the convergence of the ND model is opposite; cf. Figure 2 (bottom). This is explained by looking at the stationary states of the latter model; see Figures 5 and 6. The stationary solution reaches, in the case of nonlinear diffusion, higher values than in the linear diffusion case. Indeed, from (5.2) and (5.4) it follows that $\|\ell_\infty\|_\infty = \delta \ (\exp(\nu/(\delta D)) - 1)$ for the ND model which clearly exceeds by far the L^∞ norm ν/D of ℓ_∞ of the LD model. The closer s_∞ gets to 1, the smaller the diffusion, and thus it takes more time for the solution to evolve to the steady state.

It is important to notice that due to the nonlinear effects, both models cannot converge to (ℓ_∞, s_∞) arbitrarily fast, as can be concluded from the bottom parts of Figures 1–4. For each of varying parameters, there is some threshold value of ζ, which cannot be surpassed regardless of the value of the parameter.

Let us observe that for sufficiently large values of D, ε, δ, or $1/\nu$, the convergence speed of both models stabilizes at almost an identical level. On one hand, the similarity in the convergence rates can easily be explained by noticing that the s_∞-component of the steady state of the nonlinear model is so close to zero that the fluxes for both models are nearly the same. On the other hand, the convergence rate stabilization effect is in full agreement with our previous discussion of the Liapunov function for the LD model in the case of varying D. Indeed, denoting by (ℓ^D, s^D) the solution to the LD model for some prescribed D, it follows from (4.6) and (4.7) that

$\ell^D \to 0$ in $L^2(0,T;H^1(0,1))$ as $D \to \infty$. Hence we deduce that s^D converges towards S, solving $\partial_t S = -(\delta+\varepsilon)S$, which exponentially decays to zero at rate $\delta+\varepsilon$ (cf. Figure 3, where $\delta+\varepsilon=1$), to be compared with the L^2 norm decay rate $\chi/2 = \delta/4+\varepsilon/2$ obtained in the limit $D \to \infty$ from (4.13).

5.2. Dependence of the stationary solutions on the parameters. As confirmed by Figures 5–6 and experiments for other parameters not reported here (see also [7]), for identical set of parameters, the two models produce stationary solutions of quite similar shape. There is a difference in how the ν parameter affects the stationary states of the two models; cf. Figures 5–6. The ℓ_∞-component of the stationary solution of the nonlinear diffusion model shows more sensitivity to large values of ν than its linear diffusion counterpart. This is because for large ν, the x-derivative of ℓ is much larger in the nonlinear diffusion case, due to higher saturation level of s. The stationary states of the two models are much less sensitive to the changes in the degradation rate ε. There is only a limited range of ε, close to 1, for which large relative change in ε results in a similar large change in the stationary solutions.

Acknowledgments. Part of this research was carried out during P. Krzyżanowski's and D. Wrzosek's visits to the Institut de Mathématiques de Toulouse and during Ph. Laurençot's visit to the Institute of Applied Mathematics at Warsaw University. The authors thank both institutions for their hospitality. They also thank an anonymous referee for pointing out a gap in the original proof of Theorem 1.

REFERENCES

[1] H. Amann, *Highly degenerate quasilinear parabolic systems*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 18 (1991), pp. 135–166.

[2] Ph. Bénilan, M. G. Crandall, and P. Sacks, *Some L^1 existence and dependence results for semilinear elliptic equations under nonlinear boundary conditions*, Appl. Math. Optim., 17 (1988), pp. 203–224.

[3] T. Bollenbach, K. Kruse, P. Pantazis, M. González-Gaitán, and F. Jülicher, *Robust formation of morphogen gradients*, Phys. Rev. Lett., 94 (2005), article 018103.

[4] H. Brezis and W. A. Strauss, *Semi-linear second-order elliptic equations in L^1*, J. Math. Soc. Japan, 25 (1973), pp. 565–590.

[5] E. V. Entchev and M. A. González-Gaitán, *Morphogen gradient formation and vesicular trafficking*, Traffic, 3 (2002), pp. 98–109.

[6] M. Kerszberg and L. Wolpert, *Mechanisms for positional signalling by morphogen transport: A theoretical study*, J. Theoret. Biol., 191 (1998), pp. 103–114.

[7] P. Krzyżanowski, Ph. Laurençot, and D. Wrzosek, *Mathematical models of receptor-mediated transport of morphogens*, submitted.

[8] A. D. Lander, Q. Nie, and F. Y. M. Wan, *Do morphogen gradients arise by diffusion?*, Dev. Cell, 2 (2002), pp. 785–796.

[9] J. H. Merkin, D. J. Needham, and B. D. Sleeman, *A mathematical model for the spread of morphogens with density dependent chemosensitivity*, Nonlinearity, 18 (2005), pp. 2745–2773.

[10] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, New York, 1983.

[11] J.-P. Vincent and L. Dubois, *Morphogen transport along epithelia, an integrated trafficking problem*, Dev. Cell, 3 (2002), pp. 615–623.

SIAM J. MATH. ANAL.
Vol. 40, No. 5, pp. 1750–1782

ON THE WHITHAM EQUATIONS FOR THE DEFOCUSING COMPLEX MODIFIED KDV EQUATION*

YUJI KODAMA†, V. U. PIERCE†, AND FEI-RAN TIAN†

Abstract. We study the Whitham equations for the defocusing complex modified KdV (mKdV) equation. These Whitham equations are quasi-linear hyperbolic equations and describe the averaged dynamics of the rapid oscillations which appear in the solution of the mKdV equation when the dispersive parameter is small. The oscillations are referred to as dispersive shocks. The Whitham equations for the mKdV equation are neither strictly hyperbolic nor genuinely nonlinear. We are interested in the solutions of the Whitham equations when the initial values are given by a step-like function. We also compare the results with those of the defocusing nonlinear Schrödinger (NLS) equation. For the NLS equation, the Whitham equations are strictly hyperbolic and genuinely nonlinear. We show that the weak hyperbolicity of the mKdV–Whitham equations is responsible for some new structure in the dispersive shocks which has not been found in the NLS case.

Key words. Whitham equations, non-strictly hyperbolic equations, dispersive shocks

AMS subject classifications. 35L65, 35L67, 35Q05, 35Q15, 35Q53, 35Q58

DOI. 10.1137/070705131

1. Introduction.

In [11, 12], Pierce and Tian studied the self-similar solutions of the Whitham equations which describe the zero dispersion limits of the KdV hierarchy. The main feature of the Whitham equations for the higher members of the hierarchy, of which the KdV equation is the first member, is that these Whitham equations are neither strictly hyperbolic nor genuinely nonlinear. This is in sharp contrast to the case of the KdV equation whose Whitham equations are strictly hyperbolic and genuinely nonlinear [8]. In this paper, we extend their studies to the case of the complex modified KdV (mKdV) equation, which is the second member of the defocusing nonlinear Schrödinger (NLS) hierarchy. The Whitham equations for the defocusing NLS equation are strictly hyperbolic and genuinely nonlinear, and they have been studied extensively (see, for example, [4, 6, 7, 10, 14]). However, for the mKdV equation, the Whitham equations are neither strictly hyperbolic nor genuinely nonlinear.

Let us begin with a brief description of the zero dispersion limit of the solution of the NLS equation

$$\sqrt{-1}\,\epsilon\frac{\partial\psi}{\partial t}+2\epsilon^2\frac{\partial^2\psi}{\partial x^2}-4|\psi|^2\psi=0, \tag{1.1}$$

with the initial data

$$\psi(x,0)=A_0(x)\exp\left(\sqrt{-1}\,\frac{S_0(x)}{\epsilon}\right).$$

Here $A_0(x)$ and $S_0(x)$ are real functions that are independent of ϵ. Writing the solution $\psi(x,t;\epsilon)=A(x,t;\epsilon)\exp(\sqrt{-1}\,\frac{S(x,t;\epsilon)}{\epsilon})$ and using the notation $\rho(x,t;\epsilon)=A^2(x,t;\epsilon)$,

*Received by the editors October 11, 2007; accepted for publication (in revised form) July 8, 2008; published electronically December 17, 2008.
http://www.siam.org/journals/sima/40-5/70513.html
†Department of Mathematics, Ohio State University, 231 W. 18th Avenue, Columbus, OH 43210 (kodama@math.ohio-state.edu, vpierce@math.ohio-state.edu, tian@math.ohio-state.edu). The first and third authors were supported in part by NSF grant DMS-0404931. The second author was supported in part by NSF grant DMS-0135308.

$v(x,t;\epsilon) = \partial S(x,t;\epsilon)/\partial x$, one obtains the conservation form of the defocusing NLS equation,

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \dfrac{\partial}{\partial x}(4\rho v) = 0, \\ \dfrac{\partial}{\partial t}(\rho v) + \dfrac{\partial}{\partial x}\left(4\rho v^2 + 2\rho^2\right) = \epsilon^2 \dfrac{\partial}{\partial x}\left(\rho \dfrac{\partial^2}{\partial x^2} \ln \rho\right). \end{cases} \tag{1.2}$$

The mass density $\rho = |\psi|^2$ and momentum density $\rho v = \frac{\sqrt{-1}}{2}(\psi\psi_x^* - \psi^*\psi_x)$ have weak limits as $\epsilon \to 0$ [6]. These limits satisfy a 2×2 system of hyperbolic equations

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \dfrac{\partial}{\partial x}(4\rho v) = 0, \\ \dfrac{\partial}{\partial t}(\rho v) + \dfrac{\partial}{\partial x}\left(4\rho v^2 + 2\rho^2\right) = 0 \end{cases} \tag{1.3}$$

until its solution develops a shock. System (1.3) can be rewritten in the diagonal form for $\rho \neq 0$,

$$\frac{\partial}{\partial t}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} + 2\begin{pmatrix} 3\alpha + \beta & 0 \\ 0 & \alpha + 3\beta \end{pmatrix} \frac{\partial}{\partial x}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = 0, \tag{1.4}$$

where the Riemann invariants α and β are given by

$$\alpha = \frac{v}{2} + \sqrt{\rho}, \quad \beta = \frac{v}{2} - \sqrt{\rho}. \tag{1.5}$$

As a simple example, we consider the case with $\alpha = a =$ constant. System (1.4) reduces to a single equation

$$\frac{\partial \beta}{\partial t} + 2(a + 3\beta)\frac{\partial \beta}{\partial x} = 0. \tag{1.6}$$

The solution is given by the implicit form

$$\beta(x,t) = h(x - 2(a + 3\beta)t),$$

where $h(x) = \beta(x,0)$ is the initial function for β. One can easily see that if $\beta(x,0)$ decreases in some region, then $\beta(x,t)$ develops a shock in a finite time; i.e., $\partial\beta/\partial x$ becomes singular.

After the shock formation in the solution of (1.3) or (1.4), the weak limits are described by the NLS–Whitham equations, which can also be put in the Riemann invariant form [4, 6, 7, 10]

$$\frac{\partial u_i}{\partial t} + \lambda_{g,i}(u_1, \ldots, u_{2g+2})\frac{\partial u_i}{\partial x} = 0, \quad i = 1, 2, \ldots, 2g+2, \tag{1.7}$$

where $\lambda_{g,i}$ are expressed in terms of complete hyperelliptic integrals of genus g [8]. Here the number g is exactly the number of phases in the NLS oscillations with small dispersion. Accordingly, the zero phase $g = 0$ corresponds to no oscillations, and single and higher phases $g \geq 1$ correspond to the NLS oscillations. System (1.4) is viewed as the zero phase Whitham equations. The solution of the Whitham equations (1.7) for $g \geq 1$ then describes the averaged motion of the oscillations appearing in the solution of (1.1) (see, e.g., [7]).

Let us discuss the most important $g = 1$ case in more detail. We note that it is well known that the KdV oscillatory solution, in the single phase regime, can be approximately described by the KdV periodic solution when the dispersive parameter is small [1, 5, 16]. It is very possible to use the method of [1, 16] to show that the solution of the NLS equation (1.1) for small ϵ can be approximately described, in the single phase regime, by the periodic solution of the NLS equation. The NLS periodic solution has the form

$$\tilde{\rho}(x,t;\epsilon) = \rho_3 + (\rho_2 - \rho_3)\,\mathrm{sn}^2\left(\sqrt{\rho_1 - \rho_3}\;\frac{\theta(x,t)}{\epsilon}, s\right) \tag{1.8}$$

with $\theta(x,t) = x - V_1 t - Q$, where $V_1 = V_1(\rho_1,\rho_2,\rho_3)$ and Q is the phase shift. Here ρ_i's are determined by the equation obtained from (1.2)

$$\frac{\epsilon^2}{4}\left(\frac{d\rho}{d\theta}\right)^2 = (\rho-\rho_1)(\rho-\rho_2)(\rho-\rho_3)$$

with $\rho_1 > \rho_2 > \rho_3$, and $\mathrm{sn}(z,s)$ is the Jacobi elliptic function with the modulus $s = (\rho_2 - \rho_3)/(\rho_1 - \rho_3)$. We can also write ρ_i's as [3]

$$\begin{cases} \rho_1 = \dfrac{1}{4}(u_1 + u_2 - u_3 - u_4)^2, \\ \rho_2 = \dfrac{1}{4}(u_1 - u_2 + u_3 - u_4)^2, \\ \rho_3 = \dfrac{1}{4}(u_1 - u_2 - u_3 + u_4)^2 \end{cases} \tag{1.9}$$

with $u_1 > u_2 > u_3 > u_4$. The velocity V_1 is then given by

$$V_1 = 2(u_1 + u_2 + u_3 + u_4).$$

It can also be shown that the flow velocity of the periodic solution (1.8) is given by [3]

$$\tilde{v}(x,t;\epsilon) = \frac{V_1}{4} + \frac{\sqrt{\rho_1\rho_2}}{2\rho}\,(u_1 - u_2 - u_3 + u_4). \tag{1.10}$$

For constants u_1, u_2, u_3, and u_4, formula (1.8) gives the well-known elliptic solution of the NLS equation. To describe the solution $\rho(x,t;\epsilon)$ of the NLS equation (1.2), the quantities u_1, u_2, u_3, and u_4, are instead functions of x and t and evolve according to the single phase Whitham equations (1.7) for $g = 1$. The phase shift Q is also a function of u_1, u_2, u_3, u_4, and depends on the initial values of system (1.2). Suppose the initial function for system (1.4) is given by

$$\alpha(x,0) = a, \quad \beta(x,0) = \beta_0(x), \tag{1.11}$$

where a is a constant and $\beta_0(x)$ is a monotone function. The Whitham solution of (1.7) has the property that $u_1 = a$ and that only u_2, u_3, u_4 are nontrivial functions of (x,t). Then the phase shift $Q(u_1,u_2,u_3,u_4)$ with $u_1 = a$ is the unique solution of the boundary value problem for the Euler–Poisson–Darboux equations

$$2(u_i - u_j)\frac{\partial^2 Q}{\partial u_i \partial u_j} = \frac{\partial Q}{\partial u_i} - \frac{\partial Q}{\partial u_j}, \quad i,j = 2,3,4, \tag{1.12}$$

$$Q(a,u,u,u) = f(u), \tag{1.13}$$

where boundary value $f(u)$ is the inverse of the initial function $\beta_0(x)$ (see Appendix C for details). If, instead, $\beta(x,0)$ is a constant and $\alpha(x,0)$ is given by a monotone function, the phase shift Q is determined by a similar boundary value problem for the Euler–Poisson–Darboux equations.

The weak limit of $\rho(x,t;\epsilon)$ of the NLS equation (1.1) as $\epsilon \to 0$ can be expressed in terms of ρ_1, ρ_2, ρ_3, and ρ_4 [6] as

$$\overline{\rho(x,t)} = \rho_1 - (\rho_1 - \rho_3)\frac{E(s)}{K(s)}, \tag{1.14}$$

where $K(s)$ and $E(s)$ are the complete elliptic integrals of the first and second kind, respectively. This weak limit can also be viewed as the average value of the periodic solution $\tilde{\rho}(x,t;\epsilon)$ of (1.8) over its period $L = 2\epsilon K(s)/\sqrt{\rho_1 - \rho_3}$.

In order to see how a single phase Whitham solution appears, we consider the following step-like initial data for system (1.4):

$$\alpha(x,0) = a, \quad \beta(x,0) = \begin{cases} b, & x < 0, \\ c, & x > 0, \end{cases} \tag{1.15}$$

where $a > b$, $a > c$, $b \neq c$. The solution of (1.4) develops a shock if and only if $b > c$ (cf. (1.6)). After the formation of a shock, the Whitham equations (1.7) with $g = 1$ kick in. For instance, we consider the Whitham equations with the initial data [7]

$$u_1(x,0) = a, \quad u_2(x,0) = b, \quad u_3(x,0) = \begin{cases} b, & x < 0, \\ c, & x > 0, \end{cases} \quad u_4(x,0) = c. \tag{1.16}$$

Now notice that the Whitham equations (1.7) for $g = 1$ with the initial data (1.16) can be reduced to a single equation $u_{3t} + \lambda_{1,3}(a,b,u_3,c)u_{3x} = 0$. The equation has a global self-similar solution, which is implicitly given by $x/t = \lambda_3(a,b,u_3,c)$. The x-t plane is then divided into *three* parts:

$$(1)\ \frac{x}{t} < \gamma, \quad (2)\ \gamma < \frac{x}{t} < 2a + 4b + 2c, \quad (3)\ \frac{x}{t} > 2a + 4b + 2c,$$

where $\gamma = 2(a+b+2c) - 8(a-c)(b-c)/(a+b-2c)$ (see (2.7) and (2.8) for the derivation). The solution of system (1.4) occupies the first and third parts; i.e.,

(1) for $x/t < \gamma$,

$$\alpha(x,t) = a, \quad \beta(x,t) = b,$$

(3) for $x/t > 2a + 4b + 2c$,

$$\alpha(x,t) = a, \quad \beta(x,t) = c.$$

The Whitham solution of (1.7) with $g = 1$ lives in the second part; i.e.,

(2) for $\gamma < x/t < 2a + 4b + 2c$,

$$u_1(x,t) = a, \quad u_2(x,t) = b, \quad \frac{x}{t} = \lambda_{1,3}(a,b,u_3,c), \quad u_4(x,t) = c,$$

where the solution u_3 can be obtained as a function of the self-similarity variable x/t if

$$\frac{\partial \lambda_{1,3}}{\partial u_3}(a,b,u_3,c) \neq 0.$$

Indeed, it has been shown that the Whitham equations (1.7) are genuinely nonlinear [6, 7]; i.e.,

$$\frac{\partial \lambda_{1,i}}{\partial u_i}(u_1, u_2, u_3, u_4) > 0, \quad i = 1, 2, 3, 4, \tag{1.17}$$

for $u_1 > u_2 > u_3 > u_4$.

In Figure 1.1, we plot the self-similar solution of the Whitham equations (1.7) with $g = 1$ for the NLS equation and the corresponding periodic oscillatory solution (1.8) for the initial data (1.15) with $a = 4$, $b = 1$, and $c = -1$. The oscillations describe a dispersive shock of the NLS equation under a small dispersion. Note here that the oscillations have uniform structure, which is due to an almost linear profile of the Whitham solution u_3. This will be seen to be in sharp contrast to the case of the mKdV equation, which we will discuss later (cf. Figure 1.2).

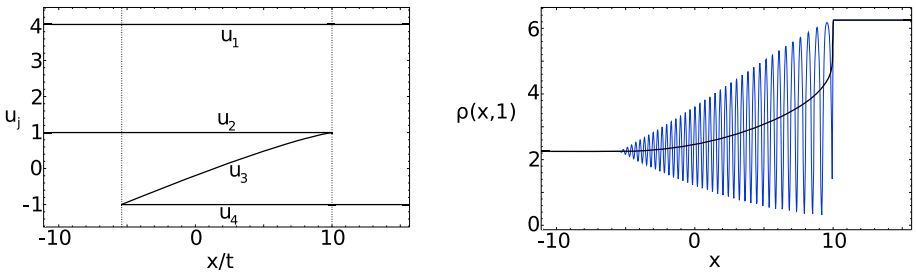


FIG. 1.1. *Self-similar solution of the NLS–Whitham equation* (1.7) *with* $g = 1$ *and the corresponding oscillatory solution* (1.8) *of the NLS equation with* $\epsilon = 0.3$ *at* $t = 1$. *The dark line in the middle of the oscillations is the weak limit* $\overline{\rho(x,t)}$ *given by* (1.14) *at* $t = 1$. *The initial data are given by* (1.15) *with* $a = 4$, $b = 1$, *and* $c = -1$.

Most of the figures in this paper have the same form: On the left-hand side is a plot of the solution of the Whitham equations as a function of the self-similarity variable x/t, which is exact, other than a numerical method used to implement the inverse function theorem. On the right-hand side is the oscillatory solution given by (1.8) at $t = 1$ (respectively, (1.25) for mKdV), while the dark plot is the weak limit (1.14) of the oscillatory solution; both plots on the right are also exact. For all the step-like initial data that we study in this paper, the resulting phase shift Q, which is determined by (1.12) and (1.13), is always zero for both NLS and mKdV. This is due to the fact that our step-like initial function has a jump discontinuity at the origin, which implies that $f \equiv 0$.

In the plots on the left-hand sides of Figures 1.1 and 1.2, we demark the region where the single phase Whitham equations govern the solution and label the four functions $u_1 > u_2 > u_3 > u_4$. The demarcation and labeling are similar in the other figures, and we will omit them for brevity.

The defocusing NLS equation is just the first member of the defocusing NLS hierarchy; the second is the (defocusing) complex mKdV equation

$$\frac{\partial \psi}{\partial t} + \frac{3}{2}\,|\psi|^2 \frac{\partial \psi}{\partial x} - \frac{\epsilon^2}{4}\frac{\partial^3 \psi}{\partial x^3} = 0. \tag{1.18}$$

We again use $\psi(x,t;\epsilon) = A(x,t;\epsilon)\exp(\sqrt{-1}\,\frac{S(x,t;\epsilon)}{\epsilon})$ and notation $\rho(x,t;\epsilon) = A^2(x,t;\epsilon)$, $v(x,t;\epsilon) = \partial S(x,t;\epsilon)/\partial x$ to obtain the conservation form of the mKdV

equation

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \dfrac{\partial}{\partial x}\left(\dfrac{3}{4}\rho^2 + \dfrac{3}{4}\rho v^2\right) = \epsilon^2 \dfrac{\partial}{\partial x}\left(\rho^{3/4}\dfrac{\partial^2}{\partial x^2}\rho^{1/4}\right), \\ \dfrac{\partial}{\partial t}(\rho v) + \dfrac{\partial}{\partial x}\left(\dfrac{3}{2}\rho^2 v + \dfrac{3}{4}\rho v^3\right) = \dfrac{\epsilon^2}{4}\ \dfrac{\partial}{\partial x}\left[\dfrac{\partial^2}{\partial x^2}(\rho v) - \dfrac{3}{2}R\right], \end{cases} \tag{1.19}$$

where

$$R = \frac{3v}{2\rho}\left(\frac{\partial \rho}{\partial x}\right)^2 + \frac{\partial v}{\partial x}\frac{\partial \rho}{\partial x} - v\frac{\partial^2 \rho}{\partial x^2}.$$

The mass density ρ and momentum density ρv for the mKdV equation also have weak limits as $\epsilon \to 0$ [6]. As in the NLS case, the weak limits satisfy

$$\begin{cases} \dfrac{\partial}{\partial t}\rho + \dfrac{\partial}{\partial x}\left(\dfrac{3}{4}\rho^2 + \dfrac{3}{4}\rho v^2\right) = 0, \\ \dfrac{\partial}{\partial t}(\rho v) + \dfrac{\partial}{\partial x}\left(\dfrac{3}{2}\rho^2 v + \dfrac{3}{4}\rho v^3\right) = 0, \end{cases} \tag{1.20}$$

until the solution of (1.20) forms a shock. One can rewrite equations (1.20) as

$$\frac{\partial}{\partial t}\begin{pmatrix}\alpha \\ \beta\end{pmatrix} + \frac{3}{8}\begin{pmatrix}5\alpha^2 + 2\alpha\beta + \beta^2 & 0 \\ 0 & \alpha^2 + 2\alpha\beta + 5\beta^2\end{pmatrix}\frac{\partial}{\partial x}\begin{pmatrix}\alpha \\ \beta\end{pmatrix} = 0, \tag{1.21}$$

where the Riemann invariants α and β are again given by formula (1.5).

Let us again consider the simplest case $\alpha(x,0) = a$, where a is constant, to see how the solution of system (1.21) develops a shock. In this case, the system reduces to a single equation

$$\frac{\partial \beta}{\partial t} + \frac{3}{8}\left(a^2 + 2a\beta + 5\beta^2\right)\frac{\partial \beta}{\partial x} = 0.$$

As in the NLS case, we consider the initial function given by $\beta(x,0) = b$ for $x < 0$ and $\beta(x,0) = c$ for $x > 0$. We recall that, in the NLS case, the zero phase solution of (1.4) develops a shock if and only if $b > c$. However, the solution in the mKdV case develops a shock for $b > c$ if and only if $a + 5b > 0$. In addition, if $b < c$, the solution in the mKdV case develops a shock if and only if $a + 5b < 0$. These differences between the mKdV and NLS cases are due to the weak hyperbolicity of the system (1.21) (note that, for the eigenspeed $\lambda = \frac{3}{8}(\alpha^2 + 2\alpha\beta + 5\beta^2)$ for β, we have $\partial\lambda/\partial\beta = \frac{3}{4}(\alpha + 5\beta)$ which can change sign). As will be shown below, this leads to new structure in the dispersive shocks for the mKdV case.

As in the case of the NLS equation, immediately after the shock formation in the solution of (1.20), the weak limits are described by the mKdV–Whitham equations

$$\frac{\partial u_i}{\partial t} + \mu_{g,i}(u_1, \ldots, u_{2g+2})\frac{\partial u_i}{\partial x} = 0, \quad i = 1, 2, \ldots, 2g+2, \tag{1.22}$$

where $\mu_{g,i}$ can also be expressed in terms of complete hyperelliptic integrals of genus g [6].

In this paper, we study the solution of the Whitham equations (1.22) with $g = 1$ when the initial mass density $\rho(x,0)$ and momentum density $\rho(x,0)v(x,0)$ are

step-like functions. In view of (1.5), this amounts to requiring α and β of system (1.21) to have step-like initial data. We are interested in the following two cases:

(i) $\alpha(x,0)$ is a constant, and

$$\alpha(x,0)=a,\quad \beta(x,0)=\begin{cases} b, & x<0,\\ c, & x>0,\end{cases}\qquad a>b,\ a>c,\ b\neq c, \tag{1.23}$$

(ii) $\beta(x,0)$ is a constant, and

$$\alpha(x,0)=\begin{cases} b, & x<0,\\ c, & x>0,\end{cases}\quad \beta(x,0)=a,\quad b>a,\ c>a,\ b\neq c. \tag{1.24}$$

In the case of the NLS equation, the genuine nonlinearity of the single phase Whitham equations (see (1.17)) warrants that the solution is found by the implicit function theorem. However, the mKdV–Whitham equations (1.22) generally are not genuinely nonlinear; that is, a property like (1.17) is not available (see Appendix B). Our construction of solutions of the Whitham equations (1.22) with $g=1$ makes use of the non-strict hyperbolicity of the equations. For the NLS case, it is known from [6, 7] that the Whitham equations (1.7) with $g=1$ are strictly hyperbolic; that is,

$$\lambda_{1,1}>\lambda_{1,2}>\lambda_{1,3}>\lambda_{1,4}$$

for $u_1>u_2>u_3>u_4$. For the mKdV–Whitham equations (1.22) with $g=1$, the eigenspeeds $\mu_{1,i}(u_1,u_2,u_3,u_4)$ may coalesce in the region $u_1>u_2>u_3>u_4$.

Let us now describe one of our main results (see Theorem 3.1) for the single phase mKdV–Whitham equations with step-like initial function (1.23) for $a=4$, $b=1$, and $c=-1$. In this case, the space time is divided into *four* regions (see Figure 1.2) instead of *three* in the case of the NLS equation (cf. Figure 1.1):

$$(1)\ \frac{x}{t}<c_1,\quad (2)\ c_1<\frac{x}{t}<c_2,\quad (3)\ c_2<\frac{x}{t}<c_3,\quad (4)\ \frac{x}{t}>c_3,$$

where c_1, c_2, and c_3 are some constants. In the first and fourth regions, the solution of the 2×2 system (1.21) governs the evolution:

(1) for $x/t<c_1$,

$$\alpha(x,t)=4,\quad \beta(x,t)=1;$$

(4) for $x/t>c_3$,

$$\alpha(x,t)=4,\quad \beta(x,t)=-1.$$

The Whitham solution of the 4×4 system (1.22) with $g=1$ lives in the second and third regions:

(2) for $c_1<x/t<c_2$,

$$u_1(x,t)=4,\quad u_2(x,t)=1,\quad \frac{x}{t}=\mu_{1,3}(4,1,u_3,u_4),\quad \frac{x}{t}=\mu_{1,4}(4,1,u_3,u_4);$$

(3) for $c_2<x/t<c_3$,

$$u_1(x,t)=4,\quad u_2(x,t)=1,\quad \frac{x}{t}=\mu_{1,3}(4,1,u_3,-1),\quad u_4(x,t)=-1.$$

Note that, in the second region, we have

$$\mu_{1,3}(4,1,u_3,u_4) = \mu_{1,4}(4,1,u_3,u_4)$$

on a curve in the region $-1 < u_4 < u_3 < 1$. This implies the non-strict hyperbolicity of the mKdV–Whitham equations (1.22) for $g = 1$.

It is again possible to use the method of [1, 16] to show that the solution of the mKdV equation (1.18) can be approximately described, in the single phase regime, by the periodic solution of the mKdV when ϵ is small. The periodic solution has the same form as (1.8) of the NLS, i.e.,

$$\tilde{\rho}(x,t;\epsilon) = \rho_3 + (\rho_2 - \rho_3)\,\mathrm{sn}^2\left(\sqrt{\rho_1 - \rho_3}\;\frac{\theta(x,t)}{\epsilon}, s\right). \tag{1.25}$$

However, $\theta(x,t)$ is now given by $\theta = x - V_2 t - Q$ with the velocity V_2 (see, e.g., [9]):

$$V_2 = \frac{3}{8}\sigma_1^2 - \frac{1}{2}\sigma_2,$$

where $\sigma_1 = \sum_{j=1}^{4} u_j$ and $\sigma_2 = \sum_{i<j} u_i u_j$ are the elementary symmetric functions of degree one and two, respectively. The functions ρ_1, ρ_2, and ρ_3 are also given by formula (1.9), and the flow velocity formula (1.10) is still valid for the mKdV. If u_1, u_2, u_3, and u_4 are constants, formula (1.25) gives the periodic solution of the mKdV equation. To describe the solution $\rho(x,t;\epsilon)$ of the mKdV equation (1.18), the quantities u_1, u_2, u_3, and u_4 must satisfy the single phase mKdV-Whitham equations (1.22) for $g = 1$. The phase shift Q is still determined by (1.12) and (1.13) if the initial function of system (1.21) is given by (1.11). The weak limit of $\rho(x,t;\epsilon)$ of the mKdV equation is also given by formula (1.14).

In Figure 1.2, we plot the self-similar solution of the Whitham equations (1.22) for $g = 1$ and the corresponding periodic oscillatory solution (1.25). We note here that the pattern of the oscillations in this case has two different kinds of structure: one corresponds to the region (2), $c_1 < x/t < c_2$, and the other corresponds to the region (3), $c_2 < x/t < c_3$. In Figure 1.2, those regions are separated by a dashed-dotted line. This is in sharp contrast to the NLS case where the oscillations have uniform structure (cf. Figure 1.1). We also note that the weak limit $\overline{\rho(x,t)}$ and the envelope of the oscillations are not C^1 smooth at $x/t = c_2 \approx 3.67$, while they are smooth (even analytic) everywhere within the oscillatory region in the NLS case. Finally, the shapes of the envelopes of the oscillations are quite different: the one on the right of Figure 1.2 looks like a Bordeaux glass and that of Figure 1.1 resembles a martini glass.

As we will show below, for other values of a, b, and c, the solutions of (1.21) and (1.22) with $g = 1$ will be seen to be quite different from the above.

The Whitham equations (1.22) with $g = 1$ for the mKdV equation are analogous to the Whitham equations for the fifth order KdV equation [11]; both Whitham equations are neither strictly hyperbolic nor genuinely nonlinear. For all the step-like initial shock data, the single phase Whitham solutions for the fifth order KdV are also constructed using the non-strict hyperbolicity of the equations. In the case of KdV, the Whitham equations are strictly hyperbolic and genuinely nonlinear, and the oscillations (dispersive shock) have uniform structure. However, in the case of the fifth order KdV, new structure has been found in the dispersive shocks. This structure is similar to the one found here in the dispersive shocks of the mKdV equation.

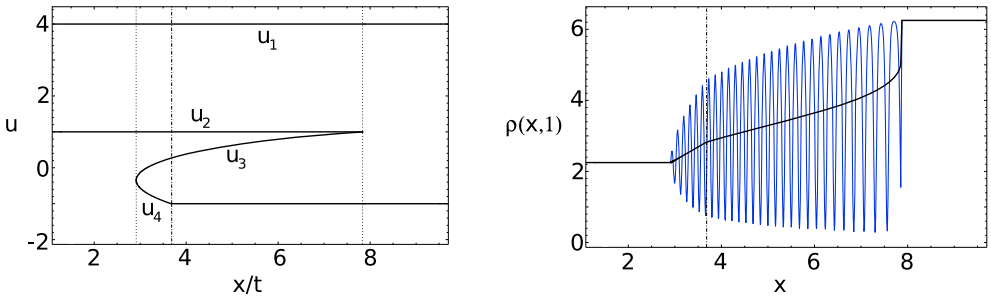


FIG. 1.2. *Self-similar solution of the mKdV–Whitham equation* (1.22) *with* $g = 1$ *and the corresponding oscillatory solution* (1.25) *of the mKdV equation with* $\epsilon = 0.1$. *The initial data are given by* (1.23) *with* $a = 4$, $b = 1$, *and* $c = -1$. *There are two different kinds of structure in the oscillations, and they are separated by* $x/t \approx 3.67$. *Both the weak limit and the envelope of the oscillations have a drastic change at* $x/t \approx 3.67$.

There are still some significant differences between the mKdV oscillations and the fifth order KdV oscillations. The biggest difference lies in the fact that, in the former case, the zero phase Whitham equations (1.21) form a system of two equations, while in the latter case, the zero phase Whitham equation is a scalar equation. In this paper, we consider initial functions (1.23) and (1.24) in which either $\alpha(x,0)$ or $\beta(x,0)$ is a constant; thus equations (1.21) reduce to a single equation. However, the constant initial function $\alpha(x,0) = a$ of (1.23) or $\beta(x,0) = a$ of (1.24) still has a significant effect on the behavior of the Whitham solutions (cf. Figures 3.1 and 4.1). Another difference is that mKdV solution ρ can have points (x,t) at which $\rho(x,t;\epsilon) = 0$ for a sequence of vanishing ϵ (see section 5). These points put the mKdV dispersive approximation (1.19) at stake. This phenomenon does not occur in the fifth order KdV case.

The Whitham equations for the higher members of the KdV hierarchy are again neither strictly hyperbolic nor genuinely nonlinear. However, step-like initial shock data will, in general, generate a mixture of single, double, or even higher phases in the higher order KdV oscillations. It has been an open problem to construct such multiphase Whitham solutions. Only rather special step-like initial shock data will produce merely single phase oscillations in the higher order KdV case. Some of these initial data have been studied in [12]. The analogues of inequalities (2.15) and (2.16), which play a crucial role in the fifth order KdV case, are not valid anymore in the higher order KdV case. As a consequence, the approach in [12] is quite different from that in [11]. Indeed, the calculations in [12] are considerably more difficult than in [11].

The Whitham equations for the higher members of the KdV hierarchy are supposed to be analogous to the Whitham equations for the higher members of the (defocusing) NLS hierarchy. It would be very interesting to see how step-like initial shock data generate multiphases in the higher order NLS oscillations.

The organization of the paper is as follows. In section 2, we will study the eigenspeeds $\mu_{g,1}$, $\mu_{g,2}$, $\mu_{g,3}$, and $\mu_{g,4}$ of the Whitham equations (1.22) for $g = 1$. In section 3, we will construct the self-similar solutions of the single phase Whitham equations for the initial function (1.23) with $a > b > c$. In section 4, we will construct the self-similar solution of the Whitham equations for the initial function (1.23) with $a > c > b$. In section 5, we will study those points (x,t) at which the mKdV solution $\rho(x,t;\epsilon) = 0$ for a sequence of vanishing ϵ. In section 6, we will briefly discuss how to handle the other step-like initial data (1.24).

2. The Whitham equations. In this section we define the eigenspeeds $\lambda_{g,i}$ and $\mu_{g,i}$ of the Whitham equations (1.7) and (1.22) with $g = 1$ for the NLS and the mKdV equations. For simplicity, we suppress the subscript $g = 1$ in the notation $\lambda_{g,i}$ and $\mu_{g,i}$ in the rest of the paper.

We first introduce the polynomials of ξ for $n = 0, 1, 2, \dots$ [2, 4, 10]:

$$P_n(\xi, u_1, u_2, u_3, u_4) = \xi^{n+2} + a_{n,1}\xi^{n+1} + \cdots + a_{n,n+2}, \tag{2.1}$$

where the coefficients $a_{n,1}, a_{n,2}, \dots, a_{n,n+2}$ are uniquely determined by the two conditions

$$\frac{P_n(\xi, u_1, u_2, u_3, u_4)}{\sqrt{(\xi - u_1)(\xi - u_2)(\xi - u_3)(\xi - u_4)}} = \xi^n + \mathcal{O}(\xi^{-2}) \quad \text{for large } |\xi|$$

and

$$\int_{u_2}^{u_1} \frac{P_n(\xi, u_1, u_2, u_3, u_4)}{\sqrt{(u_1 - \xi)(\xi - u_2)(\xi - u_3)(\xi - u_4)}} d\xi = 0.$$

The coefficients of P_n can be expressed in terms of complete elliptic integrals.

The eigenspeeds of the Whitham equations (1.7) with $g = 1$ for the NLS equation are defined in terms of P_0 and P_1 of (2.1) [4, 6, 10],

$$\lambda_i(u_1, u_2, u_3, u_4) = 8\, \frac{P_1(u_i, u_1, u_2, u_3, u_4)}{P_0(u_i, u_1, u_2, u_3, u_4)}, \quad i = 1, 2, 3, 4,$$

which give

$$\lambda_i(u_1, u_2, u_3, u_4) = 2\left(\sigma_1(u_1, u_2, u_3, u_4) - \frac{I(u_1, u_2, u_3, u_4)}{\partial_{u_i} I(u_1, u_2, u_3, u_4)}\right). \tag{2.2}$$

Here $\sigma_1 := \sum_{j=1}^4 u_j$, and $I(u_1, u_2, u_3, u_4)$ is given by a complete elliptic integral [14]

$$I(u_1, u_2, u_3, u_4) = \int_{u_2}^{u_1} \frac{d\eta}{\sqrt{(u_1 - \eta)(\eta - u_2)(\eta - u_3)(\eta - u_4)}}. \tag{2.3}$$

The function I can be rewritten as a contour integral. Hence,

$$2(u_i - u_j)\frac{\partial^2 I}{\partial u_i \partial u_j} = \frac{\partial I}{\partial u_i} - \frac{\partial I}{\partial u_j}, \quad i, j = 1, 2, 3, 4, \tag{2.4}$$

since the integrand satisfies the same equations for each $\eta \neq u_i$ (cf. (1.12)). This contour integral connection also allows us to give another formulation of I,

$$I(u_1, u_2, u_3, u_4) = \int_{u_4}^{u_3} \frac{d\eta}{\sqrt{(u_1 - \eta)(u_2 - \eta)(u_3 - \eta)(\eta - u_4)}}. \tag{2.5}$$

It follows from (2.2), (2.3), and (2.5) that

$$\lambda_4 - 2\sigma_1 < \lambda_3 - 2\sigma_1 < 0 < \lambda_2 - 2\sigma_1 < \lambda_1 - 2\sigma_1 \tag{2.6}$$

for $u_4 < u_3 < u_2 < u_1$. This implies the strict hyperbolicity of the NLS–Whitham equation (1.7) for $g = 1$.

The eigenspeeds λ_i have the following values [14]: At $u_3 = u_4$, we have

$$\begin{cases} \lambda_1 = 6u_1 + 2u_2, \\ \lambda_2 = 2u_1 + 6u_2, \\ \lambda_3 = \lambda_4 = 2(u_1 + u_2 + 2u_4) - \frac{8(u_1-u_4)(u_2-u_4)}{u_1+u_2-2u_4}, \end{cases} \tag{2.7}$$

and at $u_2 = u_3$,

$$\begin{cases} \lambda_1 = 6u_1 + 2u_4, \\ \lambda_2 = \lambda_3 = 2u_1 + 4u_3 + 2u_4, \\ \lambda_4 = 2u_1 + 6u_4. \end{cases} \tag{2.8}$$

Notice that the eigenspeed $\lambda_2 = \lambda_3$ at $u_2 = u_3$ is the same as the velocity of the periodic solution (1.8), i.e., $V_1 = 2\sigma_1 = 2(u_1 + 2u_3 + u_4)$.

The eigenspeeds of the mKdV–Whitham equations (1.22) with $g = 1$ are [6]

$$\mu_i(u_1, u_2, u_3, u_4) = 3\,\frac{P_2(u_i, u_1, u_2, u_3, u_4)}{P_0(u_i, u_1, u_2, u_3, u_4)}, \quad i = 1, 2, 3, 4. \tag{2.9}$$

They can be expressed in terms of λ_1, λ_2, λ_3, and λ_4 of the NLS–Whitham equations (1.7) with $g = 1$.

LEMMA 2.1. *The eigenspeeds $\mu_i(u_1, u_2, u_3, u_4)$ of* (2.9) *can be expressed in the form*

$$\mu_i = \frac{1}{2}\,(\lambda_i - 2\sigma_1)\,\frac{\partial q}{\partial u_i} + q, \quad i = 1, 2, 3, 4, \tag{2.10}$$

where $\sigma_1 = \sum_{j=1}^{4} u_j$ and $q = q(u_1, u_2, u_3, u_4)$ is the solution of the boundary value problem of the Euler–Poisson–Darboux equations (cf. (1.12) *and* (2.4)*)*

$$2(u_i - u_j)\frac{\partial^2 q}{\partial u_i \partial u_j} = \frac{\partial q}{\partial u_i} - \frac{\partial q}{\partial u_j}, \quad i, j = 1, 2, 3, 4, \tag{2.11}$$

$$q(u, u, u) = 3u^2.$$

Also the μ_i satisfy the overdetermined systems

$$\frac{1}{\mu_i - \mu_j}\,\frac{\partial \mu_i}{\partial u_j} = \frac{1}{\lambda_i - \lambda_j}\,\frac{\partial \lambda_i}{\partial u_j}, \qquad i \neq j. \tag{2.12}$$

We omit the proof since it is very similar to the proof of an analogous result for the KdV hierarchy [13].

The boundary value problem (2.11) has a unique solution. The solution is a symmetric quadratic function of u_1, u_2, u_3, and u_4:

$$q = \frac{3}{8}\sigma_1^2 - \frac{1}{2}\sigma_2, \tag{2.13}$$

where $\sigma_2 = \sum_{i>j} u_i u_j$ is the elementary symmetric polynomial of degree two. Notice that q gives the velocity of the periodic solution (1.25) for the mKdV equation, i.e., $V_2 = q$.

For NLS, λ_i satisfy [14]

$$\frac{\partial\lambda_4}{\partial u_4} < \frac{3}{2}\frac{\lambda_3-\lambda_4}{u_3-u_4} < \frac{\partial\lambda_3}{\partial u_3} \tag{2.14}$$

for $u_4 < u_3 < u_2 < u_1$. Similar results also hold for the mKdV–Whitham equations (1.22) with $g = 1$.

LEMMA 2.2.

$$\frac{\partial\mu_3}{\partial u_3} > \frac{3}{2}\ \frac{\mu_3-\mu_4}{u_3-u_4} \quad \textit{if}\ \ \frac{\partial q}{\partial u_3} > 0, \tag{2.15}$$

$$\frac{\partial\mu_4}{\partial u_4} < \frac{3}{2}\ \frac{\mu_3-\mu_4}{u_3-u_4} \quad \textit{if}\ \ \frac{\partial q}{\partial u_4} > 0 \tag{2.16}$$

for $u_4 < u_3 < u_2 < u_1$.

Proof. We use (2.10) and (2.14) to obtain

$$\begin{aligned}\frac{\partial\mu_3}{\partial u_3} &= \frac{1}{2}\frac{\partial\lambda_3}{\partial u_3}\frac{\partial q}{\partial u_3} + \frac{1}{2}(\lambda_3-2\sigma_1)\frac{\partial^2 q}{\partial u_3^2}\\ &> \frac{3}{4}\frac{\lambda_3-\lambda_4}{u_3-u_4}\frac{\partial q}{\partial u_3} + \frac{1}{2}(\lambda_4-2\sigma_1)\frac{\partial^2 q}{\partial u_3^2}\end{aligned} \tag{2.17}$$

and

$$\begin{aligned}\mu_3-\mu_4 &= \frac{1}{2}\,(\lambda_3-\lambda_4)\,\frac{\partial q}{\partial u_3} + \frac{1}{2}(\lambda_4-2\sigma_1)\left(\frac{\partial q}{\partial u_3}-\frac{\partial q}{\partial u_4}\right)\\ &= \frac{1}{2}\,(\lambda_3-\lambda_4)\,\frac{\partial q}{\partial u_3} + (\lambda_4-2\sigma_1)(u_3-u_4)\frac{\partial^2 q}{\partial u_3\partial u_4}\\ &= \frac{2}{3}(u_3-u_4)\left(\frac{3}{4}\,\frac{\lambda_3-\lambda_4}{u_2-u_3}\,\frac{\partial q}{\partial u_3} + \frac{3}{2}(\lambda_4-2\sigma_1)\frac{\partial^2 q}{\partial u_3\partial u_4}\right),\end{aligned} \tag{2.18}$$

where we have used (2.11)

$$\frac{\partial q}{\partial u_3}-\frac{\partial q}{\partial u_4} = 2(u_3-u_4)\frac{\partial^2 q}{\partial u_3\partial u_4}.$$

Differentiating this equation with respect to u_3 yields

$$\frac{\partial^2 q}{\partial u_3^2} - 3\frac{\partial^2 q}{\partial u_3\partial u_4} = 2(u_3-u_4)\frac{\partial^3 q}{\partial u_3^2\partial u_4}.$$

It is here we exploit the fact that q is a quadratic polynomial and conclude that

$$3\frac{\partial^2 q}{\partial u_3\partial u_4} = \frac{\partial^2 q}{\partial u_3^2},$$

which, along with (2.17) and (2.18), proves (2.15). Inequality (2.16) can be proved in the same way. ∎

The following calculations will be useful in the subsequent sections. Using formula (2.10) for μ_3 and μ_4 and formula (2.2) for λ_3 and λ_4, we obtain

$$\begin{aligned}\mu_3 - \mu_4 &= \frac{I}{(\partial_{u_3} I)(\partial_{u_4} I)}\left[\frac{\partial q}{\partial u_4}\frac{\partial I}{\partial u_3} - \frac{\partial q}{\partial u_3}\frac{\partial I}{\partial u_4}\right] \\ &= \frac{I}{(\partial_{u_3} I)(\partial_{u_4} I)}\left[\frac{\partial q}{\partial u_4}\left(\frac{\partial I}{\partial u_3} - \frac{\partial I}{\partial u_4}\right) - \left(\frac{\partial q}{\partial u_3} - \frac{\partial q}{\partial u_4}\right)\frac{\partial I}{\partial u_4}\right] \\ &= \frac{2I(u_3 - u_4)}{(\partial_{u_3} I)(\partial_{u_4} I)}\ M, \qquad (2.19)\end{aligned}$$

where

$$M = \frac{\partial q}{\partial u_4}\ \frac{\partial^2 I}{\partial u_3 \partial u_4} - \frac{\partial^2 q}{\partial u_3 \partial u_4}\ \frac{\partial I}{\partial u_4}.$$

Here we have used equations (2.4) for I and equations (2.11) for q in equality (2.19). Since q of (2.13) is quadratic, we obtain

$$\frac{\partial M}{\partial u_3} = \frac{\partial q}{\partial u_4}\ \frac{\partial^3 I}{\partial u_3^2 \partial u_4}. \qquad (2.20)$$

We note that another expression for M is

$$M = \frac{\partial q}{\partial u_3}\ \frac{\partial^2 I}{\partial u_3 \partial u_4} - \frac{\partial^2 q}{\partial u_3 \partial u_4}\ \frac{\partial I}{\partial u_3}.$$

Hence, we get

$$\frac{\partial M}{\partial u_4} = \frac{\partial q}{\partial u_3}\ \frac{\partial^3 I}{\partial u_3 \partial u_4^2}. \qquad (2.21)$$

We next evaluate $M(u_1, u_2, u_3, u_4)$ when $u_3 = u_4$. Using the integral formula (2.3) for the function I and applying the change of variable $\eta = (u_1 - u_2)\nu + u_2$, we obtain

$$\begin{aligned}M\Big|_{u_3=u_4} &= \frac{\frac{\partial q}{\partial u_4}}{4(u_2 - u_4)^3}\int_0^1 \frac{d\nu}{(1 + \frac{u_1 - u_2}{u_2 - u_4}\nu)^3\sqrt{\nu(1-\nu)}} \\ &\quad - \frac{\frac{\partial^2 q}{\partial u_3 \partial u_4}}{2(u_2 - u_4)^2}\int_0^1 \frac{d\nu}{(1 + \frac{u_1 - u_2}{u_2 - u_4}\nu)^2\sqrt{\nu(1-\nu)}}.\end{aligned}$$

The two integrals can be evaluated exactly as

$$\int_0^1 \frac{d\nu}{(1+\gamma\nu)^3\sqrt{\nu(1-\nu)}} = \frac{\pi(8 + 8\gamma + 3\gamma^2)}{8(1+\gamma)^{\frac{5}{2}}}, \quad \int_0^1 \frac{d\nu}{(1+\gamma\nu)^2\sqrt{\nu(1-\nu)}} = \frac{\pi(2+\gamma)}{2(1+\gamma)^{\frac{3}{2}}}$$

for $\gamma > -1$. We finally get

$$M\Big|_{u_3=u_4} = \frac{\pi U(u_1, u_2, u_4)}{128[(u_2 - u_4)(u_1 - u_4)]^{\frac{5}{2}}}, \qquad (2.22)$$

where

$$U(u_1,u_2,\xi) = [8(u_2-\xi)^2 + 8(u_2-\xi)(u_1-u_2) + 3(u_1-u_2)^2](u_1+u_2+4\xi)$$

$$- 8(u_1-\xi)(u_2-\xi)(u_1+u_2-2\xi). \tag{2.23}$$

Similarly to (2.19) for μ_3 and μ_4, we have

$$\mu_2 - \mu_3 = \frac{2I(u_2-u_3)}{(\partial_{u_2} I)(\partial_{u_3} I)}\ N, \tag{2.24}$$

where

$$N = \frac{\partial q}{\partial u_2}\ \frac{\partial^2 I}{\partial u_2 \partial u_3} - \frac{\partial^2 q}{\partial u_2 \partial u_3}\ \frac{\partial I}{\partial u_2}.$$

Since q of (2.13) is quadratic, we obtain

$$\frac{\partial N}{\partial u_3} = \frac{\partial q}{\partial u_2}\ \frac{\partial^3 I}{\partial u_2 \partial u_3^2}. \tag{2.25}$$

Finally, we use (2.7) and (2.10) to calculate

$$(\mu_2-\mu_3)\Big|_{u_3=u_4} = \frac{1}{2}[\lambda_2 - 2(u_1+u_2+2u_4)]\frac{\partial q}{\partial u_2} - \frac{1}{2}[\lambda_3 - 2(u_1+u_2+2u_4)]\frac{\partial q}{\partial u_3}$$

$$= \frac{(u_2-u_4)}{2(u_1+u_2-2u_4)}\ V(u_1,u_2,u_4), \tag{2.26}$$

where

$$V(u_1,u_2,u_4) = 3u_1^2 + 3u_2^2 - 12u_4^2 + 6u_1u_2 + 6u_1u_4 - 6u_2u_4. \tag{2.27}$$

3. Self-similar solutions. In this section, we construct self-similar solutions of the Whitham equations (1.22) with $g=1$ for the initial function (1.23) with $a>b>c$. The case with $a>c>b$ will be studied in next section. The solution of the zero phase Whitham equations (1.21) does not develop a shock when $a+5b\le 0$. We are therefore interested only in the case $a+5b>0$.

We first study the ξ-zero of the cubic polynomial equation

$$U(a,b,\xi) = 0, \tag{3.1}$$

where U is given by (2.23). It is easy to prove that, for each pair of a and b satisfying $a>b$ and $a+5b>0$, $U(a,b,\xi)=0$ has only one simple real root. Denoting this zero by $\xi(a,b)$, we then deduce that $U(a,b,\xi)$ is positive for $\xi>\xi(a,b)$ and negative for $\xi<\xi(a,b)$. Since $U(a,b,-(a+b)/4)<0$ in view of (2.23), we must have

$$\xi(a,b) > -\frac{a+b}{4}. \tag{3.2}$$

For initial function (1.23) with $a>b>c$ and $a+5b>0$, we now classify the resulting Whitham solutions into four types:

I. $\xi(a,b)\le c$ with any $a>b>c$,
II. $\xi(a,b)>c$ with $a+5b>3(b-c)>0$,
III. $\xi(a,b)>c$ with $a+5b=3(b-c)>0$,
IV. $\xi(a,b)>c$ with $0<a+5b<3(b-c)$.

In Figure 3.1, we illustrate this classification by scaling $a>0$ to $a=4$ and plotting regions for types I–IV in the b-c plane.

We will study the second type first.

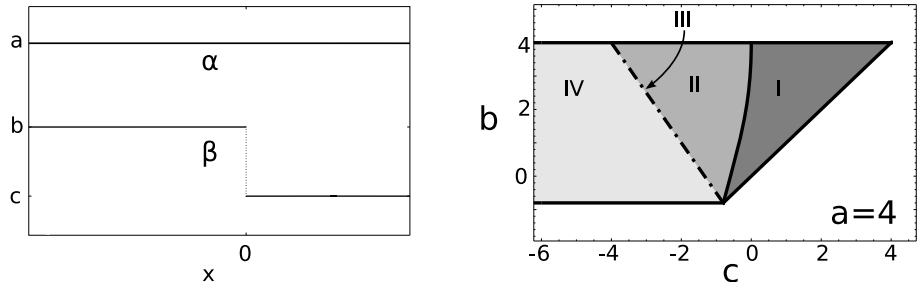


FIG. 3.1. *Graph of the step-like initial shock data* (1.23) *with* $a > b > c$ *and* $a + 5b > 0$*, and diagram of regions for types* I–IV*, where* $a > 0$ *has been scaled to* $a = 4$*. The region for type* IV *is unbounded from the left.*

3.1. Type II. Here we consider the step-like initial function (1.23) satisfying $\xi(a, b) > c$ and $a + 5b > 3(b - c) > 0$.

THEOREM 3.1 (see Figure 1.2). *For the step-like initial data* (1.23) *with* $a > b > c$*,* $a + 5b > 3(b - c)$*, and* $\xi(a, b) > c$*, the solution* (α, β) *of the zero phase Whitham equations* (1.21) *and the solution* (u_1, u_2, u_3, u_4) *of the single phase Whitham equations* (1.22) *with* $g = 1$ *are given as follows:*

(1) *For* $x/t \le \mu_3(a, b, \xi(a, b), \xi(a, b))$,

$$\alpha = a, \quad \beta = b. \tag{3.3}$$

(2) *For* $\mu_3(a, b, \xi(a, b), \xi(a, b)) < x/t < \mu_3(a, b, u^{**}, c)$,

$$u_1 = a, \quad u_2 = b, \quad \frac{x}{t} = \mu_3(a, b, u_3, u_4), \quad \frac{x}{t} = \mu_4(a, b, u_3, u_4), \tag{3.4}$$

where u^{**} *is the unique solution* u_3 *of* $\mu_3(a, b, u_3, c) = \mu_4(a, b, u_3, c)$ *in the interval* $c < u_3 < b$.

(3) *For* $\mu_3(a, b, u^{**}, c) \le x/t < \mu_3(a, b, b, c)$,

$$u_1 = a, \quad u_2 = b, \quad \frac{x}{t} = \mu_3(a, b, u_3, c), \quad u_4 = c. \tag{3.5}$$

(4) *For* $x/t \ge \mu_3(a, b, b, c)$,

$$\alpha = a, \quad \beta = c. \tag{3.6}$$

The boundaries $x/t = \mu_3(a, b, \xi(a, b), \xi(a, b))$ and $x/t = \mu_3(a, b, b, c)$ are called the trailing and leading edges, respectively. They separate the solutions of the single phase Whitham equations (1.22) with $g = 1$ and the zero phase Whitham equations (1.21). The single phase Whitham solution matches the zero phase Whitham solution in the following fashion (see Figure 1.2):

$$(u_1, u_2) = \text{the solution } (\alpha, \beta) \text{ of (1.21) defined outside the region,} \tag{3.7}$$

$$u_3 = u_4, \tag{3.8}$$

at the trailing edge;

$$(u_1, u_4) = \text{the solution } (\alpha, \beta) \text{ of (1.21) defined outside the region,} \tag{3.9}$$

$$u_2 = u_3, \tag{3.10}$$

at the leading edge.

The proof of Theorem 3.1 is based on a series of lemmas: We first show that the solutions defined by formulae (3.4) and (3.5) indeed satisfy the Whitham equations (1.22) for $g = 1$ [2, 15].

LEMMA 3.2.

(1) *The functions u_1, u_2, u_3, and u_4 determined by equations* (3.4) *give a solution of the Whitham equations* (1.22) *with $g = 1$ as long as u_3 and u_4 can be solved from* (3.4) *as functions of x and t.*

(2) *The functions u_1, u_2, u_3, and u_4 determined by equations* (3.5) *give a solution of the Whitham equations* (1.22) *with $g = 1$ as long as u_3 can be solved from* (3.5) *as a function of x and t.*

Proof. (1) u_1 and u_2 obviously satisfy the first two equations of (1.22) for $g = 1$. To verify the third and fourth equations, we observe that

$$\frac{\partial \mu_3}{\partial u_4} = \frac{\partial \mu_4}{\partial u_3} = 0 \tag{3.11}$$

on the solution of (3.4). To see this, we use (2.12) to calculate

$$\frac{\partial \mu_3}{\partial u_4} = \frac{\frac{\partial \lambda_3}{\partial u_4}}{\lambda_3 - \lambda_4} (\mu_3 - \mu_4) = 0.$$

The second part of (3.11) can be shown in the same way. We then calculate the partial derivatives of the third equation of (3.4) with respect to x and t,

$$1 = \frac{\partial \mu_3}{\partial u_3} tu_{3x}, \quad 0 = \frac{\partial \mu_3}{\partial u_3} tu_{3t} + \mu_3,$$

which give the third equation of (1.22) with $g = 1$. The fourth equation of (1.22) with $g = 1$ can be verified in the same way.

(2) The second part of Lemma 3.2 can easily be proved. □

We now determine the trailing edge. Eliminating x and t from the last two equations of (3.4) yields

$$\mu_3(a, b, u_3, u_4) - \mu_4(a, b, u_3, u_4) = 0. \tag{3.12}$$

Since it degenerates at $u_3 = u_4$, we replace (3.12) by

$$F(a, b, u_3, u_4) := \frac{\mu_3(a, b, u_3, u_4) - \mu_4(a, b, u_3, u_4)}{u_3 - u_4} = 0. \tag{3.13}$$

Therefore, at the trailing edge where $u_3 = u_4$, (3.13), in view of formulae (2.19) and (2.22), reduces to

$$U(a, b, u_4) = 0. \tag{3.14}$$

Noting that $\xi(a, b)$ is the unique solution of (3.1), we then deduce that $u_4 = \xi(a, b)$.

LEMMA 3.3. *Equation* (3.13) *has a unique solution satisfying $u_3 = u_4$. The solution is $u_3 = u_4 = \xi(a, b)$. The rest of equations* (3.4) *at the trailing edge are $u_1 = a$, $u_2 = b$, and $x/t = \mu_3(a, b, \xi(a, b), \xi(a, b))$.*

Having located the trailing edge, we now solve equations (3.4) in the neighborhood of the trailing edge. We first consider (3.13). We use (2.19) to write F of (3.13) as

$$F(a, b, u_3, u_4) = \frac{2I}{(\partial_{u_3} I)(\partial_{u_3} I)} M(a, b, u_3, u_4).$$

We note that, at the trailing edge $u_3 = u_4 = \xi(a,b)$, we have $M(a,b,\xi(a,b),\xi(a,b)) = 0$ because of (2.22) and (3.14). We then use (2.20) and (2.21) to differentiate F at the trailing edge

$$\frac{\partial F(a,b,\xi(a,b),\xi(a,b))}{\partial u_3} = \frac{\partial F(a,b,\xi(a,b),\xi(a,b))}{\partial u_4}$$

$$= \frac{I}{2(\partial_{u_3} I)(\partial_{u_3} I)}\ [a+b+4\xi(a,b)]\ \frac{\partial^3 I}{\partial u_3^2 \partial u_4} > 0,$$

where we have used the expression (2.13) for q in the last equation and (3.2) in the inequality. These show that (3.13) or, equivalently, (3.12) can be inverted to give u_4 as a decreasing function of u_3,

$$u_4 = A(u_3), \tag{3.15}$$

in a neighborhood of $u_3 = u_4 = \xi(a,b)$.

We now extend the solution $A(u_3)$ of (3.12) in the region $c < u_4 < \xi(a,b) < u_3 < b$ as far as possible. We first claim that

$$\frac{\partial q(a,b,u_3,u_4)}{\partial u_3} > 0, \quad \frac{\partial q(a,b,u_3,u_4)}{\partial u_4} > 0 \tag{3.16}$$

on the extension. To see this, we first observe that inequalities (3.16) are true at the trailing edge $u_3 = u_4 = \xi(a,b)$. This follows from (2.13) and (3.2). Therefore, inequalities (3.16) hold in a neighborhood of the trailing edge. To prove that (3.16) remains true on the extension, we use formula (2.10) for μ_3 and μ_4 to rewrite (3.12) as

$$\frac{1}{2}[\lambda_3 - 2(a+b+u_3+u_4)]\frac{\partial q}{\partial u_3} = \frac{1}{2}[\lambda_4 - 2(a+b+u_3+u_4)]\frac{\partial q}{\partial u_4}.$$

Since the two terms in the two sets of parentheses are both negative in view of (2.6) and since $\frac{\partial q}{\partial u_3} - \frac{\partial q}{\partial u_4} = (u_3 - u_4)/4 > 0$ in view of (2.13), neither $\frac{\partial q}{\partial u_3}$ nor $\frac{\partial q}{\partial u_4}$ can vanish on the extension. This proves inequalities (3.16).

We deduce from Lemma 2.2 that

$$\frac{\partial \mu_3}{\partial u_3} > 0, \quad \frac{\partial \mu_4}{\partial u_4} < 0 \tag{3.17}$$

on the solution of (3.12). Because of (3.11) and (3.17), solution (3.15) of (3.12) can be extended as long as $c < u_4 < \xi(a,b) < u_3 < b$.

There are two possibilities: (1) u_3 touches b before or simultaneously as u_4 reaches c and (2) u_4 touches c before u_3 reaches b. It follows from (2.8), (2.10), and (2.13) that

$$\mu_3(a,b,b,u_4) - \mu_4(a,b,b,u_4) = \frac{1}{2}\ (b-u_4)(a+2b+3u_4) > 0 \quad \text{for } c \le u_4 < b, \tag{3.18}$$

where we have used $a+2b+3c > 0$ in the inequality. This shows that (1) is unattainable. Hence, u_4 will touch c before u_3 reaches b. When this happens, (3.12) becomes

$$\mu_3(a,b,u_3,c) = \mu_4(a,b,u_3,c). \tag{3.19}$$

LEMMA 3.4. *Equation* (3.19) *has a simple zero in the interval* $c < u_3 < b$, *counting multiplicities. If we denote the zero by* u^{**}, *then* $\mu_3(a,b,u_3,c) - \mu_4(a,b,u_3,c)$ *is positive for* $u_3 > u^{**}$ *and negative for* $u_3 < u^{**}$.

Proof. We use (2.19) and (2.20) to prove the lemma. In both formulae, $\partial_{u_3} I$, $\partial_{u_4} I$, and $\partial^2_{u_3 u_4} I$ are all positive functions. By (2.20),

$$\frac{\partial M(a,b,u_3,c)}{\partial u_3} = \frac{(a+b+u_3+3c)}{4} \frac{\partial^3 I}{\partial u_3^2 \partial u_4} \quad \text{for } c < u_3 < b. \tag{3.20}$$

We claim that

$$M(a,b,u_3,c) < 0 \quad \text{when } u_3 = c \qquad \text{and} \qquad M(a,b,u_3,c) > 0 \quad \text{for } u_3 \text{ near } b.$$

The second inequality follows from (2.19) and (3.18). The first inequality can be deduced from formula (2.22):

$$M(a,b,c,c) = \frac{\pi U(a,b,c)}{128[(b-c)(a-c)]^{\frac{5}{2}}} < 0 \quad \text{for } c < \xi(a,b).$$

Therefore, $M(a,b,u_3,c)$ has a zero in the interval $c < u_3 < b$. The uniqueness of the zero follows from (3.20) in that $M(a,b,u_3,c)$ increases or changes from decreasing to increasing as u_3 increases. This zero is exactly u^{**}, and the rest of the theorem can be proved easily. □

Having solved (3.12) for u_4 as a decreasing function of u_3 for $c < u_4 < \xi(a,b) < u_3 < b$, we turn to equations (3.4). Because of (3.11) and (3.17), the third equation of (3.4) gives u_3 as an increasing function of x/t for $\mu_3(a,b,\xi(a,b),\xi(a,b)) < x/t < \mu_3(a,b,u^{**},c)$. Consequently, u_4 is a decreasing function of x/t in the same interval.

LEMMA 3.5. *The last two equations of* (3.4) *can be inverted to give* u_3 *and* u_4 *as increasing and decreasing functions, respectively, of the self-similarity variable* x/t *in the interval* $\mu_3(a,b,\xi(a,b),\xi(a,b)) < x/t < \mu_3(a,b,u^{**},c)$, *where* u^{**} *is given in Lemma* 3.4.

We now turn to equations (3.5). We want to solve the third equation when $x/t > \mu_3(a,b,u^{**},c)$ or, equivalently, when $u_3 > u^{**}$. According to Lemma 3.4, $\mu_3(a,b,u_3,c) - \mu_4(a,b,u_3,c) > 0$ for $u^{**} < u_3 < b$. In view of (3.16), $\partial_{u_3} q(a,b,u_3,c) = (a+b+3u_3+c)/4$ is positive at $u_3 = u^{**}$, and, hence, it remains positive for $u_3 > u^{**}$. By (2.15), we have

$$\frac{\partial \mu_3(a,b,u_3,c)}{\partial u_3} > 0.$$

Hence, the third equation of (3.5) can be solved for u_3 as an increasing function of x/t as long as $u^{**} < u_3 < b$. When u_3 reaches b, we have $x/t = \mu_3(a,b,b,c)$. We have therefore proved the following result.

LEMMA 3.6. *The third equation of* (3.5) *can be inverted to give* u_3 *as an increasing function of* x/t *in the interval* $\mu_3(a,b,u^{**},c) \le x/t \le \mu_3(a,b,b,c)$.

We are ready to conclude the proof of Theorem 3.1. The solutions (3.3) and (3.6) are obvious. According to Lemma 3.5, the last two equations of (3.4) determine u_3 and u_4 as functions of x/t in the region $\mu_3(a,b,\xi(a,b),\xi(a,b)) \le x/t \le \mu_3(a,b,u^{**},c)$. By the first part of Lemma 3.2, the resulting u_1, u_2, u_3, and u_4 satisfy the Whitham equations (1.22) with $g = 1$. Furthermore, the boundary conditions (3.7) and (3.8) are satisfied at the trailing edge $x = \mu_3(a,b,\xi(a,b),\xi(a,b))$.

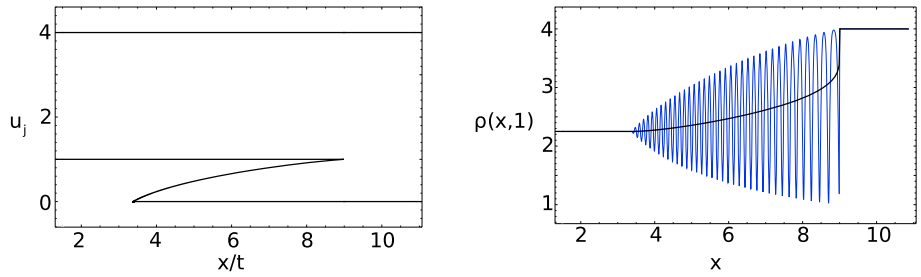


FIG. 3.2. *Self-similar solution of the Whitham equations* (1.22) *with* $g = 1$ *and the corresponding oscillatory solution* (1.25) *of the mKdV equation with* $\epsilon = 0.07$. *The initial data are given by* (1.23) *with* $a = 4$, $b = 1$, *and* $c = 0$ *of type* I.

Similarly, by Lemma 3.6, the third equation of (3.5) determines u_3 as a function of x/t in the region $\mu_3(a, b, u^{**}, c) \le x/t \le \mu_3(a, b, b, c)$. It then follows from the second part of Lemma 3.2 that u_1, u_2, u_3, and u_4 of (3.5) satisfy the Whitham equations (1.22) for $g = 1$. They also satisfy the boundary conditions (3.9) and (3.10) at the leading edge $x/t = \mu_3(a, b, b, c)$. We have therefore completed the proof of Theorem 3.1.

A graph of the Whitham solution (u_1, u_2, u_3, u_4) is given in Figure 1.2. It is obtained by plotting the exact solutions of (3.4) and (3.5).

3.2. Type I. Here we consider the initial function (1.23) satisfying $\xi(a, b) \le c$ with $b > c$ and $a + 5b > 0$.

We will present our proofs only briefly, since they are, more or less, similar to those in section 3.1. The main feature of this case is that the ξ-zero point does not appear in the solution u_3, and the Whitham equations (1.22) with $g = 1$ are strictly hyperbolic on the solution.

THEOREM 3.7 (see Figure 3.2). *For the step-like initial data* (1.23) *with* $a > b > c$, $a + 5b > 0$, *and* $\xi(a, b) \le c$, *the solution of the Whitham equations* (1.22) *with* $g = 1$ *is given by*

$$u_1 = a, \quad u_2 = b, \quad \frac{x}{t} = \mu_3(a, b, u_3, c), \quad u_4 = c$$

for $\mu_3(a, b, c, c) < x/t < \mu_3(a, b, b, c)$. *Outside this interval, the solution of* (1.21) *is given by*

$$\alpha = a, \quad \beta = b \qquad \textit{for} \quad \frac{x}{t} \le \mu_3(a, b, c, c)$$

and

$$\alpha = a, \quad \beta = c \qquad \textit{for} \quad \frac{x}{t} \ge \mu_3(a, b, b, c).$$

Proof. It suffices to show that $\mu_3(a, b, u_3, c)$ is an increasing function of u_3 for $c < u_3 < b$. Substituting (2.13) for q into (2.20) yields

$$\frac{\partial M(a, b, u_3, c)}{\partial u_3} = \frac{1}{4}[a + b + u_3 + 3c] \, \frac{\partial^3 I}{\partial u_3^2 \partial u_4} \ge \frac{1}{4}[a + b + 4\xi(a, b)] \, \frac{\partial^3 I}{\partial u_3^2 \partial u_4} > 0$$

for $c < u_3 < b$, where we have used $\xi(a, b) \le c$ in the first inequality and (3.2) in the second. We now use formula (2.22) to calculate the value of $M(a, b, u_3, c)$ at

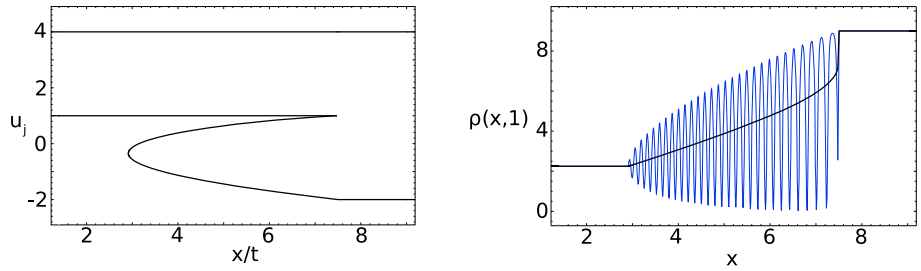


FIG. 3.3. *Self-similar solution of the Whitham equations* (1.22) *with* $g = 1$ *and the corresponding oscillatory solution* (1.25) *of the mKdV equation with* $\epsilon = 0.1$. *The initial data are given by* (1.23) *with* $a = 4$, $b = 1$, $c = -2$ *of type* III.

$u_3 = c$:

$$M(a,b,c,c) = \frac{\pi U(a,b,c)}{128[(b-c)(a-c)]^{\frac{5}{2}}} \geq 0 \quad \text{for } \xi(a,b) \leq c$$

because $U(a,b,\xi) \geq 0$ for $\xi \geq \xi(a,b)$. Therefore, $M(a,b,u_3,c) > 0$ for $c < u_3 < b$. It then follows from (2.19) that $\mu_3(a,b,u_3,c) - \mu_4(a,b,u_3,c) > 0$. Since $\frac{\partial q}{\partial u_3}(a,b,u_3,c) = (a+b+3u_3+c)/4 > (a+b+3\xi(a,b))/4 > 0$ because of (3.2), we conclude from Lemma 2.2 that

$$\frac{d\mu_3(a,b,u_3,c)}{du_3} > 0$$

for $c < u_3 < b$. □

3.3. Type III. Here we consider the step-like initial function (1.23) satisfying $\xi(a,b) > c$ with $a + 5b = 3(b-c) > 0$.

THEOREM 3.8 (see Figure 3.3). *For the step-like initial data* (1.23) *with* $a > b > c$, $\xi(a,b) > c$, *and* $a + 5b = 3(b-c)$, *the solution of the* $g = 1$ *Whitham equations* (1.22) *with* $g = 1$ *is given by*

$$u_1 = a, \quad u_2 = b, \quad \frac{x}{t} = \mu_3(a,b,u_3,u_4), \quad \frac{x}{t} = \mu_4(a,b,u_3,u_4)$$

for $\mu_3(a,b,\xi(a,b),\xi(a,b)) < x/t < \mu_3(a,b,b,c)$. *Outside the region, the solution of* (1.21) *is given by*

$$\alpha = a, \quad \beta = b \quad \textit{for} \quad \frac{x}{t} \leq \mu_3(a,b,\xi(a,b),\xi(a,b))$$

and

$$\alpha = a, \quad \beta = c \quad \textit{for} \quad \frac{x}{t} \geq \mu_3(a,b,b,c).$$

Proof. It suffices to show that u_3 and u_4 of $\mu_2(a,b,u_3,u_4) - \mu_3(a,b,u_3,u_4) = 0$ reaches b and c, respectively, simultaneously. To see this, we deduce from calculation (3.18) that

$$\mu_3(a,b,b,u_4) - \mu_4(a,b,b,u_4) = \frac{1}{2}(b-u_4)(a+2b+3u_4) \tag{3.21}$$

vanishes at $u_4 = (-a-2b)/3 = c$. □

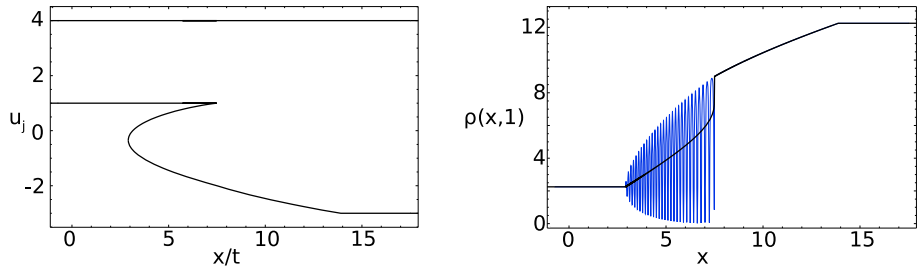


FIG. 3.4. *Self-similar solution of the Whitham equations* (1.22) *with* $g = 1$ *and the corresponding oscillatory solution* (1.25) *of the mKdV equation with* $\epsilon = 0.1$. *The initial data are given by* (1.23) *with* $a = 4$, $b = 1$, *and* $c = -3$ *of type* IV. *The solution in the region* $7.5 < x/t < 111/8$ *represents a rarefaction wave. The weak limit* $\overline{\rho(x,t)}$ *is not* C^1 *smooth at* $x/t = 111/8$.

3.4. Type IV. Here we consider the step-like initial function (1.23) satisfying $\xi(a,b) > c$ with $0 < a + 5b < 3(b-c)$.

THEOREM 3.9 (see Figure 3.4). *For the step-like initial data* (1.23) *with* $a > b > c$, $\xi(a,b) > c$, *and* $0 < a+5b < 3(b-c)$, *the solution of the Whitham equations* (1.18) *is given by*

$$u_1 = a, \quad u_2 = b, \quad \frac{x}{t} = \mu_2(a,b,u_3,u_4), \quad x = \mu_3(a,b,u_3,u_4)\ t$$

for $\mu_3(a,b,\xi(a,b),\xi(a,b)) < x/t < \mu_3(a,b,b,-(a+2b)/3)$. *Outside the region, the solution of* (1.21) *is divided into the three regions:*

(1) *For* $x/t \le \mu_3(a,b,\xi(a,b),\xi(a,b))$,

$$\alpha = a, \quad \beta = b.$$

(2) *For* $\mu_3(a,b,b,-(a+2b)/3) \le \frac{x}{t} \le \frac{3}{8}\left(a^2+2ac+5c^2\right)$,

$$\alpha = a, \quad \beta = -\frac{1}{5}a - \sqrt{\frac{8}{15}\frac{x}{t} - \frac{4}{25}a^2}.$$

(3) *For* $x/t \ge \frac{3}{8}\left(a^2+2ac+5c^2\right)$,

$$\alpha = a, \quad \beta = c.$$

Proof. By calculation (3.21), when u_3 of $\mu_3(a,b,u_3,u_4) - \mu_4(a,b,u_3,u_4) = 0$ touches b, the corresponding u_4 reaches $-(a+2b)/3$, which is above c. Hence, the equations

$$\frac{x}{t} = \mu_3(a,b,u_3,u_4), \quad \frac{x}{t} = \mu_4(a,b,u_3,u_4)$$

can be inverted to give u_3 and u_4 as functions of x/t in the region $\mu_3(a,b,\xi(a,b),\xi(a,b)) < x/t < \mu_3(a,b,b,-(a+2b)/3)$. In region (2), (1.21) has a rarefaction wave solution. ∎

4. More self-similar solutions. In this section, we construct self-similar solutions of the $g = 1$ Whitham equations (1.22) for the initial function (1.23) with $a > c > b$. The solution of (1.21) does not develop a shock for $a + 5b \ge 0$. We are therefore interested only in the case $a + 5b < 0$. We classify the resulting Whitham solution into four types:

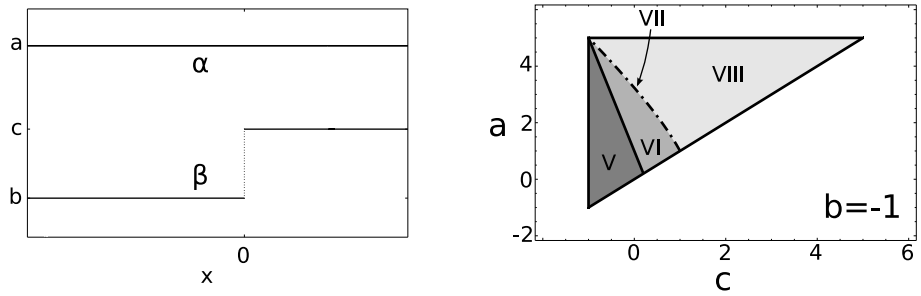


FIG. 4.1. *Graph of the step-like initial shock data* (1.23) *with* $a > c > b$ *and* $a + 5b < 0$, *and diagram of regions for types* V–VIII, *where* $b < 0$ *has been scaled to* $b = -1$.

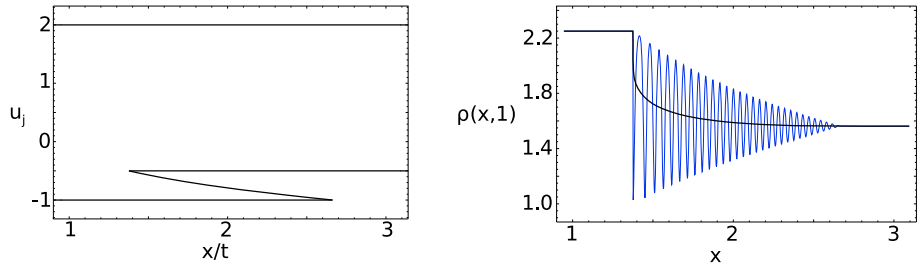


FIG. 4.2. *Self-similar solution of the Whitham equations* (1.22) *with* $g = 1$ *and the corresponding oscillatory solution* (1.25) *of the mKdV equation with* $\epsilon = 0.014$. *The initial data are given by* (1.23) *with* $a = 2$, $b = -1$, *and* $c = -1/2$ *of type* V.

V. $a + 5b \le -4(c - b) < 0$,
VI. $0 > a + 5b > -4(c - b)$ with $V(a, c, b) < 0$,
VII. $0 > a + 5b > -4(c - b)$ with $V(a, c, b) = 0$,
VIII. $0 > a + 5b > -4(c - b)$ with $V(a, c, b) > 0$,

where V is a quadratic polynomial given by (2.27). In Figure 4.1, we illustrate this classification by scaling $b < 0$ to $b = -1$ and plotting regions for types V–VIII in the a-c plane.

4.1. Type V. Here we consider the step-like initial function (1.23) satisfying $a + 5b \le -4(c - b) < 0$.

THEOREM 4.1 (see Figure 4.2). *For the step-like initial data* (1.23) *with* $a > c > b$, $a+5b \le -4(c-b)$, *the solution of the Whitham equations* (1.22) *with* $g = 1$ *is given by*

$$u_1 = a, \quad u_2 = c, \quad \frac{x}{t} = \mu_3(a, c, u_3, b), \quad u_4 = b$$

for $\mu_3(a, c, c, b) < x/t < \mu_3(a, c, b, b)$. *Outside this interval, the solution of* (1.21) *is given by*

$$\alpha = a, \quad \beta = b \quad \textit{for} \quad \frac{x}{t} \le \mu_3(a, c, c, b)$$

and

$$\alpha = a, \quad \beta = c \quad \textit{for} \quad \frac{x}{t} \ge \mu_3(a, c, b, b).$$

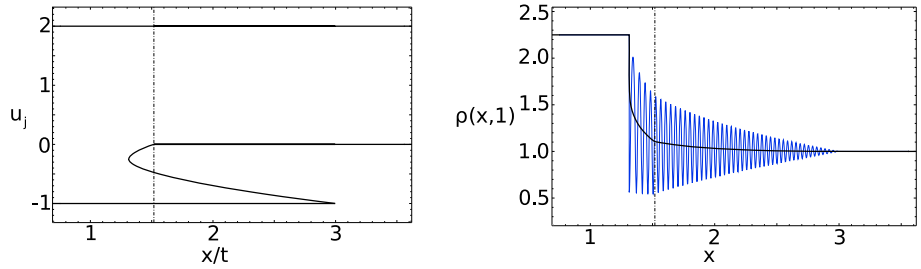


FIG. 4.3. *Self-similar solution of the Whitham equations* (1.22) *with* $g = 1$ *and the corresponding periodic oscillatory solution* (1.25) *of the mKdV equation with* $\epsilon = 0.018$. *The initial data are given by* (1.23) *with* $a = 2$, $b = -1$, *and* $c = 0$ *of type* VI. *The oscillations have two different kinds of structure, which are separated by* $x/t \approx 1.51$. *The weak limit* $\overline{\rho(x,t)}$ *and the envelope of the oscillations have noticeable corners at* $x/t \approx 1.51$.

Proof. It suffices to show that $\mu_3(a, c, u_3, b)$ is a decreasing function of u_3 for $b < u_3 < c$. By (2.10), we have

$$\frac{\partial \mu_3(a,c,u_3,b)}{\partial u_3} = \frac{1}{2}\frac{\partial \lambda_3}{\partial u_3}\frac{\partial q}{\partial u_3} + \frac{1}{2}[\lambda_3 - 2(a+c+u_3+b)]\frac{\partial^2 q}{\partial u_3^2}.$$

The second term is negative because of (2.6) and $\frac{\partial^2 q}{\partial u_2^2} = 3/8 > 0$. The first term is also negative: Its first factor is positive in view of (1.17), while its second factor is

$$\frac{\partial q}{\partial u_3} = \frac{1}{4}(a + c + 3u_3 + b) < 0$$

for $b < u_3 < c$, as we have that $a + b + 4c \le 0$. ∎

4.2. Type VI. Here we consider the step-like initial function (1.23) satisfying $0 > a + 5b > -4(c - b)$ with $V(a, c, b) < 0$.

THEOREM 4.2 (see Figure 4.3). *For the step-like initial data* (1.23) *with* $0 > a + 5b > -4(c - b)$ *and* $V(a, c, b) < 0$, *the solution of the Whitham equations* (1.22) *with* $g = 1$ *is given by*

$$u_1 = a, \quad \frac{x}{t} = \mu_2(a, u_2, u_3, b), \quad \frac{x}{t} = \mu_3(a, u_2, u_3, b), \quad u_4 = b \tag{4.1}$$

for $\mu_3(a, -(a+b)/4, -(a+b)/4, b) < x/t \le \mu_3(a, u^{***}, u^{***}, b)$ *and by*

$$u_1 = a, \quad u_2 = c, \quad \frac{x}{t} = \mu_3(a, c, u_3, b), \quad u_4 = b \tag{4.2}$$

for $\mu_3(a, u^{***}, u^{***}, b) \le x/t < \mu_3(a, c, b, b)$, *where* u^{***} *is the unique solution* u_3 *of* $\mu_2(a, c, u_3, b) = \mu_3(a, c, u_3, b)$ *in the interval* $b < u_3 < c$. *Outside the region* $\mu_3(a, -(a+b)/4, -(a+b)/4, b) < x/t < \mu_3(a, c, b, b)$, *the solution of* (1.21) *is given by*

$$\alpha = a, \quad \beta = b \quad \textit{for} \quad \frac{x}{t} \le \mu_3(a, -(a+b)/4, -(a+b)/4, b)$$

and

$$\alpha = a, \quad \beta = c \quad \textit{for} \quad \frac{x}{t} \ge \mu_3(a, c, b, b).$$

Proof. We first locate the "leading" edge, i.e., the solution of (4.1) at $u_2 = u_3$. Eliminating x/t from the first two equations of (4.1) yields

$$\mu_2(a, u_2, u_3, b) - \mu_3(a, u_2, u_3, b) = 0. \tag{4.3}$$

Since it degenerates at $u_2 = u_3$, we replace (4.3) by

$$G(a, u_2, u_3, b) := \frac{\mu_2(a, u_2, u_3, b) - \mu_3(a, u_2, u_3, b)}{(u_2 - u_3)\sqrt{(u_1 - u_3)(u_2 - u_4)}I(a, u_2, u_3, b)} = 0. \tag{4.4}$$

In Appendix A, we show that, at the "leading" edge $u_2 = u_3$, we have

$$G(a, u_3, u_3, b) = 2\left(\frac{\partial q}{\partial u_2} + \frac{\partial q}{\partial u_3}\right) = 0$$

in view of (A.6), which along with (2.13) gives $u_2 = u_3 = -(a+b)/4$. Having located the "leading" edge, we solve (4.4) near $u_2 = u_3 = -(a+b)/4$. We use formula (A.8) to obtain

$$\frac{\partial G(a, -(a+b)/4, -(a+b)/4, b)}{\partial u_2} = \frac{\partial G(a, -(a+b)/4, -(a+b)/4, b)}{\partial u_3} = 2.$$

These show that (4.4) gives u_2 as a decreasing function of u_3,

$$u_2 = B(u_3), \tag{4.5}$$

in a neighborhood of $u_2 = u_3 = -(a+b)/4$.

We now extend the solution (4.5) of (4.3) as far as possible in the region $b < u_3 < -(a+b)/4 < u_2 < c$. We use formula (2.10) to obtain

$$\frac{\partial \mu_2}{\partial u_2} = \frac{1}{2}\frac{\partial \lambda_2}{\partial u_2}\frac{\partial q}{\partial u_2} + \frac{1}{2}[\lambda_2 - 2(a + u_2 + u_3 + b)]\frac{\partial^2 q}{\partial u_2^2},$$

$$\frac{\partial \mu_3}{\partial u_3} = \frac{1}{2}\frac{\partial \lambda_3}{\partial u_3}\frac{\partial q}{\partial u_3} + \frac{1}{2}[\lambda_3 - 2(a + u_2 + u_3 + b)]\frac{\partial^2 q}{\partial u_3^2}.$$

In view of (1.17) and (2.6), we have

$$\frac{\partial \mu_2}{\partial u_2} > 0 \quad \text{if} \quad \frac{\partial q}{\partial u_2} > 0,$$

$$\frac{\partial \mu_3}{\partial u_3} < 0 \quad \text{if} \quad \frac{\partial q}{\partial u_3} < 0.$$

We claim that

$$\frac{\partial q}{\partial u_2} > 0, \quad \frac{\partial q}{\partial u_3} < 0 \tag{4.6}$$

on the solution of (4.3) in the region $b < u_3 < -(a+b)/4 < u_2 < c$. To see this, we use formula (2.10) to rewrite (4.3) as

$$\frac{1}{2}[\lambda_2 - 2(a + u_2 + u_3 + b)]\frac{\partial q}{\partial u_2} = \frac{1}{2}[\lambda_3 - 2(a + u_2 + u_3 + b)]\frac{\partial q}{\partial u_3}.$$

This, together with

$$\frac{\partial q}{\partial u_2} - \frac{\partial q}{\partial u_3} = 2(u_2 - u_3)\frac{\partial^2 q}{\partial u_2 \partial u_3} = \frac{1}{2}(u_2 - u_3) > 0$$

for $u_2 > u_3$ and inequalities (2.6), proves (4.6).

Hence, the solution (4.5) can be extended as long as $b < u_3 < -(a+b)/4 < u_2 < c$. There are two possibilities: (1) u_2 touches c before u_3 reaches b, and (2) u_3 touches b before or simultaneously as u_2 reaches c.

Possibility (2) is unattainable. To see this, we use (2.26) to write

$$\mu_2(a, u_2, b, b) - \mu_3(a, u_2, b, b) = \frac{(u_2 - b)}{2(a + u_2 - 2b)}\, V(a, u_2, b), \tag{4.7}$$

which is negative for $b < u_2 \le c$ since $V(a, u_2, b)$ of (2.27) is an increasing function of u_2 and since $V(a, c, b) < 0$. Therefore, u_2 will touch c before u_3 reaches b. When this happens, we have

$$\mu_2(a, c, u_3, b) - \mu_3(a, c, u_3, b) = 0. \tag{4.8}$$

LEMMA 4.3. *Equation* (4.8) *has a simple zero, counting multiplicities, in the interval* $b < u_3 < c$. *If we denote this zero by* u^{***}, *then* $\mu_2(a, c, u_3, b) - \mu_3(a, c, u_3, b)$ *is positive for* $u_3 > u^{***}$ *and negative for* $u_3 < u^{***}$.

The proof, which involves formulae (2.24) and (2.25), is rather similar to the proof of Lemma 3.19. We will omit it.

We now continue to prove Theorem 4.2. Having solved (4.3) for u_2 as a decreasing function of u_3 for $u^{***} < u_3 < -(a+b)/4$, we can then use the middle two equations of (4.1) to determine u_2 and u_3 as functions of x/t in the interval $\mu_2(a, -(a+b)/4, -(a+b)/4, b) < x/t < \mu_2(a, c, u^{***}, b)$.

We finally turn to equations (4.2). We want to solve the third equation of (4.2), $x/t = \mu_3(a, c, u_3, b)$, for $u_3 < u^{***}$. It is enough to show that $\mu_3(a, c, u_3, b)$ is a decreasing function of u_3 for $u_3 < u^{***}$. According to Lemma 4.3, $\mu_2(a, c, u_3, b) - \mu_3(a, c, u_3, b) < 0$ for $u_3 < u^{***}$. Using formula (2.10) for μ_2 and μ_3, we have

$$\frac{1}{2}[\lambda_2 - 2(a + c + u_3 + b)]\frac{\partial q}{\partial u_2} < \frac{1}{2}[\lambda_3 - 2(a + c + u_3 + b)]\frac{\partial q}{\partial u_3}.$$

This, together with

$$\frac{\partial q}{\partial u_2} - \frac{\partial q}{\partial u_3} = \frac{1}{2}(c - u_3) > 0$$

for $u_3 < c$ and inequalities (2.6), proves that

$$\frac{\partial q(a, c, u_3, b)}{\partial u_3} < 0$$

for $u_3 < u^{***}$. Hence,

$$\frac{\partial \mu_3}{\partial u_3} = \frac{1}{2}\frac{\partial \lambda_3}{\partial u_3}\frac{\partial q}{\partial u_3} + \frac{1}{2}[\lambda_3 - 2(a + c + u_3 + b)]\frac{\partial^2 q}{\partial u_3^2} < 0,$$

where we have used inequality (1.17). □

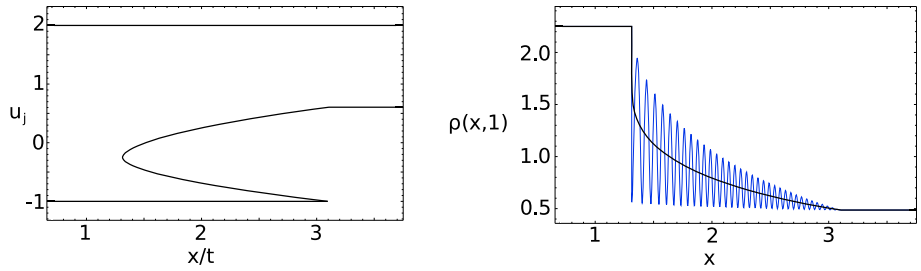


FIG. 4.4. *Self-similar solution of the Whitham equations* (1.22) *with* $g = 1$ *and the corresponding oscillatory solution* (1.25) *of the mKdV equation with* $\epsilon = 0.03$. *The initial data are given by* (1.23) *with* $a = 2$, $b = -1$, *and* $c = -3 + \sqrt{13}$ *of type* VII.

4.3. Type VII. Here we consider the step-like initial function (1.23) satisfying $0 > a + 5b > -4(c - b)$ with $V(a, c, b) = 0$.

THEOREM 4.4 (see Figure 4.4). *For the step-like initial data* (1.23) *with* $0 > a + 5b > -4(c - b)$ *and* $V(a, c, b) = 0$, *the solution of the Whitham equations* (1.22) *with* $g = 1$ *is given by*

$$u_1 = a, \quad \frac{x}{t} = \mu_2(a, u_2, u_3, b), \quad \frac{x}{t} = \mu_3(a, u_2, u_3, b), \quad u_4 = c$$

for $\mu_3(a, -(a + b)/4, -(a + b)/4, b) < x/t < \mu_3(a, c, b, b)$. *Outside the region, the solution of* (1.21) *is given by*

$$\alpha = a, \quad \beta = b \qquad \textit{for} \quad \frac{x}{t} \le \mu_3(a, -(a + b)/4, -(a + b)/4, b)$$

and

$$\alpha = a, \quad \beta = c \qquad \textit{for} \quad \frac{x}{t} \ge \mu_3(a, c, b, b).$$

Proof. It suffices to show that u_2 and u_3 of $\mu_2(a, u_2, u_3, b) - \mu_3(a, u_2, u_3, b) = 0$ reaches c and b, respectively, simultaneously. To see this, we deduce from (4.7) that

$$\mu_2(a, c, b, b) - \mu_3(a, c, b, b) = \frac{(c - b)}{2(a + c - 2b)} \, V(a, c, b) \tag{4.9}$$

vanishes when $V(a, c, b) = 0$. ∎

4.4. Type VIII. Here we consider the step-like initial function (1.23) satisfying $0 > a + 5b > -4(c - b)$ with $V(a, c, b) > 0$.

THEOREM 4.5 (see Figure 4.5). *For the step-like initial data* (1.23) *with* $0 > a + 5b > -4(c - b)$ *and* $V(a, c, b) > 0$, *the solution of the Whitham equations* (1.22) *with* $g = 1$ *is given by*

$$u_1 = a, \quad \frac{x}{t} = \mu_2(a, u_2, u_3, b), \quad \frac{x}{t} = \mu_3(a, u_2, u_3, b), \quad u_4 = b$$

for $\mu_3(a, -(a+b)/4, -(a+b)/4, b) < x/t < \mu_3(a, \hat{u}, \hat{u}, b)$, *where* $\hat{u}$ *is the unique* u_2*-zero of the quadratic polynomial* $V(a, u_2, b)$ *in the interval* $-(a + b)/4 < u_2 < c$. *Outside the region, the solution of* (1.21) *is divided into the following three regions:*

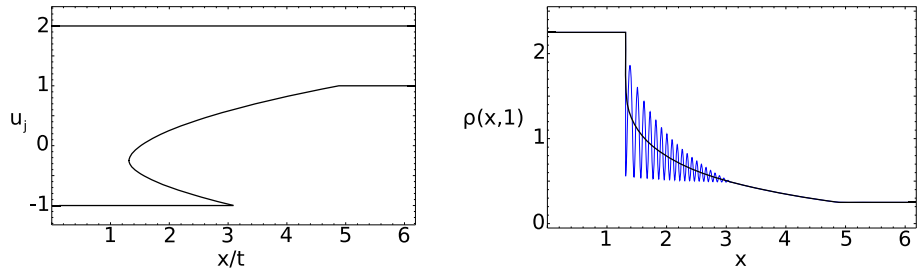


FIG. 4.5. *Self-similar solution of the Whitham equations and the corresponding oscillatory solution* (1.25) *of the mKdV equation with* $\epsilon = 0.05$. *The initial data are given by* (1.23) *with* $a = 2$, $b = -1$, *and* $c = 1$ *of type* VIII. *The solution in the region* $3.10 < x/t < 39/8$ *represents a rarefaction wave. The weak limit* $\overline{\rho(x,t)}$ *is not* C^1 *smooth at* $x/t = 39/8$.

(1) *For* $x/t \le \mu_3(a, -(a+b)/4, -(a+b)/4, b)$,

$$\alpha = a, \quad \beta = b.$$

(2) *For* $\mu_2(a, \hat{u}, b, b) \le x/t \le \frac{3}{8}\left(a^2 + 2ac + 5c^2\right)$,

$$\alpha = a, \quad \beta = -\frac{1}{5}a + \sqrt{\frac{8}{15}\frac{x}{t} - \frac{4}{25}a^2}.$$

(3) *For* $x/t \ge \frac{3}{8}\left(a^2 + 2ac + 5c^2\right)$,

$$\alpha = a, \quad \beta = c.$$

Proof. By the calculation (4.9), when u_3 of $\mu_2(a, u_2, u_3, b) - \mu_3(a, u_2, u_3, b) = 0$ touches b, the corresponding u_2 reaches $\hat{u}$, where $V(a, \hat{u}, b) = 0$. Obviously, $\hat{u} < c$. Hence, equations

$$\frac{x}{t} = \mu_2(a, u_2, u_3, b), \qquad \frac{x}{t} = \mu_3(a, u_2, u_3, b)$$

can be inverted to give u_2 and u_3 as functions of x/t in the region $\mu_2(a, -(a + b)/4, -(a+b)/4, b) < x/t < \mu_2(a, \hat{u}, \hat{u}, b)$. To the right of this region, equations (1.21) have a rarefaction wave solution. ∎

5. Vacuum points. The nonnegative function $\rho(x, t; \epsilon)$ has to be positive in order for the mKdV dispersive approximation (1.19) (or (1.2) for NLS) to make sense. It is therefore interesting to study those points (x, t) at which $\rho(x, t; \epsilon) = 0$ for a sequence of vanishing ϵ. These points are referred to as the vacuum points in [3].

Since the solution $\rho(x, t; \epsilon)$ of (1.19) can be approximated by the periodic solution (1.25) in the single phase regime, we instead study the vacuum points of the latter solution.

Since $\rho_2 \ge \rho_3$ because of (1.9) for $u_1 \ge u_2 \ge u_3 \ge u_4$, it follows from formula (1.25) that (x, t) is a vacuum point if and only if $\rho_3(x, t) = 0$, which, in view of (1.9), is equivalent to

$$u_1(x,t) - u_2(x,t) - u_3(x,t) + u_4(x,t) = 0. \tag{5.1}$$

For types I–IV, equality (5.1) can occur only if $u_1 - u_2 \le u_3 - u_4$ at the leading edge of the Whitham solution (see Figures 1.2, 3.2, 3.3, and 3.4). This inequality leads to $a + b \le 2b$ for types I and II and gives $a \le 4b$ for types III and IV. Hence,

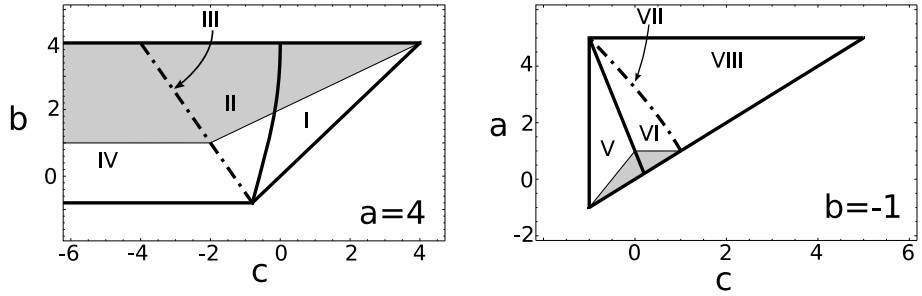


FIG. 5.1. *Diagram of regions for the initial data* (1.23) *which generate oscillatory solutions with vacuum points. The parameters* (c, b) *or* (c, a) *form the shaded region.*

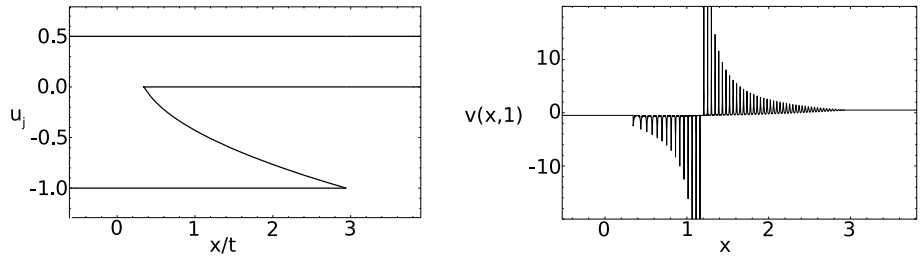


FIG. 5.2. *Self-similar solution of the Whitham equations* (1.22) *with* $g = 1$ *and the flow velocity* (1.10) *of the oscillatory solution of the mKdV equation with* $\epsilon = 0.014$. *The initial data are given by* (1.24) *of type* V *with parameters* $a = 0.5$, $b = -1$, *and* $c = 0$. *A vacuum point is located at* $x/t \approx 1.16$.

step-like initial shock data of types I–IV with $4b \geq a > b > c$ and $a + c \leq 2b$ will generate mKdV oscillatory solutions with vacuum points (cf. Figure 5.1).

Similarly, step-like initial shock data of types V and VI with $-b \geq a > c > b$ and $a + b \leq 2c$ will produce mKdV oscillatory solutions with vacuum points. Types VII and VIII oscillatory solutions cannot have any vacuum point (cf. Figure 5.1).

Types I–V oscillatory solutions have at most one vacuum point. To see this, note that the corresponding Whitham solutions have the property that $u_1(x, t)$ and $u_2(x, t)$ are constants and that $u_3(x, t) - u_4(x, t)$ are strictly monotone functions of x/t in the single phase region; so is the left-hand side of (5.1). Hence, (5.1) has at most one solution x/t. Figure 5.2 shows an example of oscillatory solutions with a unique vacuum point. Notice that the velocity $v(x, t; \epsilon)$ has a huge change at this point.

Type VI oscillatory solutions can have more than one vacuum point. They can even have a continuum of vacuum points. Namely, there exists an interval such that (5.1) holds at each point of the interval.

THEOREM 5.1. *For step-like initial data* (1.24) *of type* VI *satisfying condition* $a+b = 0$, *all the points on the closed interval* $\mu_3(a, -(a+b)/4, -(a+b)/4, b) \leq x/t \leq \mu_3(a, u^{***}, u^{***}, b)$, *where* u^{***} *is given in Theorem* 4.2, *are vacuum points.*

Proof. According to Theorem 4.2, the Whitham solution has the property that $u_1 = a$ and $u_4 = b$. Since $a + b = 0$, it follows from (5.1) that it suffices to show that $u_2 = -u_3$ over the closed interval mentioned in Theorem 5.1.

The functions u_2 and u_3 of the Whitham solution satisfies (4.3), which has a unique solution (4.5) with the property that $B(0) = 0$ in the case of $a + b = 0$. It

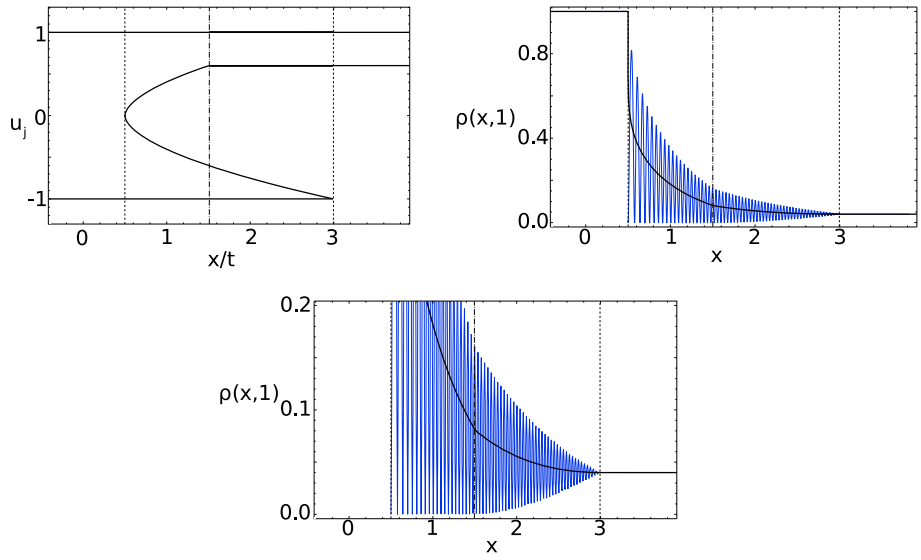


FIG. 5.3. *Self-similar solution of the Whitham equations* (1.22) *with* $g = 1$ *and the oscillatory solutions* (1.25) *of the mKdV equation with* $\epsilon = 0.04$. *The initial data are given by* (1.23) *with* $a = 1$, $b = -1$, *and* $c = 0.6$ *of type* VI. *The vacuum points occupy the whole interval* $0.5 \le \frac{x}{t} \le 1.50$.

follows from formula (2.10) for μ_2 and μ_3, formula (2.13) for q, and formulae (A.2–A.3) for λ_2 and λ_3 that $\mu_2(a, -u_3, u_3, b) = \mu_3(a, -u_3, u_3, b)$. Hence, $u_2 = -u_3$ is the unique solution of (4.3). □

Figure 5.3 is an illustration of such a solution, where rather than a vacuum point we now have a vacuum interval. It follows from formula (1.10) that $V(x, t; \epsilon) = (u_1 + u_2 + u_3 + u_4)/4 = 0$ at each point of the interval.

We close this section with a remark on the NLS case, which has been used to describe the nonlinear pulse propagation in an optical fiber [7]. The NLS oscillations with step-like initial shock data (1.15), $b > c$, have at most one vacuum point [3]. For more complicated step-like initial shock data, the NLS oscillations can also have a continuum of vacuum points. Indeed, consider the initial data

$$\alpha(x, 0) = \begin{cases} a, & x < 0, \\ b, & x > 0, \end{cases} \qquad \beta(x, 0) = \begin{cases} c, & x < 0, \\ d, & x > 0, \end{cases}$$

with $a > c > b > d$ and $a - b = c - d$. It can easily be shown that there is an interval within the single phase regime where all the u_i of the Whitham solution of (1.7) with $g = 1$ are constants. More precisely, $u_1 = a$, $u_2 = c$, $u_3 = b$, and $u_4 = d$ for $\lambda_3(a, c, b, d) \le x/t \le \lambda_2(a, c, b, d)$, where λ_2 and λ_3 are the second and third eigenspeeds of the NLS–Whitham equations (1.7) for $g = 1$. Since $a - b = c - d$, (5.1) holds at each point of the interval; hence, this is a vacuum interval. From the point of view of nonlinear optics, the optical wave has trivial frequency chirp in this region, and the region may be considered as a vacuum.

6. Other initial data. We conclude the paper by showing how to handle the initial data (1.24). Inequalities (2.14) are replaced by

$$\frac{\partial \lambda_2}{\partial u_2} < \frac{3}{2} \frac{\lambda_1 - \lambda_2}{u_1 - u_2} < \frac{\partial \lambda_1}{\partial u_1}$$

for $u_4 < u_3 < u_2 < u_1$. Results similar to those of Lemma 2.2 can then be easily proved. The rest of the calculations are similar to those in sections 3 and 4.

Appendix A. Leading edge calculations. The function $I(u_1, u_2, u_3, u_4)$ of (2.3) can be written in terms of the complete elliptic integral of the first kind $K(s)$; i.e.,

$$I = \frac{K(s)}{\sqrt{(u_1 - u_3)(u_2 - u_4)}}, \tag{A.1}$$

where

$$s = \frac{(u_1 - u_2)(u_3 - u_4)}{(u_1 - u_3)(u_2 - u_4)}.$$

Using the derivative formula

$$\frac{dK(s)}{ds} = \frac{E(s) - (1-s)K(s)}{2s(1-s)},$$

where $E(s)$ is the complete elliptic integral of the second kind, we calculate λ_2 and λ_3 of (2.2):

$$\lambda_2 = 2\sigma_1 - 4\,\frac{u_1 - u_2}{1 - \frac{u_1 - u_3}{u_2 - u_3}\,\frac{E}{K}}, \tag{A.2}$$

$$\lambda_3 = 2\sigma_1 + 4\,\frac{u_3 - u_4}{1 - \frac{u_2 - u_4}{u_2 - u_3}\,\frac{E}{K}}. \tag{A.3}$$

We then use (2.10) to write $\mu_2 - \mu_3$ as

$$-2(u_2 - u_3)K(s)\left[\frac{u_1 - u_2}{(u_2 - u_3)K - (u_1 - u_3)E}\,\frac{\partial q}{\partial u_2} + \frac{u_3 - u_4}{(u_2 - u_3)K - (u_2 - u_4)E}\,\frac{\partial q}{\partial u_3}\right].$$

Hence, $G(u_1, u_2, u_3, u_4)$ of (4.4) becomes

$$G = -2\left[\frac{u_1 - u_2}{(u_2 - u_3)K - (u_1 - u_3)E}\,\frac{\partial q}{\partial u_2} + \frac{u_3 - u_4}{(u_2 - u_3)K - (u_2 - u_4)E}\,\frac{\partial q}{\partial u_3}\right]. \tag{A.4}$$

We now use the asymptotics of $K(s)$ and $E(s)$ as s is close to 1,

$$K(s) \approx \frac{1}{2}\log\frac{16}{1-s}, \quad E(s) \approx 1 + \frac{1}{4}(1-s)\left(\log\frac{16}{1-s} - 1\right), \tag{A.5}$$

to calculate the $u_2 = u_3$ limit

$$G = 2\left(\frac{\partial q}{\partial u_2} + \frac{\partial q}{\partial u_3}\right). \tag{A.6}$$

Finally, we can also use the expression (2.13) and the derivative formulae

$$\frac{dE(s)}{ds} = \frac{E(s) - K(s)}{2s}, \quad \frac{dK(s)}{ds} = \frac{E(s) - (1-s)K(s)}{2s(1-s)} \tag{A.7}$$

to evaluate the partial derivatives of G in the $u_2 = u_3$ limit

$$\frac{\partial G}{\partial u_2}\Big|_{u_2=u_3} = 2 + \frac{(u_1-u_4)(u_1+4u_3+u_4)}{4(u_1-u_3)(u_4-u_3)}, \tag{A.8}$$

$$\frac{\partial G}{\partial u_3}\Big|_{u_2=u_3} = 2 - \frac{(u_1-u_4)(u_1+4u_3+u_4)}{4(u_1-u_3)(u_4-u_3)}.$$

Appendix B. Loss of genuine nonlinearity. In this appendix, we will show that the Whitham equations (1.22) with $g = 1$ are not genuinely nonlinear. Again, we suppress the subscript $g = 1$ in the notation $\lambda_{g,i}$ and $\mu_{g,i}$.

PROPOSITION B.1. *For fixed $u_1 > u_2 > u_4$ satisfying conditions*

$$u_1 + 4u_2 + u_4 > 0, \quad U(u_1, u_2, u_4) < 0,$$

where the function U is given in (2.23)*, the derivative*

$$\frac{\partial \mu_3(u_1, u_2, u_3, u_4)}{\partial u_3}$$

changes sign as u_3 increases from u_4 to u_2.

Proof. We first use formula (A.3) for λ_3, derivative formulae (A.7), and asymptotics (A.5) to calculate

$$\frac{\partial \lambda_3}{\partial u_3}\Big|_{u_3=u_2} = +\infty, \tag{B.1}$$

$$\frac{\partial \lambda_3}{\partial u_3}\Big|_{u_3=u_4} = \frac{24(u_2-u_4)^2 + 24(u_1-u_2)(u_2-u_4) + 9(u_1-u_2)^2}{(u_1+u_2-2u_4)^2}. \tag{B.2}$$

In the calculation of the last equation, we have also used the asymptotics of $K(s)$ and $E(s)$ near $s = 0$,

$$K(s) = \frac{\pi}{2}\left[1 + \frac{s}{4} + \frac{9}{64}s^2 + \cdots + \left(\frac{1\cdot 3\cdots(2n-1)}{2\cdot 4\cdots 2n}\right)^2 s^n + \cdots\right],$$

$$E(s) = \frac{\pi}{2}\left[1 - \frac{s}{4} - \frac{3}{64}s^2 - \cdots - \frac{1}{2n-1}\left(\frac{1\cdot 3\cdots(2n-1)}{2\cdot 4\cdots 2n}\right)^2 s^n - \cdots\right].$$

Differentiating formula (2.10) for μ_3 and using formula (2.13) for q, we obtain

$$\frac{\partial \mu_3(u_1, u_2, u_3, u_4)}{\partial u_3} = \frac{1}{8}(u_1+u_2+3u_3+u_4)\,\frac{\partial \lambda_3}{\partial u_3} + \frac{3}{8}[\lambda_3 - 2\sigma_1].$$

We then use the boundary values (2.7) and (2.8) for λ_3 and boundary values (B.1) and (B.2) for $\partial_{u_3}\lambda_3$ to calculate

$$\frac{\partial \mu_3}{\partial u_3}\Big|_{u_3=u_2} = +\infty \quad \text{if } u_1 + 4u_2 + u_4 > 0,$$

and

$$\frac{\partial \mu_3}{\partial u_3}\Big|_{u_3=u_4} = \frac{3U(u_1, u_2, u_4)}{8(u_1+u_2-2u_4)} < 0 \quad \text{if } U(u_1, u_2, u_4) < 0.$$

These two boundary values for $\partial_{u_3}\mu_3$ prove the proposition. □

The conditions of the proposition can easily be verified. For example, for the initial data (1.23) of types II and III, we have $\xi(a,b) > c$ and $a + 5b \geq 3(b-c) > 0$, which imply $U(a,b,c) < 0$ and $a + 4b + c > 0$. By Proposition B.1, the derivative

$$\frac{\partial \mu_3(a,b,u_3,c)}{\partial u_3}$$

must change sign as u_3 increases from c to b.

Appendix C. Phase shift. In this appendix, we will verify that Q given by (1.12) and (1.13) is the right phase shift for both the NLS and mKdV. We will study only the NLS case; the mKdV case can be handled in the same way.

For the NLS periodic wave train (1.8), the wave number and frequency are

$$k = \frac{\pi\sqrt{\rho_1 - \rho_3}}{\epsilon K(s)}, \quad w = kV_1.$$

The rapid phase is then given by

$$\Theta(x,t;\epsilon) = k\theta(x,t) = k[x - V_1\ t - Q]. \tag{C.1}$$

To verify that Q of (1.12) and (1.13) is indeed the phase shift, it suffices to show that Θ satisfies the generalized wave number and frequency relations

$$\frac{\partial \Theta}{\partial x} = k, \quad \frac{\partial \Theta}{\partial t} = -w. \tag{C.2}$$

We first observe that the compatibility condition for (C.2) is

$$\frac{\partial k}{\partial t} + \frac{\partial w}{\partial x} = 0,$$

which is the conservation of waves. This equation can be viewed as an additional conservation law satisfied by the solutions of the Whitham equations (1.7). Hence, the eigenspeeds λ_i of the Whitham equations can be calculated using k and w; i.e.,

$$\lambda_i = \frac{\partial_{u_i} w}{\partial_{u_i} k}. \tag{C.3}$$

We can also rewrite the wave number k in terms of I of (A.1) as

$$k = \frac{\pi}{\epsilon I},$$

which along with formula (2.2) for λ_i gives

$$\frac{k}{\partial_{u_i} k} = -\frac{I}{\partial_{u_i} I} = \frac{1}{2}(\lambda_i - 2\sigma_1). \tag{C.4}$$

Differentiating (C.1) with respect to x and using the property that $u_1 = a$, we obtain

$$\begin{aligned} \frac{\partial \Theta}{\partial x} &= k + \sum_{i=2}^{4} \left[x\frac{\partial k}{\partial u_i} - t\frac{\partial w}{\partial u_i} - k\frac{\partial Q}{\partial u_i} - Q\frac{\partial k}{\partial u_i} \right] u_{ix} \\ &= k + \sum_{i=2}^{4} \frac{\partial k}{\partial u_i}\ u_{ix} \left[x - \lambda_i t - \frac{1}{2}(\lambda_i - 2\sigma_1)\frac{\partial Q}{\partial u_i} - Q \right], \end{aligned} \tag{C.5}$$

where we have used formulae (C.3) and (C.4) in the last equality.

It is known that if $Q(a, u_2, u_3, u_4)$ satisfies (1.12) and (1.13), then

$$x = \lambda_i t + \frac{1}{2}(\lambda_i - 2\sigma_1)\frac{\partial Q}{\partial u_i} + Q$$

is the hodograph solution of the Whitham equations (1.7) [14]. Hence, we deduce from (C.5) the first equation of (C.2). The second equation of (C.2) can be shown in the same way.

Acknowledgment. We thank the referees for many valuable suggestions.

REFERENCES

[1] P. Deift, S. Venakides, and X. Zhou, *New results in small dispersion KdV by an extension of the steepest descent method for Riemann-Hilbert problems*, Internat. Math. Res. Not., No. 6 (1997), pp. 285–299.

[2] B. A. Dubrovin and S. P. Novikov, *Hydrodynamics of weakly deformed soliton lattices. Differential geometry and Hamiltonian theory*, Russian Math. Surveys, 44 (1989), pp. 35–124.

[3] G. A. El, V. V. Geogjaev, A. V. Gurevich, and A. L. Krylov, *Decay of an initial discontinuity in the defocusing NLS hydrodynamics*, Phys. D, 87 (1995), pp. 186–192.

[4] M. G. Forest and J.-E. Lee, *Geometry and modulation theory for the periodic nonlinear Schrödinger equation*, in Oscillation Theory, Computation, and Methods of Compensated Compactness, J. L. Ericksen, D. Kinderlehrer, and M. Slemrod, eds., Springer-Verlag, New York, 1986, pp. 35–69.

[5] T. Grava and C. Klein, *Numerical solution of the small dispersion limit of Korteweg-de Vries and Whitham equations*, Comm. Pure Appl. Math., 60 (2007), pp. 1623–1664.

[6] S. Jin, C. D. Levermore, and D. W. McLaughlin, *The semiclassical limit of the defocusing NLS hierarchy*, Comm. Pure Appl. Math., 52 (1999), pp. 613–654.

[7] Y. Kodama, *The Whitham equations for optical communications: Mathematical theory of NRZ*, SIAM J. Appl. Math., 59 (1999), pp. 2162–2192.

[8] P. D. Lax, C. D. Levermore, and S. Venakides, *The generation and propagation of oscillations in dispersive initial value problems and their limiting behavior*, in Important Developments in Soliton Theory 1980–1990, T. Fokas and V. E. Zakharov, eds., Springer Ser. Nonlinear Dynam., Springer-Verlag, Berlin, 1993, pp. 205–241.

[9] V. B. Matveev, *Abelian Functions and Solitons*, Preprint 373, University of Wroclaw, Wroclaw, Poland, 1976.

[10] M. V. Pavlov, *Nonlinear Schrödinger equation and the Bogolyubov-Whitham averaging method,* Theoret. and Math. Phys., 71 (1987), pp. 584–588.

[11] V. U. Pierce and F. R. Tian, *Self-similar solutions of the non-strictly hyperbolic Whitham equations*, Commun. Math. Sci., 4 (2006), pp. 799–822.

[12] V. U. Pierce and F. R. Tian, *Self-similar solutions of the non-strictly hyperbolic Whitham equations for the KdV hierarchy*, Dyn. Partial Differ. Equ., 4 (2007), pp. 263–282.

[13] F. R. Tian, *The Whitham-type equations and linear overdetermined systems of Euler-Poisson-Darboux type*, Duke Math. J., 74 (1994), pp. 203–221.

[14] F. R. Tian and J. Ye, *On the Whitham equations for the semiclassical limit of the defocusing nonlinear Schrödinger equation*, Comm. Pure Appl. Math., 52 (1999), pp. 655–692.

[15] S. P. Tsarev, *Poisson brackets and one-dimensional Hamiltonian systems of hydrodynamic type*, Soviet Math. Dokl., 31 (1985), pp. 488–491.

[16] S. Venakides, *Higher order Lax-Levermore theory*, Comm. Pure Appl. Math., 43 (1990), pp. 335–362.

SIAM J. MATH. ANAL.
Vol. 40, No. 5, pp. 1783–1814

PERSISTENCE OF ROLL WAVES FOR THE SAINT VENANT EQUATIONS*

PASCAL NOBLE†

Abstract. The purpose of the article is to study the linear and nonlinear "stability" of roll-waves that are periodic and discontinuous entropic travelling wave solutions of the Saint Venant equations. More precisely, we prove that the Cauchy problem with initial data close to a roll-wave and satisfying suitable compatibility conditions has a solution on a sufficiently small interval.

Key words. roll-waves, shallow water, Cauchy problem, hyperbolic equations

AMS subject classifications. 35L45, 35L67, 35B10, 76H05, 35Q35

DOI. 10.1137/07070810X

1. Introduction.

Roll-waves are well-known nonlinear patterns appearing in shallow waters under the effect of gravity and bottom friction. This type of flow is commonly modeled by the "shallow water equations" that can be formally derived from the Navier–Stokes system. This is a hyperbolic system for the fluid height h and the average velocity u that is similar to the isentropic Euler equations. An empiric friction term, so-called Chezy friction term, is added to the momentum equation to model the friction of the bottom; see [7] for more details on the derivation of such models. In [3], Dressler proved the existence of periodic travelling waves that are mathematical solutions of the shallow water equations: those travelling waves are discontinuous, the discontinuity being a Lax shock. Similarly to shocks in hyperbolic systems, several questions arise at this stage: first, the existence of continuous roll-wave solutions of a viscous perturbation of the shallow water equations, the convergence to the inviscid roll-waves in the vanishing viscosity limit and their stability. The existence of continuous travelling waves for viscous shallow water equations follows from a classical Hopf bifurcation argument [9]. More involved is the question of the convergence of those solutions in the vanishing viscosity limit: one can prove, using geometric singular perturbation arguments (Fenichel theorems) that continuous roll-waves are ε close to an inviscid roll-wave with spatial period T and wave speed c_0 for a suitable wave speed $c(\varepsilon, T)$ with ε, the size of the viscosity [10], [11]. Only partial results are known concerning the stability of viscous roll-waves: those patterns are proved to be spectrally stable under large wavelength and small wavelength perturbations and are linearly stable provided that they are strongly spectrally stable; see [12] for more details.

In order to obtain more information on stability of viscous roll-waves, at least in the vanishing viscosity limit, we shall consider the inviscid case. In the case of a shock wave, the connection between the viscous and inviscid shocks has been established by Rousset [15] in the one-dimensional ($1D$) case and Gues and coworkers [6] in the multidimensional case. In the hyperbolic setting, the question of the "stability" of inviscid roll-waves then arises either under $1D$ or multidimensional perturbations.

*Received by the editors November 13, 2007; accepted for publication (in revised form) September 13, 2008; published electronically January 7, 2009.
http://www.siam.org/journals/sima/40-5/70810.html
†Université de Lyon, Université Lyon 1, UMR CNRS 5208, Institut Camille Jordan, Batiment du Doyen Jean Braconnier, 43, blvd du 11 novembre 1918, F - 69622 Villeurbanne Cedex, France (noble@math.univ-lyon1.fr).

In that paper and in order to make the connection with the viscous case, we shall focus on small *smooth* perturbations of roll-waves. Due to the presence of an infinite number of shocks, even the formulation of the stability problem is a hard task. In the case of a single shock, one formulates the stability problem among the piecewise smooth functions with a single discontinuity that is close to the shock: working in the reference frame attached to that discontinuity, one can reduce the stability problem to the analysis of an initial boundary value problem. In the case of roll-waves, it is natural to work among the functions that are piecewise regular with discontinuities that are close to the discontinuities of the roll-waves. Here it is not sufficient to work in the reference frame attached to *one* discontinuity; in that case, this would mean that we restrict our attention to *periodic* perturbations of the roll-waves. In that setting, Tamada and Tougou [17] proved that, for suitable large wavelength, roll-waves are spectrally stable. In order to handle more general perturbations, we need to fix *all* the discontinuities; this is done through a Lipschitz change of variable. The author proved in that case that for large wavelength, the spectral problem has no unstable eigenvalue [13]. Due to the hyperbolic nature of the problem, it seems hard to obtain a stability result in that case; we shall consider here the question of the "persistence" of roll-waves.

Similarly to the persistence of shocks, we shall prove here that the Cauchy problem with initial data that are close to a roll-wave and satisfy suitable compatibility conditions is well posed. First, we formulate the "stability" problem among the space of piecewise regular functions with discontinuities close to the roll-waves discontinuities: the shocks are fixed through a Lipschitz change of variable, and we write both the shallow water equations and the Rankine–Hugoniot jump conditions in that setting. Then we linearize that set of equations and analyze the spectral problem. In order to tackle the Cauchy problem, we shall only consider the high-frequency perturbations; we briefly analyze an analogous equation of the Lopatinskii determinant for shocks. Then, following the approach for shocks, we derive a priori estimates on the linear problem and for a linearized problem around an approximate solution. For an initial data satisfying suitable compatibility conditions, we can construct an approximate solution of the shallow water equations on a sufficiently small time interval. We obtain the existence of a solution to the Cauchy problem through a fixed point argument around the approximate solution with the a priori estimates on the linearized problem.

This method introduced by Majda [4] for compact shocks and generalized by Métivier [5] is suitable for *multimensional perturbations*. Though *valid in the multidimensional case*, the principal results obtained in this paper are written in *the* $1D$ *setting*. Indeed, in order to simplify the presentation, we have chosen to emphasize the periodic nature of the problem and the new issues that have to be handled with, in particular, the formulation of the stability problem and the presence of a sonic point in the interior that complicates higher-order estimates. In this particular setting, the energy estimates are derived easily without the microlocal analysis. We briefly discuss at the end of the paper how to extend the analysis proposed here to multidimensional perturbations. Moreover, this approach shall be generalized and developed in a forthcoming paper to handle roll-waves in generalized hyperbolic systems obtained recently by the author in [14].

The paper is organized as follows: in section 2, we recall the Dressler construction of roll-waves and formulate the boundary value problem associated to the persistence problem. In section 3, we linearize the equations around a roll-wave and perform the spectral analysis of that linear problem. We also obtain a priori estimates for

the solutions of that problem. We then consider the linearized problem around an approximate solution: we first obtain a priori estimates on solutions of that linear problem and deduce the well posedness of those equations in suitable functional spaces. Section 4 is devoted to the analysis of the full nonlinear problem: we first obtain compatibility conditions for initial data and construct an approximate solution of the shallow water equations coupled with Rankine–Hugoniot jump conditions. The well posedness of the Cauchy problem is then obtained through a fixed point argument.

2. Formulation of the problem. In this section, we give a short proof of the existence of inviscid roll-waves. Then, we introduce the functional spaces adapted to the stability analysis of that kind of pattern: these are piecewise regular functions with discontinuities close to the discontinuities of a roll-wave. We then perform the Lipschitz change of variable that fixes *all* the shocks and write the shallow water equations and the Rankine–Hugoniot conditions in the reference frame.

2.1. Existence of roll-waves. We briefly recall Dressler analysis for the existence of roll-wave solutions of the Saint Venant system. Let us start with the nondimensional Saint Venant equations

$$\begin{aligned} &h_t + (hu)_x = 0, \\ &(hu)_t + \left(\frac{h^2}{2F} + hu^2\right)_x = h - u^2, \end{aligned} \tag{1}$$

where h is the height of the fluid, u is the average fluid speed, and F is the Froude number. In [3], Dressler looks for periodic travelling wave solutions in the form

$$(h,u)(x,t) = (H,U)(x-ct)$$

of the inviscid Saint Venant equations: these yield a first order differential system. Eliminating U from the system reduces the problem to finding periodic solutions of the scalar equation

$$H' = \frac{H - (c - \frac{q}{H})^2}{\frac{H}{F} - \frac{q^2}{H^2}} = P_1(H), \tag{2}$$

where $q = H(c-U) > 0$, a constant, is the relative discharge rate. Integrating (2) in the form $\xi = \xi_0 + \int \frac{dh}{P_1(h)}$, it is proved that (2) has no continuous periodic solution [3]. Thus, we are looking for periodic solutions with discontinuities that satisfy admissibility conditions. The hyperbolic part of the system being similar to the isentropic Euler equations and the admissibility conditions for a discontinuity located at $x = x(t)$ are the Rankine–Hugoniot conditions

$$[hu] = \dot{x}[h], \quad \left[\frac{h^2}{2F} + hu^2\right] = \dot{x}[hu], \tag{3}$$

where $[u] = \lim_{\epsilon \to 0} u(x(t)+\epsilon) - u(x(t)-\epsilon)$ and a Lax shock condition

$$u_+ + \sqrt{\frac{h_+}{F}} < \dot{x} < u_- + \sqrt{\frac{h_-}{F}}.$$

Dressler proved the following result.

THEOREM 1. *Given any $F > 4$, $L > 0$, and $c > 0$, there exists a piecewise C^1 periodic travelling wave, with wave speed c and wavelength L, entropic solution of the inviscid Saint Venant equations* (1).

We recall the proof for completeness.

Proof. Let H_- (resp. H_+) be the fluid height before (resp. after) a shock. The entropic shock must satisfy the Rankine–Hugoniot jump conditions

$$\frac{H_+^2}{2F} + \frac{q^2}{H_+} = \frac{H_-^2}{2F} + \frac{q^2}{H_-}. \tag{4}$$

The Saint Venant equations are similar to the isentropic Euler equations for an ideal gas with $\gamma = 2$, where h plays the role of the density ρ. The entropy condition is equivalent to the Lax shock condition (see [18] for more details):

$$U_+ + \sqrt{\frac{H_+}{F}} < c < U_- + \sqrt{\frac{H_-}{F}}.$$

Since the relative discharge rate $H(c-U) = q$ is a constant, we obtain

$$\sqrt{\frac{H_+}{F}} - \frac{q}{H_+} < 0 < \sqrt{\frac{H_-}{F}} - \frac{q}{H_-}.$$

Hence, $H_+ < H_-$ and there exists a "sonic" point H_0 so that $H_+ < H_0 < H_-$ and

$$\frac{H_0}{F} - \frac{q^2}{H_0^2} = 0.$$

In order to satisfy the entropy condition, we must have $\frac{dH}{d\xi}|_{H_0} > 0$ or equivalently, $F > 4$. The denominator of the fraction in (2) vanishes at point H_0. To pass continuously through this value, the numerator must also vanish at this point H_0:

$$H_0 - \left(c - \frac{q}{H_0}\right)^2 = 0.$$

We simplify the fraction in (2):

$$\frac{dH}{d\xi} = F\frac{H^2 + \left(H_0 - c^2\right)H + \frac{H_0^2}{F}}{H^2 + H_0 H + H_0^2} = P_1(H). \tag{5}$$

The construction of a periodic solution with an arbitrary wavelength L is, at this point, straightforward. Let $H(\xi)$ be the special solution of (5) so that $H(0) = H_0$: it is defined implicitly by $H(\xi) = h \Leftrightarrow \xi = f_1(h)$, where f_1 is the primitive of $\frac{1}{P_1}$ such that $f_1(H_0) = 0$. The function f_1 is given by

$$\begin{aligned} Ff_1(h) = h - H_0 &+ \frac{H_a^2 + H_0 H_a + H_0^2}{H_a - H_b} \ln\left(\frac{h - H_a}{H_0 - H_a}\right) \\ &- \frac{H_b^2 + H_0 H_b + H_0^2}{H_a - H_b} \ln\left(\frac{h - H_b}{H_0 - H_b}\right), \end{aligned}$$

and $H_a > H_b$ are the zeros of P_1. Define $H_n(\xi) = H(\xi - nL)$. We fit H_n and H_{n+1} together by means of an entropic shock. Eliminating the solution $H_+ = H_-$, the Rankine–Hugoniot condition (4) now reads

$$H_+ + H_- = \frac{2H_0^3}{H_+ H_-}. \tag{6}$$

We determine the position of the nth shock ξ_c^n, with $H_n(\xi_c^n) = H_-$ and $H_{n+1}(\xi_c^n) = H_+$. This is equivalent to the relation

$$\int_{H_0}^{H^-} \frac{dh}{P_1(h)} = L + \int_{H_0}^{H_+} \frac{dh}{P_1(h)}. \tag{7}$$

Now, eliminating H_+ between (6) and (7) yields L as a function of H_-. Then we have obtained periodic and entropic solutions of the Saint Venant equations (1). □

In what follows, we rescale space variables and unknowns by H_0 in the variable $H := H_0 h$ and $\xi := H_0 \xi$: we get the nondimensional equation of profile

$$h' = F \frac{h^2 + \left(1 - \left(1 + \frac{1}{\sqrt{F}}\right)^2\right) h + \frac{1}{F}}{h^2 + h + 1} = P(h).$$

Denote $H_+ = H_0 h_+$ and $H_- = H_0 h_-$. The nondimensional Rankine–Hugoniot conditions reads

$$h_+ + h_- = \frac{2}{h_+ h_-}.$$

We can easily describe the asymptotic profile of the roll-waves as $L \to \infty$. Indeed, when $L \to \infty$, the minimum height h_+ converges to h_a given by

$$h_a = \frac{H_a}{H_0} = \frac{1}{2F}\left(1 + 2\sqrt{F} + \sqrt{1 + 4\sqrt{F}}\right).$$

Then the family of roll-waves converges to a solitary wave with a single shock as $L \to \infty$. Denote $h_m > h_a$ so that $h_a + h_m = \frac{2}{h_a h_m}$. Before the shock, the profile $\tilde{h}$ of the solitary wave is given by the special solution $\tilde{h} = h$, $h \in (h_a, h_m)$, and after the shock, it is a constant $\tilde{h} = h_a$. The description of the roll-wave solutions of the Saint Venant system is complete, and we are now in a position to formulate the stability problem.

2.2. Formulation of the stability problem. In this section, we recall the method, introduced in [13], of studying the stability of Dressler roll-waves that possess an infinite number of shocks . These roll-waves are parametrized by the wavelength L and the wave speed c. When these parameters are fixed, denote H_+ (resp. H_-) the minimum (resp. maximum) height of the roll-wave. This is also the height after (resp. before) a shock. We first write the Saint Venant system (1) into a diagonalized form and introduce the Riemann invariants

$$r = u + 2\sqrt{\frac{h}{F}}, \quad s = u - 2\sqrt{\frac{h}{F}}.$$

The shallow water equations (1) read

$$r_t + \lambda_1(r,s) r_x = Q(r,s), \quad s_t + \lambda_2(r,s) s_x = Q(r,s), \tag{8}$$

where $\lambda_k, k = 1, 2$ and Q are defined by

$$\lambda_1(r,s) = \frac{3r}{4} + \frac{s}{4}, \quad \lambda_2(r,s) = \frac{3s}{4} + \frac{r}{4}, \quad Q(r,s) = 1 - \frac{4}{F}\left(\frac{r+s}{r-s}\right)^2.$$

For a discontinuity located at $x = x(t)$, the Rankine–Hugoniot conditions in these variables reads

$$\begin{aligned} \left[(r+s)(r-s)^2)\right] &= 2\,\dot{x}\,\left[(r-s)^2\right], \\ \left[(r-s)^4 + 8(r-s)^2(r+s)^2\right] &= 16\,\dot{x}\,\left[(r+s)(r-s)^2\right]. \end{aligned} \tag{9}$$

The derivation of the spectral problem is inspired by the method introduced by Majda for studying the stability of multidimensional shocks in hyperbolic systems [4]. In what follows, we consider a Dressler roll-wave (H, U) (or equivalently, R, S) with wavespeed c and wavelength L: the location of the shocks are given by $x_i(t) = ct + iL$, $i \in \mathbb{Z}$. We consider small perturbations of that solution: we can suppose it is piecewise C^1 functions with discontinuities at positions $X_i(t) = ct + iL + \varepsilon_i(t)$, $i \in \mathbb{Z}$. The ε_i are supposed to be small. The Rankine–Hugoniot conditions (9) are then given by

$$\begin{aligned} \left[(r+s)(r-s)^2)\right]_{X_i} &= 2\dot{X}_i\,\left[(r-s)^2\right]_{X_i}, \\ \left[(r-s)^4 + 8(r-s)^2(r+s)^2\right]_{X_i} &= 16\dot{X}_i\,\left[(r+s)(r-s)^2\right]_{X_i} \end{aligned} \qquad \forall i \in \mathbb{Z}. \tag{10}$$

In [4], working in the reference frame attached to the discontinuity made the shock steady. In our case, because of the existence of an *infinite* distribution of shocks, we have to fix *all* the discontinuities in order to make the roll-wave steady, just as Majda did in the case of the stability of planar shocks [4]. In that case, a perturbation of a planar shock has a discontinuity front at points $x_n = \psi(y,t)$, with $y = (x_1, \dots, x_{n-1})$ the transverse variable, and one introduces a new variable $\xi = x_n - \psi(y,t)$ that fixes *all* the front shock and not only a variable $\xi = x_n - \psi(t)$ that fixes only the shock at $y = 0$. In the case of roll-waves, we fix all the shocks by working in the new system of coordinates $(\xi = \xi(x,t), t)$ such that $\xi(X_i(t), t) = iL\ \forall i \in \mathbb{Z}$. The function ξ has the form

$$\forall x \in]X_i(t), X_{i+1}(t)[,\ \xi(x,t) = \frac{x - X_i(t)}{X_{i+1}(t) - X_i(t)} L + iL\ \forall i \in \mathbb{Z}. \tag{11}$$

Remark. Considering $|\varepsilon_i| << 1$, we prove that $\xi(x,t) \approx x - ct$ up to first order; it looks like a change of reference frame.

Note that in the case of the stability of roll-waves, we have a sort of multidimensional stability issue where the transverse "variable" is $i \in \mathbb{Z}$. The function ξ is Lipschitz and the system is of order one, so this change of coordinates is licit and does not change the Rankine–Hugoniot jump conditions. We make the change of coordinates $(r,s)(x,t) = (\overline{r}, \overline{s})(\xi(x,t), t)$, where $(\overline{r}, \overline{s})$ are piecewise C^1 functions with discontinuities located at points $\{iL, i \in \mathbb{Z}\}$. The derivation rules are given by

$$\begin{aligned} \frac{\partial r}{\partial x} &= \frac{\partial \overline{r}}{\partial \xi} \frac{L}{X_{i+1} - X_i}, \\ \frac{\partial r}{\partial t} &= \frac{\partial \overline{r}}{\partial t} - \frac{L}{X_{i+1} - X_i} \left(\dot{X}_i + \frac{\xi - iL}{L} \left(\dot{X}_{i+1} - \dot{X}_i \right) \right) \frac{\partial \overline{r}}{\partial \xi}. \end{aligned} \tag{12}$$

Substituting (12) into the shallow water equations (8) and dropping the overlines into the equations, the system (8) reads for all $x \in (iL, (i+1)L)$,

$$\begin{aligned} r_t + \frac{L}{X_{i+1} - X_i} \left(\lambda_1(r,s) - \left(\dot{X}_i + \gamma_i(x) \left(\dot{X}_{i+1} - \dot{X}_i \right) \right) \right) r_x &= Q(r,s), \\ s_t + \frac{L}{X_{i+1} - X_i} \left(\lambda_2(r,s) - \left(\dot{X}_i + \gamma_i(x) \left(\dot{X}_{i+1} - \dot{X}_i \right) \right) \right) s_x &= Q(r,s), \end{aligned} \tag{13}$$

with $\gamma_i(x) = \frac{x-iL}{L} \in (0,1)$. The Rankine–Hugoniot conditions (10) remain unchanged and are given by

$$[hu]_{X_i} = \dot{X}_i[h]_{X_i}, \quad \left[\frac{h^2}{2F} + hu^2\right]_{X_i} = \dot{X}_i[hu]_{X_i}, \tag{14}$$

where h, u have to be understood as functions of r, s. The particular solution $(R, S, X_i(t) = ct + iL)$ computed by Dressler is now a steady solution of (13) and (14).

In this paper, we deal with the problem of the *persistence* of a roll-wave solution: Given any initial condition (r_0, s_0) which has the same structure as a roll-wave solution with discontinuities at points x_i^0, $i \in \mathbb{Z}$, we are going to prove that, for initial conditions close to a roll-wave solution and satisfying suitable compatibility conditions, there exists a solution of the Cauchy problem (13), (14) with the initial conditions $(r,s)(.,0) = (r_0, s_0)$ on a sufficiently small interval $(0, T^*)$ that conserves a structure similar to the roll-wave structure. This result is in some sense a result of the *stability* of roll waves. Due to the hyperbolic nature of the problem, we expect the apparition of new discontinuities that breaks the roll-wave structure. To deal with this problem, we follow the approach initiated by Majda [4] to prove the existence of compact shocks and developed by Metivier [5] for the stability of multidimensonal shocks: In order to deal with the full nonlinear problem, the first task is to prove the well posedness of the linearized equations around an approximate solution, which will be assumed to be close to a roll wave solution. Then, under suitable compatibility conditions, we are able to prove the existence of an approximate solution of (13), (14), and using a fixed point argument in the neighborhood of this approximate solution, we prove the existence of a solution of the full Cauchy problem.

3. Well posedness of the linearized equations. In this section, we prove the well posedness of the linear equations obtained by the linearization of (13), (14) in the neighborhood of an approximate solution close to a roll-wave. This is done by energy estimates on the linear problem and on an adjoint problem. We start with the analysis of the spectral problem associated to the linear equations obtained by linearization of (13), (14) around the *roll-wave solution* and then perform a priori estimates of solutions of that problem and the adjoint of that problem. We prove that these estimates remain true for the solutions of the linear equations obtained by linearization around a function "close" to a roll-wave. We deduce from the energy estimates that the linearized problem around an approximate solution is well posed in suitable functional spaces.

3.1. Energy estimates on the "exact" linearized problem. In what follows, we linearize shallow water equations and Rankine–Hugoniot conditions around an exact roll-wave solution and obtain energy estimates for that linearized problem: we shall deduce estimates for the linearized problem around an approximate solution.

3.1.1. Formulation of the linearized problem. The linearization of the equation in the neighborhood of the roll-wave solution yields the following, after time rescaling $t := H_0\, t$:

$$\begin{matrix} \partial_t\, r_i + a(h)\partial_x\, r_i \\ \partial_t\, s_i + b(h)\partial_x\, s_i \end{matrix} + L^0(h) \begin{pmatrix} r_i \\ s_i \end{pmatrix} = \mathcal{F}_i \quad \forall (x,i) \in (0,L) \times \mathbb{Z}, \tag{15}$$

with $a(h) = \sqrt{h} - h^{-1}$, $b(h) = -\sqrt{h} - h^{-1}$, $(r_i, s_i) = (r, s)_{|(iL, iL+L)}$,

$$\mathcal{F}_i = \left(\dot\varepsilon_i + \frac{x}{L}\left(\dot\varepsilon_{i+1} - \dot\varepsilon_i\right)\right) P(h) \begin{pmatrix} h^{-2} + h^{-\frac12} \\ h^{-2} - h^{-\frac12} \end{pmatrix} - \frac{c(\varepsilon_{i+1} - \varepsilon_i)P(h)}{L\left(1 + \sqrt{F}\right)} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

and

$$\begin{aligned}
L^0_{11} &= \frac34\left(h^{-2} + h^{-\frac12}\right) P(h) - h^{-\frac32}\left(1 + \sqrt{F} - h^{-1}\right)\left(1 + \sqrt{F} - h^{-1} - 2h^{\frac12}\right), \\
L^0_{12} &= \frac14\left(h^{-2} + h^{-\frac12}\right) P(h) + h^{-\frac32}\left(1 + \sqrt{F} - h^{-1}\right)\left(1 + \sqrt{F} - h^{-1} + 2h^{\frac12}\right), \\
L^0_{21} &= \frac14\left(h^{-2} - h^{-\frac12}\right) P(h) - h^{-\frac32}\left(1 + \sqrt{F} - h^{-1}\right)\left(1 + \sqrt{F} - h^{-1} - 2h^{\frac12}\right), \\
L^0_{22} &= \frac34\left(h^{-2} - h^{-\frac12}\right) P(h) + h^{-\frac32}\left(1 + \sqrt{F} - h^{-1}\right)\left(1 + \sqrt{F} - h^{-1} + 2h^{\frac12}\right).
\end{aligned}$$

Here h has to be understood as $h = h(x)$, the equation of profile of the roll-wave. The linearized Rankine–Hugoniot conditions are given by

$$\begin{aligned}
2[h]_{iL}\dot\varepsilon_i &= \left[\left(h - h^{-\frac12}\right) r + \left(h + h^{-\frac12}\right) s\right]_{iL}, \\
2[h]_{iL}\dot\varepsilon_i &= \left[\left(h - h^{-\frac12} + \frac{1}{1 + \sqrt{F}}\left(h^{\frac32} + h^{-\frac32} - 2\right)\right) r\right. \\
&\qquad \left. + \left(h + h^{-\frac12} - \frac{1}{1 + \sqrt{F}}\left(h^{\frac32} + h^{-\frac32} + 2\right)\right) s\right]_{iL} \quad \forall i \in \mathbb{Z},
\end{aligned} \tag{16}$$

with $[h]_{iL} = h_+ - h_-$. In order to simplify the notations, we introduce the matrix $M(h)$ so that (16) reads

$$2[h]\dot\varepsilon \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \left[M(h) \begin{pmatrix} r \\ s \end{pmatrix}\right]_{iL}.$$

Equivalently, system (16) can be written

$$\begin{aligned}
2[h]\dot\varepsilon_i &= \left[\left(h - h^{-\frac12}\right) r + \left(h + h^{-\frac12}\right) s\right]_{iL}, \\
0 &= \left[\left(h^{\frac32} + h^{-\frac32} - 2\right) r - \left(h^{\frac32} + h^{-\frac32} + 2\right) s\right]_{iL}
\end{aligned} \qquad \forall i \in \mathbb{Z}. \tag{17}$$

In order to eliminate the derivatives of ε_i in the interior equations (15), one can, as usual, introduce the "good" unknowns:

$$\begin{aligned}
\overline{r}_i &= r_i + \left(\varepsilon_i + \frac{x}{L}(\varepsilon_{i+1} - \varepsilon_i)\right) P(h)\left(h^{-2} + h^{-\frac12}\right), \\
\overline{s}_i &= s_i + \left(\varepsilon_i + \frac{x}{L}(\varepsilon_{i+1} - \varepsilon_i)\right) P(h)\left(h^{-2} - h^{-\frac12}\right).
\end{aligned}$$

Dropping the overlines, this yields the system

$$\begin{aligned}
\partial_t r_i + a(h)\partial_x r_i &= \overline{\mathcal{F}}_i \quad \forall (x, i) \in (0, L) \times \mathbb{Z}, \\
\partial_t s_i + b(h)\partial_x s_i &
\end{aligned} \tag{18}$$

where $\overline{\mathcal{F}}_i$ contains only zeroth order terms. More precisely, $\overline{\mathcal{F}}_i$ is written

$$\begin{aligned}\overline{\mathcal{F}}_i = &- L^0 \begin{pmatrix} r_i \\ s_i \end{pmatrix} - \varepsilon_i P(h) L^0 \begin{pmatrix} h^{-2}+h^{-\frac{1}{2}} \\ h^{-2}-h^{-\frac{1}{2}} \end{pmatrix} \\ &- \frac{x}{L}(\varepsilon_{i+1}-\varepsilon_i) P(h) L^0 \begin{pmatrix} h^{-2}+h^{-\frac{1}{2}} \\ h^{-2}-h^{-\frac{1}{2}} \end{pmatrix} \\ &- \varepsilon_i \mathrm{diag}\,(a(h),b(h))\,\partial_x \begin{pmatrix} P(h)\big(h^{-2}+h^{-\frac{1}{2}}\big) \\ P(h)\big(h^{-2}-h^{-\frac{1}{2}}\big) \end{pmatrix} \\ &- (\varepsilon_{i+1}-\varepsilon_i)\mathrm{diag}(a(h),b(h))\partial_x \begin{pmatrix} \frac{x\,P(h)}{L}\big(h^{-2}+h^{-\frac{1}{2}}\big) \\ \frac{x\,P(h)}{L}\big(h^{-2}-h^{-\frac{1}{2}}\big) \end{pmatrix}.\end{aligned}$$

The Rankine–Hugoniot jump conditions are given by

$$\begin{aligned} 2[h]\dot{\varepsilon}_i &= \left[\left(h-h^{-\frac{1}{2}}\right) r + \left(h+h^{-\frac{1}{2}}\right) s\right]_{iL}, \\ -\left[P(h)\left(h-h^{-2}\right)\right]\varepsilon_i &= \left[\left(h^{\frac{3}{2}}+h^{-\frac{3}{2}}-2\right) r - \left(h^{\frac{3}{2}}+h^{-\frac{3}{2}}+2\right) s\right]_{iL}. \end{aligned} \tag{19}$$

As usual, we consider the linear boundary value problem dropping the zeroth order term from (18, 19):

$$\begin{aligned} &\partial_t r_i + a(h)\partial_x r_i = F_i^1, \\ &\partial_t s_i + b(h)\partial_x s_i = F_i^2 \end{aligned} \qquad \forall i \in \mathbb{Z}, x \in (0,L), \tag{20}$$

with the linearized Rankine–Hugoniot conditions

$$\begin{aligned} 2[h]\dot{\varepsilon}_i &= \left[\left(h-h^{-\frac{1}{2}}\right) r + \left(h+h^{-\frac{1}{2}}\right) s\right]_{iL} + a_{1,i}, \\ \left[\left(h^{\frac{3}{2}}+h^{-\frac{3}{2}}-2\right) r - \left(h^{\frac{3}{2}}+h^{-\frac{3}{2}}+2\right) s\right]_{iL} &= a_{2,i} \end{aligned} \qquad \forall i \in \mathbb{Z}. \tag{21}$$

These conditions can be interpreted as transmission conditions at the point of discontinuities between (r_i, s_i) and (r_{i+1}, s_{i+1}). Indeed, a jump of a function f of r, s at point $x = iL$ can be written

$$[f(r,s)]_{iL} = f(r_i(0), s_i(0)) - f(r_{i-1}(L), s_{i-1}(L)).$$

3.1.2. Normal mode analysis. We carry out a similar analysis made in the stability of multidimensional shocks and search for possible unstable eigenmodes. We make a Laplace transform in time and Fourier transform in discrete space (that yields Fourier series) of (20, 21):

$$\lambda \overline{r} + a(h)\overline{r}' = 0, \;\; \lambda \overline{s} + b(h)\overline{s}' = 0 \; \forall x \in (0,L),$$

with

$$\overline{r}(\lambda, \theta, x) = \sum_{k \in \mathbb{Z}} e^{ik\theta} \int_{-\infty}^{\infty} e^{\lambda t} r_k(x,t)\, dt.$$

The linearized Rankine–Hugoniot conditions reads

$$2\lambda(h_- - h_+)\overline{\varepsilon} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = e^{i\theta} M(h_-) \begin{pmatrix} \overline{r}(L) \\ \overline{s}(L) \end{pmatrix} - M(h_+) \begin{pmatrix} \overline{r}(0) \\ \overline{s}(0) \end{pmatrix}.$$

Here the space $\mathbb{E}^s$ of functions decreasing to 0 at infinity used in the analysis of multidimensional shocks is replaced by the space $\mathbb{E}^a$ of *bounded* functions at the point of singularity x_0 such that $h(x_0) = 1$. A basis of solutions for the differential system (3.1.2) is given by

$$V_1 = \left((x-x_0)^{-\frac{\lambda}{a'(x_0)}} Z(x), 0\right), \; V_2 = \left(0, e^{-\lambda \int_{x_0}^x \frac{ds}{b(h(s))}}\right),$$

where Z is an analytic (thus bounded) function on the interval $(0, L)$ that is unique if we choose $Z(0) = 1$. When $\Re(\lambda) > 0$, it is easily seen that $\mathbb{E}^a = \langle V_2 \rangle$ so that the analogous equation of the Lopatinskii determinant is an Evans function given by

$$\Delta(\lambda, \theta) = \frac{2[h]}{1+\sqrt{F}} \lambda \left(\left(h_+^{\frac{3}{2}} + h_+^{-\frac{3}{2}} + 2 \right) - \left(h_-^{\frac{3}{2}} + h_-^{-\frac{3}{2}} + 2 \right) e^{i\theta - \lambda \int_0^L \frac{ds}{b(h(s))}} \right).$$

The function $h \mapsto h^{\frac{3}{2}} + h^{-\frac{3}{2}} + 2$ is decreasing on the interval (0, 1), is increasing on (1, $+\infty$), and is invariant under the function $\mapsto \frac{1}{h}$. The Rankine–Hugoniot jump conditions for the roll-wave $h_+ + h_- = \frac{2}{h_+ h_-}$ imply that $h_- = *10 \frac{1}{2}(\sqrt{h_+^2 + \frac{8}{h_+}} - h_+) < \frac{1}{h_+}$ for any $h_+ < 1$. Then we find that $h_+^{\frac{3}{2}} + h_+^{-\frac{3}{2}} + 2 > h_-^{\frac{3}{2}} + h_-^{-\frac{3}{2}} + 2$. As a consequence, on the half plane $\Re(\lambda) > 0$, the Evans function Δ vanishes either when $\lambda = 0$ (for any θ) and for

$$-\Re(\lambda) \int_0^L \frac{ds}{b(h(s))} = \log \left(\frac{h_+^{\frac{3}{2}} + h_+^{-\frac{3}{2}} + 2}{h_-^{\frac{3}{2}} + h_-^{-\frac{3}{2}} + 2} \right), \; \theta = \Im(\lambda) \int_0^L \frac{ds}{b(h(s))} \; [2\pi].$$

We obtain a full line of unstable eigenmodes with a positive growth rate $\Re(\lambda)$ in time, since $b(h) < 0$ and $h_+^{\frac{3}{2}} + h_+^{-\frac{3}{2}} + 2 > h_-^{\frac{3}{2}} + h_-^{-\frac{3}{2}} + 2$.

Contrary to the case of multidimensional shocks where there is only one shock, this *does not* imply strong instability since the Evans function is not homogeneous; the growth rate of the unstable eigenmodes is *bounded*, whereas in the case of a single shocks, a full line of eigenvalues with an arbitrary large real part is found. This is due to the fact that in the case of shocks, the equations are invariant under the hyperbolic scaling $(x, t) \mapsto (\mu x, \mu t)$ for any $\mu > 0$. As a consequence, an unstable mode is accompanied with a full line of unstable modes that are more and more unstable as the frequency goes to infinity. This is not the case here due to the periodic nature of the problem. Moreover, it is easily seen that when $L \to \infty$, $\mathrm{Re}(\lambda) \to 0$: the zeroth order terms have an importance in the stability issue as pointed out in [13].

3.1.3. Energy estimates. In this section, we obtain energy estimates for solutions of the linear problem (20), (21). We multiply the first (resp. second) equation of (20) by r_i (resp. s_i), integrate on (0, L), and sum for $i \in \mathbb{Z}$:

$$\partial_t \sum_{i\in\mathbb{Z}} \int_0^L \left(r_i^2 + s_i^2\right) dx + \sum_{i\in\mathbb{Z}} \int_0^L a(h)\partial_x r_i^2 + b(h)\partial_x s_i^2 = 2 \sum_{i\in\mathbb{Z}} \int_0^L F_i^1 r_i + F_i^2 r_i dx.$$

An integration by parts on the second integral yields

$$\int_0^L a(h)\partial_x r_i^2 + b(h)\partial_x s_i^2 = \left[a(h) r_i^2 + b(h) s_i^2\right]_0^L - \int_0^L \partial_x a \, r_i^2 + \partial_x b \, s_i^2 \, dx.$$

We change the indices of summation in the boundary terms and find that

$$\partial_t \sum_{i\in\mathbb{Z}} \int_0^L \left(r_i^2 + s_i^2\right) dx - \sum_{i\in\mathbb{Z}} \left[a(h)r^2 + b(h)s^2\right]_{iL} = 2\sum_{i\in\mathbb{Z}} \int_0^L F_i^1 r_i + F_i^2 s_i dx + \sum_{i\in\mathbb{Z}} \int_0^L \partial_x a(h)\, r_i^2 + \partial_x b(h)\, s_i^2. \tag{22}$$

Equation (22) can be written in the form

$$\partial_t \int_{\mathbb{R}} r^2 + s^2\, dx - \sum_{i\in\mathbb{Z}} \left[a(r)\, r^2 + b(h)\, s^2\right]_{iL} = 2\int_{\mathbb{R}} F\, r + G\, s\, dx + \int_{\mathbb{R}} \partial_x\, a(h)\, r^2 + \partial_x\, b(h)\, s^2\, dx. \tag{23}$$

Next, we have to obtain an estimate on the boundary terms. For that purpose, we show that the boundary condition

$$f_- r_i^- - f_+ r_i^+ = g_- s_i^- - g_+ s_i^+,$$

with

$$f_\pm = f(h_\pm),\ g_\pm = g(h_\pm),\ f(h) = h^{\frac{3}{2}} + h^{-\frac{3}{2}} - 2,\ g(h) = h^{\frac{3}{2}} + h^{-\frac{3}{2}} + 2,$$

is maximal dissipative. More precisely, we prove that

$$B\left(r_i^+, r_i^-, s_i^+, s_i^-\right) = 0 \ \Rightarrow \mathcal{Q}\left(r_i^+, r_i^-, s_i^+, s_i^-\right) \geq 0,$$

with $Bu = \langle rh, u\rangle$, rh being the vector $rh =^t (f_+, -f_-, -g_+, g_-)$ and $\mathcal{Q}$ is the quadratic form

$$\mathcal{Q}\left(r_i^+, r_i^-, s_i^+, s_i^-\right) = a_-\left(r_i^-\right)^2 + b_-\left(s_i^-\right)^2 - a_+\left(r_i^+\right)^2 - b_+\left(s_i^+\right)^2,$$

with $a_\pm = a(h_\pm), b_k = b(h_\pm)$. This property is necessarily maximal, since $\dim(\mathrm{KerB}) = 3$ and $b_- < 0$, whereas $-b_+, a_-, -a_+ > 0$. The restriction of the quadratic form $\mathcal{Q}$ to KerB is represented by the symmetric matrix $\mathcal{A}$:

$$\mathcal{A} = \begin{pmatrix} a_- g_+^2 - b_+ f_-^2 & b_+ f_- g_- & b_+ f_+ f_- \\ b_+ f_- g_- & b_- g_+^2 - b_+ g_-^2 & -b_+ f_+ g_- \\ b_+ f_+ f_- & -b_+ f_+ g_- & -a_+ g_+^2 - b_+ f_+^2 \end{pmatrix}.$$

The matrix $\mathcal{A}$ is definite positive provided that all the principal determinants Δ_i, $i = 1, 2, 3$ are strictly nonnegative. Let us check that condition. First, $\Delta_1 = a_- g_+^2 - b_+ f_-^2 \geq 0$ since $a_-, -b_+ \geq 0$, and $\Delta_1 \neq 0$ provided that there is a real shock $h_+ < 1 < h_-$ (otherwise, it is just the stationary solution). A straightforward computation shows that

$$\Delta_2 = g_+^2\left(a_- b_- g_+^2 - b_+\left(a_- g_-^2 + b_- f_-^2\right)\right),$$

and $\Delta_2 > 0$ if and only if the following condition is satisfied:

$$h_+\left(1 + h_+^{\frac{3}{2}}\right)\left(1 + h_+^{-\frac{3}{2}}\right)^2 < 6h_- + \frac{2}{h_-^2}. \tag{24}$$

It is easily proved that this relation is satisfied for *small amplitude* roll-waves when $h_+ < 1 < h_-$ lie in the neighborhood of $h = 1$. In the large amplitude case, recall that $h_a < h_+ < 1$, where $h_a = \frac{1}{2F}(1+2\sqrt{F}+\sqrt{1+4\sqrt{F}})$ and $h_+ + h_- = \frac{2}{h_+ h_-}$. We write the Rankine–Hugoniot jump condition for the roll wave as

$$h_- = RH(h_+) = \frac{1}{2}\left(\sqrt{h_+^2 + \frac{8}{h_+}} - h_+\right)$$

and inserting that relation into (24), the problem is reduced to the analysis of the sign of

$$C(h_+) = 6RH(h_+) + \frac{2}{RH(h_+)^2} - h_+\left(1+h_+^{\frac{3}{2}}\right)\left(1+h_+^{-\frac{3}{2}}\right)^2$$

when $h_+ \in (h_a, 1)$. It is a lengthly but straightforward computation to prove that $C(h_+) > 0$ for all $h_+ \in (h_a, 1)$ provided that $4 < F \leq F_c$, with $F_c \geq 18$. Thus, when $4 < F \leq F_c$, the relation $\Delta_2 > 0$ is satisfied for *all* roll-wave solutions and for larger Froude numbers; there exists $h_c(F) > h_a$ so that $\Delta_2 > 0$ is satisfied for all $h_+ \in (h_c(F), 1)$. Next, we consider the last condition $\Delta_3 > 0$. One can show that

$$\Delta_3 = g_+^4\left(\left(a_- g_-^2 + b_- f_-^2\right)a_+ b_+ - \left(a_+ g_+^2 + b_+ f_+^2\right)a_- b_-\right)$$

and $\Delta_3 > 0$ if and only if

$$3h_- + \frac{1}{h_-^2} > 3h_+ + \frac{1}{h_+^2}.$$

Then it is an easy calculation to prove that this condition is satisfied for *all* roll-wave solutions if $4 < F \leq F_c$, with $F_c \geq 12$. As a conclusion, the matrix $\mathcal{A}$ is definite positive if $4 < F \leq F_c$ ($F_c \geq 12$). Suppose that this assumption is satisfied. The quadratic form $\mathcal{Q}_{|\mathrm{KerB}}$ is definite positive. Denote $w = (r_i^+, r_i^-, s_i^+, s_i^-)$. We deduce that there exist $\kappa > 0$ and $C > 0$ so that

$$\mathcal{Q}(w) \geq \kappa\|w\|^2 - C\|Bw\|^2$$

and substituting that relation into (23), we deduce the estimate

$$\begin{aligned}\partial_t\|(r,s)\|_2^2 + \kappa\sum_{i\in\mathbb{Z}}\left|\left(r_i^\pm, s_i^\pm\right)\right|^2 \\ \leq 2\int_{\mathbb{R}} F^1\, r + F^2\, s\, dx + C\|(r,s)\|^2 + \|a_2\|_{l^2(\mathbb{Z})}^2,\end{aligned} \tag{25}$$

where $\|(r,s)\|_2^2 = \int_{\mathbb{R}} r^2 + s^2\, dx$ is the classical $\left(L^2(\mathbb{R})\right)^2$-norm. From the Rankine–Hugoniot condition, one can easily estimate $\varepsilon_i, \dot{\varepsilon}_i$:

$$|\dot{\varepsilon}_i|^2 \leq C\left(\left|\left(r_i^\pm, s_i^\pm\right)\right|^2 + a_{1,i}^2\right).$$

We compute estimates in weighted functional spaces. For $\gamma > 0$ sufficiently large (that depends on $\|(R,S)\|_{W^{1,\infty}(0,L)}$) and using Young's type inequalities, we obtain the following proposition.

PROPOSITION 1. *Assume that the Froude number F satisfies $0 < F < F_c$ (with $F_c > 12$). Then there exists γ_c (depending on $\|(R,S)\|_{W^{1,\infty}(0,L)}$) such that for all $\gamma > \gamma_c$, the following estimate holds true:*

$$\begin{aligned}&\gamma \int_{\mathbb{R}} e^{-2\gamma t}\|(r,s)\|_2^2 + \int_{\mathbb{R}} e^{-2\gamma t}\sum_{i\in\mathbb{Z}} \left|\left(r_i^{\pm}, s_i^{\pm}\right)\right|^2 + \left(\varepsilon_i^2 + \dot\varepsilon_i^2\right)\, dt\\ &\le C\left(\frac{1}{\gamma}\int_{\mathbb{R}} e^{-2\gamma t}\left\|\left(F^1,F^2\right)\right\|_2^2 dt + \int_{\mathbb{R}} e^{-2\gamma t}\left(\|a_1\|_{l^2(\mathbb{Z})}^2 + \|a_2\|_{l^2(\mathbb{Z})}^2\right) dt\right).\end{aligned} \tag{26}$$

We are now in a position to make the energy estimates on the linearized problem around an approximate solution close to a roll-wave. In what follows, we shall assume that the assumption $0 < F < F_c$ is satisfied.

3.2. Energy estimates for the "approximate" linearized problem. In order to tackle the full nonlinear problem, we consider the linearization of the Saint Venant system (13) in the neighborhood of an approximate solution (R_i, S_i, X_i) that is close to the original roll-wave $(R, S, ct+iL)$. In what follows, we shall prove that the estimates (26) obtained in the previous section remain true for (R_i, S_i, X_i) sufficiently "close" to the roll-wave solution $(R, S, ct + iL)$.

3.2.1. Formulation of the problem. We consider an approximate "solution" $(R_i, S_i, X_i) \approx (R, S, ct + iL)$ and linearize the shallow water equations and Rankine–Hugoniot jump conditions around that function:

$$\begin{aligned}&\partial_t r_i + \frac{L}{X_{i+1} - X_i}\Lambda_1^i \partial_x r_i = \overline{F}_i^1,\\ &\partial_t s_i + \frac{L}{X_{i+1} - X_i}\Lambda_2^i \partial_x s_i = \overline{F}_i^2\end{aligned} \qquad \forall (x,i) \in (0,L)\times\mathbb{Z}, \tag{27}$$

where the function $F_k^i, k = 1, 2$ are defined by

$$\begin{aligned}\overline{F}_i^1 &= \frac{\partial Q}{\partial r}(R_i,S_i)r_i + \frac{\partial Q}{\partial s}(R_i,S_i)s_i + \frac{L(\varepsilon_{i+1}-\varepsilon_i)}{(X_{i+1}-X_i)^2}\Lambda_1^i\\ &\quad - \frac{L}{X_{i+1}-X_i}\left(\lambda_1(r_i,s_i) - \left(\dot\varepsilon_i + \frac{x}{L}\left(\dot\varepsilon_{i+1} - \dot\varepsilon_i\right)\right)\right)\partial_x R_i\\ \overline{F}_i^2 &= \frac{\partial Q}{\partial r}(R_i,S_i)r_i + \frac{\partial Q}{\partial s}(R_i,S_i)s_i + \frac{L(\varepsilon_{i+1}-\varepsilon_i)}{(X_{i+1}-X_i)^2}\Lambda_2^i\\ &\quad - \frac{L}{X_{i+1}-X_i}\left(\lambda_2(r_i,s_i) - \left(\dot\varepsilon_i + \frac{x}{L}\left(\dot\varepsilon_{i+1} - \dot\varepsilon_i\right)\right)\right)\partial_x S_i\end{aligned}$$

and $\Lambda_k^i, k = 1, 2, i \in \mathbb{Z}$ denotes $\Lambda_k^i = \lambda_k(R_i, S_i) - (\dot X_i + \frac{x}{L}(\dot X_{i+1} - \dot X_i))$. The linearized Rankine–Hugoniot jump conditions are given by

$$\begin{aligned}[\mathbb{Q}(R_i,S_i)]\dot\varepsilon_i &= \left[\left(\frac{\partial \mathbb{F}}{\partial R} - \dot X_i\frac{\partial \mathbb{Q}}{\partial R}\right)r + \left(\frac{\partial \mathbb{F}}{\partial S} - \dot X_i\frac{\partial \mathbb{Q}}{\partial S}\right)s\right]_{iL}\\ &\quad + \frac{L\varepsilon_i}{X_{i+1}-X_i}\left[\partial_x\left(\mathbb{F} - \dot X_i\mathbb{Q}\right)\right]_{iL},\\ [H(R_i,S_i)]\dot\varepsilon_i &= \left[\left(\frac{\partial \mathbb{Q}}{\partial R} - \dot X_i\frac{\partial H}{\partial R}\right)r + \left(\frac{\partial \mathbb{Q}}{\partial S} - \dot X_i\frac{\partial H}{\partial S}\right)s\right]_{iL}\\ &\quad + \frac{L\varepsilon_i}{X_{i+1}-X_i}\left[\partial_x\left(\mathbb{Q} - \dot X_i H\right)\right]_{iL},\end{aligned} \tag{28}$$

where the functions $\mathbb{Q} = HU, \mathbb{F} = \frac{H^2}{2F} + HU^2$ shall be considered as functions of R, S. One can introduce the "good" unknowns to drop the derivatives of ε_i in the interior equation (27). Similarly to the previous section, we analyse the linear problem obtained after dropping zeroth order terms in (27), (28):

$$\begin{aligned} &\partial_t r_i + \frac{L}{X_{i+1} - X_i} \Lambda_1^i \, \partial_x r_i = F_i^1, \\ &\partial_t s_i + \frac{L}{X_{i+1} - X_i} \Lambda_2^i \, \partial_x s_i = F_i^2 \end{aligned} \qquad \forall x \in (0, L) \; \forall i \in \mathbb{Z}. \tag{29}$$

The system (29) is completed with the linearized Rankine–Hugoniot conditions

$$\begin{aligned} [\mathbb{Q}(R_i, S_i)]\dot{\varepsilon}_i &= \left[\left(\frac{\partial \mathbb{F}}{\partial R} - \dot{X}_i \frac{\partial \mathbb{Q}}{\partial R} \right) r + \left(\frac{\partial \mathbb{F}}{\partial S} - \dot{X}_i \frac{\partial \mathbb{Q}}{\partial S} \right) s \right]_{iL} + a_{1,i}, \\ [H(R_i, S_i)]\dot{\varepsilon}_i &= \left[\left(\frac{\partial \mathbb{Q}}{\partial R} - \dot{X}_i \frac{\partial H}{\partial R} \right) r + \left(\frac{\partial \mathbb{Q}}{\partial S} - \dot{X}_i \frac{\partial H}{\partial S} \right) s \right]_{iL} + a_{2,i} \end{aligned} \qquad \forall i \in \mathbb{Z}. \tag{30}$$

3.2.2. Energy estimates. In what follows, we shall prove the following result.

PROPOSITION 2. *There exists $\eta > 0$ so that for all $m > 0$, there exists a constant C and γ_0 depending on m and for any Lipschitz continuous and bounded function $(R_i, S_i, X_i)_{i \in \mathbb{Z}}$ satisfying*

$$\begin{aligned} &\max_{i \in \mathbb{Z}} \left(\left\| \dot{X}_i - c \right\|_\infty + \| X_{i+1} - X_i - L \|_\infty + \| R_i - R, S_i - S \|_\infty \right) \leq \eta, \\ &\max_{i \in \mathbb{Z}} \left(|\partial_x (R_i, S_i)|_\infty + \left| \dot{X}_i \right|_\infty \right) \leq m, \end{aligned}$$

and for all $\gamma \geq \gamma_0$, $(F^1, F^2) \in \mathcal{D}(\mathbb{R}/\{iL\} \times \mathbb{R})$, $(a_1, a_2) \in l^2(\mathbb{Z}, \mathcal{D}(\mathbb{R}))$, then a solution $(r_i, s_i, \varepsilon_i)_{i \in \mathbb{Z}}$ of (29), (30) satisfies

$$\begin{aligned} &\gamma \left\| e^{-\gamma t}(r, s) \right\|^2_{L^2(\mathbb{R}/\{iL\} \times \mathbb{R})} + \left\| e^{-\gamma t} \left(r_i^\pm, s_i^\pm \right) \right\|^2_{l^2(\mathbb{Z}, L^2(\mathbb{R}))} + \left\| e^{-\gamma t} \varepsilon_i \right\|_{l^2(\mathbb{Z}, H^1(\mathbb{R}))} \\ &\leq C \left(\frac{1}{\gamma} \| e^{-\gamma t} \left(F^1, F^2 \right) \|^2_{L^2(\mathbb{R}/\{iL\} \times \mathbb{R})} + \left\| e^{-\gamma t} (a_1, a_2) \right\|_{l^2(\mathbb{Z}, L^2(\mathbb{R}))} \right). \end{aligned}$$

Proof. Similarly to the previous case, we multiply the first (resp. second) equation of (29) by r_i (resp. s_i), integrate on $(0, L)$, and sum for $i \in \mathbb{Z}$:

$$\begin{aligned} &\partial_t \int_{\mathbb{R}} r^2 + s^2 \, dx + \sum_{i \in \mathbb{Z}} \frac{L}{X_{i+1} - X_i} \left[\Lambda_1^i \, r_i^2 + \Lambda_2^i \, s_i^2 \right]_0^L \\ &= 2 \int_{\mathbb{R}} F^1 \, r + F^2 \, s \, dx + \sum_{i \in \mathbb{Z}} \frac{L}{X_{i+1} - X_i} \int_0^L \partial_x \Lambda_1^i \; r_i^2 + \partial_x \Lambda_2^i \; s_i^2 \, dx. \end{aligned}$$

It is a straightforward computation to prove that the boundary terms satisfy

$$\begin{aligned} \sum_{i \in \mathbb{Z}} \frac{L}{X_{i+1} - X_i} \left[\Lambda_1^i \, r_i^2 + \Lambda_2^i \, s_i^2 \right]_0^L \geq &- \sum_{i \in \mathbb{Z}} \left[a(h) \, r^2 + b(h) \, s^2 \right]_{iL} \\ &- \| \delta \|_{l^\infty(\mathbb{Z})} \sum_i \left| \left(r_i^\pm, s_i^\pm \right) \right|^2, \end{aligned}$$

where $(r_i^{\pm}, s_i^{\pm})$ denotes the vector $(r_i^+, r_i^-, s_i^+, s_i^-)$ and $(\delta_i)_{i\in\mathbb{Z}}$ satisfies

$$\delta_i = \mathcal{O}\left(\left\|\dot{X}_i - c\right\|_\infty + \|X_{i+1} - X_i - L\|_\infty + \|R_i - R, S_i - S\|_\infty\right).$$

On the other hand, for $\max_i(\|R-R_i, S-S_i\|_\infty)$ sufficiently small, we obtain $[H(R_i, S_i)] \neq 0$, and we can write the Rankine–Hugoniot jump conditions in the form

$$\dot{\varepsilon}_i = [\mathcal{L}_1^i(r,s)]_{iL} + L_1^i(a_i, b_i), \quad [\mathcal{L}_2^i(r,s)]_{iL} = L_2^i(a_i, b_i),$$

so that

$$[\mathcal{L}_2^i(r,s)]_{iL} = \left[\left(h^{\frac{3}{2}} + h^{-\frac{3}{2}} - 2\right) r + \left(h^{\frac{3}{2}} + h^{-\frac{3}{2}} + 2\right) s\right]_{iL} + \mu_i \left|\left(r_i^{\pm}, s_i^{\pm}\right)\right|,$$

and the real numbers μ_i are estimated by

$$\mu_i = \mathcal{O}\left(\left\|\dot{X}_i - c\right\|_\infty + \|X_{i+1} - X_i - L\|_\infty + \|R_i - R, S_i - S\|_\infty\right).$$

Moreover, the linear operators L_k^i are uniformly bounded:

$$|L_k^i(a,b)| \leq C|(a,b)| \ \forall i \in \mathbb{Z},\ k = 1, 2.$$

Choosing μ_i, δ_i sufficiently small and for

$$\max_{i\in\mathbb{Z}}\left(\|R_i, S_i\|_{W^{1,\infty}(0,L)}\right) < m, \quad \max_{i\in\mathbb{Z}}\left(\left|\dot{X}_i\right|\right) < m, \quad m < \infty,$$

we deduce the following estimate:

$$\begin{aligned}
&\gamma \int_{\mathbb{R}} e^{-2\gamma t}\|(r,s)\|_2^2\, dt + \int_{\mathbb{R}} e^{-2\gamma t} \sum_i \left|\left(r_i^{\pm}, s_i^{\pm}\right)\right|^2 + \dot{\varepsilon}_i^2 + \varepsilon_i^2\, dt \\
&\leq C\left(\frac{1}{\gamma}\int_{\mathbb{R}} e^{-2\gamma t}\left\|\left(F^1, F^2\right)\right\|_2^2\, dt + \int_{\mathbb{R}} e^{-2\gamma t}\left(\|a_1\|_{l^2(\mathbb{Z})}^2 + \|a_2\|_{l^2(\mathbb{Z})}^2\right) dt\right)
\end{aligned}$$

for any $\gamma \geq \gamma_0 > 0$, where γ_0 depends on $\|\delta\|_{l^\infty(\mathbb{Z})}$ and m. This completes the proof of the proposition. ∎

3.3. The adjoint problem. In this section, we formulate the adjoint problem of the linearized equations around the roll-wave and around an approximate solution that is close to a roll-wave. We obtain energy estimates for that adjoint problem that shall be useful to prove the existence of weak solutions to the linearized equations.

3.3.1. The adjoint problem of the "exact" linearized equations. We multiply the first equation (resp. second) of (20) by p_i (resp. q_i), integrate over $(0, L)\times\mathbb{R}$, and sum on $i \in \mathbb{Z}$:

$$\begin{aligned}
&\int_{\mathbb{R}} \sum_i \int_0^L \partial_t\, r_i\, p_i + \partial_t\, s_i\, q_i + a(h)\partial_x\, r_i\, p_i + b(h)\partial_x\, s_i\, q_i\, dxdt \\
&\qquad = \int_{\mathbb{R}} \sum_i \int_0^L F_i^1\, p_i + F_i^2\, q_i.
\end{aligned}$$

An integration by parts on the interval $(0, L)$ and $t \in \mathbb{R}$ yields

$$\begin{aligned}
&-\int_{\mathbb{R}} \sum_i \int_0^L (p_{i,t} + (a(h)p_i)_x) r_i + (q_{i,t} + (b(h)q_i)_x) s_i\, dx \\
&\quad - \int_{\mathbb{R}} \sum_i [a(h)p\, r + b(h)q\, s]_{iL} = \int_{\mathbb{R}} \sum_i \int_0^L F_i^1\, p_i + F_i^2\, q_i\, dxdt.
\end{aligned}$$

The boundary terms can be written

$$-[a(h)p\,r + b(h)q\,s]_{iL} =^t \left(p_i^\pm, q_i^\pm\right) A \left(r_i^\pm, s_i^\pm\right),$$

with $A = \mathrm{diag}(-a_+, a_-, -b_+, b_-)$. We write the Rankine–Hugoniot conditions (21) in the form

$$\begin{pmatrix} \dot{\underline{\varepsilon}}_i \\ 0 \end{pmatrix} = \underline{B}\left(r_i^\pm, s_i^\pm\right) + \begin{pmatrix} a_{1,i} \\ a_{2,i} \end{pmatrix},$$

with $\underline{\varepsilon}_i = 2[h]\varepsilon_i$ and $\underline{B}$ is defined by

$$\underline{B} = \begin{pmatrix} h_+ - h_+^{-\frac{1}{2}} & -h_- + h_-^{-\frac{1}{2}} & h_+ + h_+^{\frac{1}{2}} & -h_- - h_-^{-\frac{1}{2}} \\ 2 - h_+^{\frac{3}{2}} - h_+^{-\frac{3}{2}} & h_-^{\frac{3}{2}} + h_-^{-\frac{3}{2}} - 2 & h_+^{\frac{3}{2}} + h_+^{-\frac{3}{2}} + 2 & -h_-^{\frac{3}{2}} - h_-^{-\frac{3}{2}} - 2 \end{pmatrix}.$$

We decompose $\underline{B}$ in the form $\underline{B} = (A_1\ A_2) = \binom{B_1}{B_2}$, with $A_i \in M_2(\mathbb{R})$ and ${}^t B_i \in \mathbb{R}^4$. It is a straightforward computation to prove that

$$\det(A_1) = \left(h_- - h_-^{-\frac{1}{2}}\right)\left(h_+ - h_+^{-\frac{1}{2}}\right)\left(\left(h_+^{\frac{1}{2}} - h_+^{-1}\right) - \left(h_-^{\frac{1}{2}} - h_-^{-1}\right)\right) \neq 0,$$

provided that $h_+ < 1 < h_-$. This condition is clearly satisfied for roll-wave solutions (otherwise, there is no roll-wave). Thus, $\mathrm{Ker}\,\underline{B}$ is two-dimensional ($2D$) and possesses a basis of vectors: $\langle {}^t B_1^\perp, {}^t B_2^\perp \rangle$. We can apply Lemma 9.4, p. 255 in [1]: there exists $N_1 \in M_{2,4}(\mathbb{R})$ so that $\mathbb{R}^4 = \mathrm{Ker}\underline{B} \oplus \mathrm{Ker}N_1$ and there exist $C, N_2 \in M_{2,4}(\mathbb{R})$ so that $\mathbb{R}^4 = \mathrm{Ker}C \oplus \mathrm{Ker}N_2$ and $A =^t N_2\,\underline{B} +^t C\,N_1$. As a consequence, the boundary terms can be written in the form

$$\begin{aligned}
&\int_{\mathbb{R}} \sum_i {}^t\left(p_i^\pm, q_i^\pm\right) A \left(r_i^\pm, s_i^\pm\right) \\
&\qquad = \int_{\mathbb{R}} \sum_i {}^t\left(N_2\left(p_i^\pm, q_i^\pm\right)\right) \underline{B}\left(r_i^\pm, s_i^\pm\right) + {}^t\left(C\left(p_i^\pm, q_i^\pm\right)\right) N_1\left(r_i^\pm, s_i^\pm\right) \\
&\qquad = \int_{\mathbb{R}} \sum_i {}^t\left(C\left(p_i^\pm, q_i^\pm\right)\right) N_1\left(r_i^\pm, s_i^\pm\right) + {}^t\left(N_2\left(p_i^\pm, q_i^\pm\right)\right) \begin{pmatrix} a_{1,i} - \dot{\varepsilon}_i \\ a_{2,i} \end{pmatrix} \\
&\qquad = \int_{\mathbb{R}} \sum_i {}^t\left(N_2\left(p_i^\pm, q_i^\pm\right)\right) \begin{pmatrix} a_{1,i} \\ a_{2,i} \end{pmatrix} dt + \partial_t\left(N_2\left(p_i^\pm, q_i^\pm\right)_1\right) \varepsilon_i\, dt \\
&\qquad\quad + \int_{\mathbb{R}} \sum_i {}^t\left(C\left(p_i^\pm, q_i^\pm\right)\right) N_1\left(r_i^\pm, s_i^\pm\right) dt.
\end{aligned}$$

The adjoint problem of the linearized equations around a roll-wave is written in the form

$$\begin{array}{l} -\partial_t\, p_i - \partial_x(a(h)p_i) = 0, \\ -\partial_t\, q_i - \partial_x(b(h)q_i) = 0, \end{array} \quad \forall (x,i) \in (0,L) \times \mathbb{Z},$$

with the boundary conditions

$$\partial_t\left(N_2\left(p_i^\pm, q_i^\pm\right)_1\right) = 0, \quad C\left(p_i^\pm, q_i^\pm\right) = 0 \quad \forall i \in \mathbb{Z}.$$

In what follows, we compute energy estimates for solutions of the linear problem:

$$\begin{array}{l} -\partial_t\, p_i - \partial_x\left(a(h)p_i\right) = F_i^1, \\ -\partial_t\, q_i - \partial_x\left(b(h)q_i\right) = F_i^2 \end{array} \quad \forall i \in \mathbb{Z},\ x \in (0,L), \tag{31}$$

with the boundary equations

$$\partial_t \left(N_2 \left(p_i^\pm, q_i^\pm \right)_1 \right) = a_{1,i}, \quad C \left(p_i^\pm, q_i^\pm \right) = a_{2,i} \quad \forall i \in \mathbb{Z}. \tag{32}$$

We multiply the first equation of (31) by p_i and the second by q_i:

$$\begin{aligned} -\partial_t \int_{\mathbb{R}} p^2 + q^2 \, dx + \sum_i \left[a(h)\, p^2 + b(h)\, q^2 \right]_{iL} &= 2 \int_{\mathbb{R}} F^1\, p + F^2\, q \, dxdt \\ &+ \int_{\mathbb{R}} \partial_x a(h)\, p^2 + \partial_x b(h)\, q^2 \, dx, \end{aligned} \tag{33}$$

where (p, q) is the function $(p, q) = \sum_{i \in \mathbb{Z}} (p_i, q_i) 1_{(iL,(i+1)L}$. We note $\tilde{a}_{1,i}$, the unique exponentially decreasing solution of $\partial_t \tilde{a}_{1,i} = a_{1,i}$. The boundary conditions (32) now read

$$\begin{aligned} N_2 \left(p_i^\pm, q_i^\pm \right)_1 &= \tilde{a}_{1,i}, \\ N_2 \left(p_i^\pm, q_i^\pm \right)_2 &= N_2 \left(p_i^\pm, q_i^\pm \right)_2 \ \forall\, i \in \mathbb{Z}, \\ C \left(p_i^\pm, q_i^\pm \right) &= a_{2,i}. \end{aligned} \tag{34}$$

We recall that $A =^t N_2 \underline{B} +^t C_1$, and (34) is written

$$A \left(p_i^\pm, q_i^\pm \right) = \left(N_2 \left(p_i^\pm, q_i^\pm \right)_2 \right)^t B_2 + \tilde{a}_{1,i}\, {}^t B_1 +^t N_1\, a_{i,2}.$$

The matrix A is invertible with a bounded inverse so that

$$\left| \left(p_i^\pm, q_i^\pm \right) \right|^2 \leq C \left(\left| N_2 \left(p_i^\pm, q_i^\pm \right)_2 \right|^2 + |(\tilde{a}_{1,i}, 0, a_{2,i})|^2 \right).$$

As a consequence, there remains to estimate the term $N_2(p_i^\pm, q_i^\pm)_2$. For that purpose, we write the boundary terms in (33) in the form

$$\begin{aligned} -\mathcal{Q} \left(p_i^\pm, q_i^\pm \right) = &- \left| N_2 \left(p_i^\pm, q_i^\pm \right)_2 \right|^2 \, B_2\, A^{-1}\, {}^t B_2 \\ &- 2 N_2 \left(p_i^\pm, q_i^\pm \right)_2 \left(A^{-1}\, {}^t B_2; A^{-1} \left({}^t \underline{B}, {}^t \, N_1 \right) \begin{pmatrix} \tilde{a}_{1,i} \\ 0 \\ a_{2,i} \end{pmatrix} \right) \\ &- \left(A^{-1} \left({}^t \underline{B}, {}^t \, N_1 \right) \begin{pmatrix} \tilde{a}_{1,i} \\ 0 \\ a_{2,i} \end{pmatrix}; \left({}^t \underline{B}, {}^t \, N_1 \right) \begin{pmatrix} \tilde{a}_{1,i} \\ 0 \\ a_{2,i} \end{pmatrix} \right). \end{aligned}$$

Then, we deduce from (33) that

$$\begin{aligned} -\partial_t \int_{\mathbb{R}} p^2 + q^2 - \sum_i \left| N_2 \left(p_i^\pm, q_i^\pm \right)_2 \right|^2 \, B_2\, A^{-1}\, {}^t B_2 \\ = 2 \int_{\mathbb{R}} F^1\, p + F^2\, q + \mathcal{O} \left(\sum_i \tilde{a}_{1,i}^2 + |a_{2,i}|^2 \right) \\ + \mathcal{O} \left(\sum_i \left| N_2 \left(p_i^\pm, q_i^\pm \right)_2 \right| \, |(\tilde{a}_{1,i}, 0, a_{2,i})| \right). \end{aligned}$$

Next, we show that $B_2\, A^{-1}\, {}^t B_2 < 0$. It is an easy commputation to prove that

$$B_2\, A^{-1}\, {}^t B_2 = 2 \left(\left(3h_+ + \frac{1}{h_+^2} \right) - \left(3h_- + \frac{1}{h_-^2} \right) \right) < 0.$$

The inequality is satisfied, since we have already assumed that the Froude number F is such that $0 < F < F_c$, and thus

$$3h_- + \frac{1}{h_-} > 3h_+ + \frac{1}{h_+}$$

(see the proof of Proposition 1) so that the boundary conditions are maximal dissipative and obtain energy estimates on the linearized problem. We work in weighted spaces in time: integrating on $t \in \mathbb{R}$ and using Young's inequalities, we get for $\gamma \geq \gamma_0$ (γ_0 depending on $\|(R,S)\|_{W^{1,\infty}(0,L)}$) that

$$\begin{aligned} &\gamma \int_{\mathbb{R}} e^{2\gamma t} \|(p,q)\|_2^2 \, dt + \kappa \int_{\mathbb{R}} e^{2\gamma t} \sum_i \left| \left(p_i^{\pm}, q_i^{\pm}\right) \right|^2 \\ &\leq C \left(\frac{1}{\gamma} \int_{\mathbb{R}} e^{2\gamma t} \left\| \left(F^1, F^2\right) \right\|^2 \, dt + \int_{\mathbb{R}} e^{2\gamma t} \sum_i |a_{1,i}|^2 + |a_{2,i}|^2 \right), \end{aligned} \tag{35}$$

with $\kappa, C > 0$ constant.

3.3.2. The "approximate" adjoint problem. In this section, we compute the adjoint problem of the linearized problem around an approximate solution and a priori estimates on that linear problem. Recall that the linearized equations around an approximate solution are given by

$$\begin{aligned} &\partial_t r_i + \frac{L}{X_{i+1} - X_i} \Lambda_1^i \partial_x r_i = F_i \\ &\partial_t s_i + \frac{L}{X_{i+1} - X_i} \Lambda_2^i \partial_x s_i = G_i \end{aligned} \qquad \forall x \in (0,L), i \in \mathbb{Z}, \tag{36}$$

with $\Lambda_k^i = \lambda_k(R^i, S^i) - (\dot{X}_i + \frac{x}{L}(\dot{X}_{i+1} - \dot{X}_i))$. We multiply the first (resp. second) equation of (36) by p_i (resp. q_i), integrate on $(0,\, L) \times \mathbb{R}$, and sum on $i \in \mathbb{Z}$:

$$\begin{aligned} \int_{\mathbb{R}} \sum_i \int_0^L F_i \, p_i + G_i \, q_i = &- \int_{\mathbb{R}} \sum_i \int_0^L \partial_t \, p_i \, r_i + \partial_t \, q_i \, s_i \\ &- \int_{\mathbb{R}} \sum_i \int_0^L \partial_x \big(\Lambda_1^i \, p_i\big) r_i + \partial_x \big(\Lambda_2^i \, q_i\big) s_i \\ &+ \int_{\mathbb{R}} \sum_i {}^t \left(p_i^{\pm}, q_i^{\pm}\right) A_i \left(r_i^{\pm}, s_i^{\pm}\right), \end{aligned} \tag{37}$$

with

$$\begin{aligned} A_i = \operatorname{diag}\Bigg(&- \frac{L}{X_{i+1} - X_i} \left(\lambda_1 \left(R_i^+, S_i^+\right) - \dot{X}_i\right), \frac{L}{X_i - X_{i-1}} \left(\lambda_1(R_i^-, S_i^-) - \dot{X}_i\right), \\ &- \frac{L}{X_{i+1} - X_i} \left(\lambda_2 \left(R_i^+, S_i^+\right) - \dot{X}_i\right), \frac{L}{X_i - X_{i-1}} \left(\lambda_2 \left(R_i^-, S_i^-\right) - \dot{X}_i\right) \Bigg). \end{aligned}$$

It is a straightforward computation to prove that $A_i = A + \mathcal{O}(\eta)$, where η is given by

$$\eta = \max_{i \in \mathbb{Z}} \left(\left| \dot{X}_i - x \right|_\infty, |X_{i+1} - X_i - L|_\infty, |(R - R_i, S - S_i)|_\infty \right).$$

We consider the boundary terms in (37): recall that, under the assumption of Proposition 2, the linearized Rankine–Hugoniot conditions can be written in the form

$$\begin{pmatrix} \dot{\varepsilon}_i \\ 0 \end{pmatrix} = \underline{B}_i \left(r_i^{\pm}, s_i^{\pm}\right) + \begin{pmatrix} a_{1,i} \\ a_{2,i} \end{pmatrix}.$$

We can prove easily that $\underline{B}_i = \underline{B} + \mathcal{O}(\eta)$. Following the computation of the dual problem in the previous section, we decompose $\underline{B}_i$ in the form $\underline{B}_i = (A_{1,i}\ A_{2,i}) = \binom{B_{1,i}}{B_{2,i}}$, with $A_{k,i} \in M_2(\mathbb{R})$ and ${}^t B_{k,i} \in \mathbb{R}^4$. It is a straightforward computation to prove that, for η sufficiently small,

$$\det(A_{1,i}) = \det(A_1) + \mathcal{O}(\eta) \neq 0.$$

Thus, $\mathrm{Ker}\,\underline{B}_i$ is $2D$ and possesses a basis of vectors: $\langle {}^t B_{1,i}^{\perp}, {}^t B_{2,i}^{\perp} \rangle$. We can apply Lemma 9.4, p. 255 in [1]: there exists $N_{1,i} \in M_{2,4}(\mathbb{R})$ so that $\mathbb{R}^4 = \mathrm{Ker}\underline{B}_i \oplus \mathrm{Ker}N_{1,i}$ and there exist $C_i\,, N_{2,i} \in M_{2,4}(\mathbb{R})$ so that $\mathbb{R}^4 = \mathrm{Ker}C_i \oplus \mathrm{Ker}N_{2,i}$ and $A_i = {}^t N_{2,i}\,\underline{B}_i + {}^t C_i\,N_{1,i}$. At this stage, the formulation of the adjoint problem is straightforward, and one finds

$$\begin{aligned} &-\partial_t p_i - \frac{L}{X_{i+1} - X_i}\partial_x \left(\Lambda_i^1\, p_i\right) = 0, \\ &-\partial_t q_i - \frac{L}{X_{i+1} - X_i}\partial_x \left(\Lambda_i^2\, q_i\right) = 0 \end{aligned} \qquad \forall (x,i) \in (0,\, L) \times \mathbb{Z},$$

with the boundary conditions

$$\partial_t \left(N_{2,i}\left(p_i^{\pm}, q_i^{\pm}\right)_1\right) = 0 \quad C_i\left(p_i^{\pm}, q_i^{\pm}\right) = 0 \quad \forall i \in \mathbb{Z}.$$

Next we consider the linear problem

$$\begin{aligned} &-\partial_t p_i - \frac{L}{X_{i+1} - X_i}\partial_x (\Lambda_i^1\, p_i) = F_i^1, \\ &-\partial_t q_i - \frac{L}{X_{i+1} - X_i}\partial_x (\Lambda_i^2\, q_i) = F_i^2 \end{aligned} \qquad \forall (x,i) \in (0,\, L) \times \mathbb{Z}, \tag{38}$$

with the boundary conditions

$$\partial_t \left(N_{2,i}(p_i^{\pm}, q_i^{\pm})_1\right) = a_{1,i} \quad C_i(p_i^{\pm}, q_i^{\pm}) = a_{2,i} \quad \forall i \in \mathbb{Z}. \tag{39}$$

Following the method of proof of the energy estimates for the adjoint problem associated to the linearization of shallow water equations around roll-waves, one can prove the following result.

PROPOSITION 3. *Under the hypothesis $0 < F < F_c$, there exists $\eta > 0$ so that for all $m > 0$, there exists a constant C and γ_0 depending on m and for any Lipschitz continuous and bounded function $(R_i, S_i, X_i)_{i\in\mathbb{Z}}$ satisfying*

$$\begin{aligned} &\max_{i\in\mathbb{Z}} \left(\left\| \dot{X}_i - c \right\|_{\infty} + \|X_{i+1} - X_i - L\|_{\infty} + \|R_i - R, S_i - S\|_{\infty} \right) \leq \eta, \\ &\max_{i\in\mathbb{Z}} \left(|\partial_x (R_i, S_i)|_{\infty} + \left| \dot{X}_i \right|_{\infty} \right) \leq m, \end{aligned}$$

and for all $\gamma \geq \gamma_0$, $(F^1, F^2) \in \mathcal{D}(\mathbb{R}/\{iL\} \times \mathbb{R})$, $(a_1, a_2) \in l^2(\mathbb{Z}, \mathcal{D}(\mathbb{R}))$, then a solution $(p_i, q_i)_{i\in\mathbb{Z}}$ of (38), (39) satisfies

$$\begin{aligned} &\gamma \|e^{\gamma t}(p,q)\|^2_{L^2(\mathbb{R}/\{iL\}\times\mathbb{R})} + \|e^{-\gamma t}(p_i^{\pm}, q_i^{\pm})\|^2_{l^2(\mathbb{Z}, L^2(\mathbb{R}))} \\ &\quad \leq C \left(\frac{1}{\gamma} \left\| e^{\gamma t} \left(F^1, F^2\right) \right\|^2_{L^2(\mathbb{R}/\{iL\}\times\mathbb{R})} + \|e^{\gamma t}(a_1, a_2)\|_{l^2(\mathbb{Z}, L^2(\mathbb{R}))} \right). \end{aligned}$$

As a conclusion, we have obtained energy estimates for the linearized problem around an approximate solution, we have formulated an adjoint problem, and we computed estimates on that problem. This is the first step towards the well posedness of the linearized equations.

3.4. Well posedness of the linear differential boundary value problem. In what follows, we prove the well posedness of the linear boundary value problem without any initial condition: The method of proof is based on a weak formulation of the linearized equations that define an adjoint problem. Using the Riesz representation theorem, we prove the existence of a weak solution that is proved to enjoy more regularity using a mollifying operator in the "transverse" t-direction. At this step, there is a difference with the shock case. In that latter case, the convection matrix in front of the x-derivatives is nonsingular, and we easily get estimates on those derivatives. In the case of roll-waves, the matrix that is in front of the x-derivatives is *singular* in the neighborhood of the sonic point. As a consequence, we also need a regularization step in the interior equations in the neighborhood of the sonic point. In what follows, the presentation of the well posedness results for a linearized problem follows the presentation made in [1] for the persistence of shocks: in the proofs of those results, we shall emphasize the difference between the shock case and the roll-wave case.

In order to simplify the notations, we introduce $\tilde{\Lambda}_i^k = \frac{L}{X_{i+1}-X_i}\Lambda_i^k$ and denote $\Omega = \mathbb{R} \setminus \{iL,\, i \in \mathbb{Z}\} \times \mathbb{R}$. We shall also note $L^2_\gamma(\Omega)$ the Hilbert space of functions f so that $e^{-\gamma t} f \in L^2(\Omega)$. We first prove the following result (see Theorem 12.4, p. 360 in [1] for the shock case).

PROPOSITION 4. *There exists $\eta > 0$ and $\gamma_0 = \gamma_0(m)$ such that if*

$$\max_{i\in\mathbb{Z}} \left(\left\|\dot{X}_i - c\right\|_\infty, \|X_{i+1} - X_i - L\|_\infty, \left\|R_i - \overline{R}\right\|_\infty, \left\|S_i - \overline{S}\right\|_\infty \right) \leq \eta,$$
$$\max_{i\in\mathbb{Z}} \left(\left\|\dot{X}_{i+1} - \dot{X}_i\right\|_\infty, \|R_i\|_{W^{1,\infty}}, \|S_i\|_{W^{1,\infty}} \right) \leq m,$$

then for all $\gamma \geq \gamma_0$, $f \in L^2_\gamma(\Omega)^2$, $g \in l^2(\mathbb{Z}, L^2_\gamma(\mathbb{R}))^2$, there is one and only one $((r,s),\varepsilon) \in L^2_\gamma(\Omega)^2 \times l^2(\mathbb{Z}, H^{\frac{1}{2}}_\gamma(\mathbb{R}))$ such that

$$\mathcal{L}(r,s) = \begin{cases} \partial_t r_i + \widetilde{\Lambda}_i^1 \partial_x r_i = f_i^1, \\ \partial_t s_i + \widetilde{\Lambda}_i^2 \partial_x s_i = f_i^2 \end{cases} \qquad \forall\, i \in \mathbb{Z}, (x,t) \in]0, L[\times\mathbb{R}$$

and satisfies the boundary conditions

$$\begin{pmatrix} \dot{\varepsilon}_i \\ 0 \end{pmatrix} = \underline{B}_i \left(r_i^\pm, s_i^\pm \right) + \begin{pmatrix} g_i^1 \\ g_i^2 \end{pmatrix} \ \forall\, i \in \mathbb{Z}, t \in \mathbb{R}.$$

Furthermore, $(r_i^\pm, s_i^\pm)_{i\in\mathbb{Z}} \in l^2(\mathbb{Z}, L^2_\gamma(\mathbb{R})^4)$ and $\varepsilon \in l^2(\mathbb{Z}, H^1_\gamma(\mathbb{R}))$, and we have the estimate

$$\gamma \|(r,s)\|^2_{L^2_\gamma(\Omega)^2} + \left\| \left(r^\pm, s^\pm\right) \right\|^2_{l^2(\mathbb{Z}, L^2_\gamma(\mathbb{R})^4)} + \|\varepsilon\|^2_{l^2(\mathbb{Z}, H^1_\gamma(\mathbb{R}))}$$
$$\leq C \left(\frac{1}{\gamma} \|f\|^2_{L^2_\gamma(\Omega)} + \|g\|^2_{l^2(\mathbb{Z}, L^2_\gamma(\mathbb{R}))} \right).$$

Proof. Let us first prove the existence of a weak solution with a duality argument. With the notations of Proposition 3, denote $\mathcal{E}$ the space of functions $(p,q) \in \mathcal{D}(\Omega)^2$ so that

$$\partial_t N^t_{2,i}(p_i^\pm, q_i^\pm)_1 - \gamma\, N^t_{2,i}(p_i^\pm, q_i^\pm)_1 = 0, \quad C_i(p_i^\pm, q_i^\pm) = 0.$$

We define the adjoint operator $\mathcal{L}_\gamma^*$ of $\mathcal{L}_\gamma = \mathcal{L} + \gamma\, Id$ as

$$\mathcal{L}_\gamma^*(p,q) = \begin{cases} -\partial_t\, p_i - \partial_x \left(\widetilde{\Lambda}_i^1\, p_i\right) + \gamma\, p_i, \\ -\partial_t\, q_i - \partial_x \left(\widetilde{\Lambda}_i^2\, q_i\right) + \gamma\, q_i. \end{cases}$$

Let $(f,g) \in L^2(\Omega)^2 \times l^2(\mathbb{Z}, L^2(\mathbb{R}))$ and define a bounded linear form l on $\mathcal{L}_\gamma^*\mathcal{E}$ as

$$l\Big(\mathcal{L}_\gamma^*(p,q)\Big) = \sum_{i\in\mathbb{Z}} \int_{\mathbb{R}} \int_0^L p_i\, f_i^1 + q_i\, f_i^2\, dx + \left(N_{2,i}(p_i^\pm, q_i^\pm)\right) \begin{pmatrix} g_i^1 \\ g_i^2 \end{pmatrix} dt.$$

We deduce from Proposition 3 on the adjoint problem that $\forall (p,q) \in \mathcal{E}$,

$$\begin{aligned} \left| l\left(\mathcal{L}_\gamma^*(p,q)\right)\right| &\le C(\|f\|, \|g\|) \left(\|(p,q)\|_{L^2(\Omega)} + \|(p^\pm, q^\pm)\|_{l^2(\mathbb{Z}, L^2(\mathbb{R}))^4}\right) \\ &\le C\left(\frac{1}{\gamma}\|f\| + \frac{1}{\sqrt{\gamma}}\|g\|\right) \|\mathcal{L}_\gamma^*(p,q)\|_{L^2(\Omega)}. \end{aligned}$$

Therefore, by the Hahn–Banach theorem, the linear form l extends to a continuous form on $L^2(\Omega)^2$ and using the Riesz representation theorem, there exists $(r,s) \in L^2(\Omega)^2$ so that

$$l\left(\mathcal{L}_\gamma^*(p,q)\right) = \sum_{i\in\mathbb{Z}} \int_{\mathbb{R}} \int_0^L (r_i, s_i)^t\, \mathcal{L}_\gamma^*(p_i, q_i) dx dt \quad \forall (p,q) \in \mathcal{E}.$$

Then, by the definition of l, we find that

$$\sum_{i\in\mathbb{Z}} \int_{\mathbb{R}} \int_0^L (r_i, s_i)^t\, \mathcal{L}_\gamma^*(p_i, q_i) - \left(f_i^1, f_i^2\right)^t (p_i, q_i) dx dt = 0 \quad \forall (p,q) \in \mathcal{D}(\Omega)^2,$$

and $\mathcal{L}_\gamma(r,s) = (f^1, f^2)$ in the sense of distribution. For all $(p,q) \in \mathcal{E}$, one can prove that

$$\begin{aligned} \sum_{i\in\mathbb{Z}} \int_{\mathbb{R}} \int_0^L (r_i, s_i)^t . \mathcal{L}_\gamma^*(p_i, q_i) - \mathcal{L}_\gamma(r_i, s_i)^t . \begin{pmatrix} p_i \\ q_i \end{pmatrix} dx\, dt & \\ = \sum_{i\in\mathbb{Z}} \int_{\mathbb{R}} N_{2,i}(p_i^\pm, q_i^\pm)^t . \begin{pmatrix} g_i^1 \\ g_i^2 \end{pmatrix} dt & \\ = \sum_{i\in\mathbb{Z}} \int_{\mathbb{R}} N_{2,i}(p_i^\pm, q_i^\pm)^t . \underline{B}_i \begin{pmatrix} r_i^\pm \\ s_i^\pm \end{pmatrix} dt. & \end{aligned}$$

As a consequence, we find that

$$\sum_{i\in\mathbb{Z}} \int_{\mathbb{R}} N_{2,i}(p_i^\pm, q_i^\pm)^t \left(\underline{B}_i \left(r_i^\pm, s_i^\pm\right) - \begin{pmatrix} g_i^1 \\ g_i^2 \end{pmatrix}\right) dt = 0 \quad \forall (p,q) \in \mathcal{E}. \tag{40}$$

By density, this property holds true for all $(p,q) \in H^1(\Omega)$ such that

$$C_i^t(p_i^\pm, q_i^\pm) = 0, \quad \partial_t\, N_{2,i}(p_i^\pm, q_i^\pm)_1 - \gamma\, N_{2,i}(p_i^\pm, q_i^\pm)_1 = 0.$$

Let $(\theta^1,\theta^2)\in l^2(\mathbb{Z},H^{\frac12}(\mathbb{R})^2)$ and $(p_i^\pm,q_i^\pm)_{i\in\mathbb{Z}}\in l^2(\mathbb{Z},H^{\frac12}(\mathbb{R})^2)$ so that

$$N_{2,i}^t(p_i^\pm,q_i^\pm)=\begin{pmatrix}\theta_i^1\\ \theta_i^2\end{pmatrix},\quad C_i^t(p_i^\pm,q_i^\pm)=\begin{pmatrix}0\\0\end{pmatrix}\quad\forall i\in\mathbb{Z},t\in\mathbb{R}.$$

The sequence $(p_i^\pm,q_i^\pm)_{i\in\mathbb{Z}}$ exists, is unique, and lies in $l^2(\mathbb{Z},H^{\frac12}(\mathbb{R})^2)$, since the matrix $(N_{2,i},C_i)$ is invertible under the assumption of Proposition 3. Now with a standard trace-lifting argument, choose $(p,q)\in,H^1(\Omega)$ so that $p(iL)^\pm=p_i^\pm,q(iL)^\pm=q_i^\pm$. Then we deduce from (40) that for all $\theta\in l^2(\mathbb{Z},H^{\frac12}(\mathbb{R})^2)$ such that $\frac{d\theta_i^1}{dt}-\gamma\theta_i^1=0$,

$$\sum_{i\in\mathbb{Z}}\left\langle\begin{pmatrix}\theta_i^1\\ \theta_i^2\end{pmatrix},\underline{B}_i\left(r_i^\pm,s_i^\pm\right)-\begin{pmatrix}g_i^1\\ g_i^2\end{pmatrix}\right\rangle_{\left(H^{\frac12}(\mathbb{R}),H^{-\frac12}(\mathbb{R})\right)}=0.$$

Thus, there exists $\varepsilon\in l^2(\mathbb{Z},H^{\frac12}(\mathbb{R}))$ so that

$$\begin{pmatrix}\dot\varepsilon_i+\gamma\varepsilon_i\\0\end{pmatrix}=\underline{B}_i\left(r_i^\pm,s_i^\pm\right)+\begin{pmatrix}g_i^1\\g_i^2\end{pmatrix}\ \forall i\in\mathbb{Z}.$$

Let us now prove that $(r,s),\varepsilon$ is a strong solution. This is just a matter of smoothing the solution and using the estimates of the first part. We consider a mollifying operator R_δ^t in the t-direction, and we define a regularization (r^δ,s^δ), ε^δ of (r,s,ε). Let us first deal with the equation on s. Denote $s_i^\delta=R_\delta^t\,s_i$, we easily obtain

$$\partial_t\,s_i^\delta+\widetilde{\Lambda}_i^2\partial_x\,s_i^\delta=\widetilde{\Lambda}_i^2\left(R_\delta^t\left(\frac{f_i^2}{\widetilde{\Lambda}_i^2}\right)+\left[\frac{1}{\widetilde{\Lambda}_i^2}\partial_t,R_\delta^t\right]s_i\right).$$

Due to the presence of a *sonic point* $x_i(t)$ so that $\widetilde{\Lambda}_1^i(x_i(t),t)=0$, we cannot directly do the same regularization on r_i. In that case, we do not obtain any information on the derivatives of r_i in the x-direction. Nevertheless, under the assumptions of Proposition 3, there exists $L_1>0$ and a constant $C(m)$ so that for all $i\in\mathbb{Z}$, $iL+2L_1\leq x_i(t)\leq(i+1)L-2L_1$ and $|\widetilde{\Lambda}_1^i(x,t)|^{-1}\leq C(m)\ \forall\ x\in[0,L_1]\bigcup[L-L_1,L]$. Let $\overline{\phi}\in,C_c^\infty(0,L)$ so that $\overline{\phi}=1$ on $[2L_1,\,L-2L_1]$ and $\overline{\phi}=0$ on $[0,L_1]\bigcup[L-L_1,L]$ and define $\phi=1-\overline{\phi}$. We multiply the equation on r_i by ϕ. We find

$$\partial_t(\phi\,r_i)+\widetilde{\Lambda}_i^1\partial_x(\phi\,r_i)=\phi f_i^1+\widetilde{\Lambda}_i^1\partial_x\phi\,r_i.$$

Then we regularize that equation the same way we did with the equation on s_i:

$$\partial_t\left(\phi\,r_i\right)^\delta+\widetilde{\Lambda}_i^1\partial_x\left(\phi\,r_i\right)^\delta=\widetilde{\Lambda}_i^1\left(R_\delta^t\left(\frac{\phi\,f_i^1}{\widetilde{\Lambda}_i^1}+\partial_x\phi\,r_i\right)+\left[\frac{1}{\widetilde{\Lambda}_i^1}\partial_t,R_\delta^t\right](\phi\,r_i)\right).$$

We multiply the equation on r_i by $(1-\phi)$ and choose a mollifying operator R_δ^x in the x direction with $\delta\leq L_1$:

$$\begin{aligned}\partial_t\,R_\delta^x((1-\phi)r_i)+\widetilde{\Lambda}_i^1\partial_x\,R_\delta^x((1-\phi)r_i)&=R_\delta^x\Big((1-\phi)f_1^i-\widetilde{\Lambda}_i^1\partial_x\phi\,r_1^i\Big)\\&\quad+\left[\widetilde{\Lambda}_i^1\partial_x,R_\delta^x\right]((1-\phi)r_i).\end{aligned}$$

We define r^δ so that $r^\delta_{|[iL,(i+1)L]}=R_\delta^t(\phi\,r_i)+R_\delta^x((1-\phi)r_i)$. By construction, both r^δ and s^δ lie in $H^1(\Omega)$. On the other hand, we find that r^δ,s^δ, and ε^δ satisfy the boundary conditions

$$\begin{pmatrix}\dot\varepsilon_i^\delta\\0\end{pmatrix}=\underline{B}_i\left(r_i^{\delta,\pm},s_i^{\delta,\pm}\right)+\begin{pmatrix}R_\delta^t\,g_i^1\\R_\delta^t\,g_i^2\end{pmatrix}+\left[\underline{B}_i,R_\delta^t\right]\left(r_i^\pm,s_i^\pm\right).$$

We can apply Proposition 2 as follows:

$$\begin{aligned}\gamma\left\| \left(r^\delta, s^\delta\right)\right\|^2_{L^2_\gamma(\Omega)^2} + \left\| \left(r^{\delta,\pm}, s^{\delta,\pm}\right)\right\|^2_{l^2(\mathbb{Z},L^2_\gamma(\mathbb{R})^4)} + \left\|\varepsilon^\delta\right\|^2_{l^2(\mathbb{Z},H^1_\gamma(\mathbb{R}))}\\ \leq C\left(\frac{1}{\gamma}\left\|f^\delta\right\|^2_{L^2_\gamma(\Omega)} + \left\|g^\delta\right\|^2_{l^2(\mathbb{Z},L^2_\gamma(\mathbb{R}))}\right),\end{aligned} \tag{41}$$

with

$$\begin{aligned}f^\delta_{i,1} &= R^x_\delta\left((1-\phi)f^1_i - \widetilde{\Lambda}^1_i\partial_x\phi\, r_i\right) + \left[\widetilde{\Lambda}^1_i\partial_x, R^x_\delta\right]((1-\phi)r_i)\\ &\quad + \widetilde{\Lambda}^1_i\left(R^t_\delta\left(\frac{\phi\, f^1_i}{\widetilde{\Lambda}^1_i} + \partial_x\phi\, r_i\right) + \left[\frac{1}{\widetilde{\Lambda}^1_i}\partial_t, R^t_\delta\right](\phi\, r_i)\right),\\ f^\delta_{i,2} &= \widetilde{\Lambda}^2_i\left(R^t_\delta\left(\frac{f^2_i}{\widetilde{\Lambda}^2_i}\right) + \left[\frac{1}{\widetilde{\Lambda}^2_i}\partial_t, R^t_\delta\right](s)\right)\\ g^\delta_i &= \begin{pmatrix} R^t_\delta\, g^1_i\\ R^t_\delta\, g^2_i\end{pmatrix} + \left[\underline{B}_i, R^t_\delta\right]\left(r^\pm_i, s^\pm_i\right).\end{aligned}$$

One can prove the following convergence properties (see [1] for more details):

$$\lim_{\delta\to 0}\left\|f^\delta - f\right\|_{L^2_\gamma(\Omega)} + \left\|g^\delta - g\right\|_{l^2(\mathbb{Z},L^2_\gamma(\mathbb{R}))} = 0.$$

By linearity of the equations, these estimates also apply to

$$r^\delta - r^{\delta'}, \quad s^\delta - s^{\delta'}, \quad \varepsilon^\delta - \varepsilon^{\delta'}.$$

Together with the convergence properties, this shows that $r^{\delta,\pm}, s^{\delta,\pm}$, and ε^δ are, respectively, Cauchy sequences in $l^2(\mathbb{Z}, L^2_\gamma(\mathbb{R})^4)$ and $l^2(\mathbb{Z}, H^1_\gamma(\mathbb{R}))$. By uniqueness of the limit in the sense of distribution, the sequences $(r^{\delta,\pm}, s^{\delta,\pm})$ and ε^δ converge to $r^\pm, s^\pm$, and ε in these norms. Then passing to the limit $\delta \to 0$ in (41), one finds

$$\begin{aligned}\gamma\|(r,s)\|^2_{L^2_\gamma(\Omega)^2} + \left\| \left(r^\pm, s^\pm\right)\right\|^2_{l^2(\mathbb{Z},L^2_\gamma(\mathbb{R})^4)} + \|\varepsilon\|^2_{l^2(\mathbb{Z},H^1_\gamma(\mathbb{R}))}\\ \leq C\left(\frac{1}{\gamma}\|f\|^2_{L^2_\gamma(\Omega)} + \|g\|^2_{l^2(\mathbb{Z},L^2_\gamma(\mathbb{R}))}\right).\end{aligned}$$

This completes the proof of the proposition. ∎

We can prove more regularity on the solution (r, s, ε) under the condition that f, g, and the coefficients R_i, S_i, X_i enjoy more regularity. Similarly to the L^2 estimates, we introduce weighted Sobolev norms:

$$\|u\|_{\mathcal{H}^k_\gamma} = \sum_{|\alpha|\leq k}\gamma^{2(k-|\alpha|)}\|e^{-\gamma t}u\|_{L^2}.$$

The regularity result on (r, s, ε) is formulated as follows (see Theorem 12.5, p. 364 in [1] in the shock case).

PROPOSITION 5. *Under the assumptions of Proposition* 4 *and that* R_i, S_i, X_i *enjoy more regularity,*

$$\begin{aligned}\|R - R_i, S - S_i\|_{l^2(\mathbb{Z},H^k((0,L)\times\mathbb{R}))} &\leq \mu,\\ \|R^\pm_i - R^\pm, S^\pm_i - S^\pm\|_{l^2(\mathbb{Z},H^k(\mathbb{R}))} &\leq \mu,\\ \|\dot{X}_i - c, X_{i+1} - X_i - L\|_{l^2(\mathbb{Z},H^k(\mathbb{R}))} &\leq \mu,\end{aligned}$$

with $k > 2$, the solution (r, s, ε) found in Proposition 4 is such that $(r, s) \in H^k(\Omega)$, $(r^\pm, s^\pm) \in l^2(\mathbb{Z}, H^k(\mathbb{R}))$, and $\varepsilon \in l^2(\mathbb{Z}, H^{k+1}(\mathbb{R}))$ and, for any $\gamma \geq \gamma_k(\mu) \geq 1$, satisfy the estimate

$$\begin{aligned} \gamma\|(r,s)\|^2_{\mathcal{H}^k_\gamma(\Omega)} + \left\|\left(r^\pm, s^\pm\right)\right\|^2_{l^2(\mathbb{Z},\mathcal{H}^k_\gamma(\mathbb{R}))} + \|\varepsilon\|^2_{l^2(\mathbb{Z},\mathcal{H}^{k+1}_\gamma(\Omega))} \\ \leq C(k,\mu)\left(\frac{1}{\gamma}\|f\|^2_{\mathcal{H}^k_\gamma(\Omega)} + \|g\|^2_{l^2(\mathbb{Z},\mathcal{H}^k_\gamma(\mathbb{R}))}\right). \end{aligned} \tag{42}$$

Proof. We prove that proposition in two steps. The first step is to extend the a priori L^2 estimates obtained in Proposition 2. For all (r, s) and ε compactly supported and smooth functions, the following estimate holds true:

$$\begin{aligned} \gamma\|(r,s)\|^2_{\mathcal{H}^k_\gamma(\Omega)} + \left\|\left(r^\pm, s^\pm\right)\right\|^2_{l^2(\mathbb{Z},\mathcal{H}^k_\gamma(\mathbb{R}))} + \|\varepsilon\|^2_{l^2(\mathbb{Z},\mathcal{H}^{k+1}_\gamma(\Omega))} \\ \leq C(k,\mu)\left(\frac{1}{\gamma}\|f\|^2_{\mathcal{H}^k_\gamma(\Omega)} + \|g\|^2_{l^2(\mathbb{Z},\mathcal{H}^k_\gamma(\mathbb{R}))}\right), \end{aligned} \tag{43}$$

where $f = \mathcal{L}(r, s)$ and g is defined by

$$g_i = \begin{pmatrix} \dot{\varepsilon}_i \\ 0 \end{pmatrix} - \underline{B}_i\left(r_i^\pm, s_i^\pm\right).$$

In order to prove that estimate, we decompose in the form $r_i = \phi\, r_i + (1-\phi) r_i$, with ϕ defined in Proposition 4. Deriving the equation satisfied by $\phi r_i, s_i$ in the t-direction and applying Proposition 2, we get estimates for derivatives in the t-direction. We recover the derivatives in the x-direction with

$$\partial_x(\phi r_i) = \frac{1}{\widetilde{\lambda}_i^1}\left(f_i^1 + \partial_x\phi r_i - \partial_t(\phi r_i)\right), \quad \partial_x s_i = \frac{1}{\widetilde{\Lambda}_i^2}\left(f_i^2 - \partial_t s_i\right).$$

The obtention of estimates for $(1-\phi)r_i$ is straightforward, since it is trace-free on the discontinuity points iL, $i \in \mathbb{Z}$.

Now we come back to the solution obtained in Proposition 4. The next step is to prove enough regularity on (r, s, ε) in order to apply the a priori estimate (43). We first consider that the coefficients $R_i - R, S_i - S, X_i - (ct + iL)$ are smooth and compactly supported: the regularity of (r, s, ε) is then obtained by induction (see [1, p. 366] for the details of the proof). We then construct a family $(r^\delta, s^\delta, \varepsilon^\delta)$ of compactly supported functions that converges to (r, s, ε) in Sobolev H^k_γ norms. The regularized functions $(r^\delta, s^\delta, \varepsilon^\delta)$ satisfy the a priori estimate (43) and passing to the limit $\delta \to 0$ as in Proposition 4, we prove that (r, s, ε) satisfies the energy estimates (42). For the more general case, $R_i - R, S_i - S, X_i - (ct + iL)$ lying in Sobolev spaces, it is just a matter of smoothing the coefficients and passing to the limit. This completes the proof of that proposition. ∎

Next, we consider the well posedness of the linearized equation with zero initial data. Let $T \in \mathbb{R}$, I_T the interval $I_T =]-\infty, T]$, and $\Omega_T = \mathbb{R} \setminus \{iL,\, i \in \mathbb{Z}\} \times I_T$. One can prove that, under the assumptions of Proposition 4 and for all $f \in L^2(\mathbb{R} \setminus \{iL, i \in \mathbb{Z}\} \times I_T, \mathbb{R}^2)$ and $g \in l^2(\mathbb{Z}, L^2(I_T)^2)$ such that $f_{|t<0} = 0, \quad g_{|t<0} = 0$, there exists a unique function $(r, s) \in L^2(\Omega_T)$ and $\varepsilon \in l^2(\mathbb{Z}, H^1(I_T))$ so that

$$\begin{aligned} &\partial_t\, r_i + \widetilde{\Lambda}_i^1 \partial_x\, r_i = f_i^1, \\ &\partial_t\, s_i + \widetilde{\Lambda}_i^2 \partial_x\, s_i = f_i^2 \end{aligned} \quad \forall\, i \in \mathbb{Z}, (x,t) \in]0, L[\times I_T,$$

and

$$\begin{pmatrix} \dot{\varepsilon}_i \\ 0 \end{pmatrix} = \underline{B}_i \left(r_i^{\pm}, s_i^{\pm}\right) + \begin{pmatrix} g_i^1 \\ g_i^2 \end{pmatrix} \quad \forall\, i \in \mathbb{Z}, t \in I_T.$$

Furthermore, we have $r_{|t<0} = s_{|t<0} = 0, \quad \varepsilon_{|t<0} = 0$, $(r_i^{\pm}, s_i^{\pm})_{i\in\mathbb{Z}}$ belongs $l^2(\mathbb{Z}, L^2_\gamma(I_T)^4)$, and the solution (r, s, ε) satisfies the following estimate,

$$\begin{aligned}\gamma\|(r,s)\|^2_{L^2_\gamma(\Omega_T)^2} + \left\|\left(r^{\pm}, s^{\pm}\right)\right\|^2_{l^2(\mathbb{Z},L^2_\gamma(I_T)^4)} + \|\varepsilon\|^2_{l^2(\mathbb{Z},H^1_\gamma(I_T))} \\ \leq C\left(\frac{1}{\gamma}\|f\|^2_{L^2_\gamma(\Omega_T)} + \|g\|^2_{l^2(\mathbb{Z},L^2_\gamma(I_T))}\right).\end{aligned}$$

In the case of smoother coefficients, the solution (r, s, ε) of the linearized problem with zero initial data enjoys more regularity. One can prove the following result that shall be useful in the proof of the well posedness of the Cauchy problem for the full nonlinear problem.

PROPOSITION 6. *Assume that the hypotheses of Proposition* 4 *are satisfied, and the coefficients enjoy the following regularity properties:*

$$\begin{aligned}&\|(R_i - R, S_i - S)\|_{l^2(\mathbb{Z},H^k((0,L)\times I_T))} \leq \mu,\\ &\left\|\left(R_i^{\pm} - R^{\pm}, S_i^{\pm} - S^{\pm}\right)\right\|_{l^2(\mathbb{Z},H^k(I_T))} \leq \mu,\\ &\left\|\dot{X}_i - c, X_{i+1} - X_i - L\right\|_{l^2(\mathbb{Z},H^k(I_T))} \leq \mu.\end{aligned}$$

Furthermore, suppose that, for some $\tau < T$,

$$R_i - R_{|t<\tau} = S_i - S_{|t<\tau} = 0, \quad \left(\dot{X}_i - c, X_{i+1} - X_i - L\right)_{|t<\tau} = 0.$$

Then, for any $f \in H^k(\Omega_T)$ *and* $g \in l^2(\mathbb{Z}, H^k(I_T))$ *such that* $f_{|t<0} = 0, \quad g_{|t<0} = 0$, *the solution* (r, s) *belongs to* $H^k(\Omega_T)$, $(r^{\pm}, s^{\pm})$ *belongs to* $l^2(\mathbb{Z}, H^k(I_T))$ *and* ε *to* $l^2(\mathbb{Z}, H^{k+1}(I_T))$ *with the estimate*

$$\begin{aligned}\frac{1}{T}\|(r,s)\|^2_{H^k(\Omega_T)} + \left\|r^{\pm}, s^{\pm}\right\|_{l^2(\mathbb{Z},H^k(I_T))} + \|\varepsilon\|^2_{l^2(\mathbb{Z},H^{k+1}(I_T))} \\ \leq C(k,\mu)\left(T\,\|f\|^2_{H^k(\Omega_T)} + \|g\|^2_{l^2(\mathbb{Z},H^k(I_T))}\right).\end{aligned}$$

4. The nonlinear problem. In this section, we solve the full nonlinear Cauchy problem with prescribed initial data (r_0, s_0, X_0). To actually solve that problem, we need compatibility conditions on (r_0, s_0, X_0). Under suitable compatibility conditions, we shall prove the existence of an approximate solution $(r^{(a)}, s^{(a)}, X^{(a)})$ on a time interval $(0, T^0)$ for a sufficiently small interval. We can prove the well posedness of the full Cauchy problem through an iterative scheme based on the approximate linear problem introduced in the previous section. That scheme is proved to converge to a solution of the Cauchy problem with the energy estimates obtained previously.

4.1. Construction of an approximate solution. In what follows, we shall construct an approximate solution of the shallow water equations coupled with the Rankine–Hugoniot jump condition for an initial data (r_0, s_0, X_0) that satisfies some compatibility conditions. In what follows, we write compatibility conditions for the shallow water equations.

4.1.1. Compatibility conditions. After the change of variable introduced in section 2, the shallow water system is written as

$$\begin{aligned} &\partial_t r_i + \Lambda_i^1 \, \partial_x r_i = Q(r_i, s_i), \\ &\partial_t s_i + \Lambda_i^2 \, \partial_x s^i = Q(r_i, s_i) \end{aligned} \qquad \forall (x,i) \in (0,L) \times \mathbb{Z}, \tag{44}$$

where the Λ_i^k are defined by

$$\Lambda_i^k = \frac{L}{X_{i+1} - X_i} \left(\lambda_k(r_i, s_i) - \left(\dot{X}_i + \frac{x}{L} \left(\dot{X}_{i+1} - \dot{X}_i \right) \right) \right),$$

whereas the Rankine–Hugoniot jump conditions are given by

$$\dot{X}_i = F\left(r_i^+, r_i^-, s_i^+, s_i^-\right), \quad G\left(r_i^+, r_i^-, s_i^+, s_i^-\right) = 0 \quad \forall i \in \mathbb{Z}, \tag{45}$$

with

$$\begin{aligned} F\left(r_i^\pm, s_i^\pm\right) &= \left[(r-s)^4 + 8(r-s)^2(r+s)^2\right]_{iL} \left[(r-s)^2\right]_{iL} - 8\left[(r+s)(r-s)^2\right]_{iL}^2, \\ G\left(r_i^\pm, s_i^\pm\right) &= \frac{1}{2}\left[(r+s)(r-s)^2\right]_{iL} \left[(r-s)^2\right]_{iL}^{-1}. \end{aligned}$$

Suppose that both (r,s) and X are smooth enough: applying Faa di Bruno's formula to the jump conditions (45), we obtain

$$\frac{d^{p+1} X_i}{dt^{p+1}} = \sum_{i=1}^{p} \sum_{i_1 + \cdots + i_m = p} c_{i_1 \ldots i_m} d^m F\left(r_i^\pm, s_i^\pm\right) . \left(\frac{d^{i_1}}{dt^{i_1}} \begin{pmatrix} r_i^\pm \\ s_i^\pm \end{pmatrix}, \ldots, \frac{d^{i_m}}{dt^{i_m}} \begin{pmatrix} r_i^\pm \\ s_i^\pm \end{pmatrix} \right). \tag{46}$$

Furthermore, we derive the interior equations (44) with the Faa di Bruno formula:

$$\begin{aligned} \partial_t^{p+1} r_i &= \partial_t^p \, Q(r_i, s_i) - \sum_{l=0}^{p} C_p^l \, \partial_t^l \, \Lambda_i^1 \, \partial_x \partial_t^{p-l} r_i, \\ \partial_t^{p+1} s_i &= \partial_t^p \, Q(r_i, s_i) - \sum_{l=0}^{p} C_p^l \, \partial_t^l \, \Lambda_i^2 \, \partial_x \partial_t^{p-l} s_i. \end{aligned} \tag{47}$$

It is a straightforward computation to prove that

$$\partial_t^p \, Q(r_i, s_i) = \sum_{m=1}^{p} \sum_{i_1 + \cdots i_m = p} c_{i_1 \ldots i_m} d^m Q(r_i, s_i) . \left(\partial_t^{i_1} \begin{pmatrix} r_i \\ s_i \end{pmatrix}, \ldots, \partial_t^{i_m} \begin{pmatrix} r_i \\ s_i \end{pmatrix} \right).$$

Moreover, for $l = 1, \ldots, p-1$, we easily obtain

$$\begin{aligned} \partial_t^l \Lambda_i^k = &\sum_{m=1}^{l} \sum_{i_1 + \cdots + i_m = l} c_{i_1 \ldots i_m} d_{1,4}^m \Lambda_i^k . \left(\partial_t^{i_1} \begin{pmatrix} r_i \\ s_i \\ X_i \\ X_{i+1} \end{pmatrix}, \ldots, \partial_t^{i_m} \begin{pmatrix} r_i \\ s_i \\ X_i \\ X_{i+1} \end{pmatrix} \right) \\ &+ \sum_{m=1}^{l} \sum_{i_1 + \cdots + i_m = l} c_{i_1 \ldots i_m} d_{5,6}^m \Lambda_i^k . \left(\partial_t^{i_1} \begin{pmatrix} \dot{X}_i \\ \dot{X}_{i+1} \end{pmatrix}, \ldots, \partial_t^{i_m} \begin{pmatrix} \dot{X}_i \\ \dot{X}_{i+1} \end{pmatrix} \right), \end{aligned}$$

where $d_{1,4}\Lambda_i^k$ denotes the differential of Λ_i^k with respect to (r, s, X_i, X_{i+1}) and $d_{5,6}\Lambda_i^k$ denotes the differential with respect to $(\dot{X}_i, \dot{X}_{i+1})$. Here, the derivatives in X_i are no

larger than p. For $l = p$, one finds

$$\begin{aligned}\partial_t^p \Lambda_k^i = &\sum_{m=1}^{p} \sum_{i_1+\cdots+i_m=p} c_{i_1\dots i_m} d_{1,4}^m \Lambda_k . \left(\partial_t^{i_1} \begin{pmatrix} r_i \\ s_i \\ X_i \\ X_{i+1} \end{pmatrix}, \dots, \partial_t^{i_m} \begin{pmatrix} r_i \\ s_i \\ X_i \\ X_{i+1} \end{pmatrix} \right) \\ &+ \sum_{m=2}^{p} \sum_{i_1+\cdots+i_m=p} c_{i_1\dots i_m} d_{5,6}^m \Lambda_k . \left(\partial_t^{i_1} \begin{pmatrix} \dot{X}_i \\ \dot{X}_{i+1} \end{pmatrix}, \dots, \partial_t^{i_m} \begin{pmatrix} \dot{X}_i \\ \dot{X}_{i+1} \end{pmatrix} \right) \\ &+ \nabla_{5,6} \Lambda_k . \partial_t^{p+1} \begin{pmatrix} X_i \\ X_{i+1} \end{pmatrix}.\end{aligned}$$

The derivatives of X_i up to order $p+1$ are replaced by the expressions found in (46) involving derivatives of order $j \le p$. Following the definitions of compatibility conditions introduced in [1, p. 370] (see the definition of "compatibility conditions to order s"), we define the compatibility conditions for a pair of initial data (r_0, s_0, X_0).

We fix (r_0, s_0, X_0) an initial data and define the functions (r_p, s_p) and X_p so that

$$\begin{aligned} X_{1,i} &= F\left(r_{0,i}^{\pm}, s_{0,i}^{\pm}\right), \\ r_{1,i} &= Q(r_{0,i}, s_{0,i}) - \Lambda_{0,i}^1 \partial_x \, r_{0,i}, \\ s_{1,i} &= Q(r_{0,i}, s_{0,i}) - \Lambda_{0,i}^2 \partial_x \, s_{0,i}, \end{aligned} \tag{48}$$

and

$$\begin{aligned} X_{p+1,i} = &\sum_{i=1}^{p} \sum_{i_1+\cdots+i_m=p} c_{i_1\dots i_m} d^m \, F\left(r_{0,i}^{\pm}, s_{0,i}^{\pm}\right) . \left(\begin{pmatrix} r_{i_1,i}^{\pm} \\ s_{i_1,i}^{\pm} \end{pmatrix}, \dots, \begin{pmatrix} r_{i_m,i}^{\pm} \\ s_{i_m,i}^{\pm} \end{pmatrix} \right), \\ r_{p+1,i} = &\sum_{m=1}^{p} \sum_{i_1+\cdots i_m=p} c_{i_1\dots i_m} d^m Q(r_{0,i}, s_{0,i}) . \left(\begin{pmatrix} r_{i_1,i} \\ s_{i_1,i} \end{pmatrix}, \dots, \begin{pmatrix} r_{i_m,i} \\ s_{i_m,i} \end{pmatrix} \right) \\ &- \sum_{l=0}^{p} C_p^l \, \Lambda_{l,i}^1 \partial_x \, r_{p-l,i}, \\ s_{p+1,i} = &\sum_{m=1}^{p} \sum_{i_1+\cdots i_m=p} c_{i_1\dots i_m} d^m Q(r_{0,i}, s_{0,i}) . \left(\begin{pmatrix} r_{i_1,i} \\ s_{i_1,i} \end{pmatrix}, \dots, \begin{pmatrix} r_{i_m,i} \\ s_{i_m,i} \end{pmatrix} \right) \\ &- \sum_{l=0}^{p} C_p^l \, \Lambda_{l,i}^2 \partial_x \, s_{p-l,i}, \end{aligned} \tag{49}$$

where $\Lambda_{l,i}^k$ are defined, for any $l = 0, \dots, p-1$, by

$$\begin{aligned} \Lambda_{0,i}^k &= \frac{L}{X_{0,i+1} - X_{0,i}} \left(\lambda_k(r_{0,i}, s_{0,i}) - \left(X_{1,i} + \frac{x}{L}(X_{1,i+1} - X_{1,i}) \right) \right), \\ \Lambda_{l,i}^k &= \sum_{m=1}^{l} \sum_{i_1+\cdots+i_m=l} c_{i_1\dots i_m} d_{1,4}^m \Lambda_i^k . \left(\begin{pmatrix} r_{i_1,i} \\ s_{i_1,i} \\ X_{i_1,i} \\ X_{i_1,i+1} \end{pmatrix}, \dots, \begin{pmatrix} r_{i_m,i} \\ s_{i_m,i} \\ X_{i_m,i} \\ X_{i_m,i+1} \end{pmatrix} \right) \\ &+ \sum_{m=1}^{l} \sum_{i_1+\cdots+i_m=l} c_{i_1\dots i_m} d_{5,6}^m \Lambda_i^k . \left(\begin{pmatrix} X_{i_1+1,i} \\ X_{i_1+1,i+1} \end{pmatrix}, \dots, \begin{pmatrix} X_{i_m+1,i} \\ X_{i_m+1,i+1} \end{pmatrix} \right), \end{aligned}$$

and, for $l = p$,

$$\Lambda^k_{p,i} = \sum_{m=1}^{p} \sum_{i_1+\cdots+i_m=p} c_{i_1\dots i_m} d^m_{1,4}\Lambda^k_i . \left(\begin{pmatrix} r_{i_1,i} \\ s_{i_1,i} \\ X_{i_1,i} \\ X_{i_1,i+1} \end{pmatrix}, \dots, \begin{pmatrix} r_{i_m,i} \\ s_{i_m,i} \\ X_{i_m,i} \\ X_{i_m,i+1} \end{pmatrix} \right)$$
$$+ \sum_{m=2}^{p} \sum_{i_1+\cdots+i_m=p} c_{i_1\dots i_m} d^m_{5,6}\Lambda^k_i . \left(\begin{pmatrix} X_{i_1+1,i} \\ X_{i_1+1,i+1} \end{pmatrix}, \dots, \begin{pmatrix} X_{i_m+1,i} \\ X_{i_m+1,i+1} \end{pmatrix} \right)$$
$$+ \nabla_{5,6}\Lambda^k_i . \begin{pmatrix} X_{p+1,i} \\ X_{p+1,i+1} \end{pmatrix}.$$

For the definition of the sequence (r_p, s_p, X_p), we have not used the jump condition $G(r^\pm_i, s^\pm_i) = 0$. In order to obtain a solution to the Cauchy problem, the sequence (r_p, s_p, X_p) shall satisfy some compatible conditions. We define here those compatibility conditions.

DEFINITION 1. *A pair of initial data (r_0, s_0) and X_0 is said to be compatible up to order s if, for all $p \in \{0, \dots, s\}$, the functions (r_p, s_p, X_p) defined by (49) satisfy*

$$0 = \sum_{m=1}^{p} \sum_{i_1+\cdots+i_m=p} c_{i_1,\dots,i_m} d^m \, G\left(r^\pm_{0,i}, s^\pm_{0,i}\right) . \left(\begin{pmatrix} r^\pm_{i_1,i} \\ s^\pm_{i_1,i} \end{pmatrix}, \dots, \begin{pmatrix} r^\pm_{i_m,i} \\ s^\pm_{i_m,i} \end{pmatrix} \right).$$

4.1.2. Approximate solution. The purpose of this section is to prove that, under the compatibility conditions defined above, one can construct an approximate solution of the full Cauchy problem. Denote $\Omega = \mathbb{R} \setminus \{iL,\, i \in \mathbb{Z}\}$. We show the following result.

PROPOSITION 7. *Let (r_0, s_0) and X_0 be an initial data so that $(r_0, s_0) \in (R, S) + H^{m+\frac{1}{2}}(\Omega)$ and $(X_{0,i} - iL)_{i\in\mathbb{Z}} \in l^2(\mathbb{Z})$, with $m > 2$ an integer that is compatible up to order $m-1$ and such that*

$$\|(r_0, s_0) - (R, S)\|_{L^\infty(\Omega)} \le \rho, \quad \max_{i\in\mathbb{Z}} \left(|X_{0,i+1} - X_{0,i} - L|\right) \le \rho.$$

Then for ρ sufficiently small, there exists $T_0 > 0$ and

$$(r_a, s_a) \in (R, S) + H^{m+1}(\Omega), \quad (X_{a,i} - (ct + iL)) \in l^2\left(\mathbb{Z}, H^{m+1}(\mathbb{R})\right),$$

with both $(r_a, s_a) - (R, S)$ and $X_{a,i} - (ct + iL)$ vanishing for $|t| \ge 2T_0$, so that $(r_a, s_a)_{|t=0} = (r_0, s_0)$, $X_{a,i}(0) = X_{0,i}$ and for all $(x, t) \in \Omega \times [-T_0, T_0]$,

$$\|(r_a - r_0, s_a - s_0)\| \le \frac{\rho}{2}, \quad |X_{a,i} - X_{0,i}| \le \frac{\rho}{2}.$$

Furthermore, the functions f_a and g_a defined as

$$f^1_{a,i} = \partial_t r_{a,i} + \Lambda^1_{a,i}\partial_x r_{a,i} - Q(r_{a,i}, s_{a,i}),$$
$$f^2_{a,i} = \partial_t s_{a,i} + \Lambda^2_{a,i}\partial_x s_{a,i} - Q(r_{a,i}, s_{a,i}),$$

and $g^1_{a,i} = \dot{X}_{a,i} - F(r^\pm_{a,i}, s^\pm_{a,i})$, $g^2_{a,i} = G(r^\pm_{a,i}, s^\pm_{a,i})$ are such that

$$\partial^p_t f_{a|t=0} = 0, \quad \partial^p_t g_{a|t=0} = 0 \ \forall p \in \{0, \dots, m-1\}.$$

The function f_a belongs to $H^m(\Omega\times\mathbb{R})$, g_a lies in $l^2(\mathbb{Z},H^m(\mathbb{R}))$, and both vanish for $|t|\geq 2T_0$.

Proof. For $k=1\dots m-1$, we construct a sequence $(r_k,s_k)\in H^{m+\frac{1}{2}-k}(\Omega)$, X_k so that $X_{1,i}-c\in l^2(\mathbb{Z})$ and $X_{k,i}\in l^2(\mathbb{Z})$ $(k\geq 2)$ that satisfies the compatibility conditions (49). Then, by trace lifting, we find $(r_a,s_a)\in(R,S)+H^{m+1}(\Omega\times\mathbb{R})$ and $X_a\in(ct+iL)+l^2(\mathbb{Z},H^{m+1}(\mathbb{R}))$ so that

$$\begin{aligned}\|r_a-R,s_a-S\|_{H^{m+1}(\Omega\times\mathbb{R})}&\leq C\|r_0-R,s_0-S\|_{H^{m+1}(\Omega)},\\ \|X_a-(ct+iL)\|_{l^2(\mathbb{Z},H^{m+1}(\mathbb{R}))}&\leq C\|X_0-iL\|_{l^2(\mathbb{Z})},\end{aligned}$$

and $\partial_t^k(r_{a,i},s_{a,i})_{|t=0}=(r_{k,i},s_{k,i})$, $\partial_t^k(X_{a,i})_{|t=0}=X_{k,i}\forall k\in\{0,\dots,m-1\}$. By Sobolev embeddings, we find $T_0>0$ so that

$$\|(r_a-r_0,s_a-s_0)\|\leq\frac{\rho}{2},\quad \|X_a-X_0\|\leq\frac{\rho}{2}$$

for $|t|\leq T_0$. In order to obtain functions f_a,g_a that vanish for $|t|\geq 2T_0$, we choose $\phi\in\mathcal{D}(\mathbb{R})$ a cut-off function so that $\phi(t)=1$ if $|t|\leq T_0$ and $\phi(t)=0$ if $|t|\geq 2T_0$. Then, substitute (r_a,s_a) and X_a, respectively, by $\phi(r_a,s_a)+(1-\phi)(R,S)$ and $\phi X_{a,i}+(1-\phi)(ct+iL)$. Then, one can prove that f_a belongs to $H^m(\Omega\times\mathbb{R})$ and $g^{(a)}$ lies in $l^2(\mathbb{Z},H^m(\mathbb{R}))$, and both vanish for $|t|\geq 2T_0$. Finally, according to the compatibility conditions that are satisfied up to order $m-1$, the derivatives of these functions vanish at $t=0$. The proof of the proposition is then completed. ∎

4.2. The fixed point argument. In this section, we prove the main theorem of this paper on the persistence of roll-waves.

THEOREM 2. *Assume that the Froude number F satisfies $0<F<F_c$. There exists $\rho>0$ so that for any $(r_0,s_0)\in(R,S)+H^{m+\frac{1}{2}}(\Omega)$, with $m>2$ and $(X_{0,i}-iL)_{i\in\mathbb{Z}}\in l^2(\mathbb{Z})$, compatible up to order $m-1$ and such that*

$$\|(r_0,s_0)-(R,S)\|_{L^\infty(\Omega)}\leq\rho,\quad \|X_{0,i+1}-X_{0,i}-L\|_{l^\infty(\mathbb{Z})}\leq\rho,$$

there is $T>0$ and a solution (r,s) and X of the shallow water system (8) *coupled with the Rankine–Hugoniot jump conditions* (9) *so that*

$$(r,s)_{|t=0}=(r_0,s_0),\quad X_i(t=0)=X_{0,i},$$

and (r,s) belongs to $(R,S)+H^m(\Omega\times[0,T])$, whereas $(X_i-(ct+iL))_{i\in\mathbb{Z}}$ lies in $l^2(\mathbb{Z},H^{m+1}(0,T))$.

Proof. The method of proof is based on an interative scheme: We search the solution in the form $(r,s)=(r_a,s_a)+(\overline{r},\overline{s})$, $X=X_a+\varepsilon$, where (r_a,s_a,X_a) is the approximate solution associated to the initial data (r_0,s_0,X_0) and defined in Proposition (7). For that purpose, we introduce the iterative scheme $r^{(0)}=s^{(0)}=0$, $\varepsilon^{(0)}=0$, and $(r^{(k)},s^{(k)})$, $\varepsilon^{(k)}$ is defined inductively as the unique solution, with zero initial data, of the system

$$\partial_t r_i^{(k)}+\Lambda_i^{1,(k-1)}\partial_x r_i^{(k)}=f_i^{1,(k-1)},\quad \partial_t s_i^{(k)}+\Lambda_i^{2,(k-1)}\partial_x s_i^{(k)}=f_i^{2,(k-1)},\tag{50}$$

where the functions $f_i^{n,(k-1)}$, $n=1,2$ are defined by

$$\begin{aligned}f_i^{1,(k-1)}&=Q_i^{(k-1)}-\partial_t r_i^{(k-1)}-\Lambda_i^{1,(k-1)}\partial_x r_i^{(k-1)},\\ f_i^{2,(k-1)}&=Q_i^{(k-1)}-\partial_t s_i^{(k-1)}-\Lambda_i^{2,(k-1)}\partial_x s_i^{(k-1)},\end{aligned}$$

with $Q_i^{(k)} = Q(r_{a,i} + r^{(k)}, s_{a,i} + s^{(k)})$ and

$$\Lambda_i^{n,(k)} = \frac{L}{X_{a,i+1} - X_{a,i} + \varepsilon_{i+1}^{(k)} - \varepsilon_i^{(k)}} \Big(\lambda^n \left(r_{a,i} + r^{(k)} s_{a,i} + s^{(k)} \right) - \left(\dot{X}_{a,i} + \dot{\varepsilon}_i^{(k)} + \frac{x}{L} \left(\dot{X}_{a,i+1} - \dot{X}_{a,i} + \dot{\varepsilon}_{i+1}^{(k)} - \dot{\varepsilon}_i^{(k)} \right) \right) \Big).$$

Equations (50) are supplemented by boundary conditions that read

$$\begin{pmatrix} \dot{\varepsilon}_i^{(k)} \\ 0 \end{pmatrix} = \underline{B}_i^{(k-1)} \left(r_i^{(k),\pm}, s_i^{(k),\pm} \right) + \begin{pmatrix} g_i^{1,(k-1)} \\ g_i^{2,(k-1)} \end{pmatrix}, \tag{51}$$

where the functions $g_i^{n,(k)}$ are defined by

$$\begin{pmatrix} g_i^{1,(k)} \\ g_i^{2,(k)} \end{pmatrix} = \begin{pmatrix} -\dot{X}_{a,i} + F \left(r_{a,i}^{\pm} + r_i^{(k),\pm}, s_{a,i}^{\pm} + s_i^{(k),\pm} \right) \\ G \left(r_{a,i}^{\pm} + r_i^{(k),\pm}, s_{a,i}^{\pm} + s_i^{(k),\pm} \right) \end{pmatrix} - \underline{B}_i^{(k)} \left(r_i^{(k),\pm}, s_i^{(k),\pm} \right).$$

We estimate the nonlinear terms $f_i^{n,(k-1)}$ and $g_i^{n,(k-1)}$. One can prove that there exists $M > 0$ so that for any $T \in (0, T_0)$, and if $r^{(k-1)}, s^{(k-1)}$ belongs to $H^m(\Omega) \times I_T)$, $r^{(k-1),\pm}, s^{(k-1),\pm}$ lies in $l^2(\mathbb{Z}, H^m(I_T))$ and $\varepsilon^{(k-1)}$ to $l^2(\mathbb{Z}, H^{m+1}(I_T))$, vanish for $t < 0$ and are such that

$$\begin{aligned} \left\| r^{(k-1)}, s^{(k-1)} \right\|_{H^m(\Omega \times I_T)} &\leq M, \\ \left\| r^{(k-1),\pm}, s^{(k-1),\pm} \right\|_{l^2(\mathbb{Z}, H^m(I_T))} &\leq M, \\ \left\| \varepsilon^{(k-1)} \right\|_{l^2(\mathbb{Z}, H^{m+1}(I_T))} &\leq M, \end{aligned} \tag{52}$$

then $f^{n,(k-1)}$ lies in $H^m(\Omega \times I_T)$, $g^{n,(k-1)}$ in $l^2(\mathbb{Z}, H^m(I_T))$ and satisfy the estimates

$$\left\| f^{n,(k-1)} \right\|_{H^m(\Omega \times I_T)} \leq C(M), \quad \left\| g^{n,(k-1)} \right\|_{l^2(\mathbb{Z}, H^m(I_T))} \leq C(M)T + \eta(T),$$

with $\eta(T) \to 0$ as $T \to 0$ and $M \longmapsto C(M)$ is a continuous function. The function $(r^{(0)}, s^{(0)}), \varepsilon^{(0)}$ satisfies all the assumptions. We can apply the Proposition 6: There exists a unique solution $(r^{(1)}, s^{(1)}) \in H^m(\Omega \times I_T)$ and $\varepsilon^{(1)} \in l^2(\mathbb{Z}, H^{m+1}(I_T))$ solution of (50), (51) with $k = 1$. Moreover, diminishing T if necessary, one can prove, using the energy estimate in Proposition 6, that

$$\begin{aligned} \left\| r^{(1)}, s^{(1)} \right\|_{H^m(\Omega \times I_T)} &\leq M, \\ \left\| r^{(1),\pm}, s^{(1),\pm} \right\|_{l^2(\mathbb{Z}, H^m(I_T))} &\leq M, \\ \left\| \varepsilon^{(1)} \right\|_{l^2(\mathbb{Z}, H^{m+1}(I_T))} &\leq M. \end{aligned}$$

By induction, one can construct $(r^{(k)}, s^{(k)}) \in H^m(\Omega \times I_T)$ and $\varepsilon^{(k)} \in l^2(\mathbb{Z}, H^{m+1}(I_T))$ that satisfy (52) for $k \geq 1$. Now using the fact that $W^{1,\infty}(\mathbb{R}) \hookrightarrow H^m(\mathbb{R})$, we obtain uniform estimates on $\|(r^{(k)}, s^{(k)}\|_{W^{1,\infty}}(\mathbb{R})$. We write the equations satisfied by $(r^{(k)} - r^{(k-1)}, s^{(k)} - s^{(k-1)})$ that is denoted $u^{(k)}$ in what follows, and $\varepsilon^{(k)} - \varepsilon^{(k-1)}$. Using the L^2 energy estimate of Proposition 4, we prove that, for T sufficiently small,

$$\begin{aligned} \left\| u^{(k+1)} \right\|_{L^2(\Omega \times I_T)} &\leq \frac{1}{2} \left\| u^{(k)} \right\|_{L^2(\Omega \times I_T)}, \\ \left\| u^{(k+1),\pm} \right\|_{l^2(\mathbb{Z}, L^2(I_T))} &\leq \frac{1}{2} \left\| u^{(k),\pm} \right\|_{l^2(\mathbb{Z}, L^2(I_T))}, \\ \left\| \varepsilon^{(k+1)} - \varepsilon^{(k)} \right\|_{l^2(\mathbb{Z}, H^1(I_T))} &\leq \frac{1}{2} \left\| \varepsilon^{(k)} - \varepsilon^{(k-1)} \right\|_{l^2(\mathbb{Z}, H^1(I_T))}. \end{aligned}$$

Then, $(r^{(k)}, s^{(k)})$, $(r^{(k),\pm}, s^{(k)}, \pm)$, and $\varepsilon^{(k)}$ are, respectively, Cauchy sequences in $L^2(\Omega \times I_T)$, $l^2(\mathbb{Z}, L^2(I_T))$, and $l^2(\mathbb{Z}, H^1(I_T))$ and converge in that space. We denote $\overline{r}, \overline{s}$, and $\overline{\varepsilon}$ the limits of $(r^{(k)}, s^{(k)})$ and $\varepsilon^{(k)}$ as $k \to \infty$. Necessarily, $(\overline{r}, \overline{s})$ lies in $H^m(\Omega \times I_T)$, $(\overline{r}^{\pm}, \overline{s}^{\pm})$ in $l^2(\mathbb{Z}, H^m(I_T))$, and $\overline{\varepsilon}$ belongs to $l^2(\mathbb{Z}, H^{m+1}(I_T))$. By standard interpolation arguments, we obtain strong convergence in appropriate Sobolev spaces so that we can make $k \to \infty$ in (50),(51). As a result, the function $(r_a + \overline{r}, s_a + \overline{s})$ and $X_a + \overline{\varepsilon}$ solves the full Cauchy problem (8), (9) on the interval $(0, T)$. This concludes the proof of the theorem and the persistence of roll-waves for suitable initial data. ∎

5. Conclusion. In this paper, we proved the well posedness of the shallow water equations in the neighborhood of roll-wave solutions for initial conditions that satisfy suitable compatibility conditions. The main issue here is to formulate the stability problem in the presence of an infinite number of shocks. This is done through a Lipschitz change of variable that fixes *all* the discontinuities of functions that are close to roll-waves. Writing the shallow water equations with Riemann invariants, we diagonalize the shallow water system; in that case, it is easier to obtain estimates on the linearized problems. We prove the existence of weak solutions to the linearized problem through the formulation of an adjoint problem just as in the shock case. We then prove regularity properties on that solution. The main difference with the shock case is the presence of a singularity in the interior equations due to the presence of a sonic point: a regularization in the transverse direction is not sufficient here, and we regularize the solution also in the direction of propagation. The existence of a solution to the full nonlinear problem is then obtain through a fixed point argument. A key assumption is that the Froude number F (that measures whether a the flow of the channel is fluvial or torrential) has to be smaller than a critical value F_c: this suggests that in the super critical regime $F > F_c$, the roll-wave structure may be destroyed. This would be an interesting question to see what happens in that regime even from a numerical point of view.

Though written in a $1D$ setting, we claim that the results obtained in this paper hold true in the *multidimensional setting*. We briefly discuss the method of proof. Indeed, in the $2D$ case, one shall consider a transverse coordinate y and a transverse speed v; the hyperbolic part of the equations is nothing but the isentropic Euler equations. Here we shall assume that the shocks are located at the positions $\Psi_i(t, y) \approx iT$, $i \in \mathbb{Z}$. Next we perform a change of variable similar to the one used in this paper, and one can easily formulate the linearized equations. The main difference with the $1D$ case is the obtention of energy estimates for the linearized (exact and approximated) equations. However, once the linearization procedure is performed, we are left with an infinite but discrete number of Lax shocks that can be treated separately. For each Lax shock, one can use Kreiss symmetrizers to obtain energy estimates and then sum all the contributions over $i \in \mathbb{Z}$. The end of the proof is then completely similar to the $1D$ case presented here. This method shall be developped in a forthcoming paper on the persistence of roll-waves in general hyperbolic systems with discontinuities satisfying Lax shock conditions obtained recently by the author[14]. We shall also consider the persistence of roll-waves with characteristic discontinuities. This is of interest from a physical point of view, since they may appear when the pressure term in shallow water equations is not convex like in stratified flows (see [2] for more details).

At this stage, several questions arise. First, we have restricted our analysis to *smooth* perturbations: in the $1D$ case, because of the progress made in the analysis of the $1D$ problem, it seems relevant to consider more general perturbations. In the case of shock waves, the persistence of shocks satisfying the Majda–Liu stability condition

under small perturbations with a bounded variation (BV) using the Glimm scheme or the front tracking method has been established (see [16], [8] for more details). It is thus a natural question whether roll-waves persist under BV perturbations, though it may be a very difficult problem in the case of roll-waves. Another interesting question is the stability of viscous roll-waves that are close to Dressler roll-waves in the vanishing viscosity limit [10, 11]; that question has been treated recently in the shock case for multidimensional perturbations [6]. Moreover, the linear stability of viscous roll-waves has been proved under strong spectral stability [12]. Energy estimates have been obtained through Green functions that blow up in the vanishing viscosity case, just as in the shock case. It would be interesting here to adapt the method introduced by Metivier and coworkers [6] to the roll-wave case: the main issue would be the presence of an infinite number of viscous shocks. This question is an open problem that is postponed to a forthcoming work.

Acknowledgment. The author wishes to thank Prof. Sylvie Benzoni for many useful discussions on the multidimensional stability of shock waves.

REFERENCES

[1] S. Benzoni-Gavage and D. Serre, *Multidimensional Hyperbolic Partial Differential Equations: First-order Systems and Applications*, Oxford Math. Monogr., Clarendon Press, Oxford University Press, Oxford, 2007.

[2] A. Boudlal and V. Liapidevskii, *Multi-shock structure of roll-waves*, C.R.A.S. Mécanique, 332 (2004), pp. 659–664.

[3] R. Dressler, *Mathematical solution of the problem of roll waves in inclined open channels,* Comm. Pure Appl. Math., 2 (1949), pp. 149–190.

[4] A. Majda, *The stability of multidimensional shock fronts,* Mem. Amer. Math. Soc., 41 (1983), No. 275.

[5] G. Métivier, *Stability of multidimensional shocks,* in Advances in the Theory of Shock Waves, pp. 25–103, Progr. Nonlinear Differential Equations Appl. 47, Birkhäuser Boston, Boston 2001, pp. 25–103.

[6] O. Guès, G. Métivier, M. Williams, and K. Zumbrun, *Existence and stability of multidimensional shock fronts in the vanishing viscosity limit,* Arch. Ration. Mech. Anal., 175 (2005), pp. 151–244.

[7] J.-F. Gerbeau and B. Perthame, *Derivation of viscous Saint-Venant system for Laminar shallow-water: Numerical validation*, Discrete Contin. Dyn. Syst. Ser. B, 1 (2001), pp. 89–102.

[8] M. Lewicka and K. Trivisa, *On the L^1 well posedness of systems of conservation laws near solutions containing two large shocks,* J. Differential Equations, 179 (2002), pp. 133–177.

[9] D.J. Needham and J.H. Merkin, *On roll waves down an open inclined channel,* Proc. Roy. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 394 (1984), pp. 259–278.

[10] J. Härterich, *Existence of rollwaves in a viscous shallow water equation,* in Proceedings of EQUADIFF 2003, World Scientific, Hackensack, NJ, 2005, pp. 511–516.

[11] P. Noble, *Méthodes de Variétés Invariantes Pour les Équations de Saint Venant et les Systèmes Hamiltoniens Discrets*, Thèse Université Toulouse, Toulouse, France, 2003.

[12] P. Noble, *Linear stability of viscous roll-waves,* Comm. Partial Differential Equations, 32 (2007), pp. 1681–1713.

[13] P. Noble, *Linear stability of roll waves* Indiana Univ. Math. J., 55 (2006), pp. 795–848.

[14] P. Noble, *Roll-waves in general hyperbolic systems with source terms,* SIAM J. Appl. Math., 67 (2007), pp. 1202–1212.

[15] F. Rousset, *Viscous approximation of strong shocks of systems of conservation laws,* SIAM J. Math. Anal., 35 (2003), pp. 492–519.

[16] S. Schochet, *Sufficient conditions for local existence via Glimm's scheme for large BV data,* J. Differential Equations, 89 (1991), pp. 317–354.

[17] K. Tamada and H. Tougou, *Stability of roll-waves on thin Laminar flow down an inclined plane wall,* J. Phys. Soc. Japan, 47 (1979), pp. 1992–1998.

[18] J.P. Vila, *Sur la Théorie et l'Approximation Numérique des Problèmes Hyperboliques Non Lineaires. Application aux Équations de Saint Venant et à la Modélisation des Avalanches de neige Dense,* Thèse Univ. Paris VI, Paris, France, (1986).

SIAM J. MATH. ANAL.
Vol. 40, No. 5, pp. 1815–1851

A NONLOCAL p-LAPLACIAN EVOLUTION EQUATION WITH NONHOMOGENEOUS DIRICHLET BOUNDARY CONDITIONS*

F. ANDREU†, J. M. MAZÓN‡, J. D. ROSSI§, AND J. TOLEDO‡

Abstract. In this paper we study the nonlocal p-Laplacian-type diffusion equation $u_t(t,x) = \int_{\mathbb{R}^N} J(x-y)|u(t,y)-u(t,x)|^{p-2}(u(t,y)-u(t,x))\,dy$, $(t,x) \in]0,T[\times\Omega$, with $u(t,x) = \psi(x)$ for $(t,x) \in]0,T[\times(\mathbb{R}^N \setminus \Omega)$. If $p > 1$, this is the nonlocal analogous problem to the well-known local p-Laplacian evolution equation $u_t = \operatorname{div}(|\nabla u|^{p-2}\nabla u)$ with Dirichlet boundary condition $u(t,x) = \psi(x)$ on $(t,x) \in]0,T[\times\partial\Omega$. If $p = 1$, this is the nonlocal analogous to the total variation flow. When $p = +\infty$ (this has to be interpreted as the limit as $p \to +\infty$ in the previous model) we find an evolution problem that can be seen as a nonlocal model for the formation of sandpiles (here $u(t,x)$ stands for the height of the sandpile) with prescribed height of sand outside of Ω. We prove, as main results, existence, uniqueness, a contraction property that gives well posedness of the problem, and the convergence of the solutions to solutions of the local analogous problem when a rescaling parameter goes to zero.

Key words. nonlocal diffusion, p-Laplacian, nonhomogeneous Dirichlet boundary conditions, total variation flow, sandpiles

AMS subject classifications. 35B40, 45A07, 45G10

DOI. 10.1137/080720991

1. Introduction.

In this paper we study the nonlocal diffusion equation

$$u_t(t,x) = \int_{\mathbb{R}^N} J(x-y)|u(t,y)-u(t,x)|^{p-2}(u(t,y)-u(t,x))\,dy \qquad (t,x) \in]0,T[\times\Omega,$$

where Ω is a bounded domain and u is prescribed in $\mathbb{R}^N \setminus \Omega$ as $u(t,x) = \psi(x)$ for $(t,x) \in]0,T[\times(\mathbb{R}^N \setminus \Omega)$. We consider $1 < p < +\infty$ as well as the extreme cases $p = 1$ and the limit $p \nearrow +\infty$. Throughout the paper, we assume that $J : \mathbb{R}^N \to \mathbb{R}$ is a nonnegative, radial, continuous function, strictly positive in $B(0,1)$, vanishing in $\mathbb{R}^N \setminus B(0,1)$ and such that $\int_{\mathbb{R}^N} J(z)\,dz = 1$.

First, let us briefly introduce the prototype of nonlocal problem that will be considered along this work. Nonlocal evolution equations of the form

$$u_t(t,x) = (J * u - u)(t,x) = \int_{\mathbb{R}^N} J(x-y)u(t,y)\,dy - u(t,x), \tag{1.1}$$

and variations of it, have been recently widely used to model diffusion processes. More precisely, as stated in [31], if $u(t,x)$ is thought of as a density at the point x at time t and $J(x-y)$ is thought of as the probability distribution of jumping from location y to

*Received by the editors April 11, 2008; accepted for publication (in revised form) September 13, 2008; published electronically January 7, 2009.
http://www.siam.org/journals/sima/40-5/72099.html

†Departament de Matemàtica Aplicada, Universitat de València, Valencia, Spain (fuensanta.andreu@uv.es). This author was partially supported by the PNPGC project, reference MTM2008-03176.

‡Departament d'Anàlisi Matemàtica, Universitat de València, Valencia, Spain (mazon@uv.es, toledojj@uv.es). These authors were partially supported by the PNPGC project, reference MTM2008-03176.

§IMDEA Matematicas, C-IX, Campus Cantoblanco UAM, Madrid, Spain. On leave from Dpto. de Matemáticas, FCEyN Universidad de Buenos Aires, 1428 Buenos Aires, Argentina (jrossi@dm.uba.ar). This author was partially supported by ANPCyT PICT 5009, UBA X066, CONICET (Argentina).

location x, then $\int_{\mathbb{R}^N} J(y-x)u(t,y)\,dy = (J*u)(t,x)$ is the rate at which individuals are arriving at position x from all other places and $-u(t,x) = -\int_{\mathbb{R}^N} J(y-x)u(t,x)\,dy$ is the rate at which they are leaving location x to travel to all other sites. This consideration, in the absence of external or internal sources, leads immediately to the fact that the density u satisfies (1.1). For recent references on nonlocal diffusion, see [4], [5], [6], [9], [11], [12], [20], [21], [22], [23], [24], [25], [26], [27], [31], [33], [36] and references therein.

The first goal of this paper is to study the following nonlocal nonlinear diffusion problem:

$$P_p^J(u_0,\psi)\begin{cases} u_t(t,x) = \displaystyle\int_\Omega J(x-y)|u(t,y)-u(t,x)|^{p-2}(u(t,y)-u(t,x))\,dy \\ \qquad\qquad + \displaystyle\int_{\Omega_J\setminus\overline{\Omega}} J(x-y)|\psi(y)-u(t,x)|^{p-2}(\psi(y)-u(t,x))\,dy, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (t,x)\in]0,T[\times\Omega, \\ u(0,x) = u_0(x), \quad x\in\Omega. \end{cases}$$

Here $\Omega_J = \Omega + \operatorname{supp}(J)$ and ψ is a given function $\psi : \Omega_J \setminus \overline{\Omega} \to \mathbb{R}$.

Observe that we can rewrite $P_p^J(u_0,\psi)$, setting $u(t,x) = \psi(x)$ in $\Omega_J \setminus \overline{\Omega}$, as

$$\begin{cases} u_t(t,x) = \displaystyle\int_{\Omega_J} J(x-y)|u(t,y)-u(t,x)|^{p-2}(u(t,y)-u(t,x))\,dy, \quad (t,x)\in]0,T[\times\Omega, \\ u(t,x) = \psi(x), \quad (t,x)\in]0,T[\times\left(\Omega_J\setminus\overline{\Omega}\right), \\ u(0,x) = u_0(x), \quad x\in\Omega, \end{cases}$$

and we call it the *nonlocal p-Laplacian problem with Dirichlet boundary condition*. Note that we are prescribing the values of u outside the domain Ω and not only on its boundary. This is due to the nonlocal character of the problem.

Let us state the precise definition of solution. Solutions to $P_p^J(u_0,\psi)$ will be understood in the following sense.

DEFINITION 1.1. *Let $1 < p < +\infty$. A solution of $P_p^J(u_0,\psi)$ in $[0,T]$ is a function*

$$u \in C\left([0,T];L^1(\Omega)\right) \cap W^{1,1}\left(]0,T[;L^1(\Omega)\right),$$

which satisfies $u(0,x) = u_0(x)$ a.e. $x\in\Omega$ and

$$u_t(t,x) = \int_\Omega J(x-y)|u(t,y)-u(t,x)|^{p-2}(u(t,y)-u(t,x))\,dy + \int_{\Omega_J\setminus\overline{\Omega}} J(x-y)|\psi(y)-u(t,x)|^{p-2}(\psi(y)-u(t,x))\,dy,$$

for a.e. $t\in]0,T[$ and a.e. $x\in\Omega$.

Our first result shows existence and uniqueness of a global solution for this problem. Moreover, a contraction principle holds.

THEOREM 1.2. *Assume $p > 1$ and let $u_0 \in L^p(\Omega)$, $\psi \in L^p(\Omega_J\setminus\overline{\Omega})$. Then, there exists a unique solution to $P_p^J(u_0,\psi)$ in the sense of Definition 1.1. Moreover, if $u_{i0}\in L^1(\Omega)$ and u_i is a solution in $[0,T]$ of $P_p^J(u_{i0},\psi)$, $i=1,2$, respectively. Then*

$$\int_\Omega (u_1(t)-u_2(t))^+ \le \int_\Omega (u_{10}-u_{20})^+ \qquad \text{for every } t\in[0,T].$$

If $u_{i0} \in L^p(\Omega)$, $i = 1, 2$, *then*

$$\|u_1(t) - u_2(t)\|_{L^p(\Omega)} \le \|u_{10} - u_{20}\|_{L^p(\Omega)} \qquad \text{for every } t \in [0, T].$$

Our next step is to rescale the kernel J appropriately and take the limit as the scaling parameter goes to zero. To be more precise, for $p > 1$, we consider the local p-Laplace evolution equation with Dirichlet boundary condition

$$D_p(u_0, \tilde{\psi}) \quad \begin{cases} u_t = \Delta_p u & \text{in }]0, T[\times\Omega, \\ u = \tilde{\psi} & \text{on }]0, T[\times\partial\Omega, \\ u(0, x) = u_0(x) & \text{in } \Omega, \end{cases}$$

where the boundary datum $\tilde{\psi}$ is assumed to be the trace of a function defined in a larger domain and the operator in the equation, $\Delta_p u = \operatorname{div}(|\nabla u|^{p-2}\nabla u)$, is the usual local p-Laplacian.

We prove that the solutions of this local problem can be approximated by solutions of a sequence of nonlocal *p-Laplacian* problems of the form P_p^J. Indeed, for given $p \ge 1$ and J we consider the rescaled kernels

$$J_{p,\varepsilon}(x) := \frac{C_{J,p}}{\varepsilon^{p+N}} J\left(\frac{x}{\varepsilon}\right), \qquad \text{where} \qquad C_{J,p}^{-1} := \frac{1}{2}\int_{\mathbb{R}^N} J(z)|z_N|^p \, dz \tag{1.2}$$

is a normalizing constant in order to obtain the p-Laplacian in the limit instead of a multiple of it, and we obtain the following result.

THEOREM 1.3. *Let* Ω *be a smooth bounded domain in* $\mathbb{R}^N$ *and* $\tilde{\psi} \in W^{1/p',p}(\partial\Omega) \cap L^\infty(\partial\Omega)$. *Let* $\psi \in W^{1,p}(\Omega_J) \cap L^\infty(\Omega_J)$ *such that the trace* $\psi|_{\partial\Omega} = \tilde{\psi}$. *Assume* $J(x) \ge J(y)$ *if* $|x| \le |y|$. *Let* $T > 0$ *and* $u_0 \in L^p(\Omega)$. *Let* u_ε *be the unique solution of* $P_p^{J_{p,\varepsilon}}(u_0, \psi)$ *and* u *the unique solution of* $D_p(u_0, \tilde{\psi})$ *(see section* 2.2*). Then*

$$\lim_{\varepsilon\to 0} \sup_{t\in[0,T]} \|u_\varepsilon(t, .) - u(t, .)\|_{L^p(\Omega)} = 0. \tag{1.3}$$

Note that the above result says that P_p^J is a nonlocal problem analogous to the p-Laplacian with Dirichlet boundary condition.

The second goal of this paper is to study the Dirichlet problem for $p = 1$, called the nonlocal total variation flow, which can be written formally as

$$P_1^J(u_0, \psi) \begin{cases} u_t(t, x) = \displaystyle\int_\Omega J(x - y)\frac{u(t, y) - u(t, x)}{|u(t, y) - u(t, x)|}\, dy \\ \qquad\qquad + \displaystyle\int_{\Omega_J\setminus\overline{\Omega}} J(x - y)\frac{\psi(t, y) - u(t, x)}{|\psi(t, y) - u(t, x)|}\, dy, \quad (t, x) \in]0, T[\times\Omega, \\ u(0, x) = u_0(x), \quad x \in \Omega. \end{cases}$$

We give the following definition of what we understand by a solution of $P_1^J(u_0, \psi)$.

DEFINITION 1.4. *A solution of* $P_1^J(u_0, \psi)$ *in* $[0, T]$ *is a function*

$$u \in C\left([0, T]; L^1(\Omega)\right) \cap W^{1,1}\left(]0, T[; L^1(\Omega)\right),$$

which satisfies $u(0, x) = u_0(x)$ *a.e.* $x \in \Omega$ *and*

$$u_t(t, x) = \int_{\Omega_J} J(x - y)g(t, x, y)\, dy \quad \text{a.e. in }]0, T[\times\Omega,$$

for some $g \in L^\infty(0,T;L^\infty(\Omega_J \times \Omega))$ *with* $\|g\|_\infty \le 1$ *such that for almost every* $t \in]0,T[$, $g(t,x,y) = -g(t,y,x)$ *and*

$$J(x-y)g(t,x,y) \in J(x-y)\mathrm{sign}(u(t,y)-u(t,x)), \qquad (x,y) \in \Omega \times \Omega,$$

$$J(x-y)g(t,x,y) \in J(x-y)\mathrm{sign}(\psi(y)-u(t,x)), \qquad (x,y) \in \Omega \times \left(\Omega_J \setminus \overline{\Omega}\right).$$

Here, sign is the multivalued function defined by

$$\mathrm{sign}(r) := \begin{cases} 1 & \text{if } r > 0 \\ [-1,1] & \text{if } r = 0 \\ -1 & \text{if } r < 0. \end{cases}$$

We use sign_0 to denote the univalued function

$$\mathrm{sign}_0(r) := \begin{cases} 1 & \text{if } r > 0 \\ 0 & \text{if } r = 0 \\ -1 & \text{if } r < 0. \end{cases}$$

To get the existence and uniqueness of these kinds of solutions, the idea is to take the limit as $p \searrow 1$ of solutions to P_p^J with $p > 1$.

THEOREM 1.5. *Let* $u_0 \in L^1(\Omega)$ *and* $\psi \in L^1(\Omega_J \setminus \overline{\Omega})$. *Then, there exists a unique solution to* $P_1^J(u_0)$ *in the sense of Definition* 1.4. *Moreover, if* $u_{i0} \in L^1(\Omega)$ *and* u_i *are solutions in* $[0,T]$ *of* $P_1^J(u_{i0})$, $i = 1,2$. *Then*

$$\int_\Omega (u_1(t) - u_2(t))^+ \le \int_\Omega (u_{10} - u_{20})^+ \quad \text{for every } t \in [0,T].$$

In this case we can rescale the kernel as in (1.2) in order to obtain convergence of the solutions of the corresponding rescaled problem to the solution of the Dirichlet problem for the total variational flow, that is,

$$D_1(u_0,\tilde{\psi}) \quad \begin{cases} u_t = \mathrm{div}\left(\frac{Du}{|Du|}\right) & \text{in }]0,T[\times\Omega, \\ u = \tilde{\psi} & \text{on }]0,T[\times\partial\Omega, \\ u(0,x) = u_0(x) & \text{in } \Omega. \end{cases}$$

THEOREM 1.6. *Let* Ω *be a smooth bounded domain in* $\mathbb{R}^N$. *Assume* $J(x) \ge J(y)$ *if* $|x| \le |y|$. *Let* $T > 0$, $u_0 \in L^1(\Omega)$, $\tilde{\psi} \in L^\infty(\partial\Omega)$, *and* $\psi \in W^{1,1}(\Omega_J \setminus \overline{\Omega}) \cap L^\infty(\Omega_J \setminus \overline{\Omega})$ *such that the trace* $\psi|_{\partial\Omega} = \tilde{\psi}$. *Let* u_ε *be the unique solution of* $P_1^{J_{1,\varepsilon}}(u_0,\psi)$. *Then, if* u *is the unique solution of* $D_1(u_0,\tilde{\psi})$ *(see section* 3.2*),*

$$\lim_{\varepsilon\to 0} \sup_{t\in[0,T]} \|u_\varepsilon(t,.) - u(t,.)\|_{L^1(\Omega)} = 0.$$

Finally, the third goal of this paper is to study the limit case $p = +\infty$, which has to be understood as the limit of our nonlocal evolution problems as $p \to +\infty$ (see section 4). In this case we recover a nonlocal model for the evolution of sandpiles which is the nonlocal version of the Prigozhin model [35]. Then, the nonlocal limit problem with source for $p = +\infty$ can be written as

$$P_\infty^J(u_0,\psi,f) \quad \begin{cases} f(t,\cdot) - u_t(t,\cdot) \in \partial G^J_{\infty,\psi}(u(t)), & \text{a.e. } t \in]0,T[, \\ u(0,x) = u_0(x), \end{cases}$$

where $G^J_{\infty,\psi}$ is the functional

$$G^J_{\infty,\psi}(u) = \begin{cases} 0 & \text{if } |u(x)-u(y)| \le 1, \text{ for } x,y \in \Omega \\ & \text{and } |\psi(y)-u(x)| \le 1, \quad \text{for } x \in \Omega, y \in \Omega_J \setminus \overline{\Omega}, \\ & \text{with } x-y \in \operatorname{supp}(J) \\ +\infty & \text{in the other case,} \end{cases}$$

that is, $G^J_{\infty,\psi} = I_{K^J_{\infty,\psi}}$, the indicator function of the set

$$K^J_{\infty,\psi} := \left\{ u \in L^2(\Omega) \ : \ \begin{array}{l} |u(x)-u(y)| \le 1, x, y \in \Omega \\ \text{and } |\psi(y)-u(x)| \le 1, \quad \text{for } x \in \Omega, y \in \Omega_J \setminus \overline{\Omega}, \\ \text{with } x-y \in \operatorname{supp}(J) \end{array} \right\}.$$

More precisely, we obtain the following result.

THEOREM 1.7. *Let $\psi \in L^\infty(\Omega_J \setminus \overline{\Omega})$ such that $K^J_{\infty,\psi} \neq \emptyset$. Let $T > 0$, $f \in L^2(0,T; \cap_{q\ge 2} L^q(\Omega))$, $u_0 \in \cap_{q\ge 2} L^q(\Omega)$ such that $u_0 \in K^J_{\infty,\psi}$, and u_p, $p \ge 2$, the unique solution of the nonlocal p-Laplacian with a source term f, $P^J_p(u_0,\psi,f)$ (see section 4). Then, if u_∞ is the unique solution to $P^J_\infty(u_0,\psi,f)$,*

$$\lim_{p\to\infty} \sup_{t\in[0,T]} \|u_p(t,\cdot) - u_\infty(t,\cdot)\|_{L^2(\Omega)} = 0.$$

Our next step is to rescale the kernel J appropriately and take the limit as the scaling parameter goes to zero. We will suppose that Ω is convex and ψ verifies $\|\nabla\psi\|_\infty \le 1$. For $\varepsilon > 0$, we rescale the functional $G^J_{\infty,\psi}$ as follows:

$$G^\varepsilon_{\infty,\psi}(u) = \begin{cases} 0 & \text{if } |u(x)-u(y)| \le \varepsilon, \text{ for } x,y \in \Omega \\ & \text{and } |\psi(y)-u(x)| \le \varepsilon, \quad \text{for } x \in \Omega, y \in \Omega_J \setminus \overline{\Omega}, \\ & \text{with } |x-y| \le \varepsilon \\ +\infty & \text{in the other case,} \end{cases}$$

that is, $G^\varepsilon_{\infty,\psi} = I_{K^\varepsilon_{\infty,\psi}}$, where

$$K^\varepsilon_{\infty,\psi} := \left\{ u \in L^2(\Omega) \ : \ \begin{array}{l} |u(x)-u(y)| \le \varepsilon, x, y \in \Omega \\ \text{and } |\psi(y)-u(x)| \le \varepsilon, \quad \text{for } x \in \Omega, y \in \Omega_J \setminus \overline{\Omega}, \\ \text{with } |x-y| \le \varepsilon \end{array} \right\}.$$

Consider the gradient flow associated to the functional $G^\varepsilon_{\infty,\psi}$

$$P^\varepsilon_\infty(u_0,\psi,f) \quad \begin{cases} f(t,\cdot) - u_t(t,\cdot) \in \partial I_{K^\varepsilon_{\infty,\psi}}(u(t)), & \text{a.e. } t \in]0,T[, \\ u(0,x) = u_0(x), & \text{in } \Omega, \end{cases}$$

and the limit problem

$$P_\infty(u_0,\psi,f) \quad \begin{cases} f(t,\cdot) - u_{\infty,t} \in \partial I_{K_\psi}(u_\infty), & \text{a.e. } t \in]0,T[, \\ u_\infty(0,x) = u_0(x), & \text{in } \Omega, \end{cases}$$

where

$$K_\psi := \left\{ u \in W^{1,\infty}(\Omega) \ : \ \|\nabla u\|_\infty \le 1, \ u|_{\partial\Omega} = \psi|_{\partial\Omega} \right\}.$$

Now we state our result concerning the limit as $\varepsilon \to 0$ for the sandpile model ($p = +\infty$).

THEOREM 1.8. *Assume Ω is a convex bounded domain in $\mathbb{R}^N$. Let $T > 0$, $f \in L^2(0,T;L^2(\Omega))$, $\psi \in W^{1,\infty}(\Omega_J \setminus \overline{\Omega})$ such that $\|\nabla \psi\|_\infty \le 1$, $u_0 \in W^{1,\infty}(\Omega)$ such that $\|\nabla u_0\|_\infty \le 1$ and $u_0|_{\partial\Omega} = \psi|_{\partial\Omega}$ (this means $u_0 \in K_\psi$), and consider $u_{\infty,\varepsilon}$ the unique solution of $P^\varepsilon_\infty(u_0,\psi,f)$. Then, if v_∞ is the unique solution of $P_\infty(u_0,\psi,f)$, we have*

$$\lim_{\varepsilon\to 0} \sup_{t\in[0,T]} \|u_{\infty,\varepsilon}(t,\cdot) - v_\infty(t,\cdot)\|_{L^2(\Omega)} = 0.$$

Closely related to the present work are [5] and [6] where the homogeneous Neumann problem and its limit as p goes to infinity or to one are considered. The difference here is that we are now considering Dirichlet boundary conditions, not only the homogeneous case, but also the nonhomogeneous case, and this introduces new difficulties specially when one tries to recover the local models when $\varepsilon \to 0$. Remark that in our nonlocal formulation we are not imposing any continuity between the values of u inside Ω and outside it, ψ. However, when dealing with local problems usually the boundary datum is taken in the sense of traces, that is, $u|_{\partial\Omega} = \psi$. Recovering this condition as $\varepsilon \to 0$ is one of the main contributions of the present work.

Note that, as it happens for the local p-Laplacian, the Dirichlet problem can be written as a Neumann problem with a particular flux that depends on the solution itself. Indeed, the problem $P^J_p(u_0,\psi)$ can be written as

$$\begin{cases} u_t(t,x) = \displaystyle\int_\Omega J(x-y)|u(t,y)-u(t,x)|^{p-2}(u(t,y)-u(t,x))\,dy + \varphi(x,u(x)) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (t,x) \in]0,T[\times\Omega, \\ u(0,x) = u_0(x), \quad x \in \Omega, \end{cases}$$

where

$$\varphi(x,u(x)) = \int_{\Omega_J\setminus\overline{\Omega}} J(x-y)|\psi(t,y)-u(t,x)|^{p-2}(\psi(t,y)-u(t,x))\,dy.$$

In the homogeneous case, $\psi \equiv 0$,

$$\varphi(x,u(x)) = -\left(\int_{\Omega_J\setminus\overline{\Omega}} J(x-y)\,dy\right)|u(t,x)|^{p-2}u(t,x).$$

This problem is a nonhomogeneous Neumann problem (see [5]) with a prescribed flux given by φ.

Let us finish the introduction by collecting some notations and results that will be used in the sequel. Following [7] (see also [2]), let

$$X(\Omega) = \left\{ z \in L^\infty(\Omega,\mathbb{R}^n) \ : \ \mathrm{div}(z) \in L^1(\Omega) \right\}. \tag{1.4}$$

If $z \in X(\Omega)$ and $w \in BV(\Omega)\cap L^\infty(\Omega)$, we define the functional $(z,Dw) : C_0^\infty(\Omega) \to \mathbb{R}$ by the formula

$$\langle (z,Dw),\varphi\rangle = -\int_\Omega w\,\varphi\,\mathrm{div}(z)\,dx - \int_\Omega w\,z\cdot\nabla\varphi\,dx. \tag{1.5}$$

Then (z, Dw) is a Radon measure in Ω,

$$\int_\Omega (z, Dw) = \int_\Omega z \cdot \nabla w \, dx \tag{1.6}$$

for all $w \in W^{1,1}(\Omega) \cap L^\infty(\Omega)$ and

$$\left| \int_B (z, Dw) \right| \leq \int_B |(z, Dw)| \leq \|z\|_\infty \int_B \|Dw\| \tag{1.7}$$

for any Borel set $B \subseteq \Omega$.

In [7], a weak trace on $\partial\Omega$ of the normal component of $z \in X(\Omega)$ is defined. Concretely, it is proved that there exists a linear operator $\gamma : X(\Omega) \to L^\infty(\partial\Omega)$ such that

$$\|\gamma(z)\|_\infty \leq \|z\|_\infty,$$

and

$$\gamma(z)(x) = z(x) \cdot \nu(x) \quad \text{for all } x \in \partial\Omega \text{ if } z \in C^1\left(\overline{\Omega}, \mathbb{R}^N\right).$$

We shall denote $\gamma(z)(x)$ by $[z, \nu](x)$. Moreover, the following *Green's formula*, relating the function $[z, \nu]$ and the measure (z, Dw), for $z \in X(\Omega)$ and $w \in BV(\Omega) \cap L^\infty(\Omega)$, is established:

$$\int_\Omega w \, \operatorname{div}(z) \, dx + \int_\Omega (z, Dw) = \int_{\partial\Omega} [z, \nu] w \, d\mathcal{H}^{N-1}. \tag{1.8}$$

Organization of the paper. The rest of the paper is organized as follows. In the second section we prove the existence and uniqueness of strong solutions for the nonlocal p-Laplacian problem with Dirichlet boundary conditions for $p > 1$ and we show that our model approaches local p-Laplacian evolution equation with Dirichlet boundary condition. In section 3 we study the Dirichlet problem for the nonlocal total variation flow, proving convergence to the local model when the problem is rescaled appropriately as well. Finally, in section 4 we study the case $p = \infty$, obtaining a model for sandpiles with Dirichlet boundary conditions.

2. The case $p > 1$.

2.1. Existence of solutions for the nonlocal problems. We first study $P_p^J(u_0, \psi)$ from the point of view of nonlinear semigroup theory ([15], [28]). For that we introduce in $L^1(\Omega)$ the following operator associated with our problem.

DEFINITION 2.1. For $1 < p < +\infty$ and $\psi : \Omega_J \setminus \overline{\Omega} \to \mathbb{R}$, such that $|\psi|^{p-1} \in L^1(\Omega_J \setminus \overline{\Omega})$, we define in $L^1(\Omega)$ the operator $B_{p,\psi}^J$ by

$$\begin{aligned} B_{p,\psi}^J(u)(x) \quad &= -\int_\Omega J(x-y)|u(y)-u(x)|^{p-2}(u(y)-u(x))\,dy \\ &\quad -\int_{\Omega_J \setminus \overline{\Omega}} J(x-y)|\psi(y)-u(x)|^{p-2}(\psi(y)-u(x))\,dy, \qquad x \in \Omega. \end{aligned}$$

Remark 2.2. (i). We will set overall the section,

$$u_\psi(x) := \begin{cases} u(x) & \text{if } x \in \Omega, \\ \psi(x) & \text{if } x \in \Omega_J \setminus \overline{\Omega}, \\ 0 & \text{if } x \notin \Omega_J. \end{cases}$$

Therefore, we can rewrite

$$B^J_{p,\psi}(u)(x) = -\int_{\Omega_J} J(x-y)|u_\psi(y)-u(x)|^{p-2}(u_\psi(y)-u(x))\,dy, \qquad x\in\Omega.$$

(ii) If $\psi = 0$, then

$$B^J_{p,0}(u)(x) = -\int_{\Omega} J(x-y)|u(y)-u(x)|^{p-2}(u(y)-u(x))\,dy + \left(\int_{\Omega_J\setminus\overline{\Omega}} J(x-y)dy\right)\ |u(x)|^{p-2}u(x)\,, \qquad x\in\Omega.$$

Remark 2.3. It is easy to see that
(i) If $\psi = 0$, $B^J_{p,0}$ is positively homogeneous of degree $p-1$,
(ii) $L^{p-1}(\Omega)\subset \mathrm{Dom}(B^J_{p,\psi})$, if $p>2$.
(iii) For $1<p\le 2$, $\mathrm{Dom}(B^J_{p,\psi}) = L^1(\Omega)$ and $B^J_{p,\psi}$ is closed in $L^1(\Omega)\times L^1(\Omega)$.
We have the following monotonicity lemma, whose proof is straightforward.

LEMMA 2.4. *Let $1<p<+\infty$, $\psi:\Omega_J\setminus\overline{\Omega}\to\mathbb{R}$, $|\psi|^{p-1}\in L^1(\Omega_J\setminus\overline{\Omega})$, and $T:\mathbb{R}\to\mathbb{R}$ a nondecreasing function. Then,*

(i) *for every u, $v\in L^p(\Omega)$ such that $T(u-v)\in L^p(\Omega)$, it holds*

$$\begin{aligned}&\int_\Omega \left(B^J_{p,\psi}u(x) - B^J_{p,\psi}v(x)\right)T(u(x)-v(x))dx\\ &= \frac{1}{2}\int_{\Omega_J}\int_{\Omega_J} J(x-y)\left(T(u_\psi(y)-v_\psi(y)) - T(u_\psi(x)-v_\psi(x))\right)\\ &\qquad\times\left(|u_\psi(y)-u_\psi(x)|^{p-2}(u_\psi(y)-u_\psi(x))\right.\\ &\qquad\quad\left.-|v_\psi(y)-v_\psi(x)|^{p-2}(v_\psi(y)-v_\psi(x))\right)dydx.\end{aligned} \tag{2.1}$$

(ii) *Moreover, if T is bounded,* (2.1) *holds for u, $v\in Dom(B^J_{p,\psi})$.*

We have the following Poincaré's type inequality.

PROPOSITION 2.5. *Given Ω a bounded domain in $\mathbb{R}^N$, $J:\mathbb{R}^N\to\mathbb{R}$ a nonnegative, radial, continuous function, such that $\int_{\mathbb{R}^N} J(z)\,dz>0$, $p\ge 1$ and $\psi\in L^p(\Omega_J\setminus\overline{\Omega})$, there exists $\lambda=\lambda(J,\Omega,p)>0$ such that*

$$\lambda\int_\Omega |u(x)|^p\,dx \le \int_\Omega\int_{\Omega_J} J(x-y)|u_\psi(y)-u(x)|^p\,dy\,dx + \int_{\Omega_J\setminus\overline{\Omega}}|\psi(y)|^p\,dy \tag{2.2}$$

for all $u\in L^p(\Omega)$.

Proof. First, let us assume that there exist $r,\alpha>0$ such that $J(x)\ge\alpha$ in $B(0,r)$. Let

$$B_0 = \{x\in\Omega_J\setminus\overline{\Omega}\ :\ d(x,\Omega)\le r/2\},$$
$$B_1 = \{x\in\Omega\ :\ d(x,B_0)\le r/2\},$$
$$B_j = \left\{x\in\Omega\setminus\cup_{k=1}^{j-1}B_k\ :\ d(x,B_{j-1})\le r/2\right\},\quad j=2,3,\dots$$

Observe that we can cover Ω by a finite number of nonnull sets $\{B_j\}_{j=1}^{l_r}$. Now

$$\int_\Omega\int_{\Omega_J} J(x-y)|u_\psi(y)-u(x)|^p\,dy\,dx \ge \int_{B_j}\int_{B_{j-1}} J(x-y)|u_\psi(y)-u(x)|^p\,dy\,dx,$$

$j = 1, \dots, l_r$, and

$$\begin{aligned}
&\int_{B_j}\int_{B_{j-1}} J(x-y)|u_\psi(y)-u(x)|^p\,dy\,dx \\
&\geq \frac{1}{2^p}\int_{B_j}\int_{B_{j-1}} J(x-y)|u(x)|^p\,dy\,dx - \int_{B_j}\int_{B_{j-1}} J(x-y)|u_\psi(y)|^p\,dy\,dx \\
&= \frac{1}{2^p}\int_{B_j}\left(\int_{B_{j-1}} J(x-y)\,dy\right)|u(x)|^p\,dx - \int_{B_{j-1}}\left(\int_{B_j} J(x-y)\,dx\right)|u_\psi(y)|^p\,dy \\
&\geq \frac{1}{2^p}\min_{x\in\overline{B}_j}\int_{B_{j-1}} J(x-y)dy\int_{B_j}|u(x)|^p\,dx - \beta\int_{B_{j-1}}|u_\psi(y)|^p\,dy,
\end{aligned}$$

where $\beta = \int_{\mathbb{R}^N} J(x)dx$. Hence

$$\int_\Omega\int_{\Omega_J} J(x-y)|u_\psi(y)-u(x)|^p\,dy\,dx \geq \alpha_j\int_{B_j}|u(x)|^p\,dx - \beta\int_{B_{j-1}}|u_\psi(y)|^p\,dy,$$

where

$$\alpha_j = \frac{1}{2^p}\min_{x\in\overline{B}_j}\int_{B_{j-1}} J(x-y)dy > 0.$$

Therefore, since $u_\psi(y) = \psi(y)$ if $y \in B_0$, $u_\psi(y) = u(y)$ if $y \in B_j$, $j = 1, \dots, l_r$, $B_j \cap B_i = \emptyset$, for all $i \neq j$ and $|\Omega \setminus \cup_{j=1}^{j_r} B_j| = 0$, it is easy to see that there exists $\hat{\lambda} = \hat{\lambda}(J, \Omega, p) > 0$ such that

$$\int_\Omega |u|^p \leq \hat{\lambda}\int_\Omega\int_{\Omega_J} J(x-y)|u_\psi(y)-u(x)|^p\,dy\,dx + \hat{\lambda}\int_{B_0}|\psi|^p.$$

The proof is finished by taking $\lambda = \hat{\lambda}^{-1}$.

In the general case we have that there exist $\mathtt{a} \geq 0$ and $r, \alpha > 0$ such that

$$J(x) \geq \alpha \text{ in the annulus } A(0, \mathtt{a}, r). \tag{2.3}$$

In this case we proceed as before with the same choice of the sets B_j for $j \geq 0$ and

$$B_{-j} = \left\{x \in \Omega_J \setminus \left(\Omega \cup \cup_{k=0}^{j-1} B_{-k}\right) \;:\; d(x, B_{-j+1}) \leq r/2\right\}, \quad j = 1, 2, 3, \dots$$

Observe that for each B_j, $j \geq 1$, there exists B_{j^e} with $j^e < j$ and such that

$$|(x + A(0, \mathtt{a}, r)) \cap B_{j^e}| > 0 \quad \forall x \in \overline{B}_j. \tag{2.4}$$

With this choice of B_j and taking into account (2.3) and (2.4), as before, we obtain

$$\begin{aligned}
\int_\Omega\int_{\Omega_J} J(x-y)|u_\psi(y)-u(x)|^p\,dy\,dx &\geq \int_{B_j}\int_{B_{j^e}} J(x-y)|u_\psi(y)-u(x)|^p\,dy\,dx \\
&\geq \alpha_j\int_{B_j}|u(x)|^p\,dx - \beta\int_{B_{j^e}}|u_\psi(y)|^p\,dy,
\end{aligned}$$

$j = 1, \dots, l_r$, where

$$\alpha_j = \frac{1}{2^p}\min_{x\in\overline{B}_j}\int_{B_{j^e}} J(x-y)dy > 0$$

and $\beta = \int_{\mathbb{R}^N} J(x)dx$. And we conclude as before. ∎

Remark 2.6. Note that in [5] it is proved a Poincare's type inequality for Neumann boundary conditions, but assuming that $J(0) > 0$ (otherwise there is a counterexample). Surprisingly, for the Dirichlet problem we do not need positivity at the origin for J. This is due to the fact that for the Dirichlet problem the outside values influence the inside values.

In the next result we prove that $B^J_{p,\psi}$ is a completely accretive operator (see [14]) and verifies a range condition. In short, this means that for any $\phi \in L^p(\Omega)$ there is a unique solution of the problem $u + B^J_{p,\psi}u = \phi$ and the resolvent $(I + B^J_{p,\psi})^{-1}$ is a contraction in $L^q(\Omega)$ for all $1 \le q \le +\infty$.

THEOREM 2.7. *Let $1 < p < +\infty$. For $\psi \in L^p(\Omega_J \setminus \overline{\Omega})$, the operator $B^J_{p,\psi}$ is completely accretive and verifies the range condition*

$$L^p(\Omega) \subset Ran\left(I + B^J_{p,\psi}\right). \tag{2.5}$$

Proof. Given $u_i \in \mathrm{Dom}(B^J_{p,\psi})$, $i = 1, 2$, by the monotonicity Lemma 2.4, for any $q \in C^\infty(\mathbb{R})$, $0 \le q' \le 1$, $\mathrm{supp}(q')$ compact, $0 \notin \mathrm{supp}(q)$, we have that

$$\int_\Omega \left(B^J_{p,\psi}u_1(x) - B^J_{p,\psi}u_2(x)\right) q(u_1(x) - u_2(x))\, dx \ge 0,$$

from where it follows that $B^J_{p,\psi}$ is a completely accretive operator (see [14]).

To show that $B^J_{p,\psi}$ satisfies the range condition we have to prove that for any $\phi \in L^p(\Omega)$ there exists $u \in \mathrm{Dom}(B^J_{p,\psi})$ such that $\phi = u + B^J_{p,\psi}u$.

Assume first $p \ge 2$. Let $\phi \in L^p(\Omega)$ and set

$$K = \left\{w \in L^p(\Omega_J) \ : \ w = \psi \text{ in } \Omega_J \setminus \overline{\Omega}\right\}.$$

We consider the continuous monotone operator $A : K \to L^{p'}(\Omega_J)$ defined by

$$A(w)(x) := w(x) - \int_{\Omega_J} J(x-y)|w(y) - w(x)|^{p-2}(w(y) - w(x))\, dy.$$

A is coercive in $L^p(\Omega_J)$. In fact, by Proposition 2.5, for any $w \in K$,

$$\begin{aligned}
&\int_{\Omega_J} A(w)w = \int_{\Omega_J} w^2 - \int_{\Omega_J}\int_{\Omega_J} J(x-y)|w(y) - w(x)|^{p-2}(w(y) - w(x))\, dy w(x) dx \\
&\ge \frac{1}{2}\int_{\Omega_J}\int_{\Omega_J} J(x-y)|w(y) - w(x)|^p\, dy dx \\
&\ge \frac{1}{2}\int_{\Omega}\int_{\Omega_J} J(x-y)|w_\psi(y) - w(x)|^p\, dy dx \ge \frac{\lambda}{2}\|w\|^p_{L^p(\Omega)} - \frac{1}{2}\int_{\Omega_J\setminus\overline{\Omega}} |\psi|^p.
\end{aligned}$$

Therefore,

$$\lim_{\substack{\|w\|_{L^p(\Omega_J)} \to +\infty \\ w \in K}} \frac{\displaystyle\int_{\Omega_J} A(w)w}{\|w\|_{L^p(\Omega_J)}} = +\infty.$$

Now, since $p \ge 2$, we have the function $\phi_\psi \in L^{p'}(\Omega_J)$. Then, applying [32, Corollary III.1.8] to the operator $B(w) := A(w) - \phi_\psi$, we get there exists $w \in K$, such that

$$w(x) - \int_{\Omega_J} J(x-y)|w(y) - w(x)|^{p-2}(w(y) - w(x))\, dy = \phi_\psi(x) \qquad \text{for all } \ x \in \Omega_J.$$

Hence, $u := w_{|\Omega}$ satisfies

$$u(x) - \int_{\Omega_J} J(x-y)|u_\psi(y) - u(x)|^{p-2}(u_\psi(y) - u(x))\,dy = \phi(x) \qquad \text{for all } \ x \in \Omega,$$

and, consequently, $\phi = u + B^J_{p,\psi}u$.

Suppose now $1 < p < 2$. By the results in [5], we know that the operator

$$B^J_p u(x) = -\int_\Omega J(x-y)|u(y) - u(x)|^{p-2}(u(y) - u(x))\,dy$$

is m-accretive in $L^1(\Omega)$ and satisfies what is called property (M_0); that is, for any $q \in C^\infty(\mathbb{R})$, $0 \le q' \le 1$, $\operatorname{supp}(q')$ compact, $0 \notin \operatorname{supp}(q)$, and $(u, v) \in B^J_p$,

$$\int_\Omega q(u)v \ge 0.$$

On the other hand,

$$\varphi(x, r) = -\int_{\Omega_J \setminus \overline{\Omega}} J(x-y)|\psi(y) - r|^{p-2}(\psi(y) - r)\,dy$$

is continuous and nondecreasing in r for almost every $x \in \Omega$, and an $L^1(\Omega)$ function for all r. Therefore, by [3, Theorem 3.1], $B^J_{p,\psi}u(x) = B^J_p u(x) + \varphi(x, u(x))$ is m-accretive in $L^1(\Omega)$. □

Remark 2.8. If $\mathcal{B}^J_{p,\psi}$ denotes the closure of $B^J_{p,\psi}$ in $L^1(\Omega)$, by Theorem 2.7, we have $\mathcal{B}^J_{p,\psi}$ is m-completely accretive in $L^1(\Omega)$ (see [14]). Therefore, by the nonlinear semigroup theory (see [15] and [14]), there exists an unique mild-solution of the abstract Cauchy problem

$$\begin{cases} u'(t) + B^J_{p,\psi}u(t) = 0, & t \in (0, T), \\ u(0) = u_0, \end{cases} \tag{2.6}$$

given by the Crandall–Liggett exponential formula

$$\mathrm{e}^{-t\mathcal{B}^J_{p,\psi}}u_0 = \lim_n \left(I + \frac{t}{n}\mathcal{B}^J_{p,\psi}\right)^{-n} u_0.$$

Now, due to regularity results for mild solutions, under certain hypothesis, this mild solution is a strong solution of the abstract Cauchy problem (2.6) (see [14]) which means, for our problem $P^J_p(u_0, \psi)$, a solution in the sense of Definition 1.1.

The following result states the existence and uniqueness results for $P^J_p(u_0, \psi)$. From it, Theorem 1.2 can be derived.

THEOREM 2.9. *Assume $p > 1$. Let $T > 0$, $\psi \in L^p(\Omega_J \setminus \overline{\Omega})$, and $u_0 \in L^1(\Omega)$. Then, there exists a unique mild-solution u of* (2.6). *Moreover,*

(1) *if $u_0 \in L^p(\Omega)$, the unique mild solution u of* (2.6) *is a solution of $P^J_p(u_0, \psi)$ in the sense of Definition* 1.1. *If $1 < p \le 2$, this is true for any $u_0 \in L^1(\Omega)$ and any ψ such that $|\psi|^{p-1} \in L^1(\Omega_J \setminus \overline{\Omega})$.*

(2) *Let $u_{i0} \in L^1(\Omega)$ and u_i a solution in $[0, T]$ of $P^J_p(u_{i0})$, $i = 1, 2$. Then*

$$\int_\Omega (u_1(t) - u_2(t))^+ \le \int_\Omega (u_{10} - u_{20})^+ \quad \text{for every } t \in]0, T[.$$

Moreover, for $q \in [1, +\infty]$, if $u_{i0} \in L^q(\Omega)$, $i = 1, 2$, then

$$\|u_1(t) - u_2(t)\|_{L^q(\Omega)} \le \|u_{10} - u_{20}\|_{L^q(\Omega)} \quad \text{for every } t \in]0, T[.$$

Proof. As a consequence of Theorem 2.7 we get the existence of mild solution of (2.6) (see Remark 2.8). Now, due to the complete accretivity of $B^J_{p,\psi}$ and the range condition (2.5), by regularity results for mild solutions (see [14]), $u(t)$ is a strong solution, that is, a solution of $P^J_p(u_0, \psi)$ in the sense of Definition 1.1. Moreover, in the case $1 < p \le 2$, since $\mathrm{Dom}(B^J_{p,\psi}) = L^1(\Omega)$ and $B^J_{p,\psi}$ is closed in $L^1(\Omega) \times L^1(\Omega)$, the result holds for L^1-data. Finally, the contraction principle is a consequence of the general nonlinear semigroup theory ([15], [28]). □

2.2. Convergence to the *p*-Laplacian. Our main goal in this section is to show that the solution to the Dirichlet problem for the p-Laplacian equation $D_p(u_0, \tilde{\psi})$ can be approximated by solutions to suitable nonlocal Dirichlet problems $P^J_p(u_0, \psi)$.

Let us first recall the following result from [5]. For a function g defined in a set D, we define

$$\overline{g}(x) = \begin{cases} g(x) & \text{if } x \in D, \\ 0 & \text{otherwise,} \end{cases}$$

and we denote by χ_D the characteristic function of D.

PROPOSITION 2.10 ([5]). *Let $1 \le q < +\infty$, D a bounded domain in $\mathbb{R}^N$, $\rho : \mathbb{R}^N \to \mathbb{R}$ a nonnegative continuous radial function with compact support, nonidentically zero, and $\rho_n(x) := n^N \rho(nx)$. Let $\{f_n\}$ be a sequence of functions in $L^q(D)$ such that*

$$\int_D \int_D |f_n(y) - f_n(x)|^q \rho_n(y - x)\, dx\, dy \le M \frac{1}{n^q}. \tag{2.7}$$

1. *If $\{f_n\}$ is weakly convergent in $L^q(D)$ to f, then*
   (i) *if $q > 1$, $f \in W^{1,q}(D)$ and moreover*

$$(\rho(z))^{1/q} \chi_D \left(x + \frac{1}{n} z\right) \frac{\overline{f}_n\left(x + \frac{1}{n} z\right) - f_n(x)}{1/n} \rightharpoonup (\rho(z))^{1/q}\, z \cdot \nabla f(x)$$

weakly in $L^q(D) \times L^q(\mathbb{R}^N)$;
   (ii) *if $q = 1$, $f \in BV(D)$ and moreover*

$$\rho(z) \chi_D \left(\cdot + \frac{1}{n} z\right) \frac{\overline{f}_n\left(\cdot + \frac{1}{n} z\right) - f_n(\cdot)}{1/n} \rightharpoonup \rho(z) z \cdot Df$$

weakly as measures.

2. *Assume D is a smooth bounded domain in $\mathbb{R}^N$ and $\rho(x) \ge \rho(y)$ if $|x| \le |y|$. Then $\{f_n\}$ is relatively compact in $L^q(D)$ and, consequently, there exists a subsequence $\{f_{n_k}\}$ such that*
   (i) *if $q > 1$, $f_{n_k} \to f$ in $L^q(D)$ with $f \in W^{1,q}(D)$;*
   (ii) *If $q = 1$, $f_{n_k} \to f$ in $L^1(D)$ with $f \in BV(D)$.*

Let us now recall some results about the p-Laplacian equation

$$D_p(u_0, \tilde{\psi}) \quad \begin{cases} u_t = \Delta_p u & \text{in }]0, T[\times\Omega, \\ u = \tilde{\psi} & \text{on }]0, T[\times\partial\Omega, \\ u(0, x) = u_0(x) & \text{in } \Omega. \end{cases}$$

In the case $\tilde{\psi} \in W^{1/p',p}(\partial\Omega)$, associated to the p-Laplacian with nonhomogeneous Dirichlet boundary condition, in [2] it is defined the operator $A_{p,\psi} \subset L^1(\Omega) \times L^1(\Omega)$ as $(u,\hat{u}) \in A_{p,\tilde{\psi}}$ if and only if $\hat{u} \in L^1(\Omega)$, $u \in W^{1,p}_{\tilde{\psi}}(\Omega) := \{u \in W^{1,p}(\Omega) : u|_{\partial\Omega} = \tilde{\psi}\ \mathcal{H}^{N-1} - a.e.\ \text{on } \partial\Omega\}$ and

$$\int_\Omega |\nabla u|^{p-2}\nabla u \cdot \nabla(u-v) \le \int_\Omega \hat{u}(u-v) \quad \text{for every} \quad v \in W^{1,p}_{\tilde{\psi}}(\Omega) \cap L^\infty(\Omega).$$

This inequality is equivalent to

$$\int_\Omega |\nabla u|^{p-2}\nabla u \cdot \nabla w = \int_\Omega \hat{u}w \quad \text{for every} \quad w \in W^{1,p}_0(\Omega) \cap L^\infty(\Omega).$$

Moreover, for $\tilde{\psi} \in W^{1/p',p}(\partial\Omega) \cap L^\infty(\partial\Omega)$, $A_{p,\tilde{\psi}}$ is proved to be a completely accretive operator in $L^1(\Omega)$, satisfying the range condition $L^\infty(\Omega) \subset \mathrm{Ran}(I + A_{p,\tilde{\psi}})$, and it is easy to see that $\overline{\mathrm{D}(A_{p,\tilde{\psi}})}^{L^1(\Omega)} = L^1(\Omega)$. Therefore, its closure $\mathcal{A}_{p,\tilde{\psi}}$ in $L^1(\Omega) \times L^1(\Omega)$ is an m-completely accretive operator in $L^1(\Omega)$. Consequently, for any $u_0 \in L^1(\Omega)$ there exists a unique mild solution $u(t) = e^{-t\mathcal{A}_{p,\tilde{\psi}}}u_0$ of the abstract Cauchy problem associated to $D_p(u_0,\tilde{\psi})$, given by Crandall–Liggett's exponential formula. Due to the complete accretivity of the operator $\mathcal{A}_{p,\tilde{\psi}}$, in the case $u_0 \in D(\mathcal{A}_{p,\tilde{\psi}})$ this mild solution is the unique strong solution of problem $D_p(u_0,\tilde{\psi})$.

In the homogeneous case $\tilde{\psi} = 0$, due to the results in [13], we can say that for any $u_0 \in L^1(\Omega)$, the mild solution $u(t) = e^{-t\mathcal{A}_{p,0}}u_0$ is the unique entropy solution of problem $D_p(u_0,0)$.

For given $p > 1$ and J, we consider the rescaled kernels

$$J_{p,\varepsilon}(x) := \frac{C_{J,p}}{\varepsilon^{p+N}} J\left(\frac{x}{\varepsilon}\right), \qquad \text{where} \qquad C^{-1}_{J,p} := \frac{1}{2}\int_{\mathbb{R}^N} J(z)|z_N|^p\,dz$$

is a normalizing constant in order to obtain the p-Laplacian in the limit instead of a multiple of it.

PROPOSITION 2.11. *Let Ω be a smooth bounded domain in $\mathbb{R}^N$ and let $\tilde{\psi} \in W^{1/p',p}(\partial\Omega) \cap L^\infty(\partial\Omega)$. Let $\psi \in W^{1,p}(\Omega_J) \cap L^\infty(\Omega_J)$ such that $\psi|_{\partial\Omega} = \tilde{\psi}$. Assume $J(x) \ge J(y)$ if $|x| \le |y|$. Then, for any $\phi \in L^\infty(\Omega)$,*

$$\left(I + B^{J_{p,\varepsilon}}_{p,\psi}\right)^{-1}\phi \to \left(I + A_{p,\tilde{\psi}}\right)^{-1}\phi \quad \text{in } L^p(\Omega) \text{ as } \varepsilon \to 0. \tag{2.8}$$

Proof. We denote

$$\Omega_\varepsilon := \Omega_{J_{p,\varepsilon}} = \Omega + \mathrm{supp}(J_{p,\varepsilon}).$$

For $\varepsilon > 0$ small, let $u_\varepsilon = \left(I + B^{J_{p,\varepsilon}}_{p,\psi}\right)^{-1}\phi$. Then,

$$\int_\Omega u_\varepsilon v - \frac{C_{J,p}}{\varepsilon^{p+N}} \int_\Omega\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right)|(u_\varepsilon)_\psi(y) - u_\varepsilon(x)|^{p-2} \times ((u_\varepsilon)_\psi(y) - u_\varepsilon(x))\,dy\,v(x)\,dx = \int_\Omega \phi v \tag{2.9}$$

for every $v \in L^\infty(\Omega)$.

Let $M := \max\{\|\phi\|_{L^\infty(\Omega)}, \|\psi\|_{L^\infty(\Omega_J)}\}$. Taking $v = (u_\varepsilon - M)^+$ in (2.9), we get

$$\int_\Omega u_\varepsilon(x)(u_\varepsilon(x)-M)^+dx - \frac{C_{J,p}}{\varepsilon^{p+N}}\int_\Omega\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right)|(u_\varepsilon)_\psi(y)-u_\varepsilon(x)|^{p-2} \times ((u_\varepsilon)_\psi(y)-u_\varepsilon(x))\,dy(u_\varepsilon(x)-M)^+\,dx = \int_\Omega \phi(x)(u_\varepsilon(x)-M)^+\,dx.$$

Now,

$$-\frac{C_{J,p}}{\varepsilon^{p+N}}\int_\Omega\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right)|(u_\varepsilon)_\psi(y)-u_\varepsilon(x)|^{p-2}((u_\varepsilon)_\psi(y)-u_\varepsilon(x))\,dy \times (u_\varepsilon(x)-M)^+dx$$
$$= -\frac{C_{J,p}}{\varepsilon^{p+N}}\int_{\Omega_\varepsilon}\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right)|(u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x)|^{p-2}((u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x))\,dy \times ((u_\varepsilon)_\psi(x)-M)^+dx$$
$$= \frac{C_{J,p}}{2\varepsilon^{p+N}}\int_{\Omega_\varepsilon}\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right)|(u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x)|^{p-2}((u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x)) \times (((u_\varepsilon)_\psi(y)-M)^+-((u_\varepsilon)_\psi(x)-M)^+)\,dy\,dx$$
$$\geq 0.$$

Therefore,

$$\int_\Omega u_\varepsilon(x)(u_\varepsilon(x)-M)^+dx \leq \int_\Omega \phi(x)(u_\varepsilon(x)-M)^+\,dx.$$

Consequently, we have

$$\int_\Omega (u_\varepsilon(x)-M)(u_\varepsilon(x)-M)^+dx \leq \int_\Omega (\phi(x)-M)(u_\varepsilon(x)-M)^+\,dx \leq 0,$$

and $u_\varepsilon(x) \leq M$ for almost all $x \in \Omega$. Analogously, we can obtain $-M \leq u_\varepsilon(x)$ for almost all $x \in \Omega$. Thus

$$\|u_\varepsilon\|_{L^\infty(\Omega)} \leq M \qquad \text{for all } \epsilon > 0, \tag{2.10}$$

and, therefore, there exists a sequence $\varepsilon_n \to 0$ such that

$$u_{\varepsilon_n} \rightharpoonup u \quad \text{weakly in } L^1(\Omega).$$

Taking $v = u_\varepsilon - \psi$ in (2.9) we get

$$\int_\Omega u_\varepsilon(u_\varepsilon-\psi) - \frac{C_{J,p}}{\varepsilon^{p+N}}\int_{\Omega_\varepsilon}\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right)|(u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x)|^{p-2} \times ((u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x))\,dy\,((u_\varepsilon)_\psi(x)-\psi(x))\,dx = \int_\Omega \phi(u_\varepsilon-\psi). \tag{2.11}$$

Now, by (2.11) and (2.10),

$$\frac{C_{J,p}}{2\varepsilon^N}\int_{\Omega_\varepsilon}\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right)\frac{|(u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x)|^p}{\varepsilon^p}\,dy\,dx$$
$$\leq \frac{C_{J,p}}{2\varepsilon^N}\int_{\Omega_\varepsilon}\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right)\frac{|(u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x)|^{p-1}}{\varepsilon^{p-1}}\frac{|\psi(y)-\psi(x)|}{\varepsilon}dy\,dx + M_1.$$

Since $\psi \in W^{1,p}(\Omega_J)$, using Young's inequality, we obtain

$$\frac{1}{\varepsilon^N}\int_{\Omega_\varepsilon}\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right)\frac{|(u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x)|^p}{\varepsilon^p}\,dy\,dx \le M_2.$$

Moreover,

$$\begin{aligned}
&\int_{\Omega_J}\int_{\Omega_J}\frac{1}{\varepsilon^N}J\left(\frac{x-y}{\varepsilon}\right)\left|\frac{(u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x)}{\varepsilon}\right|^p dx\,dy\\
&=\int_{\Omega_\varepsilon}\int_{\Omega_\varepsilon}\frac{1}{\varepsilon^N}J\left(\frac{x-y}{\varepsilon}\right)\left|\frac{(u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x)}{\varepsilon}\right|^p dx\,dy\\
&\quad+2\int_{\Omega_J\setminus\overline{\Omega}_\varepsilon}\int_{\Omega_\varepsilon}\frac{1}{\varepsilon^N}J\left(\frac{x-y}{\varepsilon}\right)\left|\frac{\psi(y)-(u_\varepsilon)_\psi(x)}{\varepsilon}\right|^p dx\,dy\\
&\quad+\int_{\Omega_J\setminus\overline{\Omega}_\varepsilon}\int_{\Omega_J\setminus\overline{\Omega}_\varepsilon}\frac{1}{\varepsilon^N}J\left(\frac{x-y}{\varepsilon}\right)\left|\frac{\psi(y)-\psi(x)}{\varepsilon}\right|^p dx\,dy\\
&=\int_{\Omega_\varepsilon}\int_{\Omega_\varepsilon}\frac{1}{\varepsilon^N}J\left(\frac{x-y}{\varepsilon}\right)\left|\frac{(u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x)}{\varepsilon}\right|^p dx\,dy\\
&\quad+2\int_{\Omega_J\setminus\overline{\Omega}_\varepsilon}\int_{\Omega_\varepsilon\setminus\overline{\Omega}}\frac{1}{\varepsilon^N}J\left(\frac{x-y}{\varepsilon}\right)\left|\frac{\psi(y)-\psi(x)}{\varepsilon}\right|^p dx\,dy\\
&\quad+\int_{\Omega_J\setminus\overline{\Omega}_\varepsilon}\int_{\Omega_J\setminus\overline{\Omega}_\varepsilon}\frac{1}{\varepsilon^N}J\left(\frac{x-y}{\varepsilon}\right)\left|\frac{\psi(y)-\psi(x)}{\varepsilon}\right|^p dx\,dy \le M_3.
\end{aligned}\tag{2.12}$$

Therefore, by Proposition 2.10, there exists a subsequence, denoted as above, and $w \in W^{1,p}(\Omega_J)$ such that

$$(u_{\varepsilon_n})_\psi \to w \quad \text{strongly in } L^p(\Omega_J).$$

Hence, $w = u$ in Ω and, by [18, Proposition IX.18] and the properties of the trace, $u \in W^{1,p}_{\tilde{\psi}}(\Omega)$. Moreover, by Proposition 2.10,

$$\left(\frac{C_{J,p}}{2}J(z)\right)^{1/p}\chi_\Omega(x+\varepsilon_n z)\frac{(u_{\varepsilon_n})_\psi(x+\varepsilon_n z)-(u_{\varepsilon_n})_\psi(x)}{\varepsilon_n} \rightharpoonup \left(\frac{C_{J,p}}{2}J(z)\right)^{1/p} z\cdot\nabla u(x)\tag{2.13}$$

weakly in $L^p(\Omega)\times L^p(\mathbb{R}^N)$ (observe that $\chi_\Omega(x+\varepsilon_n z)(u_{\varepsilon_n})_\psi(x+\varepsilon_n z) = \chi_\Omega(x+\varepsilon_n z)\overline{u}_{\varepsilon_n}(x+\varepsilon_n z)$). We can also assume that

$$\begin{aligned}
(J(z))^{1/p'}&\left|\frac{(u_{\varepsilon_n})_\psi(x+\varepsilon_n z)-(u_{\varepsilon_n})_\psi(x)}{\varepsilon_n}\right|^{p-2}\chi_{\Omega_{\varepsilon_n}}(x+\varepsilon_n z)\\
&\times\frac{(u_{\varepsilon_n})_\psi(x+\varepsilon_n z)-(u_{\varepsilon_n})_\psi(x)}{\varepsilon_n} \rightharpoonup (J(z))^{1/p'}\chi(x,z)
\end{aligned}$$

weakly in $L^{p'}(\Omega_J)\times L^{p'}(\mathbb{R}^N)$, for some function $\chi \in L^{p'}(\Omega_J)\times L^{p'}(\mathbb{R}^N)$.

Passing to the limit in (2.9) for $\varepsilon = \varepsilon_n$, we get

$$\int_\Omega uv + \int_{\mathbb{R}^N}\int_\Omega \frac{C_{J,p}}{2}J(z)\chi(x,z)\,z\cdot\nabla v(x)\,dx\,dz = \int_\Omega \phi v\tag{2.14}$$

for every v smooth with support in Ω and by approximation for every $v \in W^{1,p}_0(\Omega)$.

Finally, working as in Proposition 3.3. of [5], we can prove

$$\int_{\mathbb{R}^N}\int_\Omega \frac{C_{J,p}}{2} J(z)\chi(x,z)z\cdot\nabla v(x)\,dx\,dz = \int_\Omega |\nabla u|^{p-2}\nabla u\cdot\nabla v \tag{2.15}$$

and the proof is finished. □

From the above Proposition, by the standard results of the nonlinear semigroup theory (see [19] or [15]), we obtain Theorem 1.3.

3. The nonlocal total variation flow. The case $p = 1$.

3.1. Existence of solutions for the nonlocal problem. This section deals with the existence and uniqueness of solutions for the nonlocal 1-Laplacian problem with Dirichlet boundary condition,

$$P_1^J(u_0,\psi)\quad \begin{cases} u_t(t,x) = \displaystyle\int_\Omega J(x-y)\frac{u(t,y)-u(t,x)}{|u(t,y)-u(t,x)|}\,dy. \\ \qquad\qquad + \displaystyle\int_{\Omega_J\setminus\overline{\Omega}} J(x-y)\frac{\psi(y)-u(t,x)}{|\psi(y)-u(t,x)|}\,dy, \qquad x\in\Omega. \\ u(0,x) = u_0(x). \end{cases}$$

As in the case $p > 1$, to prove existence and uniqueness of solutions of $P_1^J(u_0,\psi)$ we use the Nonlinear Semigroup Theory, so we start by introducing the following operator in $L^1(\Omega)$.

DEFINITION 3.1. *Given $\psi \in L^1(\Omega_J\setminus\overline{\Omega})$, we define the operator $B_{1,\psi}^J$ in $L^1(\Omega)\times L^1(\Omega)$ by $\hat{u}\in B_{1,\psi}^J u$ if and only if $u,\hat{u}\in L^1(\Omega)$, there exists $g\in L^\infty(\Omega_J\times\Omega_J)$, $g(x,y) = -g(y,x)$ for almost all $(x,y)\in\Omega_J\times\Omega_J$, $\|g\|_\infty\le 1$,*

$$\hat{u}(x) = -\int_{\Omega_J} J(x-y)g(x,y)\,dy \quad \text{a.e. } x\in\Omega, \tag{3.1}$$

and

$$J(x-y)g(x,y)\in J(x-y)\,\mathrm{sign}(u(y)-u(x)) \quad \text{a.e. } (x,y)\in\Omega\times\Omega, \tag{3.2}$$

$$J(x-y)g(x,y)\in J(x-y)\,\mathrm{sign}(\psi(y)-u(x)) \quad \text{a.e. } (x,y)\in\Omega\times(\Omega_J\setminus\overline{\Omega}). \tag{3.3}$$

Remark 3.2. Observe that

(i) we can rewrite (3.2) + (3.3) as

$$J(x-y)g(x,y)\in J(x-y)\,\mathrm{sign}(u_\psi(y)-u(x)) \quad \text{a.e. } (x,y)\in\Omega\times\Omega_J, \tag{3.4}$$

where we set as above, and overall the section,

$$u_\psi(x) := \begin{cases} u(x) & \text{if } x\in\Omega, \\ \psi(x) & \text{if } x\in\Omega_J\setminus\overline{\Omega}, \\ 0 & \text{if } x\notin\Omega_J. \end{cases}$$

(ii) It holds $L^1(\Omega) = \mathrm{Dom}(B_{1,\psi}^J)$ and $B_{1,\psi}^J$ is closed in $L^1(\Omega)\times L^1(\Omega)$.

(iii) It is not difficult to see that, if $g\in L^\infty(\Omega_J\times\Omega_J)$, $g(x,y) = -g(y,x)$ for almost all $(x,y)\in\Omega_J\times\Omega_J$, $\|g\|_\infty\le 1$,

$$J(x-y)g(x,y)\in J(x-y)\ \ \mathrm{sign}(z(y)-z(x)) \quad \text{a.e. } (x,y)\in\Omega_J\times\Omega_J$$

is equivalent to

$$-\int_{\Omega_J}\int_{\Omega_J} J(x-y)g(x,y)\,dy\,z(x)\,dx = \frac{1}{2}\int_{\Omega_J}\int_{\Omega_J} J(x-y)|z(y)-z(x)|\,dy\,dx.$$

THEOREM 3.3. *Let $\psi \in L^1(\Omega_J \setminus \overline{\Omega})$. The operator $B^J_{1,\psi}$ is completely accretive and satisfies the range condition*

$$L^\infty(\Omega) \subset Ran\left(I + B^J_{1,\psi}\right).$$

Proof. Let $\hat{u}_i \in B^J_{1,\psi}u_i$, $i = 1, 2$, and set $u_i(y) = \psi(y)$ in $\Omega_J \setminus \overline{\Omega}$. Then, there exist $g_i \in L^\infty(\Omega_J \times \Omega_J)$, $\|g_i\|_\infty \le 1$, $g_i(x,y) = -g_i(y,x)$, $J(x-y)g_i(x,y) \in J(x-y)\text{sign}(u_i(y)-u_i(x))$ for almost all $(x,y) \in \Omega \times \Omega_J$, such that

$$\hat{u}_i(x) = -\int_{\Omega_J} J(x-y)g_i(x,y)\,dy \quad a.e.\ x \in \Omega$$

for $i = 1, 2$. Given $q \in C^\infty(\mathbb{R})$, $0 \le q' \le 1$, $\text{supp}(q')$ compact, $0 \notin \text{supp}(q)$, we have

$$\begin{aligned}
&\int_\Omega (\hat{u}_1(x) - \hat{u}_2(x))q(u_1(x)-u_2(x))\,dx \\
&= \frac{1}{2}\int_\Omega\int_\Omega J(x-y)(g_1(x,y)-g_2(x,y))\left(q(u_1(y)-u_2(y)) - q(u_1(x)-u_2(x))\right)\,dxdy \\
&\quad - \int_\Omega\int_{\Omega_J\setminus\overline{\Omega}} J(x-y)(g_1(x,y)-g_2(x,y)\left(q(u_1(x)-u_2(x))\right)\,dx\,dy \\
&\ge \frac{1}{2}\int_\Omega\int_\Omega J(x-y)(g_1(x,y)-g_2(x,y))\left(q(u_1(y)-u_2(y)) - q(u_1(x)-u_2(x))\right)\,dxdy.
\end{aligned}$$

Now, by the mean value theorem

$$\begin{aligned}
&J(x-y)(g_1(x,y)-g_2(x,y))\left[q(u_1(y)-u_2(y)) - q(u_1(x)-u_2(x))\right] \\
&= J(x-y)(g_1(x,y)-g_2(x,y))q'(\xi)\left[(u_1(y)-u_2(y)) - (u_1(x)-u_2(x))\right] \\
&= J(x-y)q'(\xi)\left[g_1(x,y)(u_1(y)-u_1(x)) - g_1(x,y)(u_2(y)-u_2(x))\right] \\
&\qquad -J(x-y)q'(\xi)\left[g_2(x,y)(u_1(y)-u_1(x)) - g_1(x,y)(u_2(y)-u_2(x))\right] \ge 0,
\end{aligned}$$

since

$$J(x-y)g_i(x,y)(u_i(y)-u_i(x)) = J(x-y)|u_i(y)-u_i(x)|, \quad i = 1, 2,$$

and

$$-J(x-y)g_i(x,y)(u_j(y)-u_j(x)) \ge -J(x-y)|u_j(y)-u_j(x)|, \quad i \ne j.$$

Hence

$$\int_\Omega (\hat{u}_1(x) - \hat{u}_2(x))q(u_1(x)-u_2(x))\,dx \ge 0,$$

from which it follows that $B^J_{1,\psi}$ is a completely accretive operator.

To show that $B^J_{1,\psi}$ satisfies the range condition, let us see that for any $\phi \in L^\infty(\Omega)$,

$$\lim_{p\to 1+}\left(I + B^J_{p,\psi}\right)^{-1}\phi = \left(I + B^J_{1,\psi}\right)^{-1}\phi \quad \text{weakly in } L^1(\Omega).$$

We prove this in several steps.

Step 1. Let us first suppose that $\psi \in L^\infty(\Omega_J \setminus \overline{\Omega})$. For $1 < p < +\infty$, by Theorem 2.7, there is u_p such that $u_p = (I + B^J_{p,\psi})^{-1}\phi$, that is,

$$u_p(x) - \int_{\Omega_J} J(x-y)\,|(u_p)_\psi(y) - u_p(x)|^{p-2}((u_p)_\psi(y) - u_p(x))\,dy = \phi(x), \tag{3.5}$$

a.e. $x \in \Omega$. It is easy to see that $\|u_p\|_\infty \le \sup\{\|\phi\|_\infty, \|\psi\|_\infty\}$. Therefore, there exists a sequence $p_n \to 1$ such that

$$u_{p_n} \rightharpoonup u \quad \text{weakly in } L^2(\Omega).$$

On the other hand, we also have

$$\frac{1}{2}\int_{\Omega_J}\int_{\Omega_J} J(x-y)\,|(u_{p_n})_\psi(y) - (u_{p_n})_\psi(x)|^{p_n}\,dy\,dx \le M_2, \qquad \forall\, n \in \mathbb{N}.$$

Consequently, for any measurable subset $E \subset \Omega_J \times \Omega_J$, we have

$$\begin{aligned}\left|\int\int_E J(x-y)|(u_{p_n})_\psi(y) - (u_{p_n})_\psi(x)|^{p_n-2}\,((u_{p_n}(y))_\psi - (u_{p_n})_\psi(x))\right| \\ \le \int\int_E J(x-y)|(u_{p_n})_\psi(y) - (u_{p_n})_\psi(x)|^{p_n-1} \le M_2|E|^{\frac{1}{p_n}}.\end{aligned}$$

Hence, by the Dunford–Pettis theorem we may assume that there exists $g(x,y)$ such that

$$J(x-y)|(u_{p_n})_\psi(y) - (u_{p_n})_\psi(x)|^{p_n-2}\,((u_{p_n})_\psi(y) - (u_{p_n})_\psi(x)) \rightharpoonup J(x-y)g(x,y),$$

weakly in $L^1(\Omega_J \times \Omega_J)$, $g(x,y) = -g(y,x)$ for almost all $(x,y) \in \Omega_J \times \Omega_J$, and $\|g\|_\infty \le 1$.

Therefore, by (3.5),

$$u(x) - \int_{\Omega_J} J(x-y)g(x,y)\,dy = \phi(x) \qquad \text{a.e. } x \in \Omega. \tag{3.6}$$

Then, to finish the proof it is enough to show that

$$\begin{aligned}-\int_{\Omega_J}\int_{\Omega_J} J(x-y)g(x,y)\,dy\,u_\psi(x)\,dx \\ = \frac{1}{2}\int_{\Omega_J}\int_{\Omega_J} J(x-y)|u_\psi(y) - u_\psi(x)|\,dy\,dx.\end{aligned} \tag{3.7}$$

In fact, by (3.5) and (3.6),

$$\begin{aligned}
&\frac{1}{2}\int_{\Omega_J}\int_{\Omega_J} J(x-y)\,|(u_{p_n})_\psi(y)-(u_{p_n})_\psi(x)|^{p_n}\,dy\,dx = \int_\Omega \phi u_{p_n} - \int_\Omega u_{p_n}u_{p_n}\\
&-\int_{\Omega_J\setminus\overline{\Omega}}\int_{\Omega_J} J(x-y)\,|\psi(y)-(u_{p_n})_\psi(x)|^{p_n-2}\,(\psi(y)-(u_{p_n})_\psi(x))\,dy\,\psi(x)\,dx\\
&=\int_\Omega \phi u - \int_\Omega uu - \int_\Omega \phi(u-u_{p_n}) + \int_\Omega 2u(u-u_{p_n}) - \int_\Omega (u-u_{p_n})(u-u_{p_n})\\
&\qquad-\int_{\Omega_J\setminus\overline{\Omega}}\int_{\Omega_J} J(x-y)\,|\psi(y)-(u_{p_n})_\psi(x)|^{p_n-2}\,(\psi(y)-(u_{p_n})_\psi(x))\,dy\,\psi(x)\,dx\\
&\le -\int_{\Omega_J}\int_{\Omega_J} J(x-y)g(x,y)\,dy\,u(x)\,dx - \int_\Omega \phi(u-u_{p_n}) + \int_\Omega 2u(u-u_{p_n})\\
&\qquad+\int_{\Omega_J\setminus\overline{\Omega}}\int_{\Omega_J} J(x-y)g(x,y)\,dy\,\psi(x)\,dx\\
&\qquad-\int_{\Omega_J\setminus\overline{\Omega}}\int_{\Omega_J} J(x-y)\,|\psi(y)-(u_{p_n})_\psi(x)|^{p_n-2}\,(\psi(y)-(u_{p_n})_\psi(x))\,dy\,\psi(x)\,dx,
\end{aligned}$$

and so,

$$\begin{aligned}
\limsup_{n\to+\infty}\frac{1}{2}\int_{\Omega_J}\int_{\Omega_J} J(x-y)\,|u_{p_n}(y)-u_{p_n}(x)|^{p_n}\,dy\,dx\\
\le -\int_\Omega\int_\Omega J(x-y)g(x,y)\,dy\,u(x)\,dx.
\end{aligned}$$

Now, by the monotonicity Lemma 2.4, for all $\rho\in L^\infty(\Omega)$,

$$\begin{aligned}
&-\int_{\Omega_J}\int_{\Omega_J} J(x-y)|\rho(y)-\rho(x)|^{p_n-2}(\rho(y)-\rho(x))\,dy\,(u_{p_n}(x)-\rho(x))\,dx\\
&\le -\int_{\Omega_J}\int_{\Omega_J} J(x-y)|u_{p_n}(y)-u_{p_n}(x)|^{p_n-2}(u_{p_n}(y)-u_{p_n}(x))dy(u_{p_n}(x)-\rho(x))dx.
\end{aligned}$$

Taking limits,

$$\begin{aligned}
-\int_{\Omega_J}\int_{\Omega_J} J(x-y)\,\mathrm{sign}_0(\rho(y)-\rho(x))\,dy\,(u(x)-\rho(x))\,dx\\
\le -\int_{\Omega_J}\int_{\Omega_J} J(x-y)g(x,y)\,dy\,(u(x)-\rho(x))\,dx.
\end{aligned}$$

Taking now, $\rho = u\pm\lambda u$, $\lambda>0$, and letting $\lambda\to 0$, we get (3.7), and the proof is finished for this class of data.

Step 2. Let us now suppose that ψ^- is bounded. Let $\psi_n = T_n(\psi)$, n large enough such that $\psi_n^- = \psi^-$. Then, $\{\psi_n\}$ is a nondecreasing sequence that converges in L^1 to ψ. By *Step* 1, there exists $u_n = (I+B^J_{1,\psi_n})^{-1}\phi$, that is, there exists $g_n\in L^\infty(\Omega_J\times\Omega_J)$, $g_n(x,y) = -g_n(y,x)$ for almost all $(x,y)\in\Omega_J\times\Omega_J$, $\|g_n\|_\infty\le 1$,

$$u_n(x) - \int_{\Omega_J} J(x-y)g_n(x,y)\,dy = \phi(x) \qquad \text{a.e. } x\in\Omega \tag{3.8}$$

and

$$\begin{aligned}
&-\int_{\Omega_J}\int_{\Omega_J} J(x-y)g_n(x,y)\,dy\,(u_n)_{\psi_n}(x)\,dx\\
&=\frac{1}{2}\int_{\Omega_J}\int_{\Omega_J} J(x-y)|(u_n)_{\psi_n}(y)-(u_n)_{\psi_n}(x)|\,dy\,dx.
\end{aligned} \tag{3.9}$$

Therefore, by monotonicity,

$$\int_{\Omega_J}\int_{\Omega_J}\left((u_n)_{\psi_n}-(u_{n+1})_{\psi_{n+1}}\right)\left((u_n)_{\psi_n}-(u_{n+1})_{\psi_{n+1}}\right)^+\le 0,$$

which implies $u_n\le u_{n+1}$. Since $\{u_n\}$ is bounded in L^∞ we have $\{u_n\}$ converges to a function u in L^2. On the other hand, we can suppose that $J(x-y)g_n(x,y)$ converges weakly in L^2 to $J(x-y)g(x,y)$, $g(x,y)=-g(y,x)$ for almost all $(x,y)\in\Omega_J\times\Omega_J$, and $\|g\|_\infty\le 1$. Hence, passing to the limit in (3.8) and (3.9) we obtain $u=(I+B^J_{1,\psi})^{-1}\phi$.

Step 3. For a general $\psi\in L^1(\Omega_J\setminus\overline{\Omega})$, apply *Step* 2 to $\psi_n=\sup\{\psi,-n\}$ and use monotonicity in a similar way to finish the proof. ∎

Proof of Theorem 1.5. As a consequence of the above results, we have that the abstract Cauchy problem

$$\begin{cases}u'(t)+B^J_{1,\psi}u(t)\ni 0, & t\in(0,T),\\ u(0)=u_0\end{cases}\tag{3.10}$$

has a unique mild solution u for every initial datum $u_0\in L^1(\Omega)$ and $T>0$ (see [15]). Moreover, due to the complete accretivity of the operator $B^J_{1,\psi}$, the mild solution of (3.10) is a strong solution ([14]). Consequently, the proof is concluded. ∎

3.2. Convergence to the total variation flow. Let us start recalling some results from [1] (see also [2]) about the Dirichlet problem for the total variational flow, that is,

$$D_1(u_0,\tilde\psi)\quad\begin{cases}u_t=\operatorname{div}\left(\frac{Du}{|Du|}\right) & \text{in }]0,T[\times\Omega,\\ u=\tilde\psi & \text{on }]0,T[\times\partial\Omega,\\ u(0,x)=u_0(x) & \text{in } \Omega,\end{cases}$$

with $\tilde\psi\in L^1(\partial\Omega)$.

THEOREM 3.4 ([1]). *Let $T>0$ and $\tilde\psi\in L^1(\partial\Omega)$. For any $u_0\in L^1(\Omega)$ $(L^2(\Omega))$ there exists a unique entropy (strong) solution $u(t)$ of $D_1(u_0,\tilde\psi)$.*

Associated to $-\operatorname{div}(\frac{Du}{|Du|})$ with Dirichlet boundary conditions, in [1] it is defined the operator $\mathcal{A}_{\tilde\psi}\subset L^1(\Omega)\times L^1(\Omega)$ as follows: $(u,v)\in\mathcal{A}_{\tilde\psi}$ if and only if $u,v\in L^1(\Omega)$, $q(u)\in BV(\Omega)$ for all $q\in\mathcal{P}:=\{q\in W^{1,\infty}(\mathbb{R}):q'\ge 0,\operatorname{supp}(q')\text{ is compact}\}$, and there exists $\zeta\in X(\Omega)$ (where $X(\Omega)$ is defined by (1.4)), with $\|\zeta\|_\infty\le 1$, $v=-\operatorname{div}(\zeta)$ in $\mathcal{D}'(\Omega)$ such that

$$\int_\Omega (w-q(u))v\le\int_\Omega(\zeta,Dw)-|Dq(u)|+\int_{\partial\Omega}|w-q\left(\tilde\psi\right)|-\int_{\partial\Omega}|q(u)-q\left(\tilde\psi\right)|\tag{3.11}$$

for every $w\in BV(\Omega)\cap L^\infty(\Omega)$ and every $q\in\mathcal{P}$. Also in [1] it is proved that the following assertions are equivalent:

(a) $(u,v)\in\mathcal{A}_{\tilde\psi}$,

(b) $u,v\in L^1(\Omega)$, $q(u)\in BV(\Omega)$ for all $q\in\mathcal{P}$, and there exists $\zeta\in X(\Omega)$, with $\|\zeta\|_\infty\le 1$, $v=-\operatorname{div}(\zeta)$ in $\mathcal{D}'(\Omega)$ such that

$$\int_\Omega(\zeta,Dq(u))=|Dq(u)|\qquad\forall\ q\in\mathcal{P},\tag{3.12}$$

$$[\zeta,\nu]\in\operatorname{sign}\left(q\left(\tilde\psi\right)-q(u)\right)\quad\mathcal{H}^{N-1}-a.e.\ \text{ on }\ \partial\Omega,\ \ \forall\ q\in\mathcal{P}.\tag{3.13}$$

Moreover, it is shown that $\mathcal{A}_{\tilde{\psi}}$ is an m-completely accretive operator in $L^1(\Omega)$ with dense domain and that for any $u_0 \in L^1(\Omega)$, the unique entropy solution $u(t)$ of problem $D_1(u_0, \tilde{\psi})$ coincides with the unique mild solution $e^{-t\mathcal{A}_{\tilde{\psi}}}u_0$ given by Crandall–Liggett's exponential formula.

Now, given J, we consider the rescaled kernels

$$J_{1,\varepsilon}(x) := \frac{C_{J,1}}{\varepsilon^{1+N}} J\left(\frac{x}{\varepsilon}\right), \qquad \text{with} \quad C_{J,1}^{-1} := \frac{1}{2}\int_{\mathbb{R}^N} J(z)|z_N|\,dz,$$

that is, a normalizing constant in order to obtain the 1-Laplacian in the limit instead of a multiple of it.

PROPOSITION 3.5. *Let Ω be a smooth bounded domain in $\mathbb{R}^N$ and $\tilde{\psi} \in L^\infty(\partial\Omega)$. Let $\psi \in W^{1,1}(\Omega_J \setminus \overline{\Omega}) \cap L^\infty(\Omega_J \setminus \overline{\Omega})$ such that $\psi|_{\partial\Omega} = \tilde{\psi}$. Assume $J(x) \geq J(y)$ if $|x| \leq |y|$. Then, for any $\phi \in L^\infty(\Omega)$,*

$$\left(I + B_{1,\psi}^{J_{1,\varepsilon}}\right)^{-1}\phi \to \left(I + \mathcal{A}_{\tilde{\psi}}\right)^{-1}\phi \quad \textit{strongly in } L^1(\Omega) \textit{ as } \varepsilon \to 0. \tag{3.14}$$

Proof. Given $\varepsilon > 0$ small, we set $u_\varepsilon = (I + B_{1,\psi}^{J_{1,\varepsilon}})^{-1}\phi$ and denote

$$\Omega_\varepsilon := \Omega_{J_{1,\varepsilon}} = \Omega + \operatorname{supp}(J_{1,\varepsilon}).$$

Then, there exists $g_\epsilon \in L^\infty(\Omega_\varepsilon \times \Omega_\varepsilon)$, $g_\varepsilon(x,y) = -g_\varepsilon(y,x)$ for almost all $(x,y) \in \Omega_\varepsilon \times \Omega_\varepsilon$, $\|g_\varepsilon\|_\infty \leq 1$, such that

$$J\left(\frac{x-y}{\varepsilon}\right) g_\varepsilon(x,y) \in J\left(\frac{x-y}{\varepsilon}\right) \operatorname{sign}(u_\varepsilon(y) - u_\varepsilon(x)) \qquad \text{a.e. } (x,y) \in \Omega \times \Omega,$$

$$J\left(\frac{x-y}{\varepsilon}\right) g_\varepsilon(x,y) \in J\left(\frac{x-y}{\varepsilon}\right) \operatorname{sign}(\tilde{\psi}(y) - u_\varepsilon(x)) \qquad \text{a.e. } (x,y) \in \Omega \times (\Omega_\epsilon \setminus \overline{\Omega})$$

and

$$u_\varepsilon(x) - \frac{C_{J,1}}{\varepsilon^{1+N}}\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right) g_\varepsilon(x,y)\,dy = \phi(x) \quad \textit{a.e. } x \in \Omega. \tag{3.15}$$

Therefore, for $v \in L^\infty(\Omega_J)$, we can write

$$\begin{aligned}\int_\Omega u_\varepsilon(x)v(x)\,dx - \frac{C_{J,1}}{\varepsilon^{1+N}}\int_\Omega\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right) g_\varepsilon(x,y)v(x)\,dy\,dx \\ = \int_\Omega \phi(x)v(x)\,dx.\end{aligned} \tag{3.16}$$

Observe that we can extend g_ε to a function in $L^\infty(\Omega_J \times \Omega_J)$, $g_\varepsilon(x,y) = -g_\varepsilon(y,x)$ for almost all $(x,y) \in \Omega_J \times \Omega_J$, $\|g_\varepsilon\|_{L^\infty(\Omega_J)} \leq 1$, such that

$$J\left(\frac{x-y}{\varepsilon}\right) g_\varepsilon(x,y) \in J\left(\frac{x-y}{\varepsilon}\right) \operatorname{sign}((u_\varepsilon)_\psi(y) - (u_\varepsilon)_\psi(x)) \quad \text{a.e. } (x,y) \in \Omega_J \times \Omega_J.$$

Let $M := \max\{\|\phi\|_{L^\infty(\Omega)}, \|\psi\|_{L^\infty(\Omega_J \setminus \overline{\Omega})}\}$. Taking $v = (u_\varepsilon - M)^+$ in (3.16), we get

$$\begin{aligned}&\int_\Omega u_\varepsilon(x)(u_\varepsilon(x) - M)^+ dx - \frac{C_{J,1}}{\varepsilon^{1+N}}\int_\Omega\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right) g_\varepsilon(x,y)(u_\varepsilon(x) - M)^+\,dy\,dx \\ &= \int_\Omega \phi(x)(u_\varepsilon(x) - M)^+\,dx.\end{aligned}$$

Now

$$
\begin{aligned}
&-\frac{C_{J,1}}{\varepsilon^{1+N}}\int_{\Omega_\varepsilon}\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right) g_\varepsilon(x,y)((u_\varepsilon)_\psi(x)-M)^+\,dy\,dx\\
&=\frac{C_{J,1}}{2\varepsilon^{1+N}}\int_{\Omega_\varepsilon}\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right) g_\varepsilon(x,y)(((u_\varepsilon)_\psi(y)-M)^+-\left((u_\varepsilon)_\psi(x)-M\right)^+)\,dy\,dx\\
&\geq 0.
\end{aligned}
$$

Hence, we get

$$\int_\Omega u_\varepsilon(x)(u_\varepsilon(x)-M)^+dx\leq\int_\Omega \phi(x)(u_\varepsilon(x)-M)^+dx.$$

Consequently,

$$0\leq\int_\Omega (u_\varepsilon(x)-M)(u_\varepsilon(x)-M)^+dx\leq\int_\Omega(\phi(x)-M)(u_\varepsilon(x)-M)^+dx\leq 0,$$

and we deduce $u_\varepsilon(x)\leq M$ for almost all $x\in\Omega$. Analogously, we can obtain $-M\leq u_\varepsilon(x)$ for almost all $x\in\Omega$. Thus

$$\|u_\varepsilon\|_{L^\infty(\Omega)}\leq M\qquad\text{for all }\epsilon>0; \tag{3.17}$$

from here, we can assume there exists a sequence $\varepsilon_n\to 0$ such that

$$u_{\varepsilon_n}\rightharpoonup u\quad\text{weakly in } L^1(\Omega).$$

Taking $v=u_\varepsilon$ in (3.16), we have

$$\int_\Omega u_\varepsilon(x)u_\varepsilon(x)\,dx-\frac{C_{J,1}}{\varepsilon^{1+N}}\int_\Omega\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right) g_\varepsilon(x,y)\,dyu_\varepsilon(x)\,dx=\int_\Omega\phi(x)u_\varepsilon(x)dx. \tag{3.18}$$

Observe that

$$
\begin{aligned}
&-\frac{C_{J,1}}{\varepsilon^{1+N}}\int_\Omega\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right) g_\varepsilon(x,y)dyu_\varepsilon(x)\,dx\\
&=-\frac{C_{J,1}}{\varepsilon^{1+N}}\int_{\Omega_\varepsilon}\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right) g_\varepsilon(x,y)dy(u_\varepsilon)_\psi(x)\,dx\\
&\quad+\frac{C_{J,1}}{\varepsilon^{1+N}}\int_{\Omega_\varepsilon\setminus\overline{\Omega}}\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right) g_\varepsilon(x,y)dy\psi(x)\,dx.
\end{aligned}
$$

Then

$$
\begin{aligned}
&\left|\frac{C_{J,1}}{\varepsilon^{1+N}}\int_{\Omega_\varepsilon\setminus\overline{\Omega}}\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right) g_\varepsilon(x,y)dy\psi(x)dx\right|\\
&\leq\frac{C_{J,1}}{\varepsilon^{1+N}}\int_{\Omega_\varepsilon\setminus\overline{\Omega}}\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right) dy|\psi(x)|dx\\
&\leq\frac{C_{J,1}}{\varepsilon}M\int_{\Omega_\varepsilon\setminus\overline{\Omega}}\left(\frac{1}{\varepsilon^N}\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right) dy\right)dx\\
&\leq\frac{C_{J,1}}{\varepsilon}M|\Omega_\varepsilon\setminus\Omega|\leq M_1.
\end{aligned}
$$

On the other hand,

$$-\frac{C_{J,1}}{\varepsilon^{1+N}}\int_{\Omega_\varepsilon}\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right) g_\varepsilon(x,y)dy(u_\varepsilon)_\psi(x)dx$$
$$=\frac{C_{J,1}}{2\varepsilon^{1+N}}\int_{\Omega_\varepsilon}\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right)|(u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x)|\,dy\,dx.$$

Consequently, from (3.17) and (3.18), it follows that

$$\frac{C_{J,1}}{2\varepsilon^{1+N}}\int_{\Omega_\varepsilon}\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right)|(u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x)|\,dy\,dx\le M_2. \tag{3.19}$$

Let us compute,

$$\begin{aligned}
&\frac{C_{J,1}}{2\varepsilon^{1+N}}\int_{\Omega_J}\int_{\Omega_J} J\left(\frac{x-y}{\varepsilon}\right)|(u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x)|\,dy\,dx\\
&=\frac{C_{J,1}}{2\varepsilon^{1+N}}\int_{\Omega_\varepsilon}\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right)|(u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x)|\,dy\,dx\\
&\quad+\frac{C_{J,1}}{2\varepsilon^{1+N}}\int_{\Omega_\varepsilon}\int_{\Omega_J\setminus\overline{\Omega}_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right)|(u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x)|\,dy\,dx\\
&\quad+\frac{C_{J,1}}{2\varepsilon^{1+N}}\int_{\Omega_J\setminus\overline{\Omega}_\varepsilon}\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right)|(u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x)|\,dy\,dx\\
&\quad+\frac{C_{J,1}}{2\varepsilon^{1+N}}\int_{\Omega_J\setminus\overline{\Omega}_\varepsilon}\int_{\Omega_J\setminus\overline{\Omega}_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right)|(u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x)|\,dy\,dx.
\end{aligned}$$

Now, since $\psi\in W^{1,1}(\Omega_J\setminus\overline{\Omega})$, we get

$$\frac{C_{J,1}}{2\varepsilon^{1+N}}\int_{\Omega_J\setminus\overline{\Omega}_\varepsilon}\int_{\Omega_J\setminus\overline{\Omega}_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right)|(u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x)|\,dy\,dx$$
$$=\frac{C_{J,1}}{2\varepsilon^{N}}\int_{\Omega_J\setminus\overline{\Omega}_\varepsilon}\int_{\Omega_J\setminus\overline{\Omega}_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right)\frac{|\psi(y)-\psi(x)|}{\varepsilon}dydx\le M_3.$$

On the other hand, we have

$$\begin{aligned}
&\frac{C_{J,1}}{2\varepsilon^{1+N}}\int_{\Omega_J\setminus\overline{\Omega}_\varepsilon}\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right)|(u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x)|\,dy\,dx\\
&=\frac{C_{J,1}}{2\varepsilon^{N}}\int_{\Omega_J\setminus\overline{\Omega}_\varepsilon}\int_{\Omega_\varepsilon\setminus\overline{\Omega}} J\left(\frac{x-y}{\varepsilon}\right)\frac{|\psi(y)-\psi(x)|}{\varepsilon}\,dy\,dx\\
&\le M_4\frac{C_{J,1}}{2}\int_{\Omega_J\setminus\overline{\Omega}_\varepsilon}\left(\frac{1}{\varepsilon^N}\int_{\Omega_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right)dy\right)dx\le M_5.
\end{aligned}$$

With similar arguments we obtain

$$\frac{C_{J,1}}{2\varepsilon^{1+N}}\int_{\Omega_\varepsilon}\int_{\Omega_J\setminus\overline{\Omega}_\varepsilon} J\left(\frac{x-y}{\varepsilon}\right)|(u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x)|\,dy\,dx\le M_6.$$

Therefore,

$$\frac{C_{J,1}}{2\varepsilon^{1+N}}\int_{\Omega_J}\int_{\Omega_J} J\left(\frac{x-y}{\varepsilon}\right)|(u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x)|\,dy\,dx\le M_7. \tag{3.20}$$

In particular, we get

$$\int_{\Omega_J}\int_{\Omega_J}\frac{1}{2}\frac{C_{J,1}}{\varepsilon^N}J\left(\frac{x-y}{\varepsilon}\right)\left|\frac{(u_\varepsilon)_\psi(y)-(u_\varepsilon)_\psi(x)}{\varepsilon}\right|dx\,dy\le M_7\qquad\forall\, n\in\mathbb{N}.$$

By Proposition 2.10, there exists a subsequence, denote equal, and $w\in BV(\Omega_J)$ such that

$$(u_{\varepsilon_n})_\psi\to w\qquad\text{strongly in } L^1(\Omega_J)$$

and

$$\frac{C_{J,1}}{2}J(z)\chi_\Omega(\cdot+\varepsilon_n z)\frac{(u_{\varepsilon_n})_\psi(\cdot+\varepsilon_n z)-(u_{\varepsilon_n})_\psi(\cdot)}{\varepsilon_n}\rightharpoonup\frac{C_{J,1}}{2}J(z)z\cdot Dw \tag{3.21}$$

weakly as measures. Hence, it is easy to obtain that

$$w(x)=u_\psi(x)=\begin{cases}u(x) & \text{in } x\in\Omega,\\ \psi(x) & \text{in } x\in\Omega_J\setminus\overline{\Omega},\end{cases}$$

and $u\in BV(\Omega)$.

Moreover, we can also assume that

$$J(z)\chi_{\Omega_J}(x+\varepsilon_n z)\overline{g}_{\varepsilon_n}(x,x+\varepsilon_n z)\rightharpoonup\Lambda(x,z) \tag{3.22}$$

weakly* in $L^\infty(\Omega_J)\times L^\infty(\mathbb{R}^N)$ for some function $\Lambda\in L^\infty(\Omega_J)\times L^\infty(\mathbb{R}^N)$, $\Lambda(x,z)\le J(z)$ almost everywhere in $\Omega_J\times\mathbb{R}^N$. Taking in (3.16) $v\in\mathcal{D}(\Omega)$, we get for $\varepsilon=\varepsilon_n$ small enough

$$\begin{aligned}\int_\Omega u_{\varepsilon_n}(x)v(x)dx-\frac{C_{J,1}}{{\varepsilon_n}^{1+N}}\int_\Omega\int_\Omega J\left(\frac{x-y}{\varepsilon_n}\right)g_{\varepsilon_n}(x,y)v(x)\,dy\,dx\\ =\int_\Omega\phi(x)v(x)\,dx.\end{aligned} \tag{3.23}$$

Changing variables and taking into account (3.23), we can write

$$\begin{aligned}&\frac{C_{J,1}}{2}\int_{\mathbb{R}^N}\int_\Omega J(z)\chi_\Omega(x+\varepsilon_n z)\overline{g}_{\varepsilon_n}(x,x+\varepsilon_n z)\,dz\frac{\overline{v}(x+\varepsilon_n z)-v(x)}{\varepsilon_n}\,dx\\ &=-\frac{C_{J,1}}{\varepsilon_n}\int_{\mathbb{R}^N}\int_\Omega J(z)\chi_\Omega(x+\varepsilon_n z)\overline{g}_{\varepsilon_n}(x,x+\varepsilon_n z)\,dz\,v(x)\,dx\\ &=\int_\Omega(\phi(x)-u_{\varepsilon_n}(x))v(x)\,dx.\end{aligned} \tag{3.24}$$

By (3.22), passing to the limit in (3.24), we get

$$\frac{C_{J,1}}{2}\int_{\mathbb{R}^N}\int_\Omega\Lambda(x,z)z\cdot\nabla v(x)\,dx\,dz=\int_\Omega(\phi(x)-u(x))v(x)\,dx \tag{3.25}$$

for all $v\in\mathcal{D}(\Omega)$. We set $\zeta=(\zeta_1,\dots,\zeta_N)$, the vector field defined by

$$\zeta_i(x):=\frac{C_{J,1}}{2}\int_{\mathbb{R}^N}\Lambda(x,z)z_i\,dz,\qquad i=1,\dots,N.$$

Then, $\zeta \in L^\infty(\Omega_J, \mathbb{R}^N)$, and from (3.25),

$$-\mathrm{div}(\zeta) = \phi - u \quad \text{in } \mathcal{D}'(\Omega).$$

Let us see that

$$\|\zeta\|_{L^\infty(\Omega_J)} \le 1.$$

Given $\xi \in \mathbb{R}^N \setminus \{0\}$, let R_ξ be the rotation such that $R_\xi^t(\xi) = \mathbf{e}_1|\xi|$. If we make the change of variables $z = R_\xi(y)$, we obtain

$$\begin{aligned}\zeta(x)\cdot\xi &= \frac{C_{J,1}}{2}\int_{\mathbb{R}^N}\Lambda(x,z)z\cdot\xi\,dz = \frac{C_{J,1}}{2}\int_{\mathbb{R}^N}\Lambda(x,R_\xi(y))R_\xi(y)\cdot\xi\,dy\\ &= \frac{C_{J,1}}{2}\int_{\mathbb{R}^N}\Lambda(x,R_\xi(y))y_1|\xi|\,dy.\end{aligned}$$

On the other hand, since J is a radial function and $\Lambda(x,z) \le J(z)$ almost everywhere,

$$C_{J,1}{}^{-1} = \frac{1}{2}\int_{\mathbb{R}^N} J(z)|z_1|\,dz$$

and

$$|\zeta(x)\cdot\xi| \le \frac{C_{J,1}}{2}\int_{\mathbb{R}^N} J(y)|y_1|\,dy|\xi| = |\xi| \quad \text{a.e. } x \in \Omega_J.$$

Therefore, $\|\zeta\|_{L^\infty(\Omega_J)} \le 1$.

To finish the proof, that is, to show that $u = (I + \mathcal{A}_{\tilde\psi})^{-1}\phi$, since $u \in L^\infty(\Omega)$ and $\tilde\psi \in L^\infty(\partial\Omega)$, we need only to prove that

$$(\zeta, Du) = |Du| \quad \text{as measures in } \Omega \tag{3.26}$$

and

$$[\zeta,\nu] \in \mathrm{sign}\left(\tilde\psi - u\right) \quad \mathcal{H}^{N-1}\text{-a.e. on } \partial\Omega. \tag{3.27}$$

Given $0 \le \varphi \in \mathcal{D}(\Omega)$, taking $\varepsilon = \varepsilon_n$ and $v = \varphi u_{\varepsilon_n}$ in (3.16), we get

$$\begin{aligned}&-\frac{C_{J,1}}{{\varepsilon_n}^{1+N}}\int_\Omega\int_\Omega J\left(\frac{x-y}{\varepsilon_n}\right)g_{\varepsilon_n}(x,y)u_{\varepsilon_n}(x)\varphi(x)\,dy\,dx\\ &= \frac{C_{J,1}}{2{\varepsilon_n}^{1+N}}\int_\Omega\int_\Omega J\left(\frac{x-y}{\varepsilon_n}\right)g_{\varepsilon_n}(x,y)(u_{\varepsilon_n}(y)\varphi(y) - u_{\varepsilon_n}(x)\varphi(x))\,dy\,dx\\ &= \int_\Omega(\phi(x) - u_{\varepsilon_n}(x))u_{\varepsilon_n}(x)\varphi(x)\,dx.\end{aligned} \tag{3.28}$$

Now, we decompose the double integral as follows,

$$I_n := \frac{C_{J,1}}{2{\varepsilon_n}^{1+N}}\int_\Omega\int_\Omega J\left(\frac{x-y}{\varepsilon_n}\right)g_{\varepsilon_n}(x,y)(u_{\varepsilon_n}(y)\varphi(y) - u_{\varepsilon_n}(x)\varphi(x))\,dy\,dx = I_n^1 + I_n^2,$$

where

$$\begin{aligned}I_n^1 &:= \frac{C_{J,1}}{2{\varepsilon_n}^{1+N}}\int_\Omega\int_\Omega J\left(\frac{x-y}{\varepsilon_n}\right)|u_{\varepsilon_n}(y) - u_{\varepsilon_n}(x)|\varphi(y)\,dy\,dx\\ &= \frac{C_{J,1}}{2}\int_\Omega\int_\Omega J(z)\,\chi_\Omega(x+\varepsilon_n z)\frac{|\overline{u}_{\varepsilon_n}(x+\varepsilon_n z) - u_{\varepsilon_n}(x)|}{\varepsilon_n}\overline{\varphi}(x+\varepsilon_n z)\,dz\,dx\end{aligned}$$

and

$$I_n^2 := \frac{C_{J,1}}{2{\varepsilon_n}^{1+N}} \int_\Omega \int_\Omega J\left(\frac{x-y}{\varepsilon_n}\right) g_{\varepsilon_n}(x,y) u_{\varepsilon_n}(x)(\varphi(y)-\varphi(x))\,dy\,dx$$
$$= \frac{C_{J,1}}{2} \int_\Omega \int_\Omega J(z)\chi_\Omega(x+\varepsilon_n z)\overline{g}_{\varepsilon_n}(x,x+\varepsilon_n z) u_{\varepsilon_n}(x) \frac{\overline{\varphi}(x+\varepsilon_n z)-\varphi(x)}{\varepsilon_n}\,dz\,dx.$$

Having in mind (3.21), it follows that

$$\lim_{n\to\infty} I_n^1 \geq \frac{C_{J,1}}{2} \int_\Omega \int_\Omega J\left(z\right)\varphi(x)|z\cdot Du| = \int_\Omega \varphi\,|Du|.$$

On the other hand, since

$$u_{\varepsilon_n} \to u \qquad \text{strongly in } L^1(\Omega),$$

by (3.22), we get

$$\lim_{n\to\infty} I_n^2 = \frac{C_{J,1}}{2} \int_\Omega \int_{\mathbb{R}^N} u(x)\Lambda(x,z) z\cdot\nabla\varphi(x)\,dz\,dx = \int_\Omega u(x)\zeta(x)\cdot\nabla\varphi(x)\,dx.$$

Therefore, taking $n\to+\infty$ in (3.28), we obtain

$$\int_\Omega \varphi\,|Du| + \int_\Omega u(x)\zeta(x)\cdot\nabla\varphi(x)\,dx \leq \int_\Omega (\phi(x)-u(x))u(x)\varphi(x)\,dx. \tag{3.29}$$

By Green's formula,

$$\int_\Omega (\phi(x)-u(x))u(x)\varphi(x)dx = -\int_\Omega \operatorname{div}(\zeta)u\varphi\,dx = \int_\Omega (\zeta, D(\varphi u))$$
$$= \int_\Omega \varphi(\zeta,Du) + \int_\Omega u(x)\zeta(x)\cdot\nabla\varphi(x)\,dx.$$

Since $|(\zeta,Du)|\leq|Du|$, the last identity and (3.29) give (3.26).

Finally, we show that (3.27) holds. We take $w_m\in W^{1,1}(\Omega)\cap C(\Omega)$ such that $w_m=\tilde{\psi}\,\mathcal{H}^{N-1}$-a.e. on $\partial\Omega$, and $w_m\to u$ in $L^1(\Omega)$. Taking $v=v_{m,n}:=(u_{\varepsilon_n})_\psi-(w_m)_\psi$ in (3.16), we get

$$\begin{aligned}
&\int_\Omega (\phi(x)-u_{\varepsilon_n}(x))(u_{\varepsilon_n}(x)-w_m(x))\,dx\\
&= -\frac{C_{J,1}}{{\varepsilon_n}^{1+N}} \int_{\Omega_J}\int_{\Omega_J} J\left(\frac{x-y}{\varepsilon_n}\right) g_{\varepsilon_n}(x,y) v_{m,n}(x)\,dy\,dx\\
&= \frac{C_{J,1}}{2{\varepsilon_n}^{1+N}} \int_{\Omega_J}\int_{\Omega_J} J\left(\frac{x-y}{\varepsilon_n}\right) g_{\varepsilon_n}(x,y)(v_{m,n}(x)-v_{m,n}(x))\,dy\,dx\\
&= H_n^1 + H_{m,n}^1,
\end{aligned} \tag{3.30}$$

where

$$H_n^1 = \frac{C_{J,1}}{2} \int_{\Omega_J}\int_{\mathbb{R}^N} J(z)\chi_{\Omega_J}(x+\varepsilon_n z)\left|\frac{(u_{\varepsilon_n})_\psi(x+\varepsilon_n z)-(u_{\varepsilon_n})_\psi(x)}{\varepsilon_n}\right| dz\,dx$$

and

$$H_{m,n}^2 = -\frac{C_{J,1}}{2} \int_{\Omega_J}\int_{\mathbb{R}^N} J(z)\chi_{\Omega_J}(x+\varepsilon_n z)\overline{g}_{\varepsilon_n}(x,x+\varepsilon_n z)$$
$$\times \frac{(w_m)_\psi(x+\varepsilon_n z)-(w_m)_\psi(x)}{\varepsilon_n}\,dz\,dx.$$

Arguing as before,

$$\lim_{n\to\infty} H_n^1 \geq \int_{\Omega_J} |Du_\psi| = \int_\Omega |Du| + \int_{\partial\Omega} \left|u-\tilde{\psi}\right| d\mathcal{H}^{N-1} + \int_{\Omega_J\setminus\overline{\Omega}} |\nabla\psi|.$$

On the other hand, since $(w_m)_\psi \in W^{1,1}(\Omega_J)$, by (3.22),

$$\lim_{n\to\infty} H_{m,n}^2 = -\frac{C_{J,1}}{2}\int_{\Omega_J}\int_{\mathbb{R}^N} \Lambda(x,z) z\cdot\nabla(w_m)_\psi(x)\,dz\,dx = -\int_{\Omega_J} \zeta(x)\cdot\nabla(w_m)_\psi(x)\,dx.$$

Consequently, taking $n\to\infty$ in (3.30), we get

$$\begin{aligned}(3.31)\qquad &\int_\Omega (\phi(x)-u(x))(u(x)-w_m(x))\,dx\\ &\geq \int_\Omega |Du| + \int_{\partial\Omega}\left|u-\tilde{\psi}\right| d\mathcal{H}^{N-1} + \int_{\Omega_J\setminus\overline{\Omega}} |\nabla\psi| - \int_{\Omega_J}\zeta(x)\cdot\nabla(w_m)_\psi(x)\,dx.\end{aligned}$$

Now,

$$\begin{aligned}-\int_{\Omega_J}\zeta(x)\cdot\nabla(w_m)_\psi(x)\,dx &= -\int_\Omega \zeta(x)\cdot\nabla w_m(x)\,dx - \int_{\Omega_J\setminus\overline{\Omega}}\zeta(x)\cdot\nabla\psi(x)\,dx\\ &= \int_\Omega \mathrm{div}\zeta(x) w_m(x)\,dx - \int_{\partial\Omega}[\zeta,\nu]\tilde{\psi}\,d\mathcal{H}^{N-1} - \int_{\Omega_J\setminus\overline{\Omega}}\zeta(x)\cdot\nabla\psi(x)\,dx.\end{aligned}$$

Since

$$\int_{\Omega_J\setminus\overline{\Omega}} |\nabla\psi| - \int_{\Omega_J\setminus\overline{\Omega}}\zeta(x)\cdot\nabla\psi(x)\,dx \geq 0,$$

from (3.31), we have

$$\begin{aligned}&\int_\Omega (\phi(x)-u(x))(u(x)-w_m(x))\,dx\\ &\quad\geq \int_\Omega |Du| + \int_{\partial\Omega}\left|u-\tilde{\psi}\right| d\mathcal{H}^{N-1} + \int_\Omega \mathrm{div}\zeta(x) w_m(x)\,dx - \int_{\partial\Omega}[\zeta,\nu]\tilde{\psi}\,d\mathcal{H}^{N-1}.\end{aligned}$$

Letting $m\to\infty$, and using Green's formula, we deduce

$$\begin{aligned}0 &\geq \int_\Omega |Du| + \int_{\partial\Omega}\left|u-\tilde{\psi}\right| d\mathcal{H}^{N-1} + \int_\Omega \mathrm{div}\zeta(x)u(x)\,dx - \int_{\partial\Omega}[\zeta,\nu]\tilde{\psi}\,d\mathcal{H}^{N-1}\\ &= \int_\Omega |Du| + \int_{\partial\Omega}\left|u-\tilde{\psi}\right| d\mathcal{H}^{N-1} - \int_\Omega (\zeta, Du) + \int_{\partial\Omega}[\zeta,\nu]u\,d\mathcal{H}^{N-1}\\ &\qquad - \int_{\partial\Omega}[\zeta,\nu]\tilde{\psi}\,d\mathcal{H}^{N-1}.\end{aligned}$$

By (3.26), we obtain

$$\int_{\partial\Omega}\left|u-\tilde{\psi}\right| d\mathcal{H}^{N-1} \leq \int_{\partial\Omega}[\zeta,\nu]\big(\tilde{\psi}-u\big)\,d\mathcal{H}^{N-1} \leq \int_{\partial\Omega}\left|u-\tilde{\psi}\right| d\mathcal{H}^{N-1}.$$

Therefore,

$$[\zeta,\nu]\in \mathrm{sign}\big(\tilde{\psi}-u\big) \quad \mathcal{H}^{N-1}-\text{a.e. on } \partial\Omega,$$

and the proof is finished. □

From the above Proposition, by standard results of the Nonlinear Semigroup Theory (see, [19] or [15]), we obtain Theorem 1.6.

4. Limit as $p \to +\infty$. A model for sandpiles.

4.1. A model for sandpiles. Let sand be poured out onto a rigid surface, $y = u_0(x)$, given in a bounded open subset Ω of $\mathbb{R}^2$ with Lipschitz boundary $\partial\Omega$. If the support boundary is open and we assume that the angle of stability is equal to $\frac{\pi}{4}$, a model for pile surface evolution was proposed by Prigozhin [35] as

$$\partial_t u + \operatorname{div} \mathbf{q} = f, \qquad u|_{t=0} = u_0, \qquad u|_{\partial\Omega} = u_0|_{\partial\Omega}, \tag{4.1}$$

where $u(t,x)$ is the unknown pile surface, $f(t,x) \geq 0$ is the given source density, and $\mathbf{q}(t,x)$ is the unknown horizontal projection of the flux of sand pouring down the pile surface. If the support has no slopes steeper than the sand angle of repose, $\|\nabla u_0\|_\infty \leq 1$, Prigozhin ([35], see also [10], [29], and the references therein) proposed to take $\mathbf{q} = -m\nabla u$, where $m \geq 0$ is the Lagrange multiplier related to the constraint $\|\nabla u\|_\infty \leq 1$ and satisfies $m(\|\nabla u\|^2 - 1) = 0$ and reformulated this model as the following variational inequality:

$$\begin{cases} f(t,\cdot) - u_t(t) \in \partial I_{K(u_0)}(u(t)), & \text{a.e. } t \in]0,T[, \\ u(0,x) = u_0(x), \end{cases} \tag{4.2}$$

where

$$K(u_0) := \left\{ v \in W^{1,\infty}(\Omega) \ : \ \|\nabla v\|_\infty \leq 1, \ v|_{\partial\Omega} = u_0|_{\partial\Omega} \right\}.$$

Our aim is to approximate the Prigozhin model for the sandpile by a nonlocal model (Theorem 1.8) obtained as the limit as $p \to +\infty$ of the nonlocal p-Laplacian problem with Dirichlet boundary condition (Theorem 1.7).

To identify the limit as $p \to +\infty$ of the solutions u_p of problem $P_p^J(u_0, \psi)$ we will use the methods of convex analysis, and so we first recall some terminology (see [30], [17], and [8]). If H is a real Hilbert space with inner product $(\ ,\ )$ and $\Psi : H \to (-\infty, +\infty]$ is convex, then the subdifferential of Ψ is defined as the multivalued operator $\partial\Psi$ given by

$$v \in \partial\Psi(u) \iff \Psi(w) - \Psi(u) \geq (v, w-u) \quad \forall\, w \in H.$$

The epigraph of Ψ is defined by $\operatorname{Epi}(\Psi) = \{(u,\lambda) \in H \times \mathbb{R} \ : \ \lambda \geq \Psi(u)\}$.

Given K a closed convex subset of H, the indicator function of K is defined by

$$I_K(u) = \begin{cases} 0 & \text{if } \ u \in K, \\ +\infty & \text{if } \ u \notin K. \end{cases}$$

Then it is easy to see that the subdifferential is characterized as follows:

$$v \in \partial I_K(u) \iff u \in K \ \text{ and } \ (v, w-u) \leq 0 \quad \forall\, w \in K.$$

In case the convex functional $\Psi : H \to (-\infty, +\infty]$ is proper, lower-semicontinuous, and $\min \Psi = 0$, it is well known (see [17]) that the abstract Cauchy problem

$$\begin{cases} u'(t) + \partial\Psi(u(t)) \ni f(t), & \text{a.e. } t \in]0,T[, \\ u(0) = u_0, \end{cases}$$

has a unique strong solution for any $f \in L^2(0,T;H)$ and $u_0 \in \overline{D(\partial\Psi)}$.

The following convergence was studied by Mosco in [34] (see [8]). Suppose X is a metric space and $A_n \subset X$. We define

$$\liminf_{n\to\infty} A_n = \{x \in X \ : \ \exists x_n \in A_n,\ x_n \to x\}$$

and

$$\limsup_{n\to\infty} A_n = \{x \in X \ : \ \exists x_{n_k} \in A_{n_k},\ x_{n_k} \to x\}.$$

In the case X is a normed space, we note by $s-\lim$ and $w-\lim$ the above limits associated, respectively, to the strong and to the weak topology of X.

Given a sequence $\Psi_n, \Psi : H \to (-\infty, +\infty]$ of convex lower-semicontinuous functionals, we say that Ψ_n converges to Ψ in the sense of Mosco if

$$w - \limsup_{n\to\infty} \mathrm{Epi}(\Psi_n) \subset \mathrm{Epi}(\Psi) \subset s - \liminf_{n\to\infty} \mathrm{Epi}(\Psi_n). \tag{4.3}$$

It is easy to see that (4.3) is equivalent to the two following conditions:

$$\forall\, u \in D(\Psi)\ \ \exists u_n \in D(\Psi_n)\ :\ u_n \to u \ \text{ and } \ \Psi(u) \geq \limsup_{n\to\infty} \Psi_n(u_n); \tag{4.4}$$

$$\text{for every subsequence } n_k, \text{ when } u_k \rightharpoonup u, \text{ it holds } \Psi(u) \leq \liminf_{k} \Psi_{n_k}(u_k). \tag{4.5}$$

As a consequence of the results in [19] and [8] we can write the following result.

THEOREM 4.1. *Let $\Psi_n, \Psi : H \to (-\infty, +\infty]$ convex lower-semicontinuous functionals. Then the following statements are equivalent.*

(i) *Ψ_n converges to Ψ in the sense of Mosco.*

(ii) *$(I + \lambda\partial\Psi_n)^{-1}u \to (I + \lambda\partial\Psi)^{-1}u, \quad \forall\, \lambda > 0,\ u \in H.$*

Moreover, any of these two conditions (i) *or* (ii) *imply that*

(iii) *for every $u_0 \in \overline{D(\partial\Psi)}$ and $u_{0,n} \in \overline{D(\partial\Psi_n)}$ such that $u_{0,n} \to u_0$, and every $f_n, f \in L^2(0,T;H)$ with $f_n \to f$, if $u_n(t)$, $u(t)$ are the strong solutions of the abstract Cauchy problems*

$$\begin{cases} u_n'(t) + \partial\Psi_n(u_n(t)) \ni f_n, & \text{a.e. } t \in]0,T[, \\ u_n(0) = u_{0,n}, \end{cases}$$

and

$$\begin{cases} u'(t) + \partial\Psi(u(t)) \ni f, & \text{a.e. } t \in]0,T[, \\ u(0) = u_0, \end{cases}$$

respectively, then

$$u_n \to u \qquad \text{in } C([0,T] : H).$$

4.2. Limit as $p \to +\infty$. Let us consider the *nonlocal p-Laplacian evolution problem* with source

$$P_p^J(u_0,\psi,f) \begin{cases} u_t(t,x) = \displaystyle\int_\Omega J(x-y)|u(t,y)-u(t,x)|^{p-2}(u(t,y)-u(t,x))dy + f(t,x), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (t,x) \in]0,T[\times\Omega, \\ u(t,x) = \psi(x), \quad (t,x) \in]0,T[\times(\Omega_J \setminus \overline{\Omega}), \\ u(0,x) = u_0(x), \quad x \in \Omega. \end{cases}$$

This problem is associated to the energy functional

$$G^J_{p,\psi}(u) = \frac{1}{2p}\int_\Omega\int_\Omega J(x-y)|u(y)-u(x)|^p\,dy\,dx + \frac{1}{p}\int_\Omega\int_{\Omega_J\setminus\overline{\Omega}} J(x-y)|\psi(y)-u(x)|^p\,dy\,dx.$$

With a formal calculation, taking limit as $p\to+\infty$, we arrive to the functional

$$G^J_{\infty,\psi}(u) = \begin{cases} 0 & \text{if } |u(x)-u(y)|\le 1, \text{ for } x,y\in\overline{\Omega} \\ & \text{and } |\psi(y)-u(x)|\le 1, \quad \text{for } x\in\Omega, y\in\Omega_J\setminus\overline{\Omega}, \\ & \text{with } x-y\in\operatorname{supp}(J) \\ +\infty & \text{in the other case.} \end{cases}$$

Hence, if we define

$$K^J_{\infty,\psi} := \left\{ u\in L^2(\Omega) \;:\; \begin{array}{l} |u(x)-u(y)|\le 1, x,y\in\Omega \\ \text{and } |\psi(y)-u(x)|\le 1, \quad \text{for } x\in\Omega, y\in\Omega_J\setminus\overline{\Omega}, \\ \text{with } x-y\in\operatorname{supp}(J) \end{array} \right\},$$

we have that the functional $G^J_{\infty,\psi}$ is given by the indicator function of $K^J_{\infty,\psi}$; that is, $G^J_{\infty,\psi} = I_{K^J_{\infty,\psi}}$. Then, the *nonlocal limit problem* can be written as

$$P^J_\infty(u_0,\psi,f) \quad \begin{cases} f(t,\cdot) - u_t(t) \in \partial I_{K^J_{\infty,\psi}}(u(t)), & \text{a.e. } t\in]0,T[, \\ u(0,x) = u_0(x). \end{cases}$$

Proof of Theorem 1.7. Let $T>0$. By Theorem 4.1, to prove the result it is enough to show that the functionals

$$G^J_{p,\psi}(u) = \frac{1}{2p}\int_\Omega\int_\Omega J(x-y)|u(y)-u(x)|^p\,dy\,dx + \frac{1}{p}\int_\Omega\int_{\Omega_J\setminus\overline{\Omega}} J(x-y)|\psi(y)-u(x)|^p\,dy\,dx$$

converge to

$$G^J_{\infty,\psi}(u) = \begin{cases} 0 & \text{if } |u(x)-u(y)|\le 1, \text{ for } x,y\in\Omega \\ & \text{and } |\psi(y)-u(x)|\le 1, \quad \text{for } x\in\Omega, y\in\Omega_J\setminus\overline{\Omega}, \\ & \text{with } x-y\in\operatorname{supp}(J) \\ +\infty & \text{in the other case} \end{cases}$$

as $p\to+\infty$, in the sense of Mosco. First, let us check that

$$\operatorname{Epi}\left(G^J_{\infty,\psi}\right) \subset s-\liminf_{p\to+\infty}\operatorname{Epi}\left(G^J_{p,\psi}\right). \tag{4.6}$$

To this end let $(u,\lambda)\in\operatorname{Epi}(G^J_{\infty,\psi})$. We can assume that $u\in K^J_{\infty,\psi}$ and $\lambda\ge 0$ (as $G^J_{\infty,\psi}(u)=0$). Now take

$$v_p = u \qquad \text{and} \qquad \lambda_p = G^J_{p,\psi}(u)+\lambda. \tag{4.7}$$

Then, as $\lambda \geq 0$ we have $(v_p, \lambda_p) \in \mathrm{Epi}(G^J_{p,\psi})$. Obviously, $v_p = u \to u$ in $L^2(\Omega)$, and as $u \in K^J_{\infty,\psi}$,

$$\begin{aligned} G^J_{p,\psi}(u) &= \frac{1}{2p}\int_\Omega\int_\Omega J(x-y)|u(y)-u(x)|^p\,dydx \\ &\quad + \frac{1}{p}\int_\Omega\int_{\Omega_J\setminus\overline{\Omega}} J(x-y)|\psi(y)-u(x)|^p\,dy\,dx \\ &\leq \frac{1}{2p}\int_\Omega\int_\Omega J(x-y)\,dydx + \frac{1}{p}\int_\Omega\int_{\Omega_J\setminus\overline{\Omega}} J(x-y)\,dy\,dx \to 0 \end{aligned}$$

as $p \to +\infty$. Therefore, we get (4.6). Finally, let us prove that

$$w - \limsup_{p\to+\infty} \mathrm{Epi}\left(G^J_{p,\psi}\right) \subset \mathrm{Epi}\left(G^J_{\infty,\psi}\right). \tag{4.8}$$

To this end, let us consider a sequence $(u_{p_j}, \lambda_{p_j}) \in \mathrm{Epi}(G^J_{p_j,\psi})$; that is, $G^J_{p_j,\psi}(u_{p_j}) \leq \lambda_{p_j}$, with

$$u_{p_j} \rightharpoonup u, \qquad \text{and} \qquad \lambda_{p_j} \to \lambda.$$

Since, $0 \leq G^J_{p_j,\psi}(u_{p_j}) \leq \lambda_{p_j} \to \lambda$, $0 \leq \lambda$. On the other hand, we have that there exists a constant $C > 0$ such that

$$\begin{aligned} (p_jC)^{1/p_j} \geq \left(p_jG^J_{p,\psi}(u_{p_j})\right)^{1/p_j} &= \left(\frac{1}{2}\int_\Omega\int_\Omega J(x-y)|u_{p_j}(y)-u_{p_j}(x)|^{p_j}\,dy\,dx \right. \\ &\quad \left. + \int_\Omega\int_{\Omega_J\setminus\overline{\Omega}} J(x-y)|\psi(y)-u_{p_j}(x)|^{p_j}\,dy\,dx\right)^{1/p_j}. \end{aligned}$$

Then, by the above inequality,

$$\begin{aligned} &\left(\int_\Omega\int_\Omega J\,(x-y)\left|u_{p_j}(y)-u_{p_j}(x)\right|^q\,dy\,dx\right)^{1/q} \\ &\quad\leq \left(\int_\Omega\int_\Omega J\,(x-y)\,dy\,dx\right)^{(p_j-q)/p_jq} \\ &\qquad\times\left(\int_\Omega\int_\Omega J\,(x-y)\left|u_{p_j}(y)-u_{p_j}(x)\right|^{p_j}\,dy\,dx\right)^{1/p_j} \\ &\quad\leq \left(\int_\Omega\int_\Omega J\,(x-y)\,dy\,dx\right)^{(p_j-q)/p_jq}(Cp_j)^{1/p_j}. \end{aligned}$$

Hence, we can extract a subsequence (if necessary) and let $p_j \to +\infty$ to obtain

$$\left(\int_\Omega\int_\Omega J\,(x-y)\,|u(y)-u(x)|^q\,dy\,dx\right)^{1/q} \leq \left(\int_\Omega\int_\Omega J\,(x-y)\,dy\,dx\right)^{1/q}.$$

Now, just taking $q \to +\infty$, we get

$$|u(x)-u(y)| \leq 1 \qquad \text{a.e. } (x,y) \in \Omega\times\Omega,\ x-y \in \mathrm{supp}(J).$$

With a similar argument we obtain

$$|u(x)-\psi(y)| \leq 1 \qquad \text{a.e. } x \in \Omega,\ y \in \Omega_J\setminus\overline{\Omega},\ \text{with}\ \ x-y \in \mathrm{supp}(J).$$

Hence, we conclude that $u \in K^J_{\infty,\psi}$. This ends the proof. □

4.3. Rescaling. We will assume now that Ω is convex and ψ verifies $\|\nabla\psi\|_\infty \le 1$. For $\varepsilon > 0$, we rescale the functional $G^J_{\infty,\psi}$ as follows:

$$G^\varepsilon_{\infty,\psi}(u) = \begin{cases} 0 & \text{if } |u(x)-u(y)| \le \varepsilon, \text{ for } x,y \in \Omega \\ & \text{and } |\psi(y)-u(x)| \le \varepsilon, \quad \text{for } x \in \Omega, y \in \Omega_J \setminus \overline{\Omega}, \\ & \text{with } |x-y| \le \varepsilon \\ +\infty & \text{in the other case.} \end{cases}$$

In other words, $G^\varepsilon_{\infty,\psi} = I_{K^\varepsilon_{\infty,\psi}}$, where

$$K^\varepsilon_{\infty,\psi} := \left\{ u \in L^2(\Omega) \ : \ \begin{array}{l} |u(x)-u(y)| \le \varepsilon, x, y \in \Omega \\ \text{and } |\psi(y)-u(x)| \le \varepsilon, \quad \text{for } x \in \Omega, y \in \Omega_J \setminus \overline{\Omega}, \\ \text{with } |x-y| \le \varepsilon \end{array} \right\}.$$

Consider the gradient flow associated to the functional $G^\varepsilon_{\infty,\psi}$

$$P^\varepsilon_\infty(u_0,\psi,f) \quad \begin{cases} f(t,\cdot) - u_t(t,\cdot) \in \partial I_{K^\varepsilon_{\infty,\psi}}(u(t)), & \text{a.e. } t \in]0,T[, \\ u(0,x) = u_0(x), & \text{in } \Omega, \end{cases}$$

and the problem

$$P_\infty(u_0,\psi,f) \quad \begin{cases} f(t,\cdot) - u_{\infty,t} \in \partial I_{K_\psi}(u_\infty), & \text{a.e. } t \in]0,T[, \\ u_\infty(0,x) = u_0(x), & \text{in } \Omega, \end{cases}$$

where

$$K_\psi := \left\{ u \in W^{1,\infty}(\Omega) \ : \ \|\nabla u\|_\infty \le 1, \ u|_{\partial\Omega} = \psi|_{\partial\Omega} \right\}.$$

Observe that if $u \in K_\psi$, $\|\nabla u\|_\infty \le 1$. Then, since $\|\nabla\psi\|_\infty \le 1$ and Ω is convex, we have $|u(x)-u(y)| \le |x-y|$ and $|u(x)-\psi(y)| \le |x-y|$, from where it follows that $u \in K^\varepsilon_{\infty,\psi}$, that is, $K_\psi \subset K^\varepsilon_{\infty,\psi}$.

With all these definitions and notations, we can proceed with the limit as $\varepsilon \to 0$ for the sandpile model ($p = +\infty$).

Proof of Theorem 1.8. Since $u_0 \in K_\psi$, $u_0 \in K^\varepsilon_{\infty,\psi}$ for all $\varepsilon > 0$. Again we are using that $\|\nabla\psi\|_\infty \le 1$. Consequently, there exists $u_{\infty,\varepsilon}$ the unique solution of $P^\varepsilon_\infty(u_0,\psi,f)$.

By Theorem 4.1, to prove the result it is enough to show that $I_{K^\varepsilon_{\infty,\psi}}$ converges to I_{K_ψ} in the sense of Mosco. Using that $\|\nabla\psi\|_\infty \le 1$ it is easy to obtain that

$$K^{\varepsilon_1}_{\infty,\psi} \subset K^{\varepsilon_2}_{\infty,\psi}, \quad \text{if } \ \varepsilon_1 \le \varepsilon_2. \tag{4.9}$$

Since $K_\psi \subset K^\varepsilon_{\infty,\psi}$ for all $\varepsilon > 0$, we have

$$K_\psi \subset \bigcap_{\varepsilon>0} K^\varepsilon_{\infty,\psi}.$$

On the other hand, if

$$u \in \bigcap_{\varepsilon>0} K^\varepsilon_{\infty,\psi},$$

we have

$$|u(y)-u(x)| \le |y-x|, \quad \text{a.e. } x,y \in \Omega,$$

and moreover

$$|u(y)-\psi(x)| \le |y-x|, \quad \text{a.e. } x \in \Omega_J \setminus \overline{\Omega},\ y \in \Omega,$$

from where it follows that $u \in K_\psi$. Therefore, we have

$$K_\psi = \bigcap_{\varepsilon>0} K^\varepsilon_{\infty,\psi}. \tag{4.10}$$

Note that

$$\mathrm{Epi}(I_{K_\psi}) = K_\psi \times [0,\infty[, \quad \mathrm{Epi}\left(I_{K^\varepsilon_{\infty,\psi}}\right) = K^\varepsilon_{\infty,\psi} \times [0,\infty[\ \ \forall\, \varepsilon>0. \tag{4.11}$$

By (4.10) and (4.11),

$$\mathrm{Epi}\left(I_{K_\psi}\right) \subset s-\liminf_{\varepsilon\to 0} \mathrm{Epi}\left(I_{K^\varepsilon_{\infty,\psi}}\right). \tag{4.12}$$

On the other hand, given $(u,\lambda) \in w-\limsup_{\varepsilon\to 0} \mathrm{Epi}(I_{K^\varepsilon_{\infty,\psi}})$ there exists $(u_{\varepsilon_k},\lambda_k) \in K_{\varepsilon_k,\psi} \times [0,\infty[$, such that $\varepsilon_k \to 0$ and

$$u_{\varepsilon_k} \rightharpoonup u \quad \text{in } L^2(\Omega), \quad \lambda_k \to \lambda \ \text{ in } \mathbb{R}.$$

By (4.9), given $\varepsilon>0$, there exists k_0, such that $u_{\varepsilon_k} \in K^\varepsilon_{\infty,\psi}$ for all $k \ge k_0$. Then, since $K^\varepsilon_{\infty,\psi}$ is a closed convex set, we get $u \in K^\varepsilon_{\infty,\psi}$, and, by (4.10), we obtain that $u \in K_0$. Consequently,

$$w-\limsup_{n\to\infty} \mathrm{Epi}\left(I_{K^\varepsilon_{\infty,\psi}}\right) \subset \mathrm{Epi}\left(I_{K_\psi}\right). \tag{4.13}$$

Finally, by (4.12), (4.13), and having in mind (4.3), we obtain that $I_{K^\varepsilon_{\infty,\psi}}$ converges to I_{K_ψ} in the sense of Mosco. □

4.4. Explicit solutions. Our goal now is to show some explicit examples that illustrate the behavior of the solutions when $p=+\infty$.

Remark 4.2. There is a natural upper bound (and of course also a natural lower bound) for the solutions with boundary datum ψ outside Ω (regardless the source term f). Indeed, given a bounded domain $\Omega \subset \mathbb{R}^N$ let us define inductively

$$\Omega_1 = \left\{x \in \Omega \ :\ |x-y|<1 \text{ for some } y \in \Omega_J \setminus \overline{\Omega}\right\}$$

and, for $j \ge 2$,

$$\Omega_j = \left\{x \in \Omega \setminus \cup_{i=1}^{j-1}\Omega_i \ :\ |x-y|<1 \text{ for some } y \in \Omega_{j-1}\right\}.$$

Then, since $u(t) \in K^J_{\infty,\psi}$ we must have

$$u(t,x) \le \psi(y)+1 \qquad \text{if } \ |x-y| \le 1,\ x \in \Omega_1,\ y \in \Omega_J \setminus \overline{\Omega},$$

and for any $j \ge 2$

$$u(t,x) \le u(t,y)+1 \qquad \text{if } \ |x-y| \le 1,\ x \in \Omega_j,\ y \in \Omega_{j-1} \setminus \Omega_j.$$

Therefore we have an upper bound for $u(t,x)$ in the whole Ω,

$$u(t,x) \le \Psi_1(x),$$

where Ψ_1 is defined by the inductive formula,

$$\Psi_1(x) = \max\left\{\psi(y)+1 \ : \ y \in \Omega_J \setminus \overline{\Omega},\ |x-y| \le 1\right\},\ \text{for } x \in \Omega_1,$$

and

$$\Psi_1(x) = \max\left\{\Psi_1(y)+1 \ : \ y \in \Omega_{j-1},\ |x-y| \le 1\right\},\ \text{for } x \in \Omega_j,\ \text{if } j \ge 2.$$

Analogously, we can obtain a lower bound for $u(t,x)$,

$$u(t,x) \ge \Phi_1(x),$$

where Φ_1 is defined by the inductive formula,

$$\Phi_1(x) = \min\left\{\psi(y)-1 \ : \ y \in \Omega_J \setminus \overline{\Omega},\ |x-y| \le 1\right\},\ \text{for } x \in \Omega_1,$$

and

$$\Phi_1(x) = \min\left\{\Phi_1(y)-1 \ : \ y \in \Omega_{j-1},\ |x-y| \le 1\right\},\ \text{for } x \in \Omega_j,\ \text{ if } j \ge 2.$$

With this remark in mind we show some explicit examples of solutions to

$$P^J_\infty(u_0,\psi,f) \quad \begin{cases} f(t,x) - u_t(t,x) \in \partial G^J_{\infty,\psi}(u(t)), & \text{a.e. } t \in]0,T[, \\ u(0,x) = u_0(x), & \text{in } \Omega, \end{cases}$$

where

$$G^J_{\infty,\psi}(u) = \begin{cases} 0 & \text{if } u \in L^2(\Omega),\ |u(x)-u(y)| \le 1,\ \text{for } x,y \in \Omega,\ |x-y| \le 1, \\ & \text{and } |u(x)-\psi(y)| \le 1, \text{for } x \in \Omega,\ y \in \Omega_J \setminus \overline{\Omega},\ |x-y| \le 1, \\ +\infty & \text{in the other case.} \end{cases}$$

In order to verify that a function $u(t,x)$ is a solution to $P^J_\infty(u_0,\psi,f)$, we need to check that

$$G^J_{\infty,\psi}(v) \ge G^J_{\infty,\psi}(u) + \langle f - u_t,\ v-u\rangle, \qquad \text{for all } v \in L^2(\Omega). \tag{4.14}$$

To this end we can assume that $v \in K^J_{\infty,\psi}$ (otherwise $G^J_{\infty,\psi}(v) = +\infty$ and then (4.14) becomes trivial). Therefore, we need to check that

$$u(t,\cdot) \in K^J_{\infty,\psi} \tag{4.15}$$

and, by (4.14), that

$$\int_\Omega (f(t,x) - u_t(t,x))(v(x)-u(t,x))\,dx \le 0 \tag{4.16}$$

for every $v \in K^J_{\infty,\psi}$.

Example 1. Let us consider a nonnegative source f and as initial condition the upper bound defined in the previous remark, $u_0(x) = \Psi_1(x)$. Then the solution to $P^J_\infty(u_0,\psi,f)$ is given by

$$u(t,x) \equiv \Psi_1(x)$$

for every $t > 0$. Indeed, $\Psi_1(x) \in K^J_{\infty,\psi}$ and for every $v \in K^J_{\infty,\psi}$ we have that $v(x) \le \Psi_1(x)$, and therefore

$$\int_\Omega (f(t,x) - u_t(t,x))(v(x) - u(t,x))\,dx = \int_\Omega f(t,x)(v(x) - \Psi_1(x))\,dx \le 0,$$

as we have to show.

In general, given a nonnegative source f supported in $D \subset \Omega$, any initial condition $u_0 \in K^J_{\infty,\psi}$ that verifies $u_0(x) = \Psi_1(x)$ in D produces a stationary solution $u(t,x) \equiv u_0(x)$.

Analogously, it can be shown that $u(t,x) \equiv \Phi_1(x)$ when $u_0(x) = \Phi_1(x)$ and $f(t,x) \le 0$.

Example 2. Now, let us assume that we are in an interval $\Omega = (-L, L)$, $\psi = 0$, $\varepsilon = L/n$, $n \in \mathbf{N}$, $u_0 = 0$ which belongs to $K_{\varepsilon,0}$, and the source f is an approximation of a delta function,

$$f(t,x) = f_\eta(t,x) = \frac{1}{\eta}\chi_{[-\frac{\eta}{2},\frac{\eta}{2}]}(x), \quad 0 < \eta \le 2\varepsilon.$$

Using the same ideas of [6], it is easy to verify the following general formula that describes the solution of $P^\varepsilon_\infty(u_0,\psi,f)$ for every $t \ge 0$. For any given integer $l \ge 0$ we have

$$u(t,x) = \begin{cases} l\varepsilon + k_l(t - t_l), & x \in [-\frac{\eta}{2}, \frac{\eta}{2}], \\ (l-1)\varepsilon + k_l(t - t_l), & x \in [-\frac{\eta}{2} - \varepsilon, \frac{\eta}{2} + \varepsilon] \setminus [-\frac{\eta}{2}, \frac{\eta}{2}], \\ \dots & \\ k_l(t - t_l), & x \in [-\frac{\eta}{2} - l\varepsilon, \frac{\eta}{2} + l\varepsilon] \setminus [-\frac{\eta}{2} - (l-1)\varepsilon, \frac{\eta}{2} + (l-1)\varepsilon], \\ 0, & x \notin [-\frac{\eta}{2} - l\varepsilon, \frac{\eta}{2} + l\varepsilon], \end{cases}$$

for $t \in [t_l, t_{l+1})$, where

$$k_l = \frac{1}{2l\varepsilon + \eta} \quad \text{and} \quad t_{l+1} = t_l + \frac{\varepsilon}{k_l}, \quad t_0 = 0.$$

This general formula is valid until the time at which the solution verifies $u(t,x) = \Psi_\varepsilon(x)$ for $x \in [-\frac{\eta}{2}, \frac{\eta}{2}]$ (the support of f), that is, until $T = t_{l^*+1}$, where

$$l^* \text{ is the first } l \text{ such that } l\varepsilon + k_l(t_{l+1} - t_l) = \Psi_\varepsilon(0)$$

and

$$\Psi_\varepsilon \text{ is the natural upper bound defined in Remark 4.2}$$

for the corresponding rescaled kernel. Observe that for this l^*, $\frac{\eta}{2} + l^*\varepsilon \le L$. From that time on the solution is stationary, that is, $u(t,x) = u(T,x)$ for all $t > T$.

From the above formula, taking limits as $\eta \to 0$, we get that the expected solution to $P^\varepsilon_\infty(u_0,\psi,\delta_0)$ is given, for any given integer $l \ge 1$, by

$$u(t,x) = \begin{cases} (l-1)\varepsilon + k_l(t - t_l), & x \in [-\varepsilon, \varepsilon], \\ (l-2)\varepsilon + k_l(t - t_l), & x \in [-2\varepsilon, 2\varepsilon] \setminus [-\varepsilon, \varepsilon], \\ \dots & \\ k_l(t - t_l), & x \in [-l\varepsilon, l\varepsilon] \setminus [-(l-1)\varepsilon, (l-1)\varepsilon], \\ 0, & x \notin [-l\varepsilon, l\varepsilon], \end{cases} \tag{4.17}$$

for $t \in [t_l, t_{l+1})$, where $k_l = \frac{1}{2l\varepsilon}$, $t_{l+1} = t_l + \frac{\varepsilon}{k_l}$, $t_1 = 0$, until $T = t_{l^*+1}$, where

$$l^* \text{ is the first } l \text{ such that } l\varepsilon + k_l(t_{l+1} - t_l) = \Psi_\varepsilon(0).$$

And from that time on the solution is stationary, that is, $u(t,x) = u(T,x)$ for all $t > T$.

Remark that, since the space of functions $K^\varepsilon_{\infty,\psi}$ is not contained into $C(\mathbb{R})$, the formulation (4.16) with $f = \delta_0$ does not make sense. Hence the function $u(t,x)$ described by (4.17) is to be understood as a *generalized solution* to $P^\varepsilon_\infty(u_0, \psi, \delta_0)$ (it is obtained as a limit of solutions to approximating problems).

Note that the function $u(T,x)$ is a "regular and symmetric pyramid" composed by squares of side ε which is one step below the upper profile Ψ_ε.

Recovering the sandpile model as $\varepsilon \to 0$. Now, to recover the sandpile model, take the limit as $\varepsilon \to 0$ in the previous example to get that $u(t,x) \to v(t,x)$, where

$$v(t,x) = (l - |x|)^+ \qquad \text{for } t = l^2,$$

until the time at which $t = L^2$, and from that time the solution is stationary.

A similar argument shows that, for any $a \in (0, L)$, the *generalized solution* to $P^\varepsilon_\infty(0, 0, \delta_a)$ converges as $\varepsilon \to 0$ to $v(t,x)$, where

$$v(t,x) = (l - |x - a|)^+ \qquad \text{for } t = l^2,$$

until the time at which $t = (L - a)^2$, and from that time the solution is stationary.

These concrete examples illustrate the general convergence result in Theorem 1.8.

Acknowledgments. We want to thank the referees for their help to improve the manuscript. Part of this work was performed during a visit of JDR to Univ. de Valencia. He is thankful for the warm hospitality and the stimulating working atmosphere found there.

REFERENCES

[1] F. Andreu, C. Ballester, V. Caselles, and J. M. Mazón, *The Dirichlet problem for the total variation flow*, J. Funct. Anal., 180 (2001), pp. 347–403.

[2] F. Andreu, V. Caselles, and J. M. Mazón, *Parabolic quasilinear equations minimizing linear growth functionals*, Progress in Mathematics, Vol. 223, Birkhauser, 2004.

[3] F. Andreu, J. M. Mazón, and J. Toledo, *Stabilization of solutions of the filtration equation with absorption and non-linear flux*, NoDEA, 2 (1995), pp. 267–289.

[4] F. Andreu, J. M. Mazón, J. D. Rossi, and J. Toledo, *The Neumann problem for nonlocal nonlinear diffusion equations*, J. Evol. Eqn., 8 (2008), pp. 189–215.

[5] F. Andreu, J. M. Mazón, J. D. Rossi, and J. Toledo, *A nonlocal p-Laplacian evolution equation with Newmann boundary conditions*, J. Math. Pures Appl., 90 (2008), pp. 201–227.

[6] F. Andreu, J. M. Mazón, J. D. Rossi, and J. Toledo, *The limit as $p \to \infty$ in a nonlocal p-Laplacian evolution equation. A nonlocal approximation of a model for sandpiles*, Calc. Var. Partial Differential Equations, to appear.

[7] G. Anzellotti, *Pairings between measures and bounded functions and compensated compactness*, Ann. Mat. Pura Appl. IV, 135 (1983), pp. 293–318.

[8] H. Attouch, *Familles d'opérateurs maximaux monotones et mesurabilité*, Ann. Mat. Pura Appl., 120 (1979), pp. 35–111.

[9] G. Barles and C. Imbert, *Second-order elliptic integro-differential equations: Viscosity solutions theory revisited*, Ann. Inst. H. Poincaré Anal. Non Linèaire, 25 (2008), pp. 567–585.

[10] J. W. Barrett and L. Prigozhin, *Dual formulations in critical state problems*, Interfaces Free Bound., 8 (2006), pp. 349-370.

[11] P. Bates and A. Chmaj, *A discrete convolution model for phase transitions*, Arch. Ration. Mech. Anal., 150 (1999), pp. 281–305.
[12] P. Bates, P. Fife, X. Ren, and X. Wang, *Travelling waves in a convolution model for phase transitions*, Arch. Ration. Mech. Anal., 138 (1997), pp. 105–136.
[13] Ph. Bénilan, L. Boccardo, T. Gallouet, R. Gariepy, M. Pierre, and J. L. Vazquez, *An L^1-theory of existence and uniqueness of solutions of nonlinear elliptic equations*, Ann. Sc. Norm. Super. Pisa, IV, XXII (1995), pp. 241-273.
[14] Ph. Bénilan and M. G. Crandall, *Completely accretive operators*, In Semigroup Theory and Evolution Equations (Delft, 1989), Lecture Notes in Pure and Appl. Math. 135, Dekker, New York, 1991, pp. 41–75.
[15] Ph. Bénilan, M. G. Crandall, and A. Pazy, *Evolution Equations Governed by Accretive Operators*, book to appear.
[16] H. Brezis, *Équations et inéquations non linéaires dans les espaces vectoriels en dualité*, Ann. Inst. Fourier, 18 (1968), pp. 115–175.
[17] H. Brezis, *Opérateur Maximaux Monotones et Semi-groupes de Contractions dans les Espaces de Hilbert*, North-Holland, Amsterdam, 1973.
[18] H. Brezis, *Analyse fonctionnelle. Théorie et applications*, Masson, 1978.
[19] H. Brezis and A. Pazy, *Convergence and approximation of semigroups of nonlinear operators in Banach spaces*, J. Funct. Anal., 9 (1972), pp. 63–74.
[20] L. Caffarelli and L. Silvestre, *Regularity Theory for Fully Nonlinear Integro-Differential Equations*, preprint.
[21] L. Caffarelli, S. Salsa, and L. Silvestre, *Regularity estimates for the solution and the free boundary of the obstacle problem for the fractional Laplacian*, Inventiones Mathematicae, 171 (2008), pp. 425–461.
[22] C. Carrillo and P. Fife, *Spatial effects in discrete generation population models*, J. Math. Biol., 50 (2005), pp. 161–188.
[23] E. Chasseigne, M. Chaves, and J. D. Rossi, *Asymptotic behaviour for nonlocal diffusion equations*, J. Math. Pures Appl., 86 (2006), pp. 271–291.
[24] X. Chen, *Existence, uniqueness and asymptotic stability of travelling waves in nonlocal evolution equations*, Adv. Differential Equations, 2 (1997), pp. 125–160.
[25] C. Cortazar, M. Elgueta, and J. D. Rossi, *A non-local diffusion equation whose solutions develop a free boundary*, Ann. Henri Poincaré, 6 (2005), pp. 269–281.
[26] C. Cortazar, M. Elgueta, J. D. Rossi, and N. Wolanski, *Boundary fluxes for non-local diffusion*, J. Differential Equations, 234 (2007), pp. 360–390.
[27] C. Cortazar, M. Elgueta, J. D. Rossi, and N. Wolanski, *How to approximate the heat equation with Neumann boundary conditions by nonlocal diffusion problems*, Arch. Ration. Mech. Anal., 187 (2008), pp. 137–156.
[28] M. G. Crandall, *Nonlinear semigroups and evolution governed by accretive operators*, in Proceedings of the Symposium in Pure Mathematics, Part I, Vol. 45, F. Browder ed., AMS, Providence, RI, 1986, pp. 305–338.
[29] S. Dumont and N. Igbida, *On a Dual Formulation for the Growing Sandpile Model*, European J. Appl. Math., to appear.
[30] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1972.
[31] P. Fife, *Some nonclassical trends in parabolic and parabolic-like evolutions*, Trends in nonlinear analysis, Springer, Berlin, 2003, pp. 153–191.
[32] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and their Applications*, Academic Press, New York, 1980.
[33] S. Kindermann, S. Osher, and P. W. Jones, *Deblurring and denoising of images by nonlocal functionals*, Multiscale Model. Simul., 4 (2005), pp. 1091–1115.
[34] U. Mosco, *Convergence of convex sets and solutions of variational inequalities*, Adv. Math., 3 (1969), pp. 510–585.
[35] L. Prigozhin, *Variational models of sandpile growth*, Eur. J. Appl. Math., 7 (1996), pp. 225–236.
[36] L. Silvestre, *Hölder estimates for solutions of integro differential equations like the fractional Laplace*, Indiana Univ. Math. J., 55 (2006), pp. 1155–1174.

SIAM J. MATH. ANAL.
Vol. 40, No. 5, pp. 1852–1881

GROW-UP RATE AND REFINED ASYMPTOTICS FOR A TWO-DIMENSIONAL PATLAK–KELLER–SEGEL MODEL IN A DISK*

NIKOS I. KAVALLARIS† AND PHILIPPE SOUPLET‡

Abstract. We consider a special case of the Patlak–Keller–Segel system in a disc, which arises in the modeling of chemotaxis phenomena. For a critical value of the total mass, the solutions are known to be global in time but with density becoming unbounded, leading to a phenomenon of mass-concentration in infinite time. We establish the precise grow-up rate and obtain refined asymptotic estimates of the solutions. Unlike in most of the similar, recently studied, grow-up problems, the rate is neither polynomial nor exponential. In fact, the maximum of the density behaves like $e^{\sqrt{2t}}$ for large time. In particular, our study provides a rigorous proof of a behavior suggested by Sire and Chavanis [*Phys. Rev. E* (3), 66 (2002), 046133] on the basis of formal arguments.

Key words. chemotaxis system, critical mass, grow-up, sub-/supersolutions

AMS subject classifications. Primary, 35Q, 35K60, 35B40, 92C17; Secondary, 35Q72

DOI. 10.1137/080722229

1. Introduction.

1.1. The complete Patlak–Keller–Segel model. Out of the many mathematical models that have been proposed to deal with particular aspects of chemotaxis, that proposed by Patlak in 1953 (cf. [40]) and Keller and Segel in 1970 (cf. [32]) has received particular attention. The so-called (two-dimensional) Patlak–Keller–Segel model consists of two equations, describing the evolution of the population density $\rho(x,t)$ of bacteria and the concentration $c(x,t)$ of a chemical attracting substance in a bounded domain $\Omega \subset \mathbb{R}^2$ and in a time interval $[0,T]$:

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (D_1 \nabla \rho - \chi \rho \nabla c), \tag{1.1}$$

$$\theta \frac{\partial c}{\partial t} = \Delta c - ac + \rho. \tag{1.2}$$

More precisely, the first equation describes the random (Brownian) diffusion of the population of cells, which is biased in the direction of a drift velocity, proportional to the gradient of the concentration of the chemoattractant. The diffusion coefficient is denoted by $D_1 > 0$, and the proportionality coefficient of the drift (mobility parameter) is denoted by $\chi > 0$. According to the second equation, the chemoattractant, which is directly emitted by the cells, diffuses with a diffusion coefficient $D_2 = 1/\theta > 0$ on the substrate, while it is generated proportionally to the density of cells and at the same time is degraded with a rate equal to $a/\theta \geq 0$. In order for system (1.1)–(1.2) to be well posed, it should be supplemented with some initial conditions

$$\rho(x,0) = \rho_0(x) \geq 0, \quad c(x,0) = c_0(x) \geq 0, \tag{1.3}$$

*Received by the editors April 23, 2008; accepted for publication (in revised form) September 15, 2008; published electronically January 7, 2009.
http://www.siam.org/journals/sima/40-5/72222.html
†Department of Statistics and Actuarial-Financial Mathematics, University of the Aegean, Vourlioti Building, Gr-83200 Karlovassi, Samos, Greece (nkaval@aegean.gr).
‡Laboratoire Analyse Géometrie et Applications, UMR CNRS 7539, Institut Galilée, Université Paris-Nord, 99 av. J.-B. Clément, 93430 Villetaneuse, France (souplet@math.univ-paris13.fr).

along with conditions on the boundary $\partial\Omega$. A natural boundary condition, since it guarantees the conservation of total mass, is the no-flux–type condition for ρ, namely,

$$\frac{\partial \rho}{\partial \nu} - \rho \frac{\partial c}{\partial \nu} = 0 \quad \text{on } \partial\Omega, \tag{1.4}$$

where ν stands for the outer unit normal vector at $\partial\Omega$. As for c, a Dirichlet-type boundary condition is assumed, i.e.,

$$c = 0 \quad \text{on } \partial\Omega; \tag{1.5}$$

cf. [7, 46]. Note that the parabolic system (1.1)–(1.2) preserves the nonnegativity of the initial conditions, i.e., $\rho, c \geq 0$ for $t > 0$, which is also expected to be true for the physical problem. For simplicity, D_1, χ are considered to be constant and under suitable scaling can be taken as $D_1 = \chi = 1$.

In view of experimental facts, the coefficients θ and a are assumed to be small, and a simplified form of the Patlak–Keller–Segel system is obtained (in fact, this corresponds to the case when the diffusion and production of c are much faster than the dynamics of ρ and the degradation of c). Namely, by considering the limiting case $\theta, a \to 0+$, the parabolic-parabolic system (1.1)–(1.5) is reduced to the elliptic-parabolic system

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\nabla \rho - \rho \nabla c), \quad x \in \Omega,\ t \in (0, T), \tag{1.6}$$

$$-\Delta c = \rho, \quad x \in \Omega,\ t \in (0, T), \tag{1.7}$$

$$\frac{\partial \rho}{\partial \nu} - \rho \frac{\partial c}{\partial \nu} = 0, \quad x \in \partial\Omega,\ t \in (0, T), \tag{1.8}$$

$$c = 0, \quad x \in \partial\Omega,\ t \in (0, T), \tag{1.9}$$

$$\rho(x, 0) = \rho_0(x) \geq 0, \quad x \in \Omega. \tag{1.10}$$

Note that in order for (1.6)–(1.10) to be well posed, only the initial data $\rho(x, 0) = \rho_0(x)$ must be prescribed. Moreover, owing to the boundary condition (1.8), the total (mass) population of cells is conserved; that is,

$$||\rho(\cdot, t)||_1 = ||\rho_0||_1 =: \Lambda \quad \text{for } t > 0.$$

For a more detailed analysis regarding the modeling as well as the behavior of solutions of chemotaxis systems, see the review papers [24, 28, 29] as well as the monograph [48]. Here, it should be noticed that system (1.6)–(1.10) is also known as the Smoluchowski–Poisson system and can describe the motion of the mean field of many self-gravitating particles [13, 1, 2, 7, 46, 49, 50] or that of polymer molecules [16]. The behavior of the solution to (1.6)–(1.10) strongly depends on the parameter Λ. In fact, if $\Lambda > 8\pi$ and $\Omega = B(0, R)$, $R > 0$, then solutions of (1.6)–(1.10) blow up in a finite time $T^*(\rho_0)$; that is,

$$\lim_{t \to T^*} ||\rho(\cdot, t)||_{H^1} = \lim_{t \to T^*} ||\rho(\cdot, t)||_{L^p} = \lim_{t \to T^*} \int_\Omega (\rho \log \rho)(x, t)\, dx = \infty$$

for every $p > 1$; see [9, Theorem 2(i)], [2, Theorem 2]. On the other hand, for $\Lambda < 8\pi$, all solutions of system (1.6)–(1.10) are global in time; cf. [7, Theorem 2 (iv)]. In the critical case $\Lambda = 8\pi$ an infinite-time blow-up (grow-up) occurs, i.e., $||\rho(\cdot, t)||_\infty \to \infty$ as $t \to \infty$; cf. [39, Theorem 3], [5, Proposition 3.2]. Finite- or infinite-time blow-up can

be accompanied by the occurrence of a δ-function formation in the blow-up set (this represents the trend of populations to concentrate to form sporae) and is known in the literature as *chemotactic collapse*. This phenomenon was conjectured by Childress and Percus [14] and Nanjundiah [38] and was first verified, via matched asymptotics arguments, for a radially symmetric simplified Patlak–Keller–Segel system in [25]. A result regarding the infinite-time Dirac mass formation for ρ can be found in [39], where some characterization of grow-up (mass-concentration) points together with more grow-up results for different types of boundary conditions are also obtained [39, Theorems 2 and 3]. For blow-up results concerning a variation of system (1.6)–(1.10) but with Neumann boundary conditions for both ρ and c, see [30, 35, 3, 36, 44].

1.2. The simplified Patlak–Keller–Segel model in the radial case. In the case when Ω is the ball $B(0,R)$, $R>0$, and the initial data $\rho_0(x)=\rho_0(r)$ is radially symmetric, the solution of system (1.6)–(1.10) is radially symmetric, i.e., $\rho(x,t)=\rho(r,t)$, with $r=|x|$. In this case the elliptic-parabolic system (1.6)–(1.10) can be greatly simplified. Namely, by introducing the cumulative mass distribution

$$Q(r,t):=\int_{B(0,r)}\rho(x,t)\,dx=2\pi\int_0^r s\rho(s,t)\,ds,$$

which is equal to the mass contained in the sphere $B(0,r)$, the system reduces to a single equation

$$Q_t=Q_{rr}-\frac{1}{r}Q_r+\frac{1}{2\pi r}QQ_r,\quad 0<r<R,\ t>0, \tag{1.11}$$

$$Q(0,t)=0,\quad Q(R,t)=\Lambda. \tag{1.12}$$

By the definition of Q, the function

$$Q(r,0)=Q_0(r) \tag{1.13}$$

is positive nondecreasing and satisfies the compatibility conditions $Q_0(0)=0$ and $Q_0(R)=\Lambda$.

As mentioned in [4, 7], the formulation (1.11)–(1.13) allows the consideration of some initial data for the density ρ which could be either unbounded or singular (such as measures)—a case that seems rather realistic. This means that the initial data Q_0 for problem (1.11)–(1.13) could have unbounded derivatives $Q_{0,r}$ or even be discontinuous. Moreover, using formulation (1.11)–(1.13), we have the comparison principle at hand, which is not available for system (1.6)–(1.10). Due to the scaling properties of (1.11), we can assume without loss of generality that problem (1.11)–(1.13) is posed in the unit ball $B(0,1)$. (Indeed, it is easily seen that if $Q(r,t)$ is a solution of (1.11)–(1.13), then $Q(Rr,R^2t)$ is also a solution.)

Using the new variable $x=r^2$ (no confusion with the original variable x in (1.6)–(1.10) should occur) and defining $N(x,t)=Q(r,t)$, we are led to the problem

$$N_t=4xN_{xx}+\frac{1}{\pi}NN_x,\quad 0<x<1,\ t>0, \tag{1.14}$$

$$N(0,t)=0,\quad N(1,t)=\Lambda, \tag{1.15}$$

$$N(x,0)=N_0(x),\quad 0<x<1. \tag{1.16}$$

Note that (1.14) differs from the Burgers equation only by the variable coefficient x in the diffusion term. The above problem, although it degenerates at $x=0$, may be

handled more easily than (1.11)–(1.13) since it contains fewer terms and at the same time does not have any singular coefficients in the first order terms.

As expected, the behavior of the solution of problem (1.14)–(1.16) (which is well defined for suitable initial data) depends on Λ. For $\Lambda > 8\pi$ the solution N ceases to exist in a finite time T^*; more precisely, the boundary condition $N(0,t) = 0$ is no longer fulfilled at $t = T^*$. Moreover, a "gradient blow-up" occurs at $t = T^*$ in the sense that $N_x(0,t) \to \infty$ as $t \to T^*$ and the density ρ also becomes unbounded at time T^*; cf. [8, Theorem 2(i)]. On the other hand, for $0 < \Lambda < 8\pi$ and any (admissible) initial data there is a unique global-in-time solution $N \in C([0,\infty); L^2(0,1)) \cap C^{2,1}((0,1)\times(0,\infty))$. Furthermore, N converges to the unique steady state solution:

$$N(\cdot,t) \to N_d = 8\pi \frac{x}{x+d} \quad \text{as } t \to \infty \tag{1.17}$$

in $L^p(\Omega)$, $p \geq 1$, and even in $L^\infty(\Omega)$ provided that $\sup_{t\geq 0} |N_x|_\infty < \infty$, where $d = \frac{8\pi}{\Lambda} - 1 > 0$; cf. [5]. In this case the rate of the L^1-convergence of $N(\cdot,t)$ to N_d is shown to be exponential.

In the borderline case $\Lambda = 8\pi$, the situation is still different and the problem exhibits a typical critical behavior. Namely, it is proved in [5] that there exists a global-in-time solution N, which converges to the "singular" steady state $\widetilde{N}(x) \equiv 8\pi$ (note that $\widetilde{N}$ does not satisfy the boundary condition at $x = 0$). Actually, as was proven in [39, Theorem 3], an infinite-time Dirac mass formation at the origin $r = 0$ of the ball occurs in this case. However, neither an estimation of the grow-up rate nor the asymptotic profile of the grow-up are provided in [39]. On the other hand, the authors in [5, Proposition 3.2] obtain the decay estimate

$$||N(\cdot,t) - 8\pi||_{L^1} \leq \frac{8\pi}{t} \quad \text{for} \quad t \geq 1. \tag{1.18}$$

Estimate (1.18) seems to be far from optimal since formal asymptotics performed in [46] suggest a temporal decay estimate of order

$$||N(\cdot,t) - 8\pi||_{L^1} \approx O\big(e^{-\sqrt{2t}}\big) \quad \text{as } t \to \infty. \tag{1.19}$$

Remarks 1.1. (a) The (nonradial) Patlak–Keller–Segel system has also been studied in the whole plane $\mathbb{R}^2$. Again, the behavior of solutions depends on the initial mass of the system, and a dichotomy is found [11, 17]. More precisely, assuming $0 \leq (1+|x|^2)\rho_0$ and $\rho_0 \log \rho_0 \in L^1$, there exists a critical value of the mass $N_c := 8\pi$ such that if $0 < ||\rho_0||_1 < N_c$ (subcritical case), only global-in-time solutions exist, while if $||\rho_0||_1 > N_c$ (supercritical case), the solutions blow up in finite time [11, 41]. Moreover, in the subcritical case, solutions converge to a self-similar profile as $t \to \infty$ [6, 11]. Finally, for the critical case $N = N_c$, which was studied in [10], the solution is global-in-time and grows up as a Dirac mass at the center of mass as $t \to \infty$.

(b) The only previous mathematical study of grow-up rates for a system of Patlak–Keller–Segel type concerns high dimensions, namely, $n \geq 11$, and was performed recently in [45]. There, some radial global unbounded solutions were constructed in a ball and an infinite sequence of polynomial grow-up rates was obtained (for a suitable sequence of initial data). On the contrary, our results in the present paper for $n = 2$ exhibit a grow-up rate independent of the initial data.

(c) Concerning the parabolic-parabolic Patlak–Keller–Segel system in a bounded domain with Neumann conditions, interesting results can be found in [26, 27] and [19,

37], respectively, on the asymptotics of finite-time blow-up and on global existence and the convergence of global bounded solutions.

(d) Problem (1.14)–(1.16) is in fact independent of the boundary condition (1.9) for c. Of course the boundary condition for c has to be taken into account if one wants to determine c, and not only ρ, from N.

(e) In the case where the diffusion of cells is very slow as compared to the diffusion of the chemoattractant, i.e., when $D_1 \ll D_2$ in (1.1)–(1.2), the complete system (1.1)–(1.5) is reduced to a single but nonlocal equation; cf. [51]. The global existence (subcritical case) and the finite-time blow-up (supercritical case) of solutions of the derived nonlocal equation in the two-dimensional case are studied in [31].

1.3. Heuristic description of the results and methods. Our aim is to prove rigorously the decay rate (1.19) as well as to provide a refined asymptotic profile for $N(x,t)$ as $t \to \infty$. To normalize the constants arising in the calculations we shall work with the equivalent problem

$$u_t = xu_{xx} + 2uu_x, \qquad 0 < x < 1,\ t > 0, \tag{1.20}$$

$$u(0,t) = 0, \qquad u(1,t) = \xi := \frac{\Lambda}{8\pi}, \qquad t > 0, \tag{1.21}$$

$$u(x,0) = u_0(x), \qquad 0 < x < 1, \tag{1.22}$$

which is obtained from (1.14)–(1.16) by setting $N(x,t) = 8\pi u(x,4t)$.

It is clear that the stationary part $xu_{xx} + 2uu_x = 0$ of the parabolic equation (1.20) is invariant under the rescaling $u(x) \mapsto u(kx)$ $(k > 0)$. Moreover, the steady state solution for $\xi < 1$ is given by

$$U_a(x) = 1 - \frac{1}{ax+1} = \frac{ax}{ax+1}$$

for $a = U_a'(0) = \frac{\xi}{1-\xi} > 0$. Rewriting (1.17) in terms of u, we have

$$u(x,t) \to U_a(x) \quad \text{as } t \to \infty.$$

Also, observe that $U_a(x)$ converges in a monotone increasing way to the "singular" steady state $U \equiv 1$ as $a \to \infty$ $(\xi \to 1)$.

Motivated by the above considerations, we shall look for sub- and supersolutions of problem (1.20)–(1.22), which are perturbations of a moving family of steady states

$$U_{a(t)}(x), \tag{1.23}$$

the perturbation being defined in terms of the self-similar variable $y = a(t)x$. Here, a is a function of time, diverging to infinity, which is a priori unknown and will be eventually identified by a suitable "matching" procedure (see Remark 1.3). Note that such a form was used in [46] to construct "approximate solutions" leading to the formal asymptotics mentioned above. However, the expansions in [46] contained only a correction term at the first order, while those that we here construct involve first and second order terms (see Lemmas 4.1 and 4.2). This seems necessary to obtain (rigorous) sub- and supersolutions living "close" to the actual solution and eventually provides us with the desired decay rate as well as a good description of the asymptotic profile.

As a consequence of this construction, we shall derive an infinite-time boundary gradient grow-up result, with the grow-up rate

(1.24)
$$u_x(0,t) = A(t)\Big(1 + O\big(t^{-1/2}\log t\big)\Big) \quad \text{as } t \to \infty, \quad \text{where } A(t) = \exp\left[\frac{5}{2} + \sqrt{2t}\right].$$

In terms of the solution of the original system (1.6)–(1.10), (1.24) gives an estimate of the central density of bacteria, since $\rho(0,t) = 8u_x(0,4t)$. Note that (1.24) was also predicted by formal arguments in [46]. At the same time, we show the C^1 regularity of u up to the boundary for all finite time intervals, hence ruling out the possibility of $\sup_{x\in[0,1]} u_x(x,t)$ blowing up in finite time. This problem was left open in [5]. Moreover, we obtain a precise asymptotic expansion of the solution (see formula (2.8) in Theorem 2.1). It expresses the solution as the sum of a quasi-stationary profile (cf. (1.23)) and a correction term which becomes significant only for x bounded away from 0. As a consequence, we obtain the decay

$$\|u(\cdot,t) - 1\|_{L^1(0,1)} = \Big(1 + O\big(t^{-1/2}\log t\big)\Big)\sqrt{2t}\exp\left[-\frac{5}{2} - \sqrt{2t}\right] \quad \text{as } t \to \infty,$$

again in accordance with the predictions in [46].

Remark 1.2. The study of unbounded global classical solutions of superlinear parabolic problems and their asymptotic behavior has recently attracted substantial mathematical interest. Particular effort has been devoted to the reaction-diffusion equation

$$u_t - \Delta u = u^p, \tag{1.25}$$

where such solutions are known to exist for suitable $p > 1$. Let us mention the works [33, 18, 22] for the Dirichlet problem in a ball and [42, 20, 21, 34] for the Cauchy problem. See also [43, sections 22 and 29] for related questions. The case of the Frank–Kamenetskii equation (with nonlinearity e^u instead of u^p) is also studied in [18]. As for the diffusive Hamilton–Jacobi equation

$$u_t - u_{xx} = |u_x|^p, \tag{1.26}$$

with $p > 2$, results of this kind can be found in [47]. A common feature in all these examples is the stabilization of the solution to a singular steady state, the growing-up quantity being $\|u(t)\|_\infty$ (resp., $\|u_x(t)\|_\infty$) for (1.25) (resp., (1.26)). It has to be noted that the grow-up rates, as $t \to \infty$, usually behave like either $e^{\mu t}$ or t^k. The only known exception (see [22]) seems to be the case of (1.25) with zero boundary conditions, in a ball of $\mathbb{R}^4$ with critical Sobolev exponent ($p = 3$). In this situation, unlike in spatial dimensions $n \neq 4$, there holds $\log\|u(t)\|_\infty \sim 2\sqrt{t}$, thus leading to a rate similar to that in our problem.

Remark 1.3. Let us point out that our determination of the grow-up rate for problem (1.20)–(1.22) is achieved through the matching of (sub- or super-) solutions with the imposed boundary condition at the right endpoint $x = 1$, an idea also present in [22]. This is different from what is done in [18, 20, 21, 47], where the grow-up rate is determined by the matching between inner and outer (sub-/super-) solutions. In those works, the inner solution corresponds to a self-similar, quasi-stationary evolution along a continuum of regular steady states (similar to (1.23)), but the outer solution is obtained by a linearization around the singular steady state. Here, on the contrary,

the behavior of u in the inner and outer regions is unified through a single self-similar variable $y = a(t)x$ (cf. formulae (4.1) and (4.27) below). Moreover, we point out that, in our case, a linearization around the "singular" steady state $U \equiv 1$ would not give the desired grow-up rate (1.24) but only a nonoptimal exponential upper bound based on an associated eigenvalue problem.

The paper is organized as follows. In section 2 we state the main results. Section 3 contains a number of preliminary results: In subsection 3.1, we recall the basic facts concerning local existence and comparison. In subsections 3.2 and 3.3, we show that a control on the slope at $x = 0$ is enough to prevent gradient blow-up (Lemma 3.2), and we obtain preliminary estimates of solutions for small time, which will be useful in section 4 to initialize the comparison with the main sub-/supersolutions (Lemmas 3.3 and 3.4). Subsection 3.4 is devoted to the study of a second order ordinary differential operator which plays a key role in the subsequent construction of sub-/supersolutions. The proofs of the main results are given in section 4: Subsections 4.1 and 4.2 are devoted to the construction of the main sub- and supersolutions, respectively; the proofs of Theorem 2.1 and Corollary 2.2 are finally completed in subsection 4.3.

2. Main results. Consider the problem

$$u_t - xu_{xx} = 2uu_x, \quad 0 < x < 1,\ t > 0, \tag{2.1}$$

$$u(0,t) = 0, \quad t > 0, \tag{2.2}$$

$$u(1,t) = 1, \quad t > 0, \tag{2.3}$$

$$u(x,0) = u_0(x), \quad 0 \le x \le 1. \tag{2.4}$$

Concerning the initial data, we assume that

$$u_0 \in C([0,1]), \quad u_0(0) = 0, \quad u_0(1) = 1, \quad u_0 \text{ is nondecreasing}, \tag{2.5}$$

and that

$$u_0(x) \le Kx, \quad 0 < x < 1, \quad \text{for some } K \ge 1. \tag{2.6}$$

Problem (2.1)–(2.4) admits a unique global solution, with $u \in C([0,1] \times [0,\infty))$, $u \in C^{2,1}((0,1] \times (0,\infty))$ (see [5] and section 3.1 below). Moreover, it was shown in [5] that $0 \le u \le 1$, u is nondecreasing in x, and

$$\lim_{t\to\infty} u(x,t) = 1, \quad \text{uniformly for } x \text{ in compact subsets of } (0,1]. \tag{2.7}$$

Our main results are the following.

THEOREM 2.1. *Let u_0 satisfy* (2.5) *and* (2.6)*, and denote by u the global solution of problem* (2.1)–(2.4)*. Then there holds*

$$1 - u(x,t) = \frac{1 - x + O\big(t^{-1/2}\log t\big)}{1 + A(t)x} \quad \textit{uniformly in } [0,1],\ \textit{as } t \to \infty, \tag{2.8}$$

with

$$A(t) = \exp\left[\frac{5}{2} + \sqrt{2t}\right]. \tag{2.9}$$

Moreover, we have the regularity property

$$u_x \in C\big([0,1] \times (0,\infty)\big) \tag{2.10}$$

and the estimate

$$u_x(0,t) = A(t)\Big(1 + O\big(t^{-1/2}\log t\big)\Big) \quad \text{as } t \to \infty. \tag{2.11}$$

COROLLARY 2.2. *Let u_0 satisfy (2.5) and (2.6), denote by u the global solution of problem (2.1)–(2.4), and let $A(t)$ be given by (2.9).*

(i) *The solution u satisfies the inner layer expansion (quasi-stationary behavior)*

$$1 - u(x,t) = \frac{1 + o(1)}{1 + A(t)x} \quad \text{uniformly in any region } x \le o(1), \text{ as } t \to \infty.$$

(ii) *We have the L^1-decay rate*

$$\|u(\cdot,t) - 1\|_{L^1(0,1)} = \Big(1 + O\big(t^{-1/2}\log t\big)\Big)\sqrt{2t}\exp\left[-\frac{5}{2} - \sqrt{2t}\right] \quad \text{as } t \to \infty.$$

3. Preliminaries.

3.1. Local existence and comparison principle. By [5, Theorem 2.1], we know that for any u_0 satisfying (2.5), problem (2.1)–(2.4) admits a global solution u in the following sense:

$$u \in C\big([0,\infty); L^1(0,1)\big), \tag{3.1}$$

$$u \in C^{2,1}\big((0,1] \times (0,\infty)\big), \tag{3.2}$$

$$u_x \ge 0, \quad 0 < x \le 1,\ t > 0, \tag{3.3}$$

and

$$u_t - xu_{xx} = 2uu_x, \quad 0 < x < 1,\ t > 0, \tag{3.4}$$

$$\lim_{x \to 0^+} u(x,t) = 0 \quad \text{for a.e. } t > 0, \tag{3.5}$$

$$u(1,t) = 1, \quad t > 0, \tag{3.6}$$

$$u(\cdot,0) = u_0 \quad \text{in } L^1(0,1). \tag{3.7}$$

Note that the solution is obtained in [5] as a limit of solutions of regularized problems, where xu_{xx} is replaced by $(x+\varepsilon)u_{xx}$, $\varepsilon > 0$. Also, for each $T > 0$, u is the unique local solution of (3.1)–(3.7) on $(0,T)$. If, moreover, u_0 satisfies (2.6), then

$$u \in C\big([0,1] \times [0,\infty)\big);$$

see also [8, Theorem 1 (i)]. The continuity for $t > 0$ and $x = 0$ follows from [5, Propositions 2.4 and 2.5]. For $t = 0$ and $x \in [0,1]$, the continuity can be established by comparison with simple barrier functions.

The following proposition provides a comparison principle suitable to our needs.

PROPOSITION 3.1. *Let $\tau > 0$ and the functions u, v satisfy the following regularity conditions:*

$$u, v \in C\big([0,\tau); L^1(0,1)\big), \tag{3.8}$$

$$u, v \in C^{2,1}\big((0,1] \times (0,\tau)\big), \tag{3.9}$$

$$u_x, v_x \in L^1_{loc}\big([0,1] \times (0,\tau)\big), \tag{3.10}$$

$$u, v \in L^\infty_{loc}\big([0,1] \times (0,\tau)\big). \tag{3.11}$$

Assume that

$$u_t - xu_{xx} - 2uu_x \le v_t - xv_{xx} - 2vv_x, \quad 0 < x < 1,\ 0 < t < \tau, \tag{3.12}$$

$$\lim_{x\to 0^+} u(x,t) \le \lim_{x\to 0^+} v(x,t) \quad \textit{for a.e. } t \in (0,\tau), \tag{3.13}$$

$$u(1,t) \le v(1,t), \quad 0 < t < \tau, \tag{3.14}$$

$$u(\cdot,0) \le v(\cdot,0) \quad \textit{a.e. in } (0,1). \tag{3.15}$$

Then $u \le v$ *in* $(0,1)\times(0,\tau)$.

Observe that (3.10) implies that the limits in (3.13) exist for a.e. $t \in (0,\tau)$. Note also that for any u_0 satisfying (2.5), the solution u of problem (3.1)–(3.7) satisfies conditions (3.10) and (3.11) as a consequence of (3.3), (3.5), and (3.6).

Proof of Proposition 3.1. The proof is a modification of the stability proof in [5, Theorem 3.1]. Let $z = u - v$. By (3.12), we have

$$z_t \le xz_{xx} + 2uu_x - 2vv_x = \partial_x\big(xz_x + z(u+v-1)\big), \quad 0 < x < 1,\ 0 < t < \tau. \tag{3.16}$$

For $\delta \in (0,1)$ we define the following C^1 (and piecewise C^2) convex approximations of the function $s \to s_+ = \max(s,0)$:

$$\phi_\delta(s) = \begin{cases} 0 & \text{if} \quad -\infty < s \le \delta, \\ (2\delta)^{-1}(s-\delta)^2 & \text{if} \quad \delta \le s \le 2\delta, \\ s - 3\delta/2 & \text{if} \quad 2\delta < s < \infty. \end{cases}$$

Fix $0 < t_1 < t_2 < \tau$ and $\delta \in (0,1)$. Then, for any $\varepsilon \in (0,1)$, multiplying (3.16) by $\phi_\delta'(z)$, integrating by parts, and using (3.11), (3.14), $0 \le \phi_\delta' \le 1$, and $\phi_\delta'' \ge 0$, we obtain

$$\begin{aligned}
&\int_\varepsilon^1 \phi_\delta(z(x,t_2))\,dx - \int_\varepsilon^1 \phi_\delta(z(x,t_1))\,dx \\
&= \int_{t_1}^{t_2}\int_\varepsilon^1 \phi_\delta'(z) z_t\,dx\,dt \le \int_{t_1}^{t_2}\int_\varepsilon^1 \phi_\delta'(z)\partial_x\big(xz_x + z(u+v-1)\big)\,dx\,dt \\
&= \int_{t_1}^{t_2}\Big[\phi_\delta'(z)\big(xz_x + z(u+v-1)\big)\Big]_\varepsilon^1 dt - \int_{t_1}^{t_2}\int_\varepsilon^1 \phi_\delta''(z)\big(xz_x + z(u+v-1)\big)z_x\,dx\,dt \\
&\le \varepsilon\int_{t_1}^{t_2} |z_x(\varepsilon,t)|\,dt + C\int_{t_1}^{t_2}\phi_\delta'(z(\varepsilon,t))\,dt - \int_{t_1}^{t_2}\int_\varepsilon^1 \phi_\delta''(z)(u+v-1)zz_x\,dx\,dt \\
&\equiv I_\varepsilon + J_\varepsilon + K_\varepsilon,
\end{aligned}$$

where C depends on t_1, t_2 but is independent of ε (and δ). Owing to (3.10), there exists a sequence $\varepsilon_n \to 0+$ such that $\lim_{n\to\infty} I_{\varepsilon_n} = 0$. Next, since $\lim_{n\to\infty}\phi_\delta'(z(\varepsilon_n,t)) = 0$ for a.e. $t \in (t_1,t_2)$ due to (3.13) and the definition of ϕ_δ, we deduce that $\lim_{n\to\infty} J_{\varepsilon_n} = 0$ by dominated convergence. Consequently,

$$\begin{aligned}
\int_0^1 \phi_\delta(z(x,t_2))\,dx - \int_0^1 \phi_\delta(z(x,t_1))\,dx &\le -\int_{t_1}^{t_2}\int_0^1 \phi_\delta''(z)(u+v-1)zz_x\,dx\,dt \\
&\le C\int_{t_1}^{t_2}\int_0^1 |z\phi_\delta''(z)||z_x|\,dx\,dt.
\end{aligned}$$

Now, observe that $\lim_{\delta\to 0}\phi_\delta(s) = s_+$ and $\lim_{\delta\to 0} s\phi_\delta''(s) = 0$ for each $s \in \mathbb{R}$. Using $0 \le \phi_\delta(s) \le s_+$, (3.10), and (3.11), we may pass to the limit $\delta \to 0$ by dominated

convergence in the preceding inequality, and we obtain

$$\int_0^1 z_+(x,t_2)\,dx - \int_0^1 z_+(x,t_1)\,dx \le 0.$$

Letting $t_1 \to 0+$ and using (3.8) and (3.15), we conclude that $\int_0^1 z_+(x,t_2)\,dx = 0$ for all $t_2 \in (0,\tau)$; hence $u \le v$ in $(0,1)\times(0,\tau)$. □

3.2. Sufficient condition for C^1 regularity. As noted in, e.g., [12, section 2.2] and [23, section 2], by means of the transformation

$$w(r,t) := \frac{8}{r^2}u(r^2,4t) = \frac{1}{\pi r^2}\int_{B_r}\rho(y,4t)\,dy$$

(and $w_0(r) := 8r^{-2}u_0(r^2)$), problem (2.1)–(2.4) becomes equivalent to

$$w_t - \tilde{\Delta}w = w^2 + \tfrac{r}{2}\,w\,w_r, \quad 0<r<1,\ t>0, \tag{3.17}$$

$$w_r(0,t) = 0, \quad t>0, \tag{3.18}$$

$$w(1,t) = 8, \quad t>0, \tag{3.19}$$

$$w(r,0) = w_0(r), \quad 0<r<1, \tag{3.20}$$

where $\tilde{\Delta}w := w_{rr} + \frac{3}{r}w_r$, which in turn corresponds to the radial Laplacian in four space dimensions. It should be noticed that w has the same scale invariance as ρ but is smoother than ρ. Problem (3.17)–(3.20) turns out to be convenient regarding the study of C^1 regularity of u up to the boundary, which was left open in [5]. Namely, using this transformation, one can show the following additional properties for u.

LEMMA 3.2. *Let u_0 satisfy* (2.5) *and* (2.6)*, and denote by u the global solution of problem* (2.1)–(2.4)*. For any given $T>0$, if*

$$\sup_{0<x<1,\,0<t<T}\frac{u(x,t)}{x} < \infty, \tag{3.21}$$

then

$$u_x \in C\big([0,1]\times(0,T]\big) \tag{3.22}$$

and

$$u_x(0,t) > 0, \quad 0<t\le T. \tag{3.23}$$

Note that the assumption (3.21) (for all $T>0$) will be shown in section 4 (see Lemma 4.3(ii)) as a consequence of our main supersolution construction, hence leading to the global C^1 regularity property (2.10) in Theorem 2.1.

Proof of Lemma 3.2. Since the semilinear parabolic equation in (3.17) has only linear growth with respect to the gradient, standard arguments based on the variation-of-constants formula (see, e.g., [43, Example 51.30] or [15, p. 889]) show that problem (3.17)–(3.20) is locally well posed in the space of (radial, nonnegative) L^∞-functions. More precisely, for any $0 \le w_0 \in L^\infty(0,1)$, there exists a unique, maximal (radial, nonnegative) classical solution w of (3.17)–(3.20), with $w \in C^{2,1}\big([0,1]\times(0,T_m)\big)$ and $w \in C\big([0,T_m);L^q(0,1)\big)$ for all finite $q \le 1$. Moreover,

$$T_m < \infty \Longrightarrow \lim_{t\to T_m}\|w(t)\|_\infty = \infty. \tag{3.24}$$

Now, if u_0 satisfies (2.5)–(2.6), then $w_0(r) := 8r^{-2}u_0(r^2)$ verifies $0 \le w_0 \in L^\infty(0,1)$. Denote by w the corresponding maximal solution of (3.17)–(3.20), and let

$$\tilde{u}(x,t) = \frac{x}{8} w(\sqrt{x}, t/4), \qquad 0 \le x \le 1,\ 0 \le t < 4T_m.$$

We see that $\tilde{u}$ satisfies the regularity conditions in (3.8), (3.9), and (3.11) with $\tau = 4T_m$. Also, since

$$\tilde{u}_x(x,t) = \frac{1}{8} w(\sqrt{x}, t/4) + \frac{\sqrt{x}}{16}\, w_r(\sqrt{x}, t/4), \qquad 0 < x \le 1,\ 0 < t < 4T_m,$$

we have

$$\tilde{u} \in C^{1,0}([0,1] \times (0, 4T_m)) \tag{3.25}$$

and hence in particular (3.10) with $\tau = 4T_m$. Then one easily checks that $\tilde{u}$ solves (2.1)–(2.3) on $[0, 4T_m)$, along with (3.7). We may thus apply Proposition 3.1 to deduce that $\tilde{u} = u$ on $(0, T_0)$ with $T_0 = \min(T, 4T_m)$.

We claim that $T_m > T/4$. Indeed, if $T_m \le T/4$, then (3.24) implies

$$\lim_{t \to 4T_m} \sup_{0<x<1} \frac{u(x,t)}{x} = \infty,$$

contradicting (3.21). Consequently, property (3.22) follows from (3.25). Moreover, $w > 0$ in $[0,1] \times (0, T_m)$ by the strong maximum principle, which readily implies (3.23). □

3.3. Small time estimates. Throughout the paper, we denote by $\mathcal{P}$ the parabolic operator defined by

$$\mathcal{P}v := v_t - x v_{xx} - 2 v v_x. \tag{3.26}$$

The following two lemmas will be useful to initialize the comparison between u and the main sub-/supersolutions constructed in section 4.

LEMMA 3.3. *Let u_0 satisfy* (2.5) *and* (2.6)*, and denote by u the global solution of problem* (2.1)–(2.4)*. Then there exist $\tau, \eta > 0$ such that*

$$u(x,t) \le 2Kx, \qquad 0 \le x \le 1,\ 0 \le t \le \tau, \tag{3.27}$$

and

$$u(x,\tau) \le 1 - \eta(1-x), \qquad 0 \le x \le 1.$$

Proof. Define

$$\overline{v}(x,t) = \frac{Kx}{1-2Kt}, \qquad 0 \le x \le 1,\ 0 \le t < 1/2K.$$

Since

$$\mathcal{P}\overline{v} = \frac{2K^2 x}{(1-2Kt)^2} - \frac{2K^2 x}{(1-2Kt)^2} = 0$$

and $\overline{v}(1,t) \ge K \ge 1$, the comparison principle guarantees that

$$u(x,t) \le \overline{v}(x,t) \le 2Kx, \qquad 0 \le x \le 1,\ 0 < t \le 1/4K.$$

Fix $\tau = 1/4K$. By Hopf's lemma, we have $u_x(1,\tau) > 0$. Since $u(\cdot,\tau)$ is nondecreasing in x, this implies $u(x,\tau) \le 1-\eta(1-x)$, $0 \le x \le 1$, for $\eta > 0$ sufficiently small, and the conclusion follows. ∎

LEMMA 3.4. *Let u_0 satisfy* (2.5) *and* (2.6)*, and denote by u the global solution of problem* (2.1)–(2.4)*. For any given $\delta \in (0,1)$, there exists $T_\delta > 0$ such that*

$$u(x,T_\delta) \ge \min(1-\delta, x/\delta), \qquad 0 \le x \le 1.$$

Proof. By Lemma 3.2, we may fix a small $\tau > 0$ such that $u_x(\cdot,\tau) \in C([0,1])$ and $u_x(0,\tau) > 0$. Therefore, $u(x,\tau) \ge \eta x$ for all $x \in [0,1]$ and some $\eta \in (0,1)$. Since $\mathcal{P}[\eta x] \le 0$ and $\eta < 1 = u(1,t)$, the comparison principle implies that

$$u(x,t) \ge \eta x, \qquad 0 \le x \le 1,\ t \ge \tau. \tag{3.28}$$

On the other hand, by (2.7), there exists $T > \tau$ such that

$$u(x,t) \ge 1-\delta, \qquad (1-\delta)\delta \le x \le 1,\ t \ge T. \tag{3.29}$$

Define

$$\underline{v}(x,t) = \big(\eta + 2\eta^2(t-T)\big)x, \qquad 0 \le x \le 1,\ t \ge T. \tag{3.30}$$

We have

$$\mathcal{P}\underline{v} = 2\eta^2 x - 2\big(\eta + 2\eta^2(t-T)\big)^2 x \le 0 \tag{3.31}$$

and, due to (3.28),

$$u(x,T) \ge \underline{v}(x,T), \qquad 0 \le x \le 1. \tag{3.32}$$

Since $\eta < 1 < 1/\delta$, we may find $T_\delta > T$ such that

$$\eta + 2\eta^2(T_\delta - T) = 1/\delta. \tag{3.33}$$

In view of (3.29), (3.30), and (3.33), we have

$$u\big((1-\delta)\delta, t\big) \ge 1-\delta \ge \underline{v}\big((1-\delta)\delta, t\big), \qquad T \le t \le T_\delta. \tag{3.34}$$

It then follows from (3.31), (3.32), (3.34), and the comparison principle that

$$u(x,T_\delta) \ge x/\delta, \qquad 0 \le x \le (1-\delta)\delta.$$

Using $\mathcal{P}[x/\delta] \le 0$ and (3.29), we deduce that

$$u(x,t) \ge x/\delta, \qquad 0 \le x \le (1-\delta)\delta,\ t \ge T_\delta.$$

This combined with (3.29) yields the desired conclusion. ∎

3.4. ODE lemmas. We consider the differential operator

$$\mathcal{L}w := yw'' + \frac{2yw'}{1+y} + \frac{2w}{(1+y)^2}.$$

First, the expression for the operator $\mathcal{L}$ reads

$$\mathcal{L}w = \left[\left(\frac{y}{y+1}\right)^2 \left(\frac{(y+1)^2}{y} w\right)'\right]'.$$

If $\psi \in C([0,\infty))$ satisfies $\psi(y) = O(y)$ as $y \to 0$, then the problem

$$\begin{aligned} \mathcal{L}w &= \psi, \quad y > 0, \\ w(0) &= 0,\ w'(0) = 0 \end{aligned}$$

admits a unique solution w, and w can be represented as

$$w(y) = \mathcal{L}_0^{-1}\psi := \frac{y}{(y+1)^2}\int_0^y \left(\frac{t+1}{t}\right)^2 \int_0^t \psi(s)\,ds\,dt$$

(note that the integral is convergent due to the assumption on ψ). In particular,

$$\psi \geq 0 \text{ on } [0,\infty) \implies \mathcal{L}_0^{-1}\psi \geq 0 \text{ on } [0,\infty). \tag{3.35}$$

On the other hand, $w_0(y) = \frac{y}{(y+1)^2}$ solves

$$\begin{aligned} \mathcal{L}w &= 0, \quad y > 0, \\ w(0) &= 0,\ w'(0) = 1. \end{aligned}$$

In the next section, to construct our main sub- and supersolutions, we will need to know the asymptotic behavior of the action of the operator $\mathcal{L}_0$ on some particular functions as $y \to \infty$. More precisely, let us define

$$\begin{aligned} f &= \left(I + \mathcal{L}_0^{-1}\right)\left(\frac{y}{(y+1)^2}\right), \\ \tilde{f} &= 2ff' - yf' + f, \\ g &= \mathcal{L}_0^{-1}\tilde{f}, \end{aligned}$$

and

$$h = \mathcal{L}_0^{-1}(\tilde{f} + M\varphi) = g + M\mathcal{L}_0^{-1}\varphi,$$

where $M > 0$ and

$$\varphi \in C^1([0,\infty)), \qquad \varphi(0) = 0, \qquad \varphi(y) = \frac{1}{\log y},\ y \geq 2.$$

We have the following lemmas.

LEMMA 3.5. *As $y \to \infty$, the function f satisfies*

$$\text{(i)} \qquad f(y) = \log y - 2 + O\left(\frac{\log^2 y}{y}\right);$$

$$\text{(ii)} \qquad f'(y) = \frac{1}{y} + O\left(\frac{\log^2 y}{y^2}\right).$$

LEMMA 3.6. *As $y \to \infty$, the functions g and h satisfy*

$$\text{(i)} \qquad g(y) = \frac{y\log y}{2} - \frac{9y}{4} + O\left(\log^3 y\right);$$

$$\text{(ii)} \qquad g'(y) = \frac{\log y}{2} - \frac{7}{4} + O\left(\frac{\log^3 y}{y}\right);$$

$$\text{(iii)} \qquad h(y) = \frac{y\log y}{2} - \frac{9y}{4} + O\left(\frac{y}{\log y}\right);$$

$$\text{(iv)} \qquad h'(y) = \frac{\log y}{2} - \frac{7}{4} + O\left(\frac{1}{\log y}\right).$$

Proof of Lemma 3.5.

(i) Using

$$\log(y+1) = \log y + O(1/y) \quad \text{as } y \to \infty, \tag{3.36}$$

we obtain

$$\begin{aligned} f(y) &= \frac{y}{(y+1)^2}\left[1 + \int_0^y \left(\frac{t+1}{t}\right)^2 \int_0^t \left(\frac{1}{s+1} - \frac{1}{(s+1)^2}\right) ds\, dt\right] \\ &= \frac{y}{(1+y)^2}\left[1 + \int_0^y \left\{\left(1 + \frac{2}{t} + \frac{1}{t^2}\right)\left(\log(t+1) - \frac{t}{t+1}\right)\right\} dt\right] \\ &= \frac{y}{(1+y)^2}\left[y\log y - 2y + O\big(\log^2 y\big)\right] \\ &= \log y - 2 + O\left(\frac{\log^2 y}{y}\right) \quad \text{as } y \to \infty. \end{aligned}$$

(ii) Now, using (i), we deduce

$$\begin{aligned} f'(y) &= \left[\frac{1}{y} - \frac{2}{y+1}\right] f(y) + \frac{1}{y}\left(\log(y+1) - \frac{y}{y+1}\right) \\ &= \left(-\frac{1}{y} + O\left(\frac{1}{y^2}\right)\right)\left(\log y - 2 + O\left(\frac{\log^2 y}{y}\right)\right) + \frac{\log(y+1)}{y} - \frac{1}{y+1} \\ &= \frac{1}{y} + O\left(\frac{\log^2 y}{y^2}\right) \quad \text{as } y \to \infty. \qquad \square \end{aligned}$$

To show Lemma 3.6, we first note that, due to Lemma 3.5 and (3.36), we have

$$\tilde{f}(y) = f_1(y) - 3f_2(y) + O(f_3(y)) \quad \text{as } y \to \infty,$$

where

$$f_1(y) = \log(y+1), \quad f_3(y) = \frac{\log^2(y+1)}{y+1},$$

and

$$f_2 \in C^1([0,\infty)), \qquad f_2(0) = 0, \qquad f_2(y) = 1,\ y \geq 1.$$

Denote $g_i = \mathcal{L}_0^{-1} f_i$ for $i = 1,2,3$ and $g_4 = \mathcal{L}_0^{-1}\varphi$. Lemma 3.6 will then be an immediate consequence of the following lemma.

LEMMA 3.7. *As $y \to \infty$, the functions g_i satisfy*

(i) $g_1(y) = \dfrac{y\log y}{2} - \dfrac{3y}{4} + O(\log y)$ *and* $g_1'(y) = \dfrac{\log y}{2} - \dfrac{1}{4} + O\left(\dfrac{\log y}{y}\right)$;

(ii) $g_2(y) = \dfrac{y}{2} + O(1)$ *and* $g_2'(y) = \dfrac{1}{2} + O\left(\dfrac{1}{y}\right)$;

(iii) $g_3(y) = O\left(\log^3 y\right)$ *and* $g_3'(y) = O\left(\dfrac{\log^3 y}{y}\right)$;

(iv) $g_4(y) = O\left(\dfrac{y}{\log y}\right)$ *and* $g_4'(y) = O\left(\dfrac{1}{\log y}\right)$.

Proof.

(i) As $y \to \infty$, we have

$$\begin{aligned} g_1(y) &= \frac{y}{(y+1)^2}\int_0^y \left(\frac{t+1}{t}\right)^2 \big((t+1)\log(t+1)-t\big)\,dt \\ &= \frac{y}{(1+y)^2}\int_0^y \left[(t+1)\log(t+1)-t+O\left(\log(t+1)\right)\right]dt \\ &= \left[\frac{1}{y}+O\left(\frac{1}{y^2}\right)\right]\left[\frac{y^2\log y}{2}-\frac{3y^2}{4}+O\left(y\log y\right)\right] \\ &= \frac{y\log y}{2}-\frac{3y}{4}+O\left(\log y\right); \end{aligned}$$

hence

$$\begin{aligned} g_1'(y) &= \left[\frac{1}{y}-\frac{2}{y+1}\right]g_1(y)+\frac{1}{y}\int_0^y \log(s+1)\,ds \\ &= \left(-\frac{1}{y}+O\left(\frac{1}{y^2}\right)\right)\left(\frac{y\log y}{2}-\frac{3y}{4}+O\left(\log y\right)\right)+\frac{(y+1)\log(y+1)-y}{y} \\ &= \frac{\log y}{2}-\frac{1}{4}+O\left(\frac{\log y}{y}\right). \end{aligned}$$

(ii) Due to the definition of g_2, there exist constants $C_1, C_2 \in \mathbb{R}$ such that, for $y > 1$,

$$\begin{aligned} g_2(y) &= \frac{y}{(y+1)^2}\left[\int_0^1 \left(\frac{t+1}{t}\right)^2 \int_0^t f_2(s)\,ds\,dt+\int_1^y \left(\frac{t+1}{t}\right)^2 (C_1+t)\,dt\right] \\ &= \frac{y}{(y+1)^2}\left[C_2+\frac{y^2}{2}+O(y)\right] = \frac{y}{2}+O(1) \quad \text{as } y\to\infty. \end{aligned}$$

Therefore, for $y > 1$,

$$\begin{aligned} g_2'(y) &= \left[\frac{1}{y}-\frac{2}{y+1}\right]g_2(y)+\frac{y}{(y+1)^2}\left(\frac{y+1}{y}\right)^2 (C_1+y) \\ &= \left[\frac{1}{y}-\frac{2}{y+1}\right]g_2(y)+1+O\left(\frac{1}{y}\right) = \frac{1}{2}+O\left(\frac{1}{y}\right) \quad \text{as } y\to\infty. \end{aligned}$$

(iii) As $y \to \infty$, we have

$$\begin{aligned} g_3(y) &= \frac{y}{3(y+1)^2}\int_0^y \left(\frac{t+1}{t}\right)^2 \log^3(t+1)\,dt \\ &= \frac{y}{3(1+y)^2}\int_0^y O\left(\log^3(t+1)\right)dt \\ &= \frac{y}{3(1+y)^2}\,O\left((y+1)\log^3(y+1)\right) = O\left(\log^3 y\right); \end{aligned}$$

hence

$$\begin{aligned} g_3'(y) &= \left[\frac{1}{y}-\frac{2}{y+1}\right]g_3(y)+\frac{1}{y}\int_0^y \frac{\log^2(t+1)}{(t+1)}\,dt \\ &= \left(-\frac{1}{y}+O\left(\frac{1}{y^2}\right)\right)O\left(\log^3 y\right)+\frac{\log^3(y+1)}{3y} = O\left(\frac{\log^3 y}{y}\right). \end{aligned}$$

(iv) Similarly to (ii), in this case we have, for $y > 2$,

$$\begin{aligned} g_4(y) &= \frac{y}{(y+1)^2}\left[\int_0^2 \left(\frac{t+1}{t}\right)^2 \int_0^t \varphi(s)\,ds\,dt + \int_2^y \left(\frac{t+1}{t}\right)^2 \left(C_1 + \int_2^t \frac{ds}{\log s}\right) dt\right] \\ &= \frac{y}{(y+1)^2}\left[C_2 + \int_2^y \left(\frac{t+1}{t}\right)^2 O\left(\frac{t}{\log t}\right) dt\right] \\ &= \frac{y}{(y+1)^2} O\left(\frac{y^2}{\log y}\right) = O\left(\frac{y}{\log y}\right) \quad \text{as } y \to \infty. \end{aligned}$$

Therefore, for $y > 2$,

$$\begin{aligned} g_4'(y) &= \left[\frac{1}{y} - \frac{2}{y+1}\right] g_4(y) + \frac{y}{(y+1)^2}\left(\frac{y+1}{y}\right)^2 \left(C_1 + \int_2^y \frac{ds}{\log s}\right) \\ &= O\left(\frac{1}{\log y}\right) \quad \text{as } y \to \infty. \qquad \square \end{aligned}$$

4. Proof of the main results.

4.1. Construction of a subsolution. Motivated by the idea of an asymptotic expansion around (moving) steady states and based on a self-similar variable (see subsection 1.3), we make the following ansatz:

$$\underline{u}(x,t) = 1 - \frac{1}{y+1} + b(t)f(y) - b^2(t)g(y), \qquad y = a(t)x,$$

where the functions a, b, f, g have to be determined. Note that the variable y now ranges into the time-dependent interval $[0, a(t)]$. Here, a and b are expected to satisfy

$$a(t) \sim u_x(0,t), \quad \lim_{t\to\infty} a(t) = \infty, \quad \text{and} \quad \lim_{t\to\infty} b(t) = 0.$$

LEMMA 4.1. *The problem*

$$\begin{aligned} &\underline{u}_t - x\underline{u}_{xx} \le 2\,\underline{u}\,\underline{u}_x, \quad 0 < x < 1,\ t > 0, \\ &\underline{u}(0,t) = 0, \quad t \ge 0, \\ &\underline{u}(1,t) < 1, \quad t \ge 0, \end{aligned}$$

admits a solution of the form

$$\underline{u}(x,t) = 1 - \frac{1}{y+1} + b(t)f(y) - b^2(t)g(y), \qquad y = a(t)x, \tag{4.1}$$

where the smooth functions $a > 0$, $b > 0$, $f \ge 0$, and g have the following properties:

$$f(y) \sim \log y, \quad g(y) \sim \frac{y \log y}{2} \quad \text{as } y \to \infty,$$

$$f(0) = g(0) = f'(0) = g'(0) = 0, \tag{4.2}$$

$$a(t) = \left(1 + O\big(t^{-1/2}\log t\big)\right) \exp\left[\frac{5}{2} + \sqrt{2t}\right] \quad \text{as } t \to \infty, \tag{4.3}$$

$$b(t) = \frac{1 + O\big(t^{-1/2}\big)}{a(t)\log a(t)} \quad \text{as } t \to \infty, \tag{4.4}$$

and, moreover,

$$\underline{u}_x > 0, \quad 0 \le x \le 1,\ t \ge 0. \tag{4.5}$$

Proof. Step 1. *Construction of the subsolution.* In what follows we shall omit the variables t and/or y when no confusion is likely. We take $\underline{u}$ as in (4.1) where we assume

$$a > 0 \quad \text{and} \quad \lim_{t\to\infty} a(t) = \infty. \tag{4.6}$$

We compute

$$\underline{u}_x = \frac{a}{(1+y)^2} + baf'(y) - b^2 a g'(y), \tag{4.7}$$

$$\underline{u}_{xx} = \frac{-2a^2}{(1+y)^3} + ba^2 f''(y) - b^2 a^2 g''(y),$$

and

$$\underline{u}_t = \frac{a'x}{(1+y)^2} + \left[b'f(y) + ba'xf'(y) - 2bb'g(y) - b^2a'xg'(y)\right];$$

hence

$$\underline{u}_t = \frac{a'}{a}\frac{y}{(1+y)^2} + \left[b'f(y) + \frac{ba'}{a}yf'(y) - 2bb'g(y) - \frac{b^2a'}{a}yg'(y)\right].$$

Recall that the operator $\mathcal{P}$ is defined in (3.26). It follows that

$$\begin{aligned}\mathcal{P}\underline{u} = {} & \frac{a'}{a}\frac{y}{(1+y)^2} + \left[b'f + \frac{ba'}{a}yf' - 2bb'g - \frac{b^2a'}{a}yg'\right] \\ & + 2a\,\frac{y}{(1+y)^3} - ba\,yf'' + b^2a\,yg'' \\ & - 2a\left[\frac{y}{1+y} + bf - b^2g\right]\left[\frac{1}{(1+y)^2} + bf' - b^2g'\right].\end{aligned}$$

Collecting terms of the same order in b yields

$$\begin{aligned}\mathcal{P}\underline{u} = {} & \frac{a'}{a}\frac{y}{(1+y)^2} + \left[b'f + \frac{ba'}{a}yf' - 2bb'g - \frac{b^2a'}{a}yg'\right] - ab\left[yf'' + \frac{2f}{(1+y)^2} + \frac{2yf'}{1+y}\right] \\ & + ab^2\left[yg'' + \frac{2g}{(1+y)^2} + \frac{2yg'}{1+y} - 2ff'\right] + 2ab^3\left[f'g + fg'\right] - 2ab^4gg'.\end{aligned} \tag{4.8}$$

The natural scaling of the equation leads to the choice

$$b := \frac{a'}{a^2}, \tag{4.9}$$

so that in the right-hand side (RHS) of (4.8), the first term will be of the same order as the terms in the second bracket. Assuming

$$a' > 0, \tag{4.10}$$

we also denote

$$\gamma := \left(\frac{a}{a'}\right)' \tag{4.11}$$

and observe that

$$\frac{a'}{a} = ab, \qquad b' = -(1+\gamma)ab^2, \tag{4.12}$$

where the last equality comes from

$$\gamma = \left(\frac{a}{a'}\right)' = \left(\frac{1}{ab}\right)' = -\frac{b'}{ab^2} - \frac{a'}{a^2 b} = -\frac{b'}{ab^2} - 1.$$

To make things clear, let us already stress that the final choice of a will guarantee

$$\gamma \geq 0 \quad \text{and} \quad \lim_{t\to\infty} \gamma(t) = 0.$$

Using (4.12), we can recast identity (4.8) in the form

$$\begin{aligned}\mathcal{P}\underline{u} &= ab\left[\frac{y}{(1+y)^2} - yf'' - \frac{2f}{(1+y)^2} - \frac{2yf'}{1+y}\right] \\ &\quad + ab^2\left[yg'' + \frac{2g}{(1+y)^2} + \frac{2yg'}{1+y} - 2ff' + yf' - (1+\gamma)f\right] \\ &\quad + ab^3\left[2f'g + 2fg' - yg' + 2(1+\gamma)g\right] - 2ab^4 gg';\end{aligned}$$

hence

$$\begin{aligned}\mathcal{P}\underline{u} &= ab\left[\frac{y}{(1+y)^2} - \mathcal{L}f\right] + ab^2\left[\mathcal{L}g - 2ff' + yf' - (1+\gamma)f\right] \\ &\quad + ab^3\left[2f'g + 2fg' - yg' + 2(1+\gamma)g\right] - 2ab^4 gg'.\end{aligned} \tag{4.13}$$

Let us now choose

$$f := (I + \mathcal{L}_0^{-1})\left(\frac{y}{(1+y)^2}\right) \geq \frac{y}{(1+y)^2} \geq 0, \tag{4.14}$$

which solves $\mathcal{L}f = y/(1+y)^2$ for $y > 0$, with $f(0) = 0$ and $f'(0) = 1$ (cf. subsection 3.4). Dividing by ab^2, we see that $\mathcal{P}\underline{u}$ has the same sign as the quantity

$$A := \left[\mathcal{L}g - 2ff' + yf' - (1+\gamma)f\right] + b\left[2f'g + 2fg' - yg' + 2(1+\gamma)g\right] - 2b^2 gg'.$$

Next we choose

$$g := \mathcal{L}_0^{-1}\big(2ff' - yf' + f\big).$$

Therefore,

$$A = -\gamma f + b\left[2f'g + 2fg' - yg' + 2(1+\gamma)g\right] - 2b^2 gg'.$$

We now proceed to show that the quantity A is nonpositive for large t by considering separately the regions $y_0 \le y \le a(t)$ and $0 \le y \le y_0$ for some large y_0 independent of t. At this point, we make the additional assumptions that

$$\gamma \sim \frac{1}{\log a} \quad \text{as } t \to \infty \tag{4.15}$$

and

$$a' \sim \frac{a}{\log a}; \quad \text{hence} \quad b \sim \frac{1}{a \log a} \quad \text{as } t \to \infty \tag{4.16}$$

(which will be verified on the final choice of the function a; actually more precise expansions of γ and b will be needed in the final matching process). By Lemmas 3.5 and 3.6, we have

$$f \sim \log y, \quad f'g = o(g), \quad fg' = o(g), \quad \text{and} \quad yg' \sim g \sim \frac{y \log y}{2} \quad \text{as } y \to \infty.$$

Consequently, fixing $\delta > 0$ and taking y_0 and t_0 large enough, we have, for $y_0 \le y \le a(t)$ and $t \ge t_0$,

$$-\gamma f \le (-1+\delta)\frac{\log y}{\log a}, \qquad g, g' \ge 0,$$

and

$$\begin{aligned} 2f'g + 2fg' - yg' + 2(1+\gamma)g &\le \delta\, y \log y - \frac{1}{2} y \log y + 2(1+\gamma)\left(\frac{1}{2}+\delta\right) y \log y \\ &\le \left(\frac{1}{2} + 4\delta\right) y \log y. \end{aligned}$$

Taking $\delta = 1/12$ and also using (4.16), we thus obtain

$$A \le (-1+\delta)\frac{\log y}{\log a} + \frac{1}{a \log a}\left(\frac{1}{2}+5\delta\right) y \log y = \frac{\log y}{\log a}\left(-1+\delta+\left(\frac{1}{2}+5\delta\right)\frac{y}{a}\right) \le 0$$

for $y_0 \le y \le a(t)$ and $t \ge t_0$ (possibly larger). Next, for $0 \le y \le y_0$, (4.14) implies

$$f(y) \ge c_1 y, \quad 0 \le y \le y_0,$$

whereas $f(0) = g(0) = 0$ yields

$$|2f'g + 2fg' - yg'| + 4|g| + |gg'| \le c_2 y, \quad 0 \le y \le y_0,$$

with $c_1, c_2 > 0$. Consequently,

$$A \le -c_1 \gamma(t) y + c_2 b(t) y = \left[-c_1 + c_2 \frac{b(t)}{\gamma(t)}\right] \gamma(t) y \le 0$$

on $[0, y_0]$ for t large enough, due to (4.15) and (4.16).

We have thus proved that, under conditions (4.6), (4.10), (4.15), and (4.16), there holds $\mathcal{P}\underline{u} \le 0$ in $(0,1) \times (T, \infty)$ for T large enough. By a time shift we may obviously take $T = 0$.

Step 2. *Determination of $a(t)$ by matching at the outer boundary.* Next, the determination of $a(t)$ will be done by "matching" with the boundary condition at $x = 1$, i.e., by writing

$$\underline{u}(1,t) < 1,$$

which is equivalent to

$$bf(a) - b^2 g(a) < \frac{1}{a+1}. \tag{4.17}$$

It is of course sufficient to check (4.17) for large t (thanks to the possibility of shifting time). Let us first sketch the resolution of (4.17) in a rough way. Using (4.9) and applying Lemmas 3.5(i) and 3.6(i) at leading order, we are left with

$$\frac{a'}{a^2}\left(\log a - \frac{a'}{a^2}\frac{a\log a}{2}\right) \lesssim \frac{1}{a+1} \quad \text{as } t \to \infty. \tag{4.18}$$

We expect the second term in the bracket of the left-hand side (LHS) of the preceding relation to be much smaller than the first one as $a \to \infty$. If we ignore it, we obtain the differential inequality

$$a' \lesssim \frac{a}{\log a} \quad \text{as } t \to \infty, \tag{4.19}$$

which implies that

$$a(t) \lesssim e^{\sqrt{2t}} \quad \text{as } t \to \infty.$$

However, the latter estimation is not accurate enough to show the desired estimate, and we thus add a correction term in (4.19). More precisely, we look for $a(t)$ as the solution of

$$a' = \frac{a}{\log a}(1+\eta), \quad t > 0, \qquad a(0) = 2, \tag{4.20}$$

where the correction term $\eta = \eta(a)$ has the form

$$\eta = \frac{5}{2\log a} + \frac{K}{\log^2 a}, \quad K > 0. \tag{4.21}$$

Now plugging (4.20) into (4.17), recalling also (4.9), and using the exact asymptotic behavior of $f(a), g(a)$ as $a \to \infty$, provided by Lemmas 3.5(i) and 3.6(i), we are reduced to the condition

$$\frac{1+\eta}{a\log a}\left[\log a - 2 + O\left(\frac{\log^2 a}{a}\right) - \frac{1+\eta}{a\log a}\left(\frac{a\log a}{2} - \frac{9a}{4} + O\left(\log^3 a\right)\right)\right] < \frac{1}{a+1}$$

as $t \to \infty$ or

$$(1+\eta)\left[1 - \frac{5+\eta}{2\log a} + \frac{9}{4\log^2 a} + \frac{9\eta}{4\log^2 a} + O\left(\frac{\log a}{a}\right)\right] < \frac{a}{a+1} \quad \text{as } t \to \infty. \tag{4.22}$$

Plugging (4.21) into (4.22), we first obtain

$$\left(1 + \frac{5}{2\log a} + \frac{K}{\log^2 a}\right)\left[1 - \frac{5}{2\log a} + \frac{1}{\log^2 a} + O\left(\frac{1}{\log^3 a}\right)\right] < 1 - \frac{1}{a+1} \quad \text{as } t \to \infty$$

and finally

$$1+\frac{4K-21}{4\log^2 a}+O\left(\frac{1}{\log^3 a}\right)<1-\frac{1}{a+1}\quad \text{as } t\to\infty. \tag{4.23}$$

In order for (4.23) to be satisfied, we choose $K<\frac{21}{4}$, and we then obtain the following ODE for a:

$$a'=\frac{a}{\log a}\left(1+\frac{5}{2\log a}+\frac{K}{\log^2 a}\right),\quad t>0,\qquad a(0)=2. \tag{4.24}$$

Here we should mention that the same form for η, given by (4.21), will also be considered for the supersolution, with a different constant K; see the next subsection. Equation (4.24) implies that

$$a'=\frac{a}{\log a-5/2}\left(1+O\big(\log^{-2} a\big)\right)\quad \text{as } t\to\infty,$$

and, integrating with respect to t, we get

$$\log^2 a-5\log a=2t+O\left(\int_0^t\frac{ds}{\log^2(a(s))}\right)\quad \text{as } t\to\infty.$$

Solving the quadratic polynomial in $\log a$ and noting that $\log a(s)\ge\sqrt{2s}$ by (4.24), we end up with (4.3). As for (4.4), it follows from

$$\begin{aligned} b=\frac{a'}{a^2}&=\frac{1}{a\log a}\left(1+\frac{5}{2\log a}+\frac{K}{\log^2 a}\right)=\frac{1+O\left(\log^{-1} a\right)}{a\log a}\\ &=\frac{1+O\left(t^{-1/2}\right)}{a\log a}\quad \text{as } t\to\infty, \end{aligned} \tag{4.25}$$

where we used (4.24) and $\log a\ge\sqrt{2t}$ as $t\to\infty$. On the other hand, denoting $G(s)=s+(5/2)s^2+Ks^3$ and using (4.24), we see that $\gamma=(a/a')'$ satisfies

$$\gamma=\left[\frac{1}{G\big(1/\log a\big)}\right]'=\frac{a'}{a\log^2 a}\frac{G'\big(1/\log a\big)}{G^2\big(1/\log a\big)}=\frac{G'\big(1/\log a\big)}{\log^2 a\,G\big(1/\log a\big)}=H\big(1/\log a\big), \tag{4.26}$$

where $H(s)=s(1+5s+3Ks^2)(1+(5/2)s+Ks^2)^{-1}$. Finally, the assumed properties (4.6), (4.10), (4.15), and (4.16) of a,b,γ are immediate consequences of (4.24), (4.25), and (4.26).

Step 3. *Proof of* (4.5). By (4.7) we have

$$a^{-1}\underline{u}_x=\frac{1}{(1+y)^2}+bf'(y)-b^2g'(y).$$

By Lemma 3.6 and (4.16), taking y_1 and t_1 large enough, we have, for $t\ge t_1$ and $y_1\le y\le a(t)$,

$$a^{-1}\underline{u}_x\ge\frac{1}{2a^2}-\frac{\log y}{a^2\log^2 a}\ge\frac{1}{a^2}\left(\frac{1}{2}-\frac{1}{\log a}\right)>0.$$

Now, for $0\le y\le y_1$ and $t\ge t_1$ possibly larger, we get

$$a^{-1}\underline{u}_x\ge\frac{1}{(1+y_1)^2}-Cb(t)-Cb^2(t)>0.$$

By a time shift we may obviously take $t_1 = 0$, and (4.5) is proved. ∎

Remark 4.1. It is still possible to obtain a qualitatively correct subsolution with just a two-term expansion, for instance, by making the simple choice $f(y) = \log(1+y)$, $g = 0$ in (4.1). However, this yields only the lower grow-up rate up to a multiplicative constant, i.e., $u_x(0,t) \geq Ce^{\sqrt{2t}}$, and does not enable one to deduce an expansion of the form (2.8).

4.2. Construction of a supersolution. The form (4.1) does not seem to be sufficient to construct an accurate supersolution (i.e., leading to a function $a(t)$ fulfilling (4.3)). We need a slight perturbation, corresponding to the modified ansatz

$$\overline{u}(x,t) = 1 - \frac{1}{y+1} + b(t)f(y) - b^2(t)\tilde{g}(y,t), \qquad y = a(t)x, \tag{4.27}$$

where

$$\tilde{g}(y,t) = (1+\varepsilon(t))h(y), \tag{4.28}$$

and $\varepsilon(t)$ goes to 0 as $t \to \infty$.

LEMMA 4.2. *The problem*

$$\begin{aligned} &\overline{u}_t - x\overline{u}_{xx} \geq 2\,\overline{u}\,\overline{u}_x, \quad 0 < x < 1,\ t > 0,\\ &\overline{u}(0,t) = 0, \quad t \geq 0,\\ &\overline{u}(1,t) \geq 1, \quad t \geq 0, \end{aligned}$$

admits a solution of the form

$$\overline{u}(x,t) = 1 - \frac{1}{y+1} + b(t)f(y) - (1+\varepsilon(t))b^2(t)h(y), \qquad y = a(t)x,$$

where the smooth functions $a(t), b(t), f(y), h(y)$ have the following properties:

$$f(y) \sim \log y, \quad h(y) \sim \frac{y\log y}{2} \quad \text{as } y \to \infty,$$

$$f(0) = h(0) = f'(0) = h'(0) = 0,$$

$$a(t) = \Big(1 + O\big(t^{-1/2}\log t\big)\Big)\exp\left[\frac{5}{2} + \sqrt{2t}\right] \quad \text{as } t \to \infty, \tag{4.29}$$

$$b(t) = \frac{1 + O\big(t^{-1/2}\big)}{a(t)\log a(t)} \quad \text{as } t \to \infty, \tag{4.30}$$

$$\varepsilon(t) \sim (2t)^{-1/2} \quad \text{as } t \to \infty. \tag{4.31}$$

Proof. Step 1. *Construction of the supersolution.* Taking $\overline{u}$ as defined in (4.27), (4.28), the expression for $\mathcal{P}\overline{u}$ is similar to (4.13), except that g, g', g'' are now replaced with $\tilde{g}, \tilde{g}_y, \tilde{g}_{yy}$ and an additional term $-b^2\varepsilon' h$ is added, which is inherited from $\overline{u}_t$. As in the proof of Lemma 4.1, b and γ are defined through (4.9) and (4.11), and we assume (4.6), (4.10), (4.15), and (4.16). This leads to

$$\begin{aligned} \mathcal{P}\overline{u} = {}& ab\left[\frac{y}{(1+y)^2} - \mathcal{L}f\right] + ab^2\left[\mathcal{L}\tilde{g} - 2ff' + yf' - (1+\gamma)f - \frac{\varepsilon'}{a}h\right]\\ &+ ab^3\left[2f'\tilde{g} + 2f\tilde{g}_y - y\tilde{g}_y + 2(1+\gamma)\tilde{g}\right] - 2ab^4\tilde{g}\tilde{g}_y. \end{aligned}$$

Replacing $\tilde{g}(y,t)$ with $(1+\varepsilon(t))h(y)$, we obtain

$$\mathcal{P}\overline{u} = ab\left[\frac{y}{(1+y)^2} - \mathcal{L}f\right] + ab^2\left[(1+\varepsilon)\mathcal{L}h - 2ff' + yf' - (1+\gamma)f - \frac{\varepsilon'}{a}h\right] + (1+\varepsilon)ab^3\left[2f'h + 2fh' - yh' + 2(1+\gamma)h\right] - 2(1+\varepsilon)^2ab^4hh'. \tag{4.32}$$

Denote

$$B_0 := (1+\varepsilon)\mathcal{L}h - 2ff' + yf' - (1+\gamma)f - \frac{\varepsilon'}{a}h.$$

As in Lemma 4.1, we first choose

$$f := (I + \mathcal{L}_0^{-1})\left(\frac{y}{(1+y)^2}\right).$$

Dividing identity (4.32) by ab^2, we see that $\mathcal{P}\overline{u}$ has the same sign as the quantity

$$B := B_0 + (1+\varepsilon)b\left[2f'h + 2fh' - yh' + 2(1+\gamma)h\right] - 2(1+\varepsilon)^2b^2hh'.$$

Next we choose

$$h = \mathcal{L}_0^{-1}\big(2ff' - yf' + f + M\varphi\big),$$

where φ satisfies

$$\varphi(0) = 0, \qquad \varphi'(0) > 0, \qquad \varphi(y) > 0 \ \text{ for } 0 < y < 2, \qquad \varphi(y) = 1/\log y \ \text{ for } y \geq 2. \tag{4.33}$$

Note that, taking $M > 2$ suitably large, we have

$$2ff' - yf' + f + M\varphi \geq 0, \quad y \geq 0,$$

by Lemma 3.5; hence

$$h(y) \geq 0, \quad y \geq 0, \tag{4.34}$$

due to (3.35). We compute

$$\begin{aligned} B_0 &= (1+\varepsilon)\big(2ff' - yf' + f + M\varphi\big) - 2ff' + yf' - (1+\gamma)f - \frac{\varepsilon'}{a}h \\ &= \varepsilon\big(2ff' - yf'\big) + M(1+\varepsilon)\varphi + (\varepsilon - \gamma)f - \frac{\varepsilon'}{a}h. \end{aligned}$$

At this point, we choose

$$\varepsilon = \gamma, \tag{4.35}$$

where γ is defined by (4.11), and we assume again (4.15) and (4.16) along with

$$\gamma' \leq 0 \tag{4.36}$$

(these assumptions will be verified on the final choice of the function a). By (4.34) and (4.36) we have

$$B_0 \geq \gamma\big(2ff' - yf'\big) + M(1+\gamma)\varphi,$$

and it follows that

(4.37)
$$(1+\gamma)^{-1}B \ge \left[\frac{\gamma}{1+\gamma}\big(2ff'-yf'\big)+M\varphi\right]+b\Big[2f'h+2fh'-yh'+2(1+\gamma)h\Big]-2(1+\gamma)b^2hh'.$$

To show that the RHS of (4.37) is nonnegative for large t, we again consider separately the regions $y_0 \le y \le a(t)$ and $0 \le y \le y_0$ for some large y_0 independent of t. By the estimates in Lemma 3.5, we have

$$f \sim \log y, \quad f' \sim \frac{1}{y}, \quad \text{and} \quad yh' \sim h \sim \frac{y\log y}{2} \quad \text{as } y \to \infty. \tag{4.38}$$

Consequently, fixing $\delta > 0$, using (4.15), and taking y_0 and t_0 large enough, we have, for $y_0 \le y \le a(t)$ and $t \ge t_0$,

$$\frac{\gamma}{1+\gamma}\big(2ff'-yf'\big)+M\varphi \ge -\frac{3}{2}\gamma+\frac{M}{\log y} \ge -\frac{2}{\log a}+\frac{M}{\log y} \ge \frac{M-2}{\log y},$$

$$2f'h+2fh'-yh'+2(1+\gamma)h \ge -\left(\frac{1}{2}+\delta\right)y\log y+2(1+\gamma)\left(\frac{1}{2}-\delta\right)y\log y \ge \left(\frac{1}{2}-4\delta\right)y\log y,$$

and

$$2(1+\gamma)hh' \le y\log^2 y.$$

Assuming $M \ge 3$, taking $\delta = 1/8$, and also using (4.16), we infer that

$$(1+\gamma)^{-1}B \ge \frac{1}{\log a}-\frac{y\log^2 y}{a^2\log^2 a}=\frac{1}{\log a}\left(1-\frac{y\log^2 y}{a^2\log a}\right) \ge \frac{1}{\log a}\left(1-\frac{\log a}{a}\right) \ge 0$$

for $y_0 \le y \le a(t)$ and $t \ge t_0$. Next, for $0 \le y \le y_0$, (4.33) implies

$$M\varphi(y) \ge c_1 y,$$

whereas $f(0) = h(0) = 0$ yields

$$2ff'-yf' \ge -c_2 y, \quad 2f'h+2fh'-yh'+2(1+\gamma)h \ge -c_2 y, \quad 2(1+\gamma)hh' \le c_2 y,$$

with $c_1, c_2 > 0$. Therefore,

$$(1+\gamma)^{-1}B \ge -\gamma(t)c_2 y+c_1 y-b(t)c_2 y-b^2(t)c_2 y=\Big[c_1-c_2\big(\gamma(t)+b(t)+b^2(t)\big)\Big]\,y \ge 0$$

on $[0, y_0]$ for t large enough.

We have thus proved that $\mathcal{P}\overline{u} \ge 0$ in $(0,1)\times(T,\infty)$ for T large enough. By a time shift we may obviously take $T = 0$.

Step 2. *Determination of $a(t)$ by matching at the outer boundary.* The determination of $a(t)$ will be done again by "matching" with the boundary condition at $x = 1$. Imposing that

$$\underline{u}(1,t) \ge 1,$$

we obtain

$$bf(a)-b^2(1+\varepsilon)h(a) \ge \frac{1}{a+1},$$

and taking (4.9) into account, we end up with

$$\frac{a'}{a^2}\left(f(a)-\frac{a'}{a^2}(1+\varepsilon)h(a)\right)\geq\frac{1}{a+1}. \tag{4.39}$$

Again it suffices to check (4.39) for large t. Following the same reasoning as in the case of a subsolution, we again look for $a(t)$ as a solution of

$$a'=\frac{a}{\log a}(1+\eta)\quad\text{as } t\to\infty, \tag{4.40}$$

where the correction term $\eta=\eta(a)$ is given by (4.21) with K a constant to be determined.

Plugging (4.40) into (4.39) and using the exact asymptotic behavior of $f(a),h(a)$ as $a\to\infty$, given by Lemmas 3.5 and 3.6, we arrive at

$$\frac{1+\eta}{a\log a}\left[\log a-2+O\left(\frac{\log^2 a}{a}\right)-\frac{(1+\eta)(1+\varepsilon)}{a\log a}\left(\frac{a\log a}{2}-\frac{9a}{4}+O\left(\frac{a}{\log a}\right)\right)\right]\geq\frac{1}{a+1}$$

or

$$(1+\eta)\left[1-\frac{2}{\log a}+O\left(\frac{\log a}{a}\right)-\frac{(1+\eta)(1+\varepsilon)}{\log a}\left(\frac{1}{2}-\frac{9}{4\log a}+O\left(\frac{1}{\log^2 a}\right)\right)\right]\geq\frac{a}{a+1} \tag{4.41}$$

as $t\to\infty$. Denote by Γ the quantity in the bracket in the LHS of (4.41). Using (4.21) and (4.41), we obtain

$$\begin{aligned}\Gamma&=1-\frac{2}{\log a}+O\left(\frac{\log a}{a}\right)\\&\quad-\left(1+\frac{5}{2\log a}+\frac{K}{\log^2 a}\right)\left(1+\frac{1}{\log a}+\frac{5}{2\log^2 a}\right)\left(\frac{1}{2\log a}-\frac{9}{4\log^2 a}+O\left(\frac{1}{\log^3 a}\right)\right)\\&=1-\frac{5}{2\log a}+\frac{1}{2\log^2 a}+O\left(\frac{1}{\log^3 a}\right)\quad\text{as } t\to\infty.\end{aligned}$$

Then (4.41) becomes equivalent to

$$\left(1+\frac{5}{2\log a}+\frac{K}{\log^2 a}\right)\left(1-\frac{5}{2\log a}+\frac{1}{2\log^2 a}+O\left(\frac{1}{\log^3 a}\right)\right)\geq 1-\frac{1}{a+1}\quad\text{as } t\to\infty;$$

that is,

$$1+\frac{4K-23}{4\log^2 a}+O\left(\frac{1}{\log^3 a}\right)\geq 1-\frac{1}{a+1}\quad\text{as } t\to\infty. \tag{4.42}$$

For (4.42) to be satisfied we choose $K>\frac{23}{4}$, and we again take a to be the solution of the ODE (4.24). Then, by the end of step 2 of the proof of Lemma 4.1, we obtain (4.29) and (4.30) as well as the assumed properties (4.6), (4.10), (4.15), and (4.16) of a,b,γ. Finally, we note that (4.15), (4.29), and (4.35) guarantee (4.31), and that (4.26) implies

$$\gamma'=\frac{-a'}{a\log^2 a}H'\big(1/\log a\big)=\frac{-1}{\log^2 a}(H'G)\big(1/\log a\big)\sim\frac{-1}{\log^3 a}\quad\text{as } t\to\infty$$

and hence (4.36) (after a further time shift). ∎

4.3. Proofs of Theorem 2.1 and Corollary 2.2. Let u be the solution of (2.1)–(2.4) and let $\underline{u}, \overline{u}$ be the sub-/supersolutions provided by Lemmas 4.1 and 4.2. The asymptotic expansion (2.8)–(2.9) in Theorem 2.1 will be an immediate consequence of the following two Lemmas. The first one guarantees that u lies between suitable time-shifts of $\underline{u}$ and $\overline{u}$. The second one shows that (shifted versions of) $\underline{u}$ and $\overline{u}$ satisfy the required asymptotic behavior.

LEMMA 4.3.

(i) *There exists $T_1 > 0$ such that*

$$u(\cdot,t) \ge \underline{u}(\cdot,t-T_1), \qquad t \ge T_1. \tag{4.43}$$

(ii) *Let τ be as in Lemma* 3.3. *Then there exists $T_2 > 0$ such that*

$$u(\cdot,t) \le \overline{u}(\cdot,t+T_2), \qquad t \ge \tau.$$

Proof. (i) Since $\underline{u}(\cdot,0) \in C^1([0,1])$ with $\underline{u}(0,0) = 0$, $\underline{u}(1,0) < 1$, and $\underline{u}_x(\cdot,0) > 0$, it follows from Lemma 3.4 that $u(\cdot,T_1) \ge \underline{u}(\cdot,0)$ for some $T_1 > 0$. The assertion then follows from the comparison principle.

(ii) Due to (4.38), (4.14), and $h(0) = 0$, we have

$$f(y) \ge c_1 \log(y+1) \quad \text{and} \quad h(y) \le c_2\,(y+1)\log(y+1), \qquad y \ge 0,$$

for some $c_1, c_2 > 0$. This along with (4.29)–(4.31) implies that, for all $t \ge t_2$ large enough and all $x \in [0,1]$,

$$\begin{aligned}
\overline{u}(x,t) &= \frac{ax}{ax+1} + bf(ax) - (1+\varepsilon(t))b^2 h(ax) \\
&\ge \frac{ax}{ax+1} + b\log(ax+1)\big(c_1 - 2c_2(ax+1)b\big) \\
&\ge \frac{ax}{ax+1} + b\log(ax+1)\left(c_1 - 3c_2\frac{a+1}{a\log a}\right) \ge \frac{ax}{ax+1}.
\end{aligned}$$

Take τ, η as in Lemma 3.3 and set $x_0 = 1/4K$. For all $x \in [0,x_0]$ and $t \ge t_2$, with $t_2 \ge \tau$ possibly larger, we have $a/2K \ge ax+1$; hence

$$\overline{u}(x,t) \ge \frac{ax}{ax+1} \ge 2Kx \ge u(x,\tau), \qquad 0 \le x \le x_0,\ t \ge t_2.$$

On the other hand, (4.38) implies that

$$f'(y) \le c_3 \quad \text{and} \quad |h'(y)| \le c_3 + \log(y+1), \qquad y \ge 0,$$

for some $c_3 > 0$. For all $x \in (x_0,1]$ and $t \ge t_2$ (possibly larger), we thus have

$$\begin{aligned}
\overline{u}_x &= \frac{a}{(1+ax)^2} + baf'(ax) - (1+\varepsilon(t))b^2 a h'(ax) \\
&\le \frac{a}{(1+ax_0)^2} + c_3 ba + 2(c_3 + \log(a+1))b^2 a \le \eta,
\end{aligned}$$

due to (4.29) and (4.30); hence

$$\overline{u}(x,t) \ge 1 - \eta(1-x) \ge u(x,\tau), \qquad x_0 < x \le 1,\ t \ge t_2.$$

Therefore, $\overline{u}(\cdot,t_2) \ge u(\cdot,\tau)$ in $[0,1]$. The assertion, with $T_2 = t_2 - \tau$, thus follows from the comparison principle. ∎

LEMMA 4.4. *Let*

$$A(t) = \exp\left[\frac{5}{2} + \sqrt{2t}\right].$$

Then, for any $T \in \mathbb{R}$, each of the functions $w = \underline{u}$ and $w = \overline{u}$ satisfies

$$1 - w(x, t+T) = \frac{1}{1 + A(t)x}\left[1 - x + O\big(t^{-1/2}\log t\big)\right]$$

as $t \to \infty$, uniformly in $[0,1]$.

Proof. We shall give the proof for $w = \underline{u}$, the other case being completely similar. Set $\tilde{a}(t) = a(t+T)$ and $\tilde{b}(t) = b(t+T)$. We first note that, by (4.3), we have

$$\tilde{a}(t) = \Big(1 + O\big(t^{-1/2}\log t\big)\Big)A(t)$$

and

$$\log \tilde{a}(t) = \log A(t) + O\big(t^{-1/2}\log t\big) = \Big(1 + O\big(t^{-1}\log t\big)\Big)\log A(t). \tag{4.44}$$

Also, by (4.4), we have

$$\tilde{b}(t) = \frac{1 + O\big(t^{-1/2}\big)}{\tilde{a}\log\tilde{a}}. \tag{4.45}$$

Moreover, due to Lemma 3.5, there exists $C > 0$ such that

$$|f(y) - \log(1+y)| \le C, \quad y \ge 0. \tag{4.46}$$

Now, using (4.45), (4.46), and (4.44), we compute

$$\begin{aligned}
\frac{1}{1+\tilde{a}x} - \tilde{b}f(\tilde{a}x) &= \frac{1}{1+\tilde{a}x}\left[1 - \frac{(1+\tilde{a}x)f(\tilde{a}x)}{\tilde{a}\log\tilde{a}}\Big(1 + O\big(t^{-1/2}\big)\Big)\right] \\
&= \frac{1}{1+\tilde{a}x}\left[1 - \frac{(1+\tilde{a}x)\log(1+\tilde{a}x)}{\tilde{a}\log\tilde{a}}\Big(1 + O\big(t^{-1/2}\big)\Big)\right] \\
&= \frac{1}{1+\tilde{a}x}\left[1 - \frac{x\log(1+\tilde{a}x)}{\log\tilde{a}} + O\big(t^{-1/2}\big)\right] \\
&= \frac{1}{1+\tilde{a}x}\left[1 - x + R(x,t) + O\big(t^{-1/2}\big)\right],
\end{aligned}$$

where

$$R(x,t) := \frac{x\big(\log\tilde{a} - \log(1+\tilde{a}x)\big)}{\log\tilde{a}}.$$

Here and in what follows, the O's are uniform in $[0,1]$. To control R we note that, if $\tilde{a}x \ge 1$, then

$$\log(\tilde{a}x) \le \log(1+\tilde{a}x) \le \log(\tilde{a}x) + \log 2,$$

and hence

$$|R(x,t)| \le \frac{x\big(|\log x| + \log 2\big)}{\log\tilde{a}} \le \frac{C}{\log\tilde{a}},$$

whereas, if $\tilde{a}x < 1$, then $|R(x,t)| \le x \le 1/\tilde{a}$. It follows that $\sup_{x\in[0,1]} |R(x,t)| = O(t^{-1/2})$ as $t \to \infty$. Since, by (4.3), we have

$$1 + \tilde{a}x = (1 + Ax)\left[1 + \frac{(\tilde{a} - A)x}{1 + Ax}\right] = (1 + Ax)\left[1 + \frac{O\big(t^{-1/2}\log t\big)Ax}{1 + Ax}\right]$$
$$= (1 + Ax)\big(1 + O\big(t^{-1/2}\log t\big)\big),$$

we deduce that

$$\frac{1}{1 + \tilde{a}x} - \tilde{b}f(\tilde{a}x) = \frac{1 + O\big(t^{-1/2}\log t\big)}{1 + Ax}\left(1 - x + O\big(t^{-1/2}\big)\right),$$

and hence

$$\frac{1}{1 + \tilde{a}x} - \tilde{b}f(\tilde{a}x) = \frac{1 - x + O\big(t^{-1/2}\log t\big)}{1 + Ax}. \tag{4.47}$$

On the other hand, due to Lemma 3.6 and $g(0) = 0$, there exists $C_1 > 0$ such that

$$|g(y)| \le C_1(1 + y)\log(1 + y), \quad y \ge 0.$$

Consequently, using also (4.45), we obtain, for $t \to \infty$,

$$(1 + Ax)\tilde{b}^2|g(\tilde{a}x)| \le 2C_1\,\frac{(1 + \tilde{a}x)^2\log(1 + \tilde{a}x)}{\tilde{a}^2\log^2\tilde{a}} \le \frac{3C_1}{\log\tilde{a}} = O\big(t^{-1/2}\big). \tag{4.48}$$

Combining (4.47) and (4.48) finally yields

$$1 - \underline{u}(x, t + T) = \frac{1}{1 + \tilde{a}x} - \tilde{b}f(\tilde{a}x) + \tilde{b}^2 g(\tilde{a}x) = \frac{1 - x + O\big(t^{-1/2}\log t\big)}{1 + Ax}. \qquad \square$$

Proof of (2.10) *and* (2.11). First notice that estimate (3.27) in Lemma 3.3 and Lemma 4.3(ii) guarantee the control of the slope at $x = 1$, namely, (3.22), for any finite $T > 0$. The C^1 regularity property (2.10) is then a consequence of Lemma 3.2. To show (2.11), note that $\underline{u}_x(0,t) = a(t)$ due to (4.7) and (4.2). The lower estimate corresponding to (2.11) is then a consequence of (4.43) and (4.3). The proof of the upper part is similar by using $\overline{u}$. $\square$

Proof of Corollary 2.2. Assertion (i) is an immediate consequence of (2.8) and (2.9). To show (ii), it suffices to observe that, due to (2.9),

$$\int_0^1 \frac{dx}{1 + A(t)x} = \left[\frac{\log\big(1 + A(t)x\big)}{A(t)}\right]_0^1 = \frac{\log(1 + A(t))}{A(t)}$$
$$= \big(1 + O\big(t^{-1/2}\big)\big)\sqrt{2t}\exp\left[-\frac{5}{2} - \sqrt{2t}\right]$$

as $t \to \infty$ and to use

$$\int_0^1 \frac{x\,dx}{1 + A(t)x} \le \frac{1}{A(t)}. \qquad \square$$

Acknowledgments. This research was performed while Nikos I. Kavallaris was a visitor at the Laboratoire Analyse Géométrie et Applications in Université Paris-Nord. He is grateful to this institution for its hospitality and stimulating atmosphere. He would like to express his sincere thanks to Piotr Biler, who introduced him to the chemotaxis problem, and to Andrew A. Lacey for stimulating discussions.

REFERENCES

[1] F. Bavaud, *Equilibrium properties of the Vlasov functional: The generalized Poisson-Boltzmann-Emden equation*, Rev. Modern Phys., 63 (1991), pp. 129–149.
[2] P. Biler, *Existence and nonexistence of solutions for a model of gravitational interaction of particles* III, Colloq. Math., 68 (1995), pp. 229–239.
[3] P. Biler, *Local and global solvability of some parabolic systems modeling chemotaxis*, Adv. Math. Sci. Appl., 8 (1998), pp. 715–743.
[4] P. Biler, D. Hilhorst, and T. Nadzieja, *Existence and nonexistence of solutions for a model of gravitational interaction of particles* II, Colloq. Math., 67 (1994), pp. 297–308.
[5] P. Biler, G. Karch, P. Laurençot, and T. Nadzieja, *The 8π-problem for radially symmetric solutions of a chemotaxis model in a disc*, Topol. Methods Nonlinear Anal., 27 (2006), pp. 133–144.
[6] P. Biler, G. Karch, P. Laurençot, and T. Nadzieja, *The 8π-problem for radially symmetric solutions of a chemotaxis model in the plane*, Math. Methods Appl. Sci., 29 (2006), pp. 1563–1583.
[7] P. Biler and T. Nadzieja, *Existence and nonexistence of solutions for a model of gravitational interaction of particles* I, Colloq. Math., 66 (1994), pp. 319–334.
[8] P. Biler and T. Nadzieja, *Growth and accretion of mass in an astrophysical model* II, Appl. Math., 23 (1995), pp. 351–361.
[9] P. Biler and T. Nadzieja, *A nonlocal singular parabolic problem modelling gravitational interaction of particles*, Adv. Differential Equations, 3 (1998), pp. 177–197.
[10] A. Blanchet, J.A. Carrillo, and N. Masmoudi, *Infinite time aggregation for the critical two-dimensional Patlak-Keller-Segel model*, Comm. Pure Appl. Math., 61 (2008), pp. 1449–1481.
[11] A. Blanchet, J. Dolbeault, and B. Perthame, *Two dimensional Keller-Segel model: Optimal critical mass and qualitative properties of solutions*, Electron. J. Differential Equations, 44 (2006), pp. 1–32.
[12] M.P. Brenner, P. Constantin, L.P. Kadanoff, A. Schenkel, and S.C. Venkataramani, *Diffusion, attraction and collapse*, Nonlinearity, 12 (1999), pp. 1071–1098.
[13] S. Chandrasekhar, *An Introduction to the Study of Stellar Structure*, Dover, New York, 1967.
[14] S. Childress and J.K. Percus, *Nonlinear aspects of chemotaxis*, Math. Biosci., 56 (1981), pp. 217–237.
[15] M. Chipot and F.B. Weissler, *Some blowup results for a nonlinear parabolic equation with a gradient term*, SIAM J. Math. Anal., 20 (1989), pp. 886–907.
[16] M. Doi and S.F. Edwards, *The Theory of Polymer Dynamics*, Clarendon Press, Oxford, UK, 1986.
[17] J. Dolbeault and B. Perthame, *Optimal critical mass in the two-dimensional Keller-Segel model in $\mathbb{R}^2$*, C. R. Math. Acad. Sci. Paris, 339 (2004), pp. 611–616.
[18] J.W. Dold, V.A. Galaktionov, A.A. Lacey, and J.L. Vázquez, *Rate of approach to a singular steady state in quasilinear reaction-diffusion equations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 26 (1998), pp. 663–687.
[19] E. Feireisl, H. Petzeltova, and P. Laurençot, *On convergence to equilibria for the Keller-Segel chemotaxis model*, J. Differential Equations, 236 (2007), pp. 551–569.
[20] M. Fila, M. Winkler, and E. Yanagida, *Grow-up rate of solutions for a supercritical semilinear diffusion equation*, J. Differential Equations, 205 (2004), pp. 365–389.
[21] M. Fila, J.R. King, M. Winkler, and E. Yanagida, *Optimal lower bound of the grow-up rate for a supercritical parabolic equation*, J. Differential Equations, 228 (2006), pp. 339–356.
[22] V.A. Galaktionov and J.R. King, *Composite structure of global unbounded solutions of nonlinear heat equations with critical Sobolev exponents*, J. Differential Equations, 189 (2003), pp. 199–233.
[23] I.A. Guerra and M.A. Peletier, *Self-similar blow-up for a diffusion-attraction problem*, Nonlinearity, 17 (2004), pp. 2137–2162.

[24] M.A. Herrero, *The mathematics of chemotaxis*, in Handbook of Differential Equations: Evolutionary Equations, Vol. 3, Elsevier, New York, 2007, pp. 137–193.
[25] M.A. Herrero and J.J.L. Velázquez, *Singularity patterns in a chemotaxis model*, Math. Ann., 306 (1996), pp. 583–623.
[26] M.A. Herrero and J.J.L. Velázquez, *Chemotactic collapse for the Keller-Segel model*, J. Math. Biol., 35 (1996), pp. 177–194.
[27] M.A. Herrero and J.J.L. Velázquez, *A blow-up mechanism for a chemotaxis model*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 24 (1997), pp. 633–683.
[28] D. Horstmann, *From* 1970 *until present: The Keller-Segel model in chemotaxis and its consequences* I, Jahresber. Deutsch. Math.-Verein., 105 (2003), pp. 103–165.
[29] D. Horstmann, *From* 1970 *until present: The Keller-Segel model in chemotaxis and its consequences* II, Jahresber. Deutsch. Math.-Verein., 106 (2004), pp. 51–69.
[30] W. Jäger and S. Luckhaus, *On explosions of solutions to a system of partial differential equations modelling chemotaxis*, Trans. Amer. Math. Soc., 329 (1992), pp. 819–824.
[31] N.I. Kavallaris and T. Suzuki, *On the finite-time blow-up of a non-local parabolic equation describing chemotaxis*, Differential Integral Equations, 20 (2007), pp. 293–308.
[32] E.F. Keller and L.A. Segel, *Initiation of slime mold aggregation viewed as an instability*, J. Theor. Biol., 26 (1970), pp. 399–415.
[33] A.A. Lacey and D. Tzanetis, *Global existence and convergence to a singular steady state for a semilinear heat equation*, Proc. Roy. Soc. Edinburgh Sect. A, 105 (1987), pp. 289–305.
[34] N. Mizoguchi, *Growup of solutions for a semilinear heat equation with supercritical nonlinearity*, J. Differential Equations, 227 (2006), pp. 652–669.
[35] T. Nagai, *Blow-up of radially symmetric solutions to a chemotaxis system*, Adv. Math. Sci. Appl., 5 (1995), pp. 1–21.
[36] T. Nagai, *Blow-up of nonradial solutions to parabolic-elliptic systems modeling chemotaxis in two-dimensional domains*, J. Inequal. Appl., 6 (2001), pp. 37–55.
[37] T. Nagai, T. Senba, and K. Yoshida, *Application of the Trudinger-Moser inequality to a parabolic system of chemotaxis*, Funkcial. Ekvac., 40 (1997), pp. 411–433.
[38] V. Nanjundiah, *Chemotaxis, signal relaying, and aggregation morphology*, J. Theor. Biol., 42 (1973), pp. 63–105.
[39] K. Ohtsuka, T. Senba, and T. Suzuki, *Blowup in infinite time in the simplified system of chemotaxis*, Adv. Math. Sci. Appl., 17 (2007), pp. 445–472.
[40] C.S. Patlak, *Random walk with persistence and external bias*, Bull. Math. Biol. Biophys., 15 (1953), pp. 311–338.
[41] B. Perthame, *PDE models for chemotactic movements: Parabolic, hyperbolic and kinetic*, Appl. Math., 49 (2004), pp. 539–564.
[42] P. Polačik and E. Yanagida, *On bounded and unbounded global solutions of a supercritical semilinear heat equation*, Math. Ann., 327 (2003), pp. 745–771.
[43] P. Quittner and Ph. Souplet, *Superlinear Parabolic Problems. Blow-up, Global Existence and Steady States*, Birkhäuser Advanced Texts, Birkhäuser Verlag, Basel, 2007.
[44] T. Senba and T. Suzuki, *Chemotactic collapse in a parabolic-elliptic system of mathematical biology*, Adv. Differential Equations, 6 (2001), pp. 21–50.
[45] T. Senba, *Blowup in infinite time of radial solutions for a parabolic-elliptic system in high dimensional Euclidean spaces*, Nonlinear Anal., to appear.
[46] C. Sire and P.-H. Chavanis, *Thermodynamics and collapse of self-gravitating Brownian particles in D dimensions*, Phys. Rev. E (3), 66 (2002), 046133.
[47] Ph. Souplet and J.L. Vázquez, *Stabilization towards a singular steady state with gradient blow-up for a diffusion-convection problem*, Discrete Contin. Dyn. Syst., 14 (2006), pp. 221–234.
[48] T. Suzuki, *Free Energy and Self-Interacting Particles*, Birkhäuser Boston, Boston, 2005.
[49] G. Wolansky, *On steady distributions of self-attracting clusters under friction and fluctuations*, Arch. Ration. Mech. Anal., 119 (1992), pp. 355–391.
[50] G. Wolansky, *On the evolution of self-interacting clusters and applications to semilinear equations with exponential nonlinearity*, J. Anal. Math., 59 (1992), 251–272.
[51] G. Wolansky, *A critical parabolic estimate and application to nonlocal equations arising in chemotaxis*, Appl. Anal., 66 (1997), 291–321.

SIAM J. MATH. ANAL.
Vol. 40, No. 5, pp. 1882–1904

MATHEMATICAL RESULTS ON EXISTENCE FOR VISCOELASTODYNAMIC PROBLEMS WITH UNILATERAL CONSTRAINTS*

ADRIEN PETROV† AND MICHELLE SCHATZMAN‡

Abstract. This paper focuses on a damped wave equation and the evolution of a Kelvin–Voigt viscoelastic material, both problems being subject to unilateral boundary conditions. Under appropriate regularity assumptions on the initial data, both problems possess a weak solution which is obtained as the limit of a sequence of solutions of penalized problems; the functional properties of all the traces are precisely identified through Fourier analysis, and this enables us to infer the existence of a strong solution, i.e., a solution satisfying almost everywhere the unilateral conditions.

Key words. viscoelasticity, Signorini conditions, penalty method, traces, variational inequality, convolution

AMS subject classifications. 35L85, 49J40, 73D99, 73V25

DOI. 10.1137/070695101

1. Introduction and notation.

This paper aims to give some new mathematical results on existence for a damped wave equation with an obstacle and for full viscoelasticity in the particular case of a Kelvin–Voigt material with unilateral boundary conditions.

We consider in section 2 a damped wave equation taking place in a half-space, with an obstacle at the boundary. Let $u(x,t)$ be the displacement at time t of the material point of spatial coordinate $x = (x_1, x') \in (-\infty, 0] \times \mathbb{R}^{d-1}$ at rest with $d \geq 2$. We will agree that if we write a function of space and time as a function of three variables, then the first variable is the normal space variable x_1, the second variable is the tangential space variable x', and the last variable is time. Let $f(x_1, x', t)$ denote a density of external forces, depending on space and time. Define $\Omega \stackrel{\text{def}}{=} (-\infty, 0] \times \mathbb{R}^{d-1}$, and let α be a positive number. The mathematical problem is formulated as follows:

$$u_{tt} - \Delta u - \alpha \Delta u_t = f, \quad x \in \Omega, \quad t > 0, \tag{1.1}$$

with Cauchy initial data

$$u(\cdot, 0) = u_0 \quad \text{and} \quad u_t(\cdot, 0) = u_1 \tag{1.2}$$

and Signorini boundary conditions at $x_1 = 0$, $t > 0$,

$$0 \leq u \perp u_{x_1} + \alpha u_{x_1 t} \geq 0, \tag{1.3}$$

where $(\cdot)_t \stackrel{\text{def}}{=} \frac{\partial}{\partial t}(\cdot)$ and $(\cdot)_{x_1} \stackrel{\text{def}}{=} \frac{\partial}{\partial x_1}(\cdot)$. The orthogonality has the natural meaning: if we have enough regularity, it means that the product $u(u_{x_1} + \alpha u_{x_1 t})$ vanishes almost everywhere on the boundary. If we do not have enough regularity, the above inequality

*Received by the editors June 22, 2007; accepted for publication (in revised form) September 23, 2008; published electronically January 7, 2009.
http://www.siam.org/journals/sima/40-5/69510.html
†Weierstraß-Institut für Angewandte Analysis und Stochastik, Mohrenstraße 39, 10117 Berlin, Germany (petrov@wias-berlin.de).
‡CNRS, Université de Lyon, Institut Camille Jordan, 21 Avenue Claude Bernard, F-69622 Villeurbanne Cedex, France (schatz@math.univ-lyon1.fr).

is integrated on an appropriate set of test functions, yielding a weak formulation for the unilateral condition. The main result of section 2 is that, indeed, (1.3) holds almost everywhere on the boundary; i.e., we have a strong solution to our problem.

We suppose that the initial position u_0 belongs to the Sobolev space $\mathrm{H}^2(\Omega)$ and satisfies the compatibility condition $u_0(0,\cdot,\cdot) \geq 0$, the initial velocity u_1 belongs to $\mathrm{H}^1(\Omega)$, and the density of forces f belongs to $\mathrm{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$. The choice of a function f defined for all nonnegative time is justified by the use of a Fourier transform in the latter part of the article. This is not a significant restriction as we can always extend f by 0 if it is defined only for finite times.

Let us describe the weak formulation of the problem. Denote by K the convex set

$$K \stackrel{\text{def}}{=} \big\{ v \in \mathrm{H}^1_{\mathrm{loc}}(\Omega \times [0,\infty)) : \nabla v_t \in \mathrm{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega)),\ v|_{\{0\}\times\mathbb{R}^{d-1}} \geq 0 \big\}.$$

This unusual convex set has been devised in order to write a weak formulation of our problem. Since we expect to find a scalar product $(\nabla u_t, \nabla w)$, we require ∇u_t to be square integrable. Thus, the weak formulation associated to (1.1)–(1.3) is obtained by multiplying (1.1) by $v-u$, $v \in K$, and by integrating formally over $\Omega \times (0,\tau)$. Then, we get

$$\begin{cases} \text{find } u \in K \text{ such that for all } v \in K \text{ and for all } \tau \in (0,\infty), \\ \displaystyle\int_\Omega (u_t(v-u))\big|_0^\tau \,\mathrm{d}x - \int_0^\tau \int_\Omega u_t(v_t - u_t)\,\mathrm{d}x\,\mathrm{d}t \\ \displaystyle + \int_0^\tau \int_\Omega (\nabla u + \alpha \nabla u_t)(\nabla v - \nabla u)\,\mathrm{d}x\,\mathrm{d}t \geq \int_0^\tau \int_\Omega f(v-u)\,\mathrm{d}x\,\mathrm{d}t. \end{cases} \tag{1.4}$$

We approximate this problem by a regularized problem, and the bulk of our work is to get enough regularity to obtain traces of the tangential first derivatives of the solution on the boundary. This is done essentially through an analysis of the Dirichlet to Neumann operator for the damped wave equation. For this purpose, one must go to tangential Fourier variables and perform an estimate on a pseudodifferential term.

In section 3, we treat the evolution of a Kelvin–Voigt material (see [DaL90]) occupying a three dimensional half-space, satisfying Signorini conditions at the boundary and Cauchy data at $t=0$. We make the assumptions of small deformations. Let $\varepsilon_{ij}(u) \stackrel{\text{def}}{=} (u_{j,x_i} + u_{i,x_j})/2$ be the strain tensor, and let there be given two Hooke tensors, a^n_{ijkl}, $n=0,1$. We define the two stress tensors σ^n_{ij} corresponding, respectively, to the elastic and the viscous parts of the stress:

$$\sigma^n_{ij}(u) \stackrel{\text{def}}{=} a^n_{ijkl}\varepsilon_{kl}(u); \tag{1.5}$$

here, we have used the summation convention on repeated indices. The displacement field u satisfies the system

$$\rho u_{i,tt} = \sigma^0_{ij,x_j}(u) + \sigma^1_{ij,x_j}(u_t) + f_i, \quad x \in \Omega, \quad t > 0. \tag{1.6}$$

The initial data are given by

$$u(\cdot,0) = v_0 \quad \text{and} \quad u_t(\cdot,0) = v_1. \tag{1.7}$$

The components of the unit external normal are δ_{1j} (δ is the Kronecker symbol), and a basis of tangential vectors can be taken as $\tau_j = \delta_{2j}$ and $\tau'_j = \delta_{3j}$. Denote by

$\Sigma = \{0\} \times \mathbb{R}^{d-1}$ the boundary of Ω. Then, the boundary conditions on $\Sigma \times [0, \infty)$ are

$$0 \geq u_1 \perp \sigma^0_{11}(u) + \sigma^1_{11}(u_t) \leq 0, \tag{1.8a}$$

$$\sigma^0_{12}(u) + \sigma^1_{12}(u_t) = 0, \quad \text{and} \quad \sigma^0_{13}(u) + \sigma^1_{13}(u_t) = 0. \tag{1.8b}$$

One of the main results of section 3 is to show that (1.8a) holds almost everywhere, because the relevant traces exist.

In order to simplify the problem, we have considered a homogeneous and isotropic material; then, the Hooke tensors a^n_{ijkl} are defined with the help of Lamé constants λ^n and μ^n:

$$a^n_{ijkl} \stackrel{\text{def}}{=} \lambda^n \delta_{ij}\delta_{kl} + 2\mu^n \delta_{ik}\delta_{jl}, \quad n = 0, 1.$$

We define two elasticity operators A^n by

$$A^n u \stackrel{\text{def}}{=} a^n_{ijkl} \partial_j \varepsilon_{kl}(u), \quad n = 0, 1.$$

Then, the problem (1.6)–(1.8) can be rewritten as follows:

$$\rho u_{tt} - A^0 u - A^1 u_t = f, \quad x \in \Omega, \quad t > 0, \tag{1.9a}$$

$$0 \geq u_1 \perp \big(\sigma^0_{11}(u) + \sigma^1_{11}(u_t)\big) \leq 0 \quad \text{on} \quad \Sigma \times [0, \infty), \tag{1.9b}$$

$$\sigma^0_{12}(u) + \sigma^1_{12}(u_t) = 0 \quad \text{and} \quad \sigma^0_{13}(u) + \sigma^1_{13}(u_t) = 0 \quad \text{on} \quad \Sigma \times [0, \infty), \tag{1.9c}$$

$$u(\cdot, 0) = v_0 \quad \text{and} \quad u_t(\cdot, 0) = v_1. \tag{1.9d}$$

Let us describe now the functional hypotheses on the data; if X is a space of scalar functions, the bold-face notation $\mathbf{X}$ denotes the space X^d. For the final result, we require v_0 to belong to $\mathbf{H}^{5/2}(\Omega)$, v_1 to $\mathbf{H}^{3/2}(\Omega)$, and f to $\mathbf{H}^1_{\text{loc}}([0, \infty); \mathrm{L}^2(\Omega))$. The initial data must satisfy the compatibility condition $(v_0)_1(0, x') \leq 0$ for all $x' \in \Sigma$. Let K be the convex set defined by

$$K \stackrel{\text{def}}{=} \big\{ v \in \mathbf{H}^1(\Omega \times (0, \tau)) : \nabla v_t \in \mathbf{L}^2(\Omega \times (0, \tau)),\ v(0, \cdot) \leq 0 \big\}.$$

Define two bilinear forms by

$$a^0(u, v) \stackrel{\text{def}}{=} \int_\Omega a^0_{ijkl} \varepsilon_{ij}(u) \varepsilon_{kl}(v)\, \mathrm{d}x \quad \text{and} \quad a^1(u, v) \stackrel{\text{def}}{=} \int_\Omega a^1_{ijkl} \varepsilon_{ij}(u) \varepsilon_{kl}(v)\, \mathrm{d}x.$$

We obtain a weak formulation of the problem (1.9) as follows: we multiply (1.9a) by $v - u$, $v \in K$, and we formally integrate the result over $\Omega \times (0, \tau)$; we obtain then the variational inequality

$$\begin{cases} \text{find } u \in K \text{ such that for all } v \in K \text{ and for all } \tau \in (0, \infty), \\ \displaystyle\int_0^\tau \int_\Omega \rho u_{tt} \cdot (v - u)\, \mathrm{d}x\, \mathrm{d}t + \int_0^\tau a^0(u, u - v)\, \mathrm{d}t \\ \displaystyle + \int_0^\tau a^1(u_t, v - u)\, \mathrm{d}t \geq \int_0^\tau \int_\Omega f \cdot (v - u)\, \mathrm{d}x\, \mathrm{d}t. \end{cases} \tag{1.10}$$

The existence result for (1.1)–(1.3) is easily established by the penalty method and was already proved by Jarušek et al. [JM*92] in the case of distributed constraints.

Jarušek has also proved in [Jar96] an existence result for (1.9), in a much more general and complicated case, since it allows for contact, a given friction at the boundary, a nonlinear constitutive law for viscoelasticity, and a general geometry. However,

the boundary conditions must be understood in the sense of duality, since the traces themselves are defined by duality. The reader can also refer to [KuS04a]. Moreover, notice that [KuS04b] contains some regularity results for a Signorini problem with normal compliance.

The reader may wonder why our proofs of existence for weak solutions are longer than other known proofs; the reason is that we prove more estimates in order to get more information. Hence, we cannot reuse previous proofs of existence of weak solutions. In particular, we need stronger results on the regularity of solutions, and these cannot be deduced from former results on regularity, which use stronger assumptions.

In the present paper, for both problems, we penalize the obstacle constraint, we construct a solution of the penalized problem, and we show the existence of a weak solution by passing to the limit with respect to the penalty parameter. Then, under appropriate regularity conditions on the data, we prove that the penalized solution has traces, which can be estimated, and therefore the limiting weak solution that we obtained is a strong solution.

We have chosen to limit ourselves to homogeneous and, in the case of viscoelasticity, isotropic media, since we use a Fourier transform and some calculus in the dual variables for obtaining the traces. We believe that more general cases could be considered, namely, variable coefficients and curved boundary. The method is probably straightforward, but it is exacting, since one would probably have to use the full technology of pseudodifferential operators.

Observe that nothing is known about uniqueness.

These two problems are treated in the same article, because they are quite close. Proofs for the second problem are shortened when very close to proofs for the first one. Nevertheless, there are substantial differences in detail, since the second problem is much more complicated than the first. In particular, the bulk of the proof in section 3 consists in obtaining a solution of a linear system through Fourier–Laplace transform and then in estimating this solution in anisotropic Sobolev spaces.

2. The damped wave equation with Signorini boundary conditions.

2.1. The penalized problem. We approximate (1.1)–(1.3) by the penalty method. This means that we replace the rigid constraint (1.3) by a very stiff response. When the constraint is active, the response is linear, and it vanishes when the constraint is not active. More precisely, letting $r^- \stackrel{\text{def}}{=} -\min(r,0)$, we replace u by u^ϵ, which satisfies

$$u^\epsilon_{tt} - \Delta u^\epsilon - \alpha \Delta u^\epsilon_t = f, \quad x \in \Omega, \quad t > 0, \tag{2.1}$$

with initial data

$$u^\epsilon(\cdot,0) = u_0 \quad \text{and} \quad u^\epsilon_t(\cdot,0) = u_1 \tag{2.2}$$

and boundary condition

$$(u^\epsilon_{x_1} + \alpha u^\epsilon_{x_1 t})(0,\cdot,\cdot) = (u^\epsilon(0,\cdot,\cdot))^-/\epsilon. \tag{2.3}$$

Notice that (2.3) is the so-called *normal compliance condition* introduced by Martins and Oden [MaO88].

Define the following sets:

$$\forall \tau \in (0,\infty), \quad Q_\tau \stackrel{\text{def}}{=} \Omega \times (0,\tau) \quad \text{and} \quad I_\tau \stackrel{\text{def}}{=} \Sigma \times (0,\tau). \tag{2.4}$$

THEOREM 2.1. *Let* $\mathrm{W}_{\mathrm{loc}} \stackrel{\mathrm{def}}{=} \{u \in \mathrm{H}^1_{\mathrm{loc}}([0,\infty)\times\Omega) : \nabla u_t \in \mathrm{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))\}$. *Assume that* u_0 *and* u_1 *belong to* $\mathrm{H}^1(\Omega)$, *and* f *belongs to* $\mathrm{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$; *then for every* $\epsilon > 0$ *there exists a unique weak solution* $u^\epsilon \in \mathrm{W}_{\mathrm{loc}}$ *of the problem* (2.1)–(2.3) *such that*

$$u^\epsilon \in \mathrm{L}^\infty_{\mathrm{loc}}([0,\infty);\mathrm{H}^1(\Omega)), \tag{2.5a}$$

$$u^\epsilon_t \in \mathrm{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{H}^1(\Omega)), \tag{2.5b}$$

$$u^\epsilon_{tt} \in \mathrm{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega)), \tag{2.5c}$$

and for every $\tau \in (0,T)$ *and for all* $v \in \mathrm{W}_{\mathrm{loc}}$, *the following variational equality is satisfied:*

$$\begin{aligned}\int_\Omega ((u^\epsilon_t v)(\cdot,\tau) - u_1 v(\cdot,0))\,\mathrm{d}x - \int_{Q_\tau} u^\epsilon_t v_t\,\mathrm{d}x\,\mathrm{d}t + \int_{Q_\tau} \nabla u^\epsilon \nabla v\,\mathrm{d}x\,\mathrm{d}t \\ + \alpha\int_{Q_\tau} \nabla u^\epsilon_t \nabla v\,\mathrm{d}x\,\mathrm{d}t - \frac{1}{\epsilon}\int_{I_\tau} (u^\epsilon)^- v\,\mathrm{d}x'\,\mathrm{d}t = \int_{Q_\tau} f v\,\mathrm{d}x\,\mathrm{d}t.\end{aligned} \tag{2.6}$$

Proof. The theorem is proved by the standard Galerkin method, and the reader can refer to [GGZ74] or the appendix of [JM*92]. □

2.2. A priori estimates. We establish here estimates up to the boundary and interior estimates which will enable us later to infer the existence of a weak solution to (1.1)–(1.3).

LEMMA 2.2. *Assume that* f *belongs to* $\mathrm{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$, u_0 *to* $\mathrm{H}^1(\Omega)$, *and* u_1 *to* $\mathrm{L}^2(\Omega)$. *Then* u^ϵ_t *and* ∇u^ϵ *are bounded in* $\mathrm{L}^\infty_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$, ∇u^ϵ_t *is bounded in* $\mathrm{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$, *and* $(u^\epsilon(0,\cdot,\cdot))^-/\sqrt{\epsilon}$ *is bounded in* $\mathrm{L}^\infty_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\mathbb{R}^-))$ *independently of* $\epsilon > 0$. *If, moreover,* u_0 *belongs to* $\mathrm{H}^2(\Omega)$, Δu^ϵ *is bounded in* $\mathrm{L}^\infty_{\mathrm{loc}}(0,\infty;L^2(\Omega))$ *independently of* ϵ.

Proof. These estimates result from an application of the Gronwall lemma to the energy identity. We multiply (2.1) by u^ϵ_t, and we integrate this expression over Q_τ to get

$$\int_{Q_\tau} u^\epsilon_{tt} u^\epsilon_t\,\mathrm{d}x\,\mathrm{d}t - \int_{Q_\tau} \Delta u^\epsilon u^\epsilon_t\,\mathrm{d}x\,\mathrm{d}t - \alpha\int_{Q_\tau} \Delta u^\epsilon_t u^\epsilon_t\,\mathrm{d}x\,\mathrm{d}t = \int_{Q_\tau} f u^\epsilon_t\,\mathrm{d}x\,\mathrm{d}t.$$

We integrate the first integral in time in the above relation, we use Green's formula for the second and third, and, with the help of the boundary conditions (2.3), we obtain

$$\begin{aligned}&\frac{1}{2}\int_\Omega \left(|u^\epsilon_t(\cdot,\tau)|^2 + |\nabla u^\epsilon(\cdot,\tau)|^2\right)\mathrm{d}x + \alpha\int_{Q_\tau} |\nabla u^\epsilon_t|^2\,\mathrm{d}x\,\mathrm{d}t + \frac{1}{2\epsilon}\int_\Sigma \left((u^\epsilon(0,\cdot,\cdot))^-\right)^2\Big|_0^\tau \mathrm{d}x' \\ &= \int_{Q_\tau} f u^\epsilon_t\,\mathrm{d}x\,\mathrm{d}t + \frac{1}{2}\int_\Omega \left(|\nabla u_0|^2 + |u_1|^2\right)\mathrm{d}x.\end{aligned}$$

We may deduce from the classical Gronwall lemma that u^ϵ_t and ∇u^ϵ are bounded in $\mathrm{L}^\infty_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$, ∇u^ϵ_t is bounded in $\mathrm{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$, and $(u^\epsilon(0,\cdot,\cdot))^-/\sqrt{\epsilon}$ is bounded in $\mathrm{L}^\infty_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\mathbb{R}^{d-1}))$ independently of $\epsilon > 0$.

For the bound on Δu^ϵ, we multiply (2.1) by $\mathrm{e}^{t/\alpha}$, and we integrate from 0 to τ. After an integration by parts on the term

$$\int_0^\tau u^\epsilon_{tt}\mathrm{e}^{t/\alpha}\,\mathrm{d}t,$$

we find the identity

$$\begin{aligned}\Delta u^\epsilon(\cdot,\tau) &= \Delta u_0 \mathrm{e}^{-\tau/\alpha} + \alpha^{-1}\big(u_t^\epsilon(\cdot,\tau) - u_1 \mathrm{e}^{-\tau/\alpha}\big)\\ &\quad - \alpha^{-2}\int_0^\tau u_t^\epsilon(\cdot,t)\mathrm{e}^{(t-\tau)/\alpha}\,\mathrm{d}t - \alpha^{-1}\int_0^\tau f\mathrm{e}^{(t-\tau)/\alpha}\,\mathrm{d}t.\end{aligned} \tag{2.7}$$

The end of the proof is straightforward. □

Remark 2.3. If we suppose that f vanishes for t large, then, independently of $\epsilon > 0$, u_t^ϵ, ∇u^ϵ, and Δu^ϵ are bounded in $\mathrm{L}^\infty([0,\infty);\mathrm{L}^2(\Omega))$ and ∇u_t^ϵ is bounded in $\mathrm{L}^2([0,\infty);\mathrm{L}^2(\Omega))$. These properties can be proved using the arguments given in the proof of Lemma 2.2, with the origin of time moved to T if $f(\cdot,t)$ vanishes for $t \geq T$; since the integral involving f vanishes, the conclusion is clear for u_t^ϵ, ∇u^ϵ, and ∇u_t^ϵ. For Δu^ϵ, we use the identity (2.7), with 0 replaced by $T \leq \tau$, and the conclusion is clear.

LEMMA 2.4. *Assume the hypotheses of Lemma* 2.2. *Then for all nonnegative, continuously differentiable, and compactly supported* ψ *on* $\mathbb{R}^{d-1}$ *and for all* $\tau \in [0,T]$,

$$\int_{I_\tau} \frac{(u^\epsilon)^-}{\epsilon}\,\psi\,\mathrm{d}x'\,\mathrm{d}t$$

is bounded independently of $\epsilon > 0$. *In particular,* $(u^\epsilon(0,\cdot,\cdot))^-/\epsilon$ *is a bounded measure on* I_τ.

Proof. Let ϕ be a continuous function with compact support in $\mathbb{R}^d$; we multiply (2.1) by ϕ and integrate over Q_τ; thanks to the boundary conditions (2.3) and Green's formula, we obtain

$$\int_\Omega \phi u_t^\epsilon(\cdot,t)\big|_0^\tau\,\mathrm{d}x + \int_{Q_\tau}\nabla\phi(\nabla u^\epsilon + \alpha\nabla u_t^\epsilon)\,\mathrm{d}x\,\mathrm{d}t - \frac{1}{\epsilon}\int_{I_\tau}(u^\epsilon)^-\phi\,\mathrm{d}x'\,\mathrm{d}t = \int_{Q_\tau}\phi f\,\mathrm{d}x\,\mathrm{d}t.$$

Since the product $|zy|$ can be estimated by $|z|^2/2 + |y|^2/2$, we get the following inequality:

$$\begin{aligned}&\frac{1}{\epsilon}\int_{I_\tau}(u^\epsilon)^-\phi\,\mathrm{d}x'\,\mathrm{d}t \leq \frac{1}{2}\int_\Omega\big(|u_t^\epsilon(\cdot,\tau)|^2 + |u_1|^2\big)\,\mathrm{d}x + \int_\Omega|\phi|^2\,\mathrm{d}x\\ &+\int_{Q_\tau}|\nabla\phi\nabla u^\epsilon|\,\mathrm{d}x\,\mathrm{d}t + \alpha\int_{Q_\tau}|\nabla\phi\nabla u_t^\epsilon|\,\mathrm{d}x\,\mathrm{d}t + \int_{Q_\tau}|\phi f|\,\mathrm{d}x\,\mathrm{d}t.\end{aligned} \tag{2.8}$$

The right-hand side of (2.8) is bounded since f belongs to $\mathrm{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$, u_1 to $\mathrm{L}^2(\Omega)$, and u_t^ϵ, ∇u^ϵ, and ∇u_t^ϵ to $\mathrm{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$. Moreover, $(u^\epsilon(0,\cdot,\cdot))^-$ is nonnegative; if the trace ψ of ϕ over Σ is nonnegative, the inequality is clear. The last statement of the theorem is obtained by a classical approximation argument. Write $\mu^\epsilon = (u^\epsilon(0,\cdot,\cdot))^-/\epsilon$. Let ψ_n be an increasing sequence of nonnegative, continuously differentiable, and compactly supported functions on Σ, which are at most equal to ψ. Then the integrals of ψ_n against μ^ϵ converge to the integral of $\lim_n \psi_n$ against μ^ϵ, so that the integral of any nonnegative, continuous, and compactly supported function against μ^ϵ is nonnegative, and this is precisely the definition of a nonnegative measure on Σ. □

Let us turn now to interior estimates.

LEMMA 2.5. *Assume the hypotheses of Lemma* 2.2. *Then for all* $\beta > 0$, u_{tt}^ϵ *and* Δu_t^ϵ *are bounded in* $\mathrm{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2((-\infty,-\beta)\times\Sigma))$, *independently of* $\epsilon > 0$.

Proof. The idea of the proof is twofold: we multiply u^ϵ by a truncation function $\varphi \in \mathrm{C}_0^\infty(\mathbb{R})$, and we define $v^\epsilon \overset{\text{def}}{=} \varphi u^\epsilon$; we will observe that $w^\epsilon \overset{\text{def}}{=} v_t^\epsilon$ satisfies a heat equation, whose right-hand side will be estimated thanks to the previous lemmas. Let us now go into the details.

Let φ be a truncation function which is equal to 1 if $x_1 \leq -\beta$ and to 0 if $x_1 \geq -\beta/2$ ($\beta > 0$). Then, we multiply u^ϵ by φ, which enables us to forget about the strongly nonlinear boundary conditions. Define

$$v^\epsilon(x_1, \cdot, \cdot) \overset{\text{def}}{=} \varphi(x_1) u^\epsilon(x_1, \cdot, \cdot). \tag{2.9}$$

The derivatives of v^ϵ are given by

$$v_{tt}^\epsilon = \varphi u_{tt}^\epsilon, \tag{2.10a}$$

$$\Delta v^\epsilon = \varphi \Delta u^\epsilon + 2\varphi_{x_1} \nabla u^\epsilon + \varphi_{x_1 x_1} u^\epsilon, \tag{2.10b}$$

$$\Delta v_t^\epsilon = \varphi \Delta u_t^\epsilon + 2\varphi_{x_1} \nabla u_t^\epsilon + \varphi_{x_1 x_1} u_t^\epsilon. \tag{2.10c}$$

Observe that, thanks to relations (2.1) and (2.10), we have

$$v_{tt}^\epsilon - \Delta v^\epsilon - \alpha \Delta v_t^\epsilon = \widetilde{g}^\epsilon, \tag{2.11}$$

where $\widetilde{g}^\epsilon \overset{\text{def}}{=} \varphi f - 2\varphi_{x_1}(\nabla u^\epsilon + \alpha \nabla u_t^\epsilon) - \varphi_{x_1 x_1}(u^\epsilon + \alpha u_t^\epsilon)$. Since f, u_t^ϵ, ∇u^ϵ, ∇u_t^ϵ, and u^ϵ are bounded in $\mathrm{L}^2_{\mathrm{loc}}([0,\infty); \mathrm{L}^2(\Omega))$, $\widetilde{g}^\epsilon$ is bounded in $\mathrm{L}^2_{\mathrm{loc}}([0,\infty); \mathrm{L}^2(\Omega))$. Let us define

$$w^\epsilon \overset{\text{def}}{=} v_t^\epsilon \quad \text{and} \quad g^\epsilon \overset{\text{def}}{=} \widetilde{g}^\epsilon + \Delta v^\epsilon. \tag{2.12}$$

Substituting (2.12) into (2.11), we obtain

$$w_t^\epsilon - \alpha \Delta w^\epsilon = g^\epsilon. \tag{2.13}$$

Let us prove now that w_t^ϵ is bounded in $\mathrm{L}^2_{\mathrm{loc}}([0,\infty); \mathrm{L}^2(\Omega))$. For this purpose, we multiply (2.13) by w_t^ϵ; we integrate this expression over Ω:

$$\int_\Omega |w_t^\epsilon|^2 \, \mathrm{d}x - \alpha \int_\Omega \Delta w^\epsilon w_t^\epsilon \, \mathrm{d}x = \int_\Omega g^\epsilon \, w_t^\epsilon \, \mathrm{d}x.$$

We use Green's formula in the second term on the left-hand side of the above expression, thus getting the following equality:

$$\int_\Omega |w_t^\epsilon|^2 \, \mathrm{d}x + \alpha \int_\Omega \nabla w_t^\epsilon \nabla w^\epsilon \, \mathrm{d}x = \int_\Omega g^\epsilon w_t^\epsilon \, \mathrm{d}x. \tag{2.14}$$

We integrate (2.14) over $(0,\tau)$, we observe that the product $|g^\epsilon w_t^\epsilon|$ can be estimated by $|g^\epsilon|^2/2 + |w_t^\epsilon|^2/2$, and we obtain

$$\begin{aligned} &\frac{1}{2} \int_{Q_\tau} |w_t^\epsilon|^2 \, \mathrm{d}x \, \mathrm{d}t + \alpha \int_\Omega |\nabla w^\epsilon(\cdot, \tau)|^2 \, \mathrm{d}x \\ &\leq \alpha \int_\Omega |\nabla w^\epsilon(\cdot, 0)|^2 \, \mathrm{d}x + \frac{1}{2} \int_{Q_\tau} |g^\epsilon|^2 \, \mathrm{d}x \, \mathrm{d}t. \end{aligned} \tag{2.15}$$

Since u_1 belongs to $\mathrm{H}^1(\Omega)$ and φ belongs to $\mathrm{C}_0^\infty(\mathbb{R})$, $\nabla w^\epsilon(\cdot, 0) = \varphi_{x_1} u_1 + \varphi \nabla u_1$ is bounded in $\mathrm{L}^2(\Omega)$. Moreover, g^ϵ is bounded in $\mathrm{L}^2_{\mathrm{loc}}([0,\infty); \mathrm{L}^2(\Omega))$ because Δv^ϵ and $\widetilde{g}^\epsilon$ are bounded in $\mathrm{L}^2_{\mathrm{loc}}([0,\infty); \mathrm{L}^2(\Omega))$. Therefore, (2.9), (2.12), and (2.15) enable us to

deduce that u^ϵ_{tt} is bounded in $\mathrm{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2((-\infty,-\beta)\times\Sigma))$. We use analogous arguments to show that Δu^ϵ_t is bounded in $\mathrm{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2((-\infty,-\beta)\times\Sigma))$. We multiply (2.13) by Δw^ϵ, we integrate over Q_τ, and, thanks to Green's formula, we obtain

$$-\frac{1}{2}\int_\Omega |\nabla w^\epsilon|^2\big|_0^\tau \,\mathrm{d}x - \alpha\int_{Q_\tau}|\Delta w^\epsilon|^2\,\mathrm{d}x\,\mathrm{d}t = \int_{Q_\tau} g^\epsilon \Delta w^\epsilon\,\mathrm{d}x\,\mathrm{d}t. \tag{2.16}$$

The product $|g^\epsilon\Delta w^\epsilon|$ can be estimated by $|g^\epsilon|^2/(2\gamma)+\gamma|\Delta w^\epsilon|^2/2$, and if we choose $\gamma\in(0,2\alpha)$, we obtain the following inequality:

$$\Big(\alpha-\frac{\gamma}{2}\Big)\int_{Q_\tau}|\Delta w^\epsilon|^2\,\mathrm{d}x\,\mathrm{d}t \le \frac{1}{2\gamma}\int_{Q_\tau}|g^\epsilon|^2\,\mathrm{d}x\,\mathrm{d}t + \frac{1}{2}\int_\Omega|\nabla w^\epsilon(\cdot,0)|^2\,\mathrm{d}x. \tag{2.17}$$

Since g^ϵ and $\nabla w^\epsilon(\cdot,0)$ are, respectively, bounded in $\mathrm{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$ and $\mathrm{L}^2(\Omega)$, according to (2.9), (2.12), and (2.17), we infer that Δu^ϵ_t is bounded in the space $\mathrm{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2((-\infty,-\beta)\times\Sigma))$. ∎

2.3. Existence of a weak solution. In this section, we show that it is possible to pass to the limit in the variational formulation of the penalized problem to obtain a weak solution of (1.1)–(1.3). There is a minor subtlety due to the unboundedness of Ω.

THEOREM 2.6. *Assume the hypotheses of Lemma* 2.2. *Then there exists a solution of the variational inequality* (1.4)*; this solution can be obtained as a limit of a subsequence of the penalty approximation defined by* (2.1)–(2.3).

Proof. Let v belong to K and φ be a function belonging to $\mathrm{C}^\infty_0(\bar\Omega\times[0,\infty))$, which takes its values in $[0,1]$. Multiplying (2.1) by $(v-u^\epsilon)\varphi$ and integrating over Q_τ and then observing that

$$\int_{I_\tau}\big((u^\epsilon)^-\varphi(v-u^\epsilon)\big)\,\mathrm{d}x'\,\mathrm{d}t = \int_{I_\tau}\big(\big((u^\epsilon)^-\big)^2\varphi\big)\,\mathrm{d}x'\,\mathrm{d}t + \int_{I_\tau}\big((u^\epsilon)^-\varphi v\big)\,\mathrm{d}x'\,\mathrm{d}t$$

is nonnegative, we may deduce the following inequality:

$$\begin{aligned}&\int_\Omega u^\epsilon_t\varphi(v-u^\epsilon)\big|_0^\tau\,\mathrm{d}x - \int_{Q_\tau}u^\epsilon_t(\varphi(v-u^\epsilon))_t\,\mathrm{d}x\,\mathrm{d}t\\ &+\int_{Q_\tau}(\nabla u^\epsilon+\alpha\nabla u^\epsilon_t)\nabla(\varphi(v-u^\epsilon))\,\mathrm{d}x\,\mathrm{d}t \ge \int_{Q_\tau} f\varphi(v-u^\epsilon)\,\mathrm{d}x\,\mathrm{d}t.\end{aligned} \tag{2.18}$$

We infer from Lemmas 2.2 and 2.4 that it is possible to extract a subsequence, still denoted by u^ϵ, such that

$$u^\epsilon \rightharpoonup u \quad\text{in}\quad \mathrm{L}^\infty_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega)) \quad\text{weak}\quad *, \tag{2.19a}$$

$$u^\epsilon_t \rightharpoonup u_t \quad\text{in}\quad \mathrm{L}^\infty_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega)) \quad\text{weak}\quad *, \tag{2.19b}$$

$$\nabla u^\epsilon \rightharpoonup \nabla u \quad\text{in}\quad \mathrm{L}^\infty_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega)) \quad\text{weak}\quad *, \tag{2.19c}$$

$$\Delta u^\epsilon \rightharpoonup \Delta u \quad\text{in}\quad \mathrm{L}^\infty_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega)) \quad\text{weak}\quad *, \tag{2.19d}$$

$$\nabla u^\epsilon_t \rightharpoonup \nabla u_t \quad\text{in}\quad \mathrm{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega)) \quad\text{weak}. \tag{2.19e}$$

Define the set $Q_R \overset{\text{def}}{=} \{x : x_1 < 0,\ |x'|\le R\}\times[0,R]$. Thanks to the classical compactness properties of injections of Sobolev spaces on bounded open sets, we see that for all $R>0$, the restrictions of u^ϵ and ∇u^ϵ to Q_R converge strongly to their

respective limits in $\mathrm{L}^2(Q_R)$; therefore, we can pass to the limit in all the terms of (2.18) except possibly the first two terms.

Let us prove that u_t is continuous from $[0,\infty)$ to $\mathrm{L}^2(\Omega)$ equipped with the weak topology: we infer from the estimates of Lemma 2.5 that for all $\beta > 0$, u^ϵ_{tt} restricted to $x_1 < -\beta$ is bounded in $\mathrm{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2((-\infty,-\beta]\times\Sigma))$; therefore, u^ϵ_t converges to a function u_t whose restriction to $x_1 < -\beta$ is continuous from $[0,\infty)$ to $\mathrm{L}^2((-\infty,-\beta]\times\Sigma)$. Let $t_j \in [0,\infty)$ be a sequence converging to $t_\infty < \infty$; as u_t belongs to $\mathrm{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$, we may extract a subsequence, still denoted by t_j, such that

$$u^\epsilon_t(\cdot,t_j) \rightharpoonup z \quad \text{in} \quad \mathrm{L}^2(\Omega) \quad \text{weak}.$$

But since, for all $\beta > 0$,

$$u^\epsilon_t(\cdot,t_j)1_{\{x_1<-\beta\}} \to u_t(\cdot,t_\infty)1_{\{x_1<-\beta\}} \quad \text{in} \quad \mathrm{L}^2(\Omega),$$

we see that z must coincide with $u_t(\cdot,t_\infty)$ and that the whole sequence converges strongly to $u_t(\cdot,t_\infty)$; this proves that u_t is continuous from $[0,\infty)$ to $\mathrm{L}^2(\Omega)$ weak.

Let us prove now that $u^\epsilon_t(\cdot,t)$ converges weakly to $u_t(\cdot,t)$ for all $t > 0$: let γ be an arbitrary positive number; let z belong to $\mathrm{L}^2(\Omega)$; and denote by C_1 an upper bound for $|u^\epsilon_t|_{\mathrm{L}^\infty([0,T];\mathrm{L}^2(\Omega))}$ with T fixed. We choose β so small that

$$\left(\int_{-\beta<x_1<0} |z|^2\,\mathrm{d}x\right)^{1/2} \le \frac{\gamma}{4C_1};$$

then, for $t \in [0,T]$,

$$\begin{aligned}\left|\int_\Omega (u^\epsilon_t(\cdot,t)-u_t(\cdot,t))z\,\mathrm{d}x\right| &\le \left|\int_{x_1<-\beta} (u^\epsilon_t(\cdot,t)-u_t(\cdot,t))z\,\mathrm{d}x\right| \\ &+ \left(\int_{-\beta<x_1<0} |z|^2\,\mathrm{d}x\right)^{1/2}\left(\int_{-\beta<x_1<0} |u^\epsilon_t(\cdot,t)-u_t(\cdot,t)|^2\,\mathrm{d}x\right)^{1/2}.\end{aligned} \tag{2.20}$$

By definition of C_1, the second term on the right-hand side of (2.20) is estimated by $C_1\gamma/(2C_1) = \gamma/2$. As $u^\epsilon_t|_{(-\infty,-\beta)\times I_T}$ is bounded in $\mathrm{H}^1((-\infty,-\beta)\times I_T)$, we see that

$$\int_{-\infty}^{-\beta}\int_\Sigma u^\epsilon_t z\,\mathrm{d}x \quad \text{converges} \quad \text{to} \quad \int_{-\infty}^{-\beta}\int_\Sigma u_t z\,\mathrm{d}x$$

uniformly with respect to $t \in [0,T]$. It suffices therefore to choose ϵ so small that the first term on the right-hand side of (2.20) is estimated by $\gamma/2$. This proves that the convergence of $\int_\Omega u^\epsilon_t z\,\mathrm{d}x$ to $\int_\Omega u_t z\,\mathrm{d}x$ is uniform on compact sets in time. In particular, as ϵ tends to 0, it is plain that for all $\tau > 0$,

$$\int_\Omega u^\epsilon_t\varphi(v-u^\epsilon)\,\mathrm{d}x \to \int_\Omega u_t\varphi(v-u)\,\mathrm{d}x.$$

Let us turn now to the term

$$\int_{Q_\tau} u^\epsilon_t(\varphi_t(v-u^\epsilon)+\varphi(v_t-u^\epsilon_t))\,\mathrm{d}x\,\mathrm{d}t.$$

It is clear that

$$\int_{Q_\tau} u^\epsilon_t(\varphi_t(v-u^\epsilon)+\varphi v_t)\,\mathrm{d}x\,\mathrm{d}t \to \int_{Q_\tau} u_t(\varphi_t(v-u)+\varphi v_t)\,\mathrm{d}x\,\mathrm{d}t.$$

There remains to prove the convergence

$$\int_{Q_\tau} |u_t^\epsilon|^2 \varphi \,\mathrm{d}x\,\mathrm{d}t \to \int_{Q_\tau} |u_t|^2 \varphi \,\mathrm{d}x\,\mathrm{d}t.$$

We observe that

$$\int_{Q_\tau} |u_t^\epsilon - u_t|^2 \varphi \,\mathrm{d}x\,\mathrm{d}t \le \int_0^\tau \int_\Sigma \int_{x_1 \le -\beta} |u_t^\epsilon - u_t|^2 \varphi \,\mathrm{d}x\,\mathrm{d}t + \int_0^\tau \int_\Sigma \int_{-\beta \le x_1 \le 0} |u_t^\epsilon - u_t|^2 \varphi \,\mathrm{d}x\,\mathrm{d}t.$$

Let γ be any positive number. One can deduce from the estimates of $|u_t^\epsilon|_{\mathrm{L}^2(I_\tau)}$ and $|\nabla u_t^\epsilon|_{\mathrm{L}^2(Q_\tau)}$ that there exists a constant C_2 independent from ϵ such that

$$|u_t^\epsilon(x_1, \cdot, \cdot)|_{\mathrm{L}^2(\Sigma \times (0,\tau))} \le C_2.$$

Therefore,

$$\int_0^\tau \int_\Sigma \int_{-\beta \le x_1 \le 0} |u_t^\epsilon - u_t|^2 \varphi \,\mathrm{d}x\,\mathrm{d}t \le C_2^2 \beta.$$

We choose β so small that $C_2^2 \beta \le \gamma/2$; then we know from the estimates of Lemmas 2.2 and 2.5 that the restriction of u^ϵ to $\{x_1 < -\beta\}$ intersected with a ball containing the support of φ is bounded in H^2 of that set; therefore, for ϵ small enough,

$$\int_{Q_\tau} |u_t^\epsilon - u_t|^2 \varphi \,\mathrm{d}x\,\mathrm{d}t \le \frac{\gamma}{2},$$

and the convergence of the first two terms of (2.18) is proved. We observe now that since u, u_t, ∇u, and ∇u_t belong to $\mathrm{L}^2_{\mathrm{loc}}([0,\infty); \mathrm{L}^2(\Omega))$, we may replace φ by φ_R in the variational inequality where φ_R is equal to 1 over the set Q_R and vanishes outside of Q_{R+1}. It is plain that as $R \to \infty$ all the terms in (2.18) converge to their limit; thus we have proved the existence of the desired weak solution. ∎

Remark 2.7. Nothing is known about uniqueness.

2.4. Auxiliary results on the damped wave equation with Dirichlet boundary conditions. We establish a priori estimates on the damped wave equation with Dirichlet boundary conditions. These estimates will enable us to give some properties on the trace spaces which we use in the next subsection.

LEMMA 2.8. *Assume u_0 belongs to $\mathrm{H}^{5/2}(\Omega)$; then, there exists a function $z \in \mathrm{H}^3(\Omega \times [0,\infty))$ with compact support in t such that the trace of z on Ω is equal to u_0.*

Proof. We extend u_0 into a function belonging to $\mathrm{H}^{5/2}(\mathbb{R}^d)$: as the boundary of Ω is smooth, this extension is a consequence of classical results on Sobolev spaces. Then there exists a function Z belonging to $\mathrm{H}^3(\mathbb{R}^d \times [0,\infty))$ whose trace is u_0. It suffices now to select a cut-off function $\varphi \in \mathrm{C}^\infty([0,\infty))$ which is equal to 1 on $[0,1]$ and to 0 on $[2,\infty)$ and to define z as the restriction of φZ to $\Omega \times [0,\infty)$. ∎

LEMMA 2.9. *Assume u_0 belongs to $\mathrm{H}^{5/2}(\Omega)$, u_1 belongs to $\mathrm{H}^1(\Omega)$, and f belongs to $\mathrm{L}^2_{\mathrm{loc}}([0,\infty); \mathrm{L}^2(\Omega))$. Define z as in Lemma 2.8, and let $\bar{u}$ be the solution of (1.1) with the initial data (1.2) and boundary condition $\bar{u}(0,\cdot,\cdot) = z(0,\cdot,\cdot)$. Then the trace $\bar{g} \overset{\mathrm{def}}{=} -(\bar{u}_{x_1} + \alpha \bar{u}_{x_1 t})(0,\cdot,\cdot)$ is well defined and belongs to $\mathrm{L}^2_{\mathrm{loc}}([0,\infty); \mathrm{L}^2(\mathbb{R}^{d-1}))$. Moreover, if f is compactly supported in time,*

$$\int_0^\tau |\bar{g}(\cdot,t)|^2_{\mathrm{L}^2(\Sigma)} \,\mathrm{d}t$$

increases at most polynomially with respect to τ.

Proof. The function $\zeta \overset{\text{def}}{=} \bar{u} - z$ satisfies the equation

$$\zeta_{tt} - \Delta\zeta - \alpha\Delta\zeta_t = F, \quad x \in \Omega, \quad t > 0, \tag{2.21}$$

where $F \overset{\text{def}}{=} f - z_{tt} + \Delta z + \alpha\Delta z_t$, with initial data

$$\zeta(\cdot, 0) = 0 \quad \text{and} \quad \zeta_t(\cdot, 0) = u_1$$

and the Dirichlet boundary condition $\zeta(0, \cdot, \cdot) = 0$. If we multiply (2.21) by ζ_t and integrate, and if we suppose that f and F are compactly supported in time, we may easily deduce that ζ_t, $\nabla\zeta$ are bounded in $\mathrm{L}^\infty_{\mathrm{loc}}([0,\infty); \mathrm{L}^2(\Omega))$ and $\nabla\zeta_t$ is bounded in $\mathrm{L}^2_{\mathrm{loc}}([0,\infty); \mathrm{L}^2(\Omega))$. In order to get more information, we multiply (2.21) by $\Delta\zeta_t$; since the boundary term vanishes, we get immediately the identity

$$\begin{aligned}&\alpha \int_{Q_\tau} |\Delta\zeta_t|^2 \,\mathrm{d}x\,\mathrm{d}t + \frac{1}{2}\int_\Omega |\Delta\zeta(\cdot,\tau)|^2 \,\mathrm{d}x + \int_\Omega |\nabla\zeta_t(\cdot,\tau)|^2 \,\mathrm{d}x \\ &= \int_\Omega |\nabla\zeta_t(\cdot,0)|^2 \,\mathrm{d}x - \int_{Q_\tau} F\Delta\zeta_t \,\mathrm{d}x\,\mathrm{d}t.\end{aligned}$$

The product $F\Delta\zeta_t$ can be estimated by $\alpha|\Delta\zeta_t|^2/2 + |F|^2/(2\alpha)$; then $\Delta\zeta_t$ is bounded in $\mathrm{L}^2_{\mathrm{loc}}([0,\infty); \mathrm{L}^2(\Omega))$, and $\Delta\zeta$ and $\nabla\zeta_t$ are bounded in $\mathrm{L}^\infty_{\mathrm{loc}}([0,\infty); \mathrm{L}^2(\Omega))$. In particular, if the support in time of F is bounded, $\Delta\zeta_t$ is bounded in $\mathrm{L}^2([0,\infty); \mathrm{L}^2(\Omega))$ and $\Delta\zeta$ and $\nabla\zeta_t$ are bounded in $\mathrm{L}^\infty([0,\infty); \mathrm{L}^2(\Omega))$. Hence $\zeta_{x_1 t}(0,\cdot,\cdot)$ and $\zeta_{x_1}(0,\cdot,\cdot)$ belong, respectively, to $\mathrm{L}^2_{\mathrm{loc}}([0,\infty); \mathrm{H}^{1/2}(\mathbb{R}^{d-1}))$ and to $\mathrm{L}^\infty_{\mathrm{loc}}([0,\infty); \mathrm{H}^{1/2}(\mathbb{R}^{d-1}))$, and if the support in time of f is bounded, the local character of these spaces may be removed. ∎

2.5. Regularity of the trace. We characterize the trace spaces using Fourier analysis, and we prove that u is a strong solution of (1.1)–(1.3). Here, we mean by strong solution that all the traces can be defined.

Let ν be a positive number. Denote by $v^\epsilon \overset{\text{def}}{=} \mathrm{e}^{-\nu t}(u^\epsilon - \bar{u})$ a solution of

$$(\nu + \partial_t)^2 v^\epsilon - (1 + \alpha(\nu + \partial_t))\Delta v^\epsilon = 0, \quad x \in \Omega, \quad t > 0, \tag{2.22a}$$

$$(1 + \alpha(\nu + \partial_t))v^\epsilon_{x_1}(0,\cdot,\cdot) = \mathrm{e}^{-\nu t}\bar{g} - (v^\epsilon(0,\cdot,\cdot) + \mathrm{e}^{-\nu t}\bar{u}(0,\cdot,\cdot))^-/\epsilon, \tag{2.22b}$$

$$v^\epsilon(\cdot, 0) = 0 \quad \text{and} \quad v^\epsilon_t(\cdot, 0) = 0. \tag{2.22c}$$

We denote by $\xi \overset{\text{def}}{=} (\xi_2, \dots, \xi_d)^{\mathsf{T}}$ and ω, respectively, the dual variables to $x' \overset{\text{def}}{=} (x_2, \dots, x_d)^{\mathsf{T}}$ and t. The Fourier transform of $u(0, x', t)$ is $\widehat{u}(0, \xi, \omega)$, where the convention for the Fourier transform is

$$\widehat{u}(0,\xi,\omega) = \int_{\mathbb{R}^d} \mathrm{e}^{-i(\xi\cdot x' + \omega t)} u(0, x', t) \,\mathrm{d}x'\,\mathrm{d}t.$$

Then $u(0, x', t)$ belongs to the Sobolev space $\mathrm{H}^{a,b}_{\mathrm{loc}}(\mathbb{R}^{d-1}\times[0,\infty))$, $(a, b) \in \mathbb{R}^2$, iff $|\xi|^a\widehat{u}(0,\xi,\omega)$ and $|\omega|^b\widehat{u}(0,\xi,\omega)$ belong to $\mathrm{L}^2(\mathbb{R}^d)$.

We apply a partial Fourier transform in the tangential variable to (2.22a), and we get the following differential equation:

$$\widehat{v}^\epsilon_{x_1 x_1} = \left(|\xi|^2 + \frac{(\nu + i\omega)^2}{1 + \alpha(\nu + i\omega)}\right)\widehat{v}^\epsilon. \tag{2.23}$$

Define $\widehat{\lambda}$ to be

$$\widehat{\lambda}(\xi,\omega) \stackrel{\text{def}}{=} \sqrt{|\xi|^2 + \frac{(\nu+i\omega)^2}{1+\alpha(\nu+i\omega)}};$$

thus $\widehat{\lambda}$ is holomorphic in the lower half-plane $\Im(\omega) < 0$ and $\Re\widehat{\lambda} \geq 0$ for $\Im(\omega) = 0$. The general solution of (2.23) is given by $\widehat{a}^\epsilon \mathrm{e}^{\widehat{\lambda}x_1} + \widehat{b}^\epsilon \mathrm{e}^{-\widehat{\lambda}x_1}$; since we performed a Fourier transform on v^ϵ, we assumed implicitly that v^ϵ and $\widehat{v}^\epsilon$ are tempered, respectively, in (x',t) and (ξ,ω). We remark that the term $\widehat{b}^\epsilon \mathrm{e}^{-\widehat{\lambda}x_1}$ can be tempered only if $\widehat{b}^\epsilon$ decays at infinity very quickly, and since this must be true for all x_1, it implies that $\widehat{b}^\epsilon$ vanishes; the proof is similar to that given in [PeS02]. We deduce that the solution of (2.23) is $\widehat{a}^\epsilon \mathrm{e}^{\widehat{\lambda}x_1}$. In particular,

$$((1+\alpha(\nu+\partial_t))v^\epsilon_{x_1})\widehat{\ }(0,\xi,\omega) = \widehat{\lambda}_1 \widehat{v}^\epsilon(0,\xi,\omega), \tag{2.24}$$

where $\widehat{\lambda}_1 \stackrel{\text{def}}{=} (1+\alpha(\nu+i\omega))\widehat{\lambda}$. Define

$$g(x',t) \stackrel{\text{def}}{=} \mathrm{e}^{-\nu t}\bar{g}(x',t) \quad \text{and} \quad h(x',t) \stackrel{\text{def}}{=} \mathrm{e}^{-\nu t}\bar{u}(0,x',t).$$

If we let $w^\epsilon(x',t)$ be the trace $v^\epsilon(0,x',t)$, (2.22) can now be written as

$$\lambda_1 * w^\epsilon = g + (w^\epsilon + h)^-/\epsilon, \tag{2.25}$$

where w^ϵ vanishes for all $t \leq 0$.

Remark 2.10. It is clear that $\widehat{\lambda}$ is a holomorphic function in $\Im(\omega) < 0$, and thus we may deduce that λ_1 is a causal distribution.

LEMMA 2.11. *Let u^ϵ be the solution of* (2.1)–(2.3). *Then we may extract a subsequence, still denoted by u^ϵ, such that*

$$u^\epsilon(0,\cdot,\cdot) \rightharpoonup u(0,\cdot,\cdot) \quad \textit{weakly} \quad \textit{in} \quad \mathrm{H}^{1/2,5/4}_{\mathrm{loc}}(\mathbb{R}^{d-1}\times[0,\infty)).$$

Moreover, u is a strong solution of (1.1)–(1.3).

Proof. Formally, we multiply (2.25) by $\alpha(\nu w^\epsilon + w^\epsilon_t) + w^\epsilon$, and we estimate the pseudodifferential term in the Fourier variable; we obtain

$$\begin{aligned} &\frac{1}{(2\pi)^d}\,\Re\int_{\mathbb{R}^d} \widehat{\lambda}_1 \widehat{w}^\epsilon \overline{(1+\alpha(\nu+i\omega))\widehat{w}^\epsilon}\,\mathrm{d}\omega\,\mathrm{d}\xi \\ &= \frac{1}{(2\pi)^d}\,\Re\int_{\mathbb{R}^d} \widehat{g}\overline{(1+\alpha(\nu+i\omega))\widehat{w}^\epsilon}\,\mathrm{d}\omega\,\mathrm{d}\xi \\ &+ \frac{1}{\epsilon}\int_0^\infty \int_{\mathbb{R}^{d-1}} (w^\epsilon+h)^-(1+\alpha(\nu+\partial_t))w^\epsilon\,\mathrm{d}x'\,\mathrm{d}t. \end{aligned} \tag{2.26}$$

Since $(u^\epsilon(0,\cdot,\cdot))^-/\sqrt{\epsilon}$ is bounded in $\mathrm{L}^\infty_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\mathbb{R}^{d-1}))$, the absolute value of the second integral in the right-hand side of (2.26) is bounded, and we infer that there exists $C_1 > 0$ such that

$$\Re\int_{\mathbb{R}^d} \widehat{\lambda}_1 |\widehat{w}^\epsilon|^2 \overline{(1+\alpha(\nu+i\omega))}\,\mathrm{d}\omega\,\mathrm{d}\xi \leq C_1 + \Re\int_{\mathbb{R}^d} \widehat{g}\overline{(1+\alpha(\nu+i\omega))\widehat{w}^\epsilon}\,\mathrm{d}\omega\,\mathrm{d}\xi. \tag{2.27}$$

On the other hand, we have

$$\Re\widehat{\lambda}^2 = |\xi|^2 + \frac{\nu^2(1+\alpha\nu) + (-1+\alpha\nu)\omega^2}{|1+\alpha(\nu+i\omega)|^2} \quad \text{and} \quad \Im\widehat{\lambda}^2 = \frac{2\nu\omega + \alpha\omega(\nu^2+\omega^2)}{|1+\alpha(\nu+i\omega)|^2}.$$

We may choose ν such that $\nu\alpha = 1$; we get then

$$\Re\widehat{\lambda}^2 = |\xi|^2 + \frac{2}{\alpha^2|2+i\alpha\omega|^2} \quad \text{and} \quad \Im\widehat{\lambda}^2 = \frac{\omega(3+\alpha^2\omega^2)}{\alpha|2+i\alpha\omega|^2}. \tag{2.28}$$

Therefore, we infer that

$$\arg\widehat{\lambda} = \frac{1}{2}\arctan\left(\frac{|\xi|^2|2+i\alpha\omega|^2+2}{\alpha\omega(3+\alpha^2\omega^2)}\right).$$

According to (2.28), $\arg\widehat{\lambda}$ belongs to $[0,\pi/4]$, and since $\widehat{\lambda}$ is never equal to zero, we get for $|\xi|+|\omega| \gg 1$ the following inequality:

$$\Re\widehat{\lambda} \geq C\big(1+|\xi|+\sqrt{|\omega|}\big). \tag{2.29}$$

Therefore, we obtain

$$C\int_{\mathbb{R}^d} |2+i\alpha\omega|^2\big(1+|\xi|+\sqrt{|\omega|}\big)|\widehat{w}^\epsilon|^2\,\mathrm{d}\omega\,\mathrm{d}\xi \leq C_1 + \int_{\mathbb{R}^d} |2+i\alpha\omega||\widehat{g}||\widehat{w}^\epsilon|\,\mathrm{d}\omega\,\mathrm{d}\xi.$$

We estimate the product $|zy|$ by $|z|^2/(2\gamma)+\gamma|y|^2/2$, $\gamma>0$; we see that

$$\begin{aligned}&\Big(C-\frac{\gamma}{2}\Big)\int_{\mathbb{R}^d} |2+i\alpha\omega|^2\big(1+|\xi|+\sqrt{|\omega|}\big)|\widehat{w}^\epsilon|^2\,\mathrm{d}\omega\,\mathrm{d}\xi\\ &\leq C_1 + \frac{1}{2\gamma}\int_{\mathbb{R}^d}\frac{|\widehat{g}|^2}{1+|\xi|+\sqrt{|\omega|}}\,\mathrm{d}\omega\,\mathrm{d}\xi.\end{aligned} \tag{2.30}$$

We choose γ such that $\gamma < 2C$; since g belongs to $\mathrm{L}^2([0,\infty);\mathrm{H}^{1/2}(\mathbb{R}^{d-1}))$, then it is easy to deduce from (2.30) that $u^\epsilon(0,\cdot,\cdot)$ is bounded in $\mathrm{H}^{1/2,5/4}_{\mathrm{loc}}(\mathbb{R}^{d-1}\times[0,\infty))$. Moreover, it is clear that $(u_{x_1}+\alpha u_{x_1 t})(0,\cdot,\cdot)$ is bounded in $\mathrm{H}^{-1/2,-1/4}_{\mathrm{loc}}(\mathbb{R}^{d-1}\times[0,\infty))$. Therefore, all the traces are defined, and we may deduce that u is a strong solution of (2.1)–(2.3). ∎

Remark 2.12. We have been unable to establish that the energy loss is purely viscous as in the case of the one dimensional viscously damped wave equation on the half-line and with unilateral boundary conditions [PeS02, PeS08].

3. The evolution of a Kelvin–Voigt material with Signorini boundary conditions. As for the damped wave equation with unilateral boundary conditions, a priori estimates on the penalized problem and care relative due to the unboundedness of Ω enable us to pass to the limit in the penalized variational formulation and to deduce the existence of a solution to (1.9). Korn's inequality plays an important role here. If we denote by $\bar{u}$ the solution of (1.9a) with initial data (1.9d) and Dirichlet boundary data at $x_1 = 0$, then we establish that the trace $-(a^0_{11kl}\varepsilon_{kl}(\bar{u}) + a^1_{11kl}\varepsilon_{kl}(\bar{u}_t))|_{\Sigma\times[0,\infty)}$ increases exponentially with time in $\mathrm{L}^2_{\mathrm{loc}}(\Sigma\times[0,\infty))$ and not polynomially as in the case of the damped wave equation with Dirichlet boundary conditions studied in subsection 2.4. We determine the trace spaces using analogous techniques already developed in section 2.5, but here we perform a Fourier transform in the tangential variables (x_2,x_3,t) and a Laplace transform in x_1.

3.1. The penalized problem. We approximate (1.9) as in section 2.1. More precisely, letting $r^+ \stackrel{\mathrm{def}}{=} \max(r,0)$, we replace u by u^ϵ which is a solution of the following penalized problem:

$$\rho u^\epsilon_{tt} - A^0 u^\epsilon - A^1 u^\epsilon_t = f, \quad x\in\Omega, \quad t>0, \tag{3.1}$$

with initial data

$$u^\epsilon(\cdot,0) = v_0 \quad \text{and} \quad u_t^\epsilon(\cdot,0) = v_1 \tag{3.2}$$

and boundary conditions

$$a^0_{11kl}\varepsilon_{kl}(u^\epsilon) + a^1_{11kl}\varepsilon_{kl}(u_t^\epsilon) = -(u_1^\epsilon)^+/\epsilon, \tag{3.3a}$$

$$a^0_{12kl}\varepsilon_{kl}(u^\epsilon) + a^1_{12kl}\varepsilon_{kl}(u_t^\epsilon) = 0, \quad \text{and} \quad a^0_{13kl}\varepsilon_{kl}(u^\epsilon) + a^1_{13kl}\varepsilon_{kl}(u_t^\epsilon) = 0. \tag{3.3b}$$

Recall that Q_τ and I_τ were defined by (2.4).

THEOREM 3.1. *Let* $\mathrm{W} \stackrel{\text{def}}{=} \{u \in \mathbf{H}^1_{\mathrm{loc}}([0,\infty)\times\Omega) : \nabla u_t \in \mathbf{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))\}$. *Then for each* $\epsilon > 0$ *there exists a unique weak solution* $u^\epsilon \in \mathrm{W}$ *of the problem* (3.1)–(3.3) *such that*

$$\begin{aligned} u^\epsilon &\in \mathbf{L}^\infty_{\mathrm{loc}}([0,\infty);\mathrm{H}^1(\Omega)),\\ u_t^\epsilon &\in \mathbf{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{H}^1(\Omega)),\\ u_{tt}^\epsilon &\in \mathbf{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega)), \end{aligned}$$

and for every $\tau \in (0,T)$ *and for all* $v \in \mathrm{W}$, *the following variational equality is satisfied:*

$$\begin{aligned} &\int_{Q_\tau} \rho u_{tt}^\epsilon \cdot v \,\mathrm{d}x\,\mathrm{d}t + \int_0^\tau \left(a^0(u^\epsilon, v) + a^1(u_t^\epsilon, v)\right)\mathrm{d}t\\ &+ \int_{I_\tau} \frac{(u_1^\epsilon)^+}{\epsilon} v_1 \,\mathrm{d}x'\,\mathrm{d}t \geq \int_{Q_\tau} f\cdot v\,\mathrm{d}x\,\mathrm{d}t. \end{aligned} \tag{3.4}$$

Proof. We leave the verification of the proof to the reader as it is analogous to the one developed in [Jar96]. ∎

3.2. Estimates on the penalized solution. We establish a priori estimates which are essential to prove the existence of a weak solution to (3.1)–(3.3). These estimates are obtained thanks to the techniques already developed in section 2.2 for the damped wave equation and to Korn's inequality.

LEMMA 3.2. *Assume that* f *belongs to* $\mathbf{L}^2_{\mathrm{loc}}\big([0,\infty);\mathrm{L}^2(\Omega)\big)$, v_0 *to* $\mathbf{H}^1(\Omega)$, *and* v_1 *to* $\mathbf{L}^2(\Omega)$. *Then* u_t^ϵ *and* ∇u^ϵ *are bounded in* $\mathbf{L}^\infty_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$, ∇u_t^ϵ *is bounded in* $\mathbf{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$, *and* $(u_1^\epsilon(0,\cdot,\cdot))^+/\sqrt{\epsilon}$ *is bounded in* $\mathrm{L}^\infty_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\mathbb{R}^{d-1}))$, *independently of* $\epsilon > 0$.

Proof. These estimates are a simple application of the Gronwall lemma to the energy estimate. We multiply (3.1) by u_t^ε and integrate this expression over Q_τ to get

$$\begin{aligned} &\frac{1}{2}\int_\Omega \left(\rho|u_t^\epsilon|^2 + a^0_{ijkl}\varepsilon_{ij}(u^\epsilon)\varepsilon_{kl}(u^\epsilon)\right)\Big|_0^\tau \mathrm{d}x + \int_{Q_\tau} a^1_{ijkl}\varepsilon_{ij}(u_t^\epsilon)\varepsilon_{kl}(u_t^\epsilon)\,\mathrm{d}x\,\mathrm{d}t\\ &+ \frac{1}{2\epsilon}\int_\Sigma \left((u_1^\epsilon)^+\right)^2\Big|_0^\tau \mathrm{d}x' = \int_{Q_\tau} f\cdot u^\epsilon\,\mathrm{d}x\,\mathrm{d}t. \end{aligned} \tag{3.5}$$

According to Korn's inequality, it is possible to infer that there exist two positive constants C_1 and C_2 such that

$$\int_\Omega a^n_{ijkl}\varepsilon_{kl}(z)\varepsilon_{ij}(z)\,\mathrm{d}z \geq C_1\int_\Omega |\nabla z|^2\,\mathrm{d}z - C_2\int_\Omega |z|^2\,\mathrm{d}z, \quad n = 0,1.$$

As fu_t^ϵ can be estimated by $|f|^2/(2\gamma)+\gamma|u_t^\epsilon|^2/2$, $\gamma>0$, and using the above inequality, we deduce from (3.5) that

$$\begin{aligned}&\frac{1}{2}\int_\Omega \big(\rho|u_t^\epsilon|^2+C_1|\nabla u^\epsilon|^2\big)(\cdot,\tau)\,\mathrm{d}x+C_1\int_{Q_\tau}|\nabla u_t^\epsilon|^2\,\mathrm{d}x\,\mathrm{d}t+\frac{1}{2\epsilon}\int_\Sigma\big((u_1^\epsilon)^+\big)\big|_0^\tau\,\mathrm{d}x'\\&\le\frac{C_2}{2}\int_\Omega|u^\epsilon(\cdot,\tau)|^2\,\mathrm{d}x+\Big(C_2+\frac{\gamma}{2}\Big)\int_{Q_\tau}|u_t^\epsilon|^2\,\mathrm{d}x\,\mathrm{d}t+\frac{1}{2\gamma}\int_{Q_\tau}|f|^2\,\mathrm{d}x\,\mathrm{d}t\\&+\frac{1}{2}\int_\Omega\big(\rho|v_1|^2+a^0_{ijkl}\varepsilon_{ij}(v_0)\varepsilon_{kl}(v_0)\big)\,\mathrm{d}x.\end{aligned}$$

A classical Gronwall lemma enables us to deduce that u_t^ϵ and ∇u^ϵ are bounded in the space $\mathbf{L}^\infty_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$, ∇u_t^ϵ is bounded in $\mathbf{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$, and $(u_1^\epsilon(0,\cdot,\cdot))^+/\sqrt{\epsilon}$ is bounded in $\mathrm{L}^\infty_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\mathbb{R}^{d-1}))$. ∎

Remark 3.3. If we suppose that f vanishes for large t, then, independently of $\epsilon>0$,

$$\operatorname*{ess\,sup}_{0\le t\le T}|u^\epsilon(\cdot,t)|_{\mathrm{H}^1}\le C(1+T)$$

and

$$\left(\int_0^T|u_t^\epsilon(\cdot,t)|^2_{\mathrm{H}^1}\,\mathrm{d}t\right)^{1/2}\le C(1+T).$$

These properties can be proved using the arguments given in the proof of Lemma 3.2, with the origin of time moved to T if f vanishes for $t\ge T$; since the integral involving f vanishes, the conclusion is clear.

LEMMA 3.4. *Assume that f belongs to $\mathbf{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$, v_0 to $\mathbf{H}^1(\Omega)$, and v_1 to $\mathbf{L}^2(\Omega)$. Then, independently of $\epsilon>0$, the trace $(u_1^\epsilon(0,\cdot,\cdot))^+/\epsilon$ is bounded in the space of measures on I_T.*

Proof. Let φ be a cut-off function which belongs to $\mathbf{C}^1(\mathbb{R}^{d-1})$, is equal to 1 in the sphere of center 0 and radius $R>0$, and vanishes outside of a sphere of radius $R+1$. We multiply (3.1) by φ and we integrate over Q_τ; due to the boundary conditions (3.3), we obtain

$$\begin{aligned}&\int_\Omega\rho u_t^\epsilon\cdot\varphi\big|_0^\tau\,\mathrm{d}x+\frac{1}{\epsilon}\int_{I_\tau}(u_1^\epsilon)^+\varphi_1\,\mathrm{d}x'\,\mathrm{d}t+\int_{Q_\tau}\sigma^0_{ij}(u^\epsilon)\varepsilon_{ij}(\varphi)\,\mathrm{d}x\,\mathrm{d}t\\&+\int_{Q_\tau}\sigma^1_{ij}(u_t^\epsilon)\varepsilon_{ij}(\varphi)\,\mathrm{d}x\,\mathrm{d}t=\int_{Q_\tau}f\cdot\varphi\,\mathrm{d}x\,\mathrm{d}t.\end{aligned}$$

As the product $|zy|$ can be estimated by $|z|^2/2+|y|^2/2$, we get the following inequality:

$$\begin{aligned}&\frac{1}{\epsilon}\int_{I_\tau}(u_1^\epsilon)^+\varphi_1\,\mathrm{d}x'\,\mathrm{d}t\le\frac{\rho}{2}\int_\Omega\big(|u_t^\epsilon(\cdot,\tau)|^2+|v_1|^2\big)\,\mathrm{d}x+\rho\int_\Omega|\varphi|^2\,\mathrm{d}x\\&+\int_{Q_\tau}\big|\big(\sigma^0_{ij}(u^\epsilon)+\sigma^1_{ij}(u_t^\epsilon)\big)\varepsilon_{ij}(\varphi)\big|\,\mathrm{d}x\,\mathrm{d}t+\int_{Q_\tau}|f\cdot\varphi|\,\mathrm{d}x\,\mathrm{d}t.\end{aligned}\tag{3.6}$$

We may deduce that the right-hand side of (3.6) is bounded using the Lemma 3.2. Since $(u_1^\epsilon(0,\cdot,\cdot))^+$ is nonnegative, the conclusion is clear. ∎

LEMMA 3.5. *Assume that f, v_0, and v_1 belong, respectively, to $\mathbf{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$, $\mathbf{H}^2(\Omega)$, and $\mathbf{L}^2(\Omega)$. Then A^0u^ϵ and A^1u^ϵ are bounded in $\mathbf{L}^\infty_{\mathrm{loc}}([0,\infty),\mathrm{L}^2(\Omega))$, independently of $\epsilon>0$.*

Proof. Once again, we use energy techniques, but now we multiply relation (3.1) by A^1u^ϵ and we integrate over Q_τ to obtain

$$\begin{aligned}\frac{1}{2}\int_\Omega |A^1u^\epsilon(\cdot,\tau)|^2\,\mathrm{d}x &= \frac{1}{2}\int_\Omega |A^1v_0|^2\,\mathrm{d}x + \int_{Q_\tau}\rho u^\epsilon_{tt}\cdot(A^1u^\epsilon)\,\mathrm{d}x\,\mathrm{d}t\\ &- \int_{Q_\tau}(A^0u^\epsilon)\cdot(A^1u^\epsilon)\,\mathrm{d}x\,\mathrm{d}t - \int_{Q_\tau} f\cdot(A^1u^\epsilon)\,\mathrm{d}x\,\mathrm{d}t.\end{aligned} \tag{3.7}$$

Notice that

$$\begin{aligned}\int_{Q_\tau}\rho u^\epsilon_{tt}\cdot(A^1u^\epsilon)\,\mathrm{d}x\,\mathrm{d}t &= \rho\int_\Omega u^\epsilon_t\cdot(A^1u^\epsilon)\big|_0^\tau\,\mathrm{d}x\\ &- \rho\int_{I_\tau} u^\epsilon_{1,t}\sigma^1_{1j}(u^\epsilon)\,\mathrm{d}x'\,\mathrm{d}t + \int_{Q_\tau} a^1_{ijkl}\varepsilon_{ij}(u^\epsilon_t)\varepsilon_{kl}(u^\epsilon_t)\,\mathrm{d}x\,\mathrm{d}t.\end{aligned} \tag{3.8}$$

Carrying (3.8) into (3.7) and using the boundary conditions (3.3), we obtain

$$\begin{aligned}\frac{1}{2}\int_\Omega |A^1u^\epsilon(\cdot,\tau)|^2\,\mathrm{d}x &= \frac{1}{2}\int_\Omega |A^1v_0|^2\,\mathrm{d}x - \int_{Q_\tau}(A^0u^\epsilon)\cdot(A^1u^\epsilon)\,\mathrm{d}x\,\mathrm{d}t\\ &- \int_{Q_\tau} f\cdot(A^1u^\epsilon)\,\mathrm{d}x\,\mathrm{d}t + \rho\int_\Omega u^\epsilon_t\cdot(A^1u^\epsilon)\big|_0^\tau\,\mathrm{d}x + \frac{\rho}{\epsilon}\int_{I_\tau} u^\epsilon_{1,t}(u^\epsilon_1)^+\,\mathrm{d}x'\,\mathrm{d}t\\ &+ \rho\int_{I_\tau} u^\epsilon_{1,t}\sigma^0_{1j}(u^\epsilon)\,\mathrm{d}x'\,\mathrm{d}t + \int_{Q_\tau} a^1_{ijkl}\varepsilon_{ij}(u^\epsilon_t)\varepsilon_{kl}(u^\epsilon_t)\,\mathrm{d}x\,\mathrm{d}t.\end{aligned} \tag{3.9}$$

On the other hand, we observe that

$$\int_{I_\tau}|\sigma^0_{1j}(u^\epsilon)|^2\,\mathrm{d}x'\,\mathrm{d}t \le C\left(\int_{Q_\tau}|u^\epsilon|^2\,\mathrm{d}x\,\mathrm{d}t + \int_{Q_\tau}|A^1u^\epsilon|^2\,\mathrm{d}x\,\mathrm{d}t\right), \tag{3.10}$$

and for all v belonging to $\mathbf{H}^1(\Omega)$ and A^1v belonging to $\mathbf{L}^2(\Omega)$, we get

$$|A^0v|_{\mathrm{L}^2(\Omega)} \le C|v|_{\mathrm{L}^2(\Omega)} + |A^1v|_{\mathrm{L}^2(\Omega)}. \tag{3.11}$$

Define

$$F(t) \stackrel{\mathrm{def}}{=} \int_\Omega |A^1u^\epsilon(\cdot,t)|^2\,\mathrm{d}x. \tag{3.12}$$

According to (3.10)–(3.12) and since $u^\epsilon_t\cdot(A^1u^\epsilon)$ can be estimated by $|u^\epsilon_t|^2/(2\gamma) + \gamma|A^1u^\epsilon|^2/2$, $\gamma>0$, it is possible to infer from (3.9) the following inequality:

$$\begin{aligned}&\left(\frac{1}{2}-\frac{\rho\gamma}{2}\right)F(\tau) \le \frac{1}{2}F(0) + (2+C)\int_0^\tau F(t)\,\mathrm{d}t + \frac{1}{2}\int_{Q_\tau}|f|^2\,\mathrm{d}x\,\mathrm{d}t\\ &+ \frac{\rho}{2\gamma}\int_\Omega |u^\epsilon_t(\cdot,\tau)|^2\,\mathrm{d}x + \rho\int_\Omega |v_1\cdot(A^1v_0)|\,\mathrm{d}x + \int_{Q_\tau} a^1_{ijkl}\varepsilon_{ij}(u^\epsilon_t)\varepsilon_{kl}(u^\epsilon_t)\,\mathrm{d}x\,\mathrm{d}t\\ &+ \frac{\rho}{2\varepsilon}\int_\Sigma \big(u^\epsilon_1(\cdot,\tau)\big)^+\,\mathrm{d}x' + (1+C)\int_{Q_\tau}|u^\epsilon|^2\,\mathrm{d}x\,\mathrm{d}t + \int_{I_\tau}|u^\epsilon_t|^2\,\mathrm{d}x'\,\mathrm{d}t.\end{aligned}$$

If we choose γ such that $\rho\gamma<1$, we may infer using Lemma 3.2 and a classical Gronwall inequality that F is bounded in $\mathrm{L}^\infty_{\mathrm{loc}}([0,\infty))$. This proves the lemma. ∎

Remark 3.6. If we suppose that f vanishes for t large, then, independently of $\epsilon > 0$, $A^0 u^\epsilon$ and $A^1 u^\epsilon$ increase polynomially. These properties can be proved using the arguments given in Remark 3.3.

Let us turn now to interior estimates.

LEMMA 3.7. *Assume that f belongs to $\mathbf{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$, v_0 to $\mathbf{H}^2(\Omega)$, and v_1 to $\mathbf{L}^2(\Omega)$. Then for all $\beta > 0$, u^ϵ_{tt} and $A^1 u^\epsilon_t$ are bounded in $\mathbf{L}^2([0,\infty);\mathrm{L}^2((-\infty,-\beta)\times\Sigma))$, independently of $\epsilon > 0$.*

Proof. As for the proof of Lemma 2.5, we use a truncation function which enables us to forget about the strongly nonlinear boundary conditions. More precisely, we multiply u^ϵ by a cut-off function $\varphi(x_1) \in \mathrm{C}^\infty([0,\infty))$ which is equal to 0 on $x_1 \le -\beta$ and to 1 on $x_1 \ge -\beta/2$, $\beta > 0$. Define

$$v^\epsilon(x_1,\cdot,\cdot) \stackrel{\mathrm{def}}{=} \varphi(x_1) u^\epsilon(x_1,\cdot,\cdot). \tag{3.13}$$

The derivatives of v^ϵ are given by

$$v^\epsilon_{tt} = \varphi u^\epsilon_{tt}, \tag{3.14a}$$

$$\varepsilon_{kl,x_j}(v^\epsilon) = \varphi\varepsilon_{kl,x_j}(u^\epsilon) + 2\varphi_{x_1}\varepsilon_{kl}(u^\epsilon) + \varphi_{x_1x_1}u^\epsilon_k, \tag{3.14b}$$

$$\varepsilon_{kl,x_j}(v^\epsilon_t) = \varphi\varepsilon_{kl,x_j}(u^\epsilon_t) + 2\varphi_{x_1}\varepsilon_{kl}(u^\epsilon_t) + \varphi_{x_1x_1}u^\epsilon_{k,t}. \tag{3.14c}$$

Notice that thanks to relations (3.1) and (3.14), we have

$$v^\epsilon_{tt} - A^0 v^\epsilon - A^1 v^\epsilon = \widetilde{g}^\epsilon, \tag{3.15}$$

where $\widetilde{g}^\varepsilon_i \stackrel{\mathrm{def}}{=} \varphi f_i - 2\varphi_{x_1}(a^0_{ijkl}\varepsilon_{kl}(u^\epsilon) + a^1_{ijkl}\varepsilon_{kl}(u^\epsilon_t)) - \varphi_{x_1x_1}(a^0_{ijkl}u^\epsilon_k + a^1_{ijkl}u^\epsilon_{k,t})$. Thanks to Lemma 3.2, we deduce that $\widetilde{g}^\varepsilon$ is bounded in $\mathbf{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$. Define

$$w^\epsilon \stackrel{\mathrm{def}}{=} v^\epsilon_t \quad \text{and} \quad g^\epsilon \stackrel{\mathrm{def}}{=} \widetilde{g}^\epsilon + A^0 v^\epsilon. \tag{3.16}$$

We substitute (3.16) into (3.15) and obtain

$$w^\epsilon_t - A^1 w^\epsilon = g^\epsilon. \tag{3.17}$$

We will prove that w^ε_t is bounded in $\mathbf{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$. For this purpose, we multiply (3.17) by w^ε_t; we integrate this expression over Q_τ to obtain

$$\int_{Q_\tau} |w^\epsilon_t|^2 \,\mathrm{d}x\,\mathrm{d}t - \int_{Q_\tau} (A^1 w^\epsilon)\cdot w^\epsilon_t \,\mathrm{d}x\,\mathrm{d}t = \int_{Q_\tau} g^\epsilon \cdot w^\epsilon_t \,\mathrm{d}x\,\mathrm{d}t.$$

As

$$\int_{Q_\tau} (A^1 w^\epsilon)\cdot w^\epsilon_t \,\mathrm{d}x\,\mathrm{d}t = -\frac{1}{2}\int_\Omega a^1_{ijkl}\varepsilon_{ij}(w^\epsilon)\varepsilon_{kl}(w^\epsilon)\big|_0^\tau \,\mathrm{d}x, \tag{3.18}$$

we infer that

$$\begin{aligned} &\int_{Q_\tau} |w^\epsilon_t|^2 \,\mathrm{d}x\,\mathrm{d}t + \frac{1}{2}\int_\Omega a^1_{ijkl}\varepsilon_{ij}(w^\epsilon)\varepsilon_{kl}(w^\epsilon)\big|_{t=\tau}\,\mathrm{d}x \\ &= \frac{1}{2}\int_\Omega a^1_{ijkl}\varepsilon_{ij}(w^\epsilon)\varepsilon_{kl}(w^\epsilon)\big|_{t=0}\,\mathrm{d}x + \int_{Q_\tau} g^\epsilon\cdot w^\epsilon_t \,\mathrm{d}x\,\mathrm{d}t. \end{aligned} \tag{3.19}$$

According to Korn's inequality, we infer that there exist C_1 and C_2 such that

$$\int_\Omega a^1_{ijkl}\varepsilon_{ij}(w^\epsilon)\varepsilon_{kl}(w^\epsilon)\,\mathrm{d}x \ge C_1\int_\Omega |\nabla w^\epsilon|^2\,\mathrm{d}x - C_2\int_\Omega |w^\epsilon|^2\,\mathrm{d}x. \tag{3.20}$$

Carrying the above inequality into (3.19) and observing that $g^\epsilon \cdot w_t^\epsilon$ can be estimated by $|g^\epsilon|^2/2 + |w_t^\epsilon|^2/2$, we get

$$\begin{aligned}&\int_{Q_\tau} |w_t^\epsilon|^2 \,\mathrm{d}x\,\mathrm{d}t + C_1 \int_\Omega |\nabla w^\epsilon(\cdot,\tau)|^2 \,\mathrm{d}x \\ &\le \int_\Omega a^1_{ijkl}\varepsilon_{ij}(w^\epsilon)\varepsilon_{kl}(w^\epsilon)\big|_{t=0}\,\mathrm{d}x + C_2 \int_\Omega |w^\epsilon(\cdot,\tau)|^2 \,\mathrm{d}x + \int_{Q_\tau} |g^\epsilon|^2 \,\mathrm{d}x\,\mathrm{d}t.\end{aligned} \tag{3.21}$$

As v_0 belongs to $\mathbf{H}^2(\Omega)$, v_1 belongs to $\mathbf{H}^1(\Omega)$, φ belongs to $\mathrm{C}_0^\infty(\mathbb{R})$, and g^ε is bounded in $\mathbf{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$, we infer that the right-hand side of (3.21) is bounded. Therefore, using identities (3.13) and (3.16), it is possible to deduce that u^ε_{tt} is bounded in $\mathbf{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2((-\infty,-\beta)\times\Sigma))$.

We will show that $A^1 w^\epsilon$ is bounded in $\mathbf{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$ using an analogous method. We multiply (3.17) by $A^1 w^\epsilon$, we integrate over Q_τ, and we obtain

$$\int_{Q_\tau} w_t^\epsilon \cdot (A^1 w^\epsilon)\,\mathrm{d}x\,\mathrm{d}t - \int_{Q_\tau} |A^1 w^\epsilon|^2 \,\mathrm{d}x\,\mathrm{d}t = \int_{Q_\tau} g^\epsilon \cdot (A^1 w^\epsilon)\,\mathrm{d}x\,\mathrm{d}t. \tag{3.22}$$

Carrying (3.18) and (3.20) into (3.22), $g^\epsilon \cdot (A^1 w^\epsilon)$ being estimated by $|g^\epsilon|^2/2 + |A^1 w^\epsilon|^2/2$, we obtain

$$\begin{aligned}&\int_{Q_\tau} |A^1 w^\epsilon|^2 \,\mathrm{d}x\,\mathrm{d}t + C_1 \int_\Omega |\nabla w^\epsilon(\cdot,\tau)|^2 \,\mathrm{d}x \\ &\le \int_\Omega a^1_{ijkl}\varepsilon_{ij}(w^\epsilon)\varepsilon_{kl}(w^\epsilon)\big|_{t=0}\,\mathrm{d}x + C_2 \int_\Omega |w^\epsilon(\cdot,\tau)|^2 \,\mathrm{d}x + \int_{Q_\tau} |g^\epsilon|^2 \,\mathrm{d}x\,\mathrm{d}t.\end{aligned} \tag{3.23}$$

Thanks to (3.13) and (3.16), we may deduce from (3.23) that $A^1 u_t^\epsilon$ is bounded in $\mathbf{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2((-\infty,-\beta)\times\Sigma))$. □

3.3. Existence of a weak solution. Thanks to the estimates obtained in section 3.2, we are able to pass to the limit in the variational formulation associated to the penalized problem (3.1)–(3.3). Therefore, it is routine to deduce that there exists a solution to (1.9).

Because Ω is an unbounded set, the proof will be technical but similar to the one developed in section 2.3.

THEOREM 3.8. *Assume that f belongs to $\mathbf{L}^2_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\Omega))$, v_0 to $\mathbf{H}^2(\Omega)$, and v_1 to $\mathbf{L}^2(\Omega)$. Then there exists a solution to the variational inequality* (1.10)*; this solution is the limit of a subsequence of the penalty approximation defined by* (3.1)–(3.3).

Proof. Let $\varphi \in \mathrm{C}_0^\infty(\bar\Omega \times [0,\infty))$ be a function which takes its values between 0 and 1. We suppose here that v belongs to K. Multiplying (3.1) by $(v - u^\epsilon)\varphi$ and integrating over Q_τ,

$$\int_{I_\tau} (u_1^\epsilon)^+(\varphi(v_1 - u_1^\epsilon))\,\mathrm{d}x'\,\mathrm{d}t = -\int_{I_\tau} \big((u_1^\epsilon)^+\big)^2\varphi\,\mathrm{d}x'\,\mathrm{d}t + \int_{I_\tau} (u_1^\epsilon)^+\varphi v_1\,\mathrm{d}x'\,\mathrm{d}t$$

being negative, we get the following inequality:

$$\begin{aligned}&\int_\Omega \rho u_t^\epsilon \cdot (\varphi(v-u^\epsilon))|_0^\tau\,\mathrm{d}x - \int_{Q_\tau} \rho u_t^\epsilon \cdot (\varphi(v-u^\epsilon))_t\,\mathrm{d}x\,\mathrm{d}t \\ &+ \int_{Q_\tau} (a^0_{ijkl}\varepsilon_{kl}(u^\epsilon)\varepsilon_{ij}(u^\epsilon) + a^1_{ijkl}\varepsilon_{kl}(u_t^\epsilon)\varepsilon_{ij}(u^\epsilon))(\varphi(v_i - u_i^\epsilon))\,\mathrm{d}x\,\mathrm{d}t \\ &\ge \int_{Q_\tau} f \cdot \big(\varphi(v-u^\epsilon)\big)\,\mathrm{d}x\,\mathrm{d}t.\end{aligned} \tag{3.24}$$

We may deduce from Lemmas 3.2 and 3.5 that there exists a subsequence, still denoted by u^ϵ, such that

$$u^\epsilon \rightharpoonup u \quad \text{in} \quad \mathbf{L}^2_{\text{loc}}([0,\infty);\mathrm{L}^2(\Omega)) \quad \text{weak} \quad *, \tag{3.25a}$$

$$u^\epsilon_t \rightharpoonup u_t \quad \text{in} \quad \mathbf{L}^\infty_{\text{loc}}([0,\infty);\mathrm{L}^2(\Omega)) \quad \text{weak} \quad *, \tag{3.25b}$$

$$\nabla u^\epsilon \rightharpoonup \nabla u \quad \text{in} \quad \mathbf{L}^2_{\text{loc}}([0,\infty);\mathrm{L}^2(\Omega)) \quad \text{weak} \quad *, \tag{3.25c}$$

$$A^n u^\epsilon \rightharpoonup A^n u \quad \text{in} \quad \mathbf{L}^\infty_{\text{loc}}([0,\infty);\mathrm{L}^2(\Omega)) \quad \text{weak} \quad *, \quad n=0,1, \tag{3.25d}$$

$$\nabla u^\epsilon_t \rightharpoonup \nabla u_t \quad \text{in} \quad \mathbf{L}^2_{\text{loc}}([0,\infty);\mathrm{L}^2(\Omega)) \quad \text{weak} \quad *. \tag{3.25e}$$

Thanks to the classical compactness properties of Sobolev space injections on bounded open sets, we see that for all $R>0$, the restrictions of u^ϵ and $a^n_{ijkl}\varepsilon_{kl}(u^\epsilon)$, $n=0,1$, to $Q_R=\{x : x_1<0,\ |x'|\le R\}\times[0,R]$ (a set which has already been defined in section 2.3) converge strongly to their respective limits in $\mathbf{L}^2(Q_R)$. On the other hand, using the same techniques as those of section 2.3, we may prove that u^ϵ converges strongly to u in $\mathbf{L}^2_{\text{loc}}([0,\infty);\mathrm{L}^2(\Omega))$. The complete proof can be found in [Pet02, pp. 113–115].

We observe now that since u and u_t belong to $\mathbf{L}^2_{\text{loc}}([0,\infty);\mathrm{L}^2(\Omega))$, we may replace φ by φ_R in the variational inequality where φ_R is equal to 1 over the set Q_R and vanishes outside of Q_{R+1}. When R tends to infinity all the terms in (3.24) converge to their limit; thus we have proved the existence of a weak solution. ∎

Remark 3.9. As for the damped wave equation with Signorini boundary conditions, the uniqueness is still an open problem.

3.4. Preliminary results. In this section, we establish estimates on the problem (1.9a) with initial data (1.9d) and the Dirichlet boundary condition which enable us to characterize the trace spaces in the next section.

LEMMA 3.10. *Assume v_0 and v_1 belong, respectively, to $\mathbf{H}^{5/2}(\Omega)$ and $\mathbf{H}^{3/2}(\Omega)$; then, there exists a function with compact support in t such that the traces of z and z_t on Σ are, respectively, v_0 and v_1.*

Proof. We extend v_0 and v_1 into functions belonging, respectively, to $\mathbf{H}^{5/2}(\mathbb{R}^d)$ and $\mathbf{H}^{3/2}(\mathbb{R}^d)$. Then there exists a function Z belonging to $\mathbf{H}^3(\mathbb{R}^d\times[0,\infty))$ such that $Z|_{\mathbb{R}^d\times\{0\}}=v_0$ and $Z_t|_{\mathbb{R}^d\times\{0\}}=v_1$. We select a cut-off function $\varphi\in\mathrm{C}^\infty([0,\infty))$ which is equal to 1 on $[0,1]$ and to 0 on $[2,\infty)$, and we define z as the restriction of $\varphi(x)Z(x,t)$ to $\Omega\times[0,\infty)$. ∎

LEMMA 3.11. *Assume v_0 belongs to $\mathbf{H}^{5/2}(\Omega)$, v_1 belongs to $\mathbf{H}^1(\Omega)$, and f belongs to $\mathbf{L}^2_{\text{loc}}([0,\infty);\mathrm{L}^2(\Omega))$. Define z as in Lemma* 3.10*, and let $\bar u$ be the solution of* (1.9a) *with initial data* (1.9d) *and boundary condition $\bar u(0,\cdot,\cdot)=z(0,\cdot,\cdot)$. Then the trace $\bar g \stackrel{\text{def}}{=} -(a^0_{11kl}\varepsilon_{kl}(\bar u)+a^1_{11kl}\varepsilon_{kl}(\bar u_t))\big|_{\Sigma\times[0,\infty)}$ is well defined and belongs to the space $\mathrm{L}^2_{\text{loc}}([0,\infty);\mathrm{L}^2(\Sigma))$. Moreover, there exists $K>0$ such that $\mathrm{e}^{-Kt}\bar g\in\mathrm{L}^2(\Sigma\times[0,\infty))$.*

Proof. Let $\zeta\stackrel{\text{def}}{=}\bar u - z$ be the solution of the following problem:

$$\rho\zeta_{tt}-A^0\zeta-A^1\zeta_t=F,\quad x\in\Omega,\quad t>0, \tag{3.26}$$

where $F\stackrel{\text{def}}{=} f-\rho z_{tt}+A^0z+A^1z_t$ with initial data $\zeta(\cdot,0)=\zeta_t(\cdot,0)=0$ and boundary condition $\zeta(0,\cdot,\cdot)=0$. Multiplying (3.26) by ζ_t and integrating over Q_τ, Korn's inequality enables us to deduce that ζ_t and $\nabla\zeta$ are bounded in $\mathbf{L}^\infty_{\text{loc}}([0,\infty);\mathrm{L}^2(\Omega))$ and $\nabla\zeta_t$ is bounded in $\mathbf{L}^2_{\text{loc}}([0,\infty);\mathrm{L}^2(\Omega))$. If we multiply (3.26) by $A^1\zeta$, we may deduce that $A^0\zeta$ and $A^1\zeta$ are bounded in $\mathbf{L}^\infty_{\text{loc}}([0,\infty);\mathrm{L}^2(\Omega))$, arguing as in the proof of Lemma 3.5. On the other hand, we have

$$\int_{Q_\tau} \zeta_{tt}\cdot(A^1\zeta_t)\,\mathrm{d}x\,\mathrm{d}t = -\frac{1}{2}\int_\Omega a^1_{ijkl}\varepsilon_{kl}(\zeta_t)\varepsilon_{ij}(\zeta_t)\big|_0^\tau\,\mathrm{d}x.$$

Therefore, we multiply (3.26) by $A^1\zeta_t$, we integrate over Q_τ, and, thanks to the above identity, we get

$$\begin{aligned}&\frac{\rho}{2}\int_\Omega a^1_{ijkl}\varepsilon_{kl}(\zeta_{tt})\varepsilon_{ij}(\zeta_t)\big|_{t=\tau}\,\mathrm{d}x + \int_{Q_\tau}|A^1\zeta_t|^2\,\mathrm{d}x\,\mathrm{d}t + \frac{1}{2}\int_\Omega (A^0\zeta)\cdot(A^1\zeta)\big|_0^\tau\,\mathrm{d}x\\ &= \frac{\rho}{2}\int_\Omega a^1_{ijkl}\varepsilon_{kl}(\zeta_t)\varepsilon_{ij}(\zeta_t)\big|_{t=0}\,\mathrm{d}x - \int_{Q_\tau} F\cdot(A^1\zeta_t)\,\mathrm{d}x\,\mathrm{d}t.\end{aligned}$$

According to Gronwall's lemma, there exists $K>0$ such that

$$\int_{Q_\tau}|A^1\zeta_t|^2\,\mathrm{d}x\,\mathrm{d}t \le C\mathrm{e}^{K\tau}\left(|F|^2_{\mathrm{L}^2(0,\tau;\mathrm{L}^2(\Omega))} + |\xi_t(\cdot,0)|^2_{\mathrm{H}^1(\Omega)} + |\xi|^2_{\mathrm{L}^2(0,\tau;\mathrm{L}^2(\Omega))}\right).$$

The lemma is now clear. □

3.5. The trace spaces. We proceed as in section 2.5. A Fourier–Laplace transform and Lemma 3.11 enable us to infer that all the traces can be defined. Therefore, it is plain that a weak solution of (1.9) is also a strong one.

Let us remark first that the problem (3.3) can be written under an equivalent form: let us extend by 0 for $t\le 0$ the difference $v^\epsilon \stackrel{\text{def}}{=} \mathrm{e}^{-\nu t}(u^\epsilon - \bar u)$; then it satisfies

$$\begin{aligned}&\rho(\nu+\partial_t)^2 v_i^\epsilon - ((\lambda^0+\mu^0)+(\lambda^1+\nu^1)(\nu+\partial_t))\,\mathrm{div}\,v^\epsilon\\ &-(\mu^0+\mu^1(\nu+\partial_t))\Delta v_i^\epsilon = 0,\quad x\in\Omega,\quad t>0,\end{aligned}\tag{3.27}$$

with boundary conditions at $\{x_1=0\}$

$$(\mu^0+\nu\mu^1)(v^\epsilon_{j,x_1}+v^\epsilon_{1,x_j}) + \mu^1(v^\epsilon_{j,x_1t}+v^\epsilon_{1,x_jt}) = 0,\quad j=2,3,\tag{3.28a}$$

$$(\lambda^0+\lambda^1(\nu+\partial_t))\,\mathrm{div}\,v^\epsilon + 2(\mu^0+\mu^1(\nu+\partial_t))v^\epsilon_{x_1} = \mathrm{e}^{-\nu t}\bar g - \frac{(v_1^\epsilon - \mathrm{e}^{-\nu t}\bar u)^+}{\epsilon}\tag{3.28b}$$

and with initial data

$$v^\epsilon(\cdot,0)=0 \quad\text{and}\quad v^\epsilon_t(\cdot,0)=0.\tag{3.29}$$

If v^ϵ is a tempered distribution, we may perform a Fourier transform in the tangential variable (x',t) and a Laplace transform in x_1. Denoting by ξ and ω the dual variables of x' and t and by η the dual variable of x_1, we are led to the system

$$\begin{aligned}&\rho(\nu+i\omega)^2\widehat{v}^\epsilon - ((\lambda^0+\mu^0)+(\lambda^1+\mu^1)(\nu+i\omega))\begin{pmatrix}\eta\\ i\xi\end{pmatrix}(\eta, i\xi^{\mathsf T})\widetilde{v}^\epsilon\\ &+(\mu^0+\mu^1(\nu+i\omega))(|\xi|^2-\eta^2)\widetilde{v}^\epsilon = 0.\end{aligned}\tag{3.30}$$

Equation (3.30) is a linear system of equations; we seek its eigenvalues η_i and its eigenvectors ϕ_i:

$$\eta_1^2 = |\xi|^2 + \frac{\rho(\nu+i\omega)^2}{\mu^0+\mu^1(\nu+i\omega)}\quad\text{and}\quad \phi_1 = \begin{pmatrix}0\\ i\xi^\perp\end{pmatrix},\tag{3.31a}$$

$$\eta_2^2 = |\xi|^2 + \frac{\rho(\nu+i\omega)^2}{\mu^0+\mu^1(\nu+i\omega)}\quad\text{and}\quad \phi_2 = \begin{pmatrix}1\\ 0\end{pmatrix},\tag{3.31b}$$

$$\eta_3^2 = |\xi|^2 + \frac{\rho(\nu+i\omega)^2}{\lambda^0+2\mu^0+(\lambda^1+2\mu^1)(\nu+i\omega)}\quad\text{and}\quad \phi_3 = \begin{pmatrix}\eta_3\\ i\xi\end{pmatrix},\tag{3.31c}$$

where $\xi^\perp$ is obtained from ξ by a rotation of $\pi/2$. We choose η_i to be the causal determination of the square root of η_i^2; therefore, η_i is holomorphic in the lower half-plane $\Im(\omega) < 0$. Let us denote by $\widehat{v}^\epsilon$ the partial Fourier transform of v^ϵ with respect to the tangential variables. As v^ϵ and $\widetilde{v}^\epsilon$ are tempered distributions, $\widehat{v}^\epsilon$ is also tempered; therefore, it can include only factors of the form $\mathrm{e}^{\eta_i x_1}$, and thus, it must be of the form

$$\widehat{v}^\epsilon(x_1,\xi,\omega) = \sum_{i=1}^{3} \theta_i(\xi,\omega)\phi_i \mathrm{e}^{\eta_i x_1}. \tag{3.32}$$

Our goal now is to determine the θ_i's. Define $v^\epsilon \overset{\text{def}}{=} (v_1^\epsilon, (v^\epsilon)')$. If we apply a partial Fourier transform in the tangential variable to the boundary condition (3.28a), we obtain

$$(\widehat{v}^\epsilon)'_{x_1}(0,\xi,\omega) = -i\xi\widehat{v}_1^\epsilon(0,\xi,\omega). \tag{3.33}$$

Carrying (3.32) into (3.33), we infer that at $x_1 = 0$,

$$i\xi^\perp\eta_2\theta_1 + i\xi\eta_3\theta_3 = -i\xi(\theta_2 + \eta_3\theta_3);$$

thus it is clear that $\theta_1 = 0$ and $\theta_2 = -2\eta_3\theta_3$. Furthermore, relation (3.32) taken at $x_1 = 0$ enables us to deduce that $\theta_3 = -\widehat{v}_1^\epsilon(0,\xi,\omega)/\eta_3$. Finally, we obtain

$$\widehat{v}^\epsilon(x_1,\xi,\omega) = 2\widehat{v}_1^\epsilon(0,\xi,\omega)\phi_2\mathrm{e}^{\eta_2 x_1} - \widehat{v}_1^\epsilon(0,\xi,\omega)\phi_3\mathrm{e}^{\eta_3 x_1}/\eta_3. \tag{3.34}$$

At last, using (3.34), we can write the left-hand side of (3.28b) as a product of convolution: if we perform a Fourier transform of the left-hand side of (3.28b) and since

$$\widehat{v}^\epsilon_{1,x_1}(0,\xi,\omega) = (2\eta_2 - \eta_3)\widehat{v}_1^\epsilon(0,\xi,\omega) \quad \text{and} \quad (\widehat{v}^\epsilon)'(0,\xi,\omega) = -i\xi\widehat{v}_1^\epsilon(0,\xi,\omega)/\eta_3,$$

we obtain

$$((\lambda^0 + \lambda^1(\nu+\partial_t))\operatorname{div} v^\epsilon + 2(\mu^0 + \mu^1(\nu+\partial_t))v^\epsilon_{1,x_1})\widehat{\ }(0,\xi,\omega) = \widehat{b}\widehat{v}_1^\epsilon(0,\xi,\omega),$$

where

$$\widehat{b} \overset{\text{def}}{=} (\lambda^0 + 2\mu^0 + (\lambda^1 + 2\mu^1)(\nu + i\omega))(2\eta_2 - \eta_3) + (\lambda^0 + \lambda^1(\nu+i\omega))|\xi|^2/\eta_3.$$

Let $w^\epsilon(x',t)$ be the trace $v^\epsilon(0,x',t)$; then (3.28b) can now be written as

$$b * w_1^\epsilon = \mathrm{e}^{-\nu t}\bar{g} - \frac{\left(w_1^\epsilon - \mathrm{e}^{-\nu t}\bar{u}_1(0,\cdot,\cdot)\right)^+}{\epsilon}. \tag{3.35}$$

LEMMA 3.12. *Let $u^\epsilon = (u_1^\epsilon, u_2^\epsilon, u_3^\epsilon)^\mathsf{T}$ be the solution of* (3.1)–(3.3a). *Then we may extract a subsequence, still denoted by u_1^ϵ, such that*

$$u_1^\epsilon(0,\cdot,\cdot) \rightharpoonup u_1(0,\cdot,\cdot) \quad \textit{weakly} \quad \textit{in} \quad \mathrm{H}^{1/2,5/4}_{\mathrm{loc}}(\mathbb{R}^{d-1}\times[0,\infty)).$$

Moreover, u is a strong solution of (1.9).

Proof. We denote by $\widehat{\psi}$ and $\widehat{g}$ the respective Fourier transforms of $\psi \stackrel{\text{def}}{=} \lambda^0+2\mu^0+(\lambda^1+2\mu^1)(\nu+\partial_t)$ and $g \stackrel{\text{def}}{=} \mathrm{e}^{-\nu t}\bar{g}$. Multiplying (3.35) by ψw_1^ϵ and using the Plancherel identity, we obtain

$$\begin{aligned}&\frac{1}{(2\pi)^d}\,\Re\int_{\mathbb{R}^d}\overline{\widehat{\psi}}\,\widehat{b}|\widehat{w}_1^\epsilon|^2\,\mathrm{d}\xi\,\mathrm{d}\omega\\&=\frac{1}{(2\pi)^d}\,\Re\int_{\mathbb{R}^d}\widehat{g}\overline{\widehat{\psi}\widehat{w}_1^\epsilon}\,\mathrm{d}\xi\,\mathrm{d}\omega-\int_0^\infty\int_{\mathbb{R}^{d-1}}\frac{\bigl(w_1^\epsilon-\mathrm{e}^{-\nu t}\bar{u}_1(0,\cdot,\cdot)\bigr)^+}{\epsilon}\,\psi w_1^\epsilon\,\mathrm{d}x'\,\mathrm{d}t.\end{aligned}$$

According to the Cauchy–Schwarz inequality and since $(u_1^\epsilon(0,\cdot,\cdot))^+/\sqrt{\epsilon}$ is bounded in $\mathrm{L}^\infty_{\mathrm{loc}}([0,\infty);\mathrm{L}^2(\mathbb{R}^{d-1}))$, the absolute value of the second integral on the right-hand side of the above inequality is bounded by C_1; therefore, we get

$$\frac{1}{(2\pi)^d}\,\Re\int_{\mathbb{R}^d}\overline{\widehat{\psi}}\,\widehat{b}|\widehat{w}_1^\epsilon|^2\,\mathrm{d}\xi\,\mathrm{d}\omega\le C_1+\frac{1}{(2\pi)^d}\,\Re\int_{\mathbb{R}^d}\widehat{g}\overline{\widehat{\psi}\widehat{w}_1^\epsilon}\,\mathrm{d}\xi\,\mathrm{d}\omega.\tag{3.36}$$

Define

$$\begin{aligned}\kappa&\stackrel{\text{def}}{=}\frac{-\rho(\nu+i\omega)^2-2|\xi|^2(\mu^0+(\nu+i\omega)\mu^1)}{-\rho(\nu+i\omega)^2-|\xi|^2\bigl(\lambda^0+2\mu^0+(\nu+i\omega)(\lambda^1+2\mu^1)\bigr)},\\x_0&\stackrel{\text{def}}{=}\sqrt{\frac{2\rho(\lambda^0+\nu\lambda^1)}{4\lambda^1\mu^1+(\lambda^1)^2}}.\end{aligned}$$

Then $\widehat{b}=\widehat{\psi}(2\eta_2-\kappa\eta_3)$, and we remark also that it is sufficient to find a function h which depends on ξ and ω such that $\Re(2\eta_2-\kappa\eta_3)\ge|h|$. If we assume that $|\xi|+|\omega|\gg1$, we have two cases to consider according to the values taken by $|\xi|$. We suppose first that $|\xi|^2+2\rho(\nu\mu^1-\mu^0)/(\mu^1)^2\ge0$; then η_2 can be approximated by $\widetilde{\eta}_2$ defined as follows:

$$|\widetilde{\eta}_2^2|^2=\left(|\xi|^2+\frac{\rho(\nu\mu^1-\mu^0)}{(\mu^1)^2}\right)^2+\left(\frac{\rho\mu^1\omega}{(\mu^1)^2}\right)^2.$$

Therefore, it is easy to deduce that $|\widetilde{\eta}_2^2|^2\ge|\xi|^4/4+\rho^2\omega^2/(\mu^1)^2$, and then, in the case $|\xi|^2+2\rho(\nu\mu^1-\mu^0)/(\mu^1)^2\ge0$, we obtain the following estimate:

$$\Re\eta_2\ge\cos\bigl(\pi/4\bigr)|\eta_2|\ge\frac{1}{\sqrt{2}}\left(\frac{|\xi|^4}{4}+\frac{\rho^2\omega^2}{(\mu^1)^2}\right)^{1/4}.\tag{3.37}$$

In the other case, we suppose $|\xi|^2+2\rho(\nu\mu^1-\mu^0)/(\mu^1)^2\le0$; then, it is plain that

$$|\Re\eta_2^2|\le\frac{3\rho(\nu\mu^1+\mu^0)}{(\mu^1)^2}\quad\text{and}\quad|\Im\eta_2^2|\ge\frac{\rho\mu^1|\omega|}{(\mu^1+\nu\mu^0)^2+(\mu^0)^2},$$

which implies that there exists $C>0$ such that

$$|\operatorname{arc\,cotan}\eta_2^2|\le\frac{3(\nu\mu^1+\mu^0)\bigl((\mu^1+\nu\mu^0)^2+(\mu^0)^2\bigr)}{(\mu^1)^3|\omega|}\le\frac{C}{|\omega|}.$$

We deduce from the above inequality and from $|\eta_2|^2\ge C|\omega|$ that $|\arg\eta_2^2|\le\pi/2+C/|\omega|$ and thus $\cos(\arg\eta_2)\ge1/2$. In the case $|\xi|^2+2\rho(\nu\mu^1-\mu^0)/(\mu^1)^2\le0$, we get

$$\Re\eta_2\ge C\sqrt{|\omega|}/2.\tag{3.38}$$

Therefore, in both cases, we infer from (3.37) and (3.38) that there exists $M > 0$ such that

$$\Re\eta_2 \geq M\left(\omega^2 + |\xi|^4\right)^{1/4}. \tag{3.39}$$

Furthermore, there exists $C > 0$ such that $|\kappa|^2 \leq 1 + C1_{\{|\xi|\leq x_0\}}/|\omega|^2$ and for $|\xi|$ large enough, $|\eta_2| \geq |\eta_3|$. Then (3.39) enables us to deduce

$$\Re(2\eta_2 - \kappa\eta_3) \geq M\left(\omega^2 + |\xi|^4\right)^{1/4}. \tag{3.40}$$

Carrying (3.40) into (3.36), we obtain

$$M\int_{\mathbb{R}^d} |\widehat{\psi}|^2\left(\omega^2 + |\xi|^4\right)^{1/4} |\widehat{w}_1^\varepsilon|^2\,\mathrm{d}\xi\,\mathrm{d}\omega \leq C_1 + \int_{\mathbb{R}^d} |\widehat{g}||\widehat{\psi}||\widehat{w}_1^\epsilon|\,\mathrm{d}\xi\,\mathrm{d}\omega. \tag{3.41}$$

We estimate the product zy by $|z|^2/(2\gamma) + \gamma|y|^2/2$, $\gamma > 0$, and we see that

$$\left(M - \frac{\gamma}{2}\right)\int_{\mathbb{R}^d} |\omega|^2\left(\omega^2 + |\xi|^4\right)^{1/4} |\widehat{w}_1^\varepsilon|^2\,\mathrm{d}\xi\,\mathrm{d}\omega \leq C_1 + \frac{1}{2\gamma}\int_{\mathbb{R}^d} \frac{|\widehat{g}|^2}{\left(\omega^2 + |\xi|^4\right)^{1/4}}\,\mathrm{d}\xi\,\mathrm{d}\omega.$$

We choose γ such that $\gamma < 2M$. On the other hand, $\mathrm{e}^{-Kt}\bar{g}(\cdot,t)$ is bounded in $\mathrm{L}^2(\Sigma\times[0,\infty))$, so that $g(\cdot,t)$ is bounded in $\mathrm{L}^2(\Sigma\times[0,\infty))$ if we choose $\nu > K$. Therefore, u_1^ϵ is bounded in $\mathrm{H}^{1/2,5/4}_{\mathrm{loc}}(\Sigma\times[0,\infty))$. In particular, $(\lambda^0+\lambda^1(\nu+\partial_t))\,\mathrm{div}\,v^\epsilon + 2(\mu^0+\mu^1(\nu+\partial_t))v^\epsilon_{x_1}$ is bounded in $\mathrm{H}^{-1/2,-1/4}_{\mathrm{loc}}(\Sigma\times[0,\infty))$. We conclude that u is a strong solution of (1.9) because all the traces can be defined. ∎

Acknowledgment. We would like to thank M. Chambat for reading this work and for providing helpful comments.

REFERENCES

[DaL90] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology, Vol.* 3, Springer-Verlag, Berlin, 1990.

[GGZ74] H. Gajewski, K. Gröger, and K. Zacharias, *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen*, Mathematische Lehrbücher und Monographien, II. Abteilung, Mathematische Monographien, Band 38, Akademie-Verlag, Berlin, 1974.

[Jar96] J. Jarušek, *Dynamic contact problems with given friction for viscoelastic bodies*, Czechoslovak Math. J., 46 (1996), pp. 475–487.

[JM*92] J. Jarušek, J. Málek, J. Nečas, and V. Šverák, *Variational inequality for a viscous drum vibrating in the presence of an obstacle*, Rend. Mat. Appl. (7), 12 (1992), pp. 943–958.

[KuS04a] K. L. Kuttler and M. Shillor, *Dynamic contact with Signorini's condition and slip rate dependent friction*, Electron. J. Differential Equations, 83 (2004).

[KuS04b] K. L. Kuttler and M. Shillor, *Regularity of solutions to a dynamic frictionless contact problem with normal compliance*, Nonlinear Anal., 59 (2004), pp. 1063–1075.

[MaO88] J. A. C. Martins and J. T. Oden, *Corrigendum: Existence and uniqueness results for dynamic contact problems with nonlinear normal and friction interface laws*, Nonlinear Anal., 12 (1988), p. 747.

[Pet02] A. Petrov, *Modélisation mathématique de procédés d'usinage: abrasion et mouillage [Mathematical Modelization of Maching Process: Abrasion and Wetting]*, Ph.D. thesis, University Claude Bernard (Lyon 1), pp. 199–2002; available online at http://www.wias-berlin.de/people/petrov/pub.html, 2002.

[PeS02] A. Petrov and M. Schatzman, *Viscoélastodynamique monodimensionnelle avec conditions de Signorini*, C. R. Math. Acad. Sci. Paris, 334 (2002), pp. 983–988.

[PeS08] A. Petrov and M. Schatzman, *A pseudodifferential linear complementarity problem related to one dimensional viscoelastic model with Signorini conditions*, Arch. Ration. Mech. Anal., to appear.

SIAM J. MATH. ANAL.
Vol. 40, No. 5, pp. 1905–1922

L^2 DECAY OF SOLUTIONS TO A MICRO-MACRO MODEL FOR POLYMERIC FLUIDS NEAR EQUILIBRIUM*

LINGBING HE† AND PING ZHANG‡

Abstract. In this paper, we consider the long time decay of the L^2 norm to the global solutions (u, ψ) constructed in [F.-H. Lin, C. Liu, and P. Zhang, *Comm. Pure Appl. Math.*, 60 (2007), pp. 838–866] for a micro-macro model of polymeric fluids near equilibrium $(0, M)$. Under the additional assumption that $u_0 \in \dot{H}^{-\kappa}(\mathbf{R}^3)$, $(\psi_0 - M)/\sqrt{M} \in L^2(\mathbf{R}^3_q; \dot{H}^{-\kappa}(\mathbf{R}^3_x))$, we prove that $(u(t), \psi(t))$ tends to $(0, M)$ as t goes to infinity with decaying rate $\|u(t)\|_{L^2} \le C(1+t)^{-\frac{b}{2}}$ and $\|\frac{\psi(t)-M}{\sqrt{M}}\|_{L^2} \le C(1+t)^{-\frac{b+1}{2}}$ for $b = \min(\kappa, \frac{3}{2})$. In general, without this additional assumption, we shall present an explicit long time decaying formula for $\|u(t)\|_{L^2}$ and $\|\frac{\psi(t)-M}{\sqrt{M}}\|_{L^2}$.

Key words. L^2 decay, micro-macro model, Fourier transform

AMS subject classifications. 74A25, 76D99

DOI. 10.1137/07712031

1. Introduction.

In this paper, we consider the long time behavior to the global solutions constructed in [10] for a coupled microscopic-macroscopic model for polymeric fluid near equilibrium. The micro-mechanical models for polymeric liquids usually consist of beads joined by springs or rods [1, 6]. In the simplest case, a molecule configuration can be described by its end-to-end vector q. Taking into account the elastic effect together with the thermo-fluctuation, the distribution function $\psi(t, x, q)$ of molecule orientations q satisfies a Fokker–Planck equation. The convection velocity u satisfies the Navier–Stokes equations with an elastic stress which reflects the microscopic contribution of the polymer molecules to the overall macroscopic flow fields. Mathematically, this system reads (one may check [10] for a formal energetic variational derivation) as

$$\begin{cases} u_t + u \cdot \nabla u + \nabla p = \mu \triangle u + \nabla \cdot \tau, \quad x \in \mathbf{R}^3, \\ \nabla \cdot u = 0, \quad x \in \mathbf{R}^3, \\ \psi_t + u \cdot \nabla \psi = \sigma \triangle_q \psi - \nabla_q \cdot (\nabla u q \psi - \sigma \nabla_q U \psi), \quad (x, q) \in \mathbf{R}^3 \times \mathbf{R}^3, \end{cases} \tag{1.1}$$

where the polymer stress τ is given by

$$\tau = \int_{\mathbf{R}^3} \nabla_q U \otimes q \psi dq, \tag{1.2}$$

with $U(q) = U(|q|^2)$ being the potential function.

Existence results for micro-macro models of polymeric fluids are usually limited to small time existence [7, 17, 13] and uniqueness of strong solutions or global existence

*Received by the editors December 29, 2007; accepted for publication (in revised form) August 15, 2008; published electronically January 14, 2009.
http://www.siam.org/journals/sima/40-5/71203.html
†Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China (lbhe@math.tsinghua.edu.cn).
‡Academy of Mathematics & Systems Science, CAS, Beijing 100080, China (zp@amss.ac.cn). This author was partially supported by NSF of China under grant 10525101 and 10421101, National 973 project, and the innovation grant from Chinese Academy of Sciences.

of weak solutions [12]. In the setting when the last equation of (1.1) is formulated as a stochastic PDE, we refer to [7] (see also [17] for a polynomial force). Concerning the general coupled PDE system, some preliminary studies were made in the earlier work [13].

In a recent work [10], Lin, Liu, and Zhang studied the global existence of smooth solutions to (1.1) near equilibrium, which is a sort of extension to a related result of the Oldroyd model [9, 3], and which corresponds to the Hooke dumbbell model. In two space dimensions, Constantin et al. [4] proved the global existence of smooth solutions to a coupled nonlinear Fokker–Planck and Navier–Stokes system when the convection velocity u in the Fokker–Planck equation is replaced by a sort of time averaged one. Later this assumption was removed by Constantin and Masmoudi [5]. Meanwhile Lin, Zhang, and Zhang [11] independently proved the global regularity for the two-dimensional corotational FENE model with smooth initial data.

On the other hand, in [8], Jourdain et al. investigated the long time behavior of both Hookean models and FENE models for various special flows in a bounded domain with suitable boundary conditions. The main aim of this paper is to consider the long time decay rate for the global solutions to (1.1) constructed in [10].

Before we proceed, let us introduce some basic notation that will be used throughout this paper. Similarly to [10], after renormalization, we assume that $\int_{\mathbf{R}^3} \exp\{-U\}\, dq = 1$; furthermore,

$$
\begin{aligned}
&|q| \le C(|\nabla_q U| + 1), \quad \Delta_q U \le C + \delta |\nabla_q U|^2, \text{ with } \delta < 1,\\
&\int_{\mathbf{R}^3} |\nabla_q U|^2 e^{-U}\, dq \le C, \quad \int_{\mathbf{R}^3} |q|^4 e^{-U}\, dq \le C,
\end{aligned} \tag{1.3}
$$

and

$$
\begin{aligned}
&|\nabla_q^k (q \nabla_q U)| \le C(|q||\nabla_q U| + 1), \quad \int_{\mathbf{R}^3} |\nabla_q^k (q \nabla_q U e^{-\frac{U}{2}})|^2\, dq \le C,\\
&\left| \nabla_q^k \left(\Delta_q U - \frac{|\nabla_q U|^2}{2} \right) \right| \le C(1 + |\nabla_q U|^2),
\end{aligned} \tag{1.4}
$$

where the positive integer $0 \le k \le s$, which will be fixed in the later sections.

In what follows, for any given function $\phi(x, q)$, we denote

$$
|\phi|_{H^s} = \left\{ \int_{\mathbf{R}^3} \int_{\mathbf{R}^3} \sum_{|\alpha| \le s} |\nabla_x^\alpha \phi|^2\, dq dx \right\}^{\frac{1}{2}},
$$

and $\|\phi\|_{H^s}$ the standard Sobolev norm of ϕ. We shall use the convention (f, g) to stand for both the inner product on $\mathbf{R}^3$, $\int_{\mathbf{R}^3} fg\, dx$, and on $\mathbf{R}^3 \times \mathbf{R}^3$, $\int_{\mathbf{R}^3} \int_{\mathbf{R}^3} fg\, dq dx$. And we will denote ∇^s to be any of ∇_x^α, where α is any multiple indices with $|\alpha| = s$. And we use similar notation for ∇_q^s.

It is easy to check that $(0, e^{-U})$ is a stationary solution to the system (1.1). In [10], Lin, Liu, and Zhang considered solutions of (1.1) with ψ being of the form

$$
\psi = e^{-U} + e^{-U/2} f \overset{\text{def}}{=} M + \sqrt{M} f. \tag{1.5}
$$

As we assume that $\int_{\mathbf{R}^3} \psi_0(x, q) dq = 1$, there holds $\int_{\mathbf{R}^3} \psi(t, x, q) dq = 1$, which together with (1.5) yields

$$
\int_{\mathbf{R}^3} \sqrt{M} f\, dq = 0. \tag{1.6}
$$

Moreover, plugging (1.5) into (1.1), we obtain the following system for (u, f):

$$
\begin{cases}
u_t + u\cdot\nabla u + \nabla p = \mu\triangle u + \nabla\cdot\left(\int_{\mathbf{R}^3}\sqrt{M}\nabla_q U\otimes q f\, dq\right), & x\in\mathbf{R}^3,\\
f_t + u\cdot\nabla f + \nabla u q\cdot\nabla_q f - \sigma\left(\triangle_q f + \frac{\triangle_q U}{2} f - \frac{|\nabla_q U|^2}{4} f\right)\\
\qquad = \sqrt{M}\nabla u q\cdot\nabla_q U + \frac{1}{2}\nabla u q\cdot\nabla_q U f, & (x,q)\in\mathbf{R}^3\times\mathbf{R}^3,\\
\text{div } u = 0, \quad x\in\mathbf{R}^3,
\end{cases}
\tag{1.7}
$$

together with the initial conditions

$$
u|_{t=0} = u_0, \qquad f|_{t=0} = f_0, \quad \text{for} \quad (x,q)\in\mathbf{R}^3\times\mathbf{R}^3,
\tag{1.8}
$$

where f_0 satisfies (1.6) and div $u_0 = 0$.

For the reader's convenience, we first recall the following result from [10].

THEOREM 1.1. *Let $s \geq 7$ be an integer, f_0 satisfy* (1.6), *and $\psi_0 = M + \sqrt{M} f_0 > 0$. Then, there exists a sufficiently small constant ε such that if*

$$
\frac{1}{2}\|u_0\|_{L^2}^2 + \int_{\mathbf{R}^3}\int_{\mathbf{R}^3}[\psi_0\ln\psi_0 + U\psi_0]\,dq\,dx \leq \varepsilon^{1+a}\mu\sigma\min(\mu,\sigma),
$$

$$
\|u_0\|_{H^s}^2 + \|f_0\|_{H^s}^2 + |q f_0|_{H^4}^2 \leq \varepsilon\min(\mu,\sigma),
$$

where $a > 0$ is a small positive constant, then (1.7) *and* (1.8) *has a unique global classical solution (u,ψ) with $\psi = M + \sqrt{M} f > 0$, and*

$$
\begin{aligned}
&\sup_{0\leq t<\infty}\left(\|u(t)\|_{H^s}^2 + |qf|_{H^4}^2 + \|f(t)\|_{H^s}^2\right)\\
&\quad + \int_0^t\left[\mu\|\nabla u\|_{H^s}^2 + \sigma\left(\left\|q\right|\left(\nabla_q f + \frac{1}{2}\nabla_q U f\right)\right|_{H^4}^2 + \left\|\nabla_q f + \frac{1}{2}\nabla_q U f\right\|_{H^s}^2\right)\right]d\tau\\
&\quad \leq C\varepsilon\min(\mu,\sigma).
\end{aligned}
\tag{1.9}
$$

The main aim of this paper is to prove that the solution (u, f) constructed in Theorem 1.1 decays to $(0, 0)$ as t goes to ∞. More precisely, motivated by [14, 15] and with the additional assumption on the low frequency part of (u_0, f_0), we obtain the following decay rates for the L^2 norm of (u, f).

THEOREM 1.2. *Under the assumptions of Theorem* 1.1, *we assume further that there exists $\kappa, C > 0$ and a small enough number $c > 0$ such that*

$$
\int_{|\xi|\leq c}|\xi|^{-2\kappa}|\hat{u}_0|^2\,d\xi, \quad \int_{\mathbf{R}^3}\int_{|\xi|\leq c}|\xi|^{-2\kappa}|\mathcal{F}_x(f_0)(\xi,q)|^2\,d\xi\,dq \leq C.
\tag{1.10}
$$

Then the L^2 norm of (u, f) decays to $(0, 0)$ according to

$$
\|u(t)\|_{L^2} \leq C(1+t)^{-\frac{b}{2}}, \quad \|f(t)\|_{L^2} \leq C(1+t)^{-\frac{b+1}{2}},
\tag{1.11}
$$

with $b = \min(\kappa, \frac{3}{2})$.

In general, without the additional assumptions (1.10), to study the precise long time decay rates for (u, f), we need some basic facts from the Littlewood–Paley theory. One may check [2] and [16] for more details.

Let $\mathcal{C} \overset{\text{def}}{=} \{\xi \in \mathbf{R}^3,\ \frac{3}{4} \le |\xi| \le \frac{8}{3}\}$. Let $\varphi \in C_c^\infty(\mathcal{C})$, which satisfies

$$\sum_{j\in\mathbf{Z}} \varphi(2^{-j}\xi) = 1 \quad \forall\, \xi \in \mathbf{R}^3 \setminus \{0\}.$$

We denote $h \overset{\text{def}}{=} \mathcal{F}^{-1}\varphi$; then the Littlewood–Paley operators $\dot{\Delta}_j$ and $\dot{\mathcal{S}}_j$ can be defined as follows:

$$\dot{\Delta}_j f \overset{\text{def}}{=} \varphi(2^{-j}D)f = 2^{3j}\int_{\mathbf{R}^3} h(2^j y) f(x-y)dy \quad \text{for} \quad j \in \mathbf{Z},$$

$$\dot{\mathcal{S}}_j f \overset{\text{def}}{=} \sum_{j' \le j-1} \dot{\Delta}_{j'} f. \tag{1.12}$$

With the introduction of $\dot{\Delta}_j$, we present the following decay estimates for the solution (u, f) constructed in Theorem 1.1.

THEOREM 1.3. *Under the assumptions of Theorem* 1.1, *the global solutions* (u, f) *constructed in Theorem* 1.1 *decay to* $(0, 0)$ *by*

$$\|u(t)\|_{L^2}^2 \le C\Big[(1+t)^{-\frac{1}{2}} + \sum_{j<0}(\|\dot{\Delta}_j u_0\|_{L^2} + \|\dot{\Delta}_j f_0\|_{L^2})e^{-c2^{2j}t}\Big], \tag{1.13}$$

$$\|f(t)\|_{L^2}^2 \le C\Big[(1+t)^{-\frac{3}{2}} + \sum_{j<0}(\|\dot{\Delta}_j u_0\|_{L^2} + \|\dot{\Delta}_j f_0\|_{L^2})2^{2j}e^{-c2^{2j}t}\Big] \tag{1.14}$$

for some $c > 0$. *Moreover, for any* $j \in \mathbf{Z}$,

$$\|(1-\dot{\mathcal{S}}_j)u(t)\|_{L^2}^2 \le C_j(1+t)^{-\frac{1}{2}}, \qquad \|(1-\dot{\mathcal{S}}_j)f(t)\|_{L^2}^2 \le C_j(1+t)^{-\frac{3}{2}}. \tag{1.15}$$

Remark 1.1. (i) The estimates (1.13) and (1.14) can be improved when $\|u(t)\|_{L^2} \le C(1+t)^{-\frac{1}{4}}$. The main reason lies in the fact that in this case $\|u(t)\|_{L^2}$ belongs to L^p for $p > 4$. One may check the proof of Proposition 2.1 for details. Then under the assumption (1.10), we can recover Theorem 1.2 from Theorem 1.3 by using the fact that

$$\sup_{t\ge 0}\left\{\sum_{j\in\mathbf{Z}}(2^{2j}t)^\kappa e^{-c2^{2j}t}\right\} \le C.$$

(ii) The estimate (1.13) implies that $\|u(t)\|_{L^2}$ tends to 0 as t goes to ∞. While thanks to [2, 16], we have

$$\|f(t)\|_{L^2}^2 \le \frac{\varrho(t)}{1+t},$$

where $\varrho(t) \to 0$ as $t \to \infty$.

(iii) Without (1.10), Theorem 1.3 provides a more precise decay estimate for the L^2 norm of $u(t)$ and $f(t)$. For instance, let ϕ be a smooth function with low frequency part satisfying

$$(\mathcal{F}\phi)(\xi)\big|_{|\xi|\le\frac{1}{2}} = \big|\log|\xi|\big|^{-1}|\xi|^{-\frac{3}{2}}. \tag{1.16}$$

It is easy to check that $\phi \notin \dot{H}^{-\kappa}(\mathbf{R}^3)$ for any positive κ. While for any $j < 0$, we have

$$\|\dot{\Delta}_j \phi\|_{L^2}^2 \approx j^{-2}.$$

Let $\omega(t) \overset{\text{def}}{=} \sum_{j>0} j^{-2} e^{-c2^{-2j}t}$ and $\varpi(t) \overset{\text{def}}{=} \sum_{j>0} j^{-2} 2^{-2j} e^{-c2^{-2j}t}$; then there exist $N, C_N > 0$ such that

$$\begin{aligned}
\omega(2^{2k}) &= \sum_{j>0} j^{-2} e^{-c2^{2(k-j)}} \\
&= \sum_{j>k} j^{-2} e^{-c2^{2(k-j)}} + \sum_{0<j\le k} 2^{-2\log_2 j} e^{-c2^{2(k-j)}} \\
&\le \int_k^\infty x^{-2} dx + k^{-2} C_N \sum_{0<j\le k} 2^{2(\log_2 k - \log_2 j)} 2^{-2N(k-j)} \\
&\le C_N k^{-1}.
\end{aligned}$$

By the monotone property of $\omega(t)$ and $\varpi(t)$, we conclude that

$$\omega(t) \le C(\log t)^{-1}, \quad \varpi(t) \le C(1+t)^{-1}(\log t)^{-\frac{3}{2}}.$$

On the other hand, it is easy to construct examples so that the low frequency part of u_0 coincides with $\varepsilon\phi$, and the low frequency part of f_0 coincides with $\varepsilon\phi(x)\psi(q)$ with $\psi \in \mathcal{S}(\mathbf{R}^3)$. Then for ε sufficiently small, let (u, f) be the unique solution of (1.7) and (1.8) obtained in Theorem 1.1. The above calculations show that $u(t)$ has the logarithm decay, while $f(t)$ has a better decay rate than $(1+t)^{-1}(\log t)^{-1}$, which can be obtained by directly using Schonbek's devices [14, 15].

(iv) Under the assumptions of Theorem 1.1, we can provide an explicit decay rate for $\|u(t)\|_{L^p}$ and $\|f(t)\|_{L^p}$ for $p > 2$ (see Corollary 2.1).

Remark 1.2. We can study the decay rates for the global smooth solutions to the Oldroyd model in [10, 3] by using exactly the same procedure as that in this paper. We omit the details here.

2. The proof of Theorem 1.2. Motivated by [14, 15], we shall present the proof of Theorem 1.2 in this section. However, compared with [14, 15], the source term in the Navier–Stokes equations of (1.7) does not decay as fast as what is required in [14, 15]. To deal with this term, we need to use the coupling effect between u and f in (1.7).

Let us first recall the following lemma from [10], which shall be of constant use in what follows.

LEMMA 2.1. *Let f satisfy $\int_{\mathbf{R}^3} f\sqrt{M}\, dq = 0$, and let U satisfy* (1.3). *Then there hold*

$$\begin{aligned}
&\|f\|_{L^2} \le C \left\| \nabla_q f + \frac{1}{2} \nabla_q U f \right\|_{L^2}, \\
&\|\nabla_q U f\|_{L^2} \le C \left\| \nabla_q f + \frac{1}{2} \nabla_q U f \right\|_{L^2}, \\
&\|\nabla_q U q f\|_{L^2} \le C \|(1+|q|^2)^{\frac{1}{2}} \left(\nabla_q f + \frac{1}{2} \nabla_q U f \right) \|_{L^2}.
\end{aligned}$$

Moreover, for any integer $s_1 \geq 0$ and $s_2 \geq 1$, there holds

$$\|\nabla_q U \nabla^{s_1} \nabla_q^{s_2} f\|_{L^2} \leq C \sum_{k=0}^{s_2} \left\| \nabla_q \nabla^{s_1} \nabla_q^k f + \frac{1}{2} \nabla_q U \nabla^{s_1} \nabla_q^k f \right\|_{L^2},$$

$$\|\nabla^{s_1} \nabla_q^{s_2} f\|_{L^2} \leq C \sum_{k=0}^{s_2-1} \left\| \nabla_q \nabla^{s_1} \nabla_q^k f + \frac{1}{2} \nabla_q U \nabla^{s_1} \nabla_q^k f \right\|_{L^2}.$$

As a first step in the proof of Theorem 1.2, we shall prove the following L^2 decay estimate for (u, f).

PROPOSITION 2.1. *Under the assumptions of Theorem* 1.2, *the L^2 norm of the global classical solutions (u, f) to* (1.7) *constructed in Theorem* 1.1 *decay to* $(0, 0)$ *as t goes to ∞ according to*

$$\|u(t)\|_{L^2}^2 + \|f(t)\|_{L^2}^2 \leq C(1+t)^{-b}, \tag{2.1}$$

with $b = \min(\kappa, \frac{3}{2})$.

Proof. We first get by using standard energy estimation to (1.7) that

$$\begin{aligned} &\frac{d}{dt}(\|u\|_{L^2}^2 + \|f\|_{L^2}^2) + 2\mu\|\nabla u\|_{L^2}^2 + 2\sigma \left\| \nabla_q f + \frac{1}{2} \nabla_q U f \right\|_{L^2}^2 \\ &\quad = (\nabla u q \cdot \nabla_q U f,\ f) \leq |qf|_{H^4} \|\nabla u\|_{L^2} \|\nabla_q U f\|_{L^2}, \end{aligned}$$

which together with Lemma 2.1 and Theorem 1.1 applied gives

$$\frac{d}{dt}(\|u\|_{L^2}^2 + \|f\|_{L^2}^2) + \mu\|\nabla u\|_{L^2}^2 + \sigma \left\| \nabla_q f + \frac{1}{2} \nabla_q U f \right\|_{L^2}^2 \leq 0. \tag{2.2}$$

Thanks to (2.2), to use Schonbek's strategy in [14, 15], we need to split the phase space into two time-dependent domains. More precisely, we decompose $\|\nabla u\|_{L^2}^2$ as

$$\|\nabla u\|_{L^2}^2 = \int_{S(t)} |\xi|^2 |\hat{u}(\xi)|^2 d\xi + \int_{S(t)^c} |\xi|^2 |\hat{u}(\xi)|^2 d\xi,$$

where $S(t) \overset{\text{def}}{=} \{\xi : |\xi| \leq C^{\frac{1}{2}}(1+t)^{-\frac{1}{2}}\}$ and the constant C will be chosen later on. Then, thanks to (2.2), we have

$$\frac{d}{dt}(\|u\|_{L^2}^2 + \|f\|_{L^2}^2) + \frac{C\mu}{1+t}\|u\|_{L^2}^2 + \sigma \left\| \nabla_q f + \frac{1}{2} \nabla_q U f \right\|_{L^2}^2 \leq \frac{C\mu}{1+t} \int_{S(t)} |\hat{u}(\xi)|^2 d\xi. \tag{2.3}$$

In what follows, we shall focus on the L^2 estimate to the low frequency part of u. In order to do so, we take Fourier transform with respect to x variables in (1.7) to get

$$\begin{cases} \hat{u}_t + \mathcal{F}_x(u \cdot \nabla u) + i\xi\, \hat{p} = -\mu|\xi|^2 \hat{u} + i\xi \cdot (\int_{\mathbf{R}^3} \sqrt{M} \nabla_q U \otimes q \hat{f}\ dq), \\ \hat{f}_t + \mathcal{F}_x(u \cdot \nabla f) + \mathcal{F}_x(\nabla u q \cdot \nabla_q f) - \sigma(\triangle_q \hat{f} + \frac{\triangle_q U}{2} \hat{f} - \frac{|\nabla_q U|^2}{4} \hat{f}) \\ \qquad = i(\xi_j \hat{u}_i)(\nabla_q U \otimes q\sqrt{M}) + \frac{1}{2}\mathcal{F}_x(\nabla u q \cdot \nabla_q U f), \\ \xi \cdot \hat{u} = \xi \cdot \bar{\hat{u}} = 0, \end{cases} \tag{2.4}$$

where $\hat{f}(\xi,q,t) \overset{\text{def}}{=} \mathcal{F}_x(f)(\xi,q,t)$. Notice that $\xi\cdot\hat{u}=\xi\cdot\bar{\hat{u}}=0$. By multiplying the first equation of (2.4) by $\bar{\hat{u}}(t,\xi)$ and taking the resulting real part we get that

$$\begin{aligned}(2.5)\qquad &\frac{1}{2}\frac{d}{dt}|\hat{u}(\xi,t)|^2+\mathcal{R}e[\mathcal{F}_x(u\cdot\nabla u)\cdot\bar{\hat{u}}(\xi,t)]\\ &\quad=-\mu|\xi|^2|\hat{u}(\xi,t)|^2+\mathcal{R}e\left[i\xi\otimes\bar{\hat{u}}(\xi,t):\left(\int_{\mathbf{R}^3}\sqrt{M}\nabla_q U\otimes q\hat{f}\;dq\right)\right].\end{aligned}$$

A similar process applied to the microscopic equation of (2.4) gives

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int_{\mathbf{R}^3}|\hat{f}(\xi,q,t)|^2dq+\mathcal{R}e\left[\int_{\mathbf{R}^3}\left(\mathcal{F}_x(u\cdot\nabla f)\bar{\hat{f}}(\xi,q,t)+\mathcal{F}_x(\nabla uq\cdot\nabla_q f)\bar{\hat{f}}(\xi,q,t)\right)dq\right]\\ &(2.6)\quad+\sigma\left\|\nabla_q\hat{f}+\frac{1}{2}\nabla_q U\hat{f}\right\|^2_{L^2_q}\\ &=\mathcal{R}e\left[i\xi\otimes\hat{u}(\xi,t):\left(\int_{\mathbf{R}^3}\sqrt{M}\nabla_q U\otimes q\bar{\hat{f}}(\xi,q,t)\;dq\right)\right]\\ &\quad+\frac{1}{2}\mathcal{R}e\left[\int_{\mathbf{R}^3}\mathcal{F}_x(\nabla uq\cdot\nabla_q Uf)\,\bar{\hat{f}}(\xi,q,t)dq\right].\end{aligned}$$

Thanks to (2.5)–(2.6), we obtain that for any $\delta>0$, there is a C_δ such that

$$\begin{aligned}&\frac{d}{dt}\left(|\hat{u}(\xi,t)|^2+\int_{\mathbf{R}^3}|\hat{f}(\xi,q,t)|^2dq\right)+\mu|\xi|^2|\hat{u}(\xi,t)|^2+2\sigma\left\|\nabla_q\hat{f}+\frac{1}{2}\nabla_q U\hat{f}\right\|^2_{L^2_q}\\ &\quad\le C|\mathcal{F}_x(u\otimes u)|^2+C_\delta\int_{\mathbf{R}^3}(|\mathcal{F}_x(u\cdot\nabla f)|^2+|\mathcal{F}_x(\nabla uq\cdot\nabla_q f)|^2\\ &\qquad+|\mathcal{F}_x(\nabla uq\cdot\nabla_q Uf)|^2)\,dq+\delta\int_{\mathbf{R}^3}|\bar{\hat{f}}(\xi,q,t)|^2dq,\end{aligned}$$

from which we deduce that

$$\begin{aligned}&|\hat{u}(\xi,t)|^2+\int_{\mathbf{R}^3}|\hat{f}(\xi,q,t)|^2dq+2\sigma\int_0^t\int_{\mathbf{R}^3}e^{-\mu|\xi|^2(t-s)}\left|\nabla_q\hat{f}+\frac{1}{2}\nabla_q U\hat{f}\right|^2dqds\\ &\quad\le e^{-\mu|\xi|^2t}\left(|\hat{u_0}(\xi)|^2+\int_{\mathbf{R}^3}|\hat{f}_0(\xi,q)|^2dq\right)\\ &\qquad+\int_0^t\int_{\mathbf{R}^3}e^{-\mu|\xi|^2(t-s)}(\mu|\xi|^2+\delta)|\hat{f}(\xi,q,s)|^2dqds\\ &\quad+C\int_0^t|\mathcal{F}_x(u\otimes u)|^2ds+C_\delta\int_0^t\int_{\mathbf{R}^3}(|\mathcal{F}_x(u\cdot\nabla f)(\xi,q,s)|^2\\ &(2.7)\qquad+|\mathcal{F}_x(\nabla uq\cdot\nabla_q f)(\xi,q,s)|^2+|\mathcal{F}_x(\nabla uq\cdot\nabla_q Uf)(\xi,q,s)|^2)\,dq\,ds.\end{aligned}$$

Now we are in a position to estimate the right-hand side of (2.3). First, it is easy to observe that

$$\int_0^t \int_{\mathbf{R}^3} \int_{S(t)} e^{-|\xi|^2(t-s)}(\mu|\xi|^2+\delta)|\hat{f}(\xi,q,s)|^2 d\xi\, dq\, ds$$

$$\leq \left(\frac{C\mu}{1+t}+\delta\right)\int_0^t \int_{\mathbf{R}^3} \int_{S(t)} e^{-\mu|\xi|^2(t-s)}|\hat{f}(\xi,q,s)|^2 d\xi dq ds.$$

While noticing that $\int_{\mathbf{R}^3} \sqrt{M}\hat{f} dq = 0$, applying Lemma 2.1 gives

$$\int_{\mathbf{R}^3} \int_{S(t)} e^{-\mu|\xi|^2(t-s)}|\hat{f}(\xi,q,s)|^2 d\xi dq = \int_{S(t)} e^{-\mu|\xi|^2(t-s)} \int_{\mathbf{R}^3} |\hat{f}|^2\, dq\, d\xi$$

$$\leq C\int_{\mathbf{R}^3} \int_{S(t)} e^{-\mu|\xi|^2(t-s)} \left|\nabla_q \hat{f} + \frac{1}{2}\nabla_q U \hat{f}\right|^2 d\xi dq.$$

To deal with the remaining terms (2.7), we introduce $\phi \in \mathcal{S}(\mathbf{R}^3)$, the Fourier transform of which satisfies

$$\hat{\phi}(\xi) = \begin{cases} 1, & |\xi| \leq 2C^{\frac{1}{2}}, \\ 0, & |\xi| \geq 3C^{\frac{1}{2}}, \end{cases}$$

and we denote $\phi_t(x) \overset{\text{def}}{=} (1+t)^{-\frac{3}{2}}\phi((1+t)^{-\frac{1}{2}}x)$. Then, by using Young's inequality and (1.9) we get that

$$\int_0^t \int_{S(t)} |\mathcal{F}_x(u\otimes u)|^2 d\xi ds \leq \int_0^t \int_{\mathbf{R}^3} |(u\otimes u) * \phi_t|^2\, dx\, ds$$

$$\leq C(1+t)^{-\frac{3}{2}} \int_0^t \|u\otimes u\|_{L^1}^2 ds \leq C(1+t)^{-\frac{3}{2}} \int_0^t \|u\|_{L^2}^4 ds$$

$$\leq C(1+t)^{-\frac{1}{2}}. \tag{2.8}$$

While thanks to Minkowski's inequality and (1.9), we obtain

$$\int_0^t \int_{\mathbf{R}^3} \int_{S(t)} |\mathcal{F}_x(u\cdot\nabla f)|^2\, d\xi\, dq\, ds \leq \int_0^t \int_{\mathbf{R}^3} \int_{\mathbf{R}^3} |(u\cdot\nabla f) * \phi_t|^2 dx dq ds$$

$$\leq \int_0^t \int_{\mathbf{R}^3} \Big| (|u|\|\nabla f\|_{L^2_q}) * |\phi_t| \Big|^2 dx ds \leq C(1+t)^{-\frac{3}{2}} \int_0^t \Big\| |u|\|\nabla f\|_{L^2_q} \Big\|_{L^1_x}^2 ds$$

$$\leq C(1+t)^{-\frac{3}{2}} \int_0^t \|u\|_{L^2}^2 \|\nabla f\|_{L^2}^2 ds \leq C(1+t)^{-\frac{3}{2}} \int_0^t |\nabla_q f + \frac{1}{2}\nabla_q U f|_{H^4}^2 ds$$

$$\leq C(1+t)^{-\frac{3}{2}}.$$

A similar argument for the last two terms in (2.7) yields

$$\int_0^t \int_{\mathbf{R}^3} \int_{S(t)} |\mathcal{F}_x(\nabla u q \cdot \nabla_q f)|^2 d\xi dq ds \le C(1+t)^{-\frac{3}{2}} \int_0^t \Big\| |\nabla u| \|q \nabla_q f\|_{L^2_q} \Big\|^2_{L^1_x} ds$$

$$\le C(1+t)^{-\frac{3}{2}} \int_0^t \|q\nabla_q f\|^2_{L^2} ds \le C(1+t)^{-\frac{3}{2}} \int_0^t \Big\| (1+|q|^2)^{\frac{1}{2}} \Big(\nabla_q f + \frac{1}{2} \nabla_q U f \Big) \Big\|^2_{L^2} ds$$

$$\le C(1+t)^{-\frac{3}{2}}$$

and

$$\int_0^t \int_{\mathbf{R}^3} \int_{S(t)} |\mathcal{F}_x(\nabla u q \cdot \nabla_q U f)|^2 d\xi dq ds \le C(1+t)^{-\frac{3}{2}}.$$

Therefore, thanks to (2.7), we obtain

$$\int_{S(t)} |\hat{u}(\xi,t)|^2 d\xi \le \int_{S(t)} e^{-\mu|\xi|^2 t} |\hat{u}(\xi,0)|^2 d\xi + \int_{S(t)} \int_{\mathbf{R}^3} e^{-\mu|\xi|^2 t} |\hat{f}(\xi,q,0)|^2 d\xi dq$$
$$+ C(1+t)^{-\frac{1}{2}}$$

for t large enough. Then with the additional assumption (1.10), we conclude

$$\int_{S(t)} |\hat{u}(\xi,t)|^2 d\xi \le C(1+t)^{-\kappa} + C(1+t)^{-\frac{1}{2}}, \tag{2.9}$$

for t large enough. This together with (2.3) ensures that

$$\frac{d}{dt}(\|u\|^2_{L^2} + \|f\|^2_{L^2}) + \frac{C\mu}{1+t}\|u\|^2_{L^2} + \sigma \left\| \nabla_q f + \frac{1}{2}\nabla_q U f \right\|^2_{L^2} \le \frac{C\mu}{1+t}(1+t)^{-b},$$

where $b = \min(\kappa, \frac{1}{2})$, from which we deduce that

$$\|u(t)\|^2_{L^2} + \|f(t)\|^2_{L^2} \le C(1+t)^{-b}. \tag{2.10}$$

With the above decay estimate for $\|u(t)\|_{L^2}$, we can improve the estimate in (2.8). Suppose that $\kappa > \frac{1}{2}$; then $b = \frac{1}{2}$ in (2.10) and

$$\int_0^t \int_{S(t)} |\mathcal{F}_x(u \otimes u)|^2 d\xi ds \le C(1+t)^{-\frac{3}{2}} \int_0^t \|u\|^4_{L^2} ds$$

$$\le C(1+t)^{-\frac{7}{6}} \left(\int_0^t \|u\|^6_{L^2} ds \right)^{\frac{2}{3}}$$

$$\le C(1+t)^{-\frac{7}{6}}.$$

Then the proof of (2.9) implies that

$$\int_{S(t)} |\hat{u}(\xi,t)|^2 d\xi \le C(1+t)^{-\kappa} + C(1+t)^{-\frac{7}{6}},$$

for t large enough, from which, with the proof of (2.10), we deduce that

$$\|u(t)\|^2_{L^2} + \|f(t)\|^2_{L^2} \le C(1+t)^{-b}, \tag{2.11}$$

with $b=\min(\kappa,\frac{7}{6})$. Now if $\kappa>\frac{7}{6}$, then by using (2.11) we get that

$$\int_0^t\int_{S(t)}|\mathcal{F}_x(u\otimes u)|^2d\xi ds\le C(1+t)^{-\frac{3}{2}}\int_0^t\|u\|_{L^2}^4ds$$
$$\le C(1+t)^{-\frac{3}{2}}.$$

Then the proof of (2.9) yields

$$\int_{S(t)}|\hat{u}(\xi,t)|^2d\xi\le C(1+t)^{-\kappa}+C(1+t)^{-\frac{3}{2}},$$

from which, with the proof of (2.10), we obtain

$$\|u(t)\|_{L^2}^2+\|f(t)\|_{L^2}^2\le C(1+t)^{-b},$$

with $b=\min(\kappa,\frac{3}{2})$. This completes the proof of the proposition. □

With Proposition 2.1, to complete the proof of Theorem 1.2, we still need to improve the decay rate for the L^2 norm of f. In order to do so, we shall first prove the L^2 decays for the derivatives of u and f, which, in particular, implies a sort of L^p decay estimate for u,f without any additional assumption on the initial data.

PROPOSITION 2.2. *Under the assumptions of Theorem* 1.1, *we have the following decay estimate for* $\nabla u,\nabla f$, *and* f*:*

$$\|\nabla u(t)\|_{L^2}^2+\|\nabla f(t)\|_{L^2}^2+\|f(t)\|_{L^2}^2\le C(1+t)^{-1}. \tag{2.12}$$

Proof. Taking ∇ to (1.7), by a standard energy estimation we get that

$$\frac{d}{dt}(\|\nabla u\|_{L^2}^2+\|\nabla f\|_{L^2}^2)+2\mu\|\nabla^2u\|_{L^2}^2+2\sigma\left\|\nabla_q\nabla f+\frac{1}{2}\nabla_qU\nabla f\right\|_{L^2}^2$$
$$=-2(\nabla(u\cdot\nabla u),\nabla u)-2(\nabla(u\cdot\nabla f),\nabla f)$$
$$-2(\nabla(\nabla uq\cdot\nabla_qf),\nabla f)+(\nabla(\nabla uq\cdot\nabla_qUf),\ \nabla f) \tag{2.13}$$

Thanks to (1.9), we have

$$|(\nabla(u\cdot\nabla u),\nabla u)|=|(\nabla(u\otimes u),\nabla\nabla u)|$$
$$\le C\|u\|_{L^3}\|\nabla u\|_{L^6}\|\nabla^2u\|_{L^2}\le C\epsilon\|\nabla^2u\|_{L^2}^2.$$

Similarly, as $\operatorname{div}u=0$ and $\operatorname{div}_q(\nabla uq)=0$, we have

$$|(\nabla(u\cdot\nabla f),\nabla f)|=|((\nabla u)\cdot\nabla f,\nabla f)|\le\|\nabla u\|_{H^2}\|\nabla f\|_{L^2}^2,$$
$$|(\nabla(\nabla uq\cdot\nabla_qf),\nabla f)|=|((\nabla\nabla u)q\cdot\nabla_qf,\nabla f)|$$
$$\le\|\nabla_qf\|_{H^4}\|\nabla^2u\|_{L^2}\|q\nabla f\|_{L^2},$$

and

$$|(\nabla(\nabla uq\cdot\nabla_qUf),\nabla f)|$$
$$=|(\nabla(\nabla u)q\cdot\nabla_qUf,\nabla f)+(\nabla uq\cdot\nabla_qU(\nabla f),\nabla f)|$$
$$\le|qf|_{H^4}\|\nabla^2u\|_{L^2}\|\nabla_qU\nabla f\|_{L^2}+\|\nabla u\|_{H^2}\|q\nabla f\|_{L^2}\|\nabla_qU\nabla f\|_{L^2}.$$

Then we deduce from (1.9), (2.13) that

$$\frac{d}{dt}(\|\nabla u\|_{L^2}^2+\|\nabla f\|_{L^2}^2)+\mu\|\nabla^2u\|_{L^2}^2+\sigma\left\|\nabla_q\nabla f+\frac{1}{2}\nabla_qU\nabla f\right\|_{L^2}^2\le0. \tag{2.14}$$

Then Schonbek's strategy [14, 15] applied to (2.14) gives

$$\frac{d}{dt}(\|\nabla u\|_{L^2}^2+\|\nabla f\|_{L^2}^2)+\frac{C\mu}{1+t}\|\nabla u\|_{L^2}^2 + \sigma\left\|\nabla_q(\nabla f)+\frac{1}{2}\nabla_q U(\nabla f)\right\|_{L^2}^2 \le \frac{C\mu}{1+t}\int_{S(t)}|\xi|^2|\hat{u}(\xi,t)|^2 d\xi, \tag{2.15}$$

from which we deduce that

$$\|\nabla u(t)\|_{L^2}^2+\|\nabla f(t)\|_{L^2}^2 \le C(1+t)^{-1}. \tag{2.16}$$

With (2.16), we now turn to the proof of the L^2 decay of f. First by a standard energy estimation we get that

$$\frac{1}{2}\frac{d}{dt}\|f\|_{L^2}^2+\sigma\left\|\nabla_q f+\frac{1}{2}\nabla_q U f\right\|_{L^2}^2 \le \|\nabla u\|_{L^2}\|\sqrt{M}\nabla_q U\|_{L^2}\|qf\|_{L^2}+|qf|_{H^4}\|\nabla u\|_{L^2}\|\nabla_q U f\|_{L^2}.$$

Then thanks to Lemma 2.1 and (1.9), we obtain

$$\frac{d}{dt}\|f\|_{L^2}^2+\sigma\|f\|_{L^2}^2 \le C\|\nabla u\|_{L^2}^2 \le C(1+t)^{-1}, \tag{2.17}$$

from which we deduce that

$$\|f(t)\|_{L^2}^2 \le C\|f_0\|_{L^2}^2 e^{-\sigma t}+\int_0^t e^{-\sigma(t-s)}(1+s)^{-1}ds. \le C(1+t)^{-1},$$

which together with (2.16) gives (2.12). This completes the proof of Proposition 2.2. □

Remark 2.1. Assume that $\|u(t)\|_{L^2}^2 \le C(1+t)^{-b}$; then, thanks to (2.15), we can improve (2.16) to $\|\nabla u(t)\|_{L^2}^2+\|\nabla f(t)\|_{L^2}^2 \le C(1+t)^{-b-1}$, which, in turn, implies that $\|f(t)\|_{L^2}^2 \le C(1+t)^{-b-1}$.

An immediate corollary of Proposition 2.2 and Sobolev imbedding is the following corollary.

COROLLARY 2.1. *Under the assumption of Theorem* 1.1, *we have*

$$\|u(t)\|_{L^p}^2+\|f(t)\|_{L_q^2(L_x^p)}^2 \le \begin{cases} C_p(1+t)^{-\frac{6}{p}} & \text{if} \quad 6\le p\le\infty, \\ C_p(1+t)^{-\frac{3(p-2)}{2p}} & \text{if} \quad 2<p\le 6. \end{cases}$$

Now we are in a position to complete the proof of Theorem 1.2.

Proof of Theorem 1.2. Combining Proposition 2.1 and Remark 2.1, we complete the proof of the theorem. □

3. The proof of Theorem 1.3. In this section, we consider the L^2 decay of u without the additional assumption (1.10). It is easy to observe from the proof of Theorem 1.2 that the low frequency part of initial data makes this problem difficult. To get rid of this difficulty, we are going to study the decay of solution to the linearized

problem of (1.7) first, then we estimate the difference between the solution of the linearized equation (3.1) and that of (1.7).

The linearized equation of (1.7) reads as

$$(3.1)\quad \begin{cases} v_t + \nabla h = \mu\triangle v + \nabla\cdot(\int_{\mathbf{R}^3}\sqrt{M}\nabla_q U\otimes qg\ dq), & x\in\mathbf{R}^3,\\ g_t - \sigma(\triangle_q g + \frac{\triangle_q U}{2}g - \frac{|\nabla_q U|^2}{4}g) = \sqrt{M}\nabla v q\cdot\nabla_q U, & (x,q)\in\mathbf{R}^3\times\mathbf{R}^3,\\ \text{div } v = 0, \quad x\in\mathbf{R}^3, \end{cases}$$

together with the initial condition

$$(3.2)\qquad v|_{t=0} = u_0, \qquad g|_{t=0} = f_0 \quad \text{for} \quad (x,q)\in\mathbf{R}^3\times\mathbf{R}^3\,.$$

Notice from (3.1) that

$$\frac{d}{dt}\int_{\mathbf{R}^3} g\sqrt{M}dq = 0,$$

which together with the fact that $\int_{\mathbf{R}^3} f_0\sqrt{M}dq = 0$ ensures that

$$\int_{\mathbf{R}^3} g\sqrt{M}dq = 0,$$

so that Lemma 2.1 can still be applied for g.

PROPOSITION 3.1. *Under the assumptions of Theorem* 1.1, (3.1)–(3.2) *has a unique solution* (v,g) *so that*

$$(3.3)\quad \sup_{0\le t<\infty}\Big(\|v(t)\|_{H^s}^2 + |qg|_{H^4}^2 + \|g(t)\|_{H^s}^2\Big) + \int_0^t\Big[\mu\|\nabla v\|_{H^s}^2 + \sigma\Big(||q|\Big(\nabla_q g + \frac{1}{2}\nabla_q U g\Big)|_{H^4}^2 + \Big\|\nabla_q g + \frac{1}{2}\nabla_q U g\Big\|_{H^s}^2\Big)\Big]d\tau \le C\varepsilon\min(\mu,\sigma).$$

Furthermore, there holds

$$\||D|^\lambda v\|_{L^2}^2 \le C_\lambda(1+t)^{-\lambda},$$

$$(3.4)\qquad \||D|^\lambda g\|_{L^2}^2 + \||D|^\lambda\nabla_q g\|_{L^2}^2 + \||q||D|^\lambda g\|_{L^2}^2 \le C_\lambda(1+t)^{-\lambda-1},$$

for $\lambda\in[0,s]$, *where* $|D|^\lambda$ *is the Fourier multiplier with the symbol* $|\xi|^\lambda$.

Proof. A standard functional analysis method can be applied to prove the global existence of a smooth solution to (3.1). Furthermore, a similar proof of (1.9) in [10] ensures that (v,g) satisfies (3.3). We omit the details here.

Now let us turn to the long time behavior of (v,g). First, let $\lambda\in\mathbf{N}$. We get by using standard energy estimation that

$$\frac{d}{dt}(\|\nabla^\lambda v\|_{L^2}^2 + \|\nabla^\lambda g\|_{L^2}^2) + 2\mu\|\nabla\nabla^\lambda v\|_{L^2}^2 + 2\sigma\Big\|\nabla_q(\nabla^\lambda g) + \frac{1}{2}\nabla_q U(\nabla^\lambda g)\Big\|_{L^2}^2 = 0.$$

Schonbek's strategy applied gives

$$\frac{d}{dt}(\|\nabla^\lambda v\|_{L^2}^2 + \|\nabla^\lambda g\|_{L^2}^2) + \frac{\lambda+1}{1+t}\|\nabla^\lambda v\|_{L^2}^2 + 2\sigma\Big\|\nabla_q(\nabla^\lambda g) + \frac{1}{2}\nabla_q U(\nabla^\lambda g)\Big\|_{L^2}^2 \le \frac{\lambda+1}{1+t}\int_{|\xi|\le(\lambda+1)^{\frac12}(1+t)^{-\frac12}}|\mathcal{F}(\nabla^\lambda v)(\xi)|^2 d\xi \le C_\lambda(1+t)^{-\lambda-1},$$

from which we deduce that

$$\|\nabla^\lambda v\|_{L^2}^2 + \|\nabla^\lambda g\|_{L^2}^2 \le C_\lambda (1+t)^{-\lambda}. \tag{3.5}$$

On the other hand, by applying ∇^λ and $\nabla^\lambda \nabla_q$ to the microscopic equation (3.1) and taking the L^2 inner product of the resulting equations with $\nabla^\lambda g$ and $\nabla^\lambda \nabla_q g$, respectively, we obtain

$$\frac{d}{dt}\|\nabla^\lambda g\|_{L^2}^2 + \sigma \left\| \nabla_q(\nabla^\lambda g) + \frac{1}{2}\nabla_q U(\nabla^\lambda g)\right\|_{L^2}^2 \le C\|\nabla\nabla^\lambda v\|_{L^2}^2$$

and

$$\begin{aligned}
&\frac{d}{dt}\|\nabla^\lambda \nabla_q g\|_{L^2}^2 + \sigma \left\| \nabla_q(\nabla^\lambda \nabla_q g) + \frac{1}{2}\nabla_q U(\nabla^\lambda \nabla_q g)\right\|_{L^2}^2 \\
&\quad \le C(\|\nabla_q U \nabla^\lambda g\|_{L^2}\|\nabla_q U \nabla^\lambda \nabla_q g\|_{L^2} + \|\nabla^\lambda g\|_{L^2}\|\nabla^\lambda \nabla_q g\|_{L^2}) \\
&\qquad + \|\nabla\nabla^\lambda v\|_{L^2}\|\nabla^\lambda \nabla_q g\|_{L^2},
\end{aligned}$$

where we used (1.3) and (1.4) in the estimate of the second inequality.

In order to get the estimate for $\||q|\nabla^\lambda g\|_{L^2}$, we apply ∇^λ to the microscopic equation (3.1) and then take the L^2 inner product of the resulting equation with $|q|^2\nabla^\lambda g$ to get

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\||q|\nabla^\lambda g\|_{L^2}^2 + \sigma \left\| |q|\left(\nabla_q(\nabla^\lambda g) + \frac{1}{2}\nabla_q U(\nabla^\lambda g)\right)\right\|_{L^2}^2 \\
&\quad \le \|\nabla^\lambda g\|_{L^2}^2 + \int_{\mathbf{R}^3\times\mathbf{R}^3} |\nabla_q U q||\nabla^\lambda g|^2 dqdx \\
&\qquad + \|\nabla^{1+\lambda} v\|_{L^2}\left(\int_{\mathbf{R}^3} (M|q|^2)dq\right)^{\frac{1}{2}} \|q\nabla_q U\nabla^\lambda g\|_{L^2}.
\end{aligned}$$

Therefore, by using Lemma 2.1 we get that

$$\begin{aligned}
&\frac{d}{dt}\left(\|\nabla^\lambda g\|_{L^2}^2 + c\|\nabla^\lambda \nabla_q g\|_{L^2}^2 + c\||q|\nabla^\lambda g\|_{L^2}^2\right) + \sigma\left(\left\| \nabla_q(\nabla^\lambda g) + \frac{1}{2}\nabla_q U(\nabla^\lambda g)\right\|_{L^2}^2 \right. \\
&\quad \left. + c\left\| \nabla_q(\nabla^\lambda \nabla_q g) + \frac{1}{2}\nabla_q U(\nabla^\lambda \nabla_q g)\right\|_{L^2}^2 + c\left\| |q|\left(\nabla_q(\nabla^\lambda g) + \frac{1}{2}\nabla_q U(\nabla^\lambda g)\right)\right\|_{L^2}^2\right) \\
&\le C\|\nabla\nabla^\lambda v\|_{L^2}^2,
\end{aligned}$$

by taking c sufficiently small, which implies

$$\|\nabla^\lambda g(t)\|_{L^2}^2 + \|\nabla^\lambda \nabla_q g(t)\|_{L^2}^2 + \||q|\nabla^\lambda g(t)\|_{L^2}^2 \le C_\lambda (1+t)^{-\lambda-1}. \tag{3.6}$$

With (3.5) and (3.6), we complete the proof of (3.4) via a trivial interpolation argument. □

Remark 3.1. Plugging the second inequality of (3.4) into the v equation of (3.1) and using Schonbek's strategy in [14, 15], we cannot obtain an improved decay estimate for v. In particular, the decay rate of $\|v(t)\|_{L^2}$ is not as good as that in [14, 15].

The main reason lies in the fact that elastic stress τ defined in (3.1) does not decay as fast as the source term in [14, 15]. A similar remark applies to (1.7) as well.

PROPOSITION 3.2. *Let (u, f) be the unique solution of* (1.7) *and* (1.8) *obtained in Theorem* 1.1, *and let (v, g) be the unique solution of* (3.1) *and* (3.2) *obtained in Proposition* 3.1. *Then there hold*

$$\|(u-v)(t)\|_{L^2}^2 \le C(1+t)^{-\frac{1}{2}},$$

$$\|(f-g)(t)\|_{L^2}^2 \le C(1+t)^{-\frac{3}{2}}. \tag{3.7}$$

Proof. Thanks to (1.7) and (3.1), we obtain

(3.8)
$$\begin{cases} (u-v)_t + u\cdot\nabla u + \nabla(p-h) \\ \quad = \mu\triangle(u-v) + \nabla\cdot\left(\displaystyle\int_{\mathbf{R}^3}\sqrt{M}\nabla_q U\otimes q(f-g)\,dq\right), \quad x\in\mathbf{R}^3, \\ (f-g)_t + u\cdot\nabla f + \nabla u q\cdot\nabla_q f - \sigma\left(\triangle_q(f-g) + \dfrac{\triangle_q U}{2}(f-g) - \dfrac{|\nabla_q U|^2}{4}(f-g)\right) \\ \quad = \sqrt{M}\nabla(u-v)q\cdot\nabla_q U + \dfrac{1}{2}\nabla u q\cdot\nabla_q U f, \qquad (x,q)\in\mathbf{R}^3\times\mathbf{R}^3, \\ \text{div }(u-v) = 0, \quad x\in\mathbf{R}^3. \end{cases}$$

Note that as $\operatorname{div} u = 0$ and $\operatorname{div} v = 0$, by standard energy estimates we get that

$$\begin{aligned} &\frac{d}{dt}(\|u-v\|_{L^2}^2 + \|f-g\|_{L^2}^2) + 2\mu\|\nabla(u-v)\|_{L^2}^2 + 2\sigma\left\|\nabla_q(f-g) + \frac{1}{2}\nabla_q U(f-g)\right\|_{L^2}^2 \\ &\quad = -2(u\cdot\nabla u, u-v) - 2(u\cdot\nabla f, f-g) - 2(\nabla u q\cdot\nabla_q f, f-g) \\ &\qquad + (\nabla u q\cdot\nabla_q U f,\ f-g) \\ &\quad = 2(u\cdot\nabla v, u-v) + 2(u\cdot\nabla g, f-g) + 2(\nabla u q\cdot\nabla_q g, f-g) \\ &\qquad + (\nabla u q\cdot\nabla_q U(f-g),\ f-g) + (\nabla u q\cdot\nabla_q U g,\ f-g). \end{aligned}$$

Thanks to (1.9) and Lemma 2.1, we have

$$\begin{aligned} |(\nabla u q\cdot\nabla_q U(f-g),\ f-g)| &\le \|\nabla u\|_{H^2}\|q(f-g)\|_{L^2}\|\nabla_q U(f-g)\|_{L^2} \\ &\le \epsilon\min\{\mu,\sigma\}\left\|\nabla_q(f-g) + \frac{1}{2}\nabla_q U(f-g)\right\|_{L^2}^2. \end{aligned}$$

Then by taking ϵ small enough we get

$$\begin{aligned} &\frac{d}{dt}(\|u-v\|_{L^2}^2 + \|f-g\|_{L^2}^2) + \mu\|\nabla(u-v)\|_{L^2}^2 + \sigma\left\|\nabla_q(f-g) + \frac{1}{2}\nabla_q U(f-g)\right\|_{L^2}^2 \\ &\quad \le 2|(u\cdot\nabla v, u-v)| + 2|(u\cdot\nabla g, f-g)| + 2|(\nabla u q\cdot\nabla_q g, f-g)| \\ &\qquad + |(\nabla u q\cdot\nabla_q U g,\ f-g)|. \end{aligned} \tag{3.9}$$

In what follows, we shall deal with the right-hand side of (3.9) term by term. Our main idea is to make full use of the decay estimate to the solutions of the linear system

(3.4). First of all, thanks to (1.9), (2.12), and (3.4), we have

$$\begin{aligned}|(u\cdot\nabla v, u-v)| &\le C\|u\|_{L^4}\|v\|_{L^4}\|\nabla(u-v)\|_{L^2}\\ &\le C_\delta\|u\|^2_{\dot H^{\frac34}}\|v\|^2_{\dot H^{\frac34}}+\delta\|\nabla(u-v)\|^2_{L^2}\\ &\le C_\delta(1+t)^{-\frac32}+\delta\|\nabla(u-v)\|^2_{L^2},\end{aligned}$$

for any $\delta>0$. A similar procedure applied to the remaining terms in (3.9) gives

$$\begin{aligned}|(u\cdot\nabla g, f-g)| &\le C\int_{\mathbf{R}^3}\|u\|_{L^4}\|f-g\|_{L^2_x}\|\nabla g\|_{L^4_x}dq\\ &\le C\|u\|_{\dot H^{\frac34}}\|f-g\|_{L^2}|g|_{\dot H^{\frac74}}\le C_\delta(1+t)^{-\frac52}+\delta\|f-g\|^2_{L^2},\end{aligned}$$

$$\begin{aligned}|(\nabla uq\cdot\nabla_q g, f-g)| &\le C\|\nabla u\|_{\dot H^{\frac34}}|\nabla_q g|_{\dot H^{\frac34}}\|q(f-g)\|_{L^2}\\ &\le C_\delta(1+t)^{-2}+\delta\|q(f-g)\|^2_{L^2},\end{aligned}$$

and

$$\begin{aligned}|(\nabla uq\cdot\nabla_q Ug, f-g)| &\le C\|\nabla u\|_{\dot H^{\frac34}}|qg|_{\dot H^{\frac34}}\|\nabla_q U(f-g)\|_{L^2}\\ &\le C_\delta(1+t)^{-2}+\delta\|\nabla_q U(f-g)\|^2_{L^2}.\end{aligned}$$

Then taking δ small enough, we deduce from (3.9) and Schonbek's strategy in [14, 15] that

$$\begin{aligned}&\frac{d}{dt}(\|u-v\|^2_{L^2}+\|f-g\|^2_{L^2})+\frac{C\mu}{t+1}\|u-v\|^2_{L^2}+\sigma\left\|\nabla_q(f-g)+\frac12\nabla_q U(f-g)\right\|^2_{L^2}\\ &\le\frac{C\mu}{t+1}\int_{S(t)}|\mathcal{F}(u-v)(\xi)|^2d\xi+C(1+t)^{-\frac32}.\end{aligned}\tag{3.10}$$

Now we repeat the procedure in section 2 to estimate the low frequency of $u-v$. First, thanks to (3.8), we have

$$\begin{aligned}&\frac{d}{dt}|\mathcal{F}(u-v)(\xi,t)|^2+\frac{d}{dt}\int_{\mathbf{R}^3}|\mathcal{F}_x(f-g)(\xi,q,t)|^2dq\\ &\quad+\mu|\xi|^2|\mathcal{F}(u-v)(\xi,t)|^2+2\sigma\left|\nabla_q\mathcal{F}_x(f-g)+\frac12\nabla_q U\mathcal{F}_x(f-g)\right|^2_{L^2_q}\\ &\le C|\mathcal{F}(u\otimes u)|^2+C_\delta\int_{\mathbf{R}^3}\Big(|\mathcal{F}_x(u\cdot\nabla f)|^2+|\mathcal{F}_x(\nabla uq\cdot\nabla_q f)|^2\\ &\quad+|\mathcal{F}_x(\nabla uq\cdot\nabla_q Uf)|^2\Big)dq+\delta\int_{\mathbf{R}^3}|\mathcal{F}(f-g)(\xi,q,t)|^2dq,\end{aligned}$$

which implies

$$
\begin{aligned}
&|\mathcal{F}(u-v)(\xi,t)|^2+\int_{\mathbf{R}^3}|\mathcal{F}_x(f-g)(\xi,q,t)|^2dq\\
&\quad+2\sigma\int_0^t\int_{\mathbf{R}^3}e^{-\mu|\xi|^2(t-s)}\left|\nabla_q\mathcal{F}_x(f-g)+\frac{1}{2}\nabla_qU\mathcal{F}_x(f-g)\right|^2dqds\\
&\le\int_0^t\int_{\mathbf{R}^3}e^{-|\xi|^2(t-s)}(\delta+\mu|\xi|^2)|\mathcal{F}(f-g)(\xi,q,s)|^2\,dq\,ds+C\int_0^t|\mathcal{F}(u\otimes u)|^2ds\\
&\quad+C\int_0^t\int_{\mathbf{R}^3}(|\mathcal{F}_x(u\cdot\nabla f)(\xi,q,s)|^2+|\mathcal{F}_x(\nabla uq\cdot\nabla_qf)(\xi,q,s)|^2\\
&\quad+|\mathcal{F}_x(\nabla uq\cdot\nabla_qUf)(\xi,q,s)|^2)\,dqds,
\end{aligned}
$$

from which, with the proof of Proposition 2.1, we deduce that

$$
\int_{S(t)}|\mathcal{F}(u-v)(\xi,t)|^2d\xi\le C(1+t)^{-\frac{1}{2}}. \tag{3.11}
$$

Plugging (3.11) into (3.10), we get for large enough t that

$$
\frac{d}{dt}\left((1+t)^{C\mu}\|u-v\|_{L^2}^2+(1+t)^{C\mu}\|f-g\|_{L^2}^2\right)\le C(1+t)^{C\mu-\frac{3}{2}}.
$$

Let us choose C in the definition of $S(t)$ to be large enough so that $C\mu\ge 2$. Then, we obtain

$$
\|(u-v)(t)\|_{L^2}^2+\|(f-g)(t)\|_{L^2}^2\le C(1+t)^{-\frac{1}{2}}. \tag{3.12}
$$

A similar procedure applied to $\nabla(u-v)$ and $\nabla(f-g)$ gives

$$
\|\nabla(u-v)(t)\|_{L^2}^2+\|\nabla(f-g)(t)\|_{L^2}^2\le C(1+t)^{-\frac{3}{2}}. \tag{3.13}
$$

With (3.12) and (3.13), we are in a position to prove (3.7). In fact, standard energy estimation applied to the microscopic equation of (3.8) yields

$$
\begin{aligned}
&\frac{d}{dt}\|f-g\|_{L^2}^2+2\sigma\left\|\nabla_q(f-g)+\frac{1}{2}\nabla_qU(f-g)\right\|_{L^2}^2\\
&\quad=-2(u\cdot\nabla f,f-g)+2(\sqrt{M}\nabla(u-v)q\cdot\nabla_qU,f-g)\\
&\qquad-2(\nabla uq\cdot\nabla_qf,f-g)+(\nabla uq\cdot\nabla_qUf,\ f-g)\\
&\quad=2(u\cdot\nabla g,f-g)+2(\sqrt{M}\nabla(u-v)q\cdot\nabla_qU,f-g)+(\nabla uq\cdot\nabla_qg,f-g)\\
&\qquad+(\nabla uq\cdot\nabla_qU(f-g),\ f-g)+(\nabla uq\cdot\nabla_qUg,\ f-g),
\end{aligned}
$$

from which, with the arguments from (3.9) to (3.10), we obtain

$$
\|(f-g)(t)\|_{L^2}^2\le C(1+t)^{-\frac{3}{2}}.
$$

This together with (3.12) completes the proof of (3.7). □

Now we present the proof of Theorem 1.3.

Proof of Theorem 1.3. Thanks to Proposition 3.2, we only need to take care of the decay estimate of the L^2 norm of (v, g). In order to do so, we first act $\dot{\Delta}_j$ to (3.1), then use the micro-local energy estimate to get

$$\frac{d}{dt}(\|\dot{\Delta}_j v\|_{L^2}^2 + \|\dot{\Delta}_j g\|_{L^2}^2) + 2\mu\|\nabla\dot{\Delta}_j v\|_{L^2}^2 + 2\sigma\left\|\nabla_q(\dot{\Delta}_j g) + \frac{1}{2}\nabla_q U(\dot{\Delta}_j g)\right\|_{L^2}^2 = 0,$$

from which we deduce that there exists some small positive constant c such that

$$\frac{d}{dt}(\|\dot{\Delta}_j v\|_{L^2}^2 + \|\dot{\Delta}_j g\|_{L^2}^2) + c\|\dot{\Delta}_j v\|_{L^2}^2 + c\|\dot{\Delta}_j g\|_{L^2}^2 \le 0 \qquad \text{for} \quad j \ge 0,$$

and

$$\frac{d}{dt}(\|\dot{\Delta}_j v\|_{L^2}^2 + \|\dot{\Delta}_j g\|_{L^2}^2) + c2^{2j}\|\dot{\Delta}_j v\|_{L^2}^2 + c2^{2j}\|\dot{\Delta}_j g\|_{L^2}^2 \le 0 \qquad \text{for} \quad j < 0.$$

Then applying Gronwall's inequality gives

$$\|\dot{\Delta}_j v(t)\|_{L^2}^2 + \|\dot{\Delta}_j g(t)\|_{L^2}^2 \le (\|\dot{\Delta}_j u_0\|_{L^2}^2 + \|\dot{\Delta}_j f_0\|_{L^2}^2)e^{-ct} \quad \text{for} \quad j \ge 0,$$

$$\|\dot{\Delta}_j v(t)\|_{L^2}^2 + \|\dot{\Delta}_j g(t)\|_{L^2}^2 \le (\|\dot{\Delta}_j u_0\|_{L^2}^2 + \|\dot{\Delta}_j f_0\|_{L^2}^2)e^{-c2^{2j}t} \quad \text{for} \quad j < 0. \tag{3.14}$$

Next, let us turn to the decay estimate of $\|\Delta_j g(t)\|_{L^2}$. First by using a micro-local energy estimate to the microscopic equation of (3.1) we get that

$$\frac{d}{dt}\|\dot{\Delta}_j g\|_{L^2}^2 + 2\sigma\left\|\nabla_q(\dot{\Delta}_j g) + \frac{1}{2}\nabla_q U(\dot{\Delta}_j g)\right\|_{L^2}^2 \le \|\nabla(\dot{\Delta}_j v)\|_{L^2}\|q\sqrt{M}\nabla_q U\|_{L^2_q}\|\dot{\Delta}_j g\|_{L^2},$$

from which, with (3.14) and (2.1), we deduce that

$$\|\dot{\Delta}_j g(t)\|_{L^2}^2 \le C(\|\dot{\Delta}_j u_0\|_{L^2}^2 + \|\dot{\Delta}_j f_0\|_{L^2}^2)2^{2j}e^{-c2^{2j}t}. \tag{3.15}$$

On the other hand, thanks to [2, 16], we have $\|u\|_{L^2}^2 \equiv \sum_{j\in\mathbf{Z}}\|\Delta_j u\|_{L^2}^2$. Therefore, thanks to (3.7) and (3.14), we obtain

$$\begin{aligned}\|u(t)\|_{L^2}^2 &\le C\Big(\|u(t) - v(t)\|_{L^2}^2 + \sum_{j\in\mathbf{Z}}\|\Delta_j v(t)\|_{L^2}^2\Big)\\ &\le C\Big[(1+t)^{-\frac{1}{2}} + \sum_{j\le 0}(\|\dot{\Delta}_j u_0\|_{L^2}^2 + \|\dot{\Delta}_j f_0\|_{L^2}^2)e^{-c2^{2j}t}\Big].\end{aligned}$$

This proves (1.13) and the first part of (1.15). A similar argument gives (1.14) and the second part of (1.15). This completes the proof of Theorem 1.3. ∎

Acknowledgments. Both authors would like to thank the support of Morningside Center of Mathematics of the Chinese Academy of Sciences. The authors also thank the anonymous referees for valuable suggestions.

REFERENCES

[1] R. B. Bird, C. F. Curtis, R. C. Armstrong, and O. Hassager, *Dynamics of Polymeric Liquids, Volume 2: Kinetic Theory*, 2nd ed., Wiley Interscience, New York, 1987.

[2] J.-Y. Chemin, *Localization in Fourier space and Navier-Stokes system*, in Phase Space Analysis of Partial Differential Equations, Vol. 1, Pubbl. Cent. Ric. Mat. Ennio Giorgi, Scuola Norm. Sup., Pisa, 2004, pp. 53–135.

[3] Y. Chen and P. Zhang, *The global existence of small solutions to the incompressible viscoelastic fluid system in general space dimensions*, Comm. Partial Differential Equations, 31 (2006), pp. 1793–1810.

[4] P. Constantin, C. Fefferman, E. Titi, and A. Zarnescu, *Regularity for coupled two-dimensional nonlinear Fokker-Planck and Navier-Stokes systems*, Comm. Math. Phys., 270 (2007), pp. 789–811.

[5] P. Constantin and N. Masmoudi, *Global well-posedness for a Smoluchowski equation coupled with Navier-Stokes equations in* $2D$, Comm. Math. Phys., 278 (2008), pp. 179–191.

[6] M. Doi and S. F. Edwards, *The Theory of Polymer Dynamics*, Oxford Science Publications, Oxford, UK, 1986.

[7] B. Jourdain, T. Lelièvre, and C. Le Bris, *Existence of solution for a micro-macro model of polymeric fluid: The FENE model*, J. Funct. Anal., 209 (2004), pp. 162–193.

[8] B. Jourdain, T. Leliévre, C. Le Bris, and F. Otto, *Long-time asymptotics of a multiscale model for polymeric fluid flows*, Arch. Ration. Mech. Anal., 181 (2006), pp. 97–148.

[9] F.-H. Lin, C. Liu, and P. Zhang, *On hydrodynamics of viscoelastic fluids*, Comm. Pure Appl. Math., 58 (2005), pp. 1437–1471.

[10] F.-H. Lin, C. Liu, and P. Zhang, *On a micro-macro model for polymeric fluids near equilibrium*, Comm. Pure Appl. Math., 60 (2007), pp. 838–866.

[11] F.-H. Lin, P. Zhang, and Z. Zhang, *On the global existence of smooth solution to the* 2*-D FENE dumbbell model*, Comm. Math. Phys., 277 (2008), pp. 531–553.

[12] P.-L. Lions and N. Masmoudi, *Global existence of weak solutions to micro-macro models*, C. R. Math. Acad. Sci. Paris, 345 (2007), pp. 15–20.

[13] M. Renardy, *An existence theorem for model equations resulting from kinetic theories of polymer solutions*, SIAM J. Math. Anal., 22 (1991), pp. 313–327.

[14] M. E. Schonbek, L^2 *decay for weak solutions of the Navior-Stokes equations*, Arch. Rational Mech. Anal., 88 (1985), pp. 209–222.

[15] M. E. Schonbek, *Large time behavior of solutions to the Navior-Stokes equations*, Comm. Partial Differential Equations, 11 (1986), pp. 733–763.

[16] H. Triebel, *Theory of Function Spaces*, Monogr. Math. 78, Birkhäuser Verlag, Basel, 1983.

[17] E. Weinan, T. J. Li, and P. W. Zhang, *Well-posedness for the dumbbell model of polymeric fluids*, Comm. Math. Phys., 248 (2004), pp. 409–427.

SIAM J. MATH. ANAL.
Vol. 40, No. 5, pp. 1923–1951

ANISOTROPIC INHOMOGENEOUS RECTANGULAR THIN-WALLED BEAMS*

LORENZO FREDDI†, FRANÇOIS MURAT‡, AND ROBERTO PARONI§

Abstract. This paper is devoted to the asymptotic analysis of the problem of linear elasticity for an anisotropic and inhomogeneous body occupying, in its reference configuration, a cylindrical domain with a rectangular cross section with sides proportional to ε and ε^2 and clamped on one of its bases. The sequence of solutions u^ε of the equilibrium problem is shown to converge in an appropriate topology, as ε goes to zero, to the solution of a problem for a beam in which the extensional, flexural, and torsional effects are all coupled together.

Key words. asymptotic analysis, calculus of variations, thin-walled beams, dimension reduction, variational convergence, linear elasticity

AMS subject classifications. 49J45, 74K10, 74B05

DOI. 10.1137/080720279

1. Introduction.

Geometrically, a thin-walled beam is a slender structural element whose length is much larger than the diameter of the cross section which, on its hand, is larger than the thickness of the thin wall. These kinds of beams have been used for a long time in civil and mechanical engineering and, most of all, in flight vehicle structures because of their high ratio between maximum strength and weight. More recently, their importance has increased because of the introduction of fiber-reinforced composite materials in structural components. These materials are finding more and more applications for their high resistance to corrosion and high strength. Composite beams are usually made up by fiber-reinforced laminates and, hence, are anisotropic and inhomogeneous, even in cross-section planes. These peculiarities make classical thin-walled beam theories not applicable. The problem though has attracted the interest of several researchers and by now a huge number of articles can be found on the subject; see, for instance, [11] and the references therein. This is strongly remarked also in the first sentences of the abstract of [12]: "There is no lack of composite beam theories. Quite to the contrary, there might be too many of them. Different approaches, notations, etc., are used by the authors of those theories, so it is not always straightforward to compare the assumptions made and to assess the quantitative consequences of those assumptions."

The problem under study has a huge technological interest. One very suggestive, mentioned in [2], concerns the rotor blades of helicopters. The blades are composite beams and, hence, anisotropic and inhomogeneous. The anisotropy and the inhomogeneity introduce, as we shall also deduce, structural couplings between bending, extension, and twisting behaviors. It has been observed experimentally that these

*Received by the editors April 4, 2008; accepted for publication (in revised form) August 15, 2008; published electronically January 14, 2009. The first and third authors were partially supported by the Italian Ministry of University and Research within the Project PRIN 2005.
http://www.siam.org/journals/sima/40-5/72027.html

†Dipartimento di Matematica e Informatica, Università di Udine, via delle Scienze 206, 33100 Udine, Italy (freddi@dimi.uniud.it). This author was supported by INDAM with a GNAMPA research project 2008.

‡Laboratoire Jacques-Louis Lions, Université Paris VI, Boîte courrier 187, 75252 Paris Cedex 05, France (murat@ann.jussieu.fr).

§Dipartimento di Architettura e Pianificazione, Università degli Studi di Sassari, Palazzo del Pou Salit, Piazza Duomo, 07041 Alghero, Italy (paroni@uniss.it).

couplings have a powerful influence on blade dynamics including vibrations and the aeroelastic stability; see [2]. If a model of composite beams that accurately describes the structural couplings was at our disposal, then we could try to vary the anisotropy and inhomogeneity so as to minimize undesired effects like, for instance, vibrations. Through the control of lamination parameters (ply orientation and stacking sequence), it would then be possible for industry to minimize the undesired effects.

Our aim here is to deduce a "rigorous" model for a composite thin-walled beam, that is, a inhomogeneous and anisotropic beam. We shall achieve our goal by means of well-established asymptotic methods starting from the three-dimensional linear theory of elasticity.

This paper is devoted to the asymptotic analysis of the linearized system of equilibrium equations of a body which occupies, in its reference configuration, a cylindrical domain with a rectangular cross section with sides proportional to ε and ε^2 and clamped on one of its two bases. In particular, we study the compactness properties of the sequence of solutions u^ε of the equilibrium problems and, letting ε go to zero, we are concerned with the identification of the limit problem. The same problem has been studied from the point of view of Γ-convergence: in [4] in the simpler setting of homogeneous and isotropic material and in [3] in the case of an anisotropic material which is inhomogeneous only along the longitudinal axis and subject to residual stress. Trabucho and Viano [9] also studied the same problem by superimposing two asymptotic analyses where the lengths of the two sides of the cross section go to zero independently.

Besides the material properties of the body, our treatment differs from the preceding works also in the topology used in the passage from the three-dimensional problem to the one-dimensional: the one used in the present paper delivers much more information on the deformation of the beam. The approach is close to the one developed in a recent paper of Murat and Sili [8] for a thin cylinder of radius ε. The lack of isotropy or homogeneity assumptions leads to a limit problem where the extensional, flexural, and torsional effects are coupled together. In fact, we prove that the limit problem can be written as a system of five equations in a 5-tuple of unknowns (u, v, w, p, q) (see Theorem 6.1) and that $u^\varepsilon - (u + \varepsilon v + \varepsilon^2 w + \varepsilon^3 p + \varepsilon^4 q)$ converges strongly to zero in $H^1(\Omega)$, under some regularity assumptions on v, w, p, and q (see Corollary 6.1). We also derive the set of Euler equations of the variational limit problem, that is, the system of equilibrium differential equations, in the fully general case. Then we show that a strong simplification and a partial decoupling occurs when the material is homogeneous, and a complete decoupling is obtained for a homogeneous orthotropic material.

Notation. Throughout this paper Ω_1, Ω_2, and Ω_3 will denote the following three intervals:

$$\Omega_\alpha := (-a_\alpha/2, +a_\alpha/2) \quad \text{for } \alpha = 1, 2 \quad \text{and } \Omega_3 := (0, \ell),$$

where a_1, a_2, and ℓ are three positive real numbers. Also, for $i, j = 1, 2, 3$ we set

$$\Omega_{ij} := \Omega_i \times \Omega_j$$

and

$$\Omega := \Omega_1 \times \Omega_2 \times \Omega_3.$$

Unless otherwise specified, we use the Einstein summation convention. Moreover, we use the following convention for indexing vector and tensor components: Greek indices α and β take their values in the set $\{1,2\}$ and Latin indices i, j, and k in the set $\{1,2,3\}$. With a little abuse of notation, and because this is a common practice and does not give rise to any mistakes, we call "sequences" even those families indicized by a continuous parameter $\varepsilon \in (0,1)$. The component k of a vector v will be denoted either with $(v)_k$ or v_k, and an analogous notation will be used to denote tensor components. $\mathcal{E}_{\alpha\beta}$ denotes the Ricci's symbol, that is, $\mathcal{E}_{11} = \mathcal{E}_{22} = 0$, $\mathcal{E}_{12} = 1$, and $\mathcal{E}_{21} = -1$. Since usually $x = (x_1, x_2, x_3)$, we shall then denote by $x' := (x_1, x_2)$. A wide use will be made of vector valued distributions and Sobolev spaces; for a brief account of which and for the current notation we refer the reader to the book of Le Dret [5]. Throughout this paper C will denote a constant which may change line by line.

2. The three-dimensional problem. We consider a body which occupies, in its reference configuration, the region

$$\Omega_\varepsilon := \varepsilon^2\Omega_1 \times \varepsilon\Omega_2 \times \Omega_3 \subset \mathbb{R}^3.$$

We denote by $E(u)$ the strain of the displacement u, whose components are

$$E_{ij}(u) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right).$$

The elasticity tensor, with respect to the reference configuration Ω_ε, of the material will be denoted by $\mathbb{C}^\varepsilon$. We assume it to be essentially bounded,

$$\mathbb{C}^\varepsilon \in L^\infty(\Omega_\varepsilon; \mathbb{R}^{3\times3\times3\times3});$$

to have the minor symmetries,

$$\mathbb{C}^\varepsilon_{ijkl} = \mathbb{C}^\varepsilon_{jikl} = \mathbb{C}^\varepsilon_{ijlk};$$

and to be positive definite. That is, there exists a constant $c > 0$ such that

$$\mathbb{C}^\varepsilon A \cdot A \geq c|A|^2, \tag{1}$$

for all three by three symmetric matrices A and for all ε. We consider the body clamped on $\Gamma^\varepsilon_b := \partial\Omega_\varepsilon \cap \{x_3 = 0\}$ and we denote by

$$H^1_{dn}(\Omega_\varepsilon; \mathbb{R}^3) := \left\{\varphi \in H^1(\Omega_\varepsilon; \mathbb{R}^3) : \varphi = 0 \text{ on } \Gamma^\varepsilon_b\right\}.$$

The weak form of the equilibrium problem can be written as

$$\begin{cases} \tilde{u}^\varepsilon \in H^1_{dn}(\Omega_\varepsilon; \mathbb{R}^3), \\ \displaystyle\int_{\Omega_\varepsilon} \mathbb{C}^\varepsilon E(\tilde{u}^\varepsilon) \cdot E(\varphi)\, dv = \int_{\Omega_\varepsilon} \tilde{F}^\varepsilon \cdot E(\varphi)\, dv \quad \forall \varphi \in H^1_{dn}(\Omega_\varepsilon; \mathbb{R}^3), \end{cases} \tag{2}$$

where the matrix field $\tilde{F}^\varepsilon$, which takes into account the presence of external forces, is assumed to be an element of $L^2(\Omega_\varepsilon; \mathbb{R}^{3\times3}_{\text{sym}})$.

If $\tilde{F}^\varepsilon$ is not just in $L^2(\Omega_\varepsilon; \mathbb{R}^{3\times3}_{\text{sym}})$ but in

$$H(\text{div}, \Omega_\varepsilon) := \{T \in L^2(\Omega_\varepsilon; \mathbb{R}^{3\times3}_{\text{sym}}) : \text{div}\, T \in L^2(\Omega_\varepsilon; \mathbb{R}^3)\},$$

then the previous problem can be seen as the weak form of the following problem:

$$\begin{cases} \operatorname{div} \mathbb{C}^\varepsilon E(\tilde{u}^\varepsilon) + \tilde{b}^\varepsilon = 0 & \text{in } \Omega_\varepsilon, \\ \mathbb{C}^\varepsilon E(\tilde{u}^\varepsilon) n^\varepsilon = \tilde{c}^\varepsilon & \text{on } \Gamma_c^\varepsilon, \\ \tilde{u}^\varepsilon = 0 & \text{on } \Gamma_b^\varepsilon, \end{cases} \tag{3}$$

where $\Gamma_c^\varepsilon := \partial\Omega_\varepsilon \backslash \Gamma_b^\varepsilon$, and n^ε denotes the unit outward normal vector to Ω_ε, while the *body loads* $\tilde{b}^\varepsilon$ and the *contact loads* $\tilde{c}^\varepsilon$ are simply given by

$$\tilde{b}^\varepsilon := -\operatorname{div} \tilde{F}^\varepsilon \quad \text{in } \Omega_\varepsilon, \qquad \tilde{c}^\varepsilon := \tilde{F}^\varepsilon n^\varepsilon \quad \text{in } \Gamma_c^\varepsilon. \tag{4}$$

Note that given $\tilde{b}^\varepsilon \in L^2(\Omega_\varepsilon; \mathbb{R}^3)$ and $\tilde{c}^\varepsilon \in H^{-1/2}(\partial\Omega_\varepsilon; \mathbb{R}^3)$ it is always possible to find an $\tilde{F}^\varepsilon \in H(\operatorname{div}, \Omega_\varepsilon)$ which satisfies (4).

3. The rescaled problem. It is convenient to work with the domain Ω instead of the domain Ω_ε. We therefore rescale the problem by means of the scaling map $s_\varepsilon : \Omega \to \Omega_\varepsilon$,

$$s_\varepsilon(x_1, x_2, x_3) = (\varepsilon^2 x_1, \varepsilon x_2, x_3).$$

Let $\tilde{u}^\varepsilon$ be the solution of (2); then we define the "rescaled solution" u^ε by

$$u_1^\varepsilon := \varepsilon^2 \tilde{u}_1^\varepsilon \circ s_\varepsilon, \quad u_2^\varepsilon := \varepsilon \tilde{u}_2^\varepsilon \circ s_\varepsilon, \quad u_3^\varepsilon := \tilde{u}_3^\varepsilon \circ s_\varepsilon.$$

Let E^ε be the "rescaled strain" defined by

$$E^\varepsilon(\varphi) := \begin{pmatrix} \frac{1}{\varepsilon^4} E_{11}(\varphi) & \frac{1}{\varepsilon^3} E_{12}(\varphi) & \frac{1}{\varepsilon^2} E_{13}(\varphi) \\ \frac{1}{\varepsilon^3} E_{21}(\varphi) & \frac{1}{\varepsilon^2} E_{22}(\varphi) & \frac{1}{\varepsilon} E_{23}(\varphi) \\ \frac{1}{\varepsilon^2} E_{31}(\varphi) & \frac{1}{\varepsilon} E_{32}(\varphi) & E_{33}(\varphi) \end{pmatrix}. \tag{5}$$

It follows that $E^\varepsilon(u^\varepsilon) = E(\tilde{u}^\varepsilon) \circ s_\varepsilon$.

We further assume that there exists a $\mathbb{C} \in L^\infty(\Omega; \mathbb{R}^{3\times3\times3\times3})$ such that

$$\mathbb{C}^\varepsilon = \mathbb{C} \circ s_\varepsilon,$$

and we denote with $F^\varepsilon = \tilde{F}^\varepsilon \circ s_\varepsilon^{-1} \in L^2(\Omega; \mathbb{R}^{3\times3}_{\mathrm{sym}})$. With this notation u^ε turns out to be the unique solution of

$$\begin{cases} u^\varepsilon \in H^1_{dn}(\Omega; \mathbb{R}^3), \\ \displaystyle\int_\Omega \mathbb{C} E^\varepsilon(u^\varepsilon) \cdot E^\varepsilon(\varphi)\, dx = \int_\Omega F^\varepsilon \cdot E^\varepsilon(\varphi)\, dx \quad \forall \varphi \in H^1_{dn}(\Omega; \mathbb{R}^3), \end{cases} \tag{6}$$

where $\Gamma_b := \partial\Omega \cap \{x_3 = 0\}$ and

$$H^1_{dn}(\Omega; \mathbb{R}^3) := \Big\{ \varphi \in H^1(\Omega; \mathbb{R}^3) : \varphi = 0 \text{ on } \Gamma_b \Big\}.$$

By taking $\varphi = u^\varepsilon$ and using (1), we find

$$c \|E^\varepsilon(u^\varepsilon)\|_{L^2(\Omega)} \le \|F^\varepsilon\|_{L^2(\Omega)}. \tag{7}$$

Thus a uniform bound on $\|F^\varepsilon\|_{L^2(\Omega)}$ would lead to rescaled strains uniformly bounded in ε. We augment this requirement by assuming

$$F^\varepsilon \to F \quad \text{in } L^2(\Omega; \mathbb{R}^{3\times3}_{\mathrm{sym}}). \tag{8}$$

Remark 3.1. Let $S^\varepsilon := \operatorname{diag}(1/\varepsilon^2, 1/\varepsilon, 1)$; then $E^\varepsilon(\varphi) = S^\varepsilon E(\varphi) S^\varepsilon$. If we assume $F^\varepsilon \in H(\operatorname{div}, \Omega)$, then we can write

$$\begin{aligned}\int_\Omega F^\varepsilon \cdot E^\varepsilon(\varphi)\,dx &= \int_\Omega S^\varepsilon F^\varepsilon S^\varepsilon \cdot E(\varphi)\,dx \\ &= -\int_\Omega \operatorname{div}(S^\varepsilon F^\varepsilon S^\varepsilon)\cdot\varphi\,dx + \langle S^\varepsilon F^\varepsilon S^\varepsilon n, \varphi\rangle_{\partial\Omega},\end{aligned}$$

for all $\varphi \in H^1_{dn}(\Omega;\mathbb{R}^3)$, and conclude that instead of considering F^ε we could have been using the following body and contact forces

$$b^\varepsilon := -\operatorname{div}(S^\varepsilon F^\varepsilon S^\varepsilon) \quad \text{in } \Omega \quad \text{and} \quad c^\varepsilon := S^\varepsilon F^\varepsilon S^\varepsilon n \quad \text{in } \Gamma_c.$$

Note that if $F^\varepsilon = S^{\varepsilon^{-1}} F^{(0)} S^{\varepsilon^{-1}}$ for some $F^{(0)} \in H(\operatorname{div}, \Omega)$ the body and the contact forces would be independent of ε. Since $S^{\varepsilon^{-1}} = \operatorname{diag}(\varepsilon^2, \varepsilon, 1)$, the sequence $\{S^{\varepsilon^{-1}} F^{(0)} S^{\varepsilon^{-1}}\}$, with $F^{(0)} \in H(\operatorname{div}, \Omega)$, strongly converges in $L^2(\Omega; \mathbb{R}^{3\times 3}_{\mathrm{sym}})$ and hence satisfies assumption (8). Assumption (8) though allows us to consider "stronger" forces than $F^\varepsilon = S^{\varepsilon^{-1}} F^{(0)} S^{\varepsilon^{-1}}$, like $F^\varepsilon = F^{(0)}$ or, more generally,

$$F^\varepsilon = F^{(0)} + \varepsilon F^{(1)} + \varepsilon^2 F^{(2)} + \varepsilon^3 F^{(3)} + \varepsilon^4 F^{(4)},$$

with $F^{(i)} \in H(\operatorname{div}, \Omega)$ for $i = 0, 1, \dots, 4$, where, for instance, the term $\varepsilon^4 F^{(4)}_{11}$ would lead to the definition of body and contact forces independent of ε.

4. Partial Korn's inequalities. In this section we state and prove several Korn's inequalities. The proofs of Theorems 4.2 and 4.3 follow some of the lines of that of Theorem 4.4, which is due to Monneau, Murat, and Sili [7].

THEOREM 4.1. *There exists a constant C such that*

$$\Big\| u_1 - \fint_{\Omega_1} u_1\,dx_1 \Big\|_{L^2(\Omega)} \le C \Big\| \frac{\partial u_1}{\partial x_1} \Big\|_{L^2(\Omega)}$$

for every $u_1 \in H^1(\Omega_1; L^2(\Omega_{23}))$.

Proof. By density we may restrict ourselves to considering $u_1 \in C^1(\bar\Omega)$. For every x_2 and x_3 there exists a $\xi = \xi(x_2, x_3)$ such that $u_1(\xi, x_2, x_3) = \fint_{\Omega_1} u_1(s, x_2, x_3)\,ds$ and

$$u_1(x_1, \cdot, \cdot) - u_1(\xi, \cdot, \cdot) = \int_\xi^{x_1} \frac{\partial u_1}{\partial x_1}(s, \cdot, \cdot)\,ds.$$

Taking squares and applying Jensen's inequality we conclude the proof. ∎

THEOREM 4.2. *There exists a constant C such that*

$$\begin{aligned}\Big\| u_2 - \Big(\fint_{\Omega_1} u_2\,dx_1 - x_1 \frac{\partial}{\partial x_2} \fint_{\Omega_1} u_1\,dx_1 \Big) \Big\|_{H^{-1}(\Omega_2; L^2(\Omega_{13}))} \\ \le C \left(\big\| E_{11}(u) \big\|_{L^2(\Omega)} + \big\| E_{12}(u) \big\|_{L^2(\Omega)} \right)\end{aligned}$$

for every $u \in H^1(\Omega; \mathbb{R}^2)$.

Proof. Let

$$\bar u_1 := u_1 - \fint_{\Omega_1} u_1\,dx_1, \quad \bar u_2 := u_2 - \fint_{\Omega_1} u_2\,dx_1 + x_1 \frac{\partial}{\partial x_2} \fint_{\Omega_1} u_1\,dx_1,$$

and note that $⨍_{\Omega_1} \bar{u}_2\,dx_1 = 0$ and $\bar{u}_2 \in H^1(\Omega_1; L^2(\Omega_{23}))$. Let $\psi \in H^1_0(\Omega_2)$; then

$$\begin{aligned}\frac{\partial}{\partial x_1}\int_{\Omega_2}\psi\bar{u}_2\,dx_2 &= \int_{\Omega_2}\psi\left(\frac{\partial u_2}{\partial x_1}+\frac{\partial}{\partial x_2}⨍_{\Omega_1}u_1\,dx_1\right)dx_2\\ &= \int_{\Omega_2}\psi\left(2E_{12}(u)-\frac{\partial\bar{u}_1}{\partial x_2}\right)dx_2\\ &= \int_{\Omega_2}2\psi E_{12}(u)+\bar{u}_1\frac{d\psi}{dx_2}\,dx_2.\end{aligned}$$

Since $\int_{\Omega_2}\psi\bar{u}_2\,dx_2 \in H^1(\Omega_1; L^2(\Omega_{23}))$ and $\int_{\Omega_1}\int_{\Omega_2}\psi\bar{u}_2\,dx_2dx_1 = 0$, by Theorem 4.1 and the above equation we deduce

$$\begin{aligned}\left\|\int_{\Omega_2}\psi\bar{u}_2\,dx_2\right\|_{L^2(\Omega)} &\le C\left\|\frac{\partial}{\partial x_1}\int_{\Omega_2}\psi\bar{u}_2\,dx_2\right\|_{L^2(\Omega)}\\ &\le C\,\|\psi\|_{H^1(\Omega_2)}(\|E_{12}(u)\|_{L^2(\Omega)}+\|\bar{u}_1\|_{L^2(\Omega)})\\ &\le C\,\|\psi\|_{H^1(\Omega_2)}(\|E_{12}(u)\|_{L^2(\Omega)}+\|E_{11}(u)\|_{L^2(\Omega)}).\end{aligned}$$

Let $\varphi \in L^2(\Omega_{13})$; then

$$\begin{aligned}\left|\int_\Omega\varphi\psi\bar{u}_2\,dx\right| &\le \|\varphi\|_{L^2(\Omega_{13})}\left\|\int_{\Omega_2}\psi\bar{u}_2\,dx_2\right\|_{L^2(\Omega_{13})}\\ &\le C\,\|\varphi\|_{L^2(\Omega_{13})}\|\psi\|_{H^1(\Omega_2)}(\|E_{11}(u)\|_{L^2(\Omega_{13})}+\|E_{12}(u)\|_{L^2(\Omega)}).\end{aligned}$$

A density argument concludes the proof. Indeed, let $\{\varphi_n\}$ be an orthonormal basis of $L^2(\Omega_{13})$ and for any $v \in H^1_0(\Omega_2; L^2(\Omega_{13}))$, let

$$\psi_n(x_2) := \int_{\Omega_{13}}\varphi_n(x_1,x_3)v(x)\,dx_1dx_3 \in H^1_0(\Omega_2).$$

Then for $v_N := \sum_{n=1}^N \psi_n\varphi_n$ we have

$$\left|\int_\Omega v_N\bar{u}_2\,dx\right| \le \|v_N\|_{H^1_0(\Omega_2;L^2(\Omega_{13}))}(\|E_{11}(u)\|_{L^2(\Omega)}+\|E_{12}(u)\|_{L^2(\Omega)}),$$

and letting N go to infinity we conclude the proof. □

Remark 4.1. In spite of the fact that the left-hand side belongs to $L^2(\Omega)$, the inequality of Theorem 4.2 does not hold true if one replaces the norm H^{-1} with the norm of L^2, because of the following counterexample, which is inspired by an example contained in [7] in a quite similar framework.

Consider two scalar smooth functions $\varphi_\alpha \in C^\infty(\bar{\Omega}_\alpha)$ with φ_1 satisfying

$$\int_{\Omega_1}\varphi_1\,dx_1 = \int_{\Omega_1}\frac{\partial\varphi_1}{\partial x_1}\,dx_1 = 0.$$

Define

$$u_1 := -\varphi_2\frac{\partial\varphi_1}{\partial x_1},\quad u_2 := \varphi_1\frac{\partial\varphi_2}{\partial x_2},\quad u_3 := 0.$$

Then $u \in H^1(\Omega;\mathbb{R}^3)$, and the inequality of Theorem 4.2 reduces to

$$\left\|\varphi_1\frac{\partial\varphi_2}{\partial x_2}\right\|_{H^{-1}(\Omega_2;L^2(\Omega_1))} \le C\left\|\varphi_2\frac{\partial^2\varphi_1}{\partial x_1^2}\right\|_{L^2(\Omega)},$$

which cannot be true if we replace H^{-1} by L^2 because in such a case, taking $\|\varphi_1\|_{L^2(\Omega_1)} = \|\partial^2\varphi_1/\partial x_1^2\|_{L^2(\Omega_1)}$ would imply that

$$\left\|\frac{\partial\varphi_2}{\partial x_2}\right\|_{L^2(\Omega)} \le C\|\varphi_2\|_{L^2(\Omega)}$$

for any $\varphi_2 \in C^\infty(\bar\Omega_2)$, which is clearly impossible.

Define (using the summation convention)

$$rd_2 = \{r \in L^2(\Omega_{12};\mathbb{R}^2) : \exists\, c \in \mathbb{R},\, d \in \mathbb{R}^2 \text{ such that } r_\alpha(y) = \mathcal{E}_{\beta\alpha}\, x_\beta\, d + c_\alpha\},$$

where $\mathcal{E}$ denotes the Ricci's symbol. The elements of rd_2 are the infinitesimal rigid displacements on Ω_{12}. It is easy to see that $rd_2 \subset H^1(\Omega_{12};\mathbb{R}^2)$; moreover, being finite-dimensional, it is closed in $L^2(\Omega_{12};\mathbb{R}^2)$. Thus, the orthogonal projection operator of $L^2(\Omega_{12};\mathbb{R}^2)$ on rd_2, which will be denoted by $\wp$, is well defined. Given a vector function $v \in L^2(\Omega_{12},\mathbb{R}^m)$ with $m \ge 2$, we define

$$\vartheta(v) := \frac{1}{I_O}\int_{\Omega_{12}} (x_1 v_2 - x_2 v_1)\, dx', \text{ where } I_O = \int_{\Omega_{12}} (x_1^2 + x_2^2)\, dx'. \tag{9}$$

If $v \in L^2(\Omega_{12};\mathbb{R}^2)$, then the components of $\wp$ turn out to be

$$\wp_\alpha(v) = \mathcal{E}_{\beta\alpha}\, x_\beta \vartheta(v) + \fint_{\Omega_{12}} v_\alpha\, dx'. \tag{10}$$

Furthermore, the two-dimensional Korn's inequality can be written as

$$\|v - \wp(v)\|_{H^1(\Omega_{12};\mathbb{R}^2)} \le C \sum_{\alpha,\beta} \|E_{\alpha\beta}(v)\|_{L^2(\Omega_{12};\mathbb{R}^{2\times 2})} \tag{11}$$

for all $v \in H^1(\Omega_{12};\mathbb{R}^2)$ with a constant C which is independent of v.

Similarly, given a vector function $u \in L^2(\Omega_3; L^2(\Omega_{12},\mathbb{R}^m))$ with $m \ge 2$, we analogously define $\vartheta(u) \in L^2(\Omega_3)$ and $\wp_\alpha(u)$. The operator $\wp$ associates to any $u \in L^2(\Omega_3; L^2(\Omega_{12},\mathbb{R}^m))$ a function $\wp(u) \in L^2(\Omega_3; L^2(\Omega_{12},\mathbb{R}^2))$ which is an infinitesimal rigid displacement on Ω_{12} for almost every $x_3 \in \Omega_3$.

Let us observe that the orthogonal complement, with respect to the $L^2(\Omega_{12};\mathbb{R}^2)$ inner product, of rd_2 in $H^1(\Omega_{12};\mathbb{R}^2)$ can be then characterized as

$$rd_2^\perp = \{v \in H^1(\Omega_{12};\mathbb{R}^2) \ :\ \wp(v) = 0\} = \{v \in H^1_m(\Omega_{12};\mathbb{R}^2) \ :\ \vartheta(v) = 0\}.$$

Moreover, we denote by

$$RD_2^\perp(\Omega) = \{v \in L^2(\Omega_3; H^1_m(\Omega_{12};\mathbb{R}^2)) \ :\ \vartheta(v) = 0 \text{ a.e. } x_3 \in \Omega_3\}. \tag{12}$$

Hereafter, for any $u \in L^2(\Omega_3; H^1(\Omega_{12};\mathbb{R}^m))$, $m \ge 2$, we set

$$\tilde u_\alpha := u_\alpha - \wp_\alpha(u). \tag{13}$$

Of course $\tilde u \in L^2(\Omega_3; H^1(\Omega_{12};\mathbb{R}^2))$ and $\fint_{\Omega_{12}} \tilde u\, dx_1 dx_2 = 0$ and $\vartheta(\tilde u) = 0$ a.e. in Ω_3, where the latter follows from the linearity of ϑ and the fact that $\vartheta(u) = \vartheta(\wp(u))$. Thus $\tilde u \in RD_2^\perp(\Omega)$.

LEMMA 4.1. *There exists a constant C such that*

$$\|\tilde u\|_{L^2(\Omega_3;H^1(\Omega_{12};\mathbb{R}^2))} \le C \sum_{\alpha,\beta} \|E_{\alpha\beta}(u)\|_{L^2(\Omega_3;L^2(\Omega_{12}))}$$

for every $u \in L^2(\Omega_3; H^1(\Omega_{12};\mathbb{R}^m))$, $m \ge 2$.

Proof. Since $u_\alpha(\cdot,\cdot,x_3) \in L^2(\Omega_{12})$ for almost every $x_3 \in \Omega_3$, from (10) and (11) we have that the relations

$$\wp_\alpha(u) = \mathcal{E}_{\beta\alpha}\, x_\beta\, \vartheta(u) + \frac{1}{|\Omega_{12}|}\int_{\Omega_{12}} u_\alpha\, dx', \tag{14}$$

$$\|\tilde{u}\|_{H^1(\Omega_{12};\mathbb{R}^2)} \le C \sum_{\alpha,\beta} \|E_{\alpha\beta}(u)\|_{L^2(\Omega_{12})} \tag{15}$$

hold for almost every x_3 in Ω_3, and the claimed inequality follows by integration. □

A different proof of the lemma above can be found in Le Dret [6].

PROPOSITION 4.1. *$RD_2^\perp(\Omega)$ is a Hilbert space with the norm*

$$\|v\|_{RD_2^\perp(\Omega)} := \left(\sum_{\alpha,\beta} \|E_{\alpha\beta}(v)\|^2_{L^2(\Omega)}\right)^{1/2}.$$

Proof. We have only to prove that $\|v\|_{RD_2^\perp(\Omega)}$ is equivalent to the norm induced on $RD_2^\perp(\Omega)$ by $L^2(\Omega_3;H^1(\Omega_{12};\mathbb{R}^2))$ since the former space is a closed subspace of the latter. For any $v \in RD_2^\perp(\Omega)$, recalling that $\wp(v) = 0$, so that $v = \tilde{v}$, and using Lemma 4.1, we then have

$$\|v\|_{L^2(\Omega_3;H^1(\Omega_{12};\mathbb{R}^2))} = \|\tilde{v}\|_{L^2(\Omega_3;H^1(\Omega_{12};\mathbb{R}^2))} \le C\|v\|_{RD_2^\perp(\Omega)}$$

while the opposite inequality is trivially satisfied. □

THEOREM 4.3. *There exists a constant C such that*

$$\left\| u_3 - \left(x_1 x_2 \frac{d\vartheta(u)}{dx_3} - x_1 \frac{d}{dx_3} \fint_{\Omega_{12}} u_1\, dx' + \fint_{\Omega_1} u_3\, dx_1 \right) \right\|_{H^{-1}(\Omega_3;L^2(\Omega_{12}))} \le C \left(\sum_{\alpha,\beta} \|E_{\alpha\beta}(u)\|_{L^2(\Omega)} + \|E_{13}(u)\|_{L^2(\Omega)} \right)$$

for every $u \in H^1(\Omega;\mathbb{R}^3)$.

Proof. Let $\tilde{u}_\alpha := u_\alpha - \wp_\alpha(u)$ as in (13) and

$$\tilde{u}_3 := u_3 - \left(x_1 x_2 \frac{d\vartheta(u)}{dx_3} - x_1 \frac{d}{dx_3} \fint_{\Omega_{12}} u_1\, dx' + \fint_{\Omega_1} u_3\, dx_1 \right).$$

Since $E_{13}(u) = E_{13}(\tilde{u})$,

$$\frac{\partial \tilde{u}_3}{\partial x_1} = 2E_{13}(u) - \frac{\partial \tilde{u}_1}{\partial x_3}.$$

Let $\psi \in H_0^1(\Omega_3)$; then

$$\begin{aligned} \frac{\partial}{\partial x_1}\int_{\Omega_3} \psi \tilde{u}_3\, dx_3 &= \int_{\Omega_3} \psi\Big(2E_{13}(u) - \frac{\partial \tilde{u}_1}{\partial x_3}\Big)\, dx_3 \\ &= \int_{\Omega_3} 2\Big(\psi E_{13}(u) + \tilde{u}_1 \frac{d\psi}{dx_3}\Big)\, dx_3. \end{aligned}$$

Since $\int_{\Omega_3} \psi\tilde{u}_3\,dx_3 \in H^1(\Omega)$ and $\int_{\Omega_1}\int_{\Omega_3} \psi\tilde{u}_3\,dx_3dx_1 = 0$, by Theorem 4.1, Lemma 4.1, and the previous equality we deduce

$$\begin{aligned}\left\|\int_{\Omega_3}\psi\tilde{u}_3\,dx_3\right\|_{L^2(\Omega)} &\le C\left\|\frac{\partial}{\partial x_1}\int_0^\ell \psi\tilde{u}_3\,dx_3\right\|_{L^2(\Omega)}\\ &\le C\|\psi\|_{H^1(\Omega_3)}(\|E_{13}(u)\|_{L^2(\Omega)} + \|\tilde{u}_1\|_{L^2(\Omega)})\\ &\le C\|\psi\|_{H^1(\Omega_3)}(\|E_{13}(u)\|_{L^2(\Omega)} + \textstyle\sum_{\alpha,\beta}\|E_{\alpha\beta}(u)\|_{L^2(\Omega)}).\end{aligned}$$

Let $\varphi \in L^2(\Omega_{12})$; then

$$\begin{aligned}\left|\int_\Omega \varphi\psi\tilde{u}_3\,dx\right| &= \left|\int_{\Omega_{12}}\varphi\int_{\Omega_3}\psi\tilde{u}_3\,dx_3dx'\right| \le \|\varphi\|_{L^2(\Omega_{12})}\left\|\int_{\Omega_3}\psi\tilde{u}_3\,dx_3\right\|_{L^2(\Omega_{12})}\\ &\le C\|\varphi\|_{L^2(\Omega_{12})}\|\psi\|_{H^1(\Omega_3)}\left(\|E_{13}(u)\|_{L^2(\Omega)} + \sum_{\alpha,\beta}\|E_{\alpha\beta}(u)\|_{L^2(\Omega)}\right).\end{aligned}$$

Arguing as in the proof of Theorem 4.2, we conclude the proof. □

Remark 4.2. As in Remark 4.1, in spite of the fact that the left-hand side belongs to $L^2(\Omega)$, the inequality of Theorem 4.3 does not hold true if one replaces the norm H^{-1} with the norm of L^2, because of the following counterexample.

Consider three scalar smooth functions $\varphi_i \in C^\infty(\bar{\Omega}_i)$ with φ_α satisfying

$$\int_{\Omega_\alpha}\varphi_\alpha\,dx_\alpha = \int_{\Omega_\alpha}\frac{\partial\varphi_\alpha}{\partial x_\alpha}\,dx_\alpha = 0.$$

Define

$$u_1 := -\varphi_2\frac{\partial\varphi_1}{\partial x_1}\varphi_3, \quad u_2 := -\varphi_1\frac{\partial\varphi_2}{\partial x_2}\varphi_3, \quad u_3 := +\varphi_1\varphi_2\frac{\partial\varphi_3}{\partial x_3}.$$

Then $u \in H^1(\Omega;\mathbb{R}^3)$, and the inequality of Theorem 4.3 reduces to

$$\left\|\varphi_1\varphi_2\frac{\partial\varphi_3}{\partial x_3}\right\|_{H^{-1}(\Omega_2;L^2(\Omega_1))} \le C\left\|\varphi_3\left(\frac{\partial^2\varphi_1}{\partial x_1^2}\varphi_2 + \varphi_1\frac{\partial^2\varphi_2}{\partial x_2^2}\right)\right\|_{L^2(\Omega)},$$

which cannot be true if we replace H^{-1} by L^2, because in such a case taking $\|\varphi_\alpha\|_{L^2(\Omega_\alpha)} = \|\partial^2\varphi_\alpha/\partial x_1^\alpha\|_{L^2(\Omega_\alpha)}$ would imply that

$$\left\|\frac{\partial\varphi_3}{\partial x_3}\right\|_{L^2(\Omega)} \le C\|\varphi_3\|_{L^2(\Omega)}$$

for any $\varphi_3 \in C^\infty(\bar{\Omega}_3)$, which is clearly impossible.

The next partial Korn's inequality is proved in Monneau, Murat, and Sili [7].

THEOREM 4.4. *There exists a constant C such that*

$$\begin{aligned}&\left\|u_3 - \left(\fint_{\Omega_{12}} u_3\,dx' - x_\alpha\frac{d}{dx_3}\fint_{\Omega_{12}} u_\alpha\,dx'\right)\right\|_{H^{-1}(\Omega_3;L^2(\Omega_{12}))}\\ &\qquad\le C\left(\sum_{\alpha\beta}\|E_{\alpha\beta}(u)\|_{L^2(\Omega)} + \sum_\alpha\|E_{\alpha3}(u)\|_{L^2(\Omega)}\right)\end{aligned}$$

for every $u \in H^1_{dn}(\Omega;\mathbb{R}^3)$.

5. Limit strain characterization. Let

$$\begin{aligned} H^1_m(\Omega_{12}) &:= \{z \in H^1(\Omega_{12}) : \textstyle\int_{\Omega_{12}} z\,dx' = 0\},\\ H^1_{m_1}(\Omega_1; L^2(\Omega_{23})) &:= \{z \in H^1(\Omega_1; L^2(\Omega_{23})) : \textstyle\int_{\Omega_1} z\,dx_1 = 0 \text{ a.e. in } \Omega_{23}\},\\ H^{-1}_{m_1}(\Omega_3; L^2(\Omega_{12})) &:= \{v \in H^{-1}(\Omega_3; L^2(\Omega_{12})) : \langle z_3, \varphi\rangle = 0 \\ &\qquad\qquad \forall\, \varphi \in H^1_0(\Omega_3; L^2(\Omega_2))\}, \end{aligned}$$

where the bracket in the last definition has to be understood in the sense of the duality $H^{-1}(\Omega_3; L^2(\Omega_{12})) \times H^1_0(\Omega_3; L^2(\Omega_{12}))$.

Let (u^ε) be the sequence of solutions to problems (6). From (7) and assumption (8) it follows that

$$\sup_\varepsilon \|E^\varepsilon(u^\varepsilon)\|_{L^2(\Omega)} < +\infty. \tag{16}$$

Hence, possibly passing to a subsequence, we have that

$$E^\varepsilon(u^\varepsilon) \rightharpoonup E \quad \text{in } L^2(\Omega; \mathbb{R}^{3\times 3}_{\rm sym})$$

for some $E \in L^2(\Omega; \mathbb{R}^{3\times 3}_{\rm sym})$.

In this section we characterize the limit strain E. For clarity we state several lemmas.

LEMMA 5.1 (component 33). *There exists a function $\bar u$ in the set of the so-called Bernoulli–Navier displacements*

$$\mathscr{U} := \{u \in H^1_{dn}(\Omega; \mathbb{R}^3) : E_{\alpha i}(u) = 0\}$$

such that

$$E_{33} = \frac{\partial \bar u_3}{\partial x_3} = E_{33}(\bar u).$$

Proof. From (16), the structure of E^ε, and Korn's inequality, we have that

$$C \geq \|E^\varepsilon(u^\varepsilon)\|_{L^2(\Omega)} \geq \|E(u^\varepsilon)\|_{L^2(\Omega)} \geq C_K \|u^\varepsilon\|_{H^1(\Omega)},$$

where C is a constant independent of ε and C_K is Korn's constant. Hence, up to a subsequence $u^\varepsilon \rightharpoonup \bar u$ in $H^1(\Omega; \mathbb{R}^3)$, for some $\bar u \in H^1_{dn}(\Omega; \mathbb{R}^3)$. The claim follows by noticing that $\|E_{\alpha i}(u^\varepsilon)\|_{L^2(\Omega)} \leq C\varepsilon$, and $E^\varepsilon_{33}(u^\varepsilon) = \partial u^\varepsilon_3/\partial x_3$. In fact it follows that $\bar u \in \{u \in H^1_{dn}(\Omega; \mathbb{R}^3) : E_{\alpha i}(u) = 0\}$. ∎

Remark 5.1 (representation of the space $\mathscr{U}$). It is well known (see, for instance, Le Dret [5]) that the space of Bernoulli–Navier displacements admits the following representation:

$$\mathscr{U} := \Big\{ u \in H^2_{dn}(\Omega)^2 \times H^1_{dn}(\Omega) : \text{exists } \zeta \in H^2_{dn}(\Omega_3)^2 \times H^1_{dn}(\Omega_3) \text{ such that } u_1 = \zeta_1,\ u_2 = \zeta_2,\ u_3 = \zeta_3 - x_1 \frac{d\zeta_1}{dx_3} - x_2 \frac{d\zeta_2}{dx_3} \Big\}.$$

Moreover, $\mathscr{U}$ is a Hilbert space with the norm

$$\|u\|_{\mathscr{U}} := \|E_{33}(u)\|_{L^2(\Omega)},$$

which is equivalent to that induced by $H^1_{dn}(\Omega; \mathbb{R}^3)$ (see [8]).

LEMMA 5.2 (component 23). *There exists a function $\bar{v}$ in the space*

$$\mathscr{V} := \Big\{ v \in H^1_{dn}(\Omega)^2 \times L^2(\Omega_3; H^1_m(\Omega_{12})) : \textit{exist } \vartheta \in H^1(\Omega_3) \textit{ such that } \vartheta(0) = 0$$
$$\textit{and } \varrho \in L^2(\Omega_3; H^1(\Omega_2)) \textit{ such that } v_1(x) = -x_2\vartheta(x_3),\ v_2(x) = x_1\vartheta(x_3),$$
$$v_3(x) = x_1x_2\frac{d\vartheta}{dx_3}(x_3) + \varrho(x_2, x_3)\Big\}$$

such that

$$E_{23} = E_{23}(\bar{v}).$$

Proof. Let $v_i^\varepsilon := \frac{1}{\varepsilon}u_i^\varepsilon$ and $\vartheta^\varepsilon := \vartheta(v^\varepsilon)$ (see (9)).

First of all, by adapting an argument of [4] and Lemmas 4.4 and 4.5, we prove that there exists $\vartheta \in H^1(\Omega_3)$ with $\vartheta(0) = 0$, such that, up to subsequences,

$$\vartheta^\varepsilon \to \vartheta \text{ in } L^2(\Omega_3). \tag{17}$$

Applying Lemma 4.1, there exists a constant C such that

$$\|\tilde{v}^\varepsilon\|_{L^2(\Omega_3;H^1(\Omega_{12};\mathbb{R}^2))} \le C\sum_{\alpha,\beta}\|E_{\alpha\beta}(v^\varepsilon)\|_{L^2(\Omega_3;L^2(\Omega_{12}))}. \tag{18}$$

Since, furthermore,

$$E_{11}(v^\varepsilon)_{11} = \varepsilon^3E^\varepsilon_{11}(u^\varepsilon),\ E_{12}(v^\varepsilon) = \varepsilon^2E^\varepsilon_{12}(u^\varepsilon),\ E_{22}(v^\varepsilon) = \varepsilon E^\varepsilon_{22}(u^\varepsilon),$$

and using (16) we have

$$\|E_{\alpha\beta}(v^\varepsilon)\|_{L^2(\Omega)} \le \varepsilon\|E^\varepsilon_{\alpha\beta}(u^\varepsilon)\|_{L^2(\Omega)} \le C\,\varepsilon. \tag{19}$$

From (18) we get

$$\|\tilde{v}^\varepsilon\|_{L^2(\Omega_3;H^1(\Omega_{12};\mathbb{R}^2))} \le C\varepsilon; \tag{20}$$

hence

$$\tilde{v}^\varepsilon \to 0 \text{ in } L^2(\Omega_3; H^1(\Omega_{12};\mathbb{R}^2)). \tag{21}$$

Now let $\eta \in C^\infty_c(\Omega_{12})$ be such that

$$\int_{\Omega_{12}} \eta\, dx' = -\frac{I_O}{2}.$$

Then, taking into account (14), we have

$$
\begin{aligned}
I_O\vartheta^\varepsilon &= -2\vartheta^\varepsilon \int_{\Omega_{12}} \eta\, dx' = -\vartheta^\varepsilon \int_{\Omega_{12}} \eta D_\alpha x_\alpha\, dx' \\
&= \vartheta^\varepsilon \int_{\Omega_{12}} D_\alpha\eta\, x_\alpha\, dx' = \vartheta^\varepsilon \int_{\Omega_{12}} \mathcal{E}_{\alpha\gamma}\mathcal{E}_{\beta\gamma} D_\alpha\eta\, x_\beta\, dx' \\
&= \int_{\Omega_{12}} \mathcal{E}_{\alpha\gamma} D_\alpha\eta \mathcal{E}_{\beta\gamma} x_\beta \vartheta^\varepsilon\, dx' \\
&= \int_{\Omega_{12}} \mathcal{E}_{\alpha\gamma} D_\alpha\eta \Big(\wp_\gamma(v^\varepsilon) - \frac{1}{|\Omega_{12}|}\int_{\Omega_{12}} v^\varepsilon_\gamma\, dx'\Big)\, dx' \\
&= \int_{\Omega_{12}} \mathcal{E}_{\alpha\gamma} D_\alpha\eta\, \wp_\gamma(v^\varepsilon)\, dx' \\
&= \int_{\Omega_{12}} \mathcal{E}_{\alpha\gamma} D_\alpha\eta\, v^\varepsilon_\gamma\, dx' - \int_{\Omega_{12}} \mathcal{E}_{\alpha\gamma} D_\alpha\eta \big(v^\varepsilon - \wp_\gamma(v^\varepsilon)\big)\, dx'.
\end{aligned}
$$

Hence, denoting by

$$
\tilde\vartheta^\varepsilon = \frac{1}{I_O}\int_{\Omega_{12}} \mathcal{E}_{\alpha\gamma} D_\alpha\eta\, w^\varepsilon_\gamma\, dx'
$$

and recalling (21), we find

$$
\vartheta^\varepsilon - \tilde\vartheta^\varepsilon \to 0 \ \text{ in } L^2(\Omega_3). \tag{22}
$$

We now show that $D_3\tilde\vartheta^\varepsilon$ is bounded in L^2. Since $\mathcal{E}_{\alpha\gamma}D_\alpha D_\gamma\eta = 0$ everywhere in Ω_3 and $D_\alpha\eta = 0$ on $\partial\Omega_3$, we have

$$
I_O D_3\tilde\vartheta^\varepsilon = \int_{\Omega_{12}} \mathcal{E}_{\alpha\gamma} D_\alpha\eta\, D_3 v^\varepsilon_\gamma\, dx' = 2\int_{\Omega_{12}} \mathcal{E}_{\alpha\gamma} D_\alpha\eta\, E_{\gamma 3}(v^\varepsilon)\, dx',
$$

but $E_{13}(v^\varepsilon) = \varepsilon E^\varepsilon_{13}(u^\varepsilon)$ and $E_{23}(v^\varepsilon) = E^\varepsilon_{23}(u^\varepsilon)$, and therefore $D_3\tilde\vartheta^\varepsilon$ is bounded in $L^2(\Omega_3)$. Since $\tilde\vartheta^\varepsilon(0) = 0$, $\tilde\vartheta^\varepsilon$ is then bounded in $H^1(\Omega_3)$ so that there exists $\vartheta \in H^1(\Omega_3)$ with $\vartheta(0) = 0$, such that, up to subsequences,

$$
\tilde\vartheta^\varepsilon \rightharpoonup \vartheta \text{ in } H^1(\Omega_3).
$$

Thus, from (22) we obtain (17).

Let us now set

$$
\bar v^\varepsilon_1 := v^\varepsilon_1 - \fint_{\Omega_{12}} v^\varepsilon_1\, dx', \quad \bar v^\varepsilon_2 := v^\varepsilon_2 - \fint_{\Omega_{12}} v^\varepsilon_2\, dx',
$$

and

$$
\bar v^\varepsilon_3 := v^\varepsilon_3 - \Big(\fint_{\Omega_{12}} v^\varepsilon_3\, dx' - x_\alpha \frac{d}{dx_3}\fint_{\Omega_{12}} v^\varepsilon_\alpha\, dx'\Big).
$$

Observing that, by the definitions,

$$
\bar v^\varepsilon_1 = \tilde v^\varepsilon_1 - x_2\vartheta^\varepsilon, \quad \bar v^\varepsilon_2 = \tilde v^\varepsilon_2 + x_1\vartheta^\varepsilon,
$$

from (17) and (21) we have that

$$\bar{v}_1^\varepsilon \to -x_2\vartheta, \qquad \bar{v}_2^\varepsilon \to +x_1\vartheta \qquad \text{in } L^2(\Omega_3; H^1(\Omega_{12})).$$

By Theorem 4.4, we have that

$$\|\bar{v}_3^\varepsilon\|_{H^{-1}(\Omega_3;L^2(\Omega_{12}))} \le C,$$

and hence there exists $\bar{v}_3 \in H^{-1}(\Omega_3; L^2(\Omega_{12}))$ such that, up to subsequences,

$$\bar{v}_3^\varepsilon \rightharpoonup : \bar{v}_3 \text{ in } H^{-1}(\Omega_3; L^2(\Omega_{12})). \tag{23}$$

Moreover, let us set

$$\bar{v}_1 = -x_2\vartheta, \quad \bar{v}_2 = +x_1\vartheta,$$

and check that the vector field $\bar{v}$ so defined satisfies the properties claimed in the statement of Lemma 4.1. A simple computation shows that

$$E_{13}(\bar{v}^\varepsilon) = E_{13}(v^\varepsilon) \quad \text{and} \quad E_{23}(\bar{v}^\varepsilon) = E_{23}(v^\varepsilon).$$

Noticing that $E_{13}(v^\varepsilon) = \varepsilon E_{13}^\varepsilon(u^\varepsilon) \to 0$ in $L^2(\Omega)$ and $E_{13}(\bar{v}^\varepsilon) \to E_{13}(\bar{v})$ in $\mathscr{D}'(\Omega)$, we obtain that $E_{13}(\bar{v}) = 0$. Hence,

$$\frac{\partial \bar{v}_3}{\partial x_1} = -\frac{\partial \bar{v}_1}{\partial x_3} = x_2 \frac{d\vartheta}{dx_3}$$

and, integrating with respect to x_1,

$$\bar{v}_3 = x_1 x_2 \frac{d\vartheta}{dx_3} + \varrho(x_2, x_3)$$

for some function $\varrho \in L^2(\Omega_3; H^1(\Omega_2))$. Moreover, $\bar{v}_3 \in L^2(\Omega_3; H^1(\Omega_{12}))$ and, from (23) and the fact that $\bar{v}_3$, $\bar{v}_3^\varepsilon \in L^2(\Omega)$, we then obtain easily that $⨍_{\Omega_{12}} \bar{v}_3 dx' = 0$, which concludes the proof. □

LEMMA 5.3 (characterization of the space $\mathscr{V}$). *The space $\mathscr{V}$ admits the following characterization:*

$$\begin{aligned}\mathscr{V} = \{v \in H^1_{dn}(\Omega)^2 \times L^2(\Omega_3; H^1_m(\Omega_{12})) : E_{\alpha\beta}(v) = 0,\ E_{13}(v) = 0,\\ E_{23}(v) \in L^2(\Omega) \text{ and } \textstyle\int_{\Omega_{12}} v_\alpha\, dx' = 0 \text{ a.e.}\}.\end{aligned} \tag{24}$$

Moreover, it is a Hilbert space with the norm

$$\|v\|_{\mathscr{V}} := \|W_{13}(v)\|_{L^2(\Omega)} + \|E_{23}(v)\|_{L^2(\Omega)},$$

where

$$W_{13}(v) = \frac{1}{2}\left(\frac{\partial v_1}{\partial x_3} - \frac{\partial v_3}{\partial x_1}\right).$$

Proof. Let us call V the space at the right-hand side of equality (24) and let $\mathscr{V}$ be as in the statement of Lemma 5.2. It is trivial to check that $\mathscr{V} \subseteq V$. Let us prove the opposite inclusion. Let $v \in V$. Since $v_\alpha \in H^1_{dn}(\Omega)$ and $E_{\alpha\beta}(v) = 0$, by integration there exists $\vartheta \in L^2(\Omega_3)$ such that

$$v_1 = -x_2\vartheta(x_3) + a_1(x_3), \quad v_2 = x_1\vartheta(x_3) + a_2(x_3),$$

and since $\int_{\Omega_{12}} v_\alpha \, dx' = 0$, we have $a_1 = a_2 = 0$ a.e.. From the resulting expression of v_α it follows that $\vartheta \in H^1_{dn}(\Omega)$ and $\vartheta(0) = 0$.

Since $E_{13}(v) = 0$ we obtain that $\partial v_3/\partial x_1 = x_2 \vartheta'(x_3)$. Then there exists $\varrho \in L^2(\Omega_{23})$ such that

$$v_3 = x_1 x_2 \vartheta'(x_3) + \varrho(x_2, x_3),$$

from which it follows also that $\varrho \in L^2(\Omega_3; H^1_m(\Omega_2))$ and that $E_{23}(v) \in L^2(\Omega)$.

The last part of the claim follows from the fact that $L^2(\Omega_3; H^1_m(\Omega_{12})$ is a Hilbert space with the scalar product

$$\langle u_3, v_3 \rangle_{L^2(\Omega_3; H^1_m(\Omega_{12}))} = \int_{\Omega_3} \langle u_3(x_3), v_3(x_3) \rangle_{H^1_m(\Omega_{12})} \, dx_3,$$

and $\mathscr{V}$ is a closed subspace of $H^1_{dn}(\Omega)^2 \times L^2(\Omega_3; H^1_m(\Omega_{12}))$ which is Hilbert with the product norm

$$\|v_1\|_{H^1_{dn}(\Omega)} + \|v_2\|_{H^1_{dn}(\Omega)} + \|v_3\|_{L^2(\Omega_3; H^1_m(\Omega_{12}))}$$

induced by $H^1_{dn}(\Omega)^2 \times L^2(\Omega_3; H^1_m(\Omega_{12}))$. The proof that this norm is equivalent to $\|v\|_{\mathscr{V}}$ is an easy consequence of the Poincaré inequality, the representation lemma, Lemma 5.2, and the characterization of the space $\mathscr{V}$ proved above. ∎

LEMMA 5.4 (components 22 and 13). *There exists a function $\bar{w}$ in the space*

$$\mathscr{W} = \{ w \in RD^\perp_2(\Omega) \times H^{-1}_{m_1}(\Omega_3; L^2(\Omega_{12})) \ : \ E_{11}(w) = E_{12}(w) = 0, \ E_{13}(w) \in L^2(\Omega) \}$$

such that

$$E_{22} = \frac{\partial \bar{w}_2}{\partial x_2} = E_{22}(\bar{w}), \quad E_{13} = \frac{1}{2}\left(\frac{\partial \bar{w}_1}{\partial x_3} + \frac{\partial \bar{w}_3}{\partial x_1} \right) = E_{13}(\bar{w}).$$

Moreover, $\mathscr{W}$ is a Hilbert space with the norm

$$\|w\|_{\mathscr{W}} := \|E_{22}(w)\|_{L^2(\Omega)} + \|E_{13}(w)\|_{L^2(\Omega)} + \|w_3\|_{H^{-1}_{m_1}(\Omega_3; L^2(\Omega_{12}))}.$$

Proof. Let

$$w^\varepsilon := \frac{1}{\varepsilon^2} u^\varepsilon;$$

then

$$\begin{aligned} &\|E_{11}(w^\varepsilon)\|_{L^2(\Omega)} \le C\varepsilon^2, \qquad \|E_{12}(w^\varepsilon)\|_{L^2(\Omega)} \le C\varepsilon, \\ &\|E_{22}(w^\varepsilon)\|_{L^2(\Omega)} \le C, \qquad \|E_{13}(w^\varepsilon)\|_{L^2(\Omega)} \le C. \end{aligned} \tag{25}$$

Let us recall that $\tilde{w}^\varepsilon_\alpha := w^\varepsilon_\alpha - \wp_\alpha(w^\varepsilon)$. By Lemma 4.1 and using (25), we have

$$\|\tilde{w}^\varepsilon\|_{L^2(\Omega_3; H^1(\Omega_{12}; \mathbb{R}^2))} \le C \sum_{\alpha,\beta} \|E_{\alpha\beta}(w^\varepsilon)\|_{L^2(\Omega)} \le C,$$

$⨍_{\Omega_{12}} \tilde{w}^\varepsilon \, dx' = 0$, and $\vartheta(\tilde{w}^\varepsilon) = 0$. Thus, up to a subsequence,

$$\tilde{w}^\varepsilon_\alpha \rightharpoonup \bar{w}_\alpha \quad \text{in } L^2(\Omega_3; H^1(\Omega_{12})) \text{ for } \alpha = 1, 2; \tag{26}$$

moreover, a.e.,

$$\fint_{\Omega_{12}} \bar{w}_\alpha dx' = 0,\ \vartheta(\bar{w}) = 0. \tag{27}$$

Hence $(\bar{w}_1, \bar{w}_2) \in RD_2^\perp(\Omega)$ (see (12)). Using (25) we also obtain

$$E_{11}(\bar{w}) = E_{12}(\bar{w}) = 0.$$

Moreover,

$$E_{22}^\varepsilon(u^\varepsilon) = E_{22}(w^\varepsilon) = E_{22}(\tilde{w}^\varepsilon) \rightharpoonup E_{22}(\bar{w}) \text{ in } L^2(\Omega).$$

Let

$$\tilde{w}_3^\varepsilon := w_3^\varepsilon - \left(x_1 x_2 \frac{d\vartheta(w^\varepsilon)}{dx_3} - x_1 \frac{d}{dx_3} \fint_{\Omega_{12}} w_1^\varepsilon \, dx' + \fint_{\Omega_1} w_3^\varepsilon \, dx_1 \right).$$

By Theorem 4.3 we have that

$$\|\tilde{w}_3^\varepsilon\|_{H^{-1}(\Omega_3; L^2(\Omega_{12}))} \le C \left(\sum_{\alpha,\beta} \|E_{\alpha\beta}(w^\varepsilon)\|_{L^2(\Omega)} + \|E_{13}(w^\varepsilon)\|_{L^2(\Omega)} \right) \le C.$$

Hence, up to a subsequence,

$$\tilde{w}_3^\varepsilon \rightharpoonup w_3 \quad \text{ in } H^{-1}(\Omega_3; L^2(\Omega_{12})).$$

Taking into account that (see (16))

$$\wp_1(w^\varepsilon) = \fint_{\Omega_{12}} w_1^\varepsilon \, dx' - x_2 \vartheta(w^\varepsilon), \quad \wp_2(w^\varepsilon) = \fint_{\Omega_{12}} w_2^\varepsilon \, dx' + x_1 \vartheta(w^\varepsilon),$$

we easily deduce that $E_{13}^\varepsilon(u^\varepsilon) = E_{13}(w^\varepsilon) = E_{13}(\tilde{w}^\varepsilon)$ and hence that

$$E_{13} = \frac{1}{2} \left(\frac{\partial \bar{w}_1}{\partial x_3} + \frac{\partial \bar{w}_3}{\partial x_1} \right).$$

Finally, since $\fint_{\Omega_1} \tilde{w}_3^\varepsilon \, dx_1 = 0$, we have that

$$\langle \bar{w}_3, \varphi \rangle_{H^{-1}(\Omega_3; L^2(\Omega_{12})) \times H_0^1(\Omega_3; L^2(\Omega_{12}))} = 0$$

for all $\varphi \in H_0^1(\Omega_3; L^2(\Omega_2))$.

The last part of the claim follows from the fact that $\mathscr{W}$ is a closed subspace of

$$\{ z \in RD_2^\perp(\Omega) \times H_{m_1}^{-1}(\Omega_3; L^2(\Omega_{12})) \ : \ E_{13}(z) \in L^2(\Omega) \}$$

which, in turn, is a Hilbert space under the scalar product

$$\langle z, \zeta \rangle := \langle z, \zeta \rangle_{RD_2^\perp(\Omega) \times H_{m_1}^{-1}(\Omega_3; L^2(\Omega_{12}))} + \int_\Omega E_{13}(z) E_{13}(\zeta) \, dx$$

and from an application of Proposition 4.1. □

LEMMA 5.5 (representation of the space $\mathscr{W}$). *The space $\mathscr{W}$ admits the following representation:*

$$\mathscr{W} = \Big\{ w \in L^2(\Omega_3; H^1(\Omega_{12}))^2 \times H^{-1}_{m_1}(\Omega_3; L^2(\Omega_{12})) \ : \ E_{13}(w) \in L^2(\Omega),$$

there exists $\eta_1 \in L^2(\Omega_3; H^2(\Omega_2))$ and $\eta_2 \in L^2(\Omega_3; H^1(\Omega_2))$ such that

$$w_1(x) = \eta_1(x_2, x_3), \quad w_2(x) = -x_1 \frac{\partial \eta_1}{\partial x_2}(x_2, x_3) + \eta_2(x_2, x_3), \tag{28}$$

$$\int_{\Omega_2} \eta_\alpha \, dx_2 = 0, \quad \int_{\Omega_2} \left(\frac{a_1^2}{12} \frac{\partial \eta_1}{\partial x_2} + x_2 \eta_1 \right) dx_2 = 0 \Big\}.$$

Proof. Let us call W the set on the right-hand side of equality (28), and let $\mathscr{W}$ be as in the statement of Lemma 5.4. Then it is trivial to check that $W \subseteq \mathscr{W}$. Let us prove the converse inequality. Let $w \in \mathscr{W}$ as in Lemma 5.4. Since $E_{11}(w) = E_{12}(w) = 0$, by integration we deduce that there exist $\eta_1 \in L^2(\Omega_3; H^2(\Omega_2))$ and $\eta_2 \in L^2(\Omega_3; H^1(\Omega_2))$ such that

$$w_1(x) = \eta_1(x_2, x_3), \quad w_2(x) = -x_1 \frac{\partial \eta_1}{\partial x_2}(x_2, x_3) + \eta_2(x_2, x_3).$$

Since

$$⨍_{\Omega_{12}} w_\alpha dx' = 0, \quad \vartheta(w) = 0, \tag{29}$$

we have that

$$\int_{\Omega_2} \eta_\alpha \, dx_2 = 0 \quad \text{a.e. for } \alpha = 1, 2,$$

and

$$\int_{\Omega_2} \left(\frac{a_1^2}{12} \frac{\partial \eta_1}{\partial x_2} + x_2 \eta_1 \right) dx_2 = 0,$$

a.e.. □

LEMMA 5.6 (component 12). *There exists a vector function $\bar{p}$ in the set*

$$\mathscr{P} = \{0\} \times H^1_{m_1}(\Omega_1; L^2(\Omega_{23})) \times \{0\}$$

such that

$$E_{12} = \frac{1}{2} \frac{\partial \bar{p}_2}{\partial x_1} = E_{12}(\bar{p}). \tag{30}$$

Moreover, $\mathscr{P}$ is a Hilbert space with the norm

$$\|p\|_{\mathscr{P}} := \|E_{12}(p)\|.$$

Proof. Let

$$p^\varepsilon_\alpha := \frac{1}{\varepsilon^3} u^\varepsilon_\alpha \quad \text{for } \alpha = 1, 2,$$

and

$$\bar{p}^\varepsilon_1 := p^\varepsilon_1 - ⨍_{\Omega_1} p^\varepsilon_1 \, dx_1, \quad \bar{p}^\varepsilon_2 := p^\varepsilon_2 - ⨍_{\Omega_1} p^\varepsilon_2 \, dx_1 + x_1 \frac{\partial}{\partial x_2} ⨍_{\Omega_1} p^\varepsilon_1 \, dx_1.$$

Then $E_{11}(\bar{p}^\varepsilon) = E_{11}(p^\varepsilon) = \varepsilon E^\varepsilon_{11}(u^\varepsilon)$; hence, by Theorem 4.1 we have

$$\bar{p}^\varepsilon_1 \to 0, \quad \frac{\partial \bar{p}^\varepsilon_1}{\partial x_1} \to 0 \quad \text{in } L^2(\Omega).$$

Since $E_{12}(\bar{p}^\varepsilon) = E_{12}(p^\varepsilon) = E^\varepsilon_{12}(u^\varepsilon)$, by Theorem 4.2 we also have that, up to subsequences,

$$\bar{p}^\varepsilon_2 \rightharpoonup \bar{p}_2 \text{ in } H^{-1}(\Omega_2; L^2(\Omega_{13})).$$

Setting $\bar{p} := (0, \bar{p}_2, 0)$, we have

$$E_{12} = E_{12}(\bar{p}) = \frac{1}{2}\frac{\partial \bar{p}_2}{\partial x_1},$$

that is, (30). It remains then to prove that $\bar{p}_2 \in H^1(\Omega_1; L^2(\Omega_{23}))$ and that $⨍_{\Omega_1} \bar{p}_2 \, dx_1 = 0$.

Since $\bar{p}_2 \in H^{-1}(\Omega_2; L^2(\Omega_{13}))$, for any $\varphi \in H^1_0(\Omega_2)$ and any $\psi \in L^2(\Omega_{13})$ the product $\varphi \otimes \psi$ belongs to $H^1_0(\Omega_2; L^2(\Omega_{13}))$, and the linear map

$$P_\varphi : L^2(\Omega_{13}) \to \mathbb{R}, \quad \langle P_\varphi, \psi \rangle := \langle \bar{p}_2, \varphi \otimes \psi \rangle$$

satisfies the estimate

$$|\langle P_\varphi, \psi \rangle| \le \|\bar{p}_2\|_{H^{-1}} \|\varphi\|_{H^1_0} \|\psi\|_{L^2}.$$

Thus $P_\varphi \in L^2(\Omega_{13})$. Moreover, from the definition of P_φ and the fact that $\frac{\partial \bar{p}_2}{\partial x_1} \in L^2(\Omega)$, we obtain that $\frac{\partial P_\varphi}{\partial x_1} \in L^2(\Omega_{13})$ and also that

$$\frac{\partial P_\varphi}{\partial x_1} = \int_{\Omega_2} \frac{\partial \bar{p}_2}{\partial x_1} \varphi \, dx_2. \tag{31}$$

Since $⨍_{\Omega_1} \bar{p}^\varepsilon_2 \, dx_1 = 0$ it follows from the definitions that

$$⨍_{\Omega_1} P_\varphi(s, x_3) \, ds = \langle \bar{p}_2, \varphi \rangle = 0 \tag{32}$$

for almost every $x_3 \in \Omega_3$; using this fact, the following Poincaré inequality holds:

$$\|P_\varphi(\cdot, x_3)\|_{L^2(\Omega_1)} \le C \left\| \frac{\partial P_\varphi}{\partial x_1}(\cdot, x_3) \right\|_{L^2(\Omega_1)},$$

where the constant C depends only on the domain Ω_1 and is therefore independent of x_3. By substituting (31) inside the Poincaré inequality, we obtain that

$$\|P_\varphi\|_{L^2(\Omega_{13})} \le C \left\| \frac{\partial \bar{p}_2}{\partial x_1} \right\|_{L^2(\Omega)} \|\varphi\|_{L^2(\Omega_2)}. \tag{33}$$

Using the density of $C^\infty_c(\Omega_2) \otimes C^\infty_c(\Omega_{13})$ in $L^2(\Omega)$ (see, for instance, Treves [10, Theorem 39.2 and subsequent Corollary 3]), the fact that $P_\varphi \in L^2(\Omega_{13})$, and inequality (33), we have that

$$\begin{aligned}
\|\bar{p}_2\|_{L^2(\Omega)} &= \sup_{\varphi \in C^\infty_c(\Omega_2),\ \psi \in C^\infty_c(\Omega_{13})} \frac{|\langle \bar{p}_2, \varphi\psi \rangle|}{\|\varphi\|_{L^2(\Omega_2)} \|\psi\|_{L^2(\Omega_{13})}} \\
&= \sup_{\varphi \in C^\infty_c(\Omega_2),\ \psi \in C^\infty_c(\Omega_{13})} \frac{|\langle P_\varphi, \psi \rangle|}{\|\varphi\|_{L^2(\Omega_2)} \|\psi\|_{L^2(\Omega_{13})}} \\
&\le \sup_{\varphi \in C^\infty_c(\Omega_2)} \frac{\|P_\varphi\|_{L^2(\Omega_{13})}}{\|\varphi\|_{L^2(\Omega_2)}} \le C \left\| \frac{\partial \bar{p}_2}{\partial x_1} \right\|_{L^2(\Omega)};
\end{aligned}$$

hence $\bar{p}_2 \in L^2(\Omega)$. Thus $\bar{p}_2 \in H^1(\Omega_1; L^2(\Omega_{23}))$, and (32) implies $⨍_{\Omega_1} \bar{p}_2\, dx_1 = 0$.

The last part of the claim follows from the fact that $H^1_{m_1}(\Omega_1; L^2(\Omega_{23}))$ is a Hilbert space with the norm

$$\|p\|_{H^1_{m_1}(\Omega_1;L^2(\Omega_{23}))} := \left\| \frac{\partial p_2}{\partial x_1} \right\|_{L^2(\Omega)},$$

which, by Theorem 4.1, turns out to be equivalent to the canonical one. ∎

LEMMA 5.7 (component 11). *There exists a function $\bar{q}$ in the space*

$$\mathscr{Q} := H^1_{m_1}(\Omega_1; L^2(\Omega_{23})) \times \{0\}^2$$

such that

$$E_{11} = \frac{\partial \bar{q}_1}{\partial x_1} = E_{11}(\bar{q}).$$

Moreover, $\mathscr{Q}$ is a Hilbert space with the norm

$$\|q\|_{\mathscr{Q}} := \|E_{11}(q)\|.$$

Proof. Let

$$q_1^\varepsilon := \frac{1}{\varepsilon^4}\left(u_1^\varepsilon - ⨍_{\Omega_1} u_1^\varepsilon dx_1 \right);$$

then

$$\sup_\varepsilon \left\| \frac{\partial q_1^\varepsilon}{\partial x_1} \right\|_{L^2(\Omega)} = \sup_\varepsilon \|E_{11}^\varepsilon(u^\varepsilon)\|_{L^2(\Omega)} \le C,$$

and by Theorem 4.1 we have $\sup_\varepsilon \|q_1^\varepsilon\|_{L^2(\Omega)} \le C$. Then, up to a subsequence, $q_1^\varepsilon \rightharpoonup \bar{q}_1$ in $L^2(\Omega)$, and $E_{11}^\varepsilon(u^\varepsilon) = \partial \bar{q}_1^\varepsilon / \partial x_1 \rightharpoonup \partial \bar{q}_1 / \partial x_1$ in $L^2(\Omega)$ for some $\bar{q}_1 \in L^2(\Omega)$.

The last part of the claim follows from the fact that $H^1_{m_1}(\Omega_1; L^2(\Omega_{23}))$ is a Hilbert space with the norm

$$\|q\|_{H^1_{m_1}(\Omega_1;L^2(\Omega_{23}))} := \|q_{1,1}\|_{L^2(\Omega)},$$

which, by Theorem 4.1, turns out to be equivalent to the canonical one. ∎

6. The limit problem. Let us consider the space $\mathscr{A} := \mathscr{U} \times \mathscr{V} \times \mathscr{W} \times \mathscr{P} \times \mathscr{Q}$. According to the notation and the results proved in the previous section, $\mathscr{A}$ is a Hilbert space when endowed with the product norm

$$\|(u, v, w, p, q)\|_{\mathscr{A}} := \|u\|_{\mathscr{U}} + \|v\|_{\mathscr{V}} + \|w\|_{\mathscr{W}} + \|p\|_{\mathscr{P}} + \|q\|_{\mathscr{Q}}.$$

Given a 5-tuple of vector valued distributions $(u, v, w, p, q) \in \mathcal{D}'(\Omega; \mathbb{R}^3)^5$, let us define

$$E(u, v, w, p, q) := \begin{pmatrix} E_{11}(q) & E_{12}(p) & E_{13}(w) \\ & E_{22}(w) & E_{23}(v) \\ \text{Sym.} & & E_{33}(u) \end{pmatrix}. \tag{34}$$

We are now in a position to state the main result of this paper.

THEOREM 6.1. *Let $\mathbb{C}$ be a positive definite fourth order tensor field on Ω with the minor symmetries, i.e., $\mathbb{C}_{ijkl} = \mathbb{C}_{jikl} = \mathbb{C}_{ijlk}$. Let F^ε be a second order symmetric tensor field which belongs to $L^2(\Omega;\mathbb{R}^{3\times 3})$. Then problem* (6)*, that is,*

$$\text{(35)}\qquad \begin{cases} u^\varepsilon \in H^1_{dn}(\Omega;\mathbb{R}^3), \\ \displaystyle\int_\Omega \mathbb{C}E^\varepsilon(u^\varepsilon)\cdot E^\varepsilon(\varphi)\,dx = \int_\Omega F^\varepsilon \cdot E^\varepsilon(\varphi)\,dx \quad \forall \varphi \in H^1_{dn}(\Omega;\mathbb{R}^3), \end{cases}$$

admits a unique solution u^ε. Moreover, if $F^\varepsilon \to F$ in $L^2(\Omega;\mathbb{R}^{3\times 3})$, then we have the following:

1. *the problem*

$$\text{(36)}\qquad \int_\Omega \mathbb{C}E(\bar u,\bar v,\bar w,\bar p,\bar q)\cdot E(u,v,w,p,q)\,dx = \int_\Omega F\cdot E(u,v,w,p,q)\,dx \quad \forall\,(u,v,w,p,q)\in \mathscr{A}$$

   admits a unique solution $(\bar u,\bar v,\bar w,\bar p,\bar q)\in\mathscr{A}$;
2. *$u^\varepsilon \rightharpoonup \bar u$ in $H^1(\Omega;\mathbb{R}^3)$;*
3. *$E^\varepsilon(u^\varepsilon)\to E(\bar u,\bar v,\bar w,\bar p,\bar q)$ in $L^2(\Omega;\mathbb{R}^{3\times 3})$.*

The following corollary can be seen as a corrector result.

COROLLARY 6.1. *If the solution $(\bar u,\bar v,\bar w,\bar p,\bar q)$ of problem* (36) *is such that*

$$\text{(37)}\qquad \frac{\partial \bar v_3}{\partial x_3},\ E_{23}(\bar w),\ \frac{\partial \bar w_3}{\partial x_3},\ \frac{\partial \bar q_1}{\partial x_2},\ \frac{\partial \bar q_1}{\partial x_3},\ \frac{\partial \bar p_2}{\partial x_2},\ \frac{\partial \bar p_2}{\partial x_3} \in L^2(\Omega),$$

then

$$\|E^\varepsilon(u^\varepsilon) - E^\varepsilon(\bar u^\varepsilon)\|_{L^2(\Omega;\mathbb{R}^{3\times 3})}\to 0,$$

where

$$\bar u^\varepsilon := \bar u + \varepsilon\bar v + \varepsilon^2\bar w + \varepsilon^3\bar p + \varepsilon^4\bar q.$$

Proof. Since

$$E^\varepsilon(\bar u^\varepsilon) = E(\bar u,\bar v,\bar w,\bar p,\bar q) + \varepsilon\begin{pmatrix} 0 & E_{12}(\bar q) & 0 \\ & E_{22}(\bar p) & E_{23}(\bar w) \\ \text{Sym.} & & E_{33}(\bar v)\end{pmatrix} + \varepsilon^2\begin{pmatrix} 0 & 0 & E_{13}(\bar q) \\ & 0 & E_{23}(\bar p) \\ \text{Sym.} & & E_{33}(\bar w)\end{pmatrix}$$

the additional regularity assumptions imply that $E^\varepsilon(\bar u^\varepsilon)\in L^2(\Omega;\mathbb{R}^{3\times 3})$ and

$$\|E^\varepsilon(\bar u^\varepsilon) - E(\bar u,\bar v,\bar w,\bar p,\bar q)\|_{L^2(\Omega;\mathbb{R}^{3\times 3})}\to 0.$$

Then the claim follows from step 3 of Theorem 6.1. ∎

In order to prove Theorem 6.1, we introduce the subspaces of $H^1_{dn}(\Omega;\mathbb{R}^3)$,

$$\hat{\mathscr{U}} := \{u\in H^1_{dn}(\Omega;\mathbb{R}^3)\ :\ E_{\alpha\beta}(u) = E_{\alpha 3}(u) = 0\},$$
$$\hat{\mathscr{V}} := \{v\in H^1_{dn}(\Omega;\mathbb{R}^3)\ :\ E_{\alpha\beta}(v) = E_{13}(v) = 0\},$$
$$\hat{\mathscr{W}} := \{w\in H^1_{dn}(\Omega;\mathbb{R}^3)\ :\ E_{1\beta}(w) = 0\},$$
$$\hat{\mathscr{P}} := \{p\in H^1_{dn}(\Omega;\mathbb{R}^3)\ :\ E_{11}(p) = 0\},$$
$$\hat{\mathscr{Q}} := H^1_{dn}(\Omega;\mathbb{R}^3),$$

and define $\hat{\mathscr{A}} := \hat{\mathscr{U}} \times \hat{\mathscr{V}} \times \hat{\mathscr{W}} \times \hat{\mathscr{P}} \times \hat{\mathscr{Q}}$.

Let us note that $\mathscr{U} = \hat{\mathscr{U}}$, but similar equalities are not true for the spaces $\mathscr{V}$, $\mathscr{W}$, $\mathscr{P}$, and $\mathscr{Q}$. Nevertheless, for such spaces we can prove the following approximation lemma.

LEMMA 6.1. *For every $v \in \mathscr{V}$, $w \in \mathscr{W}$, $p \in \mathscr{P}$, and $q \in \mathscr{Q}$ there exist sequences $(\hat{v}^n)$ in $\hat{\mathscr{V}}$, $(\hat{w}^n)$ in $\hat{\mathscr{W}}$, $(\hat{p}^n)$ in $\hat{\mathscr{P}}$, and $(\hat{q}^n)$ in $\hat{\mathscr{Q}}$ such that the following convergences hold in the norm of $L^2(\Omega)$:*

(i) $E_{23}(\hat{v}^n) \to E_{23}(v)$,
(ii) $E_{13}(\hat{w}^n) \to E_{13}(w)$ *and* $E_{22}(\hat{w}^n) \to E_{22}(w)$,
(iii) $E_{12}(\hat{p}^n) \to E_{12}(p)$,
(iv) $E_{11}(\hat{q}^n) \to E_{11}(q)$.

Proof. Let us prove (i). Since $v \in \mathscr{V}$ (see Lemma 5.2), there exist $\vartheta \in H^1(\Omega_3)$ with $\vartheta(0) = 0$ and $\varrho \in L^2(\Omega_3; H^1(\Omega_2))$ such that

$$v_1(x) = -x_2\vartheta(x_3), \quad v_2(x) = x_1\vartheta(x_3),$$
$$v_3(x) = x_1x_2\frac{d\vartheta}{dx_3}(x_3) + \varrho(x_2, x_3).$$

Then there exist sequences $\vartheta_n \in H^2_{dn}(\Omega_3)$ and $\varrho_n \in H^1_{dn}(\Omega_{23})$ such that

$$\vartheta_n \to \vartheta \text{ in } H^1(\Omega_3), \quad \varrho_n \to \varrho \text{ in } L^2(\Omega_3; H^1(\Omega_2)).$$

Setting

$$\hat{v}^n_1(x) = -x_2\vartheta_n(x_3), \quad \hat{v}^n_2(x) = x_1\vartheta_n(x_3),$$
$$\hat{v}^n_3(x) = x_1x_2\frac{d\vartheta_n}{dx_3}(x_3) + \varrho_n(x_2, x_3),$$

we obtain the claim.

Let us prove (ii). As $w \in \mathscr{W}$ (see Lemma 5.4), there exist $\eta_1 \in L^2(\Omega_3; H^2(\Omega_2))$ and $\eta_2 \in L^2(\Omega_3; H^1(\Omega_2))$ such that

$$w_1(x) = \eta_1(x_2, x_3), \quad w_2(x) = -x_1\frac{\partial\eta_1}{\partial x_2}(x_2, x_3) + \eta_2(x_2, x_3).$$

Then, there exist a sequence $\eta^n_1 \in H^2_{dn}(\Omega)$ with $\partial\eta^n_1/\partial x_1 = 0$ such that

$$\eta^n_1 \to \eta_1 \quad \text{in } L^2(\Omega_3; H^2(\Omega_2))$$

and a sequence $\eta^n_2 \in H^1_{dn}(\Omega)$ with $\partial\eta^n_2/\partial x_1 = 0$ such that

$$\eta^n_2 \to \eta_2 \quad \text{in } L^2(\Omega_3; H^1(\Omega_2)).$$

Since $\partial w_3/\partial x_1 = 2E_{13}(w) - \partial\eta_1/\partial x_3$ and $E_{13}(w) \in L^2(\Omega)$, by integration, there exists $G_{13} \in H^1(\Omega_1; L^2(\Omega_{23}))$ such that

$$\frac{\partial G_{13}}{\partial x_1} = E_{13}(w) \quad \text{and} \quad w_3 = 2G_{13} - x_1\frac{\partial\eta_1}{\partial x_3},$$

where we have also used the fact that η_1 does not depend on x_1.

We may also find a sequence $G^n_{13} \in H^1_{dn}(\Omega)$ such that

$$G^n_{13} \to G_{13} \quad \text{in } H^1(\Omega_1; L^2(\Omega_{23}))$$

and define

$$\hat{w}_1^n := \eta_1^n, \quad \hat{w}_2^n := -x_1 \frac{\partial \eta_1^n}{\partial x_2} + \eta_2^n, \quad \hat{w}_3^n := G_{13}^n - x_1 \frac{\partial \eta_1^n}{\partial x_3}.$$

Then $\hat{w}^n \in H^1_{dn}(\Omega)$, $E_{11}(\hat{w}^n) = \partial \eta_1^n / \partial x_1 = 0$, and $E_{12}(\hat{w}^n) = 0$ so that $\hat{w}^n \in \hat{\mathscr{W}}$. Moreover,

$$E_{13}(\hat{w}^n) = \frac{\partial G_{13}^n}{\partial x_1} \rightarrow \frac{\partial G_{13}}{\partial x_1} = E_{13}(w) \quad \text{in } L^2(\Omega),$$

and

$$E_{22}(\hat{w}^n) = \frac{\partial \hat{w}_2^n}{\partial x_2} = -x_1 \frac{\partial^2 \eta_1^n}{\partial x_2^2} + \frac{\partial \eta_2^n}{\partial x_2} \rightarrow \frac{\partial \hat{w}_2}{\partial x_2} = E_{22}(\hat{w}) \quad \text{in } L^2(\Omega).$$

To prove (iii) it is enough to consider, for a given $p = (0, p_2, 0) \in \mathscr{P}$, a sequence $\hat{p}_2^n \in H^1_{dn}(\Omega)$, which converges to p_2 in the norm of $H^1(\Omega_1; L^2(\Omega_{23}))$, and set $\hat{p}^n = (0, \hat{p}_2^n, 0)$. Finally, claim (iv) simply follows from the density of $H^1_{dn}(\Omega)$ in $H^1_{m_1}(\Omega_1; L^2(\Omega_{23}))$. □

Proof of Theorem 6.1. The existence and uniqueness of the solution of problem (36) follows from an application of the Lax–Milgram lemma to the symmetric bilinear form defined on $\mathscr{A}$ by

$$a[(u, v, w, p, q), (\hat{u}, \hat{v}, \hat{w}, \hat{p}, \hat{q})] := \int_\Omega \mathbb{C} E(u, v, w, p, q) \cdot E(\hat{u}, \hat{v}, \hat{w}, \hat{p}, \hat{q})\, dx,$$

which is continuous and coercive with respect to the Hilbertian norm on $\mathscr{A}$ defined at the beginning of this section.

Part 2 of the statement of Theorem 6.1 is actually a consequence of step 3. Let us now prove part 3. According to the results proved in the previous section, we have

$$E^\varepsilon(u^\varepsilon) \rightharpoonup E \text{ in } L^2(\Omega; \mathbb{R}^{3\times 3}), \tag{38}$$

and there exists a $(\bar{u}, \bar{v}, \bar{w}, \bar{p}, \bar{q}) \in \mathscr{A}$ such that

$$E = \begin{pmatrix} E_{11}(\bar{q}) & E_{12}(\bar{p}) & E_{13}(\bar{w}) \\ & E_{22}(\bar{w}) & E_{23}(\bar{v}) \\ \text{Sym.} & & E_{33}(\bar{u}) \end{pmatrix} = E(\bar{u}, \bar{v}, \bar{w}, \bar{p}, \bar{q}).$$

The result will be achieved in two steps: (i) we prove that $(\bar{u}, \bar{v}, \bar{w}, \bar{p}, \bar{q})$ satisfies equality (36) and therefore coincides with the unique solution of the variational problem, and (ii) we show that the convergence in (38) is indeed strong.

Let $(\hat{u}, \hat{v}, \hat{w}, \hat{p}, \hat{q}) \in \hat{\mathscr{A}}$ and set

$$\hat{\varphi}^\varepsilon := \hat{u} + \varepsilon \hat{v} + \varepsilon^2 \hat{w} + \varepsilon^3 \hat{p} + \varepsilon^4 \hat{q};$$

then $\hat{\varphi}^\varepsilon \in H^1_{dn}(\Omega; \mathbb{R}^3)$ and an easy computation shows that, as $\varepsilon \rightarrow 0$,

$$E^\varepsilon(\hat{\varphi}^\varepsilon) \rightarrow \begin{pmatrix} E_{11}(\hat{q}) & E_{12}(\hat{p}) & E_{13}(\hat{w}) \\ & E_{22}(\hat{w}) & E_{23}(\hat{v}) \\ \text{Sym.} & & E_{33}(\hat{u}) \end{pmatrix} = E(\hat{u}, \hat{v}, \hat{w}, \hat{p}, \hat{q}) \tag{39}$$

in the norm convergence of $L^2(\Omega;\mathbb{R}^{3\times 3})$. Taking $\varphi = \hat{\varphi}^\varepsilon$ in (35) and passing to the limit we find

$$\int_\Omega \mathbb{C}E(\bar{u},\bar{v},\bar{w},\bar{p},\bar{q})\cdot E(\hat{u},\hat{v},\hat{w},\hat{p},\hat{q})\,dx = \int_\Omega F\cdot E(\hat{u},\hat{v},\hat{w},\hat{p},\hat{q})\,dx,$$

for every $(\hat{u},\hat{v},\hat{w},\hat{p},\hat{q}) \in \hat{\mathscr{A}}$. This equality holds in fact for any $(u,v,w,p,q) \in \mathscr{A}$ in place of $(\hat{u},\hat{v},\hat{w},\hat{p},\hat{q}) \in \hat{\mathscr{A}}$ because of the approximation Lemma 6.1 which ensures that there exists a sequence $(u,\hat{v}^n,\hat{w}^n,\hat{p}^n,\hat{q}^n) \in \hat{\mathscr{A}}$ such that

$$\|(u,\hat{v}^n,\hat{w}^n,\hat{p}^n,\hat{q}^n) - (u,v,w,p,q)\|_{\mathscr{A}} \to 0.$$

To show that the convergence in (38) is indeed strong, it suffices to prove that $\lim_{\varepsilon\to 0}\|E^\varepsilon(u^\varepsilon)\|_{L^2(\Omega;\mathbb{R}^{3\times 3})} = \|E(\bar{u},\bar{v},\bar{w},\bar{p},\bar{q})\|_{L^2(\Omega;\mathbb{R}^{3\times 3})}$ or, equivalently, that

$$\begin{aligned}\lim_{\varepsilon\to 0}\int_\Omega \mathbb{C}E^\varepsilon(u^\varepsilon)\cdot E^\varepsilon(u^\varepsilon)\,dx &= \lim_{\varepsilon\to 0}\int_\Omega F^\varepsilon\cdot E^\varepsilon(u^\varepsilon)\,dx\\ &= \int_\Omega F\cdot E(\bar{u},\bar{v},\bar{w},\bar{p},\bar{q})\,dx\\ &= \int_\Omega \mathbb{C}E(\bar{u},\bar{v},\bar{w},\bar{p},\bar{q})\cdot E(\bar{u},\bar{v},\bar{w},\bar{p},\bar{q})\,dx,\end{aligned}$$

where we passed to the limit thanks to the strong convergence of F^ε. □

7. Equilibrium differential equations. In this last section we derive the differential formulation of the limit problem. For simplicity we assume here that the elasticity tensor also satisfies the major symmetries; that is, $\mathbb{C}_{ijkl} = \mathbb{C}_{klij}$ for any i,j,k,l. Nevertheless, the same computation can be performed also in the general case.

To make (36) more explicit and to keep the notation compact, in writing the elasticity tensor components $\mathbb{C}_{ijkl}$ we associate to a pair of components ij a single component s following the rule $11\mapsto 1, 22\mapsto 2, 33\mapsto 3, 23\mapsto 4, 13\mapsto 5, 12\mapsto 6$, and we write, for instance, c_{14} for $\mathbb{C}_{1123}$; see Auld [1] for more details on the notation used. Clearly $c_{ij} = c_{ji}$. Still, for brevity, define $\bar{e}_1 = E_{11}(\bar{q})$, $\bar{e}_2 = E_{22}(\bar{w})$, $\bar{e}_3 = E_{33}(\bar{u})$, $\bar{e}_4 = 2E_{23}(\bar{v})$, $\bar{e}_5 = 2E_{13}(\bar{w})$, $\bar{e}_6 = 2E_{12}(\bar{p})$. Letting

$$\mathscr{L}(u,v,w,p,q) := \int_\Omega F\cdot E(u,v,w,p,q)\,dx,$$

we can then rewrite (36) as

$$\begin{aligned}\int_\Omega \sum_{j=1}^6 \big[&c_{1j}\bar{e}_j E_{11}(q) + c_{2j}\bar{e}_j E_{22}(w) + c_{3j}\bar{e}_j E_{33}(u) + 2c_{4j}\bar{e}_j E_{23}(v)\\ &+ 2c_{5j}\bar{e}_j E_{13}(w) + 2c_{6j}\bar{e}_j E_{12}(p)\big]\,dx = \mathscr{L}(u,v,w,p,q),\end{aligned}$$

for every $(u,v,w,p,q)\in\mathscr{A}$. Thus

$$\int_\Omega \sum_{j=1}^{6} c_{1j}\bar{e}_j E_{11}(q)\,dx = \mathscr{L}(0,0,0,0,q), \tag{40}$$

$$\int_\Omega \sum_{j=1}^{6} 2c_{6j}\bar{e}_j E_{12}(p)\,dx = \mathscr{L}(0,0,0,p,0), \tag{41}$$

$$\int_\Omega \sum_{j=1}^{6} \left[c_{2j}\bar{e}_j E_{22}(w) + 2c_{5j}\bar{e}_j E_{13}(w)\right] dx = \mathscr{L}(0,0,w,0,0), \tag{42}$$

$$\int_\Omega \sum_{j=1}^{6} 2c_{4j}\bar{e}_j E_{23}(v)\,dx = \mathscr{L}(0,v,0,0,0), \tag{43}$$

$$\int_\Omega \sum_{j=1}^{6} c_{3j}\bar{e}_j E_{33}(u)\,dx = \mathscr{L}(u,0,0,0,0). \tag{44}$$

In this section, for simplicity, we assume

$$\begin{gathered}\mathscr{L}(0,0,\cdot,\cdot,\cdot)=0,\quad \mathscr{L}(0,v,0,0,0)=\int_0^\ell m(x_3)\vartheta(x_3)\,dx_3,\\ \mathscr{L}(u,0,0,0,0)=\int_\Omega b\cdot u\,dx+\int_{\Gamma_\ell} s\cdot u\,d\mathscr{H}^2,\end{gathered} \tag{45}$$

where $\Gamma_\ell=\partial\Omega\cap\{x_3=\ell\}$; $u\in\mathscr{U}$; $b\in L^2(\Omega)$; $s\in L^2(\Gamma_\ell)$; $v\in\mathscr{V}$; $\vartheta\in H^1(\Omega_3)$, $\vartheta(0)=0$, is related to v as in Lemma 5.2; and $m\in L^2(\Omega_3)$. Such assumptions are quite often satisfied in engineering applications.

We now derive the equilibrium equations in differential form. Let $\psi\in L^2(\Omega)$ and define

$$q_1 := \int_{-a_1/2}^{x_1} \psi(s,\cdot,\cdot)\,ds - \fint_{\Omega_1}\int_{-a_1/2}^{x_1} \psi(s,\cdot,\cdot)\,ds\,dx_1.$$

Then $q:=(q_1,0,0)\in\mathscr{Q}$ and $E_{11}(q)=\psi$; hence, from (40) and (45) we deduce

$$\sum_{j=1}^{6} c_{1j}\bar{e}_j = 0 \quad \text{a.e..} \tag{46}$$

With the same argument it follows from (41) that

$$\sum_{j=1}^{6} c_{6j}\bar{e}_j = 0 \quad \text{a.e..} \tag{47}$$

From (46) and (47) we deduce, since $c_{11}c_{66}-c_{16}^2>0$, that

$$\bar{e}_1 = E_{11}(\bar{q}) = -\sum_{j=2}^{5} \frac{c_{66}c_{1j}-c_{16}c_{6j}}{c_{11}c_{66}-c_{16}^2}\,\bar{e}_j \quad \text{a.e.,} \tag{48}$$

$$\bar{e}_6 = 2E_{12}(\bar{p}) = -\sum_{j=2}^{5} \frac{c_{11}c_{6j}-c_{16}c_{1j}}{c_{11}c_{66}-c_{16}^2}\,\bar{e}_j \quad \text{a.e..} \tag{49}$$

Using (45), (48), and (49) we can rewrite (42), after setting

$$\tilde{c}_{ij} = c_{ij} - c_{i1}\frac{c_{66}c_{1j} - c_{16}c_{6j}}{c_{11}c_{66} - c_{16}^2} - c_{i6}\frac{c_{11}c_{6j} - c_{16}c_{1j}}{c_{11}c_{66} - c_{16}^2},$$

for $i, j = 2, \dots, 5$, as

$$\int_\Omega \sum_{j=2}^{5} \left[\tilde{c}_{2j}\bar{e}_j E_{22}(w) + 2\tilde{c}_{5j}\bar{e}_j E_{13}(w)\right] dx = \mathscr{L}(0,0,w,0,0) = 0. \tag{50}$$

Since $w \in \mathscr{W}$, it then admits the representation given in Lemma 5.5 in terms of functions η_1 and η_2. Choosing $\eta_1 = \eta_2 = 0$, so that $E_{22}(w) = 0$, and w_3 like it has been chosen q_1 previously, we find from (50) that

$$\sum_{j=2}^{5} \tilde{c}_{5j}\bar{e}_j = 0 \quad \text{a.e..} \tag{51}$$

Let $\psi \in L^2(\Omega_{23})$. Taking $\eta_1 = w_3 = 0$ and

$$\eta_2 := \int_{-a_2/2}^{x_2} \psi(s,\cdot)\, ds - \fint_{\Omega_2} \int_{-a_2/2}^{x_2} \psi(s,\cdot)\, ds\, dx_2$$

so that $E_{22}(w) = \psi$, we find from (50) that

$$\sum_{j=2}^{5} \int_{\Omega_1} \tilde{c}_{2j}\bar{e}_j\, dx_1 = 0 \quad \text{a.e..} \tag{52}$$

Taking instead $\eta_2 = w_3 = 0$ and

$$\eta_1 := \int_{-a_2/2}^{x_2} \int_{-a_2/2}^{t} \psi(s,\cdot)\, ds\, dt - K_1 x_2 - K_2,$$

where the constants K_1 and K_2 are chosen in order to satisfy the mean integral conditions required on η_1 by (28), we have $E_{22}(w) = -x_1\psi$, and hence, from (50) we deduce

$$\sum_{j=2}^{5} \int_{\Omega_1} x_1\tilde{c}_{2j}\bar{e}_j\, dx_1 = 0 \quad \text{a.e..} \tag{53}$$

From (51), and observing that the positive definiteness of the elastic tensor implies $\tilde{c}_{55} > 0$, we find

$$\bar{e}_5 = 2E_{13}(\bar{w}) = -\sum_{j=2}^{4} \frac{\tilde{c}_{5j}}{\tilde{c}_{55}}\, \bar{e}_j \quad \text{a.e..} \tag{54}$$

To solve (52) and (53) we need to write explicitly $\bar{e}_2 = E_{22}(\bar{w})$. Since $\bar{w} \in \mathscr{W}$, by Lemma 5.5, we can write

$$\bar{w}_1(x) = \bar{\eta}_1(x_2, x_3), \quad \bar{w}_2(x) = -x_1\frac{\partial\bar{\eta}_1}{\partial x_2}(x_2, x_3) + \bar{\eta}_2(x_2, x_3),$$

where $\bar{\eta}_1$ and $\bar{\eta}_2$ belong to the appropriate spaces. Since

$$\bar{e}_2 = -x_1 \frac{\partial^2 \bar{\eta}_1}{\partial x_2^2} + \frac{\partial \bar{\eta}_2}{\partial x_2} \tag{55}$$

and using (54), we can rewrite (52) and (53) as

$$s_{1;22} \frac{\partial^2 \bar{\eta}_1}{\partial x_2^2} - s_{0;22} \frac{\partial \bar{\eta}_2}{\partial x_2} = \sum_{j=3}^{4} \fint_{\Omega_1} \hat{c}_{2j} \bar{e}_j \, dx_1,$$

$$s_{2;22} \frac{\partial^2 \bar{\eta}_1}{\partial x_2^2} - s_{1;22} \frac{\partial \bar{\eta}_2}{\partial x_2} = \sum_{j=3}^{4} \fint_{\Omega_1} x_1 \hat{c}_{2j} \bar{e}_j \, dx_1,$$

where we have set

$$\hat{c}_{ij} := \frac{\tilde{c}_{55}\tilde{c}_{ij} - \tilde{c}_{i5}\tilde{c}_{j5}}{\tilde{c}_{55}}, \qquad s_{k;ij} := \fint_{\Omega_1} x_1^k \, \hat{c}_{ij} \, dx_1, \tag{56}$$

for $i, j = 2, 3, 4$ and $k = 0, 1, 2$. From these equations we find

$$\frac{\partial^2 \bar{\eta}_1}{\partial x_2^2} = \frac{1}{s_{0;22}s_{2;22} - s_{1;22}^2} \biggl(s_{0;22} \sum_{j=3}^{4} \fint_{\Omega_1} x_1 \hat{c}_{2j} \bar{e}_j \, dx_1 - s_{1;22} \sum_{j=3}^{4} \fint_{\Omega_1} \hat{c}_{2j} \bar{e}_j \, dx_1 \biggr),$$

$$\frac{\partial \bar{\eta}_2}{\partial x_2} = \frac{1}{s_{0;22}s_{2;22} - s_{1;22}^2} \biggl(s_{1;22} \sum_{j=3}^{4} \fint_{\Omega_1} x_1 \hat{c}_{2j} \bar{e}_j \, dx_1 - s_{2;22} \sum_{j=3}^{4} \fint_{\Omega_1} \hat{c}_{2j} \bar{e}_j \, dx_1 \biggr),$$

and then by integration $\bar{\eta}_1$ and $\bar{\eta}_2$ (the fact that $s_{0;22}s_{2;22} - s_{1;22}^2 > 0$ can be checked, for instance, by using Hölder's inequality; see Wheeden and Zygmund [13, Chapter 8, Exercise 4]).

According to Remark 5.1 and Lemma 5.2, we let

$$\bar{e}_3 = \bar{\zeta}_3' - x_1 \bar{\zeta}_1'' - x_2 \bar{\zeta}_2'', \qquad \bar{e}_4 = 2x_1 \bar{\vartheta}' + \frac{\partial \bar{\varrho}}{\partial x_2}. \tag{57}$$

Setting

$$S_{mpqr}^{ijkl} := \frac{s_{i;2j}s_{k;2l} - s_{m;2p}s_{q;2r}}{s_{0;22}s_{2;22} - s_{1;22}^2},$$

we then have

$$\frac{\partial^2 \bar{\eta}_1}{\partial x_2^2} = S_{0312}^{0213} \bigl(\bar{\zeta}_3' - x_2 \bar{\zeta}_2'' \bigr) - S_{1213}^{0223} \bar{\zeta}_1'' + 2S_{1214}^{0224} \bar{\vartheta}' + S_{0412}^{0214} \frac{\partial \bar{\varrho}}{\partial x_2},$$

$$\frac{\partial \bar{\eta}_2}{\partial x_2} = S_{0322}^{1213} \bigl(\bar{\zeta}_3' - x_2 \bar{\zeta}_2'' \bigr) - S_{2213}^{1223} \bar{\zeta}_1'' + 2S_{1422}^{1224} \bar{\vartheta}' + S_{0422}^{1214} \frac{\partial \bar{\varrho}}{\partial x_2},$$

and taking into account the relations (48), (49), (54), (55), and (57), we find

$$\begin{aligned} \sum_{j=1}^{6} c_{ij} \bar{e}_j = {} & \bigl(\hat{c}_{i3} - x_1 \hat{c}_{i2} S_{0312}^{0213} + \hat{c}_{i2} S_{0322}^{1213} \bigr) \bigl(\bar{\zeta}_3' - x_2 \bar{\zeta}_2'' \bigr) \\ & - \bigl(x_1 \hat{c}_{i3} - x_1 \hat{c}_{i2} S_{1213}^{0223} + \hat{c}_{i2} S_{2213}^{1223} \bigr) \bar{\zeta}_1'' \\ & + 2 \bigl(x_1 \hat{c}_{i4} - x_1 \hat{c}_{i2} S_{1214}^{0224} + \hat{c}_{i2} S_{1422}^{1224} \bigr) \bar{\vartheta}' \\ & + \bigl(\hat{c}_{i4} - x_1 \hat{c}_{i2} S_{0412}^{0214} + \hat{c}_{i2} S_{0422}^{1214} \bigr) \frac{\partial \bar{\varrho}}{\partial x_2}, \end{aligned} \tag{58}$$

for $i = 3, 4$. Now let $v \in \mathscr{V}$ and ϑ and ϱ be as in Lemma 5.2. With $\psi \in L^2(\Omega_{23})$ and $\vartheta = 0$ and

$$\varrho := \int_{-a_2/2}^{x_2} \psi(s,\cdot)\,ds - \fint_{\Omega_2} \int_{-a_2/2}^{x_2} \psi(s,\cdot)\,ds\,dx_2,$$

we find from (43) that

$$\fint_{\Omega_1} \sum_{j=1}^{6} c_{4j}\bar{e}_j\,dx_1 = 0.$$

It then follows that

$$\begin{aligned} 0 = {} & \big(s_{0;43} - s_{1;42}S^{0213}_{0312} + s_{0;42}S^{1213}_{0322}\big)\big(\bar{\zeta}_3' - x_2\bar{\zeta}_2''\big) \\ & -\big(s_{1;43} - s_{1;42}S^{0223}_{1213} + s_{0;42}S^{1223}_{2213}\big)\bar{\zeta}_1'' \\ & +2\big(s_{1;44} - s_{1;42}S^{0224}_{1214} + s_{0;42}S^{1224}_{1422}\big)\bar{\vartheta}' \\ & +\big(s_{0;44} - s_{1;42}S^{0214}_{0412} + s_{0;42}S^{1214}_{0422}\big)\frac{\partial\bar{\varrho}}{\partial x_2}, \end{aligned}$$

and provided that the last coefficient $s_{0;44} - s_{1;42}S^{0214}_{0412} + s_{0;42}S^{1214}_{0422} \neq 0$, one finds

$$\begin{aligned} \frac{\partial\bar{\varrho}}{\partial x_2} = {} & \frac{-1}{s_{0;44} - s_{1;42}S^{0214}_{0412} + s_{0;42}S^{1214}_{0422}} \\ & \cdot\Big[\big(s_{0;43} - s_{1;42}S^{0213}_{0312} + s_{0;42}S^{1213}_{0322}\big)\big(\bar{\zeta}_3' - x_2\bar{\zeta}_2''\big) \\ & - \big(s_{1;43} - s_{1;42}S^{0223}_{1213} + s_{0;42}S^{1223}_{2213}\big)\bar{\zeta}_1'' \\ & + \big(s_{1;44} - s_{1;42}S^{0224}_{1214} + s_{0;42}S^{1224}_{1422}\big)2\bar{\vartheta}'\Big]. \end{aligned}$$

We may rewrite (58) as

$$\sum_{j=1}^{6} c_{ij}\bar{e}_j = F_{i3}\bar{\zeta}_3' - F_{i2}\bar{\zeta}_2'' - F_{i1}\bar{\zeta}_1'' + F_{i4}\bar{\vartheta}' \tag{59}$$

for $i = 3, 4$, where

$$\begin{aligned} F_{i1} = {} & (x_1\hat{c}_{i3} - x_1\hat{c}_{i2}S^{0223}_{1213} + \hat{c}_{i2}S^{1223}_{2213}) \\ & -\frac{(\hat{c}_{i4} - x_1\hat{c}_{i2}S^{0214}_{0412} + \hat{c}_{i2}S^{1214}_{0422})\big(s_{1,43} - s_{1,42}S^{0223}_{1213} + s_{0,42}S^{1223}_{2213}\big)}{s_{0,44} - s_{1,42}S^{0214}_{0412} + s_{0,42}S^{1214}_{0422}}, \\ F_{i2} = {} & x_2F_{i3}, \\ F_{i3} = {} & (\hat{c}_{i3} - x_1\hat{c}_{i2}S^{0213}_{0312} + \hat{c}_{i2}S^{1213}_{0322}) \\ & -\frac{(\hat{c}_{i4} - x_1\hat{c}_{i2}S^{0214}_{0412} + \hat{c}_{i2}S^{1214}_{0422})(s_{0,43} - s_{1,42}S^{0213}_{0312} + s_{0,42}S^{1213}_{0322})}{s_{0,44} - s_{1,42}S^{0214}_{0412} + s_{0,42}S^{1214}_{0422}}, \\ F_{i4} = {} & 2(x_1\hat{c}_{i4} - x_1\hat{c}_{i2}S^{0224}_{1214} + \hat{c}_{i2}S^{1224}_{1422}) \\ & -2\frac{(\hat{c}_{i4} - x_1\hat{c}_{i2}S^{0214}_{0412} + \hat{c}_{i2}S^{1214}_{0422})(s_{1,44} - s_{1,42}S^{0224}_{1214} + s_{0,42}S^{1224}_{1422})}{s_{0,44} - s_{1,42}S^{0214}_{0412} + s_{0,42}S^{1214}_{0422}}. \end{aligned}$$

Let

$$A_{ij}(x_3) := \int_{\Omega_{12}} F_{ij}(\cdot,\cdot,x_3)\,dx_1dx_2$$

and

$$K_{ij}(x_3) := \int_{\Omega_{12}} x_1F_{ij}(\cdot,\cdot,x_3)\,dx_1dx_2, \quad L_{ij}(x_3) := \int_{\Omega_{12}} x_2F_{ij}(\cdot,\cdot,x_3)\,dx_1dx_2.$$

Then, from (43), (44), and (59) we finally deduce the following system of equilibrium differential equations:

$$\left\{\begin{array}{l} (A_{33}\bar{\zeta}_3' - A_{31}\bar{\zeta}_1'' - A_{32}\bar{\zeta}_2'' + A_{34}\bar{\vartheta}')' = -p_3,\\ (K_{33}\bar{\zeta}_3' - K_{31}\bar{\zeta}_1'' - K_{32}\bar{\zeta}_2'' + K_{34}\bar{\vartheta}')'' = -p_1,\\ (L_{33}\bar{\zeta}_3' - L_{31}\bar{\zeta}_1'' - L_{32}\bar{\zeta}_2'' + L_{34}\bar{\vartheta}')'' = -p_2,\\ 2(K_{43}\bar{\zeta}_3' - K_{41}\bar{\zeta}_1'' - K_{42}\bar{\zeta}_2'' + K_{44}\bar{\vartheta}')' = -m, \end{array}\right. \tag{60}$$

where

$$\begin{aligned} p_1 &= \left(\int_{\Omega_{12}} x_1b_3\,dx_1dx_2\right)' + \int_{\Omega_{12}} b_1\,dx_1dx_2,\\ p_2 &= \left(\int_{\Omega_{12}} x_2b_3\,dx_1dx_2\right)' + \int_{\Omega_{12}} b_2\,dx_1dx_2,\\ p_3 &= \int_{\Omega_{12}} b_3\,dx_1dx_2. \end{aligned}$$

The system (60) should then be completed with the suitable boundary conditions.

7.1. The homogeneous beam. In the general inhomogeneous and anisotropic case, the torsional, flexional, and extensional problems are all coupled together in the equilibrium differential system (60). A partial decoupling occurs already in the homogeneous fully anisotropic case. Indeed, in this case, from (56), we have

$$s_{0;ij} = \hat{c}_{ij}, \qquad s_{1;ij} = 0, \qquad s_{2;ij} = \frac{a_1^3}{12}\hat{c}_{ij},$$

and therefore

$$S_{0312}^{0213} = S_{2213}^{1223} = S_{0412}^{0214} = S_{1422}^{1224} = 0, \qquad S_{1213}^{0223} = \frac{\hat{c}_{23}}{\hat{c}_{22}}, \qquad S_{0322}^{1213} = -\frac{\hat{c}_{23}}{\hat{c}_{22}}, \qquad S_{1422}^{1214} = \frac{\hat{c}_{24}}{\hat{c}_{22}},$$

which causes many of the coefficients of the system A_{ij}, K_{ij}, and L_{ij} to be zero. In this case, the system (60) simply rewrites as

$$\left\{\begin{array}{l} A_{33}\bar{\zeta}_3'' = -p_3,\\ (-K_{31}\bar{\zeta}_1'' + K_{34}\bar{\vartheta}')'' = -p_1,\\ -L_{32}\bar{\zeta}_2^{(iv)} = -p_2,\\ 2(-K_{41}\bar{\zeta}_1'' + K_{44}\bar{\vartheta}')' = -m, \end{array}\right. \tag{61}$$

where

$$A_{33} = a_1 a_2 \left[\frac{\hat{c}_{22}\hat{c}_{33} - \hat{c}_{23}^2}{\hat{c}_{22}} + \frac{(\hat{c}_{22}\hat{c}_{34} - \hat{c}_{23}\hat{c}_{24})^2}{\hat{c}_{22}(\hat{c}_{22}\hat{c}_{44} - \hat{c}_{24}^2)} \right],$$

$$K_{31} = \frac{a_1^3 a_2}{12} \frac{\hat{c}_{22}\hat{c}_{33} - \hat{c}_{23}^2}{\hat{c}_{22}}, \; K_{34} = \frac{a_1^3 a_2}{6} \frac{\hat{c}_{22}\hat{c}_{34} - \hat{c}_{23}\hat{c}_{24}}{\hat{c}_{22}},$$

$$K_{41} = \frac{a_1^3 a_2}{12} \frac{\hat{c}_{22}\hat{c}_{34} - \hat{c}_{23}\hat{c}_{24}}{\hat{c}_{22}}, \; K_{44} = \frac{a_1^3 a_2}{6} \frac{\hat{c}_{22}\hat{c}_{44} - \hat{c}_{24}^2}{\hat{c}_{22}},$$

$$L_{32} = \frac{a_1 a_2^3}{12} \left[\frac{\hat{c}_{22}\hat{c}_{33} - \hat{c}_{23}^2}{\hat{c}_{22}} + \frac{(\hat{c}_{22}\hat{c}_{34} - \hat{c}_{23}\hat{c}_{24})^2}{\hat{c}_{22}(\hat{c}_{22}\hat{c}_{44} - \hat{c}_{24}^2)} \right].$$

Thus for a fully anisotropic but homogeneous beam there is only coupling between twisting and bending in direction 1.

7.2. The homogeneous orthotropic/isotropic beam. When the material is orthotropic a complete decoupling occurs. Indeed, for orthotropic material we have that $c_{ki} = 0$ for $k = 1, 2, 3$ and $i = 4, 5, 6$, and $c_{45} = c_{46} = c_{56} = 0$. It then follows that $K_{34} = K_{41} = 0$ and the system (61) reduces to

$$\text{(62)} \qquad \begin{cases} A_{33}\bar{\zeta}_3'' = -p_3, \\ -K_{31}\bar{\zeta}_1^{(iv)} = -p_1, \\ -L_{32}\bar{\zeta}_2^{(iv)} = -p_2, \\ 2K_{44}\bar{\vartheta}'' = -m. \end{cases}$$

Finally, if the material is isotropic, that is, if it is orthotropic and $c_{12} = c_{23} = c_{13} =: \lambda$, $c_{44} = c_{55} = c_{66} =: \mu$, and $c_{11} = c_{22} = c_{33} = \lambda + 2\mu$, where λ and μ are the Lamé moduli, then we have that

$$A_{33} = a_1 a_2 E, \quad K_{31} = \frac{a_1^3 a_2}{12} E, \quad 2K_{44} = \frac{a_1^3 a_2}{3}\mu, \quad L_{32} = \frac{a_1 a_2^3}{12} E,$$

where $E := \mu\frac{3\lambda+2\mu}{\lambda+\mu}$ is the Young modulus of the material. Hence, in the isotropic and homogeneous case we recover the usual form of the differential system of equilibrium equations.

REFERENCES

[1] B. A. Auld, *Acoustic Fields and Waves in Solids*, Vol. 1, John Wiley and Sons, New York, 1973.

[2] R. Chandra, A. D. Stemple, and I. Chopra, *Thin-walled composite beams under bending, torsional, and extensional effects*, J. Aircraft, 27 (1990), pp. 619–626.

[3] L. Della Longa and A. Londero, *Thin Walled Beams with Residual Stress*, submitted.

[4] L. Freddi, A. Morassi, and R. Paroni, *Thin-walled beams: The case of the rectangular cross-section*, J. Elasticity, 76 (2004), pp. 45–66.

[5] H. Le Dret, *Problèmes Variationnels dans les Multi-Domaines*, Recherches en Mathématiques Appliquées 19, Masson, Paris, 1991.

[6] H. Le Dret, *Convergence of displacements and stresses in linearly elastic slender rods as the thickness goes to zero*, Asymptot. Anal., 10 (1995), pp. 367–402.

[7] R. Monneau, F. Murat, and A. Sili, *A Partial Korn's Inequality and Error Estimates for the 3d-1d Dimension Reduction in Anisotropic Heterogeneous Linearized Elasticity*, in preparation.

[8] F. Murat and A. Sili, *Anisotropic, Heterogeneous, Linearized Elasticity Problems in Thin Cylinders*, in preparation.

[9] L. Trabucho and J. M. Viano, *Mathematical modelling of rods*, in Handbook of Numerical Analysis, Vol. 4, North–Holland, Amsterdam, 1996, pp. 487–974.

[10] F. Treves, *Topological Vector Spaces, Distributions and Kernels*, Academic Press, New York, London, 1967.

[11] V. V. Volovoi and D. H. Hodges, *Theory of anisotropic thin-walled beams*, J. Appl. Mech., 67 (2000), pp. 453–459.

[12] V. V. Volovoi, D. H. Hodges, C. E. S. Cesnik, and B. Popescu, *Assessment of beam modeling methods for rotor blade applications*, Math. Comput. Modelling, 33 (2001), pp. 1099–1112.

[13] R. L. Wheeden and A. Zygmund, *Measure and Integral. An Introduction to Real Analysis*, Monogr. Textbooks Pure Appl. Math. 43, Marcel Dekker, New York, Basel, 1977.

SIAM J. MATH. ANAL.
Vol. 40, No. 5, pp. 1952–1978

CORRECTORS FOR THE HOMOGENIZATION OF A CLASS OF HYPERBOLIC EQUATIONS WITH IMPERFECT INTERFACES*

PATRIZIA DONATO†, LUISA FAELLA‡, AND SARA MONSURRÒ§

Abstract. We present here some corrector results for the homogenization of the wave equation in a two-component composite with ε-periodic connected inclusions which complete the homogenization results proved in [P. Donato, L. Faella, and S. Monsurrò, *J. Math. Pures Appl.*, 87 (2007), pp. 119–143] by the authors. On the interface separating the two components we prescribe a jump of the solution proportional to the conormal derivatives via a function of order ε^γ, with $-1 < \gamma \leq 1$. Due to different expressions of the energies of the limit problems, the cases $-1 < \gamma < 1$ and $\gamma = 1$ need to be treated separately. The second one, where a memory effect appears in the homogenized problem, is the most interesting. For this critical case, displaying lack of compactness, we in particulur establish the central upper semicontinuity type inequalities by splitting a related energy term into a compact part and a part vanishing in appropriate norms.

Key words. homogenization, correctors, hyperbolic equations

AMS subject classifications. 35B27, 35L05, 82B24

DOI. 10.1137/080712684

1. Introduction.

In this paper we prove some corrector results for the homogenization of a linear hyperbolic problem in a domain Ω of $\mathbb{R}^n$ made up of two components, a connected one $\Omega_{1\varepsilon}$ and a disconnected one $\Omega_{2\varepsilon}$, which is a union of ε-periodic connected inclusions of size ε. The conditions prescribed on the interface $\Gamma^\varepsilon := \partial\Omega_{2\varepsilon}$ between the two components are the continuity of the conormal derivatives and a jump of the solution proportional to the conormal derivatives via a function of order ε^γ. We suppose here that $-1 < \gamma \leq 1$. This work completes the corresponding homogenization results proved by the authors in [9].

Let A be a periodic, symmetric, bounded, and elliptic matrix field and h a bounded and periodic function. We consider the following problem:

$$\begin{cases} u_\varepsilon'' - \operatorname{div}(A^\varepsilon \nabla u_\varepsilon) = f_\varepsilon \quad \text{in } (\Omega_{1\varepsilon} \times \Omega_{2\varepsilon}) \times]0,T[, \\ [A^\varepsilon \nabla u_\varepsilon] \cdot n_{1\varepsilon} = 0 \quad \text{on } \Gamma^\varepsilon \times]0,T[, \\ A^\varepsilon \nabla u_{1\varepsilon} \cdot n_{1\varepsilon} = -\varepsilon^\gamma h^\varepsilon [u_\varepsilon] \quad \text{on } \Gamma^\varepsilon \times]0,T[, \\ u_\varepsilon = 0 \quad \text{on } \partial\Omega \times]0,T[, \\ u_\varepsilon(0) = U_\varepsilon^0 \quad \text{in } \Omega, \\ u_\varepsilon'(0) = U_\varepsilon^1 \quad \text{in } \Omega, \end{cases} \tag{1.1}$$

where $A^\varepsilon(x) := A(x/\varepsilon)$, $h^\varepsilon(x) := h(x/\varepsilon)$, $u_\varepsilon = (u_{1\varepsilon}, u_{2\varepsilon})$ is defined in $\Omega_{1\varepsilon} \times \Omega_{2\varepsilon}$, $[\,\cdot\,]$ denotes the jump trough Γ^ε, and $n_{i\varepsilon}$ is the unitary outward normal to $\Omega_{i\varepsilon}$, $i = 1, 2$.

*Received by the editors January 7, 2008; accepted for publication (in revised form) September 2, 2008; published electronically January 21, 2009.
http://www.siam.org/journals/sima/40-5/71268.html

†Laboratoire de Mathématiques Raphaël Salem, Université de Rouen, Tecnopôle du Madrillet, Avenue de l'Université, B.P. 12, 76801 Saint Etienne du Rouvray, France (Patrizia.Donato@univ-rouen.fr).

‡Dipartimento di Automazione, Elettromagnetismo, Ingegneria, dell'Informazione e Matematica Industriale, Università di Cassino, via G. Di Biasio n.43, 03043 Cassino (FR), Italy (l.faella@unicas.it).

§Dipartimento di Matematica ed Informatica, Università degi Studi di Salerno, via Ponte don Melillo, 84084, Fisciano (SA), Italy (smonsurro@unisa.it).

This problem models the wave propagation in a medium made up of two materials with very different coefficients of propagation, which gives rise to the jump in the boundary condition on the interface. For the physical model we refer to [4], where these kinds of interface conditions are derived.

The homogenization results proved in [9] (recalled in section 2) show two different asymptotic behaviors for $-1 < \gamma < 1$ and $\gamma = 1$, the last case being the most interesting one, since a memory effect appears at the limit (see also Remark 2.7).

When $-1 < \gamma < 1$, under suitable assumptions on the data (see Theorem 2.3), we have, as $\varepsilon \to 0$, the following convergences:

$$\begin{cases} P_1^\varepsilon u_{1\varepsilon} \rightharpoonup u_1 & \text{weakly* in } L^\infty\left(0,T;H_0^1(\Omega)\right), \\ P_1^\varepsilon u'_{1\varepsilon} \rightharpoonup u'_1 & \text{weakly* in } L^\infty\left(0,T;L^2(\Omega)\right), \end{cases}$$

while

$$\begin{cases} \widetilde{u_{2\varepsilon}} \rightharpoonup \theta_2 u_1 & \text{weakly* in } L^\infty\left(0,T;L^2(\Omega)\right), \\ \widetilde{u'_{2\varepsilon}} \rightharpoonup \theta_2 u'_1 & \text{weakly* in } L^\infty\left(0,T;L^2(\Omega)\right), \end{cases}$$

and

$$\begin{cases} A^\varepsilon \widetilde{\nabla u_{1\varepsilon}} \rightharpoonup A^0 \nabla u_1 & \text{weakly* in } L^\infty\left(0,T;\left[L^2(\Omega)\right]^n\right), \\ A^\varepsilon \widetilde{\nabla u_{2\varepsilon}} \rightharpoonup 0 & \text{weakly* in } L^\infty\left(0,T;\left[L^2(\Omega)\right]^n\right), \end{cases} \tag{1.2}$$

where P_1^ε is a suitable extension operator, θ_2 is the proportion of material occupying the inclusions, $\widetilde{\ }$ denotes the zero extension to the whole of Ω, and u_1 is the solution of the homogenized problem

$$\begin{cases} u_1'' - \operatorname{div}\left(A^0 \nabla u_1\right) = f_1 + f_2 & \text{in } \Omega\times]0,T[, \\ u_1 = 0 & \text{on } \partial\Omega\times]0,T[, \\ u_1(0) = U_0 & \text{in } \Omega, \\ u_1'(0) = U_1 & \text{in } \Omega. \end{cases}$$

The matrix field A^0 is the same constant positive definite matrix obtained by Cioranescu and Saint Jean Paulin (see [7] and also [6]) for the homogenization of the elliptic problem in the perforated domain $\Omega_{1\varepsilon}$, with a Neumann condition on the boundary of the holes.

All of these convergences are weak, but sufficient, to obtain the homogenized problem. At this point, the question is how to improve them, in particular give more information on the gradient and on the first time derivative of the solution. Such results are known in standard homogenization as corrector results. The aim of this paper is precisely to answer the above questions.

The corrector result given in Theorem 2.4 states that, under some additional assumptions,

$$\begin{cases} \widetilde{u'_{1\varepsilon}} + \widetilde{u'_{2\varepsilon}} \to u'_1 \text{ strongly in } C^0(0,T;L^2(\Omega)), \\ \lim\limits_{\varepsilon\to 0} \|\nabla u_{1\varepsilon} - C^\varepsilon \nabla u_1\|_{C^0(0,T;[L^1(\Omega_{1\varepsilon})]^n)} = 0, \\ \lim\limits_{\varepsilon\to 0} \|\nabla u_{2\varepsilon}\|_{C^0(0,T;[L^2(\Omega_{2\varepsilon})]^n)} = 0, \end{cases}$$

where C^ε is the same corrector matrix of the problem studied in [7].

The additional assumptions concern the data. We assume strong convergence for both f_ε and U_ε^1, and we suppose that U_ε^0 solves the particular elliptic equation given by (2.10). These assumptions are needed for the convergence of the energy of problem (1.1) to that of the homogenized one. This convergence is necessary to prove the corrector results, as already evidenced in the classical homogenization of the wave equation (see [2]) or in a perforated domain with a Neumann condition on the boundaries of the holes (see [18]).

If $\gamma = 1$, the situation is more complicated since the limit problem consists in a p.d.e. coupled with an o.d.e., as proved in [9]. One still has the convergences

$$\begin{cases} P_1^\varepsilon u_{1\varepsilon} \rightharpoonup u_1 & \text{weakly* in } L^\infty\left(0,T;H_0^1(\Omega)\right), \\ P_1^\varepsilon u'_{1\varepsilon} \rightharpoonup u'_1 & \text{weakly* in } L^\infty\left(0,T;L^2(\Omega)\right), \end{cases}$$

and

$$\begin{cases} \widetilde{u_{2\varepsilon}} \rightharpoonup u_2 & \text{weakly* in } L^\infty\left(0,T;L^2(\Omega)\right), \\ \widetilde{u'_{2\varepsilon}} \rightharpoonup u'_2 & \text{weakly* in } L^\infty\left(0,T;L^2(\Omega)\right), \end{cases}$$

but u_2 is not $\theta_2 u'_1$ anymore. In this case (see Theorem 2.6), (u_1, u_2) is the unique solution of the coupled problem

$$\begin{cases} \theta_1 u_1'' - \operatorname{div}\left(A^0 \nabla u_1\right) + c_h(\theta_2 u_1 - u_2) = f_1 & \text{in } \Omega\times]0,T[, \\ u_2'' - c_h(\theta_2 u_1 - u_2) = f_2 & \text{in } \Omega\times]0,T[, \\ u_1 = 0 & \text{on } \partial\Omega\times]0,T[, \\ u_1(x,0) = U^0, \quad u_2(x,0) = \theta_2 U^0 & \text{in } \Omega, \\ u'_1(x,0) = U^1, \quad u'_2(x,0) = \theta_2 U^1 & \text{in } \Omega, \end{cases}$$

where $\theta_1 = 1 - \theta_2$ and c_h is a constant depending on the function h. Furthermore, convergences (1.2) still hold. Let us mention that in [9, Remark 2.11] the limit function u_2 is explicitly computed in terms of u_1 and it is shown that u_1 is the solution of a linear hyperbolic problem with a linear memory effect.

Under the same assumptions as in the previous case, the corrector result for $\gamma = 1$ (Theorem 2.8) states the convergences

$$\begin{cases} \lim\limits_{\varepsilon\to 0} \|u'_{1\varepsilon} - u'_1\|_{C^0(0,T;L^2(\Omega_{1\varepsilon}))} = 0, \\ \lim\limits_{\varepsilon\to 0} \|u'_{2\varepsilon} - \theta_2^{-1} u'_2\|_{C^0(0,T;L^2(\Omega_{2\varepsilon}))} = 0, \\ \lim\limits_{\varepsilon\to 0} \|\nabla u_{1\varepsilon} - C^\varepsilon \nabla u_1\|_{C^0(0,T;[L^1(\Omega_{1\varepsilon})]^n)} = 0, \\ \lim\limits_{\varepsilon\to 0} \|\nabla u_{2\varepsilon}\|_{C^0(0,T;[L^2(\Omega_{2\varepsilon})]^n)} = 0. \end{cases}$$

The corrector results are proved in section 5 for $-1 < \gamma < 1$ and in section 6 for $\gamma = 1$.

Let us point out the main difficulties of our situation due to the presence of the interface. The first concerns the study of the convergence of the energy associated to problem (1.1). This is in general straightforward, but here the expression of the energy of problem (1.1) is more complicated than usual, since it contains a boundary term. Moreover, for $\gamma = 1$, the energy of the homogenized coupled system contains also a zero order term which is a linear combination of u_1 and u_2. This can be seen comparing formulas (4.5) (case $-1 < \gamma < 1$) and (4.8) (case $\gamma = 1$). We can prove in section 4 (Theorems 4.3 and 4.4) that in both cases the energy converges to those of the homogenized problems.

The second and more difficult point consists in proving two upper semicontinuity type inequalities given for the two cases in Propositions 5.1 and 6.1, respectively. They require specific arguments and are crucial, since they allow us to end, by a density argument, the proof of the corrector result. Let us mention that usually in the literature one obtains a convergence, which gives rise to an equality for the limit (see Remark 5.2), but here we can only obtain an inequality on an upper limit. Nevertheless, this is enough to conclude.

Let us briefly describe Proposition 6.1 concerning the case $\gamma = 1$, which is the most interesting one. It states that if we set, for $\Phi, \Psi \in C^\infty([0,T], \mathcal{D}(\Omega))$,

$$\hat{X}_\varepsilon = \frac{1}{2}\Big[\int_{\Omega_{1\varepsilon}} |u'_{1\varepsilon} - \Phi'|^2 \, dx + \int_{\Omega_{2\varepsilon}} |u'_{2\varepsilon} - \Psi'|^2 \, dx + \int_{\Omega_{1\varepsilon}} A^\varepsilon \left(\nabla u_{1\varepsilon} - C^\varepsilon \nabla \Phi\right)\left(\nabla u_{1\varepsilon} - C^\varepsilon \nabla \Phi\right) \, dx + \int_{\Omega_{2\varepsilon}} A^\varepsilon \nabla u_{2\varepsilon} \nabla u_{2\varepsilon} \, dx \Big],$$

then

$$\limsup_{\varepsilon \to 0} \left\| \hat{X}_\varepsilon \right\|_{C^0([0,T])} \le \left\| \hat{X} \right\|_{C^0([0,T])}, \tag{1.3}$$

where

$$\hat{X} = \frac{1}{2}\left[\theta_1 \| u'_1 - \Phi' \|^2_{L^2(\Omega)} + \theta_2^{-1} \| u'_2 - \theta_2 \Psi' \|^2_{L^2(\Omega)} + \int_\Omega A^0 (\nabla u_1 - \nabla \Phi)(\nabla u_1 - \nabla \Phi) \, dx \right].$$

The main difficulty when proving this result is due to the boundary term in the energy associated to problem (1.1). Indeed, we cannot have estimates for the time derivative of $\hat{X}_\varepsilon$ so that we cannot derive any compactness of $\hat{X}_\varepsilon$ in $C^0([0,T])$. To overcome this difficulty, in Step 4 of the proof of Proposition 6.1 we decompose $\hat{X}_\varepsilon$ as a sum of two terms and we show that the first one is compact and the second one goes to zero in $C^0([0,T])$ (see also Remarks 6.2 and 6.3 for details). This allows us to prove (1.3).

The correctors for the corresponding elliptic case have been studied by the first author in [8]. The first homogenization result for these kinds of boundary conditions was done in the elliptic case, for some values of the parameter γ, by Auriault and Ene [1] by the multiple scales method. We refer to Lipton [15] for the study of the limit problem when $\gamma = 0$, to S. Monsurrò [17] for the case $\gamma \le -1$, and to [10] for the case $\gamma > -1$. For similar elliptic homogenization problems we refer also to [3], [12], [13], [16], [19].

2. Formulation of the problem and main results. Along this paper, Ω will denote an open bounded subset of $\mathbb{R}^n$ and $\{\varepsilon\}$ a sequence of positive real numbers converging to zero.

Let $Y =]0, l_1[\times \cdots \times]0, l_n[$ be a reference cell in $\mathbb{R}^n$ and Y_1 and Y_2 be two nonempty open and disjoint subsets of Y such that

$$Y = Y_1 \cup \overline{Y_2},$$

with Y_1 connected and $\Gamma := \partial Y_2$ of class C^2.

For any $k \in Z^n$ we denote

$$Y_i^k := \; k_l + Y_i, \qquad \Gamma_k := \; k_l + \Gamma,$$

where $k_l = (k_1 l_1, \ldots, k_n l_n)$ and $i = 1, 2$, and we suppose that

$$\partial\Omega \cap \left(\cup_{k \in Z^n} (\varepsilon \Gamma_k) \right) := \emptyset. \tag{2.1}$$

We introduce then, for any ε, the two components of Ω and the interface, respectively, by

$$\Omega_{i\varepsilon} := \Omega \cap \left\{\cup_{k\in K_\varepsilon} \varepsilon Y_i^k\right\},\ i=1,2, \quad \text{and} \quad \Gamma^\varepsilon := \partial\Omega_{2\varepsilon},$$

where K_ε is the set of the n-tuples such that $\varepsilon\Gamma_k$ is included in Ω, namely,

$$K_\varepsilon := \{k \in Z^n |\ \varepsilon\Gamma_k \cap \Omega \neq \emptyset\},$$

so that $\partial\Omega \cap \Gamma^\varepsilon = \emptyset$.

Due to (2.1), the set $\Omega_{1\varepsilon}$ is connected while $\Omega_{2\varepsilon}$ is a union of $c\varepsilon^{-n}$ disjoint translated sets of εY_2, c being a constant independent of ε.

In what follows, we will denote by χ_ω the characteristic function of any open set $\omega \subset \mathbb{R}^n$. We know that (see, for instance, [6])

$$\chi_{\Omega_{i\varepsilon}} \rightharpoonup \theta_i := \frac{|Y_i|}{|Y|} \quad \text{weakly in } L^2(\Omega), \qquad i=1,2. \tag{2.2}$$

We define the normed space V^ε by

$$V^\varepsilon := \{v_1 \in H^1(\Omega_{1\varepsilon}) |\ v_1 = 0\}$$

endowed with the norm

$$\|v_1\|_{V^\varepsilon} := \|\nabla v_1\|_{L^2(\Omega_{1\varepsilon})}.$$

Remark 2.1. Since we do not assume any regularity on $\partial\Omega$, the condition on $\partial\Omega$ in the definition of V^ε has to be understood in a density sense. More precisely, V^ε is the closure with respect to the $H^1(\Omega_{1\varepsilon})$-norm of the set of the functions of $C^\infty(\Omega_{1\varepsilon})$ with a compact support contained in Ω. This make sense because of (2.1).

As in [9], we also define, for every $\gamma \in \mathbb{R}$,

$$H^\varepsilon_\gamma := \left\{v = (v_1, v_2) \mid v_1 \in V^\varepsilon \quad \text{and} \quad v_2 \in H^1(\Omega_{2\varepsilon})\right\}$$

equipped with the norm

$$\|v\|^2_{H^\varepsilon_\gamma} := \|\nabla v_1\|^2_{L^2(\Omega_{1\varepsilon})} + \|\nabla v_2\|^2_{L^2(\Omega_{2\varepsilon})} + \varepsilon^\gamma \|v_1 - v_2\|^2_{L^2(\Gamma^\varepsilon)}.$$

Obviously, if $0 < \varepsilon < 1$ and $\gamma_1 \leq \gamma_2$, then

$$\|v\|_{H^\varepsilon_{\gamma_2}} \leq \|v\|_{H^\varepsilon_{\gamma_1}}.$$

Moreover, as shown in [17, Lemmas 2.7 and 2.8 and their proofs], for every fixed ε the norms of H^ε_γ and $V^\varepsilon \times H^1(\Omega_{2\varepsilon})$ are equivalent.

To introduce the coefficient matrix, we define, for $\alpha, \beta \in \mathbb{R}$ with $0 < \alpha < \beta$, the set $M(\alpha, \beta, Y)$ of the $n \times n$ Y-periodic matrix-valued functions in $L^\infty(Y)$ such that

$$(A(x)\lambda, \lambda) \geq \alpha|\lambda|^2, \quad |A(x)\lambda| \leq \beta\lambda$$

for any $\lambda \in \mathbb{R}^n$ and a.e. in Y.

We assume that

$$A \in M(\alpha, \beta, Y), \quad A \text{ symmetric}, \tag{2.3}$$

and we set, for any $\varepsilon > 0$,

$$A^\varepsilon(x) := A(x/\varepsilon). \tag{2.4}$$

Moreover, we consider a Y-periodic function h such that

$$h \in L^\infty(\Gamma) \text{ and } \exists\, h_0 \in \mathbb{R} \text{ such that } 0 < h_0 < h(y) \text{ a.e. in } \Gamma$$

and set

$$h^\varepsilon(x) := h\left(\frac{x}{\varepsilon}\right). \tag{2.5}$$

Throughout this work, we suppose $T > 0$ be given and $-1 < \gamma \leq 1$. We consider the following problem:

$$\begin{cases} u''_{1\varepsilon} - \operatorname{div}(A^\varepsilon \nabla u_{1\varepsilon}) = f_{1\varepsilon} \quad \text{in } \Omega_{1\varepsilon} \times]0,T[, \\ u''_{2\varepsilon} - \operatorname{div}(A^\varepsilon \nabla u_{2\varepsilon}) = f_{2\varepsilon} \quad \text{in } \Omega_{2\varepsilon} \times]0,T[, \\ A^\varepsilon \nabla u_{1\varepsilon} \cdot n_{1\varepsilon} = -A^\varepsilon \nabla u_{2\varepsilon} \cdot n_{2\varepsilon} \quad \text{on } \Gamma^\varepsilon \times]0,T[, \\ A^\varepsilon \nabla u_{1\varepsilon} \cdot n_{1\varepsilon} = -\varepsilon^\gamma h^\varepsilon (u_{1\varepsilon} - u_{2\varepsilon}) \quad \text{on } \Gamma^\varepsilon \times]0,T[, \\ u_{1\varepsilon} = 0 \quad \text{on } \partial\Omega \times]0,T[, \\ u_{1\varepsilon}(0) = U^0_{1\varepsilon} \quad \text{in } \Omega_{1\varepsilon}, \qquad u_{2\varepsilon}(0) = U^0_{2\varepsilon} \quad \text{in } \Omega_{2\varepsilon}, \\ u'_{1\varepsilon}(0) = U^1_{\varepsilon\,|\Omega_{1\varepsilon}} \quad \text{in } \Omega_{1\varepsilon}, \qquad u'_{2\varepsilon}(0) = U^1_{\varepsilon\,|\Omega_{2\varepsilon}} \quad \text{in } \Omega_{2\varepsilon}, \end{cases} \tag{2.6}$$

where $n_{i\varepsilon}$ is the unitary outward normal to $\Omega_{i\varepsilon}$, $i = 1, 2$, and

$$\begin{cases} \text{(i) } f_\varepsilon := (f_{1\varepsilon|\Omega_{1\varepsilon}}, f_{2\varepsilon|\Omega_{2\varepsilon}}) \text{ with } f_{i\varepsilon} \in L^2\left(0,T; L^2(\Omega)\right),\ i = 1,2, \\ \text{(ii) } U^0_\varepsilon := \left(U^0_{1\varepsilon}, U^0_{2\varepsilon}\right) \in V^\varepsilon \times H^1(\Omega_{2\varepsilon}), \\ \text{(iii) } U^1_\varepsilon \in L^2(\Omega). \end{cases} \tag{2.7}$$

Its variational (weak) formulation is

$$\begin{cases} \text{find } u_\varepsilon = (u_{1\varepsilon}, u_{2\varepsilon}) \text{ in } L^2(0,T;\ V^\varepsilon) \times L^2(0,T;\ H^1(\Omega_{2\varepsilon})) \\ \text{such that } u'_{1\varepsilon} \in L^2(0,T; L^2(\Omega_{1\varepsilon})),\ u'_{2\varepsilon} \in L^2(0,T; L^2(\Omega_{2\varepsilon})) \text{ and} \\ \langle u''_{1\varepsilon}, v_1 \rangle_{(V^\varepsilon)',V^\varepsilon} + \langle u''_{2\varepsilon}, v_2 \rangle_{(H^1(\Omega_{2\varepsilon}))',H^1(\Omega_{2\varepsilon})} \\ + \displaystyle\int_{\Omega_{1\varepsilon}} A^\varepsilon \nabla u_{1\varepsilon} \cdot \nabla v_1 \, dx + \int_{\Omega_{2\varepsilon}} A^\varepsilon \nabla u_{2\varepsilon} \cdot \nabla v_2 \, dx \\ + \varepsilon^\gamma \displaystyle\int_{\Gamma^\varepsilon} h^\varepsilon (u_{1\varepsilon} - u_{2\varepsilon})(v_1 - v_2) \, d\sigma_x = \int_{\Omega_{1\varepsilon}} f_{1\varepsilon} v_1 \, dx + \int_{\Omega_{2\varepsilon}} f_{2\varepsilon} v_2 \, dx \\ \text{for every } (v_1, v_2) \in V^\varepsilon \times H^1(\Omega_{2\varepsilon}) \text{ in } \mathcal{D}'(0,T), \\ u_{1\varepsilon}(0) = U^0_{1\varepsilon} \quad \text{in } \Omega_{1\varepsilon}, \qquad u_{2\varepsilon}(0) = U^0_{2\varepsilon} \quad \text{in } \Omega_{2\varepsilon}, \\ u'_{1\varepsilon}(0) = U^1_{\varepsilon\,|\Omega_{1\varepsilon}} \quad \text{in } \Omega_{1\varepsilon}, \qquad u'_{2\varepsilon}(0) = U^1_{\varepsilon\,|\Omega_{2\varepsilon}} \quad \text{in } \Omega_{2\varepsilon}. \end{cases} \tag{2.8}$$

The homogenization of this problem has been studied in [9] under more general assumptions than (2.7), obtaining different limit problems according to different values of γ. We prove here some corrector results for the two cases $-1 < \gamma < 1$ and $\gamma = 1$.

It is well known that (see, for instance, [2] and [18]) corrector results for the wave equation need stronger assumptions than that of the convergence results. In particular, we suppose that

$$\begin{cases} \text{(i) } (f_{1\varepsilon}, f_{2\varepsilon}) \to \left(\theta_1^{-1} f_1, \theta_2^{-1} f_2\right) \text{ strongly in } \left[L^2(0,T;\ L^2(\Omega))\right]^2, \\ \text{(ii) } U^1_\varepsilon \to U^1 \text{ strongly in } L^2(\Omega) \end{cases} \tag{2.9}$$

and, concerning the initial condition U^0_ε, we assume that it is the unique solution of the problem

$$\begin{cases} -\mathrm{div}\left(A^\varepsilon \nabla U^0_{1\varepsilon}\right) = Q_1^{\varepsilon *}\left(-\mathrm{div}\ \left(A^0 \nabla U^0\right)\right) & \text{in } \Omega_{1\varepsilon}, \\ -\mathrm{div}\left(A^\varepsilon \nabla U^0_{2\varepsilon}\right) = 0 \quad \text{in } \Omega_{2\varepsilon}, \\ A^\varepsilon \nabla U^0_{1\varepsilon} \cdot n_{1\varepsilon} = -A^\varepsilon \nabla U^0_{2\varepsilon} \cdot n_{2\varepsilon} \quad \text{on } \Gamma^\varepsilon, \\ A^\varepsilon \nabla U^0_{1\varepsilon} \cdot n_{1\varepsilon} = -\varepsilon^\gamma h^\varepsilon \left(U^0_{1\varepsilon} - U^0_{2\varepsilon}\right) \quad \text{on } \Gamma^\varepsilon, \\ U^0_{1\varepsilon} = 0 \quad \text{on } \partial\Omega, \end{cases} \tag{2.10}$$

where U^0 is given in $H^1_0(\Omega)$ and $Q_1^{\varepsilon *}$ is the adjoint of a suitable extension operator Q_1^ε, defined in section 3.

Remark 2.2. The homogenization result proved in [8] for the elliptic case (recalled in section 3 for the reader's convenience) applies to (2.10) and gives (see the proof of Lemma 4.2 for details) the following convergences:

$$\begin{cases} \text{(i) } Q_1^\varepsilon U^0_{1\varepsilon} \rightharpoonup U^0 \quad \text{weakly in } H^1_0(\Omega), \\ \text{(ii) } \widetilde{A^\varepsilon \nabla U^0_{1\varepsilon}} \rightharpoonup A^0 \nabla U^0 \quad \text{weakly in } \left[L^2(\Omega)\right]^n, \\ \text{(iii) } \widetilde{A^\varepsilon \nabla U^0_{2\varepsilon}} \rightharpoonup 0 \quad \text{weakly in } \left[L^2(\Omega)\right]^n, \\ \text{(iv) } \left\|U^0_\varepsilon\right\|_{H^\varepsilon_\gamma} \le C \text{ with } C \text{ independent of } \varepsilon, \\ \text{(v) } \widetilde{U^0_{2\varepsilon}} \rightharpoonup \theta_2 U^0 \quad \text{weakly in } H^1_0(\Omega), \end{cases} \tag{2.11}$$

where $\widetilde{\ }$ denotes the zero extension to the whole of Ω and the homogenized matrix A^0 is defined by

$$A^0 \lambda = \frac{1}{|Y|} \int_{Y_1} A \nabla w_\lambda \, dy \tag{2.12}$$

with $w_\lambda \in H^1(Y_1)$ solution, for any $\lambda \in \mathbb{R}^n$, of

$$\begin{cases} -\mathrm{div}\ \ (A \nabla w_\lambda) = 0 \quad \text{in } Y_1, \\ (A \nabla w_\lambda) \cdot n_1 = 0 \quad \text{on } \Gamma, \\ w_\lambda - \lambda \cdot y \qquad Y\text{-periodic}, \\ \dfrac{1}{|Y_1|} \displaystyle\int_{Y_1} (w_\lambda - \lambda \cdot y)\, dy = 0. \end{cases} \tag{2.13}$$

Moreover, from (2.9) and (2.11)(i) and (v) one has

$$\begin{cases} \text{(i) } (\chi_{\Omega_{1\varepsilon}} f_{1\varepsilon}, \chi_{\Omega_{2\varepsilon}} f_{2\varepsilon}) \rightharpoonup (f_1, f_2) \text{ weakly in } \left[L^2(0,T;\ L^2(\Omega))\right]^2, \\ \text{(ii) } \left(\chi_{\Omega_{1\varepsilon}} U^1_\varepsilon, \chi_{\Omega_{2\varepsilon}} U^1_\varepsilon\right) \rightharpoonup \left(\theta_1 U^1, \theta_2 U^1\right) \text{ weakly in } L^2(\Omega) \times L^2(\Omega), \\ \text{(iii) } \left(\widetilde{U^0_{1\varepsilon}}, \widetilde{U^0_{2\varepsilon}}\right) \rightharpoonup \left(\theta_1 U^0, \theta_2 U^0\right) \quad \text{weakly in } L^2(\Omega) \times L^2(\Omega). \end{cases}$$

Let us describe first the case $-1 < \gamma < 1$, for which the homogenization result proved in [9] applies and, under the above assumptions, reads as follows.

THEOREM 2.3 (see [9]). *For $-1 < \gamma < 1$, let A^ε and h^ε be defined by* (2.4) *and* (2.5), *respectively, suppose that* (2.7), (2.9), *and* (2.10) *hold, and let u_ε be the solution of problem* (2.6). *Then, there exist a constant $C > 0$ (independent of ε) and an*

extension operator $P_1^\varepsilon \in \mathcal{L}(L^\infty(0,T;H^k(\Omega_{1\varepsilon}));L^\infty(0,T;H^k(\Omega)))$, *for* $k=1,2$, *such that*

$$\begin{cases} \text{(i) } P_1^\varepsilon u_{1\varepsilon} \rightharpoonup u_1 \quad \textit{weakly* in } L^\infty\left(0,T;H_0^1(\Omega)\right), \\ \text{(ii) } \widetilde{u_{2\varepsilon}} \rightharpoonup u_2 := \theta_2 u_1 \quad \textit{weakly* in } L^\infty\left(0,T;L^2(\Omega)\right), \\ \text{(iii) } \varepsilon^{\gamma/2}\|u_{1\varepsilon}-u_{2\varepsilon}\|_{L^\infty(0,T;L^2(\Gamma^\varepsilon))} < C, \end{cases} \tag{2.14}$$

$$\begin{cases} \text{(i) } P_1^\varepsilon u'_{1\varepsilon} \rightharpoonup u'_1 \quad \textit{weakly* in } L^\infty\left(0,T;L^2(\Omega)\right), \\ \text{(ii) } \widetilde{u'_{2\varepsilon}} \rightharpoonup u'_2 = \theta_2 u'_1 \quad \textit{weakly* in } L^\infty\left(0,T;L^2(\Omega)\right), \end{cases} \tag{2.15}$$

$$\begin{cases} \text{(i) } A^\varepsilon \widetilde{\nabla u_{1\varepsilon}} \rightharpoonup A^0\nabla u_1 \quad \textit{weakly* in } L^\infty\left(0,T;\left[L^2(\Omega)\right]^n\right), \\ \text{(ii) } A^\varepsilon \widetilde{\nabla u_{2\varepsilon}} \rightharpoonup 0 \quad \textit{weakly* in } L^\infty\left(0,T;\left[L^2(\Omega)\right]^n\right), \end{cases} \tag{2.16}$$

where θ_2 *is given by* (2.2) *and* u_1 *is the unique solution in* $L^2(0,T;H_0^1(\Omega))$, *with* u'_1 *in* $L^2(0,T;L^2(\Omega))$, *of the problem*

$$\begin{cases} u''_1 - \operatorname{div}\left(A^0\nabla u_1\right) = f_1+f_2 & \textit{in } \Omega\times]0,T[, \\ u_1 = 0 & \textit{on } \partial\Omega\times]0,T[, \\ u_1(0) = U^0 & \textit{in } \Omega, \\ u'_1(0) = U^1 & \textit{in } \Omega, \end{cases} \tag{2.17}$$

with A^0 *defined in* (2.12).

Let $(e_j)_{j=1,\dots,n}$ be the canonical basis and $w_j \in H^1(Y_1)$ be the solution of problem (2.13), written for $\lambda = e_j$, for $j=1,\dots,n$. The corrector matrix C^ε is defined, for any ε, by

$$C^\varepsilon(x) = \widetilde{C}\left(\frac{x}{\varepsilon}\right), \tag{2.18}$$

where

$$C_{ij}(y) := \frac{\partial w_j}{\partial y_i}(y), \quad \text{for } i,j=1,\dots,n,$$

and $\widetilde{\ }$ denotes the zero extension to the whole reference cell Y. Let us mention that, as in the elliptic case (see [8]), C^ε is the same corrector obtained in the case of a periodic perforated domain with an homogeneous Neumann condition on the holes studied in [7]. Here, due to the symmetry of A, the matrix C^ε is symmetric too.

Our first corrector result is as follows.

THEOREM 2.4 (correctors for $-1<\gamma<1$). *Let* A^ε *and* h^ε *be defined by* (2.4) *and* (2.5), *respectively, and suppose that* (2.7), (2.9), *and* (2.10) *hold. If* u_ε *is the solution of problem* (2.6) *with* $-1<\gamma<1$ *and* u_1 *is the solution of problem* (2.17), *then*

$$\begin{cases} \text{(i) } \widetilde{u'_{1\varepsilon}} + \widetilde{u'_{2\varepsilon}} \to u'_1 \quad \textit{strongly in } C^0\left(0,T;L^2(\Omega)\right), \\ \text{(ii) } \lim\limits_{\varepsilon\to 0}\|\nabla u_{1\varepsilon} - C^\varepsilon\nabla u_1\|_{C^0(0,T;[L^1(\Omega_{1\varepsilon})]^n)} = 0, \\ \text{(iii) } \lim\limits_{\varepsilon\to 0}\|\nabla u_{2\varepsilon}\|_{C^0(0,T;[L^2(\Omega_{2\varepsilon})]^n)} = 0. \end{cases} \tag{2.19}$$

This theorem will be proved in section 5.

Remark 2.5. Convergence (2.19)(i) is equivalent to

$$\lim_{\varepsilon\to 0}\|u'_{i\varepsilon} - u'_1\|_{C^0(0,T;L^2(\Omega_{i\varepsilon}))} = 0, \ i=1,2,$$

since

$$\int_\Omega |\widetilde{u'_{1\varepsilon}} + \widetilde{u'_{2\varepsilon}} - u'_1|^2\,dx = \int_{\Omega_{1\varepsilon}} |u'_{1\varepsilon} - u'_1|^2\,dx + \int_{\Omega_{2\varepsilon}} |u'_{2\varepsilon} - u'_1|^2\,dx,$$

$\Omega_{1\varepsilon}$ and $\Omega_{2\varepsilon}$ being disjoint sets.

We present now the more complicated case $\gamma = 1$, which is also the more interesting, since the limit function u_2 is not $\theta_2 u_1$ anymore and the limit problem satisfied by (u_1, u_2) is a coupled system consisting in a p.d.e. and an o.d.e. Indeed, we have the following theorem.

THEOREM 2.6 (see [9]). *For $\gamma = 1$, let A^ε and h^ε be defined by* (2.4) *and* (2.5), *respectively, suppose that* (2.7), (2.10), *and* (2.9) *hold, and let u_ε be the solution of the problem* (2.6). *Then, there exist a constant $C > 0$ (independent of ε) and an extension operator $P_1^\varepsilon \in \mathcal{L}(L^\infty(0,T;H^k(\Omega_{1\varepsilon}));L^\infty(0,T;H^k(\Omega)))$, for $k = 1, 2$, such that*

$$\begin{cases} \text{(i) } P_1^\varepsilon u_{1\varepsilon} \rightharpoonup u_1 \quad \textit{weakly* in } L^\infty\left(0,T;H_0^1(\Omega)\right), \\ \text{(ii) } \widetilde{u_{2\varepsilon}} \rightharpoonup u_2 \quad \textit{weakly* in } L^\infty\left(0,T;L^2(\Omega)\right), \\ \text{(iii) } \varepsilon^{1/2}\|u_{1\varepsilon} - u_{2\varepsilon}\|_{L^\infty(0,T;L^2(\Gamma^\varepsilon))} < C, \end{cases} \tag{2.20}$$

$$\begin{cases} \text{(i) } P_1^\varepsilon u'_{1\varepsilon} \rightharpoonup u'_1 \quad \textit{weakly* in } L^\infty\left(0,T;L^2(\Omega)\right), \\ \text{(ii) } \widetilde{u'_{2\varepsilon}} \rightharpoonup u'_2 \quad \textit{weakly* in } L^\infty\left(0,T;L^2(\Omega)\right), \end{cases} \tag{2.21}$$

$$\begin{cases} \text{(i) } A^\varepsilon \widetilde{\nabla u_{1\varepsilon}} \rightharpoonup A^0 \nabla u_1 \quad \textit{weakly* in } L^\infty\left(0,T;\left[L^2(\Omega)\right]^n\right), \\ \text{(ii) } A^\varepsilon \widetilde{\nabla u_{2\varepsilon}} \rightharpoonup 0 \quad \textit{weakly* in } L^\infty\left(0,T;\left[L^2(\Omega)\right]^n\right), \end{cases} \tag{2.22}$$

where the couple (u_1, u_2) is the unique solution $L^2(0,T;H_0^1(\Omega)) \times L^2(0,T;L^2(\Omega))$, with (u'_1, u'_2) in $L^2(0,T;L^2(\Omega)) \times L^2(0,T;L^2(\Omega))$, of the problem

$$\begin{cases} \theta_1 u''_1 - \text{div}\,\left(A^0 \nabla u_1\right) + c_h(\theta_2 u_1 - u_2) = f_1 & \textit{in } \Omega\times]0,T[, \\ u''_2 - c_h(\theta_2 u_1 - u_2) = f_2 & \textit{in }]0,T[\textit{ for a.e. } x \in \Omega, \\ u_1 = 0 & \textit{on } \partial\Omega\times]0,T[, \\ u_1(0) = U^0, \quad u_2(0) = \theta_2 U^0 & \textit{in } \Omega, \\ u'_1(0) = U^1, \quad u'_2(0) = \theta_2 U^1 & \textit{in } \Omega, \end{cases} \tag{2.23}$$

where θ_i, for $i = 1, 2$, are given by (2.2), *$c_h = \frac{1}{|Y_2|}\int_\Gamma h(y)d\sigma_y > 0$, and the homogenized matrix A^0 is defined by* (2.12).

Remark 2.7. (i) In [9], it is shown that u_1 is the solution of the equation

$$\theta_1 u''_1 - \text{div}\,\left(A^0 \nabla u_1\right) + c_h\theta_2 u_1 - c_h^2\theta_2 \int_0^t \mathcal{K}(t,s)u_1(s)\,ds = F \text{ in } \Omega\times]0,T[,$$

containing a periodic memory kernel $\mathcal{K}$; moreover, the limit solution u_2 and the function F are explicitly computed. In this paper we will need only to use the limit problem under the form (2.23), which is well adapted to the study of the energies.

(ii) Observe that convergences (2.14)–(2.15) and (2.20)–(2.21) imply that for $-1 < \gamma \le 1$

$$\chi_{\Omega_{1\varepsilon}} P_1^\varepsilon u'_{1\varepsilon} \rightharpoonup \theta_1 u'_1 \quad \text{weakly* in } L^\infty\left(0,T;L^2(\Omega)\right).$$

Moreover, from classical compactness results (see [14])

$$P_1^\varepsilon u_{1\varepsilon} \to u_1 \quad \text{strongly in } C^0\left(0,T;L^2(\Omega)\right).$$

Let us state the corrector result associated to the case $\gamma = 1$.

THEOREM 2.8 (correctors for $\gamma = 1$). *Let A^ε and h^ε be defined by* (2.4) *and* (2.5), *respectively, and suppose that* (2.7), (2.9), *and* (2.10) *hold. If u_ε is the solution of problem* (2.6) *with $\gamma = 1$ and (u_1, u_2) is the solution of problem* (2.23), *then*

$$\begin{cases} \text{(i) } \lim\limits_{\varepsilon\to 0} \|u'_{1\varepsilon} - u'_1\|_{C^0(0,T;L^2(\Omega_{1\varepsilon}))} = 0, \\ \text{(ii) } \lim\limits_{\varepsilon\to 0} \|u'_{2\varepsilon} - \theta_2^{-1}u'_2\|_{C^0(0,T;L^2(\Omega_{2\varepsilon}))} = 0, \\ \text{(iii) } \lim\limits_{\varepsilon\to 0} \|\nabla u_{1\varepsilon} - C^\varepsilon \nabla u_1\|_{C^0(0,T;[L^1(\Omega_{1\varepsilon})]^n)} = 0, \\ \text{(iv) } \lim\limits_{\varepsilon\to 0} \|\nabla u_{2\varepsilon}\|_{C^0(0,T;[L^2(\Omega_{2\varepsilon})]^n)} = 0. \end{cases} \tag{2.24}$$

Remark 2.9. Observe that in both cases, due to convergences (2.19)(iii) and (2.24)(iv), respectively, the additional assumptions (2.9) and (2.10) imply that convergences (2.16)(ii) and (2.22)(ii) are actually strong.

Theorem 2.4 will be proved in section 6. Its proof is more delicate than that of the case $-1 < \gamma < 1$, due to the particular form of the limit problem, and requires specific arguments (see Remark 6.3 for details).

3. Some preliminary results. In this section we will recall some technical results needed in the proofs of Theorems 2.4 and 2.8 as well as the homogenization result proved in [8] and [10] for the elliptic problem.

LEMMA 3.1 (see [7]). (i) *There exists a linear continuous extension operator Q_1 belonging to $\mathcal{L}(H^1(Y_1);H^1(Y)) \cap \mathcal{L}(L^2(Y_1);L^2(Y))$ such that, for some positive constant C,*

$$\|\nabla Q_1 v_1\|_{L^2(Y)} \le C\|\nabla v_1\|_{L^2(Y_1)}$$

for every $v_1 \in H^1(Y_1)$.

(ii) *There exists an extension operator Q_1^ε belonging to $\mathcal{L}(L^2(\Omega_{1\varepsilon});L^2(\Omega)) \cap \mathcal{L}(V^\varepsilon;H_0^1(\Omega))$ such that, for some positive constant C (independent of ε),*

$$\|Q_1^\varepsilon v_1\|_{L^2(\Omega)} \le C\|v_1\|_{L^2(\Omega_{1\varepsilon})}$$

and

$$\|\nabla Q_1^\varepsilon v_1\|_{L^2(\Omega)} \le C\|\nabla v_1\|_{L^2(\Omega_{1\varepsilon})}$$

for every $v_1 \in V^\varepsilon$.

LEMMA 3.2 (see [5]). *There exists a linear continuous extension operator P_1^ε belonging to $\mathcal{L}(L^\infty(0,T;H^k(\Omega_{1\varepsilon}));L^\infty(0,T;H^k(\Omega)))$, $k = 1, 2$, such that, for some positive constant C (independent of ε),*

$$\begin{cases} \text{(i) } P_1^\varepsilon\varphi = \varphi \text{ in } \Omega_{1\varepsilon}\times]0,T[, \\ \text{(ii) } P_1^\varepsilon\varphi' = (P^\varepsilon\varphi)' \text{in } \Omega\times]0,T[, \\ \text{(iii) } \|P_1^\varepsilon\varphi\|_{L^\infty(0,T;L^2(\Omega))} \le C\|\varphi\|_{L^\infty(0,T;L^2(\Omega_{1\varepsilon}))}, \\ \text{(iv) } \|P_1^\varepsilon\varphi'\|_{L^\infty(0,T;L^2(\Omega))} \le C\|\varphi'\|_{L^\infty(0,T;L^2(\Omega_{1\varepsilon}))}, \\ \text{(v) } \|\nabla(P_1^\varepsilon\varphi)\|_{L^\infty(0,T;[L^2(\Omega)]^n)} \le C\|\nabla\varphi\|_{L^\infty(0,T;[L^2(\Omega_{1\varepsilon})]^n)} \end{cases}$$

for any $\varphi \in L^\infty(0,T;H^k(\Omega_{1\varepsilon}))$.

The following homogenization result concerns the elliptic problem associated to problem (2.6).

THEOREM 3.3 (see [6] and [8]). *Let $-1 < \gamma \leq 1$, A^ε and h^ε be defined by* (2.4) *and* (2.5)*, respectively, and u_ε be the solution of problem*

$$\begin{cases} -\operatorname{div}(A^\varepsilon \nabla u_{1\varepsilon}) = b_1^\varepsilon + Q_1^{\varepsilon *}(g) & \text{in } \Omega_{1\varepsilon}, \\ -\operatorname{div}(A^\varepsilon \nabla u_{2\varepsilon}) = b_2^\varepsilon & \text{in } \Omega_{2\varepsilon}, \\ A^\varepsilon \nabla u_{1\varepsilon} \cdot n_{1\varepsilon} = -A^\varepsilon \nabla u_{2\varepsilon} \cdot n_{2\varepsilon} & \text{on } \Gamma^\varepsilon, \\ A^\varepsilon \nabla u_{1\varepsilon} \cdot n_{1\varepsilon} = -\varepsilon^\gamma h^\varepsilon (u_{1\varepsilon} - u_{2\varepsilon}) & \text{on } \Gamma^\varepsilon, \\ u_{1\varepsilon} = 0 & \text{on } \partial\Omega, \end{cases}$$

where $Q_1^\varepsilon \in \mathcal{L}(H^{-1}(\Omega); V^\varepsilon)$ is the adjoint of the extension operator given in Lemma 3.1(ii)*. Suppose that $b_1^\varepsilon \in L^2(\Omega_{1\varepsilon})$ and $b_2^\varepsilon \in L^2(\Omega_{2\varepsilon})$ satisfy*

$$\begin{cases} \widetilde{b_1^\varepsilon} \rightharpoonup \theta_1 b_1 & \text{weakly in } L^2(\Omega), \\ \widetilde{b_2^\varepsilon} \rightharpoonup \theta_2 b_2 & \text{weakly in } L^2(\Omega), \end{cases}$$

and let g be given in $L^2(\Omega)$. Then, there exists a positive constant C (independent of ε) such that

$$\begin{cases} Q_1^\varepsilon u_{1\varepsilon} \rightharpoonup u_1 & \text{weakly in } H_0^1(\Omega), \\ \widetilde{A^\varepsilon \nabla u_{1\varepsilon}} \rightharpoonup A^0 \nabla u_1 & \text{weakly in } [L^2(\Omega)]^n, \\ \|u_{1\varepsilon} - u_{2\varepsilon}\|_{L^2(\Gamma^\varepsilon)} < C\varepsilon^{-\gamma/2}, \end{cases}$$

and the following convergences hold:

$$\begin{cases} \widetilde{u_{2\varepsilon}} \rightharpoonup u_2 & \text{weakly in } L^2(\Omega), \\ \widetilde{A^\varepsilon \nabla u_{2\varepsilon}} \rightharpoonup 0 & \text{weakly in } [L^2(\Omega)]^n. \end{cases}$$

The function u_1 is the unique solution in $H_0^1(\Omega)$ of the problem

$$\begin{cases} -\operatorname{div}\left(A^0 \nabla u_1\right) = \theta_1 b_1 + \theta_2 b_2 + g & \text{in } \Omega, \\ u_1 = 0 & \text{on } \partial\Omega, \end{cases}$$

with θ_i, $i = 1, 2$, given by (2.2) *and A^0 by* (2.12)*.*

Moreover, for $-1 < \gamma < 1$, one has

$$\begin{cases} u_2 = \theta_2 u_1, \\ \|Q_1^\varepsilon u_{1\varepsilon} - u_{2\varepsilon}\|^2_{L^2(\Omega_{2\varepsilon})} \to 0, \end{cases}$$

while, for $\gamma = 1$,

$$u_2 = \theta_2 \left(u_1 + c_h^{-1} b_2\right),$$

where $c_h = \frac{1}{|Y_2|} \int_\Gamma h(y) d\sigma_y$.

The following lemma is a straightforward adaptation of Lemma 3.3 of [8] to the time-dependent case.

LEMMA 3.4. *Suppose that Γ is of class C^2. Let $g \in L^\infty(\Gamma)$, and set $c_g := \frac{1}{|Y_2|} \int_\Gamma g(y) d\sigma_y$.*

(i) *There exists* $\psi_g \in W^{1,\infty}(Y_2)$ *such that for every* $v_\varepsilon \in L^2(0,T;W^{1,1}(\Omega_{2\varepsilon}))$ *one has*

$$\varepsilon \int_{\Gamma^\varepsilon} g(x/\varepsilon) v_\varepsilon(x,t)\, d\sigma_x \ = c_g \int_{\Omega_{2\varepsilon}} v_\varepsilon(x,t)\, dx \ + \varepsilon \int_{\Omega_{2\varepsilon}} \nabla_y \psi_g(x/\varepsilon) \nabla_x v_\varepsilon(x,t)\, dx$$

for all $t \in [0,T]$.

(ii) *If for some positive constant* C *(independent of* ε*) one has*

$$||v_\varepsilon||_{L^2(0,T;W^{1,1}(\Omega_{2\varepsilon}))} \le C,$$

then

$$\liminf_{\varepsilon\to 0} \varepsilon \int_{\Gamma^\varepsilon} g(x/\varepsilon) v_\varepsilon(x,t)\, d\sigma_x \ = \liminf_{\varepsilon\to 0} c_g \int_{\Omega_{2\varepsilon}} v_\varepsilon(x,t)\, dx \ .$$

We will also need the following result.

LEMMA 3.5 (see [11]). *Let* $\mathcal{O}$ *be an open set of* $\mathbb{R}^n$ *and* $\{\mathcal{O}_\varepsilon\}_\varepsilon \subset \mathcal{O}$ *a sequence of open subsets of* $\mathcal{O}$. *Suppose that* $\{v_\varepsilon\}_\varepsilon \subset L^p(0,T;L^p(\mathcal{O}_\varepsilon))$, $p>1$, *is such that*

$$\begin{cases} \chi_{\mathcal{O}_\varepsilon} \rightharpoonup \chi_0 & \text{in } L^\infty(\mathcal{O}) \text{ weakly } *, \\ \widetilde{v_\varepsilon} \rightharpoonup \chi_0\, v & \text{weakly in } L^p\left(0,T;L^p(\mathcal{O})\right). \end{cases}$$

Then,

$$\liminf_{\varepsilon\to 0} \int_{\mathcal{O}_\varepsilon} |v_\varepsilon(t)|^p\, dx \ge \int_{\mathcal{O}} \chi_0 |v(t)|^p\, dx.$$

4. Asymptotic behavior of the energy. In this section we study the convergence of the energy associated to problem (2.6) which, for every ε, is defined by

(4.1)
$$E^\varepsilon(t) := \frac{1}{2}\Big[\int_{\Omega_{1\varepsilon}} |u'_{1\varepsilon}(t)|^2\, dx + \int_{\Omega_{2\varepsilon}} |u'_{2\varepsilon}(t)|^2\, dx + \int_{\Omega_{1\varepsilon}} A^\varepsilon \nabla u_{1\varepsilon}(t) \nabla u_{1\varepsilon}(t)\, dx + \int_{\Omega_{2\varepsilon}} A^\varepsilon \nabla u_{2\varepsilon}(t) \nabla u_{2\varepsilon}(t)\, dx + \varepsilon^\gamma \int_{\Gamma^\varepsilon} h^\varepsilon |u_{1\varepsilon}(t) - u_{2\varepsilon}(t)|^2\, d\sigma_x \Big].$$

LEMMA 4.1. *For* $-1 < \gamma \le 1$, *let* A^ε *and* h^ε *be defined by* (2.4) *and* (2.5), *respectively, and suppose that* (2.7) *holds. If* u_ε *is the solution of problem* (2.6), *then*

$$\begin{aligned} E^\varepsilon(t) = \frac{1}{2}\Big[& \int_{\Omega_{1\varepsilon}} \left|U^1_\varepsilon\right|^2 dx + \int_{\Omega_{2\varepsilon}} \left|U^1_\varepsilon\right|^2 dx + \int_{\Omega_{1\varepsilon}} A^\varepsilon \nabla U^0_{1\varepsilon} \nabla U^0_{1\varepsilon}\, dx \\ & + \int_{\Omega_{2\varepsilon}} A^\varepsilon \nabla U^0_{2\varepsilon} \nabla U^0_{2\varepsilon}\, dx + \varepsilon^\gamma \int_{\Gamma^\varepsilon} h^\varepsilon \left|U^0_{1\varepsilon} - U^0_{2\varepsilon}\right|^2 d\sigma_x \Big] \\ & + \int_0^t \int_{\Omega_{1\varepsilon}} f_{1\varepsilon} u'_{1\varepsilon}\, dx\, d\tau + \int_0^t \int_{\Omega_{2\varepsilon}} f_{2\varepsilon} u'_{2\varepsilon}\, dx\, d\tau \ \forall\, t \in [0,T]. \end{aligned} \tag{4.2}$$

Proof. If we take $(u'_{1\varepsilon}, u'_{2\varepsilon})$ as test functions in the variational formulation of problem (2.6) (actually a standard density argument has to be used, see, for instance,

[4] and [20]) we get

$$\begin{aligned}
&\int_0^t \langle u''_{1\varepsilon}, u'_{1\varepsilon}\rangle_{(V^\varepsilon)',V^\varepsilon}\, d\tau + \int_0^t \langle u''_{2\varepsilon}, u'_{2\varepsilon}\rangle_{(H^1(\Omega_{2\varepsilon}))',H^1(\Omega_{2\varepsilon})}\, d\tau \\
&+ \int_0^t \int_{\Omega_{1\varepsilon}} A^\varepsilon \nabla u_{1\varepsilon} \nabla u'_{1\varepsilon}\, dx\, d\tau + \int_0^t \int_{\Omega_{2\varepsilon}} A^\varepsilon \nabla u_{2\varepsilon} \nabla u'_{2\varepsilon}\, dx\, d\tau \\
&\qquad + \varepsilon^\gamma \int_0^t \int_{\Gamma^\varepsilon} h^\varepsilon (u_{1\varepsilon} - u_{2\varepsilon})(u'_{1\varepsilon} - u'_{2\varepsilon})\, d\sigma_x\, d\tau \\
&\qquad = \int_0^t \int_{\Omega_{1\varepsilon}} f_{1\varepsilon} u'_{1\varepsilon}\, dx\, d\tau + \int_0^t \int_{\Omega_{2\varepsilon}} f_{2\varepsilon} u'_{2\varepsilon}\, dx\, d\tau.
\end{aligned} \tag{4.3}$$

Moreover,

$$\int_0^t \langle u''_{i\varepsilon}, u'_{i\varepsilon}\rangle_{(V^\varepsilon)',V^\varepsilon}\, d\tau = \frac{1}{2}\int_0^t \frac{d}{d\tau} \int_{\Omega_{i\varepsilon}} |u'_{i\varepsilon}|^2\, dx\, d\tau, i = 1, 2.$$

Similarly, since the matrix A is symmetric we get

$$\int_0^t \int_{\Omega_{i\varepsilon}} A^\varepsilon \nabla u_{i\varepsilon} \cdot \nabla u'_{i\varepsilon}\, dx\, d\tau = \frac{1}{2}\int_0^t \frac{d}{d\tau} \int_{\Omega_{i\varepsilon}} A^\varepsilon \nabla u_{i\varepsilon} \cdot \nabla u_{i\varepsilon}\, dx\, d\tau, \quad i = 1, 2.$$

On the other hand,

$$\varepsilon^\gamma \int_0^t \int_{\Gamma^\varepsilon} h^\varepsilon (u_{1\varepsilon} - u_{2\varepsilon})(u'_{1\varepsilon} - u'_{2\varepsilon})\, d\sigma_x\, d\tau = \frac{\varepsilon^\gamma}{2}\int_0^t \frac{d}{d\tau} \int_{\Gamma^\varepsilon} h^\varepsilon |u_{1\varepsilon} - u_{2\varepsilon}|^2\, d\sigma_x\, d\tau.$$

Therefore, identity (4.3) can be rewritten, using (4.1), as

$$E^\varepsilon(t) - E^\varepsilon(0) = \int_0^t \frac{dE^\varepsilon}{d\tau}\, d\tau = \int_0^t \int_{\Omega_{1\varepsilon}} f_{1\varepsilon} u'_{1\varepsilon}\, dx\, d\tau + \int_0^t \int_{\Omega_{2\varepsilon}} f_{2\varepsilon} u'_{2\varepsilon}\, dx\, d\tau.$$

Hence, again by (4.1) and (2.6), we have (4.2). ∎

LEMMA 4.2. *For $-1 < \gamma \le 1$, let A^ε and h^ε be defined by* (2.4) *and* (2.5), *respectively, and suppose that* (2.7), (2.9), *and* (2.10) *hold. If u_ε is the solution of problem* (2.6), *then*

$$\begin{aligned}
\lim_{\varepsilon\to 0} E^\varepsilon(t) = \frac{1}{2}\left[\int_\Omega \left|U^1\right|^2 dx + \int_\Omega A^0 \nabla U^0 \nabla U^0\, dx\right] \\
+ \int_0^t \int_\Omega f_1 u'_1\, dx\, d\tau + {\theta_2}^{-1} \int_0^t \int_\Omega f_2 u'_2\, dx\, d\tau \quad \forall t \in [0, T],
\end{aligned}$$

where we have $u'_2 = \theta_2 u'_1$, if $-1 < \gamma < 1$.

Proof. We make use of Lemma 4.1, passing to the limit in (4.2) for any fixed $t \in [0, T]$.

For the first two integrals, using (2.9), we have

$$\lim_{\varepsilon\to 0}\left(\int_{\Omega_{1\varepsilon}} \left|U^1_\varepsilon\right|^2 dx + \int_{\Omega_{2\varepsilon}} \left|U^1_\varepsilon\right|^2 dx\right) = \int_\Omega \left|U^1\right|^2 dx.$$

Observe now that (2.11) holds in both cases $-1 < \gamma < 1$ and $\gamma = 1$. Indeed, Theorem 3.3 written for $u_{1\varepsilon} = U^0_{1\varepsilon}$, $u_{2\varepsilon} = U^0_{2\varepsilon}$, $b_{1\varepsilon} = b_{2\varepsilon} = 0$, and $g = -\mathrm{div}\, A^0 \nabla U^0$

gives, by the uniqueness of the solution of the limit problem, $u_1 = U^0$. Hence, if we take $(U^0_{1\varepsilon}, U^0_{2\varepsilon})$ as test functions in (2.10), in view of (2.11)(i) we have

$$\lim_{\varepsilon\to 0}\left[\int_{\Omega_{1\varepsilon}} A^\varepsilon \nabla U^0_{1\varepsilon}\nabla U^0_{1\varepsilon}\,dx + \int_{\Omega_{2\varepsilon}} A^\varepsilon \nabla U^0_{2\varepsilon}\nabla U^0_{2\varepsilon}\,dx + \varepsilon^\gamma \int_{\Gamma^\varepsilon} h^\varepsilon \left(U^0_{1\varepsilon} - U^0_{2\varepsilon}\right)^2 d\sigma_x\right]$$
$$= \lim_{\varepsilon\to 0}\left\langle -\operatorname{div}\left(A^0\nabla U^0\right), Q^\varepsilon_1 U^0_{1\varepsilon}\right\rangle = \int_\Omega A^0\nabla U^0\nabla U^0\,dx.$$

On the other hand, (2.9) and Remark 2.7(ii) imply that

$$\lim_{\varepsilon\to 0}\int_0^t\int_{\Omega_{1\varepsilon}} f_{1\varepsilon}u'_{1\varepsilon}\,dx\,d\tau + \int_0^t\int_{\Omega_{2\varepsilon}} f_{2\varepsilon}u'_{2\varepsilon}\,dx\,d\tau$$
$$= \int_0^t\int_\Omega f_1u'_1\,dx\,d\tau + {\theta_2}^{-1}\int_0^t\int_\Omega f_2u'_2\,dx\,d\tau,$$

where for $-1 < \gamma < 1$ and $\gamma = 1$ we used convergences (2.15) and (2.21), respectively. □

From now on, the convergence of the energy in the two cases $-1 < \gamma < 1$ and $\gamma = 1$ need to be studied separately.

THEOREM 4.3 (convergence of energy for $-1 < \gamma < 1$). *Let A^ε and h^ε be defined by* (2.4) *and* (2.5)*, respectively, and suppose that* (2.7), (2.9), *and* (2.10) *hold. If u_ε is the solution of problem* (2.6) *with $-1 < \gamma < 1$, then*

$$E^\varepsilon \to E \text{ in } C^0([0,T]), \tag{4.4}$$

where

$$E(t) := \frac{1}{2}\left[\int_\Omega |u'_1(t)|^2\,dx + \int_\Omega A^0\nabla u_1(t)\nabla u_1(t)\,dx\right] \tag{4.5}$$

is the energy associated to problem (2.17).

Proof. Taking u'_1 as test function in problem (2.17) and arguing as in the proof of Lemma 4.1, we obtain

$$E(t) = \frac{1}{2}\left[\int_\Omega \left|U^1\right|^2 dx + \int_\Omega A^0\nabla U^0\nabla U^0\,dx\right] + \int_0^t\int_\Omega (f_1+f_2)u'_1\,dx\,d\tau.$$

Hence, by Lemma 4.2 we obtain

$$\lim_{\varepsilon\to 0} E^\varepsilon(t) = E(t) \qquad \forall t\in[0,T], \tag{4.6}$$

since in this case $u'_2 = \theta_2 u'_1$.

Using Lemma 4.1, Theorem 2.3, (2.7), (2.9), and (2.10) we deduce that there exists a constant C (independent of ε) such that

$$|E^\varepsilon(t)| \le C \quad \forall\, t\ \in[0,T]$$

and, from the Hölder inequality,

$$|E^\varepsilon(t+\sigma) - E^\varepsilon(t)|$$
$$= \left|\int_t^{t+\sigma}\int_{\Omega_{1\varepsilon}} f_{1\varepsilon}u'_{1\varepsilon}\,dx\,d\tau + \int_t^{t+\sigma}\int_{\Omega_{2\varepsilon}} f_{2\varepsilon}u'_{2\varepsilon}\,dx\,d\tau\right| \le \sigma^{1/2}C.$$

Consequently, by the Ascoli–Arzelà theorem, there exist a subsequence, still denoted by E^ε, and a function ζ in $C^0([0,T])$ such that

$$E^\varepsilon \to \zeta \text{ in } C^0([0,T]). \tag{4.7}$$

Hence, from (4.6) and (4.7), by uniqueness, one has $\zeta = E$, which concludes the proof. ∎

THEOREM 4.4 (convergence of energy for $\gamma = 1$). *Let A^ε and h^ε be defined by* (2.4) *and* (2.5), *respectively, and suppose that* (2.7), (2.9), *and* (2.10) *hold. If u_ε is the solution of problem* (2.6) *with $\gamma = 1$, then*

$$E^\varepsilon \to \hat{E} \text{ in } C^0([0,T]),$$

where

$$\begin{aligned}\hat{E}(t) := \frac{1}{2}\Big[&\theta_1 \int_\Omega |u_1'(t)|^2\,dx + {\theta_2}^{-1}\int_\Omega |u_2'(t)|^2\,dx \\ &+ \int_\Omega A^0\nabla u_1(t)\nabla u_1(t)\,dx + c_h{\theta_2}^{-1}\int_\Omega |\theta_2 u_1(t) - u_2(t)|^2\,dx\Big]\end{aligned} \tag{4.8}$$

is the energy associated to problem (2.23).

Proof. Let us prove first that

$$\begin{aligned}\hat{E}(t) = \frac{1}{2}&\left[\int_\Omega \left|U^1\right|^2 dx + \int_\Omega A^0\nabla U^0\nabla U^0\,dx\right] \\ &+ \int_0^t\int_\Omega f_1u_1'\,dx\,d\tau + \theta_2^{-1}\int_0^t\int_\Omega f_2u_2'\,dx\,d\tau \quad \forall t\in[0,T].\end{aligned} \tag{4.9}$$

Let us take u_1' as test function in the first equation of (2.23). Multiplying the second equation by $\theta_2^{-1}u_2'$ and integrating over $\Omega\times[0,t]$ we have by summation

$$\begin{aligned}&\int_0^t \langle\theta_1u_1'', u_1'\rangle_{L^2(\Omega),L^2(\Omega)}\,d\tau + \int_0^t \langle u_2'', \theta_2^{-1}u_2'\rangle_{L^2(\Omega),L^2(\Omega)}\,d\tau \\ &\qquad + \int_0^t\int_\Omega A^0\nabla u_1\nabla u_1'\,dx\,d\tau + c_h\int_0^t\int_\Omega(\theta_2u_1 - u_2)u_1'\,dx\,d\tau \\ &\qquad - c_h\int_0^t\int_\Omega(\theta_2u_1 - u_2)\theta_2^{-1}u_2'\,dx\,d\tau \\ &\qquad = \int_0^t\int_\Omega f_1u_1'\,dx\,d\tau + {\theta_2}^{-1}\int_0^t\int_\Omega f_2u_2'\,dx\,d\tau \quad \forall t\in[0,T].\end{aligned} \tag{4.10}$$

Now, observe that

$$\begin{aligned}\int_0^t \langle\theta_1u_1'', u_1'\rangle_{L^2(\Omega),L^2(\Omega)}\,d\tau &= \frac{1}{2}\theta_1\int_0^t\frac{d}{d\tau}\int_\Omega |u_1'|^2\,dx\,d\tau, \\ \int_0^t \langle u_2'', \theta_2^{-1}u_2'\rangle_{L^2(\Omega),L^2(\Omega)}\,d\tau &= \frac{1}{2}\theta_2^{-1}\int_0^t\frac{d}{d\tau}\int_\Omega |u_2'|^2\,dx\,d\tau.\end{aligned}$$

Similarly, taking into account that the matrix A^0 is symmetric we get

$$\int_0^t\int_\Omega A^0\nabla u_1\nabla u_1'\,dx\,d\tau = \frac{1}{2}\int_0^t\frac{d}{d\tau}\int_\Omega A^0\nabla u_1\cdot\nabla u_1\,dx\,d\tau.$$

On the other hand,

$$c_h \int_0^t \int_\Omega (\theta_2 u_1 - u_2) u_1' \, dx \; d\tau - c_h \int_0^t \int_\Omega (\theta_2 u_1 - u_2) \theta_2^{-1} u_2' \, dx \; d\tau$$
$$= c_h \theta_2^{-1} \int_0^t \int_\Omega (\theta_2 u_1 - u_2)(\theta_2 u_1' - u_2') \, dx \, d\tau = \frac{c_h \theta_2^{-1}}{2} \int_0^t \frac{d}{d\tau} \int_\Omega (\theta_2 u_1 - u_2)^2 dx \, d\tau.$$

Therefore, taking into account (4.8) and identity (4.10) we obtain

$$\hat{E}(t) - \hat{E}(0) = \int_0^t \frac{d\hat{E}}{d\tau} \, d\tau = \int_0^t \int_\Omega f_1 u_1' \, dx \, d\tau + \theta_2{}^{-1} \int_0^t \int_\Omega f_2 u_2' \, dx \, d\tau \quad \forall t \in [0,T].$$

Using (2.23) and again (2.8), we deduce that

$$\begin{aligned} \hat{E}(0) &= \frac{1}{2} \left[\theta_1 \int_\Omega \left| U^1 \right|^2 dx + \theta_2^{-1} \int_\Omega \left| \theta_2 U^1 \right|^2 dx + \int_\Omega A^0 \nabla U^0 \nabla U^0 \, dx \right] \\ &\quad + c_h \theta_2^{-1} \int_\Omega \left| \theta_2 U^0 - \theta_2 U^0 \right|^2 dx \\ &= \frac{1}{2} \left[\int_\Omega \left| U^1 \right|^2 dx + \int_\Omega A^0 \nabla U^0 \nabla U^0 \, dx \right]. \end{aligned}$$

This gives (4.9) which, together with Lemma 4.2, implies that

$$\lim_{\varepsilon \to 0} E^\varepsilon(t) \to \hat{E}(t) \quad \forall t \in [0,T].$$

Now, arguing as in the proof of Theorem 4.3, but using here Theorem 2.6 instead of Theorem 2.3, we conclude the proof. □

5. Proof of the corrector result in the case $-1 < \gamma < 1$. Let Q_1 be defined by Lemma 3.1 and introduce the test functions associated to the solution w_λ of problem (2.13) by

$$\chi_{1\lambda} = \lambda \cdot y - w_\lambda(y), \;\; w_\lambda^\varepsilon(x) := \; \lambda \cdot x - \varepsilon(Q_1(\chi_{1\lambda})(x/\varepsilon)). \tag{5.1}$$

A simple change of scale gives that

$$\int_{\Omega_{1\varepsilon}} A^\varepsilon \nabla w_\lambda^\varepsilon \cdot \nabla v_1 \, dx = 0 \tag{5.2}$$

for every $v_1 \in V^\varepsilon$ and (see [7]) the following convergences hold true:

$$\begin{cases} w_\lambda^\varepsilon \to \lambda \cdot x & \text{weakly in } H^1(\Omega), \\ w_\lambda^\varepsilon \to \lambda \cdot x & \text{strongly in } L^2(\Omega), \\ \chi_{\Omega_{1\varepsilon}} A^\varepsilon \nabla w_\lambda^\varepsilon \rightharpoonup A^0 \lambda & \text{weakly in } \; [L^2(\Omega)]^n, \end{cases} \tag{5.3}$$

where A^0 is given by (2.12).

From now on, we will adopt the Einstein summation convention. The proof of the corrector result is based on the following proposition.

PROPOSITION 5.1. *Suppose that the assumptions of Theorem* 2.4 *are fulfilled. Set for any* $\Phi \in C^\infty([0,T], \mathcal{D}(\Omega))$

$$X_\varepsilon(t) = \frac{1}{2} \Big[\int_\Omega \left| \widetilde{u_{1\varepsilon}'}(t) + \widetilde{u_{2\varepsilon}'}(t) - \Phi'(t) \right|^2 \, dx$$
$$+ \int_{\Omega_{1\varepsilon}} A^\varepsilon (\nabla u_{1\varepsilon}(t) - C^\varepsilon \nabla \Phi(t))(\nabla u_{1\varepsilon}(t) - C^\varepsilon \nabla \Phi(t)) \, dx + \int_{\Omega_{2\varepsilon}} A^\varepsilon \nabla u_{2\varepsilon}(t) \nabla u_{2\varepsilon}(t) \, dx \Big],$$

where C^ε is given in (2.18). *Then*

$$\limsup_{\varepsilon\to 0} \|X_\varepsilon\|_{C^0([0,T])} \le \|X\|_{C^0([0,T])}, \tag{5.4}$$

where

$$X(t) = \frac{1}{2}\left[\|u_1'(t) - \Phi'(t)\|^2_{L^2(\Omega)} + \int_\Omega A^0(\nabla u_1(t) - \nabla\Phi(t))(\nabla u_1(t) - \nabla\Phi(t))\,dx\right].$$

Remark 5.2. Observe that in general in literature (see, for instance, [5], [17], [18]) one can introduce suitable quantities X_ε and X such that

$$X_\varepsilon \to X \quad \text{in } C^0([0,T]).$$

Here, we can prove only (5.4), which is just an inequality on the upper limit of $\|X_\varepsilon\|_{C^0([0,T])}$. Nevertheless, this will be sufficient to prove the corrector result.

Proof of Proposition 5.1. Using the symmetry of A^ε and C^ε, one has

$$\begin{aligned} X_\varepsilon(t) = \frac{1}{2}\Bigg[&\int_\Omega \left|\widetilde{u_{1\varepsilon}'}(t) + \widetilde{u_{2\varepsilon}'}(t) - \Phi'(t)\right|^2 \\ &+ \int_{\Omega_{1\varepsilon}} A^\varepsilon \nabla u_{1\varepsilon}(t)\nabla u_{1\varepsilon}(t)\,dx - 2\int_{\Omega_{1\varepsilon}} A^\varepsilon C^\varepsilon \nabla\Phi(t)\nabla u_{1\varepsilon}(t)\,dx \\ &+ \int_{\Omega_{1\varepsilon}} A^\varepsilon C^\varepsilon \nabla\Phi(t) C^\varepsilon\nabla\Phi(t)\,dx + \int_{\Omega_{2\varepsilon}} A^\varepsilon \nabla u_{2\varepsilon}(t)\nabla u_{2\varepsilon}(t)\,dx\Bigg]. \end{aligned}$$

Set, for $t\in[0,T]$,

$$\begin{aligned} X_{1\varepsilon}(t) &= \frac{1}{2}\Bigg[\int_\Omega \left|\widetilde{u_{1\varepsilon}'}(t) + \widetilde{u_{2\varepsilon}'}(t)\right|^2 + \int_{\Omega_{1\varepsilon}} A^\varepsilon \nabla u_{1\varepsilon}(t)\nabla u_{1\varepsilon}(t)\,dx \\ &\qquad + \int_{\Omega_{2\varepsilon}} A^\varepsilon \nabla u_{2\varepsilon}(t)\nabla u_{2\varepsilon}(t)\,dx\Bigg], \\ X_{2\varepsilon}(t) &= \int_\Omega \left(\widetilde{u_{1\varepsilon}'}(t) + \widetilde{u_{2\varepsilon}'}(t)\right)\Phi'(t)\,dx + \int_{\Omega_{1\varepsilon}} A^\varepsilon C^\varepsilon\nabla\Phi(t)\nabla u_{1\varepsilon}(t)\,dx, \end{aligned}$$

and

$$X_{3\varepsilon}(t) = \frac{1}{2}\left[\int_\Omega |\Phi'(t)|^2 + \int_{\Omega_{1\varepsilon}} A^\varepsilon C^\varepsilon\nabla\Phi(t) C^\varepsilon\nabla\Phi(t)\,dx\right]$$

so that

$$X_\varepsilon = X_{1\varepsilon} - X_{2\varepsilon} + X_{3\varepsilon}.$$

For $X_{1\varepsilon}$, recalling (4.1), we have

$$X_{1\varepsilon}(t) \le E^\varepsilon(t) \quad \forall\, t\in[0,T]. \tag{5.5}$$

For $X_{2\varepsilon}$, taking into account (2.18), (5.1) written for $\lambda = e_i$, and (5.2) we obtain

$$\begin{aligned} X_{2\varepsilon}(t) &= \int_\Omega \left(\widetilde{u'_{1\varepsilon}}(t) + \widetilde{u'_{2\varepsilon}}(t)\right) \Phi'(t)\, dx + \int_{\Omega_{1\varepsilon}} A^\varepsilon \nabla w_i^\varepsilon \frac{\partial \Phi}{\partial x_i}(t) \nabla u_{1\varepsilon}(t)\, dx \\ &= \int_\Omega \left(\widetilde{u'_{1\varepsilon}}(t) + \widetilde{u'_{2\varepsilon}}(t)\right) \Phi'(t)\, dx + \int_{\Omega_{1\varepsilon}} A^\varepsilon \nabla w_i^\varepsilon \nabla \left[\frac{\partial \Phi}{\partial x_i}(t)\, u_{1\varepsilon}(t)\right] dx \\ &\quad - \int_{\Omega_{1\varepsilon}} A^\varepsilon \nabla w_i^\varepsilon \nabla \left[\frac{\partial \Phi}{\partial x_i}(t)\right] u_{1\varepsilon}(t)\, dx \\ &= \int_\Omega \left(\widetilde{u'_{1\varepsilon}}(t) + \widetilde{u'_{2\varepsilon}}(t)\right) \Phi'(t)\, dx \\ &\quad - \int_\Omega \chi_{\Omega_{1\varepsilon}} A^\varepsilon \nabla w_i^\varepsilon \nabla \left[\frac{\partial \Phi}{\partial x_i}(t)\right] P_1^\varepsilon u_{1\varepsilon}(t)\, dx \ \forall\, t \in [0,T]. \end{aligned} \tag{5.6}$$

Hence, by (2.14), (2.15), and (5.3) we conclude that

$$\begin{aligned} X_{2\varepsilon} &\to \int_\Omega (\theta_1 u'_1 + \theta_2 u'_1)\Phi'\, dx - \int_\Omega A^0 e_i \nabla \left[\frac{\partial \Phi}{\partial x_i}\right] u_1\, dx \\ &= \int_\Omega u'_1 \Phi'\, dx + \int_\Omega A^0 \nabla \Phi \nabla u_1\, dx \text{ in } \mathcal{D}'(0,T). \end{aligned} \tag{5.7}$$

Moreover, taking into account (5.6), by Theorem 2.3, (5.2), and (5.3) we obtain that there exists a constant C (independent of ε) such that

$$\|X_{2\varepsilon}\|_{L^\infty(0,T)} \le C. \tag{5.8}$$

Consequently, to prove that convergence (5.7) takes place in $C^0([0,T])$ it is enough to show that

$$\left\| \frac{\partial X_{2\varepsilon}}{\partial t} \right\|_{L^2(0,T)} \le C. \tag{5.9}$$

Using the variational formulation (2.8), we have

$$\begin{aligned} \frac{\partial}{\partial t} \int_\Omega \left(\widetilde{u'_{1\varepsilon}}(t) + \widetilde{u'_{2\varepsilon}}(t)\right) \Phi'(t)\, dx &= \int_\Omega \left(\widetilde{u'_{1\varepsilon}}(t) + \widetilde{u'_{2\varepsilon}}(t)\right) \Phi''(t) \\ &\quad - \int_{\Omega_{1\varepsilon}} A^\varepsilon \nabla u_{1\varepsilon}(t) \cdot \nabla \Phi'(t)\, dx - \int_{\Omega_{2\varepsilon}} A^\varepsilon \nabla u_{2\varepsilon}(t) \cdot \nabla \Phi'(t)\, dx \\ &\quad + \int_{\Omega_{1\varepsilon}} f_{1\varepsilon}(t)\Phi'(t)\, dx + \int_{\Omega_{2\varepsilon}} f_{2\varepsilon}(t)\Phi'(t)\, dx, \end{aligned}$$

which is bounded in $L^2(0,T)$ due to (2.15), (2.16), and (2.7).

Moreover, we have

$$\begin{aligned} \frac{\partial}{\partial t} \int_\Omega \chi_{\Omega_{1\varepsilon}} A^\varepsilon \nabla w_i^\varepsilon \nabla \left[\frac{\partial \Phi}{\partial x_i}(t)\right] P_1^\varepsilon u_{1\varepsilon}(t)\, dx &= \int_\Omega \chi_{\Omega_{1\varepsilon}} A^\varepsilon \nabla w_i^\varepsilon \nabla \left[\frac{\partial \Phi'}{\partial x_i}(t)\right] P_1^\varepsilon u_{1\varepsilon}(t)\, dx \\ &\quad + \int_\Omega \chi_{\Omega_{1\varepsilon}} A^\varepsilon \nabla w_i^\varepsilon \nabla \left[\frac{\partial \Phi}{\partial x_i}(t)\right] P_1^\varepsilon u'_{1\varepsilon}(t)\, dx, \end{aligned}$$

which is bounded in $L^2(0,T)$ as a consequence of (5.3) and Theorem 2.3.

Thus, from (5.6) we obtain (5.9). Due to classical compactness results, from this inequality and (5.8) it turns out that $X_{2\varepsilon}$ is relatively compact in $C^0([0,T])$. This, together with (5.7), gives

$$X_{2\varepsilon} \to X_2 := \int_\Omega u_1'\Phi'\,dx + \int_\Omega A^0\nabla\Phi\nabla u_1\,dx \text{ strongly in } C^0([0,T]). \tag{5.10}$$

For $X_{3\varepsilon}$, by the smoothness of Φ we have that it is bounded in $L^\infty(0,T)$ as well as its time derivative. Moreover, by (5.2) we deduce that

$$\begin{aligned}\int_{\Omega_{1\varepsilon}} A^\varepsilon C^\varepsilon\nabla\Phi(t)C^\varepsilon\nabla\Phi(t)\,dx &= \int_{\Omega_{1\varepsilon}} A^\varepsilon\nabla w_i^\varepsilon\frac{\partial\Phi}{\partial x_i}(t)\nabla w_j^\varepsilon\frac{\partial\Phi}{\partial x_j}(t)\,dx\\ &= -\int_{\Omega_{1\varepsilon}} A^\varepsilon\nabla w_i^\varepsilon w_j^\varepsilon\nabla\left[\frac{\partial\Phi}{\partial x_i}(t)\frac{\partial\Phi}{\partial x_j}(t)\right]\,dx.\end{aligned}$$

Thus, from (5.3)

$$\begin{aligned}\lim_{\varepsilon\to0} X_{3\varepsilon}(t) &= \lim_{\varepsilon\to0}\frac12\left[\|\Phi'(t)\|^2_{L^2(\Omega)} + \int_{\Omega_{1\varepsilon}} A^\varepsilon C^\varepsilon\nabla\Phi(t)C^\varepsilon\nabla\Phi(t)\,dx\right]\\ &= \frac12\left[\|\Phi'(t)\|^2_{L^2(\Omega)} - \int_\Omega A^0 e_i x_j\nabla\left[\frac{\partial\Phi}{\partial x_i}(t)\frac{\partial\Phi}{\partial x_j}(t)\right]\,dx\right]\\ &= \frac12\left[\|\Phi'(t)\|^2_{L^2(\Omega)} + \int_\Omega A^0\nabla\Phi(t)\nabla\Phi(t)\,dx\right]\ \forall\, t\in[0,T],\end{aligned}$$

which gives

$$X_{3\varepsilon} \to X_3 := \frac12\left[\|\Phi'\|^2_{L^2(\Omega)} + \int_\Omega A^0\nabla\Phi\nabla\Phi\,dx\right] \text{ strongly in } C^0([0,T]). \tag{5.11}$$

Convergence (5.11), together with (5.10), (4.4), and (4.5), gives

$$E^\varepsilon - X_{2\varepsilon} + X_{3\varepsilon} \to E - X_2 + X_3 = X \quad \text{strongly in } C^0([0,T]). \tag{5.12}$$

Since from (5.5) one has

$$0 \le X_\varepsilon(t) \le E^\varepsilon(t) - X_{2\varepsilon}(t) + X_{3\varepsilon}(t) \quad \forall\, t\in[0,T],$$

by (5.12) we conclude that

$$\limsup_{\varepsilon\to0}\|X_\varepsilon\|_{C^0([0,T])} \le \lim_{\varepsilon\to0}\|E^\varepsilon - X_{2\varepsilon} + X_{3\varepsilon}\|_{C^0([0,T])} = \|X\|_{C^0([0,T])}. \qquad \square$$

Remark 5.3. Let us point out that the main difficulty when proving Proposition 5.1 is due to the fact that, as a consequence of the presence of the boundary term in the energy, we have only $X_{1\varepsilon} \le E^\varepsilon$ and not $X_{1\varepsilon} = E^\varepsilon$ as in the classical cases.

Nevertheless, the fact that $E^\varepsilon - X_{2\varepsilon} + X_{3\varepsilon} \to X$ in $C^0([0,T])$ allows us to prove the result.

Proof of Theorem 2.4. The function $u_1 \in L^2(0,T;H^1_0(\Omega))\cap C^0([0,T];L^2(\Omega))$ and its derivative $u_1' \in C^0([0,T];L^2(\Omega))$; thus, from classical density results, for any $\delta>0$ there exists $\Phi\in C^\infty([0,T];\mathcal{D}(\Omega))$ such that

$$\begin{cases}\|u_1' - \Phi'\|_{C^0([0,T];L^2(\Omega))} \le \delta,\\ \|\nabla u_1 - \nabla\Phi\|_{L^2(0,T;L^2(\Omega))} \le \delta.\end{cases} \tag{5.13}$$

Therefore,

$$\begin{aligned}
&\left\|\widetilde{u'_{1\varepsilon}}+\widetilde{u'_{2\varepsilon}}-u'_1\right\|^2_{C^0(0,T;L^2(\Omega))}\\
&\qquad\le 2\Big(\left\|\widetilde{u'_{1\varepsilon}}+\widetilde{u'_{2\varepsilon}}-\Phi'\right\|^2_{C^0(0,T;L^2(\Omega))}+\|\Phi'-u'_1\|^2_{C^0(0,T;L^2(\Omega))}\Big)\\
&\qquad\le 2\left\|\widetilde{u'_{1\varepsilon}}+\widetilde{u'_{2\varepsilon}}-\Phi'\right\|^2_{C^0(0,T;L^2(\Omega))}+2\delta^2.
\end{aligned}\tag{5.14}$$

On the other hand, due to the Cauchy–Schwarz inequality, definition (2.18) of C^ε, and (5.13) there exist two constants C_1, C_2 such that

$$\begin{aligned}
&\|\nabla u_{1\varepsilon}-C^\varepsilon\nabla u_1\|^2_{C^0([0,T];[L^1(\Omega_{1\varepsilon})]^n)}+\|\nabla u_{2\varepsilon}\|^2_{C^0([0,T];[L^2(\Omega_{2\varepsilon})]^n)}\\
&\quad\le 2\|\nabla u_{1\varepsilon}-C^\varepsilon\nabla\Phi\|^2_{C^0([0,T];[L^1(\Omega_{1\varepsilon})]^n)}+\|\nabla u_{2\varepsilon}\|^2_{C^0([0,T];[L^2(\Omega_{2\varepsilon})]^n)}\\
&\qquad+2\|C^\varepsilon(\nabla\Phi-\nabla u_1)\|^2_{C^0([0,T];[L^1(\Omega_{1\varepsilon})]^n)}\\
&\quad\le 2C_1\|\nabla u_{1\varepsilon}-C^\varepsilon\nabla\Phi\|^2_{C^0([0,T];[L^2(\Omega_{1\varepsilon})]^n)}+\|\nabla u_{2\varepsilon}\|^2_{C^0([0,T];[L^2(\Omega_{2\varepsilon})]^n)}+2C_2\delta^2.
\end{aligned}\tag{5.15}$$

From properties (2.3) and definition (2.4) of A^ε, we have

$$\begin{aligned}
&\frac12\left\|\widetilde{u'_{1\varepsilon}}(t)+\widetilde{u'_{2\varepsilon}}(t)-\Phi'(t)\right\|^2_{L^2(\Omega)}\\
&\qquad+\frac{\alpha}{2}\left(\|\nabla u_{1\varepsilon}(t)-C^\varepsilon\nabla\Phi(t)\|^2_{[L^2(\Omega_{1\varepsilon})]^n}+\|\nabla u_{2\varepsilon}(t)\|^2_{[L^2(\Omega_{2\varepsilon})]^n}\right)\le X_\varepsilon(t)\quad\forall t\in[0,T];
\end{aligned}$$

moreover, there exists a constant C_3

$$\|X\|_{C^0([0,T])}\le C_3\delta^2,$$

where X_ε and X are defined in Proposition 5.1 and are applied to the function Φ, given by (5.13).

From this last inequality and by Proposition 5.1, we conclude that

$$\begin{aligned}
\limsup_{\varepsilon\to0}&\left\{\frac12\left\|\widetilde{u'_{1\varepsilon}}+\widetilde{u'_{2\varepsilon}}-\Phi'\right\|^2_{C^0([0,T];L^2(\Omega))}\right.\\
&\left.+\frac{\alpha}{2}\left(\|\nabla u_{1\varepsilon}-C^\varepsilon\nabla\Phi\|^2_{C^0([0,T];[L^2(\Omega_{1\varepsilon})]^n)}+\|\nabla u_{2\varepsilon}\|^2_{C^0([0,T];[L^2(\Omega_{2\varepsilon})]^n)}\right)\right\}\\
&\qquad\le\limsup_{\varepsilon\to0}\|X_\varepsilon\|_{C^0([0,T])}\le\|X\|_{C^0([0,T])}\le C_3\delta^2.
\end{aligned}$$

This, together with (5.14) and (5.15), gives convergences (2.19), δ being arbitrary. □

6. Proof of the corrector result in the case $\gamma = 1$. The proof is based on the result below, which is analogous to that shown for $-1<\gamma<1$ in Proposition 5.1. Let us emphasize that here the proof is more delicate and contains some technically specific parts.

PROPOSITION 6.1. *Suppose that the assumptions of Theorem* 2.8 *are fulfilled. Set for any* $\Phi, \Psi \in C^\infty([0,T], \mathcal{D}(\Omega))$

$$\begin{aligned}\hat{X}_\varepsilon(t) = \frac{1}{2}\int_{\Omega_{1\varepsilon}} |u'_{1\varepsilon}(t) - \Phi'(t)|^2\,dx + \int_{\Omega_{2\varepsilon}} |u'_{2\varepsilon}(t) - \Psi'(t)|^2\,dx \\ + \int_{\Omega_{1\varepsilon}} A^\varepsilon(\nabla u_{1\varepsilon}(t) - C^\varepsilon\nabla\Phi(t))(\nabla u_{1\varepsilon}(t) - C^\varepsilon\nabla\Phi(t))\,dx \\ + \int_{\Omega_{2\varepsilon}} A^\varepsilon\nabla u_{2\varepsilon}(t)\nabla u_{2\varepsilon}(t)\,dx\Big].\end{aligned}$$

Then,

$$\limsup_{\varepsilon\to 0} \big\|\hat{X}_\varepsilon\big\|_{C^0([0,T])} \le \big\|\hat{X}\big\|_{C^0([0,T])},$$

where

$$\begin{aligned}\hat{X}(t) = \frac{1}{2}\Big[\theta_1\|u'_1(t) - \Phi'(t)\|^2_{L^2(\Omega)} + \theta_2^{-1}\|u'_2(t) - \theta_2\Psi'(t)\|^2_{L^2(\Omega)} \\ + \int_\Omega A^0(\nabla u_1(t) - \nabla\Phi(t))(\nabla u_1(t) - \nabla\Phi(t))\,dx\Big].\end{aligned} \tag{6.1}$$

Remark 6.2. Here also (see Remark 5.2) we have only an inequality on the upper limit of $\|\hat{X}_\varepsilon\|_{C^0([0,T])}$. Due to the special form of the energy, to prove Proposition 6.1 we need to introduce more complex arguments than those used in the proof of Proposition 5.1.

Proof of Proposition 6.1. By the symmetry of A^ε and C^ε, one has

$$\begin{aligned}\hat{X}_\varepsilon(t) = \frac{1}{2}\Big[\int_{\Omega_{1\varepsilon}} |u'_{1\varepsilon}(t) - \Phi'(t)|^2\,dx + \int_{\Omega_{2\varepsilon}} |u'_{2\varepsilon}(t) - \Psi'(t)|^2\,dx \\ + \int_{\Omega_{1\varepsilon}} A^\varepsilon\nabla u_{1\varepsilon}(t)\nabla u_{1\varepsilon}(t)\,dx - 2\int_{\Omega_{1\varepsilon}} A^\varepsilon C^\varepsilon\nabla\Phi(t)\nabla u_{1\varepsilon}(t)\,dx \\ + \int_{\Omega_{1\varepsilon}} A^\varepsilon C^\varepsilon\nabla\Phi(t)C^\varepsilon\nabla\Phi(t)\,dx + \int_{\Omega_{2\varepsilon}} A^\varepsilon\nabla u_{2\varepsilon}(t)\nabla u_{2\varepsilon}(t)\,dx\Big].\end{aligned}$$

Decompose $\hat{X}_\varepsilon$ as

$$\hat{X}_\varepsilon = X_{1\varepsilon} - \hat{X}_{2\varepsilon} + \hat{X}_{3\varepsilon},$$

where, as in the previous case,

$$\begin{aligned}X_{1\varepsilon}(t) = \frac{1}{2}\Big[\int_{\Omega_{1\varepsilon}} |u'_{1\varepsilon}(t)|^2\,dx + \int_{\Omega_{2\varepsilon}} |u'_{2\varepsilon}(t)|^2\,dx \\ + \int_{\Omega_{1\varepsilon}} A^\varepsilon\nabla u_{1\varepsilon}(t)\nabla u_{1\varepsilon}(t)\,dx + \int_{\Omega_{2\varepsilon}} A^\varepsilon\nabla u_{2\varepsilon}(t)\nabla u_{2\varepsilon}(t)\,dx\Big],\end{aligned}$$

while

$$\begin{aligned}\hat{X}_{2\varepsilon}(t) = \int_{\Omega_{1\varepsilon}} u'_{1\varepsilon}(t)\Phi'(t)\,dx + \int_{\Omega_{2\varepsilon}} u'_{2\varepsilon}(t)\Psi'(t)\,dx \\ + \int_{\Omega_{1\varepsilon}} A^\varepsilon C^\varepsilon\nabla\Phi(t)\nabla u_{1\varepsilon}(t)\,dx\end{aligned}$$

and

$$\hat{X}_{3\varepsilon}(t) = \frac{1}{2}\left[\int_{\Omega_{1\varepsilon}} |\Phi'(t)|^2\, dx + \int_{\Omega_{2\varepsilon}} |\Psi'(t)|^2\, dx + \int_{\Omega_{1\varepsilon}} A^\varepsilon C^\varepsilon \nabla\Phi C^\varepsilon \nabla\Phi\, dx\right].$$

We proceed in several steps.

Step 1. Let us prove that

$$\begin{aligned}\limsup_{\varepsilon\to 0} X_{1\varepsilon}(t) \le \hat{X}_1 := \frac{1}{2}\Big[&\theta_1 \int_\Omega |u_1'(t)|^2\, dx + \theta_2^{-1}\int_\Omega |u_2'(t)|^2\, dx \\ &+ \int_\Omega A^0\nabla u_1(t)\nabla u_1(t)\, dx\Big] \quad \forall\, t\in[0,T].\end{aligned} \tag{6.2}$$

Since from (4.1) we have

$$X_{1\varepsilon} = E^\varepsilon - \frac{1}{2}\left(\varepsilon\int_{\Gamma^\varepsilon} h^\varepsilon |u_{1\varepsilon} - u_{2\varepsilon}|^2 d\sigma_x\right), \tag{6.3}$$

following the same ideas of [8] we show first that

$$\liminf_{\varepsilon\to 0}\left(\varepsilon\int_{\Gamma^\varepsilon} h^\varepsilon |u_{1\varepsilon}(t) - u_{2\varepsilon}(t)|^2 d\sigma_x\right) \ge c_h\theta_2^{-1}\int_\Omega |\theta_2 u_1(t) - u_2(t)|^2\, dx \tag{6.4}$$

for any $t\in[0,T]$.

To do that, we apply Lemma 3.4(ii) with $g = h$ and $v_\varepsilon = (P_1^\varepsilon u_{1\varepsilon} - u_{2\varepsilon})^2$ getting

$$\begin{aligned}&\liminf_{\varepsilon\to 0}\left(\varepsilon\int_{\Gamma^\varepsilon} h^\varepsilon |u_{1\varepsilon}(t) - u_{2\varepsilon}(t)|^2 d\sigma_x\right)\\ &\quad= \liminf_{\varepsilon\to 0}\left(\varepsilon\int_{\Gamma^\varepsilon} h^\varepsilon |P_1^\varepsilon u_{1\varepsilon}(t) - u_{2\varepsilon}(t)|^2 d\sigma_x\right)\\ &\quad= \liminf_{\varepsilon\to 0} c_h\int_{\Omega_{2\varepsilon}} |P_1^\varepsilon u_{1\varepsilon}(t) - u_{2\varepsilon}(t)|^2 dx \quad \forall\, t\in[0,T].\end{aligned} \tag{6.5}$$

Observe now that, thanks to (2.2), (2.20), and Remark 2.7(ii), we have

$$\widetilde{P_1^\varepsilon u_{1\varepsilon}}_{|\Omega_{2\varepsilon}}(t) - \widetilde{u_{2\varepsilon}}(t) = \chi_{|\Omega_{2\varepsilon}} P_1^\varepsilon u_{1\varepsilon}(t) - \widetilde{u_{2\varepsilon}}(t) \rightharpoonup \theta_2\left(u_1(t) - \theta_2^{-1}u_2(t)\right) \text{ weakly in } L^2(\Omega)$$

for any $t\in[0,T]$.

Hence, we can apply Lemma 3.5 with $p = 2$, $\mathcal{O}_\varepsilon = \Omega_{2\varepsilon}$, $\chi_0 = \theta_2$, $v_\varepsilon = P_1^\varepsilon u_{1\varepsilon}{}_{|\Omega_{2\varepsilon}} - u_{2\varepsilon}$, and $v = (u_1 - \theta_2^{-1}u_2)$. We have

$$\liminf_{\varepsilon\to 0} c_h\int_{\Omega_{2\varepsilon}} |P_1^\varepsilon u_{1\varepsilon}{}_{|\Omega_{2\varepsilon}}(t) - u_{2\varepsilon}(t)|^2 dx \;\ge c_h\int_\Omega \theta_2\left|u_1(t) - \theta_2^{-1}u_2(t)\right|^2\, dx\ \forall\, t\in[0,T],$$

which, together with (6.5), gives (6.4).

By (6.3), (6.4), and Theorem 4.4 we conclude that

$$\begin{aligned}\limsup_{\varepsilon\to 0} X_{1\varepsilon}(t) &= \limsup_{\varepsilon\to 0}\left[E^\varepsilon(t) - \frac{1}{2}\left(\varepsilon\int_{\Gamma^\varepsilon} h^\varepsilon |u_{1\varepsilon}(t)-u_{2\varepsilon}(t)|^2 d\sigma_x\right)\right]\\ &= \lim_{\varepsilon\to 0} E^\varepsilon(t) - \liminf_{\varepsilon\to 0}\left[\frac{1}{2}\left(\varepsilon\int_{\Gamma^\varepsilon} h^\varepsilon |u_{1\varepsilon}(t)-u_{2\varepsilon}(t)|^2 d\sigma_x\right)\right]\\ &\le \frac{1}{2}\Bigg[\theta_1\int_\Omega |u_1'(t)|^2\,dx + {\theta_2}^{-1}\int_\Omega |u_2'(t)|^2\,dx\\ &\quad + \int_\Omega A^0\nabla u_1(t)\nabla u_1(t)\,dx + c_h{\theta_2}^{-1}\int_\Omega |\theta_2 u_1(t)-u_2(t)|^2\,dx\\ &\quad - c_h{\theta_2}^{-1}\int_\Omega |\theta_2 u_1(t)-u_2(t)|^2\,dx\Bigg]\ \forall\, t\in[0,T],\end{aligned}$$

which proves (6.2).

Step 2. Observe now that, due to the presence of the boundary term in (6.3), we cannot have a priori estimates on the time derivative of $X_{1\varepsilon}$. Hence, we cannot deduce any compactness of $X_{1\varepsilon}$ in $C^0([0,T])$. To overcome this difficulty, we decompose $X_{1\varepsilon}$ as a sum of a compact part $E^\varepsilon - Y_\varepsilon$ and a rest $\hat{Y}_\varepsilon$, which goes to zero in $C^0([0,T])$ as $\varepsilon\to 0$. To do that we use Lemma 3.4(i) in (6.3) to obtain

$$X_{1\varepsilon} = E^\varepsilon - Y_\varepsilon - \hat{Y}_\varepsilon,$$

where

$$Y_\varepsilon(t) := c_h\int_{\Omega_{2\varepsilon}} |P_1^\varepsilon u_{1\varepsilon}(t) - u_{2\varepsilon}(t)|^2\,dx$$

and

$$\hat{Y}_\varepsilon(t) := \varepsilon\int_{\Omega_{2\varepsilon}} \nabla_y\psi_h(x/\varepsilon)\nabla_x |P_1^\varepsilon u_{1\varepsilon}(t) - u_{2\varepsilon}(t)|^2\,dx\ \forall\, t\ \in[0,T]. \tag{6.6}$$

Let us show first that

$$E^\varepsilon - Y_\varepsilon \text{ is compact in } C^0([0,T]). \tag{6.7}$$

Indeed, from Theorem 4.4 E^ε converges in $C^0([0,T])$. Moreover, Y_ε is bounded in $L^\infty(0,T)$ and

$$\left\|\frac{\partial Y_\varepsilon}{\partial t}\right\|_{L^\infty(0,T)} \le C\|P_1^\varepsilon u_{1\varepsilon}' - u_{2\varepsilon}'\|_{L^\infty(0,T;L^2(\Omega_{2\varepsilon}))}\|P_1^\varepsilon u_{1\varepsilon} - u_{2\varepsilon}\|_{L^\infty(0,T;L^2(\Omega_{2\varepsilon}))} \le C.$$

This gives (6.7).

To conclude this step, let us prove that

$$\hat{Y}_\varepsilon \to 0 \text{ in } C^0([0,T]). \tag{6.8}$$

Indeed,

$$\begin{aligned}&\varepsilon\left|\int_{\Omega_{2\varepsilon}} \nabla_y\psi_h(x/\varepsilon)\nabla_x |P_1^\varepsilon u_{1\varepsilon}(t) - u_{2\varepsilon}(t)|^2\,dx\right|\\ &\quad\le 2\varepsilon\|\nabla_y\psi_h\|_{L^\infty(\mathbb{R}^n)}\,\|\nabla(P_1^\varepsilon u_{1\varepsilon} - u_{2\varepsilon})\,(P_1^\varepsilon u_{1\varepsilon} - u_{2\varepsilon})\|_{L^\infty(0,T;L^1(\Omega_{2\varepsilon}))}\\ &\quad\le C_3\varepsilon\|P_1^\varepsilon u_{1\varepsilon} - u_{2\varepsilon}\|_{C^0(0,T;H^1(\Omega_{2\varepsilon}))} \le C\varepsilon\ \forall\, t\ \in[0,T]\end{aligned}$$

with C independent of ε and t. Therefore,

$$\varepsilon \left| \int_{\Omega_{2\varepsilon}} \nabla_y \psi_h(x/\varepsilon) \nabla_x (P_1^\varepsilon u_{1\varepsilon} - u_{2\varepsilon})^2 \, dx \right| \to 0 \text{ in } C^0([0,T]),$$

which gives (6.8).

Step 3. We prove here the convergences of $\hat{X}_{2\varepsilon}$ and $\hat{X}_{3\varepsilon}$ in $C^0([0,T])$.

Consider first $\hat{X}_{2\varepsilon}$. From (2.18), (5.1) written for $\lambda = e_i$, and (5.2) we have

(6.9)
$$\begin{aligned}
\hat{X}_{2\varepsilon}(t) &= \int_{\Omega_{1\varepsilon}} u'_{1\varepsilon}(t)\Phi'(t)\,dx + \int_{\Omega_{2\varepsilon}} u'_{2\varepsilon}(t)\Psi'(t)\,dx + \int_{\Omega_{1\varepsilon}} A^\varepsilon \nabla w_i^\varepsilon \nabla u_{1\varepsilon}(t) \frac{\partial \Phi}{\partial x_i}(t)\,dx \\
&= \int_\Omega \widetilde{u'_{1\varepsilon}}(t)\Phi'(t)\,dx + \int_\Omega \widetilde{u'_{2\varepsilon}}(t)\Psi'(t)\,dx + \int_{\Omega_{1\varepsilon}} A^\varepsilon \nabla w_i^\varepsilon \nabla \left[u_{1\varepsilon}(t) \frac{\partial \Phi}{\partial x_i}(t) \right] dx \\
&\quad - \int_{\Omega_{1\varepsilon}} A^\varepsilon \nabla w_i^\varepsilon \nabla \left[\frac{\partial \Phi}{\partial x_i}(t) \right] u_{1\varepsilon}(t)\,dx = \int_\Omega \widetilde{u'_{1\varepsilon}}(t)\Phi'(t)\,dx + \int_\Omega \widetilde{u'_{2\varepsilon}}(t)\Psi'(t)\,dx \\
&\quad - \int_\Omega \chi_{\Omega_{1\varepsilon}} A^\varepsilon \nabla w_i^\varepsilon \nabla \left[\frac{\partial \Phi}{\partial x_i}(t) \right] P_1^\varepsilon u_{1\varepsilon}(t)\,dx.
\end{aligned}$$

Due to Theorem 2.6, Remark 2.7(ii), and (5.3) one concludes that

(6.10)
$$\begin{aligned}
\hat{X}_{2\varepsilon} &\to \int_\Omega \theta_1 u'_1 \Phi'\,dx + \int_\Omega u'_2 \Psi'\,dx - \int_\Omega A^0 e_i \nabla \left[\frac{\partial \Phi}{\partial x_i} \right] u_1\,dx \\
&= \int_\Omega \theta_1 u'_1 \Phi'\,dx + \int_\Omega u'_2 \Psi'\,dx + \int_\Omega A^0 \nabla \Phi \nabla u_1\,dx \text{ in } \mathcal{D}'(0,T).
\end{aligned}$$

We have to prove that this convergence takes place in $C^0([0,T])$. From (6.9), Theorem 2.6, (5.2), and (5.3) there exists a constant C (independent of ε) such that

$$\|\hat{X}_{2\varepsilon}\|_{L^\infty(0,T)} \le C. \tag{6.11}$$

Let us show that

$$\left\| \frac{\partial \hat{X}_{2\varepsilon}}{\partial t} \right\|_{L^2(0,T)} \le C. \tag{6.12}$$

The variational formulation (2.8) written for $v_1 = \Phi$ and $v_2 = \Psi$ reads

$$\begin{aligned}
&\frac{\partial}{\partial t} \left(\int_\Omega u'_{1\varepsilon}(t)\Phi'(t)\,dx + \int_\Omega u'_{2\varepsilon}(t)\Psi'(t)\,dx \right) \\
&\quad = \int_\Omega \widetilde{u'_{1\varepsilon}}(t)\Phi''(t) + \int_\Omega \widetilde{u'_{2\varepsilon}}(t)\Psi''(t) - \int_{\Omega_{1\varepsilon}} A^\varepsilon \nabla u_{1\varepsilon}(t) \cdot \nabla \Phi'(t)\,dx \\
&\qquad - \int_{\Omega_{2\varepsilon}} A^\varepsilon \nabla u_{2\varepsilon}(t) \nabla \Psi'(t)\,dx - \varepsilon \int_{\Gamma^\varepsilon} h^\varepsilon (u_{1\varepsilon}(t) - u_{2\varepsilon}(t))(\Phi'(t) - \Psi'(t))\,d\sigma_x \\
&\qquad + \int_{\Omega_{1\varepsilon}} f_{1\varepsilon}(t)\Phi'(t)\,dx + \int_{\Omega_{2\varepsilon}} f_{2\varepsilon}(t)\Psi'(t)\,dx,
\end{aligned}$$

which is bounded in $L^2(0,T)$ due to Theorems 2.6 and (2.9)(i).

Moreover, we have

$$\frac{\partial}{\partial t}\int_\Omega \chi_{\Omega_{1\varepsilon}} A^\varepsilon \nabla w_i^\varepsilon \nabla \left[\frac{\partial \Phi}{\partial x_i}(t)\right] P_1^\varepsilon u_{1\varepsilon}(t)\,dx = \int_\Omega \chi_{\Omega_{1\varepsilon}} A^\varepsilon \nabla w_i^\varepsilon \nabla \left[\frac{\partial \Phi'}{\partial x_i}(t)\right] P_1^\varepsilon u_{1\varepsilon}(t)\,dx + \int_\Omega \chi_{\Omega_{1\varepsilon}} A^\varepsilon \nabla w_i^\varepsilon \nabla \left[\frac{\partial \Phi}{\partial x_i}(t)\right] P_1^\varepsilon u'_{1\varepsilon}(t)\,dx,$$

and the right-hand side of this equality is bounded in $L^2(0,T)$ in view of (5.3) and Theorem 2.6. Hence (6.12) holds; this, together with (6.11), shows that $\hat{X}_{2\varepsilon}$ is relatively compact in $C^0([0,T])$ so that from (6.10) we obtain

(6.13)
$$\hat{X}_{2\varepsilon} \to \hat{X}_2 := \int_\Omega \theta_1 u_1' \Phi'\,dx + \int_\Omega u_2' \Psi'\,dx + \int_\Omega A^0 \nabla\Phi\nabla u_1\,dx \text{ strongly in } C^0([0,T]).$$

Concerning $\hat{X}_{3\varepsilon}$, a similar argument to that used in the proof of Proposition 5.1 to show (5.11) gives

$$\hat{X}_{3\varepsilon} \to \hat{X}_3 := \frac{1}{2}\left[\theta_1\|\Phi'\|^2_{L^2(\Omega)} + \theta_2\|\Psi'\|^2_{L^2(\Omega)} + \int_\Omega A^0\nabla\Phi\nabla\Phi\,dx\right] \text{ strongly in } C^0([0,T]). \tag{6.14}$$

Step 4. Observe first that from (6.2), (6.13), and (6.14) we have

$$0 \le \limsup_{\varepsilon\to 0} \hat{X}_\varepsilon(t) \le \hat{X}_1(t) - \hat{X}_2(t) + \hat{X}_3(t) = \hat{X}(t) \quad \forall\, t\in[0,T], \tag{6.15}$$

where $\hat{X}$ is given by (6.1).

On the other hand, if we write $\hat{X}_\varepsilon$ as

$$\hat{X}_\varepsilon = Z_\varepsilon - \hat{Y}_\varepsilon, \tag{6.16}$$

where

$$Z_\varepsilon := E^\varepsilon - Y_\varepsilon - \hat{X}_{2\varepsilon} + \hat{X}_{3\varepsilon} \tag{6.17}$$

and $\hat{Y}_\varepsilon$ is given by (6.6), due to (6.8) we have

$$\limsup_{\varepsilon\to 0} \left\|\hat{X}_\varepsilon\right\|_{C^0([0,T])} \le \limsup_{\varepsilon\to 0} \|Z_\varepsilon\|_{C^0([0,T])} + \lim_{\varepsilon\to 0}\left|\hat{Y}_\varepsilon\right| = \limsup_{\varepsilon\to 0}\|Z_\varepsilon\|_{C^0([0,T])}. \tag{6.18}$$

Consequently, to conclude, it is enough to prove that

$$\limsup_{\varepsilon\to 0}\|Z_\varepsilon\|_{C^0([0,T])} \le \left\|\hat{X}\right\|_{C^0([0,T])}, \tag{6.19}$$

which together with (6.18) will give the claimed result.

Let $\{\varepsilon'\}$ be a subsequence such that

$$\limsup_{\varepsilon\to 0}\|Z_\varepsilon\|_{C^0([0,T])} = \lim_{\varepsilon'\to 0}\|Z_{\varepsilon'}\|_{C^0([0,T])}. \tag{6.20}$$

From (6.7), (6.13), and (6.14) it follows that Z_ε is compact in $C^0([0,T])$. Therefore, there exists a subsequence (still denoted $\{\varepsilon'\}$) such that

$$Z_{\varepsilon'} \to Z \text{ strongly in } C^0([0,T]). \tag{6.21}$$

But, from (6.16), (6.15), and (6.8) we have

$$0 \le Z(t) = \lim_{\varepsilon' \to 0} \left[\hat{X}_{\varepsilon'}(t) + \hat{Y}_{\varepsilon'}(t) \right] = \lim_{\varepsilon' \to 0} \hat{X}_{\varepsilon'}(t) \le X(t) \quad \forall\, t \ \in [0,T]$$

so that from (6.21) one has

$$\lim_{\varepsilon' \to 0} \|Z_{\varepsilon'}\|_{C^0([0,T])} = \|Z\|_{C^0([0,T])} \le \left\|\hat{X}\right\|_{C^0([0,T])},$$

which together with (6.20) gives (6.19) and ends the proof. □

Remark 6.3. Let us emphasize the main difficulties in the proof of Proposition 6.1. First, as in the previous case (see Remark 5.3), due to the presence of the boundary term in the energy, we have only $X_{1\varepsilon} \le E^\varepsilon$ and not $X_{1\varepsilon} = E^\varepsilon$, as usual. Moreover, here the limit of $E^\varepsilon - \hat{X}_{2\varepsilon} + \hat{X}_{3\varepsilon}$ is bigger than $\hat{X}$ so that we cannot conclude as in the previous case. On the other hand, since we have no estimates for the time derivative of the boundary term of $X_{1\varepsilon}$, we cannot derive any compactness of $X_{1\varepsilon}$ in $C^0([0,T])$. This is why we introduced in Step 4 of the proof a more technical argument, which relies on the decomposition (6.17) of $\hat{X}_\varepsilon$ as a sum of a compact term $Z_\varepsilon - Y_\varepsilon$ and a rest $\hat{Y}_\varepsilon$, which goes to zero in $C^0([0,T])$ as $\varepsilon \to 0$.

Proof of Theorem 2.8. Since $u_1 \in L^2(0,T;H_0^1(\Omega)) \cap C^0([0,T];L^2(\Omega))$, $u_1' \in C^0([0,T];L^2(\Omega))$, and $u_2' \in C^0([0,T];L^2(\Omega))$, from classical density results, we have that for any $\delta > 0$ there exist $\Phi, \Psi \in C^\infty([0,T];\mathcal{D}(\Omega))$ such that

$$\begin{cases} \text{(i)} \ \|u_1' - \Phi'\|_{C^0([0,T];L^2(\Omega))} \le \delta, \\ \text{(ii)} \ \left\|\theta_2^{-1} u_2' - \Psi'\right\|_{C^0([0,T];L^2(\Omega))} \le \delta, \\ \text{(iii)} \ \|\nabla u_1 - \nabla \Phi\|_{L^2(0,T;L^2(\Omega))} \le \delta. \end{cases} \tag{6.22}$$

Therefore,

$$\begin{aligned} &\|u_{1\varepsilon}' - u_1'\|^2_{C^0(0,T;L^2(\Omega_{1\varepsilon}))} + \left\|u_{2\varepsilon}' - \theta_2^{-1} u_2'\right\|^2_{C^0(0,T;L^2(\Omega_{2\varepsilon}))} \\ &\le 2 \Big(\|u_{1\varepsilon}' - \Phi'\|^2_{C^0(0,T;L^2(\Omega_{1\varepsilon}))} + \|\Phi' - u_1'\|^2_{C^0(0,T;L^2(\Omega_{1\varepsilon}))} \\ &\quad + \|u_{2\varepsilon}' - \Psi'\|^2_{C^0(0,T;L^2(\Omega_{2\varepsilon}))} + \left\|\Psi' - \theta_2^{-1} u_2'\right\|^2_{C^0(0,T;L^2(\Omega_{2\varepsilon}))} \Big) \\ &\le 2 \Big(\|u_{1\varepsilon}' - \Phi'\|^2_{C^0(0,T;L^2(\Omega_{1\varepsilon}))} + \|u_{2\varepsilon}' - \Psi'\|^2_{C^0(0,T;L^2(\Omega_{2\varepsilon}))} \Big) + 2\delta^2. \end{aligned} \tag{6.23}$$

The same argument used to prove (5.15) gives that there exists a constant C_1 such that

$$\begin{aligned} &\left(\|\nabla u_{1\varepsilon} - C^\varepsilon \nabla u_1\|^2_{C^0([0,T];[L^1(\Omega_{1\varepsilon})]^n)} + \|\nabla u_{2\varepsilon}\|^2_{C^0([0,T];[L^2(\Omega_{2\varepsilon})]^n)} \right) \\ &\le C_1 \left(\|\nabla u_{1\varepsilon} - C^\varepsilon \nabla \Phi\|^2_{C^0([0,T];[L^2(\Omega_{1\varepsilon})]^n)} + \|\nabla u_{2\varepsilon}\|^2_{C^0([0,T];[L^2(\Omega_{2\varepsilon})]^n)} + \delta^2 \right). \end{aligned} \tag{6.24}$$

On the other hand, from the properties of A^ε,

$$\begin{aligned} \frac{1}{2} \Big[&\|u_{1\varepsilon}'(t) - \Phi'(t)\|^2_{L^2(\Omega_{1\varepsilon})} + \|u_{2\varepsilon}'(t) - \Psi'(t)\|^2_{L^2(\Omega_{2\varepsilon})} \\ &+ \alpha \left(\|\nabla u_{1\varepsilon} - C^\varepsilon \nabla \Phi\|^2_{[L^2(\Omega_{1\varepsilon})]^n} + \|\nabla u_{2\varepsilon}\|^2_{[L^2(\Omega_{2\varepsilon})]^n} \right) \Big] \le \hat{X}_\varepsilon \quad \forall\, t \in [0,T] \end{aligned}$$

and there exists a constant C_2 such that

$$\left\|\hat{X}\right\|_{C^0([0,T])} \le C_2 \delta^2,$$

where $\hat{X}_\varepsilon$ and $\hat{X}$, defined in Proposition 6.1, are applied to the function Φ and Ψ, given by (6.22).

Therefore, from Proposition 6.1 we conclude

$$\begin{aligned}\limsup_{\varepsilon\to 0}\ &\left\{\frac{1}{2}\left[\|u'_{1\varepsilon}-\Phi'\|^2_{C^0([0,T];[L^2(\Omega_{1\varepsilon})])}+\|u'_{2\varepsilon}-\Psi'\|^2_{C^0([0,T];[L^2(\Omega_{2\varepsilon})])}\right]\right.\\ &\left.+\frac{\alpha}{2}\left(\|\nabla u_{1\varepsilon}-C^\varepsilon\nabla\Phi\|^2_{C^0([0,T];[L^2(\Omega_{1\varepsilon})]^n)}+\|\nabla u_{2\varepsilon}\|^2_{C^0([0,T];[L^2(\Omega_{2\varepsilon})]^n)}\right)\right\}\\ &\le\limsup_{\varepsilon\to 0}\left\|\hat{X}_\varepsilon\right\|_{C^0([0,T])}\le\left\|\hat{X}\right\|_{C^0([0,T])}\le C_2\delta^2,\end{aligned}$$

which, together with (6.23) and (6.24), gives (2.24), since δ is arbitrary. □

REFERENCES

[1] J.L. Auriault and H. Ene, *Macroscopic modelling of heat transfer in composites with interfacial thermal barrier,* Internat. J. Heat Mass Transfer, 37 (1994), pp. 2885–2892.

[2] S. Brahim-Otsmane, G.A. Francfort, and F. Murat, *Correctors for the homogenization of the wave and heat equations*, J. Math. Pure. Appl., 71 (1992), pp. 197–231.

[3] E. Canon and J.N. Pernin, *Homogenization of diffusion in composite media with interfacial barrier*, Rev. Roumaine Math. Pures Appl., 44 (1999), pp. 23–36.

[4] H.S. Carslaw and J.C. Jaeger, *Conduction of Heat in Solids*, Clarendon Press, Oxford, 1947.

[5] D. Cioranescu and P. Donato, *Exact internal controllability in perforated domains,* J. Math. Pures Appl., 68 (1989), pp. 185–213.

[6] D. Cioranescu and P. Donato, *An Introduction to Homogenization*, Oxford Lecture Ser. Math., Appl. 17, Oxford University Press, New York, 1999.

[7] D. Cioranescu and J. Saint Jean Paulin, *Homogenization in open sets with holes*, J. Math. Anal. Appl., 71 (1979), pp. 590–607.

[8] P. Donato, *Some corrector results for composites with imperfect interface*, Rend. Mat. Ser. VII, 26 (2006), pp. 189–209.

[9] P. Donato, L. Faella, and S. Monsurrò, *Homogenization of the wave equation in composites with imperfect interface: A memory effect*, J. Math. Pures Appl., 87 (2007), pp. 119–143.

[10] P. Donato and S. Monsurrò, *Homogenization of two heat conductors with interfacial contact resistance*, Anal. Appl., 2 (2004), pp. 247–273.

[11] P. Donato and A. Nabil, *Approximate controllability of linear parabolic equations in perforated domains*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 21–38.

[12] H. Ene and D. Polisevski, *Model of diffusion in partially fissured media*, Z. Angew. Math. Phys., 53 (2002), pp. 1052–1059.

[13] H.C. Hummel, *Homogenization for heat transfer in polycristals with interfacial resistances*, Appl. Anal., 75 (2000), pp. 403–424.

[14] J.L. Lions and E. Magenes, *Problèmes aux Limites Non Homogènes et Applications*, Dunod, Paris, 1968.

[15] R. Lipton, *Heat conduction in fine scale mixtures with interfacial contact resistance*, SIAM J. Appl. Math., 58 (1998), pp. 55–72.

[16] R. Lipton and B. Vernescu, *Composite with imperfect interface*, Proc. R. Soc. Lond. Ser. A, 452 (1996), pp. 329–358.

[17] S. Monsurrò, *Homogenization of a two-component composite with interfacial thermal barrier*, Adv. Math. Sci. Appl., 13 (2003), pp. 43–63.

[18] A. Nabil, *A Corrector result for the wave equations in perforated domains*, GAKUTO Internat. Ser., Math. Sci. Appl., 9 (1997), pp. 309–321.

[19] J.N. Pernin, *Homogénéisation d'un problème de diffusion en milieu composite à deux composantes*, C.R. Acad. Sci. Paris Sér. I, 321 (1995), pp. 949–952.

[20] E. Zeidler, *Nonlinear Functional Analysis and Its Applications*, II/B. *Nonlinear Monotone Operators*, Springer-Verlag, New York, 1990.

SIAM J. MATH. ANAL.
Vol. 40, No. 5, pp. 1979–1991

RELAXATION-TIME LIMIT IN THE ISOTHERMAL HYDRODYNAMIC MODEL FOR SEMICONDUCTORS*

JIANG XU†

Abstract. This work is concerned with the relaxation-time limit in the multidimensional isothermal hydrodynamic model for semiconductors in the critical Besov space. As the initial data are sufficiently close to equilibrium, the uniform (global) classical solutions are constructed by the high- and low-frequency decomposition methods. Furthermore, it is shown that the scaled classical solutions strongly converge towards that of a drift-diffusion model, as the relaxation time tends to zero.

Key words. relaxation-time limit, isothermal hydrodynamic model, semiconductors

AMS subject classifications. 35L45, 35B25, 35M20

DOI. 10.1137/080721893

1. Introduction and main results. In this work, we are interested in the isothermal hydrodynamic model for semiconductors, which is of the form

$$\begin{cases} n_t + \nabla \cdot (n\mathbf{u}) = 0, \\ (n\mathbf{u})_t + \nabla \cdot (n\mathbf{u} \otimes \mathbf{u}) + a^2 \nabla n = n \nabla \Phi - \frac{n\mathbf{u}}{\tau}, \\ \Delta \Phi = n - \bar{n}, \quad \Phi \to 0 \ \text{ as } \ |x| \to +\infty \end{cases} \tag{1.1}$$

for $(t, x) \in [0, +\infty) \times \mathbf{R}^N (N \geq 2)$. Here $n, \mathbf{u} = (u^1, u^2, \ldots, u^N)^\top (\top$ transpose), and Φ stand for the electron density, the electron velocity, and the electrostatic potential, respectively; $a > 0$ is the sound speed; and $0 < \tau \leq 1$ is the (small) momentum relaxation time for electrons. The symbols ∇, Δ, and $\otimes$ are the gradient operator, Laplacian operator, and the tensor products of two vectors, respectively. The positive constant $\bar{n}$ stands for the density of positively charged background ions.

The system (1.1) is supplemented with the initial data

$$(n, \mathbf{u})(x, 0) = (n_0, \mathbf{u}_0)(x), \quad x \in \mathbf{R}^N. \tag{1.2}$$

Concerning the small relaxation-time analysis, we define the scaled variables as in [12]:

$$(n^\tau, \mathbf{u}^\tau, \mathbf{e}^\tau)(x, s) = \Big(n, \frac{1}{\tau}\mathbf{u}, \mathbf{e}\Big)\Big(x, \frac{s}{\tau}\Big), \quad (\mathbf{e} = \nabla \Phi). \tag{1.3}$$

Then the new variables satisfy the equations

$$\begin{cases} n_s^\tau + \nabla \cdot (n^\tau \mathbf{u}^\tau) = 0, \\ \tau^2 \mathbf{u}_s^\tau + \tau^2 (\mathbf{u}^\tau \cdot \nabla)\mathbf{u}^\tau + \mathbf{u}^\tau = \mathbf{e}^\tau - a^2 \frac{\nabla n^\tau}{n^\tau}, \\ \nabla \cdot \mathbf{e}^\tau = n^\tau - \bar{n}, \end{cases} \tag{1.4}$$

with the initial data

$$(n^\tau, \tau \mathbf{u}^\tau)(x, 0) = (n_0, \mathbf{u}_0). \tag{1.5}$$

*Received by the editors April 22, 2008; accepted for publication September 2, 2008; published electronically January 21, 2009. This work was supported by NUAA's Scientific Fund for the Introduction of Qualified Personnel.

http://www.siam.org/journals/sima/40-5/72189.html

†Department of Mathematics, Nanjing University of Aeronautics and Astronautics, Nanjing 211100, People's Republic of China (jiangxu_79@nuaa.edu.cn, jiangxu_79@yahoo.com.cn).

Let $\tau \to 0$; then, formally, we arrive at the classical drift-diffusion model

$$\begin{cases} \mathcal{N}_s + \nabla \cdot (\mathcal{N}\mathcal{E} - a^2 \nabla \mathcal{N}) = 0, \\ \nabla \cdot \mathcal{E} = \mathcal{N} - \bar{n}, \\ \mathcal{N}(x, 0) = n_0. \end{cases} \tag{1.6}$$

This relaxation-time limit for the system (1.1)–(1.2) was first studied by Marcati and Natalini [12]. They obtained the uniform weak solutions with respect to τ and proved that the scaled weak solutions converged towards that of the drift-diffusion model (1.6). Subsequently, some rigorous results related to the relaxation limit have appeared (we refer the reader to [1, 3, 7, 8, 9, 10]); however, these results are restricted to one space dimension. Physically, it is more important and more interesting to study this asymptotic limit in several space dimensions; unfortunately, due to the serious difficulties in establishing the global existence of weak or smooth solutions to (1.1)–(1.2), up to now, only partial relaxation limit results are available in several space dimensions. In [11], Lattanzio and Marcati considered the three-dimensional isentropic hydrodynamic model, assuming the existence of L^∞-solutions in an τ-independent time interval, and justified the relaxation limit. Inspired by the Maxwell iteration, Yong [15] dealt with the periodic initial-value problem for the multidimensional isentropic hydrodynamic model in the Sobolev space $H^\ell(\mathbf{T}^N)$; however, the regularity index is required to be high ($\ell > 1 + N/2$, an integer).

Recently, in [6], we first established the global existence and exponential stability of classical solutions to the (isentropic and isothermal) hydrodynamic model for semiconductors in the critical Besov space $B^{1+N/2}_{2,1}(\mathbf{R}^N)$, but it is not clear whether the results are independent of the relaxation time τ. In this paper, we construct the uniform classical solutions (close to equilibrium) to the *isothermal* hydrodynamic model (1.1)–(1.2) and justify the above formal limit rigorously. For the proof of the global result, different from that in [6], we add some "new" ideas. More concretely speaking, we take full advantage of the special structure (skew-symmetry) of (1.1)–(1.2), which can help us avoid differentiating the system with respect to variable t once, and develop some "new" frequency-localization estimates; for details, see Lemmas 4.3, 4.4, and 4.5. In this sense, the proof of the global result is shortened heavily. Here, we first state the main results as follows.

THEOREM 1.1. *Let $\bar{n} > 0$ be a constant reference density. Suppose that $n_0 - \bar{n}, \boldsymbol{u}_0$, and $\boldsymbol{e}_0 \in B^\sigma_{2,1}(\mathbf{R}^N)(\sigma = 1 + N/2)$. There exists a positive constant δ_0 independent of τ, such that if*

$$\|(n_0 - \bar{n}, \boldsymbol{u}_0, \boldsymbol{e}_0)\|_{B^\sigma_{2,1}(\mathbf{R}^N)} \le \delta_0,$$

then the system (1.1)–(1.2) *admits a unique global solution $(n, \boldsymbol{u}, \boldsymbol{e})$ satisfying*

$$(n - \bar{n}, \boldsymbol{u}, \boldsymbol{e}) \in \mathcal{C}([0, \infty), B^\sigma_{2,1}(\mathbf{R}^N)).$$

Moreover, the uniform energy estimate holds:

$$\begin{aligned} &\sup_{t \ge 0} \Big(\|(n - \bar{n}, \boldsymbol{u}, \boldsymbol{e})(\cdot, t)\|_{B^\sigma_{2,1}(\mathbf{R}^N)} \Big) \\ &+ \mu_0 \int_0^\infty \Big(\tau \|(n - \bar{n}, \boldsymbol{e})(\cdot, t)\|_{B^\sigma_{2,1}(\mathbf{R}^N)} + \|\boldsymbol{u}(\cdot, t)\|_{B^\sigma_{2,1}(\mathbf{R}^N)} \Big) dt \\ &\le C_0 \|(n_0 - \bar{n}, \boldsymbol{u}_0, \boldsymbol{e}_0)\|_{B^\sigma_{2,1}(\mathbf{R}^N)}, \end{aligned} \tag{1.7}$$

where μ_0, C_0 are the positive constants independent of τ, and $\boldsymbol{e}_0 := \nabla \Delta^{-1}(n_0 - \bar{n})$.

Remark 1.1. From the energy estimate (1.7) and the smallness of τ $(0 < \tau \le 1)$, we can obtain the uniform exponential decay of classical solution $(n, \mathbf{u}, \mathbf{e})$ near to equilibrium $(\bar{n}, 0, 0)$ in [6]:

$$\|(n-\bar{n}, \mathbf{u}, \mathbf{e})(\cdot, t)\|_{B^{\sigma}_{2,1}(\mathbf{R}^N)} \le C_0 \|(n_0-\bar{n}, \mathbf{u}_0, \mathbf{e}_0)\|_{B^{\sigma}_{2,1}(\mathbf{R}^N)} \exp(-\mu_0 \tau t),\ t \ge 0.$$

Remark 1.2. To our knowledge, Theorem 1.1 cannot directly be applied to the generally isentropic hydrodynamic model for semiconductors. As a matter of fact, for the isentropic model, we can establish a similar relaxation-limit result in stronger Besov space $B^{\sigma+\epsilon}_{2,2}(\mathbf{R}^N)(\epsilon > 0)$.

Then, by considering an "$\mathcal{O}(1/\tau)$ time scale" in (1.3), we justify the following convergence to the drift-diffusion model (1.6) by the standard weak convergence methods and the application of the compactness theorem in [14].

THEOREM 1.2. *Let* $(n, \boldsymbol{u}, \boldsymbol{e})$ *be the global solution of* (1.1)–(1.2) *given by Theorem* 1.1. *Then there exists some function* $(\mathcal{N}, \mathcal{E})$ *which is a global weak solution to* (1.6) *satisfying* $(\mathcal{N}-\bar{n}, \mathcal{E}) \in \mathcal{C}([0,\infty), B^{\sigma}_{2,1}(\mathbf{R}^N))$ *such that as* $\tau \to 0$, *it yields the following* $(1 \le p < +\infty,\ \sigma' < \sigma)$*:*

$$(n, \boldsymbol{e})\Big(x, \frac{s}{\tau}\Big) \to (\mathcal{N}, \mathcal{E})(x, s) \quad \textit{strongly in}\ \ L^p(0, T; (B^{\sigma'}_{2,1}(\mathbf{R}^N))_{\mathrm{loc}}) \quad \textit{for any}\ T > 0.$$

Remark 1.3. Let us mention that this strong convergence result is *weaker* than that obtained in [1, 4], which can be regarded as a supplement to the theory of diffusive limit for hyperbolic problems.

The rest of this paper is organized as follows. In section 2, we present some definitions and basic facts on the Littlewood–Paley decomposition theory and Besov space. In section 3, we rewrite the isothermal hydrodynamic model as a symmetric hyperbolic system and recall the local existence result of classical solutions. Then in section 4, we establish the uniform a priori estimate, which is used to derive the global existence of uniform classical solutions. Finally, in section 5, we perform the relaxation-time limit.

Throughout this paper, $f \approx g$ means that $f \le Cg$ and $g \le Cf$, where $C > 0$ is a uniform constant with respect to τ. All functional spaces are considered in $\mathbf{R}^N$, so we may omit the space dependence for simplicity.

2. Littlewood–Paley analysis. In this section, we review briefly the Littlewood–Paley decomposition theory and the characterization of Besov space; see also, e.g., [2] or [6].

Let $\mathcal{S}$ be the Schwarz class. (φ, χ) is a couple of smooth functions valued in [0,1] such that φ is supported in the shell $\mathbf{C}(0, \frac{3}{4}, \frac{8}{3}) = \{\xi \in \mathbf{R}^N | \frac{3}{4} \le |\xi| \le \frac{8}{3}\}$, χ is supported in the ball $\mathbf{B}(0, \frac{4}{3}) = \{\xi \in \mathbf{R}^N | |\xi| \le \frac{4}{3}\}$, and

$$\chi(\xi) + \sum_{q=0}^{\infty} \varphi(2^{-q}\xi) = 1, \quad q \in \mathbf{Z},\ \ \xi \in \mathbf{R}^N.$$

For $f \in \mathcal{S}'$(denote the set of temperate distributions which is the dual of $\mathcal{S}$), one can define the nonhomogeneous dyadic blocks as follows:

$$\Delta_{-1} f := \chi(D) f = \tilde{h} * f \quad \text{with}\ \ \tilde{h} = \mathcal{F}^{-1}\chi,$$

$$\Delta_q f := \varphi(2^{-q}D) f = 2^{qN} \int h(2^q y) f(x-y) dy \quad \text{with}\ \ h = \mathcal{F}^{-1}\varphi\ \ \text{if}\ \ q \ge 0,$$

where $*$, $\mathcal{F}^{-1}$ represent the convolution operator and the inverse Fourier transform, respectively. The nonhomogeneous Littlewood–Paley decomposition is

$$f = \sum_{q\geq -1} \Delta_q f \quad \text{in} \quad \mathcal{S}'.$$

Define the low-frequency cut-off by

$$S_q f := \sum_{p\leq q-1} \Delta_p f.$$

Of course, $S_0 f = \Delta_{-1} f$. The above Littlewood–Paley decomposition is almost orthogonal in L^2.

PROPOSITION 2.1. *For any $f, g \in \mathcal{S}'$, the following properties hold:*

$$\Delta_p \Delta_q f \equiv 0 \quad \textit{if} \quad |p-q| \geq 2,$$

$$\Delta_q (S_{p-1} f \Delta_p g) \equiv 0 \quad \textit{if} \quad |p-q| \geq 5.$$

Besov space can be characterized in virtue of the Littlewood–Paley decomposition.

DEFINITION 2.2. *Let $1 \leq p \leq \infty$ and $s \in \mathbf{R}$. For $1 \leq r < \infty$, the Besov spaces $B^s_{p,r}$ are defined by*

$$f \in B^s_{p,r} \Leftrightarrow \left(\sum_{q\geq -1} (2^{qs} \|\Delta_q f\|_{L^p})^r \right)^{\frac{1}{r}} < \infty$$

and $B^s_{p,\infty}$ are defined by

$$f \in B^s_{p,\infty} \Leftrightarrow \sup_{q\geq -1} 2^{qs} \|\Delta_q f\|_{L^p} < \infty.$$

LEMMA 2.3 (Bernstein's inequality). *Let $k \in \mathbf{N}$ and $0 < R_1 < R_2$. There exists a constant C depending only on R_1, R_2, and N such that for all $1 \leq a \leq b \leq \infty$ and $f \in L^a$, we have*

$$\text{Supp } \mathcal{F} f \subset \boldsymbol{B}(0, R_1\lambda) \Rightarrow \sup_{|\alpha|=k} \|\partial^\alpha f\|_{L^b} \leq C^{k+1} \lambda^{k+N(\frac{1}{a}-\frac{1}{b})} \|f\|_{L^a},$$

$$\text{Supp } \mathcal{F} f \subset \boldsymbol{C}(0, R_1\lambda, R_2\lambda) \Rightarrow C^{-k-1}\lambda^k \|f\|_{L^a} \leq \sup_{|\alpha|=k} \|\partial^\alpha f\|_{L^a} \leq C^{k+1}\lambda^k \|f\|_{L^a}.$$

Here, $\mathcal{F} f$ (or $\widehat{f} = \int_{\mathbf{R}^N} f(x) \exp(-ix\cdot\xi) dx$) represents the Fourier transform on f.

There is a compactness result in Besov space, which we show in the following proposition.

PROPOSITION 2.4. *Let $1 \leq p, r \leq \infty$, $s \in \mathbf{R}$, and $\varepsilon > 0$. For all $\phi \in C^\infty_c$, the map $f \mapsto \phi f$ is compact from $B^{s+\varepsilon}_{p,r}$ to $B^s_{p,r}$.*

Finally, we state a result of continuity for the composition to end this section.

PROPOSITION 2.5. *Let $1 \leq p, r \leq \infty$, and I be an open interval of $\mathbf{R}$. Let $s > 0$ and let n be the smallest integer such that $n \geq s$. Let $F : I \to \mathbf{R}$ satisfy $F(0) = 0$ and $F' \in W^{n,\infty}(I; \mathbf{R})$. Assume that $v \in B^s_{p,r}$ takes values in $J \subset\subset I$. Then $F(v) \in B^s_{p,r}$ and there exists a constant C depending only on s, I, J, and N such that*

$$\|F(v)\|_{B^s_{p,r}} \leq C(1 + \|v\|_{L^\infty})^n \|F'\|_{W^{n,\infty}(I)} \|v\|_{B^s_{p,r}}.$$

3. Reformulation and local existence. Let us introduce the enthalpy $\mathcal{H}(\varrho) = a^2 \ln \varrho$ $(\varrho > 0)$, and set

$$m(t,x) = (\mathcal{H}(n(t,x)) - \mathcal{H}(\bar{n}))/a. \tag{3.1}$$

Then (1.1) can be transformed into the symmetric hyperbolic form

$$\partial_t U + \sum_{j=1}^{N} A_j(\mathbf{u}) \partial_{x_j} U = \begin{pmatrix} 0 \\ -\frac{1}{\tau}\mathbf{u} + \mathbf{e} \end{pmatrix}, \tag{3.2}$$

coupled with the dynamic electron field equation

$$\mathrm{div}\mathbf{e} = h(m), \tag{3.3}$$

where

$$U = \begin{pmatrix} m \\ \mathbf{u} \end{pmatrix}, \qquad A_j(\mathbf{u}) = \begin{pmatrix} u^j & ae_j^\top \\ ae_j & u^j I_N \end{pmatrix}$$

(I_N denotes the unit matrix of order N),

and $h(m) = \bar{n}\{\exp(m/a) - 1\}$ is a smooth function on the domain $\{m | -\infty < m < +\infty\}$ satisfying $h(0) = 0$. The initial data (1.2) become

$$(m, \mathbf{u}, \mathbf{e})|_{t=0} = (a(\ln n_0 - \ln \bar{n}), \mathbf{u}_0, \mathbf{e}_0). \tag{3.4}$$

Remark 3.1. The variable change is from the open set $\{(n, \mathbf{u}, \mathbf{e}) \in (0, +\infty) \times \mathbf{R}^N \times \mathbf{R}^N\}$ to the whole space $\{(m, \mathbf{u}, \mathbf{e}) \in \mathbf{R} \times \mathbf{R}^N \times \mathbf{R}^N\}$. It is easy to show that for classical solutions $(n, \mathbf{u}, \mathbf{e})$ away from the vacuum, (1.1)–(1.2) is equivalent to (3.2)–(3.4).

Now, we recall a local existence and uniqueness result of the classical solutions to (3.2)–(3.4) which has been established in [6].

PROPOSITION 3.1. *For any fixed relaxation time $\tau > 0$, assume that $(m_0, \boldsymbol{u}_0, \boldsymbol{e}_0) \in B^\sigma_{2,1}$; then, there exist a time $T_0 > 0$ (depending only on the initial data) and a unique solution $(m, \boldsymbol{u}, \boldsymbol{e})$ to (3.2)–(3.4) such that $(m, \boldsymbol{u}, \boldsymbol{e}) \in \mathcal{C}^1([0, T_0] \times \mathbf{R}^N)$ and $(m, \boldsymbol{u}, \boldsymbol{e}) \in \mathcal{C}([0, T_0], B^\sigma_{2,1}) \cap \mathcal{C}^1([0, T_0], B^{\sigma-1}_{2,1})$.*

4. A uniform estimate and global existence. In this section, we shall establish a uniform a priori estimate, which is used to derive the global existence of classical solutions to (3.2)–(3.4).

PROPOSITION 4.1. *There exist three positive constants δ_1, C_1, and μ_1 independent of τ such that for any $T > 0$, if*

$$\sup_{0 \le t \le T} \|(m, \boldsymbol{u}, \boldsymbol{e})(\cdot, t)\|_{B^\sigma_{2,1}} \le \delta_1, \tag{4.1}$$

then

$$\begin{aligned} &\|(m, \boldsymbol{u}, \boldsymbol{e})(\cdot, t)\|_{B^\sigma_{2,1}} + \mu_1 \int_0^t \Big(\tau \|(m, \boldsymbol{e})(\cdot, \varsigma)\|_{B^\sigma_{2,1}} + \|\boldsymbol{u}(\cdot, \varsigma)\|_{B^\sigma_{2,1}}\Big) d\varsigma \\ &\le C_1 \|(m, \boldsymbol{u}, \boldsymbol{e})(\cdot, 0)\|_{B^\sigma_{2,1}}, \qquad t \in [0, T]. \end{aligned} \tag{4.2}$$

The main ingredients in the proof of Proposition 4.1 are the high-frequency ($q \ge 0$) estimates and low-frequency ($q = -1$) estimates on $(m, \mathbf{u}, \mathbf{e})$. To do this, we need to establish some lemmas.

LEMMA 4.2. *If* $(m, \boldsymbol{u}, \boldsymbol{e}) \in \mathcal{C}([0,T], B^{\sigma}_{2,1}) \cap \mathcal{C}^1([0,T], B^{\sigma-1}_{2,1})$ *is a solution of* (3.2)–(3.4) *for any given* $T > 0$*, then the following estimate holds* ($q \geq -1$)*:*

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\Big(\|\Delta_q m\|_{L^2}^2 + \|\Delta_q \boldsymbol{u}\|_{L^2}^2 + \frac{1}{\bar{n}}\|\Delta_q \boldsymbol{e}\|_{L^2}^2\Big) + \frac{1}{\tau}\|\Delta_q \boldsymbol{u}\|_{L^2}^2 \\
&\leq \frac{1}{2}\|\nabla \boldsymbol{u}\|_{L^\infty}(\|\Delta_q m\|_{L^2}^2 + \|\Delta_q \boldsymbol{u}\|_{L^2}^2) + \|[\boldsymbol{u}, \Delta_q]\cdot\nabla m\|_{L^2}\|\Delta_q m\|_{L^2} \\
&\quad + \|[\boldsymbol{u}, \Delta_q]\cdot\nabla \boldsymbol{u}\|_{L^2}\|\Delta_q \boldsymbol{u}\|_{L^2} + \frac{1}{\bar{n}}\|\Delta_q(h(m)\boldsymbol{u})\|_{L^2}\|\Delta_q \boldsymbol{e}\|_{L^2}, \qquad (4.3)
\end{aligned}$$

where the commutator $[f, g] = fg - gf$*.*

Remark 4.1. By applying the operator Δ_q to (3.2), then multiplying the resulting equations by the conjugator of $\Delta_q m(\overline{\Delta_q m})$ and $\overline{\Delta_q \mathbf{u}}$, respectively, we can directly achieve (4.3) without extra trouble; see [6].

Below, we formulate an important skew-symmetry lemma which was developed in [4, 13, 16].

LEMMA 4.3 (Shizuta–Kawashima). *For all* $\xi \in \mathbf{R}^N$, $\xi \neq 0$*, the system* (3.2) *admits a real skew-symmetric smooth matrix* $K(\xi)$ *which is defined in the unit sphere* $\boldsymbol{S}^{N-1}$*:*

$$K(\xi) = \begin{pmatrix} 0 & \frac{\xi^\top}{|\xi|} \\ -\frac{\xi}{|\xi|} & 0 \end{pmatrix}, \qquad (4.4)$$

and then

$$K(\xi)\sum_{j=1}^{N}\xi_j A_j(0) = \begin{pmatrix} a|\xi| & 0 \\ 0 & -a\frac{\xi\otimes\xi}{|\xi|} \end{pmatrix}. \qquad (4.5)$$

Thanks to the skew-symmetry of the system (3.2), we can develop some "new" frequency-localization estimates and avoid performing the t-derivative to (3.2), which is different from the estimates in [6].

LEMMA 4.4. *If* $(m, \boldsymbol{u}, \boldsymbol{e}) \in \mathcal{C}([0,T], B^{\sigma}_{2,1}) \cap \mathcal{C}^1([0,T], B^{\sigma-1}_{2,1})$ *is a solution of* (3.2)–(3.4) *for any given* $T > 0$*, then the following estimates hold:*

$$\begin{aligned}
&\frac{\tau}{2}\frac{d}{dt}\mathrm{Im}\int|\xi|\Big((\widehat{\Delta_q U})^* K(\xi)\widehat{\Delta_q U}\Big)d\xi + \tau\frac{a}{2}2^{2q}\|\Delta_q m\|_{L^2}^2 \\
&\leq \frac{C}{\tau}2^{2q}\|\Delta_q \boldsymbol{u}\|_{L^2}^2 + C\tau 2^q\|\Delta_q U\|_{L^2}\|\Delta_q \mathcal{G}\|_{L^2} \\
&\quad + C\tau\|\Delta_q(\tilde{h}(m)m)\|_{L^2}\|\Delta_q m\|_{L^2}\ (q \geq 0); \qquad (4.6)
\end{aligned}$$

$$\begin{aligned}
&\frac{\tau}{2}\frac{d}{dt}\mathrm{Im}\int|\xi|\Big((\widehat{\Delta_{-1} U})^* K(\xi)\widehat{\Delta_{-1} U}\Big)d\xi + \tau\frac{\bar{n}}{a}\|\Delta_{-1} m\|_{L^2}^2 \\
&\leq \frac{C}{\tau}\|\Delta_{-1}\boldsymbol{u}\|_{L^2}^2 + C\tau\|\Delta_{-1} U\|_{L^2}\|\Delta_{-1}\mathcal{G}\|_{L^2} \\
&\quad + C\tau\|\Delta_{-1}(\tilde{h}(m)m)\|_{L^2}\|\Delta_{-1} m\|_{L^2}, \qquad (4.7)
\end{aligned}$$

where the function $\mathcal{G}$ *is given in* (4.9)*,* $\tilde{h}(m) = \int_0^1 h'(\varsigma m)d\varsigma - \frac{\bar{n}}{a}$ *is a smooth function on* $\{m| -\infty < \varsigma m < \infty,\ \varsigma \in [0,1]\}$ *satisfying* $\tilde{h}(0) = 0$*, and* $C > 0$ *is a uniform constant independent of* τ*.*

Proof. The system (3.2) can be written as the linearized form

$$\partial_t U + \sum_{j=1}^{N} A_j(0)\partial_{x_j} U = \mathcal{G} + \begin{pmatrix} 0 \\ -\frac{1}{\tau}\mathbf{u} + \mathbf{e} \end{pmatrix}, \tag{4.8}$$

where

$$\mathcal{G} = \sum_{j=1}^{N} \Big\{ A_j(0) - A_j(\mathbf{u}) \Big\} \partial_{x_j} U. \tag{4.9}$$

Applying the operator Δ_q to the system (4.8) gives

$$\partial_t \Delta_q U + \sum_{j=1}^{N} A_j(0)\partial_{x_j} \Delta_q U = \Delta_q \mathcal{G} + \begin{pmatrix} 0 \\ -\frac{1}{\tau}\Delta_q \mathbf{u} + \Delta_q \mathbf{e} \end{pmatrix}. \tag{4.10}$$

By performing the Fourier transform with respect to the space variable x for (4.10) and multiplying the resulting equation by $-i\tau(\widehat{\Delta_q U})^* K(\xi)$(* represents transpose and conjugator), then taking the real part of each term in the equality, we can obtain

$$\begin{aligned} &\tau \mathrm{Im}\Big((\widehat{\Delta_q U})^* K(\xi) \frac{d}{dt} \widehat{\Delta_q U} \Big) + \tau (\widehat{\Delta_q U})^* K(\xi) \Big(\sum_{j=1}^{N} \xi_j A_j(0) \Big) \widehat{\Delta_q U} \\ &= \tau \mathrm{Im}\Big((\widehat{\Delta_q U})^* K(\xi) (\widehat{\Delta_q \mathcal{G}}) \Big) - \mathrm{Im}\Big((\overline{\widehat{\Delta_q m}}) \frac{\xi^\top}{|\xi|} \widehat{\Delta_q \mathbf{u}} \Big) + \tau \mathrm{Im}\Big((\overline{\widehat{\Delta_q m}}) \frac{\xi^\top}{|\xi|} \widehat{\Delta_q \mathbf{e}} \Big). \end{aligned} \tag{4.11}$$

Using the skew-symmetry of $K(\xi)$, we have

$$\mathrm{Im}\Big((\widehat{\Delta_q U})^* K(\xi) \frac{d}{dt} \widehat{\Delta_q U} \Big) = \frac{1}{2} \frac{d}{dt} \mathrm{Im}\Big((\widehat{\Delta_q U})^* K(\xi) \widehat{\Delta_q U} \Big). \tag{4.12}$$

Substituting (4.5) into the second term on the left-hand side of (4.11), it is not difficult to get

$$\begin{aligned} &\tau \mathrm{Im}\Big((\widehat{\Delta_q U})^* K(\xi) \frac{d}{dt} \widehat{\Delta_q U} \Big) + \tau (\widehat{\Delta_q U})^* K(\xi) \Big(\sum_{j=1}^{N} \xi_j A_j(0) \Big) \widehat{\Delta_q U} \\ &\geq \frac{\tau}{2} \frac{d}{dt} \mathrm{Im}\Big((\widehat{\Delta_q U})^* K(\xi) \widehat{\Delta_q U} \Big) + a\tau |\xi| |\widehat{\Delta_q U}|^2 - 2a\tau |\xi| |\widehat{\Delta_q \mathbf{u}}|^2. \end{aligned} \tag{4.13}$$

With the help of Young's inequality, the right-hand side of (4.11) can be estimated as

$$\begin{aligned} &\tau \mathrm{Im}\Big((\widehat{\Delta_q U})^* K(\xi) (\widehat{\Delta_q \mathcal{G}}) \Big) - \mathrm{Im}\Big((\overline{\widehat{\Delta_q m}}) \frac{\xi^\top}{|\xi|} \widehat{\Delta_q \mathbf{u}} \Big) + \tau \mathrm{Im}\Big((\overline{\widehat{\Delta_q m}}) \frac{\xi^\top}{|\xi|} \widehat{\Delta_q \mathbf{e}} \Big) \\ &\leq \tau \frac{a}{2} |\xi| |\widehat{\Delta_q U}|^2 + \frac{C}{\tau |\xi|} |\widehat{\Delta_q \mathbf{u}}|^2 + \tau |\widehat{\Delta_q U}| |\widehat{\Delta_q \mathcal{G}}| + \tau \mathrm{Im}\Big((\overline{\widehat{\Delta_q m}}) \frac{\xi^\top}{|\xi|} \widehat{\Delta_q \mathbf{e}} \Big), \end{aligned} \tag{4.14}$$

where we have used the uniform boundedness of the matrix $K(\xi)(\xi \neq 0)$; the positive constant C is independent of τ. Combining this with the equality (4.11) and the inequality (4.13)–(4.14), we deduce that

$$\begin{aligned} &\frac{\tau}{2} \frac{d}{dt} \mathrm{Im}\Big((\widehat{\Delta_q U})^* K(\xi) \widehat{\Delta_q U} \Big) + \tau \frac{a}{2} |\xi| |\widehat{\Delta_q U}|^2 \\ &\leq \frac{C}{\tau} \Big(|\xi| + \frac{1}{|\xi|} \Big) |\widehat{\Delta_q \mathbf{u}}|^2 + \tau |\widehat{\Delta_q U}| |\widehat{\Delta_q \mathcal{G}}| + \tau \mathrm{Im}\Big((\overline{\widehat{\Delta_q m}}) \frac{\xi^\top}{|\xi|} \widehat{\Delta_q \mathbf{e}} \Big). \end{aligned} \tag{4.15}$$

Multiplying (4.15) by $|\xi|$ and integrating it over $\mathbf{R}^N$, from Plancherel's theorem, we obtain

$$\begin{aligned}&\frac{\tau}{2}\frac{d}{dt}\mathrm{Im}\int|\xi|\Big((\widehat{\Delta_q U})^*K(\xi)\widehat{\Delta_q U}\Big)d\xi+\tau\frac{a}{2}\|\Delta_q\nabla U\|_{L^2}^2\\ &\le\frac{C}{\tau}2^{2q}\|\Delta_q\mathbf{u}\|_{L^2}^2+C\tau 2^q\|\Delta_q U\|_{L^2}\|\Delta_q\mathcal{G}\|_{L^2}+\tau\mathrm{Im}\int\Big((\overline{\widehat{\Delta_q m}})\xi^\top\widehat{\Delta_q\mathbf{e}}\Big)d\xi.\end{aligned}\tag{4.16}$$

For the third term on the right-hand side of (4.16), we have

$$\tau\mathrm{Im}\int\Big((\overline{\widehat{\Delta_q m}})\xi^\top\widehat{\Delta_q\mathbf{e}}\Big)d\xi=\tau\Big(J+\bar J\Big),\tag{4.17}$$

where

$$\begin{aligned}J+\bar J&=-\frac{i}{2}\int\Big((\overline{\widehat{\Delta_q m}})\xi^\top\widehat{\Delta_q\mathbf{e}}\Big)d\xi+\frac{i}{2}\int\Big((\widehat{\Delta_q m})\xi^\top\overline{\widehat{\Delta_q\mathbf{e}}}\Big)d\xi\\ &=\frac{1}{2}\int(\overline{\widehat{\Delta_q\nabla m}})\cdot\widehat{\Delta_q\mathbf{e}}d\xi+\frac{1}{2}\int(\widehat{\Delta_q\nabla m})\cdot\overline{\widehat{\Delta_q\mathbf{e}}}d\xi\\ &=\frac{1}{2}(2\pi)^N\Big(\int\overline{\Delta_q\nabla m}\cdot\Delta_q\mathbf{e}dx+\int\Delta_q\nabla m\cdot\overline{\Delta_q\mathbf{e}}dx\Big)\\ &=-\frac{1}{2}(2\pi)^N\Big(\int\overline{\Delta_q m}\Delta_q\mathrm{div}\mathbf{e}dx+\int\Delta_q m\overline{\Delta_q\mathrm{div}\mathbf{e}}dx\Big)\\ &=-\frac{1}{2}(2\pi)^N\Big(\int\overline{\Delta_q m}\Delta_q(h(m)-h(0))dx+\int\Delta_q m\overline{\Delta_q(h(m)-h(0))}dx\Big)\\ &=-\frac{(2\pi)^N\bar n}{a}\|\Delta_q m\|_{L^2}^2-\frac{(2\pi)^N}{2}\Big(\int\overline{\Delta_q m}\Delta_q(\tilde h(m)m)dx\\ &\qquad+\int\Delta_q m\overline{\Delta_q(\tilde h(m)m)}dx\Big).\end{aligned}\tag{4.18}$$

Here, $\tilde h(m)=\int_0^1 h'(\varsigma m)d\varsigma-\bar n/a$ is a smooth function on $\{m|-\infty<\varsigma m<\infty,\ \varsigma\in[0,1]\}$ satisfying $\tilde h(0)=0$. Therefore, from (4.16)–(4.18), we arrive at

$$\begin{aligned}&\frac{\tau}{2}\frac{d}{dt}\mathrm{Im}\int|\xi|\Big((\widehat{\Delta_q U})^*K(\xi)\widehat{\Delta_q U}\Big)d\xi+\tau\frac{a}{2}\|\Delta_q\nabla U\|_{L^2}^2+\tau\frac{(2\pi)^N\bar n}{a}\|\Delta_q m\|_{L^2}^2\\ &\le\frac{C}{\tau}2^{2q}\|\Delta_q\mathbf{u}\|_{L^2}^2+C\tau 2^q\|\Delta_q U\|_{L^2}\|\Delta_q\mathcal{G}\|_{L^2}+C\tau\|\Delta_q(\tilde h(m)m)\|_{L^2}\|\Delta_q m\|_{L^2}.\end{aligned}\tag{4.19}$$

In view of Lemma 2.3,

$$\|\Delta_q\nabla m\|_{L^2}\approx 2^q\|\Delta_q m\|_{L^2}\ (q\ge 0)$$

follows, so we get the estimates (4.6) and (4.7) immediately. □

From the proof of Lemma 4.4, we see that the Poisson equation remedies the low-frequency estimate on $\|\Delta_{-1}m\|_{L^2}^2$, which plays a key role in the uniform exponential decay of classical solutions (see Remark 1.1). This is essentially different from the Euler equations studied in [5]. On the electron field $\mathbf{e}$, we also have some "new" a priori estimates.

LEMMA 4.5. *If $(m,\boldsymbol{u},\boldsymbol{e})\in\mathcal{C}([0,T],B^\sigma_{2,1})\cap\mathcal{C}^1([0,T],B^{\sigma-1}_{2,1})$ is a solution of (3.2)–(3.4) for any given $T>0$, then*

$$2^q\|\Delta_q\boldsymbol{e}\|_{L^2}^2\le C\Big(\frac{\bar n}{a}\|\Delta_q m\|_{L^2}+\|\Delta_q(\tilde h(m)m)\|_{L^2}\Big)\|\Delta_q\boldsymbol{e}\|_{L^2}\ (q\ge 0);\tag{4.20}$$

$$
\begin{aligned}
&-\frac{d}{dt}\int \Delta_{-1}\boldsymbol{e}\cdot\overline{\Delta_{-1}\boldsymbol{u}}dx+\|\Delta_{-1}\boldsymbol{e}\|_{L^2}^2\\
&\le C(\bar n\|\Delta_{-1}\boldsymbol{u}\|_{L^2}+\|\Delta_{-1}(h(m)\boldsymbol{u})\|_{L^2})\|\Delta_{-1}\boldsymbol{u}\|_{L^2}+C\Big(a\|\Delta_{-1}\nabla m\|_{L^2}\\
&+\frac{1}{\tau}\|\Delta_{-1}\boldsymbol{u}\|_{L^2}+\|\boldsymbol{u}\|_{L^\infty}\|\Delta_{-1}\nabla\boldsymbol{u}\|_{L^2}+\|[\boldsymbol{u},\Delta_{-1}]\cdot\nabla\boldsymbol{u}\|_{L^2}\Big)\|\Delta_{-1}\boldsymbol{e}\|_{L^2},
\end{aligned}
\tag{4.21}
$$

where $C>0$ is a uniform constant independent of τ.

Proof. (I) By applying the operator Δ_q to both sides of div$\mathbf{e}=h(m)$ $(q\ge 0)$, integrating it over $\mathbf{R}^N$ after multiplying $\Delta_q$div$\mathbf{e}$, and noticing the irrotationality of $\mathbf{e}$, we can obtain (4.20) in virtue of Hölder's inequality.
(II) From (1.1) and (3.1), we get

$$
\mathbf{e}_t=-\nabla\Delta^{-1}\nabla\cdot\{h(m)\mathbf{u}+\bar n\mathbf{u}\},
\tag{4.22}
$$

where the nonlocal term $\nabla\Delta^{-1}\nabla\cdot f$ is the product of Riesz transforms on f. From (3.2) and (4.22), we have

$$
\begin{aligned}
&-\frac{d}{dt}\int\Delta_q\mathbf{e}\cdot\overline{\Delta_q\mathbf{u}}dx\\
&=-\int\Delta_q\mathbf{e}_t\overline{\Delta_q\mathbf{u}}dx-\int\Delta_q\mathbf{e}\overline{\Delta_q\mathbf{u}_t}dx\\
&=\int\nabla\Delta^{-1}\nabla\cdot\Delta_q\{h(m)\mathbf{u}+\bar n\mathbf{u}\}\overline{\Delta_q\mathbf{u}}dx\\
&-\int\Delta_q\mathbf{e}\cdot\Big(-a\overline{\Delta_q\nabla m}-\frac{1}{\tau}\overline{\Delta_q\mathbf{u}}-\mathbf{u}\overline{\Delta_q\nabla\mathbf{u}}+\overline{[\mathbf{u},\Delta_q]\cdot\nabla\mathbf{u}}+\overline{\Delta_q\mathbf{e}}\Big)dx.
\end{aligned}
\tag{4.23}
$$

Using the L^2-boundedness of the Riesz transform and Hölder's inequality, we derive (4.21) immediately. □

For those estimates of the commutators in (4.3) and (4.21), we have the following fact.

LEMMA 4.6 (see [6]). *Let $s>0$ and $1<p<\infty$; then the following inequalities are true:*

$$
\begin{aligned}
&2^{qs}\|[f,\Delta_q]\mathcal{A}g\|_{L^p}\\
&\le\begin{cases}
Cc_q\|f\|_{B^s_{p,1}}\|g\|_{B^s_{p,1}}, & f,g\in B^s_{p,1},\ s=1+N/p,\\
Cc_q\|f\|_{B^s_{p,1}}\|g\|_{B^{s+1}_{p,1}}, & f\in B^s_{p,1},\ \ g\in B^{s+1}_{p,1},\ s=N/p,\\
Cc_q\|f\|_{B^{s+1}_{p,1}}\|g\|_{B^s_{p,1}}, & f\in B^{s+1}_{p,1},\ \ g\in B^s_{p,1},\ s=N/p.
\end{cases}
\end{aligned}
\tag{4.24}
$$

In particular, if $f=g$, then

$$
2^{qs}\|[f,\Delta_q]\mathcal{A}g\|_{L^p}\le Cc_q\|\nabla f\|_{L^\infty}\|g\|_{B^s_{p,1}},\ s>0,
\tag{4.25}
$$

where the operator $\mathcal{A}=\operatorname{div}$ or ∇, C is a harmless constant, and c_q denotes a sequence such that $\|(c_q)\|_{l^1}\le 1$.

In addition, in the proof of Proposition 4.1, an elementary algebra inequality will be used. For clarity, we put it into a lemma.

LEMMA 4.7. *Assume that $f,g,h>0$ and $0<\tau\le 1$; then the following inequality is true:*

$$
f+\tau g+\tau h\le C\frac{\frac{1}{\tau}f^2+\tau g^2+\tau h^2}{(f^2+g^2+h^2)^{1/2}},
\tag{4.26}
$$

where $C>0$ is a uniform constant independent of τ.

Proof. Note that

$$(f^2+g^2+h^2)^{1/2}\approx f+g+h. \tag{4.27}$$

We need only show

$$\Big(f+\tau g+\tau h\Big)(f+g+h)\le C\bigg(\frac{1}{\tau}f^2+\tau g^2+\tau h^2\bigg), \tag{4.28}$$

which is obvious in virtue of the smallness of $\tau(0<\tau\le 1)$ and Young's inequality. □

Proof of Proposition 4.1. From the a priori estimate assumption (4.1), we have

$$\sup_{0\le t\le T}(\|(m,\mathbf{u},\mathbf{e})(\cdot,t)||_{W^{1,\infty}})\le C\delta_1. \tag{4.29}$$

In what follows, we are going to divide the proof into some lemmas.

LEMMA 4.8 ($q\ge 0$). *There exist some positive constants K_1,K_2,μ_2 independent of τ such that the following estimate holds:*

$$\begin{aligned}
&2^{q(\sigma-1)}\frac{d}{dt}\bigg\{\frac{K_1}{2}2^{2q}\bigg(\|\Delta_q m\|_{L^2}^2+\|\Delta_q\boldsymbol{u}\|_{L^2}^2+\frac{1}{\bar n}\|\Delta_q\boldsymbol{e}\|_{L^2}^2\bigg)\\
&+\frac{K_2\tau}{2}\mathrm{Im}\int|\xi|\Big((\widehat{\Delta_q U})^*K(\xi)\widehat{\Delta_q U}\Big)d\xi\bigg\}^{1/2}\\
&+\mu_2 2^{q\sigma}\Big(\tau\|\Delta_q m\|_{L^2}+\|\Delta_q\boldsymbol{u}\|_{L^2}+\tau\|\Delta_q\boldsymbol{e}\|_{L^2}\Big)\\
&\le C\{2^{q\sigma}\|\nabla\boldsymbol{u}\|_{L^\infty}(\|\Delta_q m\|_{L^2}+\|\Delta_q\boldsymbol{u}\|_{L^2})+c_q\|\boldsymbol{u}\|_{B^\sigma_{2,1}}(\|m\|_{B^\sigma_{2,1}}+\|\boldsymbol{u}\|_{B^\sigma_{2,1}})\\
&+\tau 2^{q(\sigma-1)}\|\Delta_q\mathcal{G}\|_{L^2}+2^{q\sigma}(\|\Delta_q(h(m)\boldsymbol{u})\|_{L^2}+\tau\|\Delta_q(\tilde h(m)m)\|_{L^2})\},
\end{aligned} \tag{4.30}$$

where K_1,K_2 are given in (4.32)*, and $C>0$ is a uniform constant independent of τ.*

Proof. Combining (4.3), (4.6), and (4.20), we have

$$\begin{aligned}
&\frac{d}{dt}\bigg\{\frac{K_1}{2}2^{2q}\bigg(\|\Delta_q m\|_{L^2}^2+\|\Delta_q\mathbf{u}\|_{L^2}^2+\frac{1}{\bar n}\|\Delta_q\mathbf{e}\|_{L^2}^2\bigg)\\
&+\frac{K_2\tau}{2}\mathrm{Im}\int|\xi|\Big((\widehat{\Delta_q U})^*K(\xi)\widehat{\Delta_q U}\Big)d\xi\bigg\}\\
&+\frac{K_1}{\tau}2^{2q}\|\Delta_q\mathbf{u}\|_{L^2}^2+\tau\frac{a}{2}K_2 2^{2q}\|\Delta_q m\|_{L^2}^2+K_3\tau 2^{2q}\|\Delta_q\mathbf{e}\|_{L^2}^2\\
&\le K_1 2^{2q}\bigg\{\frac{1}{2}\|\nabla\mathbf{u}\|_{L^\infty}(\|\Delta_q m\|_{L^2}^2+\|\Delta_q\mathbf{u}\|_{L^2}^2)\\
&+\|[\mathbf{u},\Delta_q]\cdot\nabla m\|_{L^2}\|\Delta_q m\|_{L^2}+\|[\mathbf{u},\Delta_q]\cdot\nabla\mathbf{u}\|_{L^2}\|\Delta_q\mathbf{u}\|_{L^2}\\
&+\frac{1}{\bar n}\|\Delta_q(h(m)\mathbf{u})\|_{L^2}\|\Delta_q\mathbf{e}\|_{L^2}\bigg\}+K_2\bigg\{\frac{C}{\tau}2^{2q}\|\Delta_q\mathbf{u}\|_{L^2}^2\\
&+C\tau 2^q\|\Delta_q U\|_{L^2}\|\Delta_q\mathcal{G}\|_{L^2}+C\tau\|\Delta_q(\tilde h(m)m)\|_{L^2}\|\Delta_q m\|_{L^2}\bigg\}\\
&+K_3 2^q\tau C\bigg\{\bigg(\frac{\bar n}{a}\|\Delta_q m\|_{L^2}+\|\Delta_q(\tilde h(m)m)\|_{L^2}\bigg)\|\Delta_q\mathbf{e}\|_{L^2}\bigg\},
\end{aligned} \tag{4.31}$$

where these positive constants K_1,K_2, and K_3 (independent of τ) satisfy

$$K_2=\frac{K_1}{4C},\quad K_3=\frac{a^3}{2C^2\bar n^2}K_2. \tag{4.32}$$

We introduce them in order to ensure that

$$\begin{aligned}&\frac{K_1}{2}2^{2q}\Big(\|\Delta_q m\|_{L^2}^2+\|\Delta_q \mathbf{u}\|_{L^2}^2+\frac{1}{\bar{n}}\|\Delta_q \mathbf{e}\|_{L^2}^2\Big)\\&+\frac{K_2\tau}{2}\mathrm{Im}\int|\xi|\Big((\widehat{\Delta_q U})^* K(\xi)\widehat{\Delta_q U}\Big)d\xi\\&\approx 2^{2q}\Big(\|\Delta_q m\|_{L^2}^2+\|\Delta_q \mathbf{u}\|_{L^2}^2+\|\Delta_q \mathbf{e}\|_{L^2}^2\Big),\end{aligned}\tag{4.33}$$

since

$$\left|\frac{K_2\tau}{2}\mathrm{Im}\int|\xi|\Big((\widehat{\Delta_q U})^* K(\xi)\widehat{\Delta_q U}\Big)d\xi\right|\le\frac{CK_2}{2}2^{2q}(\|\Delta_q m\|_{L^2}^2+\|\Delta_q \mathbf{u}\|_{L^2}^2),\tag{4.34}$$

and we eliminate the quadratic term $\frac{C}{\tau}2^{2q}\|\Delta_q \mathbf{u}\|_{L^2}^2$, $K_3 2^q\tau C\frac{\bar{n}}{a}\|\Delta_q m\|_{L^2}\|\Delta_q \mathbf{e}\|_{L^2}$ in the right-hand side of (4.31) with the aid of Young's inequality; for similar details, see [6]. Then, dividing the resulting inequality by

$$\left\{\frac{K_1}{2}2^{2q}\Big(\|\Delta_q m\|_{L^2}^2+\|\Delta_q \mathbf{u}\|_{L^2}^2+\frac{1}{\bar{n}}\|\Delta_q \mathbf{e}\|_{L^2}^2\Big)+\frac{K_2}{2}\mathrm{Im}\int|\xi|\Big((\widehat{\Delta_q U})^* K(\xi)\widehat{\Delta_q U}\Big)d\xi\right\}^{1/2}$$

and after eliminating the quadratic terms and multiplying the factor $2^{q(\sigma-1)}$ on both sides of the inequality, we arrive at (4.30) immediately with the help of Lemmas 4.6 and 4.7. □

Similarly, for the case of low frequency ($q=-1$), we have the following a priori estimate.

LEMMA 4.9 ($q=-1$). *There exist some positive constants $\bar{K}_1,\bar{K}_2,\bar{K}_3$, and μ_3 independent of τ such that the following estimate holds:*

$$\begin{aligned}&2^{-(\sigma-1)}\frac{d}{dt}\left\{\frac{\bar{K}_1}{2}2^{-2}\Big(\|\Delta_{-1} m\|_{L^2}^2+\|\Delta_{-1}\boldsymbol{u}\|_{L^2}^2+\frac{1}{\bar{n}}\|\Delta_{-1}\boldsymbol{e}\|_{L^2}^2\Big)\right.\\&\left.+\frac{\bar{K}_2\tau}{2}\mathrm{Im}\int|\xi|\Big((\widehat{\Delta_{-1}U})^* K(\xi)\widehat{\Delta_{-1}U}\Big)d\xi-\bar{K}_3\tau\int\Delta_{-1}\boldsymbol{e}\cdot\overline{\Delta_{-1}\boldsymbol{u}}dx\right\}^{1/2}\\&+\mu_3 2^{-\sigma}\Big(\tau\|\Delta_{-1} m\|_{L^2}+\|\Delta_{-1}\boldsymbol{u}\|_{L^2}+\tau\|\Delta_{-1}\boldsymbol{e}\|_{L^2}\Big)\\&\le C\{2^{-\sigma}(\|\nabla\boldsymbol{u}\|_{L^\infty}+\|\boldsymbol{u}\|_{L^\infty})(\|\Delta_{-1} m\|_{L^2}+\|\Delta_{-1}\boldsymbol{u}\|_{L^2})\\&+c_{-1}\|\boldsymbol{u}\|_{B^\sigma_{2,1}}(\|m\|_{B^\sigma_{2,1}}+\|\boldsymbol{u}\|_{B^\sigma_{2,1}})+\tau 2^{-(\sigma-1)}\|\Delta_{-1}\mathcal{G}\|_{L^2}\\&+2^{-\sigma}(\|\Delta_{-1}(h(m)\boldsymbol{u})\|_{L^2}+\tau\|\Delta_{-1}(\tilde{h}(m)m)\|_{L^2})\},\end{aligned}\tag{4.35}$$

where $C>0$ is a uniform constant independent of τ.

Remark 4.2. Similar to the proof of Lemma 4.8, the constants $\bar{K}_1,\bar{K}_2,\bar{K}_3$ are used to ensure that

$$\begin{aligned}&\frac{\bar{K}_1}{2}2^{-2}\Big(\|\Delta_{-1} m\|_{L^2}^2+\|\Delta_{-1}\mathbf{u}\|_{L^2}^2+\frac{1}{\bar{n}}\|\Delta_{-1}\mathbf{e}\|_{L^2}^2\Big)\\&+\frac{\bar{K}_2\tau}{2}\mathrm{Im}\int|\xi|\Big((\widehat{\Delta_{-1}U})^* K(\xi)\widehat{\Delta_{-1}U}\Big)d\xi-\bar{K}_3\tau\int\Delta_{-1}\mathbf{e}\cdot\overline{\Delta_{-1}\mathbf{u}}dx\\&\approx 2^{-2}\Big(\|\Delta_{-1} m\|_{L^2}^2+\|\Delta_{-1}\mathbf{u}\|_{L^2}^2+\|\Delta_{-1}\mathbf{e}\|_{L^2}^2\Big)\end{aligned}\tag{4.36}$$

and to eliminate some quadratic terms in the right-hand side of inequality (4.35).

Summing (4.30) on $q \in \mathbf{N} \cup \{0\}$ and adding with (4.35), then, according to the smallness of $\tau(0 < \tau \le 1)$, a priori assumption (4.1), (4.29), and Moser's estimates (Proposition 2.5), we obtain the following differential inequality:

$$\begin{aligned} &\frac{d}{dt}Q + \mu_4\Big(\tau\|m\|_{B^\sigma_{2,1}} + \|\mathbf{u}\|_{B^\sigma_{2,1}} + \tau\|\mathbf{e}\|_{B^\sigma_{2,1}}\Big) \\ &\le C\delta_1\Big(\tau\|m\|_{B^\sigma_{2,1}} + \|\mathbf{u}\|_{B^\sigma_{2,1}} + \tau\|\mathbf{e}\|_{B^\sigma_{2,1}}\Big), \end{aligned} \tag{4.37}$$

where

$$\begin{aligned} Q = \sum_{q\ge 0} 2^{q(\sigma-1)} \Bigg\{ &\frac{K_1}{2}2^{2q}\Big(\|\Delta_q m\|^2_{L^2} + \|\Delta_q \mathbf{u}\|^2_{L^2} + \frac{1}{\bar n}\|\Delta_q \mathbf{e}\|^2_{L^2}\Big) \\ &+ \frac{K_2\tau}{2}\mathrm{Im}\int |\xi|\Big((\widehat{\Delta_q U})^* K(\xi)\widehat{\Delta_q U}\Big)d\xi\Bigg\}^{1/2} \\ &+ \Bigg\{\frac{\bar K_1}{2}2^{-2}\Big(\|\Delta_{-1} m\|^2_{L^2} + \|\Delta_{-1}\mathbf{u}\|^2_{L^2} + \frac{1}{\bar n}\|\Delta_{-1}\mathbf{e}\|^2_{L^2}\Big) \\ &+ \frac{\bar K_2\tau}{2}\mathrm{Im}\int |\xi|\Big((\widehat{\Delta_{-1} U})^* K(\xi)\widehat{\Delta_{-1} U}\Big)d\xi - \bar K_3\tau\int \Delta_{-1}\mathbf{e}\cdot\overline{\Delta_{-1}\mathbf{u}}dx\Bigg\}^{1/2} \end{aligned} \tag{4.38}$$

and $\mu_4 > 0$ is a uniform constant independent of τ. Note that

$$Q \approx \|(m, \mathbf{u}, \mathbf{e})(\cdot, t)\|_{B^\sigma_{2,1}}, \quad t \ge 0, \tag{4.39}$$

and by choosing $\delta_1 = \frac{\mu_4}{2C}, \mu_1 = \frac{\mu_4}{2}$, we conclude the proof of Proposition 4.1.

Based on Propositions 3.1 and 4.1, we establish the global existence of classical solutions to the system (3.2)–(3.4) by virtue of the standard continuation argument. Using the imbedding property in Besov space $B^\sigma_{2,1}$, $(m, \mathbf{u}, \mathbf{e}) \in \mathcal{C}^1([0,\infty)\times\mathbf{R}^N)$ solves (3.2)–(3.4). From Remark 3.1, we know $(n, \mathbf{u}, \mathbf{e}) \in \mathcal{C}^1([0,\infty)\times\mathbf{R}^N)$ is a solution of (1.1)–(1.2) with $n > 0$. Furthermore, we attain Theorem 1.1.

5. Relaxation-time limit. In this section, we give the proof of Theorem 1.2.

Proof. From the uniform energy estimate (1.7) in Theorem 1.1 and the scaled variables (1.3), we have

$$\begin{aligned} &\sup_{s\ge 0}\Big(\|(n^\tau - \bar n, \tau\mathbf{u}^\tau, \mathbf{e}^\tau)(\cdot, s)\|_{B^\sigma_{2,1}}\Big) + \mu_0\int_0^\infty \Big(\|(n^\tau - \bar n, \mathbf{u}^\tau, \mathbf{e}^\tau)(\cdot, s)\|_{B^\sigma_{2,1}}\Big)ds \\ &\le C_0\|(n_0 - \bar n, \mathbf{u}_0, \mathbf{e}_0)\|_{B^\sigma_{2,1}}. \end{aligned} \tag{5.1}$$

According to Proposition 2.4 and the compactness theorem in [14], there exists a function $(\mathcal{N}, \mathcal{E}) \in \mathcal{C}([0,\infty), \bar n + B^\sigma_{2,1}) \times \mathcal{C}([0,\infty), B^\sigma_{2,1})$ such that, for any $T > 0$, the sequence (up to a subsequence)

$$\{n^\tau\} \to \mathcal{N} \quad \text{strongly in} \quad L^p(0, T; (B^{\sigma'}_{2,1})_{\mathrm{loc}})(1 \le p < +\infty,\ \sigma' < \sigma),$$

$$\{\mathbf{e}^\tau\} \to \mathcal{E} \quad \text{strongly in} \quad L^p(0, T; (B^{\sigma'}_{2,1})_{\mathrm{loc}})(1 \le p < +\infty,\ \sigma' < \sigma),\ \text{as}\ \ \tau \to 0.$$

From (5.1), we can derive that

$$\begin{aligned} &\mathbf{u}^\tau \ \text{ is uniformly bounded in } \ L^1(0, T; B^\sigma_{2,1}) \\ &\text{and} \quad \tau(\mathbf{u}^\tau\cdot\nabla)\mathbf{u}^\tau \quad \text{is uniformly bounded in } \ L^1(0, T; B^{\sigma-1}_{2,1}). \end{aligned}$$

Hence, in the system (1.4)–(1.5), the above convergence properties allow us to pass to the limit $\tau \to 0$ in the sense of distributions, which implies that $(\mathcal{N}, \mathcal{E})$ is a global weak solution to the drift-diffusion model (1.6). ∎

Acknowledgments. The author thanks his adviser, Professor Daoyuan Fang, for his constant support and encouragement. He also thanks the anonymous referees for their useful comments and suggestions.

REFERENCES

[1] G. Alì, D. Bini and R. Rionero, *Global existence and relaxation limit for smooth solutions to the Euler–Poisson model for semiconductors*, SIAM J. Math. Anal., 32 (2000), pp. 572–587.
[2] J. Y. Chemin, *Perfect Incompressible Fluids*, Oxford Lecture Ser. Math. Appl. 14, Oxford University Press, New York, 1998 (in English).
[3] G. Q. Chen, J. Jerome, and B. Zhang, *Particle hydrodynamic models in biology and microelectronics: Singular relaxation limits*, Nonlinear Anal., 30 (1997), pp. 233–244.
[4] J. F. Coulombel and T. Goudon, *The strong relaxation limit of the multidimensional isothermal Euler equations*, Trans. Amer. Math. Soc., 359 (2007), pp. 637–648.
[5] D. Y. Fang and J. Xu, *Existence and asymptotic behavior of C^1 solutions to the multidimensional compressible Euler equations with damping*, Nonlinear Anal., 70 (2009), pp. 244–261.
[6] D. Y. Fang, J. Xu, and T. Zhang, *Global exponential stability of classical solutions to the hydrodynamic model for semiconductors*, Math. Models Methods Appl. Sci., 17 (2007), pp. 1507–1530.
[7] L. Hsiao and K. Zhang, *The relaxation of the hydrodynamic model for semiconductors to the drift-diffusion equations*, J. Differential Equations, 165 (2000), pp. 315–354.
[8] A. Jüngel and Y. J. Peng, *A hierarchy of hydrodynamic models for plasmas: Zero-relaxation-time limits*, Comm. Partial Differential Equations, 24 (1999), pp. 1007–1033.
[9] A. Jüngel and Y. J. Peng, *Zero-relaxation-time limits in the hydrodynamic models for plasmas revisited*, Z. Angew. Math. Phys., 51 (2000), pp. 385–396.
[10] S. Junca and M. Rascle, *Relaxation of the isothermal Euler-Poisson system to the drift-diffusion equations*, Quart. Appl. Math., 58 (2000), pp. 511–521.
[11] C. Lattanzio and P. Marcati, *The relaxation to the drift-diffusion system for the 3-D isentropic Euler-Poisson model for semiconductors*, Discrete Contin. Dyn. Syst., 5 (1999), pp. 449–455.
[12] P. Marcati and R. Natalini, *Weak solutions to a hydrodynamic model for semiconductors and relaxation to the drift-diffusion equations*, Arch. Ration. Mech. Anal., 129 (1995), pp. 129–145.
[13] Y. Shizuta and S. Kawashima, *Systems of equations of hyperbolic-parabolic type with applications to the discrete Boltzmann equation*, Hokkaido Math. J., 14 (1985), pp. 249–275.
[14] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–96.
[15] W. A. Yong, *Diffusive relaxation limit of multidimensional isentropic hydrodynamical models for semiconductors*, SIAM J. Appl. Math., 64 (2004), pp. 1737–1748.
[16] W. A. Yong, *Entropy and global existence for hyperbolic balance laws*, Arch. Ration. Mech. Anal., 172 (2004), pp. 247–266.

SIAM J. MATH. ANAL.
Vol. 40, No. 5, pp. 1992–2006

EXISTENCE AND STABILITY RESULTS FOR PERIODIC STOKESIAN HELE–SHAW FLOWS*

JOACHIM ESCHER† AND BOGDAN-VASILE MATIOC†

Abstract. We consider here a 2π-periodic and two-dimensional Hele–Shaw flow modelling the motion of a viscous and incompressible fluid. The free surface is moving under the influence of gravity and is modelled by a modified Darcy law for Stokesian fluids. The bottom of the cell is assumed to be impermeable. We prove the existence of a unique classical solution if the initial data is near a constant, identify the equilibria of the flow, and study their stability.

Key words. quasi-linear elliptic equation, nonlinear parabolic equation, Hele–Shaw flow, non-Newtonian fluid, Oldroyd-B fluid, power law fluid

AMS subject classifications. 35J65, 35K55, 35R35, 42A45

DOI. 10.1137/070707671

1. Introduction.

1.1. The mathematical model. Using a fluid model with a shear-rate dependent viscosity, the authors in [12] derive a Darcy law coupling the pressure p and the velocity field v of the bulk fluid situated between the parallel plates of a Hele–Shaw cell:

$$v = -\frac{Dp}{\overline{\mu}(|Dp|^2)}. \tag{1}$$

Here $\overline{\mu}$ is the so-called *effective viscosity*. In the situation studied in [12] the cell is horizontal and the gap d between the plates is small in relation to the other two dimensions of the cell. Having averaged over the gap d they consider (1) for a two-dimensional fluid. From the conservation of mass one concludes that the pressure p of a finite patch Ω of fluid with boundary Γ must solve the Dirichlet problem

$$\operatorname{div}\left(\frac{Dp}{\overline{\mu}(|Dp|^2)}\right) = 0 \quad \text{in} \quad \Omega, \qquad p = -\sigma\kappa \quad \text{on} \quad \Gamma, \tag{2}$$

where σ is the surface tension parameter and κ the curvature of Γ. The solvability of (2) is discussed in terms of the monotonicity of $\overline{\mu}$. It should be emphasized that in [12] only problems on temporal fixed domains are rigorously investigated. However, the full problem describing the more complex situation in which the boundary Γ moves and which is obtained by coupling (2) with an evolution equation describing the motion of the fluid molecules situated on Γ is not addressed by the authors in [12].

In this paper we study the general Cauchy problem of the full system describing the flow of a non-Newtonian fluid in a vertical Hele–Shaw cell with an impermeable bottom and laterals (see Figure 1). Hence, effects caused by gravity must be incorporated into our setting. As in [12] the gap d is again small compared to the height and length of the cell. In order to avoid the contact angle problem a two-dimensional,

*Received by the editors November 8, 2007; accepted for publication (in revised form) September 13, 2008; published electronically January 21, 2009.
http://www.siam.org/journals/sima/40-5/70767.html
†Institute of Applied Mathematics, Leibniz University of Hanover, Welfengarten 1, Hanover, Germany (escher@ifam.uni-hannover.de, matioc@ifam.uni-hannover.de).

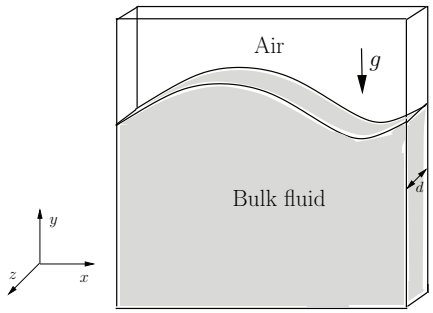


FIG. 1.

strip-like geometry will be considered. Given a positive function $f \in C^1(\mathbb{R})$, bounded away from 0, we define the set

$$\widetilde{\Omega}_f := \{(x,y) \in \mathbb{R}^2 \,:\, 0 < y < f(x)\},$$

identified as the fluid domain, and denote the components of its boundary $\partial\widetilde{\Omega}_f$ by

$$\widetilde{\Gamma}_f := \{(x, f(x)) \,:\, x \in \mathbb{R}\}, \quad \widetilde{\Gamma}_0 := \mathbb{R} \times \{0\}.$$

We denote by ν the outward-pointing normal of $\partial\widetilde{\Omega}_f$. The motion of the fluid is modelled using the Darcy law (1):

$$v = -\frac{Du}{\overline{\mu}(|Du|^2)},$$

with p replaced by the *velocity potential* u defined by

$$u(x,y) = \frac{p(x,y)}{g \cdot \rho} + y, \quad (x,y) \in \widetilde{\Omega}_f.$$

Here g is the gravity acceleration constant, ρ is the density constant, and $Du = (\partial_1 u, \partial_2 u)$ is the gradient of u. Also called *piezometric head*, u expresses the mechanical energy due to gravity and pressure of the fluid, per unit weight of fluid (see [3]).

The viscosity $\mu \in C^\infty([0,\infty),(0,\infty))$ is assumed to satisfy $\mu(r) + 2r\mu'(r) > 0$ for $r \geq 0$. Particularly, this implies that the mapping $[0,\infty) \ni r \longmapsto h(r) := r\mu^2(r)$ is invertible. The effective viscosity $\overline{\mu}$ (see [12]) is defined by

$$\frac{1}{\overline{\mu}(r)} := c \int_{-1}^{1} \frac{s^2}{\widetilde{\mu}(rs^2)}\, ds,$$

where c is a positive constant and $\widetilde{\mu} := \mu \circ h^{-1}$. The free surface $\widetilde{\Gamma}_f$ separating the fluid from air, at pressure normalized to be zero, is moving under the influence of gravity. Surface tension effects are neglected. Consequently, $p = 0$ on $\widetilde{\Gamma}_f$ and

$$u = f \quad \text{on} \quad \widetilde{\Gamma}_f.$$

On the fixed boundary $\widetilde{\Gamma}_0$ we have no flux, i.e.,

$$\partial_\nu u = 0 \quad \text{on} \quad \widetilde{\Gamma}_0.$$

Furthermore, the interface $\widetilde{\Gamma}_f$ is implicitly given by $F(t,z)=0$, where $z=(x,y)$ and $F(t,z)=y-f(t,x)$. Assuming that a particle located on the interface remains there, we obtain by differentiating this equation with respect to t the relation

$$0=\frac{d}{dt}F(t,z)=-f_t(t,x)+(-f_x,1)z'.$$

Finally, we shall make the following periodicity requirement on f and u:

$$\begin{aligned} f(t,x+2\pi) &= f(t,x) \qquad \forall x\in\mathbb{R},\, t\geq 0,\\ u(x+2\pi,y) &= u(x,y) \qquad \forall (x,y)\in\widetilde{\Omega}_{f(t)},\, t\geq 0. \end{aligned}$$

Summarizing, we arrive at the following system:

$$\begin{aligned} \operatorname{div}\left(\tfrac{Du}{\overline{\mu}(|Du|^2)}\right) &= 0 \quad \text{in} \quad \Omega_{f(t)}, \quad t\geq 0,\\ \partial_\nu u &= 0 \quad \text{on} \quad \Gamma_0, \quad t\geq 0,\\ u &= f \quad \text{on} \quad \Gamma_{f(t)}, \quad t\geq 0,\\ \partial_t f(t,\cdot)+\tfrac{\sqrt{1+\partial_x f^2(t,\cdot)}}{\overline{\mu}(|Du(\cdot,f(t,\cdot))|^2)}\partial_\nu u(\cdot,f(t,\cdot)) &= 0 \quad \text{on} \quad \mathbb{S}^1, \quad t>0,\\ f(0,\cdot) &= f_0 \quad \text{on} \quad \mathbb{S}^1, \end{aligned} \tag{3}$$

where, given $t\geq 0$, we use the notation

$$\begin{aligned} \Omega_{f(t)} &:= \{(x,y)\in\mathbb{S}^1\times\mathbb{R}\,:\,0<y<f(t,x)\},\\ \Gamma_{f(t)} &:= \{(x,f(t,x))\,:\,x\in\mathbb{S}^1\},\quad \Gamma_0=\mathbb{S}^1\times\{0\}, \end{aligned}$$

and where $\mathbb{S}^1$ stands for the unit circle. For the sake of simplicity, we identify periodic functions on $\mathbb{R}$ with functions on $\mathbb{S}^1$, and periodic functions in the x variable in $\widetilde{\Omega}_f$ with functions in Ω_f, for positive functions f on $\mathbb{S}^1$, respectively. Given $k\in\mathbb{N}$ and $\alpha\in(0,1)$, we define the so-called little Hölder spaces $h^{k+\alpha}(\mathbb{S}^1)$ as the closure of $C^\infty(\mathbb{S}^1)$ in $C^{k+\alpha}(\mathbb{S}^1)$.

For our analysis we fix $\alpha\in(0,1)$ and define

$$\mathcal{U}:=\Big\{f\in C^{2+\alpha}(\mathbb{S}^1)\,:\,\min_{x\in\mathbb{S}^1}f(x)>0\Big\},$$

respectively, $\mathcal{V}:=\mathcal{U}\cap h^{2+\alpha}(\mathbb{S}^1)$. For $f\in\mathcal{U}$ we denote by $buc^{k+\alpha}(\Omega_f)$ the closure of $BUC^\infty(\Omega_f)$ in the Hölder space $BUC^{k+\alpha}(\Omega_f)$.

A pair (u,f) is called a *classical Hölder solution* of (3) on $[0,T]$, with $T>0$, if

$$f\in C([0,T],\mathcal{V})\cap C^1([0,T],h^{1+\alpha}(\mathbb{S}^1)),$$

$$u(\cdot,t)\in buc^{2+\alpha}(\Omega_{f(t)}),\, t\in[0,T],$$

and if (u,f) satisfies the equations in (3) pointwise.

For simplicity we denote

$$\mathcal{Q}u:=\operatorname{div}\left(\frac{Du}{\overline{\mu}(|Du|^2)}\right).$$

Then $\mathcal{Q}u = a_{ij}(Du)u_{ij}$, with

$$a_{ij}(q_1,q_2) = \frac{\delta_{ij}}{\overline{\mu}(|q|^2)} - \frac{2q_iq_j\overline{\mu}'(|q|^2)}{\overline{\mu}^2(|q|^2)}, \quad 1 \le i,\, j \le 2,\, q = (q_1,q_2) \in \mathbb{R}^2.$$

The eigenvalues of the matrix $[a_{ij}(q)]_{1\le i,\,j\le 2}$, $q \in \mathbb{R}^2$, are

$$\lambda_1(q) = \frac{1}{\overline{\mu}(|q|^2)}, \qquad \lambda_2(q) = \frac{1}{\overline{\mu}(|q|^2)} - \frac{2|q|^2\overline{\mu}'(|q|^2)}{\overline{\mu}^2(|q|^2)},$$

and we have

$$c|\xi|^2 \le a_{ij}(q_1,q_2)\xi_i\xi_j \le C|\xi|^2 \quad \forall \xi = (\xi_1,\xi_2) \in \mathbb{R}^2,\, (q_1,q_2) \in \mathbb{R}^2,$$

if we assume that there exist positive constants c and C such that

$$(\mathrm{A}_1) \quad c \le \frac{1}{\overline{\mu}(r)} \le C \quad \forall\, r \ge 0,$$

$$(\mathrm{A}_2) \quad c \le \frac{1}{\overline{\mu}(r)} - \frac{2r\overline{\mu}'(r)}{\overline{\mu}^2(r)} \le C \quad \forall\, r \ge 0.$$

It can be shown (cf. [6]) that (A_1) and (A_2) hold true, provided the viscosity μ satisfies

$$(\mathrm{V}_1) \quad m \le \mu(r) \le M,$$

$$(\mathrm{V}_2) \quad m \le \mu(r) + 2r\mu'(r) \le M,$$

for all $r \ge 0$, where m and M are positive constants. If the viscosity μ is constant, then the fluid is called Newtonian. Results regarding this situation can be found in [4] and [7], [8], [9], [10]. The class of Stokesian, sometimes also called non-Newtonian, fluids satisfying (V_1) and (V_2) is quite large; see [6]. This class includes particularly numerous Oldroyd-B and power law fluids.

Under the above assumptions $\mathcal{Q}$ is a uniformly elliptic operator in $\mathbb{R}^2$. Consider now the problem

$$\begin{aligned}
\operatorname{div}\left(\tfrac{Du}{\overline{\mu}(|Du|^2)}\right) &= 0 \quad \text{in} \quad G_{f(t)}, && t \ge 0,\\
u &= f \quad \text{on} \quad \partial G_{f(t)}, && t \ge 0,\\
\partial_t f(t,\cdot) + \tfrac{\sqrt{1+\partial_x f^2(t,\cdot)}}{\overline{\mu}(|Du(\cdot,f(t,\cdot))|^2)}\partial_\nu u(\cdot, f(t,\cdot)) &= 0 \quad \text{on} \quad \mathbb{S}^1, && t > 0,\\
f(0,\cdot) &= f_0 \quad \text{on} \quad \mathbb{S}^1,
\end{aligned} \tag{4}$$

where

$$G_{f(t)} := \{(x,y) \in \mathbb{S}^1 \times \mathbb{R} : -f(t,x) < y < f(t,x)\},$$

for $t \ge 0$. A pair (u,f) is called a *classical Hölder solution* of (4) on $[0,T]$ if

$$f \in C([0,T],\mathcal{V}) \cap C^1([0,T],h^{1+\alpha}(\mathbb{S}^1)),$$

$$u(\cdot\,,t) \in buc^{2+\alpha}(G_{f(t)}),\, t \in [0,T],$$

and if (u,f) satisfies the equations in (4) pointwise.

Given $f \in \mathcal{V}$, let $u \in buc^{2+\alpha}(G_f)$ be the unique solution (existence will be discussed later) of the quasi-linear Dirichlet problem

$$\begin{aligned} \mathcal{Q}u &= 0 \quad \text{in} \quad G_f, \\ u &= f \quad \text{on} \quad \partial G_f. \end{aligned} \tag{5}$$

Because of the symmetry of the domain G_f and of the boundary conditions we obtain

$$u(x,-y) = u(x,y)$$

for all $(x,y) \in G_f$. Consequently, $\partial_\nu u = 0$ on Γ_0; thus the restriction of u to Ω_f is the unique solution of

$$\begin{aligned} \mathcal{Q}u &= 0 \quad \text{in} \quad \Omega_f, \\ \partial_\nu u &= 0 \quad \text{on} \quad \Gamma_0, \\ u &= f \quad \text{on} \quad \Gamma_f. \end{aligned} \tag{6}$$

We deduce that there exists a one-to-one correspondence between solutions to (3) and (4). Namely, if (u,f) is a solution to (3) on the interval $[0,T]$, $T>0$, then for each $t \in [0,T]$ the velocity potential $u(\cdot\,,t) \in buc^{2+\alpha}(\Omega_{f(t)})$ can be extended by reflection on the entire domain $G_{f(t)}$. We obtain in this manner a solution to (4). Moreover, if $(\widetilde{u},f)$ is a solution to (4) on the interval $[0,T]$, $T>0$, then we can restrict for $t \in [0,T]$ the function $\widetilde{u}(\cdot\,,t)$ to $\Omega_{f(t)}$, and so obtain a solution to (3).

It is more convenient to treat (4), because we have in this case a Dirichlet problem for the velocity potential, which can be solved using the techniques presented in [11].

We state now the first main result of this work.

THEOREM 1.1. *Assume that* (A_1) *and* (A_2) *hold true.*

(a) *Let c be a positive constant. Then we find an open neighborhood $\mathcal{O}$ of c in $\mathcal{V}$ such that for each $f_0 \in \mathcal{O}$, problem* (3) *has a classical Hölder solution (u,f) on an interval $[0,T]$ with $T>0$. Moreover, there exists a constant $\gamma \in (0,1)$ such that $f \in C^\gamma_\gamma((0,T],h^{2+\alpha}(\mathbb{S}^1))$.*

(b) *Let (u_1,f_1) and (u_2,f_2) be solutions of* (3) *with $f_1 \in C^\gamma_\gamma((0,T],h^{2+\alpha}(\mathbb{S}^1))$, $\gamma \in (0,1)$, and $f_2 \in C^\delta_\delta((0,T],h^{2+\alpha}(\mathbb{S}^1))$, $\delta \in (0,1)$. If $f_1([0,T]) \subset \mathcal{O}$ and $f_2([0,T]) \subset \mathcal{O}$, then $(u_1,f_1)=(u_2,f_2)$.*

For a definition of the weighted Hölder spaces $C^\gamma_\gamma((0,T],h^{2+\alpha}(\mathbb{S}^1))$, $\gamma \in (0,1)$, see [13].

In section 3 we prove that each pair $(c,c) \in buc^{2+\alpha}(\Omega_c) \times \mathcal{V}$, with $c>0$, is a stationary solution of the flow (3). Furthermore, we prove that if the initial data $f_0 \in \mathcal{O}$ is close enough to c in $h^{2+\alpha}(\mathbb{S}^1)$ and if Ω_{f_0} and Ω_c enclose the same fluid volume, then the solution corresponding to f_0 exists globally in time and is attracted at an exponential rate by c (see Theorem 3.3).

1.2. The transformed problem. In order to solve the problem we transform it on a fixed reference manifold $G := \mathbb{R} \times (0,2)$. In order to do so we define for each $f \in \mathcal{U}$ a diffeomorphism $\phi_f \in \mathit{Diff}^{\,2+\alpha}(G,G_f)$ by

$$\phi_f(x,y) := (x,(1-y)f(x)),\ (x,y) \in G,$$

and the push-forward and pull-back operators induced by ϕ_f:

$$\phi_f^* : BUC(G_f) \to BUC(G), \quad u \longmapsto u \circ \phi_f,$$

$$\phi_*^f : BUC(G) \to BUC(G_f), \quad v \longmapsto v \circ \phi_f^{-1}.$$

The transformed operators $\mathcal{A}(f)$ and $\mathcal{B}$, acting on $BUC^2(G)$, respectively, $\mathcal{U}\times BUC^{2+\alpha}(G)$, are defined by

$$\mathcal{A}(f) := \phi_f^* \circ \mathcal{Q} \circ \phi_*^f,$$
$$\mathcal{B}(f,v)(x) := \frac{D(\phi_*^f v)}{\overline{\mu}(|D(\phi_*^f v)|^2)}(x, f(x)) \cdot n(x), \quad x \in \mathbb{S}^1,$$

with $n(x) := (-f'(x), 1)$, $x \in \mathbb{S}^1$. Transformation of (4) to G yields

$$\begin{aligned} \mathcal{A}(f)v &= 0 \quad \text{in} \quad G \times [0,\infty), \\ v &= f \quad \text{on} \quad \partial G \times [0,\infty), \\ \partial_t f + \mathcal{B}(f,v) &= 0 \quad \text{on} \quad \mathbb{S}^1 \times (0,\infty), \\ f(0) &= f_0, \end{aligned} \tag{7}$$

where $v := \phi_f^* u$. A pair (v, f) is called a *classical Hölder solution* of (7) on $[0, T]$, with $T > 0$, if

$$f \in C([0,T], \mathcal{V}) \cap C^1([0,T], h^{1+\alpha}(\mathbb{S}^1)),$$
$$v(\cdot\,, t) \in buc^{2+\alpha}(G),\ t \in [0,T],$$

and if (v, f) satisfies the equations in (7) pointwise. It is obvious that problems (4) and (7) are equivalent in the following sense.

LEMMA 1.2. *Let $f_0 \in \mathcal{V}$ be given.*

(a) *If (u, f) is a classical Hölder solution of* (4)*, then $(\phi_f^* u, f)$ is a classical Hölder solution of* (7).

(b) *If (v, f) is a classical Hölder solution of* (7)*, then $(\phi_*^f v, f)$ is a classical Hölder solution of* (4).

Proof. See, for example, [6]. □

LEMMA 1.3. *Given $f \in \mathcal{U}$, we have*

$$\mathcal{A}(f)v = b_{ij}(y, f, Dv)v_{ij} + b(y, f, Dv)v_2 \quad \text{for} \quad v \in BUC^2(G),$$

where, using the notation $D_f v := (v_1 + \frac{(1-y)f'}{f} v_2, -\frac{1}{f} v_2)$ for $f \in \mathcal{U}$, $v \in BUC^2(G)$, and $y \in [0,2]$, we have

$$b_{11}(y, f, Dv) = a_{11}(D_f v),$$

$$b_{12}(y, f, Dv) = b_{21}(y, f, Dv) = \frac{(1-y)f'}{f} a_{11}(D_f v) - \frac{1}{f} a_{12}(D_f v),$$

$$b_{22}(y, f, Dv) = \frac{(1-y)^2 f'^2}{f^2} a_{11}(D_f v) - \frac{2(1-y)f'}{f^2} a_{12}(D_f v) + \frac{1}{f^2} a_{22}(D_f v),$$

$$b(y, f, Dv) = (1-y)\left(\frac{f''}{f} - \frac{2f'^2}{f^2}\right) a_{11}(D_f v) + \frac{2f'}{f^2} a_{12}(D_f v).$$

Proof. This follows by direct computation. □

We have the following existence and uniqueness theorem. The proof is similar to that of Lemma 2.2 in [6].

THEOREM 1.4. *Let $f \in \mathcal{V}$ be given. There exists a unique solution $v \in buc^{2+\alpha}(G)$ of the quasi-linear Dirichlet problem*

$$\begin{aligned} \mathcal{A}(f)v &= 0 \quad in \quad G, \\ v &= f \quad on \quad \partial G. \end{aligned} \tag{8}$$

Given $f \in \mathcal{V}$, we denote by $\mathcal{T}(f) \in buc^{2+\alpha}(G)$ the solution to (8). The mapping $[\mathcal{V} \ni f \mapsto \mathcal{T}(f) \in buc^{2+\alpha}(G)]$ is smooth.

2. The nonlinear Cauchy problem. Due to Theorem 1.4 we can reduce problem (4) to an abstract, fully nonlinear Cauchy problem on $\mathbb{S}^1$:

$$\partial_t f + \Phi(f) = 0, \qquad f(0) = f_0, \tag{9}$$

where

$$\Phi(f) := \mathcal{B}(f, \mathcal{T}(f)).$$

Note that the mapping $f \longmapsto \Phi(f)$ is a nonlinear and nonlocal operator of first order. It carries also a fully nonlinear structure in the sense that there is no leading linear part in $\mathcal{Q}$. We solve the Cauchy problem (9) by applying the theory of maximal regularity; cf. [13]. Further on we verify that the assumption in Theorem 8.1.1 in [13] are satisfied.

We first restrict our attention to the operator $\mathcal{B}$. Let $(f, v) \in \mathcal{V} \times buc^{2+\alpha}(G)$ be given. The function $\mathcal{B}(f, v)$ is given by

$$\mathcal{B}(f, v) = -\frac{1}{\overline{\mu}(|\gamma_0 D_f v|^2)} \left(f' \gamma_0 v_1 + \frac{1}{f}(1 + f'^2)\gamma_0 v_2 \right),$$

with γ_0 denoting the trace operator on Γ_0. Together with the relation

$$|\gamma_0 D_f v|^2 = \gamma_0 v_1^2 + 2\frac{f'}{f}\gamma_0 v_1 v_2 + \frac{1 + f'^2}{f^2}\gamma_0 v_2^2$$

we obtain that the operator $\mathcal{B}$ defined above is smooth.

LEMMA 2.1. *The mapping $\mathcal{B} : \mathcal{V} \times buc^{2+\alpha}(G) \to h^{1+\alpha}(\mathbb{S}^1)$ is smooth. Given $(f, v) \in \mathcal{V} \times buc^{2+\alpha}(G)$, we have*

$$\begin{aligned} \partial\mathcal{B}(f, v)[h, u] = &- \frac{1}{\overline{\mu}} \left(|\gamma_0 D_f v|^2\right) \Big[f' \gamma_0 u_1 + h' \gamma_0 v_1 + \frac{1}{f}(1 + f'^2)\gamma_0 u_2 \\ &- \left(\frac{h}{f^2} - \frac{2f'h'}{f} + \frac{hf'^2}{f^2} \right) \gamma_0 v_2 \Big] \\ &- 2 \left(\frac{1}{\overline{\mu}} \right)' \left(|\gamma_0 D_f v|^2\right) \left(f' \gamma_0 v_1 + \frac{1}{f}(1 + f'^2)\gamma_0 v_2 \right) \Big[\gamma_0 v_1 u_1 \\ &+ \frac{h'}{f}\gamma_0 v_1 v_2 + \frac{f'}{f}\gamma_0 u_1 v_2 + \frac{f'}{f}\gamma_0 v_1 u_2 - \frac{f'h}{f^2}\gamma_0 v_1 v_2 + \frac{f'h'}{f^2}\gamma_0 v_2^2 \\ &+ \frac{f'^2}{f^2}\gamma_0 v_2 u_2 - \frac{hf'^2}{f^3}\gamma_0 v_2^2 + \frac{1}{f^2}\gamma_0 v_2 u_2 - \frac{h}{f^3}\gamma_0 v_2^2 \Big] \end{aligned}$$

for all $[h, u] \in h^{2+\alpha}(\mathbb{S}^1) \times buc^{2+\alpha}(G)$.

LEMMA 2.2. *Given $f \in \mathcal{V}$ and $h \in h^{2+\alpha}(\mathbb{S}^1)$, the function $\partial\mathcal{T}(f)[h]$ is the unique solution of the linear Dirichlet problem*

$$
\begin{aligned}
&b_{ij}w_{ij} + bw_2 + D_f w \left[u_{11}\partial a_{11}(D_f u) + 2u_{12}\left(\frac{(1-y)f'}{f}\partial a_{11}(D_f u) - \frac{1}{f}\partial a_{12}(D_f u)\right)\right.\\
&+ u_{22}\left(\frac{(1-y)^2 f'^2}{f^2}\partial a_{11}(D_f u) - 2\frac{(1-y)f'}{f^2}\partial a_{12}(D_f u) + \frac{1}{f^2}\partial a_{22}(D_f u)\right)\\
&\left.+ u_2\left((1-y)\left(\frac{f''}{f} - 2\frac{f'^2}{f^2}\right)\partial a_{11}(D_f u) + 2\frac{f'}{f^2}\partial a_{12}(D_f u)\right)\right]\\
&\qquad = -u_2\left((1-y)\frac{fh' - f'h}{f^2}, \frac{h}{f^2}\right)\cdot\left[u_{11}\partial a_{11}(D_f u)\right.\\
&+ 2u_{12}\left(\frac{(1-y)f'}{f}\partial a_{11}(D_f u) - \frac{1}{f}\partial a_{12}(D_f u)\right) + u_{22}\left(\frac{(1-y)^2 f'^2}{f^2}\partial a_{11}(D_f u)\right.\\
&- 2\frac{(1-y)f'}{f^2}\partial a_{12}(D_f u) + \frac{1}{f^2}\partial a_{22}(D_f u)\Big) + u_2\left((1-y)\left(\frac{f''}{f} - 2\frac{f'^2}{f^2}\right)\partial a_{11}(D_f u)\right.\\
&\left.\left.+ 2\frac{f'}{f^2}\partial a_{12}(D_f u)\right)\right] - 2u_{12}\left((1-y)\frac{fh' - f'h}{f^2}a_{11}(D_f u) + \frac{h}{f^2}a_{12}(D_f u)\right)\\
&- 2u_{22}\left(\frac{(1-y)^2(ff'h' - f'^2 h)}{f^3}a_{11}(D_f u) - (1-y)\frac{fh' - 2f'h}{f^3}a_{12}(D_f u)\right.\\
&\left.- \frac{h}{f^3}a_{22}(D_f u)\right) - u_2\left((1-y)\left(\frac{fh'' - f''h}{f^2} - 4\frac{ff'h' - f'^2 h}{f^3}\right)a_{11}(D_f u)\right.\\
&\left.+ 2\frac{fh' - 2f'h}{f^3}a_{12}(D_f u)\right) \qquad \text{in} \quad G,
\end{aligned}
$$

$$
w = h \quad \text{on} \quad \partial G,
$$

where $u := \mathcal{T}(f)$ and $b_{ij} = b_{ij}(y, f, Du)$, $b = b(y, f, Du)$ are the coefficients computed in Lemma 1.3.

Proof. The proof follows using standard arguments and we omit it. □

Consequently, we obtain from Lemma 2.1 and Theorem 1.4 that the operator Φ is smooth. We can use the chain rule to compute its derivative. More precisely, we have the following result.

THEOREM 2.3. *The map Φ belongs to $C^\infty(\mathcal{V}, h^{1+\alpha}(\mathbb{S}^1))$ and*

$$
\partial\Phi(f) = \partial\mathcal{B}(f, \mathcal{T}(f)) \circ (\mathrm{id}_{h^{2+\alpha}(\mathbb{S}^1)}, \partial\mathcal{T}(f)) \quad \forall f \in \mathcal{V}.
$$

Our goal is to show that for $c \in \mathbb{R}_{>0}$ the Fréchet derivative $-\partial\Phi(c)$ generates a strongly continuous analytic semigroup in $\mathcal{L}(h^{1+\alpha}(\mathbb{S}^1))$. Then, using general results from the theory of maximal regularity (cf. [13]), we can prove Theorem 1.1. It is clear that we consider here $-\partial\Phi(c)$ as an unbounded operator in $h^{1+\alpha}(\mathbb{S}^1)$ with domain $h^{2+\alpha}(\mathbb{S}^1)$.

Let $c \in \mathbb{R}_{>0}$ be given. The solution $\mathcal{T}(c)$ to (8) is the constant function c on G. Given $h \in h^{2+\alpha}(\mathbb{S}^1)$, the map $\partial\mathcal{T}(c)[h]$ is the solution of the Dirichlet problem

$$\begin{aligned} c^2 w_{11} + w_{22} &= 0 \quad \text{in} \quad G, \\ w &= h \quad \text{on} \quad \partial G. \end{aligned} \tag{10}$$

Consequently,

$$-\partial\Phi(c)[h](x) = \frac{\zeta}{c} w_2(x,0), \quad x \in \mathbb{S}^1,$$

where $\zeta := 1/\overline{\mu}(0)$.

We expand h and w in the following way:

$$h(x) = \sum_{k\in\mathbb{Z}} c_k e^{ikx}, \qquad w(x,y) = \sum_{k\in\mathbb{Z}} C_k(y) e^{ikx},$$

and we substitute these expressions into (10). Comparing the coefficients of e^{ikx}, $k \in \mathbb{Z}$, we get the following equations for C_k:

$$\begin{aligned} C_k'' - c^2k^2 C_k &= 0, \quad 0 < y < 2, \\ C_k(0) &= c_k, \\ C_k(2) &= c_k, \end{aligned} \tag{11}$$

for $k \in \mathbb{Z} \setminus \{0\}$, and

$$\begin{aligned} C_0'' &= 0, \quad 0 < y < 2, \\ C_0(0) &= c_0, \\ C_0(2) &= c_0. \end{aligned} \tag{12}$$

The solution to (12) is the constant function $C_0 = c_0$, and for $k \in \mathbb{Z} \setminus \{0\}$ we have

$$C_k(y) = c_k d_k(y), \quad 0 \le y \le 2,$$

where

$$d_k(y) := \frac{e^{2ck}-1}{e^{4ck}-1} e^{cky} + \frac{e^{4ck}-e^{2ck}}{e^{4ck}-1} e^{-cky}, \quad 0 \le y \le 2.$$

Summarizing, we obtain

$$-\partial\Phi(c)\left[\sum_{k\in\mathbb{Z}} c_k e^{ikx}\right] = \sum_{k\in\mathbb{Z}} \lambda_k c_k e^{ikx}, \tag{13}$$

for all $h = \sum_{k\in\mathbb{Z}} c_k e^{ikx} \in h^{2+\alpha}(\mathbb{S}^1)$, with

$$\lambda_k := -\zeta \frac{e^{2ck}-1}{e^{2ck}+1} k, \quad k \in \mathbb{Z}. \tag{14}$$

The multiplying coefficients λ_k satisfy $\lambda_k = \lambda_{-k}$ for all $k \in \mathbb{Z}$. Furthermore,

$$0 = \lambda_0 > \lambda_{\pm 1} > \lambda_{\pm 2} > \cdots > \lambda_{\pm k} > \lambda_{\pm(k+1)} > \cdots .$$

For the purpose of proving that $-\partial\Phi(c)$ generates a strongly continuous analytic semigroup in $\mathcal{L}(h^{1+\alpha}(\mathbb{S}^1))$, i.e., $\partial\Phi(c) \in \mathcal{H}(h^{2+\alpha}(\mathbb{S}^1), h^{1+\alpha}(\mathbb{S}^1))$, it is enough (cf. [1]) to find constants $\omega > 0$ and $\kappa \geq 1$ such that

$$\lambda + \partial\Phi(c) \in \mathcal{L}is(h^{2+\alpha}(\mathbb{S}^1), h^{1+\alpha}(\mathbb{S}^1)), \tag{15}$$

$$|\lambda| \cdot \|R(\lambda, -\partial\Phi(c))\|_{\mathcal{L}(h^{1+\alpha}(\mathbb{S}^1))} \leq \kappa \tag{16}$$

for all $\operatorname{Re}\lambda \geq \omega$.

For $r \geq 0$ we introduce the Sobolev space

$$H^r(\mathbb{S}^1) := \left\{ f \in L^2(\mathbb{S}^1) \,:\, \sum_{n\in\mathbb{Z}} (1+n^2)^r |\widehat{f}(n)|^2 < \infty \right\},$$

endowed with the scalar product

$$\langle f \,,\, g \rangle := \sum_{n\in\mathbb{Z}} (1+n^2)^r \widehat{f}(n)\overline{\widehat{g}(n)}.$$

The smooth functions are dense in $H^r(\mathbb{S}^1)$ and the Sobolev embedding $H^{k+r}(\mathbb{S}^1) \hookrightarrow C^k(\mathbb{S}^1)$ holds for all $k \in \mathbb{N}$, provided $r > 1/2$. Therefore we have

$$H^{k+s}(\mathbb{S}^1) \overset{d}{\hookrightarrow} h^{k+\beta}(\mathbb{S}^1)$$

for all $k \in \mathbb{N}$, $\beta \in [0,1]$, and $s > 3/2$.

We choose $\omega = 1$ and prove that relations (15) and (16) hold.

LEMMA 2.4. *Given $r \geq 0$ and $\operatorname{Re}\lambda \geq \omega$, we have*

$$\lambda + \partial\Phi(c) \in \mathcal{L}is(H^{r+1}(\mathbb{S}^1), H^r(\mathbb{S}^1)).$$

Proof. The proof is just a consequence of the fact that

$$\lim_{k\to\infty} \frac{\lambda_k}{k} = -\zeta.$$

Given $r \geq 0$ and $\operatorname{Re}\lambda \geq \omega$, the inverse of $\lambda + \partial\Phi(c)$ is the operator $R(\lambda, -\partial\Phi(c)) \in \mathcal{L}is(H^r(\mathbb{S}^1), H^{r+1}(\mathbb{S}^1))$ defined by

$$R(\lambda, -\partial\Phi(c)) \left[\sum_{k\in\mathbb{Z}} c_k e^{ikx} \right] = \sum_{k\in\mathbb{Z}} \frac{1}{\lambda - \lambda_k} c_k e^{ikx} \tag{17}$$

for all $h := \sum_{k\in\mathbb{Z}} c_k e^{ikx} \in H^r(\mathbb{S}^1)$. □

In view of the above-mentioned facts, the multiplying operator $R(\lambda, -\partial\Phi(c))$ belongs to $\mathcal{L}(h^{1+\alpha}(\mathbb{S}^1), h^{k+\alpha}(\mathbb{S}^1))$, $k \in \{1,2\}$, if it belongs to $\mathcal{L}(C^{1+\alpha}(\mathbb{S}^1), C^{k+\alpha}(\mathbb{S}^1))$, $k \in \{1,2\}$. In the next proposition we show that, with our choice of ω, relation (15) holds; i.e., for $\operatorname{Re}\lambda \geq \omega$ it follows that λ belongs to the resolvent set $\varrho(-\partial\Phi(c))$ of the operator $-\partial\Phi(c)$.

PROPOSITION 2.5.

$$\{\lambda \in \mathbb{C} \,:\, \operatorname{Re}\lambda \geq \omega\} \subset \varrho(-\partial\Phi(c)).$$

Proof. It is sufficient to prove that $R(\lambda, -\partial\Phi(c)) \in \mathcal{L}(C^{1+\alpha}(\mathbb{S}^1), C^{2+\alpha}(\mathbb{S}^1))$ for all $\operatorname{Re}\lambda \geq \omega$. Let $\operatorname{Re}\lambda \geq \omega$. Then $R(\lambda, -\partial\Phi(c)) \in \mathcal{L}(C^{1+\alpha}(\mathbb{S}^1), C^{2+\alpha}(\mathbb{S}^1))$, if we have

$$\text{(i)} \quad \sup_{k\in\mathbb{Z}} |k||M_k| < \infty,$$

$$\text{(ii)} \quad \sup_{k\in\mathbb{Z}} |k|^2|M_{k+1} - M_k| < \infty,$$

$$\text{(iii)} \quad \sup_{k\in\mathbb{Z}} |k|^3|M_{k+2} - 2M_{k+1} + M_k| < \infty,$$

where $M_k := 1/(\lambda - \lambda_k)$ (see, e.g., [2] and [14]). The relation

$$\lim_{k\to\infty} \frac{k}{\lambda - \lambda_k} = \frac{1}{\zeta}$$

implies (i). For $k \geq 1$ we have

$$k^2|M_{k+1} - M_k| = \frac{k}{\lambda - \lambda_{k+1}} \frac{k}{\lambda - \lambda_k} |\lambda_{k+1} - \lambda_k| \underset{k\to\infty}{\longrightarrow} \frac{1}{\zeta},$$

because of $\lambda_k - \lambda_{k+1} \to \zeta$. Also

$$k^3|M_{k+2} - 2M_{k+1} + M_k| = \frac{k}{\lambda - \lambda_{k+2}} \frac{k}{\lambda - \lambda_{k+1}} \frac{k}{\lambda - \lambda_k} |\lambda(\lambda_{k+2} - 2\lambda_{k+1} + \lambda_k) + \lambda_k(\lambda_{k+1} - \lambda_{k+2}) + \lambda_{k+2}(\lambda_{k+1} - \lambda_k)|,$$

and $\lambda_{k+2} - 2\lambda_{k+1} + \lambda_k \to 0$, respectively, $\lambda_k(\lambda_{k+1} - \lambda_{k+2}) + \lambda_{k+2}(\lambda_{k+1} - \lambda_k) \to 2\zeta^2$. This completes the proof. □

PROPOSITION 2.6. *There exists $\kappa \geq 1$ such that*

$$|\lambda| \cdot \|R(\lambda, -\partial\Phi(c))\|_{\mathcal{L}(h^{1+\alpha}(\mathbb{S}^1))} \leq \kappa$$

for all $\operatorname{Re}\lambda \geq \omega$.

Proof. Given $h = \sum_{k\in\mathbb{Z}} \widehat{h}(k)e^{ikx} \in h^{1+\alpha}(\mathbb{S}^1)$, we have

$$|\lambda| R(\lambda, -\partial\Phi(c)) \left[\sum_{k\in\mathbb{Z}} \widehat{h}(k)e^{ikx}\right] = \sum_{k\in\mathbb{Z}} M_k^\lambda \widehat{h}(k)e^{ikx},$$

where

$$M_k^\lambda = \frac{|\lambda|}{\lambda - \lambda_k} \quad \forall\, k \in \mathbb{Z},\ \operatorname{Re}\lambda \geq \omega.$$

It suffices to prove (see [2], [14]) that there exist positive constants s_1, s_2, and s_3 such that

$$\text{(i)} \quad \sup_{k\in\mathbb{Z}} |M_k^\lambda| \leq s_1,$$

$$\text{(ii)} \quad \sup_{k\in\mathbb{Z}} |k||M_{k+1}^\lambda - M_k^\lambda| \leq s_2,$$

$$\text{(iii)} \quad \sup_{k\in\mathbb{Z}} |k|^2|M_{k+2}^\lambda - 2M_{k+1}^\lambda + M_k^\lambda| \leq s_3$$

for all $\operatorname{Re}\lambda \geq \omega$. For $\operatorname{Re}\lambda \geq \omega$ and $k \in \mathbb{Z}$ we have $\operatorname{Re}(\lambda - \lambda_k) \geq 1$ and $|\lambda - \lambda_k| \geq |\lambda|$. Thus, (i) holds with $s_1 = 1$. Moreover, for $k \neq 0$, we have $|\lambda - \lambda_k| \geq |\lambda_k|$ for all $\operatorname{Re}\lambda \geq \omega$; hence

$$|k||M_{k+1}^{\lambda} - M_k^{\lambda}| = \frac{|\lambda|}{|\lambda - \lambda_{k+1}|}\frac{|k|}{|\lambda - \lambda_k|}|\lambda_{k+1} - \lambda_k| \leq \frac{|k|}{|\lambda_k|}|\lambda_{k+1} - \lambda_k|,$$

and (ii) holds. For $k \geq 1$ we further have

$$\begin{aligned} k^2|M_{k+2} - 2M_{k+1} + M_k| &= \frac{|\lambda|}{|\lambda - \lambda_{k+2}|}\frac{k}{|\lambda - \lambda_{k+1}|}\frac{k}{|\lambda - \lambda_k|}|\lambda(\lambda_{k+2} - 2\lambda_{k+1} + \lambda_k) \\ &\quad + \lambda_k(\lambda_{k+1} - \lambda_{k+2}) + \lambda_{k+2}(\lambda_{k+1} - \lambda_k)| \\ &\leq \frac{k}{|\lambda_k|}|k(\lambda_{k+2} - 2\lambda_{k+1} + \lambda_k)| \\ &\quad + \frac{k}{|\lambda_k|}\frac{k}{|\lambda_{k+1}|}|\lambda_k(\lambda_{k+1} - \lambda_{k+2}) + \lambda_{k+2}(\lambda_{k+1} - \lambda_k)|. \end{aligned}$$

Taking into account

$$k(\lambda_{k+2} - 2\lambda_{k+1} + \lambda_k) \underset{k\to\infty}{\longrightarrow} 0,$$

and using the symmetry of the coefficients λ_k, we have (iii). ∎

In conclusion, given $c \in \mathbb{R}_{>0}$, we have proved that $-\partial\Phi(c)$ generates a strongly continuous analytic semigroup in $\mathcal{L}(h^{1+\alpha}(\mathbb{S}^1))$. The proof of Theorem 1.1 is similar to that of Theorem 8.1.1 in [13], and the assumptions of this theorem are all satisfied (see also Theorem 2.3).

3. Stability of the equilibria. We want to identify the steady states of this moving boundary problem, i.e., solutions which do not depend on time, and to study their stability. A pair $(u, f) \in buc^{2+\alpha}(\Omega_f) \times \mathcal{V}$ is a stationary solution of the flow (3) iff it is a solution of the free boundary problem

$$\begin{aligned} \operatorname{div}\left(\tfrac{Du}{\overline{\mu}(|Du|^2)}\right) &= 0 \quad \text{in} \quad \Omega_f, \\ \partial_\nu u &= 0 \quad \text{on} \quad \Gamma_0, \\ u &= f \quad \text{on} \quad \Gamma_f, \\ \partial_\nu u &= 0 \quad \text{on} \quad \Gamma_f. \end{aligned} \tag{18}$$

If $f \in \mathcal{V}$ is known, then $u \in buc^{2+\alpha}(\Omega_f)$ is a solution of the Neumann problem

$$\begin{aligned} \operatorname{div}\left(\tfrac{Du}{\overline{\mu}(|Du|^2)}\right) &= 0 \quad \text{in} \quad \Omega_f, \\ \partial_\nu u &= 0 \quad \text{on} \quad \partial\Omega_f. \end{aligned} \tag{19}$$

Thus there exists a constant $c \in \mathbb{R}$ such that $u = c$. The condition $u = f$ on Γ_f, together with $f \in \mathcal{V}$, implies $c \in \mathbb{R}_{>0}$. In conclusion, $(c, c) \in buc^{2+\alpha}(\Omega_c) \times \mathcal{V}$, $c > 0$, are the stationary solutions of the flow (3), or, equivalently, $(c, c) \in buc^{2+\alpha}(G) \times \mathcal{V}$, $c > 0$, are the stationary solutions of the flow (7).

Let $c \in \mathbb{R}_{>0}$ be given. We want to identify the spectrum of the operator $-\partial\Phi(c)$. Of course, we have

$$\{\lambda_k \,:\, k \in \mathbb{N}\} \subset \sigma_p(-\partial\Phi(c)) \subset \sigma(-\partial\Phi(c)).$$

Using a similar argument as in Proposition 2.5, one can in fact show that

$$\{\lambda_k \,:\, k \in \mathbb{N}\} = \sigma_p(-\partial\Phi(c)) = \sigma(-\partial\Phi(c)).$$

Thus, the spectrum of $-\partial\Phi(c)$ contains just eigenvalues. The largest eigenvalue is $\lambda_0 = 0$. This eigenvalue is simple and has a one-dimensional eigenspace consisting of constant functions. All other eigenvalues λ_k, $k \geq 1$, have a two-dimensional eigenspace spanned by $\{e^{ikx}, e^{-ikx}\}$. Compared to [5], the analysis is more involved here, since we have to take the eigenvalue $\lambda_0 = 0$ into consideration.

In order to prove the stability we transfer problem (9) into a neighborhood of the origin in $h^{2+\alpha}(\mathbb{S}^1)$. Let $\mathcal{V}_c := \mathcal{V} - c$. Then $\mathcal{V}_c$ is an open neighborhood of the origin in $h^{2+\alpha}(\mathbb{S}^1)$. We define $\psi : \mathcal{V}_c \to h^{1+\alpha}(\mathbb{S}^1)$ by $\psi(f) := -\Phi(f + c)$ and consider the following Cauchy problem:

$$\partial_t f = \psi(f), \qquad f(0) = f_0. \tag{20}$$

We have $\psi(0) = -\Phi(c) = 0$ and $\partial\psi(0) = -\partial\Phi(c)$. Thus, 0 is a stationary solution of (20), and there exists a one-to-one correspondence between solutions to (9) and (20). Namely, if f is a solution to (20) corresponding to the initial data $f_0 \in \mathcal{V}_c$, then $f + c$ is the solution to (9) corresponding to $f_0 + c \in \mathcal{V}$. Conversely, if f is a solution to (9) corresponding to the initial data $f_0 \in \mathcal{V}$, then $f - c$ is the solution to (20) corresponding to $f_0 - c \in \mathcal{V}_c$.

Hence, instead of studying the stability of the steady state c for (9), we study the stability of 0 for (20). Note that 0 is an eigenvalue of the derivative $\partial\psi(0)$. Moreover, the null solution is not asymptotically stable because each neighborhood of 0 in $\mathcal{V}_c$ contains constants, which are also steady states of this flow.

However, it is inappropriate to pose the problem in this way because, as the next lemma shows, we have conservation of the fluid volume.

We introduce first some notation. Define by $\widetilde{h}^{k+\alpha}(\mathbb{S}^1)$, $k \in \{1, 2\}$, the closure of the set of all smooth functions on $\mathbb{S}^1$ with mean 0 in $C^{k+\alpha}(\mathbb{S}^1)$. We have defined in this way closed subspaces of the little Hölder spaces. It is obvious that

$$\widetilde{h}^{2+\alpha}(\mathbb{S}^1) \overset{d}{\hookrightarrow} \widetilde{h}^{1+\alpha}(\mathbb{S}^1) \quad \text{and} \quad \widetilde{h}^{k+\alpha}(\mathbb{S}^1) = \left\{ f \in h^{k+\alpha}(\mathbb{S}^1) \,:\, \int_{\mathbb{S}^1} f\,dx = 0 \right\},\ k \in \{1, 2\}.$$

Further, let $\widetilde{\mathcal{V}} := \mathcal{V}_c \cap \widetilde{h}^{2+\alpha}(\mathbb{S}^1)$.

LEMMA 3.1 (conservation of volume). *Given $g \in \widetilde{\mathcal{V}}$, we have $\psi(g) \in \widetilde{h}^{1+\alpha}(\mathbb{S}^1)$.*

Proof. Let $g + c = f \in \mathcal{V}$ and denote by u the solution of (6). We have

$$\begin{aligned}
\int_{\mathbb{S}^1} \psi(g)dx &= -\int_{\mathbb{S}^1} \Phi(f)dx = -\int_{\mathbb{S}^1} \mathcal{B}(f, \mathcal{T}(f))dx \\
&= -\int_{\mathbb{S}^1} \frac{Du}{\overline{\mu}(|Du|^2)}(x, f(x)) \cdot n(x)\, dx = -\int_{\Gamma_f} \frac{\partial_\nu u}{\overline{\mu}(|Du|^2)}\, d\sigma \\
&= -\int_{\Omega_f} \operatorname{div}\left(\frac{Du}{\overline{\mu}(|Du|^2)}\right) dx + \int_{\Gamma_0} \frac{\partial_\nu u}{\overline{\mu}(|Du|^2)}\, d\sigma = 0. \qquad \square
\end{aligned}$$

We consider now the restriction $\widetilde{\psi}$ of ψ to $\widetilde{\mathcal{V}}$. From Lemma 3.1 and Theorem 2.3 we conclude that $\widetilde{\psi} \in C^\infty(\widetilde{\mathcal{V}}, \widetilde{h}^{1+\alpha}(\mathbb{S}^1))$. Given $h = \sum_{k\in\mathbb{Z}} \widehat{h}(k)e^{ikx} \in \widetilde{h}^{2+\alpha}(\mathbb{S}^1))$ we have

$$\partial\widetilde{\psi}(0)\left[\sum_{k\in\mathbb{Z}} \widehat{h}(k)e^{ikx}\right] = \sum_{k\in\mathbb{Z}} \lambda_k \widehat{h}(k)e^{ikx},$$

where the coefficients (λ_k) are given by (14).

THEOREM 3.2. *Consider $\partial\widetilde{\psi}(0)$ as an unbounded operator in $\widetilde{h}^{1+\alpha}(\mathbb{S}^1)$ with dense domain $\widetilde{h}^{2+\alpha}(\mathbb{S}^1)$. Then*

$$\sigma(\partial\widetilde{\psi}(0)) = \{\lambda_k \,:\, k \geq 1\}, \tag{21}$$

$$|\lambda| \cdot \|R(\lambda, -\partial\widetilde{\psi}(0))\|_{\mathcal{L}(\widetilde{h}^{1+\alpha}(\mathbb{S}^1))} \leq \kappa \quad \forall\, \mathrm{Re}\,\lambda \geq \omega, \tag{22}$$

where κ and ω are the constants from Proposition 2.6.

Proof. Given $\lambda \in \varrho(\partial\psi(0))$ and $h = \sum_{k\in\mathbb{Z}} \widehat{h}(k)e^{ikx} \in \widetilde{h}^{1+\alpha}(\mathbb{S}^1)$ there exists a unique

$$u = \sum_{k\in\mathbb{Z}} \frac{1}{\lambda - \lambda_k}\widehat{h}(k)e^{ikx} \in h^{2+\alpha}(\mathbb{S}^1),$$

such that $(\lambda - \partial\psi(0))u = h$. Since $\widehat{h}(0) = 0$ we conclude that $\widehat{u}(0) = 0$; thus $u \in \widetilde{h}^{2+\alpha}(\mathbb{S}^1)$ and $\lambda \in \varrho(\partial\widetilde{\psi}(0))$. Moreover, $\partial\widetilde{\psi}(0) \in \mathcal{L}is(\widetilde{h}^{2+\alpha}(\mathbb{S}^1), \widetilde{h}^{1+\alpha}(\mathbb{S}^1))$ and (21) holds. Relation (22) follows from Proposition 2.6, using the fact that

$$(\lambda - \partial\psi(0))[h] = (\lambda - \partial\widetilde{\psi}(0))[h],$$

for all $h \in \widetilde{h}^{2+\alpha}(\mathbb{S}^1)$ and $\mathrm{Re}\,\lambda \geq \omega$. □

In conclusion $-\partial\widetilde{\psi}(0)$ belongs to $\mathcal{H}(\widetilde{h}^{2+\alpha}(\mathbb{S}^1), \widetilde{h}^{1+\alpha}(\mathbb{S}^1)$, and therefore, if the initial value f_0 in (20) belongs to $\widetilde{h}^{2+\alpha}(\mathbb{S}^1)$, then the evolution takes place in $\widetilde{h}^{2+\alpha}(\mathbb{S}^1)$. We have the following stability result for the flow (3).

THEOREM 3.3 (exponential stability). *For any $\omega < -\lambda_1 = \tanh c/\overline{\mu}(0)$, there are positive constants r and C such that for any $f_0 \in c + \widetilde{h}^{2+\alpha}(\mathbb{S}^1)$ with $\|f_0 - c\|_{C^{2+\alpha}(\mathbb{S}^1)} \leq r$ the solution to* (3) *exists globally in time and the estimate*

$$\|f(t) - c\|_{C^{2+\alpha}(\mathbb{S}^1)} + \|f'(t)\|_{C^{1+\alpha}(\mathbb{S}^1)} \leq Ce^{-\omega t}\|f_0 - c\|_{C^{2+\alpha}(\mathbb{S}^1)} \quad \forall t \geq 0$$

holds.

Proof. Transferring problem (9) into a neighborhood of 0 in $\widetilde{h}^{2+\alpha}(\mathbb{S}^1)$ we find that all of the assumptions of Theorem 9.1.2 in [13] are satisfied (cf. Theorem 3.2). The result follows directly from this theorem. □

REFERENCES

[1] H. AMANN, *Linear and Quasilinear Parabolic Problems*, Vol. I, Birkhäuser, Basel, 1995.

[2] W. ARENDT AND S. BU, *Operator-valued Fourier multipliers on periodic Besov spaces and applications*, Proc. Edinb. Math. Soc. (2), 47 (2004), pp. 15–33.

[3] J. BEAR AND Y. BACHMAT, *Introduction to Modeling of Transport Phenomena in Porous Media*, Kluwer Academic Publishers, Boston, 1990.

[4] J. ESCHER, *On moving boundaries in deformable media*, Adv. Math. Sci. Appl., 7 (1997), pp. 275–316.

[5] J. Escher and B.-V. Matioc, *Stability of the equilibria for periodic Stokesian Hele-Shaw flows*, J. Evol. Equ., 8 (2008), pp. 513–522.
[6] J. Escher and B.-V. Matioc, *A moving boundary problem for periodic Stokesian Hele-Shaw flows*, Interfaces Free Bound., to appear.
[7] J. Escher and G. Prokert, *Stability of the equilibria for spatially periodic flows in porous media*, Nonlinear Anal., 45 (2001), pp. 1061–1080.
[8] J. Escher and G. Simonett, *Maximal regularity for a free boundary problem*, NoDEA Nonlinear Differential Equations Appl., 2 (1995), pp. 463–510.
[9] J. Escher and G. Simonett, *Analyticity of the interface in a free boundary problem*, Math. Ann., 305 (1996), pp. 435–459.
[10] J. Escher and G. Simonett, *Classical solutions of multidimensional Hele–Shaw models*, SIAM J. Math. Anal., 28 (1997), pp. 1028–1047.
[11] D. Gilbarg and T. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1997.
[12] L. Kondic, P. Palffy-Mahorny, and M. J. Shelley, *Models of non-Newtonian Hele-Shaw flow*, Phys. Rev. E (3), 54 (1996), pp. R4536–R4539.
[13] A. Lunardi, *Analytic Semigroups and Optimal Regularity in Parabolic Problems*, Birkhäuser, Basel, 1995.
[14] H.-J. Schmeisser and H. Triebel, *Topics in Fourier Analysis and Function Spaces*, John Wiley and Sons, New York, 1987.

SIAM J. MATH. ANAL.
Vol. 40, No. 5, pp. 2007–2033

CONVERGENCE TO EQUILIBRIUM FOR PARABOLIC-HYPERBOLIC TIME-DEPENDENT GINZBURG–LANDAU–MAXWELL EQUATIONS*

MAURIZIO GRASSELLI†, HAO WU‡, AND SONGMU ZHENG§

Abstract. We consider a Ginzburg–Landau–Maxwell model which describes the behavior of a two-dimensional superconducting material. The state variables are the complex-valued order parameter ψ, the magnetic potential $\tilde{\mathbf{A}}$, and the electric potential Φ. Under the choice of Coulomb (i.e., London) gauge, the resulting system is a parabolic-hyperbolic coupled system of nonlinear partial differential equations subject to suitable boundary and initial conditions. Global well-posedness results were proved in [M. Tsutsumi and H. Kasai, *Nonlinear Anal.*, 37 (1999), pp. 187–216], while the existence of global attractor and exponential attractors was proved in [V. Berti and S. Gatti, *Quart. Appl. Math.*, 64 (2006), pp. 617–639]. In this paper we use an extended Łojasiewicz–Simon approach to show that for any initial datum in certain phase space, the corresponding global solution converges to an equilibrium as time goes to infinity. Besides, we also provide an estimate on the convergence rate with respect to the phase space metric.

Key words. Ginzburg–Landau–Maxwell equations, superconductivity, Coulomb gauge, convergence to equilibrium, Łojasiewicz–Simon inequality

AMS subject classifications. Primary, 35B40; Secondary, 82D55

DOI. 10.1137/080717833

1. Introduction.

This paper is concerned with the asymptotic behavior of solutions to a two-dimensional Ginzburg–Landau–Maxwell model of superconductivity proposed in [27] (see also [2, 13, 32]). When a superconducting material is kept close to a certain critical temperature, its behavior can be macroscopically described, within the Ginzburg–Landau phase transition theory (see [18] and its references, cf. also [8, Chapter 11]), by the state variables $(\psi, \tilde{\mathbf{A}}, \Phi)$. Here ψ is the complex order parameter, whose squared modulus represents the concentration of the superconducting electrons, while $\tilde{\mathbf{A}}$ and Φ are the magnetic and the electric potentials, respectively.

On account of [32], we introduce the equations governing the evolution of $(\psi, \tilde{\mathbf{A}}, \Phi)$. Suppose that the material occupies a bounded domain $\Omega \subset \mathbb{R}^2$ with smooth boundary Γ for any time $t \geq 0$. Then the state variables $\psi : \Omega \times (0,\infty) \to \mathbb{C}$, $\tilde{\mathbf{A}} : \Omega \times (0,\infty) \to \mathbb{R}^2$ and $\Phi : \Omega \times (0,\infty) \to \mathbb{R}$ satisfy the following equations:

$$\psi_t - i\Phi\psi - D_{\tilde{\mathbf{A}}}^2\psi - \lambda^2\left(1 - |\psi|^2\right)\psi = 0, \tag{1.1}$$

$$\varepsilon\left(\tilde{\mathbf{A}}_t - \nabla\Phi\right)_t + \sigma\left(\tilde{\mathbf{A}}_t - \nabla\Phi\right) + \mathrm{curl}^2\tilde{\mathbf{A}} + \frac{i}{2}\left(\overline{\psi}D_{\tilde{\mathbf{A}}}\psi - \psi\overline{D_{\tilde{\mathbf{A}}}\psi}\right) + \mathrm{curl}\mathbf{H}_{ext} = 0, \tag{1.2}$$

in $\Omega \times (0,\infty)$, subject to the boundary conditions

$$D_{\tilde{\mathbf{A}}}\psi\cdot\mathbf{n} = 0, \quad \left(\mathrm{curl}\tilde{\mathbf{A}} + \mathbf{H}_{ext}\right)\times\mathbf{n} = 0, \quad \left(\tilde{\mathbf{A}}_t - \nabla\Phi\right)\cdot\mathbf{n} = 0, \quad \text{on } \Gamma\times(0,\infty), \tag{1.3}$$

*Received by the editors March 6, 2008; accepted for publication (in revised form) September 23, 2008; published electronically January 21, 2009.

http://www.siam.org/journals/sima/40-5/71783.html

†Dipartimento di Matematica, Politecnico di Milano, 20133 Milano, Italy (maurizio.grasselli@polimi.it). This author was partially supported by the Italian PRIN Research Project 2006 *Problemi a frontiera libera, transizioni di fase e modelli di isteresi*.

‡School of Mathematical Sciences, Fudan University, Shanghai 200433, China (haowufd@yahoo.com). This author was supported by the China Postdoctoral Science Foundation.

§Corresponding author. Institute of Mathematics, Fudan University, Shanghai 200433, China (songmuzheng@yahoo.com). This author was supported by NSF of China under the grant 10631020 and by the Chinese Ministry of Education under grant 20050246002.

where

$$D_{\tilde{\mathbf{A}}}\psi = \nabla\psi - i\tilde{\mathbf{A}}\psi, \quad D^2_{\tilde{\mathbf{A}}}\psi = \Delta\psi - 2i\tilde{\mathbf{A}}\cdot\nabla\psi - i\psi\nabla\cdot\tilde{\mathbf{A}} - |\tilde{\mathbf{A}}|^2\psi. \tag{1.4}$$

Here $\mathbf{H}_{ext}$ is the external magnetic field, which is assumed to be time independent for the sake of simplicity throughout this paper; the positive constants λ, ε, and σ represent the Ginzburg–Landau parameter, the dielectric coefficient, and the electric conductivity, respectively, and $\mathbf{n}$ denotes the unit outward normal to the boundary Γ.

We recall that (1.2) is derived from Maxwell's equations. In the so-called quasi-steady approximation, namely, when the displacement current $\varepsilon(\tilde{\mathbf{A}}_t - \nabla\Phi)_t$ is negligible, we obtain the well-known system proposed by Gor'kov and Èliashberg [13], which has been widely studied in the literature (see, e.g., [7, 9, 11, 18, 19, 22, 23, 24, 26, 27, 29, 30, 33] and reference therein). However, if surface charge must be taken into account, then the displacement current cannot be neglected (see [2]).

Similarly to the classical time-dependent Ginzburg–Landau equations, the evolution equations (1.1)–(1.2) and the boundary conditions (1.3) are invariant under the gauge transformation:

$$\mathcal{T}_\chi : \left(\psi, \tilde{\mathbf{A}}, \Phi\right) \longmapsto \left(\psi e^{i\chi}, \tilde{\mathbf{A}} + \nabla\chi, \Phi + \chi_t\right). \tag{1.5}$$

The gauge χ can be any (smooth) real scalar-valued function depending on x and t. In [3, 32], the authors choose the Coulomb gauge (also known as the London gauge). This choice entails that the following identities hold:

$$\mathrm{div}\tilde{\mathbf{A}} = 0 \quad \text{in } \Omega, \quad \int_\Omega \Phi dx = 0, \tag{1.6}$$

$$\tilde{\mathbf{A}}\cdot\mathbf{n} = 0, \quad \partial_{\mathbf{n}}\Phi = 0, \quad \text{on } \Gamma. \tag{1.7}$$

For the existence of the gauge χ corresponding to (1.6) and (1.7), one may refer, e.g., to [23, Lemma 2.2, Remark 2.4] where even more general cases have been discussed (see also [18, section 1.4.4]). Using the Coulomb gauge, $\tilde{\mathbf{A}}$ is a solenoidal vector field so that

$$D^2_{\tilde{\mathbf{A}}}\psi = \Delta\psi - 2i\tilde{\mathbf{A}}\cdot\nabla\psi - \left|\tilde{\mathbf{A}}\right|^2\psi. \tag{1.8}$$

Moreover, system (1.1)–(1.2) is subject to the boundary conditions (cf. (1.3), (1.7))

$$\partial_{\mathbf{n}}\psi = 0, \quad \tilde{\mathbf{A}}\cdot\mathbf{n} = 0, \quad \left(\mathrm{curl}\tilde{\mathbf{A}} + \mathbf{H}_{ext}\right)\times\mathbf{n} = 0, \quad \text{on } \Gamma\times(0,\infty). \tag{1.9}$$

It will be more convenient to treat problem (1.1), (1.2), (1.9) in a homogeneous form. For this purpose, we consider the following boundary value problem:

$$\begin{cases} \mathrm{curl}^{\,2}\mathbf{A}_{ext} = \mathrm{curl}\mathbf{H}_{ext}, & \mathrm{div}\mathbf{A}_{ext} = 0, & \text{in } \Omega, \\ \mathbf{A}_{ext}\cdot\mathbf{n} = 0, & (\mathrm{curl}\mathbf{A}_{ext} - \mathbf{H}_{ext})\times\mathbf{n} = 0, & \text{on } \Gamma. \end{cases} \tag{1.10}$$

Throughout this paper we always assume that $\mathbf{H}_{ext} \in H^1$. Then, it is easy to see that the convex quadratic functional

$$J(\mathbf{A}) = \int_\Omega |\mathrm{curl}\mathbf{A} - \mathbf{H}_{ext}|^2 dx \tag{1.11}$$

on the domain

$$\mathcal{D} := \left\{\mathbf{A}\in\mathbf{H}^{\,1}(\Omega) :\ \mathrm{div}\mathbf{A} = 0 \ \text{ in } \Omega, \ \ \mathbf{A}\cdot\mathbf{n}|_\Gamma = 0\right\} \tag{1.12}$$

admits a unique minimizer $\mathbf{A}_{ext} \in \mathcal{D}$, which is the (unique) solution to (1.10). The linear mapping $\mathbf{H}_{ext} \to \mathbf{A}_{ext}$ is continuous from $\mathbf{H}^{\alpha}(\Omega)$ to $\mathbf{H}^{\alpha+1}(\Omega)$ for $0 \le \alpha \le 1$ (see [11, Lemma 3, Lemma 4]; cf. also [9]).

Set now

$$\mathbf{A} = \tilde{\mathbf{A}} + \mathbf{A}_{ext}. \tag{1.13}$$

Then the Ginzburg–Landau–Maxwell equations (1.1) and (1.2) can be reduced to the following homogeneous form in terms of ψ and $\mathbf{A}$ (cf. [3,32]):

$$\psi_t - i\Phi\psi - D^2_{\tilde{\mathbf{A}}}\psi - \lambda^2\left(1 - |\psi|^2\right)\psi = 0, \tag{1.14}$$

$$\varepsilon(\mathbf{A}_t - \nabla\Phi)_t + \sigma(\mathbf{A}_t - \nabla\Phi) + \mathrm{curl}^2\mathbf{A} + \frac{i}{2}\left(\overline{\psi}D_{\tilde{\mathbf{A}}}\psi - \psi\overline{D_{\tilde{\mathbf{A}}}\psi}\right) = 0. \tag{1.15}$$

By applying the spatial divergence operator to (1.15) and taking (1.14) into account, we obtain the evolution equation for Φ, namely,

$$-\varepsilon\Delta\Phi_t - \sigma\Delta\Phi + \frac{i}{2}\left(\overline{\psi}\psi_t - \psi\overline{\psi}_t\right) + |\psi|^2\Phi = 0. \tag{1.16}$$

Moreover, we have (see (1.5))

$$\mathrm{div}\mathbf{A} = 0 \quad \text{in } \Omega, \quad \int_{\Omega}\Phi dx = 0, \tag{1.17}$$

and (1.14)–(1.17) are subject to the homogeneous boundary conditions (cf. (1.7)–(1.10))

$$\partial_{\mathbf{n}}\psi = 0, \quad \mathbf{A}\cdot\mathbf{n} = 0, \quad \mathrm{curl}\mathbf{A}\times\mathbf{n} = 0, \quad \partial_{\mathbf{n}}\Phi = 0, \quad \text{on } \Gamma\times(0,\infty), \tag{1.18}$$

as well as to the initial conditions

$$\psi|_{t=0} = \psi_0, \quad \mathbf{A}|_{t=0} = \mathbf{A}_0, \quad \mathbf{A}_t|_{t=0} = \mathbf{A}_1, \quad \Phi|_{t=0} = \Phi_0, \quad \text{in } \Omega \tag{1.19}$$

for some given initial data.

The main concern of this paper is the convergence to equilibrium of the solution to problem (1.14)–(1.19) as time goes to infinity. In what follows we recall related results in the literature on problem (1.14)–(1.19). The global well-posedness of problem (1.14)–(1.19) has been carefully investigated in [32] in both two and three dimensions, even though, physically speaking, the present model is essentially two-dimensional (see [18, sections 1.5, 1.6, 1.8]; cf. also [5,23]). The results of [32] are also concerned with the case $\varepsilon = 0$, which has been studied by many other authors (see, e.g., [18,23, 26,29,30,33] and references therein). Regarding the large time behavior of solutions, problem (1.14)–(1.19) has recently been analyzed in [3] within the theory of infinite-dimensional dissipative dynamical systems. In that paper the authors show that the quoted problem generates a strongly continuous semigroup which possesses the global attractor and an exponential attractor. However, the issue of convergence to equilibrium for single solutions to problem (1.14)–(1.19) has not been studied in [3]. This is the goal of the present contribution. More precisely, we will prove that, for any initial datum in the phase space introduced in [3], the corresponding global solution to (1.14)–(1.19) converges to a single equilibrium as time goes to infinity. Besides, an estimate of the convergence rate in the phase space metric will also be given. The key ingredients of the present paper can be summarized as follows:

(I) Since (1.15) is hyperbolic, we can no longer take advantage of the smoothing effects for solutions to parabolic equations like, e.g., in [9] where the Lorentz gauge is used. Instead, we will use a result proved in [3] to show the precompactness of solutions.

(II) We need to develop an extended Łojasiewicz–Simon approach for which it is necessary to establish a suitable Łojasiewicz–Simon-type inequality. Such an inequality is a suitable adaptation of the one proved in [9] for a problem similar to ours with $\varepsilon = 0$ but subject to the Lorentz gauge $\Phi = -\omega(\nabla \cdot \mathbf{A})$, with ω being a nonnegative constant (for $\omega = 0$, see also [24]).

(III) Due to the hyperbolic nature of (1.15), the standard Łojasiewicz–Simon approach used in the parabolic case must be modified by introducing an appropriate auxiliary functional $\mathfrak{F}$ (see section 5), which usually depends on the problem under consideration (see, e.g., [15, 19, 35] and references therein for similar problems).

(IV) In the existing literature the convergence rate in a (lower order) norm is usually obtained directly by using the Łojasiewicz–Simon approach (see, e.g., [9, 17, 36]). Then, estimates in higher order norms can be deduced by means of interpolation inequalities (cf. [17]) and, consequently, the decay exponent deteriorates. On the contrary, in this paper and some earlier papers by the authors, using the original equations and suitable energy estimates and constructing proper differential inequalities, we are able to preserve the original convergence rate when we estimate the convergence with respect to the phase space metric. This technique has been successfully applied to other equations as well (see, e.g., [14, 34, 35, 36]).

In [9] (see also [19]) the convergence to single equilibria for the case $\varepsilon = 0$ has been investigated in detail in the hypothesis of a time-dependent applied magnetic field and using the Lorentz gauge. The present contribution is a first step towards a more comprehensive analysis of the case $\varepsilon > 0$. The corresponding results for other widely used gauges (e.g., the Lorentz gauge) as well as for the case of a time-dependent applied magnetic field will be considered in the near future. We recall that the choice of the gauge plays an important role in the mathematical treatment of these models (see [10] and references therein).

This paper is organized as follows. In section 2 we introduce the notation and state the main results of this paper. Section 3 contains some preliminary results on the time-dependent Ginzburg–Landau–Maxwell model. In section 4 we state a suitable version of a Łojasiewicz–Simon-type inequality. The last two sections are devoted to the proof of our main results. More precisely, in section 5 we prove the convergence to a single equilibrium, while in section 6, we establish the convergence rate estimate.

2. Notation and main results. We use the notation introduced in [32] (see also [3]). As usual, $L^p(\Omega)$ and $W^{k,p}(\Omega)$ stand for the Lebesgue and the Sobolev spaces of real-valued functions, with the convention that $H^k(\Omega) = W^{k,2}(\Omega)$. We denote by bold letters the spaces of vector-valued functions, whereas a subscript $\mathbb{C}$ characterizes those of complex-valued functions. Without further specifications, $\|\cdot\|$ stands for the norm in $L^2_{\mathbb{C}}(\Omega), \mathbf{L}^2(\Omega)$, or $L^2(\Omega)$, according to the context. This norm is always induced by the scalar inner product

$$\langle u, v\rangle = \int_\Omega u(x)\overline{v}(x)dx,$$

where for a vector-valued function, the product $u\overline{v}$ is replaced by the Euclidean inner product $\mathbf{u}\cdot\overline{\mathbf{v}}$. Next, we introduce the function spaces

$$\mathbf{X} = \left\{\mathbf{u} \in \mathbf{H}^2(\Omega) :\ \mathrm{div}\mathbf{u} = 0 \text{ in } \Omega,\ \ \mathbf{u}\cdot\mathbf{n}|_\Gamma = 0,\ \ \mathrm{curl}\mathbf{u}\times\mathbf{n}|_\Gamma = \mathbf{0}\right\} \tag{2.1}$$

equipped with the usual $\mathbf{H}^2(\Omega)$ norm. Let $\mathbf{X}_0$ be the completion of $\mathbf{X}$ in $\mathbf{H}^1(\Omega)$ and

$$H^1_{0m}(\Omega) = \left\{ u \in H^1(\Omega) : \int_\Omega u dx = 0 \right\},$$
$$H^2_{0m}(\Omega) = \left\{ u \in H^2(\Omega) : \partial_{\mathbf{n}} u|_\Gamma = 0, \int_\Omega u dx = 0 \right\}.$$

We can thus define the phase space we shall work with:

$$\mathbb{X}^0_\infty = \left\{ \psi \in H^2_{\mathbb{C}}(\Omega) : \partial_{\mathbf{n}} \psi|_\Gamma = 0, \ \|\psi\|_{L^\infty_{\mathbb{C}}(\Omega)} \le 1 \right\} \times \mathbf{X} \times \mathbf{X}_0 \times H^2_{0m}(\Omega), \tag{2.2}$$

which is a closed subset of the Banach space

$$\mathbb{X}^0 = \left\{ \psi \in H^2_{\mathbb{C}}(\Omega) : \partial_{\mathbf{n}} \psi|_\Gamma = 0 \right\} \times \mathbf{X} \times \mathbf{X}_0 \times H^2_{0m}(\Omega). \tag{2.3}$$

The norm on $\mathbb{X}^0$ is defined by

$$\|z\|_0^2 = \|z_1\|^2_{H^2_{\mathbb{C}}(\Omega)} + \|\mathbf{z}_2\|^2_{\mathbf{H}^2(\Omega)} + \|\mathbf{z}_3\|^2_{\mathbf{H}^1(\Omega)} + \|z_4\|^2_{H^2(\Omega)} \tag{2.4}$$

for any $z = (z_1, \mathbf{z}_2, \mathbf{z}_3, z_4) \in \mathbb{X}^0$. We also introduce the Banach space

$$\mathbb{X}^{-1} = H^1_{\mathbb{C}}(\Omega) \times \mathbf{X}_0 \times \mathbf{L}^2(\Omega) \times H^1_{0m} \supset \mathbb{X}^0, \tag{2.5}$$

endowed with the norm

$$\|z\|_{-1}^2 = \|z_1\|^2_{H^1_{\mathbb{C}}(\Omega)} + \|\mathbf{z}_2\|^2_{\mathbf{H}^1(\Omega)} + \|\mathbf{z}_3\|^2_{\mathbf{L}^2(\Omega)} + \|z_4\|^2_{H^1(\Omega)} \tag{2.6}$$

for any $z = (z_1, \mathbf{z}_2, \mathbf{z}_3, z_4) \in \mathbb{X}^{-1}$.

In what follows, we will make use of the Gagliardo–Nirenberg inequalities in dimension two, namely,

$$\|u\|^2_{L^4(\Omega)} \le C\|u\|\|u\|_{H^1(\Omega)} \qquad \forall\ u \in H^1(\Omega), \tag{2.7}$$

$$\|u\|^2_{L^\infty(\Omega)} \le C\|u\|\|u\|_{H^2(\Omega)} \qquad \forall\ u \in H^2(\Omega), \tag{2.8}$$

where $C > 0$ depends only on Ω.

We also recall that there exist constants $\kappa > 0$ and $c > 0$ depending only on Ω such that (see, e.g., [25, Corollary 3.51])

$$\|\mathbf{A}\| \le \kappa \|\mathrm{curl}\mathbf{A}\| \qquad \forall\ \mathbf{A} \in \mathbf{X}_0, \tag{2.9}$$

and (cf. [12, Chapter I, (5.45), p. 92] also [32, section 2])

$$\|\mathbf{A}\|_{\mathbf{H}^1(\Omega)} \le c\|\mathrm{curl}\mathbf{A}\| \qquad \forall\ \mathbf{A} \in \mathbf{X}_0. \tag{2.10}$$

In the remaining part of the paper, we always assume that

$$\mathbf{A}_{ext} \in \mathbf{X}_0 \cap \mathbf{H}^2(\Omega). \tag{2.11}$$

We now recall the following result.

PROPOSITION 2.1 ([3, Theorem 2.1, Proposition 3.1]). *Denote* $z_0 = (\psi_0, \mathbf{A}_0, \mathbf{A}_1, \Phi_0) \in \mathbb{X}^0$. *Problem* (1.14)–(1.19) *generates a strongly continuous semigroup* $\{S(t)\}_{t\ge 0}$

on the space $\mathbb{X}^0$. *If* $z_0 = (\psi_0, \mathbf{A}_0, \mathbf{A}_1, \Phi_0) \in \mathbb{X}^0_\infty$, *then* $S(t)z_0 \in \mathbb{X}^0_\infty$. *Moreover, there exists a positive constant* C *depending only on* $\|z_0\|_0$ *such that*

$$\|S(t)z_0\|_0 \le C \qquad \forall\ t \ge 0. \tag{2.12}$$

The stationary problem corresponding to (1.14)–(1.19) is

$$-(\nabla - i(\mathbf{A}_\infty - \mathbf{A}_{ext}))^2\psi_\infty - \lambda^2\left(1 - |\psi_\infty|^2\right)\psi_\infty = 0, \tag{2.13}$$

$$\mathrm{curl}^2\mathbf{A}_\infty + \frac{i}{2}\left(\overline{\psi}_\infty\nabla\psi_\infty - \psi_\infty\nabla\overline{\psi}_\infty\right) + |\psi|^2(\mathbf{A}_\infty - \mathbf{A}_{ext}) = 0, \tag{2.14}$$

$$\mathrm{div}\mathbf{A}_\infty = 0, \tag{2.15}$$

$$\partial_{\mathbf{n}}\psi_\infty = 0, \quad \mathbf{A}_\infty\cdot\mathbf{n} = 0, \quad \mathrm{curl}\mathbf{A}_\infty\times\mathbf{n} = 0, \quad \text{on } \Gamma, \tag{2.16}$$

$$\Phi_\infty = 0. \tag{2.17}$$

Set now

$$\mathbf{A}_\infty = \tilde{\mathbf{A}}_\infty + \mathbf{A}_{ext}. \tag{2.18}$$

Then problem (2.13)–(2.17) becomes

$$-\left(\nabla - i\tilde{\mathbf{A}}_\infty\right)^2\psi_\infty - \lambda^2\left(1 - |\psi_\infty|^2\right)\psi_\infty = 0, \tag{2.19}$$

$$\mathrm{curl}^2\tilde{\mathbf{A}}_\infty + \frac{i}{2}\left(\overline{\psi}_\infty\nabla\psi_\infty - \psi_\infty\nabla\overline{\psi}_\infty\right) + |\psi_\infty|^2\tilde{\mathbf{A}}_\infty + \mathrm{curl}\mathbf{H}_{ext} = 0, \tag{2.20}$$

$$\mathrm{div}\tilde{\mathbf{A}}_\infty = 0, \tag{2.21}$$

$$\partial_{\mathbf{n}}\psi_\infty = 0, \quad \tilde{\mathbf{A}}_\infty\cdot\mathbf{n} = 0, \quad \left(\mathrm{curl}\tilde{\mathbf{A}}_\infty + \mathbf{H}_{ext}\right)\times\mathbf{n} = 0, \quad \text{on } \Gamma, \tag{2.22}$$

and

$$\Phi_\infty = 0. \tag{2.23}$$

Define the functional

$$\begin{aligned}\mathcal{E}_0\left(\psi_\infty, \tilde{\mathbf{A}}_\infty\right) = \int_\Omega \Bigg[&\frac{\lambda^2}{4}\left(1 - |\psi_\infty|^2\right)^2 + \frac{1}{2}\left|\left(\nabla - i\tilde{\mathbf{A}}_\infty\right)\psi_\infty\right|^2 \\ &+ \frac{1}{2}\left|\mathrm{curl}\tilde{\mathbf{A}}_\infty + \mathbf{H}_{ext}\right|^2\Bigg]dx,\end{aligned} \tag{2.24}$$

on $H^1_{\mathbb{C}}\times\mathbf{X}_0$. It is easy to see that a critical point $(\psi_\infty, \tilde{\mathbf{A}}_\infty)$ of $\mathcal{E}_0$ is a weak solution of (2.19)–(2.22). This stationary problem embodies the macroscopic quantum-mechanical nature of the superconducting state. There are many works on the stationary Ginzburg–Landau equations. For instance, the existence of vortex-like solutions to the time-independent Ginzburg–Landau equations was proved in [24]. An existence result for the stationary problem of the Ginzburg–Landau–Maxwell equations with Coulomb gauge in the L^p framework was recently given in [1]. An equivalence relation between solutions of the time-independent and time-dependent Ginzburg–Landau equations that describe the same physical state of a superconductor was established in [21].

We are now in a position to state our main theorem.

THEOREM 2.1. *For any initial datum* $z_0 = (\psi_0, \mathbf{A}_0, \mathbf{A}_1, \Phi_0) \in \mathbb{X}^0_\infty$, *the global solution* $z(t) = (\psi(t), \mathbf{A}(t), \mathbf{A}_t(t), \Phi(t))$ *to problem* (1.14)–(1.19) *converges in* $\mathbb{X}^0$ *to a unique equilibrium* $z_\infty = (\psi_\infty, \mathbf{A}_\infty, \mathbf{0}, 0)$, *where* $(\psi_\infty, \mathbf{A}_\infty)$ *satisfies* (2.13)–(2.16). *Moreover, the following estimate holds:*

$$\|z(t) - z_\infty\|_0 \le C(1+t)^{-\theta/(1-2\theta)} \qquad \forall\ t \ge 0, \tag{2.25}$$

where $C > 0$ *is a constant depending only on* $\|z_0\|_0$ *and* $\theta \in (0, \frac{1}{2})$ *depends on* z_∞.

Remark 2.1. We recall that θ is the so-called Łojasiewicz exponent (see Lemma 4.2 below). Note that the convergence rate is estimated in the same norm where convergence takes place. This requires some work (see section 6).

3. Preliminaries. In this section we first recall the basic results proven in [3], which entail the precompactness of the trajectory originated from an initial datum in $\mathbb{X}^0_\infty$.

Let

$$\mathbb{X}^1 = H^3_{\mathbb{C}}(\Omega) \times \left[\mathbf{X} \cap \mathbf{H}^3(\Omega)\right] \times \mathbf{X} \times \left[H^2_{0m}(\Omega) \cap H^3(\Omega)\right], \tag{3.1}$$

with the standard norm defined by

$$\|z\|_1^2 = \|z_1\|^2_{H^3_{\mathbb{C}}(\Omega)} + \|\mathbf{z}_2\|^2_{\mathbf{H}^3(\Omega)} + \|\mathbf{z}_3\|^2_{\mathbf{H}^2(\Omega)} + \|z_4\|^2_{H^3(\Omega)} \tag{3.2}$$

for any $z = (z_1, \mathbf{z}_2, \mathbf{z}_3, z_4) \in \mathbb{X}^1$.

For any given $z_0 \in \mathbb{X}^0_\infty$, let $z(t) = S(t)z_0$ be the solution to problem (1.14)–(1.19). Following [3, section 4], in order to prove the precompactness of $z(t)$, we make a decomposition of the solution z into a uniformly stable part z_d which decays to zero and a compact part z_c, that is,

$$z(t) = z_d(t) + z_c(t). \tag{3.3}$$

Here $z_d(t) = (\psi^d(t), \mathbf{A}^d(t), \mathbf{A}^d_t(t), \Phi^d(t))$ solves

$$\begin{cases} \psi^d_t - \Delta\psi^d + \psi^d = 0, \\ \varepsilon\left(\mathbf{A}^d_t - \nabla\Phi^d\right)_t + \sigma\left(\mathbf{A}^d_t - \nabla\Phi^d\right) + \operatorname{curl}^2\mathbf{A}^d = 0, \\ -\varepsilon\Delta\Phi^d_t - \sigma\Delta\Phi^d = 0, \\ \operatorname{div}\mathbf{A}^d = 0, \quad \int_\Omega \Phi^d = 0, \\ \partial_{\mathbf{n}}\psi^d\big|_\Gamma = 0, \quad \mathbf{A}^d\cdot\mathbf{n}\big|_\Gamma = 0, \ \operatorname{curl}\mathbf{A}^d\times\mathbf{n}\big|_\Gamma = 0, \quad \partial_{\mathbf{n}}\Phi^d\big|_\Gamma = 0, \\ \psi^d(0) = \psi_0, \quad \mathbf{A}^d(0) = \mathbf{A}_0, \quad \mathbf{A}^d_t(0) = \mathbf{A}_1, \quad \Phi^d(0) = \Phi_0, \end{cases} \tag{3.4}$$

and $z_c(t) = (\psi^c(t), \mathbf{A}^c(t), \mathbf{A}^c_t(t), \Phi^c(t))$ is the solution to

$$\begin{cases} \psi^c_t - \Delta\psi^c + \psi^c = F\left(\psi, \tilde{\mathbf{A}}, \Phi\right), \\ \varepsilon\left(\mathbf{A}^c_t - \nabla\Phi^c\right)_t + \sigma\left(\mathbf{A}^c_t - \nabla\Phi^c\right) + \operatorname{curl}^2\mathbf{A}^c = \mathbf{G}\left(\psi, \tilde{\mathbf{A}}\right), \\ -\varepsilon\Delta\Phi^c_t - \sigma\Delta\Phi^c = H(\psi, \Phi), \\ \operatorname{div}\mathbf{A}^c = 0, \quad \int_\Omega \Phi^c = 0, \\ \partial_{\mathbf{n}}\psi^c|_\Gamma = 0, \quad \mathbf{A}^c\cdot\mathbf{n}|_\Gamma = 0, \ \operatorname{curl}\mathbf{A}^c\times\mathbf{n}|_\Gamma = 0, \quad \partial_{\mathbf{n}}\Phi^c|_\Gamma = 0, \\ \psi^c(0) = 0, \quad \mathbf{A}^c(0) = \mathbf{0}, \quad \mathbf{A}^c_t(0) = \mathbf{0}, \quad \Phi^c(0) = 0, \end{cases} \tag{3.5}$$

where, in (3.5), we have set

$$F\left(\psi,\tilde{\mathbf{A}},\Phi\right)=i\Phi\psi-\left|\tilde{\mathbf{A}}\right|^2\psi-2i\tilde{\mathbf{A}}\cdot\nabla\psi+\lambda^2\left(1-|\psi|^2\right)\psi+\psi, \tag{3.6}$$

$$\mathbf{G}\left(\psi,\tilde{\mathbf{A}}\right)=\frac{i}{2}\left(\psi\nabla\overline{\psi}-\overline{\psi}\nabla\psi\right)-\tilde{\mathbf{A}}|\psi|^2, \tag{3.7}$$

$$H(\psi,\Phi)=-\frac{i}{2}\left(\overline{\psi}\psi_t-\psi\overline{\psi}_t\right)-|\psi|^2\Phi. \tag{3.8}$$

Then the following lemmas show that the $\mathbb{X}^0$-norm of $z_d(t)$ exponentially decays to zero as time goes to infinity, while $z_c(t)$ remains in a bounded set of $\mathbb{X}^1$ which is compact in $\mathbb{X}^0$.

LEMMA 3.1 (cf. [3, Lemma 4.2]). *There exist constants $\gamma>0$ and $\nu>0$ independent of $\|z_0\|_0$ such that*

$$\|z_d(t)\|_0^2\le\gamma^2e^{-2\nu t}\|z_0\|_0^2\qquad\forall\ t\ge0. \tag{3.9}$$

LEMMA 3.2 (cf. [3, Lemma 4.3]). *There exists a constant $C>0$ depending on $\|z_0\|_0$ such that*

$$\|z_c(t)\|_1\le C\qquad\forall\ t\ge0. \tag{3.10}$$

Let us define

$$\begin{aligned}\mathcal{E}\left(\psi,\tilde{\mathbf{A}},\tilde{\mathbf{A}}_t,\Phi\right)&=\frac{1}{2}\left\|D_{\tilde{\mathbf{A}}}\psi\right\|^2+\frac{\lambda^2}{4}\left\|1-|\psi|^2\right\|^2+\frac{1}{2}\left\|\mathrm{curl}\tilde{\mathbf{A}}+\mathbf{H}_{ext}\right\|^2\\&\quad+\frac{\varepsilon}{2}\|\tilde{\mathbf{A}}_t\|^2+\frac{\varepsilon}{2}\|\nabla\Phi\|^2.\end{aligned} \tag{3.11}$$

It is easy to see that $\mathcal{E}$ is well defined and continuous on $\mathbb{X}^{-1}$ (cf. (2.5)). In addition, we have what follows.

LEMMA 3.3. *$\mathcal{E}$ is a global Lyapunov functional for the dynamical system $(\mathbb{X}^0_\infty,S(t))$.*

Proof. We only need to prove that $\mathcal{E}$ decreases along any given trajectory $z(t)=S(t)z_0$. Here and in what follows, we will sometimes use formal arguments which can be rigorously justified, e.g., by using a Galerkin approximation scheme (see [32]).

Multiplying (1.1) by $\overline{\psi}_t$ and its conjugate by ψ_t, respectively, integrating on Ω and adding the results together, we get

$$\begin{aligned}&\frac{d}{dt}\left(\frac{1}{2}\|D_{\tilde{\mathbf{A}}}\psi\|^2+\frac{\lambda^2}{4}\left\|1-|\psi|^2\right\|^2\right)+\|\psi_t\|^2+\frac{i}{2}\left\langle\Phi,\overline{\psi}\psi_t-\psi\overline{\psi}_t\right\rangle\\&\quad-\left\langle\tilde{\mathbf{A}}\cdot\tilde{\mathbf{A}}_t,|\psi|^2\right\rangle-\frac{i}{2}\left\langle\tilde{\mathbf{A}}_t,\overline{\psi}\nabla\psi-\psi\nabla\overline{\psi}\right\rangle=0.\end{aligned} \tag{3.12}$$

On the other hand, multiplying (1.2) by $\tilde{\mathbf{A}}_t$ in $\mathbf{L}^2(\Omega)$, we have

$$\begin{aligned}&\frac{d}{dt}\left(\frac{1}{2}\left\|\mathrm{curl}\tilde{\mathbf{A}}+\mathbf{H}_{ext}\right\|^2+\frac{\varepsilon}{2}\left\|\tilde{\mathbf{A}}_t\right\|^2\right)+\sigma\left\|\tilde{\mathbf{A}}_t\right\|^2+\left\langle\tilde{\mathbf{A}}\cdot\tilde{\mathbf{A}}_t,|\psi|^2\right\rangle\\&\quad+\frac{i}{2}\left\langle\overline{\psi}\nabla\psi-\psi\nabla\overline{\psi},\tilde{\mathbf{A}}_t\right\rangle=0.\end{aligned} \tag{3.13}$$

Moreover, taking the inner product of (1.16) with Φ in $L^2(\Omega)$ yields

$$\frac{\varepsilon}{2}\frac{d}{dt}\|\nabla\Phi\|^2+\sigma\|\nabla\Phi\|^2+\left\langle\Phi^2,|\psi|^2\right\rangle+\frac{i}{2}\left\langle\overline{\psi}\psi_t-\psi\overline{\psi}_t,\Phi\right\rangle=0. \tag{3.14}$$

Hence, we infer from (3.12)–(3.14) that, for all $t \geq 0$,

$$\begin{aligned}
&\frac{d}{dt}\mathcal{E}\left(\psi, \tilde{\mathbf{A}}, \tilde{\mathbf{A}}_t, \Phi\right) \\
&= -\|\psi_t\|^2 - i\left\langle \Phi, \overline{\psi}\psi_t - \psi\overline{\psi}_t\right\rangle - \sigma\left\|\tilde{\mathbf{A}}_t\right\|^2 - \sigma\left\|\nabla\Phi\right\|^2 - \left\langle \Phi^2, |\psi|^2\right\rangle \\
&= -\|\psi_t - i\Phi\psi\|^2_{L^2_{\mathbb{C}}(\Omega)} - \sigma\left\|\tilde{\mathbf{A}}_t\right\|^2 - \sigma\|\nabla\Phi\|^2 \\
&\leq 0, \qquad (3.15)
\end{aligned}$$

and the assertion follows. □

Due to the above lemmas, we can see that $(S(t), \mathbb{X}^0_\infty)$ is a gradient system. Furthermore, on account of Lemmas 3.1, 3.2, 3.3, and some well-known results in dynamical systems (see, for instance, [4, Chapter 9]), we deduce the following.

LEMMA 3.4. *For any $z_0 \in \mathbb{X}^0_\infty$, the ω-limit set of z_0 is a nonempty compact connected subset of $\mathbb{X}^0$. Furthermore, we have*

(i) $dist_{\mathbb{X}^0}(S(t)z_0, \omega(z_0)) \to 0$ *as* $t \to \infty$;
(ii) $\omega(z_0)$ *consists of equilibria of the form* $z_\infty = (\psi_\infty, \mathbf{A}_\infty, \mathbf{0}, 0)$, *where* $(\psi_\infty, \mathbf{A}_\infty)$ *satisfies* (2.13)–(2.16);
(iii) $S(t)\omega(z_0) = \omega(z_0)\ \forall\ t \geq 0$;
(iv) $\mathcal{E}$ *is constant on* $\omega(z_0)$.

Remark 3.1. If the set of equilibria were discrete, then we could conclude immediately that each solution converges to a single equilibrium (see (i) of Lemma 3.4). However, as is well known, in more than one spatial dimension, the set of stationary solutions can be a continuum for physically reasonable nonlinearity. The reader is referred, for instance, to [16, Remark 2.3.13]), where the following two-dimensional equation $-\Delta u + u^3 - \lambda u = 0$, $\lambda > 0$, endowed with a standard Dirichlet homogeneous boundary condition, is considered.

Remark 3.2. Lemmas 3.1 and 3.2 play a basic role in [3]. In particular, they enable the authors to prove that $(S(t), \mathbb{X}^0_\infty)$ has the global attractor $\mathcal{A}_\varepsilon$ bounded in $\mathbb{X}^1$. In addition, due to Lemma 3.3, we infer that $\mathcal{A}_\varepsilon$ coincides with the unstable manifold of the set of equilibria (see, e.g., [31, Chapter 7, section 4]).

4. Extended Łojasiewicz–Simon inequality. Let us set $\psi = \psi_1 + i\psi_2$, where $\psi_1, \psi_2 : \Omega \times \mathbb{R}^+ \to \mathbb{R}$, and define $\overrightarrow{\psi} = (\psi_1, \psi_2)$. Recalling (2.24), we rewrite the functional $\mathcal{E}_0$ in the real form E_0:

$$\begin{aligned}
&E_0\left[(\psi_1, \psi_2), \tilde{\mathbf{A}}\right] = \mathcal{E}_0\left(\psi, \tilde{\mathbf{A}}\right) \\
&= \int_\Omega \left[\frac{\lambda^2}{4}\left(1 - |\psi|^2\right)^2 + \frac{1}{2}\left|\left(\nabla - i\tilde{\mathbf{A}}\right)\psi\right|^2 + \frac{1}{2}\left|\mathrm{curl}\tilde{\mathbf{A}} + \mathbf{H}_{ext}\right|^2\right] dx \\
&= \int_\Omega \left[\frac{\lambda^2}{4}\left(1 - \psi_1^2 - \psi_2^2\right)^2 + \frac{1}{2}\left|\nabla\psi_1 + \tilde{\mathbf{A}}\psi_2\right|^2 + \frac{1}{2}\left|\nabla\psi_2 - \tilde{\mathbf{A}}\psi_1\right|^2\right. \\
&\qquad \left. + \frac{1}{2}\left|\mathrm{curl}\tilde{\mathbf{A}} + \mathbf{H}_{ext}\right|^2\right] dx. \qquad (4.1)
\end{aligned}$$

In analogy to [9, Lemma 5.1], we can see that the functional E_0 is analytic on

$$\mathcal{X} = \left[H^1(\Omega)\right]^2 \times \mathbf{X}_0 \qquad (4.2)$$

in the sense of Deimling (see [6, Definition 15.1, pp. 150]). Then a direct calculation shows what follows.

LEMMA 4.1. *The Fréchet derivative DE_0 of E_0 is an analytic mapping from $\mathcal{X}$ to its dual $\mathcal{X}'$, and it is given by*

$$\left\langle DE_0[(\psi_1,\psi_2),\tilde{\mathbf{A}}],[(v_1,v_2),\mathbf{V}]\right\rangle = \langle \partial_{\psi_1}E_0, v_1\rangle + \langle \partial_{\psi_2}E_0, v_2\rangle + \langle \partial_{\tilde{\mathbf{A}}}E_0, \mathbf{V}\rangle, \tag{4.3}$$

for any $[(\psi_1,\psi_2),\tilde{\mathbf{A}}],[(v_1,v_2),\mathbf{V}]\in\mathcal{X}$, where

$$\begin{aligned}&\left\langle \partial_{\psi_1}E_0\left[(\psi_1,\psi_2),\tilde{\mathbf{A}}\right], v_1\right\rangle \\ &= \int_\Omega \left[\left(\nabla\psi_1+\tilde{\mathbf{A}}\psi_2\right)\cdot\nabla v_1 - \left(\nabla\psi_2-\tilde{\mathbf{A}}\psi_1\right)\cdot\tilde{\mathbf{A}}v_1\right]dx - \lambda^2\int_\Omega\left(1-|\psi|^2\right)\psi_1 v_1 dx,\end{aligned} \tag{4.4}$$

$$\begin{aligned}&\left\langle \partial_{\psi_2}E_0\left[(\psi_1,\psi_2),\tilde{\mathbf{A}}\right], v_2\right\rangle \\ &= \int_\Omega \left[\left(\nabla\psi_2-\tilde{\mathbf{A}}\psi_1\right)\cdot\nabla v_2 + \left(\nabla\psi_1+\tilde{\mathbf{A}}\psi_2\right)\cdot\tilde{\mathbf{A}}v_2\right]dx - \lambda^2\int_\Omega\left(1-|\psi|^2\right)\psi_2 v_2 dx,\end{aligned} \tag{4.5}$$

$$\begin{aligned}&\left\langle \partial_{\tilde{\mathbf{A}}}E_0\left[(\psi_1,\psi_2),\tilde{\mathbf{A}}\right], \mathbf{V}\right\rangle \\ &= \int_\Omega \left(\operatorname{curl}\tilde{\mathbf{A}}+\mathbf{H}_{ext}\right)\cdot(\operatorname{curl}\mathbf{V})\,dx + \int_\Omega\left[-\left(\psi_1\nabla\psi_2-\psi_2\nabla\psi_1\right)+\tilde{\mathbf{A}}|\psi|^2\right]\cdot\mathbf{V}dx.\end{aligned} \tag{4.6}$$

Let us introduce now the quadratic form $\mathcal{J}:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ given by

$$\mathcal{J}\left([(v_1,v_2),\mathbf{V}],[(w_1,w_2),\mathbf{W}]\right) := \int_\Omega(\nabla\overrightarrow{v}\cdot\nabla\overrightarrow{w}+\overrightarrow{v}\cdot\overrightarrow{w})dx + \int_\Omega(\operatorname{curl}\mathbf{V})\cdot(\operatorname{curl}\mathbf{W})dx, \tag{4.7}$$

$\forall\ [(v_1,v_2),\mathbf{V}],[(w_1,w_2),\mathbf{W}]\in\mathcal{X}$.

It follows from (2.10) that $(\mathcal{J}([(v_1,v_2),\mathbf{V}],[(v_1,v_2),\mathbf{V}]))^{1/2}$ defines an equivalent norm of $[(v_1,v_2),\mathbf{V}]$ on $\mathcal{X}$. Based on this fact, we can retrace the steps of [9, section 6] to prove the following version of the Łojasiewicz–Simon inequality, which is a slight modification of [9, Proposition 6.1]. We thus omit the details. This lemma can be viewed as an extended version of Simon's result [28] for the scalar case under the use of L^2 norm.

LEMMA 4.2 (extended Łojasiewicz–Simon inequality). *Let $[(\psi_{1\infty},\psi_{2\infty}),\tilde{\mathbf{A}}_\infty]\in\mathcal{X}$ be a critical point of E_0. Then there exist constants $\beta>0$ and $\theta\in(0,\frac{1}{2})$ depending on $[(\psi_{1\infty},\psi_{2\infty}),\tilde{\mathbf{A}}_\infty]$ such that, for any $[(\psi_1,\psi_2),\tilde{\mathbf{A}}]\in\mathcal{X}$ satisfying*

$$\left\|\left[(\psi_1,\psi_2),\tilde{\mathbf{A}}\right]-\left[(\psi_{1\infty},\psi_{2\infty}),\tilde{\mathbf{A}}_\infty\right]\right\|_{\mathcal{X}}<\beta, \tag{4.8}$$

we have

$$\left|E_0\left[(\psi_1,\psi_2),\tilde{\mathbf{A}}\right]-E_0\left[(\psi_{1\infty},\psi_{2\infty}),\tilde{\mathbf{A}}_\infty\right]\right|^{1-\theta}\le\left\|DE_0\left[(\psi_1,\psi_2),\tilde{\mathbf{A}}\right]\right\|_{\mathcal{X}'}. \tag{4.9}$$

5. Convergence to equilibrium. In this section we prove the convergence to equilibrium for the global solutions to our Ginzburg–Landau–Maxwell equations. First, we show the convergence for $\tilde{\mathbf{A}}_t(t)$ and $\Phi(t)$ to 0, which is given by the following lemma.

LEMMA 5.1. *We have*

$$\lim_{t\to+\infty}\left\|\tilde{\mathbf{A}}_t(t)\right\|_{\mathbf{X}_0}=0, \tag{5.1}$$

$$\lim_{t\to+\infty}\|\Phi(t)\|_{H^2(\Omega)}=0. \tag{5.2}$$

Proof. Integrating (3.15) from 0 to t, we obtain

$$\mathcal{E}(t) + \int_0^t \|\psi_t(\tau) - i\Phi(\tau)\psi(\tau)\|^2 d\tau + \sigma \int_0^t \left\|\tilde{\mathbf{A}}_t(\tau)\right\|^2 d\tau + \sigma \int_0^t \|\nabla\Phi(\tau)\|^2 d\tau$$
$$\leq \mathcal{E}(0) < \infty. \tag{5.3}$$

Since $\mathcal{E}$ is nonnegative, we deduce

$$\int_0^\infty \left\|\tilde{\mathbf{A}}_t(\tau)\right\|^2 d\tau < \infty, \quad \int_0^\infty \|\nabla\Phi(\tau)\|^2 d\tau < \infty, \tag{5.4}$$

and

$$\int_0^\infty \|\psi_t(\tau) - i\Phi(\tau)\psi(\tau)\|^2 d\tau < \infty. \tag{5.5}$$

From (1.16), we infer that

$$\begin{aligned} \varepsilon\|\Delta\Phi_t\| &\leq \sigma\|\Delta\Phi\| + \left\|\frac{i}{2}\left(\overline{\psi}\psi_t - \psi\overline{\psi}_t\right)\right\| + \||\psi|^2\Phi\| \\ &\leq \sigma\|\Delta\Phi\| + C\left(\|\psi_t\| + \|\nabla\Phi\|\right) \end{aligned} \tag{5.6}$$

and, on account of (1.17), $\int_\Omega \Phi_t dx = 0$. Thus we have

$$\|\nabla\Phi_t\| \leq \|\Phi_t\|_{H^2(\Omega)} \leq C\|\Delta\Phi_t\| \leq C\left(\|\Delta\Phi\| + \|\psi_t\| + \|\nabla\Phi\|\right). \tag{5.7}$$

On the other hand, from (1.1), (1.8), and the Sobolev embedding theorem, we have

$$\begin{aligned} \|\psi_t\| &\leq C\left(\|\Phi\psi\| + \|\Delta\psi\| + \left\|\tilde{\mathbf{A}}\cdot\nabla\psi\right\| + \left\|\left|\tilde{\mathbf{A}}\right|^2\psi\right\| + \left\|\left(1 - |\psi|^2\right)\psi\right\|\right) \\ &\leq C\left(\|\Phi\|\|\psi\|_{L^\infty_{\mathbb{C}}(\Omega)} + \|\psi\|_{H^2_{\mathbb{C}}(\Omega)} + \left\|\tilde{\mathbf{A}}\right\|_{\mathbf{H}^2(\Omega)}\|\nabla\psi\| + \left\|\tilde{\mathbf{A}}\right\|^2_{\mathbf{H}^2(\Omega)}\|\psi\|\right. \\ &\qquad \left. + \|\psi\|^3_{H^1_{\mathbb{C}}(\Omega)}\right). \end{aligned} \tag{5.8}$$

Then, it follows from (5.7), (5.8), and the uniform estimate (2.12) that

$$\|\nabla\Phi_t(t)\| \leq C \qquad \forall\ t \geq 0. \tag{5.9}$$

Similarly, it follows from (1.15) and estimates (2.12) and (5.9) that

$$\begin{aligned} \left\|\tilde{\mathbf{A}}_{tt}\right\| &\leq C\left(\|\nabla\Phi_t\| + \left\|\tilde{\mathbf{A}}_t\right\| + \|\nabla\Phi\| + \left\|\mathrm{curl}^2\tilde{\mathbf{A}}\right\| + \left\|\left(\overline{\psi}D_{\tilde{\mathbf{A}}}\psi - \psi\overline{D_{\tilde{\mathbf{A}}}\psi}\right)\right\|\right) \\ &\leq C\left(\|\nabla\Phi_t\| + \left\|\tilde{\mathbf{A}}_t\right\| + \|\nabla\Phi\| + \left\|\mathrm{curl}^2\tilde{\mathbf{A}}\right\| + \|\psi\|_{L^\infty_{\mathbb{C}}(\Omega)}\|\nabla\psi\|\right. \\ &\qquad \left. + \|\psi\|^2_{L^\infty_{\mathbb{C}}(\Omega)}\left\|\tilde{\mathbf{A}}\right\|\right) \leq C. \end{aligned} \tag{5.10}$$

Let $h(t) = \|\tilde{\mathbf{A}}_t(t)\|^2$. Then we have

$$\left|\frac{dh}{dt}\right| = 2\left|\left\langle\tilde{\mathbf{A}}_t, \tilde{\mathbf{A}}_{tt}\right\rangle\right| \leq 2\left\|\tilde{\mathbf{A}}_t\right\|\left\|\tilde{\mathbf{A}}_{tt}\right\| \leq C. \tag{5.11}$$

From (5.4) and (5.11), we infer that $h \in L^1(0,\infty)$, and it is globally Lipschitz on $(0,\infty)$. Hence we have

$$\lim_{t\to+\infty}\left\|\tilde{\mathbf{A}}_t(t)\right\| = 0. \tag{5.12}$$

Thus, on account of Lemmas 3.1 and 3.2, we deduce (5.1).

We recall that $\|\psi_0\|_{L^\infty_{\mathbb{C}}(\Omega)} \le 1$. Therefore, Proposition 2.1 implies

$$\|\psi(t)\|_{L^\infty_{\mathbb{C}}(\Omega)} \le 1 \qquad \forall\, t \ge 0. \tag{5.13}$$

Thus, recalling (1.6), the Poincaré inequality for Φ combined with (5.4) and (5.13) implies that

$$\int_0^\infty \| - i\Phi(\tau)\psi(\tau)\|^2 d\tau \le C \int_0^\infty \|\nabla \Phi(\tau)\|^2 d\tau < \infty. \tag{5.14}$$

This, together with (5.5), yields

$$\int_0^\infty \|\psi_t(\tau)\|^2 d\tau < \infty. \tag{5.15}$$

Observe now that

$$|i\langle \overline{\psi}\psi_t - \psi\overline{\psi}_t, \Phi\rangle| \le 2\,\|\psi\Phi\|\|\psi_t\| \le \left\langle \Phi^2, |\psi|^2\right\rangle + \|\psi_t\|^2. \tag{5.16}$$

Hence, from (3.14), the Hölder inequality, and the Cauchy–Schwarz inequality, we get

$$\begin{aligned} &\frac{\varepsilon}{2}\frac{d}{dt}\|\nabla\Phi\|^2 + \sigma\|\nabla\Phi\|^2 + \left\langle \Phi^2, |\psi|^2\right\rangle \\ &\le \left| \frac{i}{2}\left\langle \overline{\psi}\psi_t - \psi\overline{\psi}_t, \Phi\right\rangle\right| \le \|\psi\Phi\|\|\psi_t\| \le \frac{1}{2}\left\langle \Phi^2, |\psi|^2\right\rangle + \frac{1}{2}\|\psi_t\|^2. \end{aligned} \tag{5.17}$$

Setting $h_1(t) = \|\nabla\Phi(t)\|^2$, it follows from (5.4), (5.15), (5.17), and [36, Lemma 6.2.1] that

$$\lim_{t\to+\infty} h_1(t) = 0, \tag{5.18}$$

which entails

$$\lim_{t\to+\infty} \|\Phi(t)\|_{H^1(\Omega)} = 0. \tag{5.19}$$

Hence, (5.2) is a consequence of Lemma 3.1, Lemma 3.2, and (5.19). ∎

To establish the convergence for ψ and $\tilde{\mathbf{A}}$ (or $\mathbf{A}$), it is convenient to rewrite $\mathcal{E}$ in the real form. As before, we denote $\psi = \psi_1 + i\psi_2$, with ψ_1, ψ_2 being real functions. Then we have (cf. (3.11) and (4.1))

$$\begin{aligned} E\left[(\psi_1,\psi_2), \tilde{\mathbf{A}}, \Phi\right] &:= \mathcal{E}\left(\psi, \tilde{\mathbf{A}}, \tilde{\mathbf{A}}_t, \Phi\right) \\ &= E_0\left[(\psi_1,\psi_2), \tilde{\mathbf{A}}\right] + \frac{\varepsilon}{2}\left\|\tilde{\mathbf{A}}_t\right\|^2 + \frac{\varepsilon}{2}\|\nabla\Phi\|^2 \end{aligned} \tag{5.20}$$

and accordingly, (3.15) becomes

$$\frac{d}{dt}E\left[(\psi_1,\psi_2), \tilde{\mathbf{A}}, \Phi\right] = -\left\|\psi_{1t} + \Phi\psi_2\right\|^2 - \left\|\psi_{2t} - \Phi\psi_1\right\|^2 - \sigma\left\|\tilde{\mathbf{A}}_t\right\|^2 - \sigma\|\nabla\Phi\|^2 \le 0 \tag{5.21}$$

$\forall\, t \ge 0$.

It follows from (2.12) and the Poincaré inequality that

$$\|\psi_{1t}\|^2 \le 2\left(\|\psi_{1t} + \Phi\psi_2\|^2 + \|\Phi\psi_2\|^2\right) \le C_1\left(\|\psi_{1t} + \Phi\psi_2\|^2 + \|\nabla\Phi\|^2\right), \tag{5.22}$$

$$\|\psi_{2t}\|^2 \le 2\left(\|\psi_{2t}-\Phi\psi_1\|^2+\|\Phi\psi_1\|^2\right) \le C_1\left(\|\psi_{2t}-\Phi\psi_1\|^2+\|\nabla\Phi\|^2\right), \tag{5.23}$$

where $C_1>0$ is a constant. By choosing

$$C_2 \in \left(0, \min\left\{\frac{\sigma}{4C_1}, \frac{1}{2C_1}\right\}\right), \tag{5.24}$$

we infer from (5.21) that

$$\frac{d}{dt}E\left[(\psi_1,\psi_2),\tilde{\mathbf{A}},\Phi\right] \le -C_2\|\psi_{1t}\|^2 - C_2\|\psi_{2t}\|^2 - \frac{\sigma}{2}\|\tilde{\mathbf{A}}_t\|^2 - \frac{\sigma}{2}\|\nabla\Phi\|^2. \tag{5.25}$$

Next, we introduce the following auxiliary functional:

$$\mathcal{G}\left[(\psi_1,\psi_2),\tilde{\mathbf{A}},\Phi\right] = \left\langle \mathrm{curl}^2\tilde{\mathbf{A}} + \frac{i}{2}\left(\overline{\psi}D_{\tilde{\mathbf{A}}}\psi - \psi\overline{D_{\tilde{\mathbf{A}}}\psi}\right) + \mathrm{curl}\mathbf{H}_{ext},\ \tilde{\mathbf{A}}_t - \nabla\Phi\right\rangle_{\mathbf{X}_0'}. \tag{5.26}$$

Then we have what follows.

LEMMA 5.2. *The following inequality holds for all $t\ge 0$:*

$$\begin{aligned}
&\frac{d}{dt}\mathcal{G}\left[(\psi_1,\psi_2),\tilde{\mathbf{A}},\Phi\right]\\
&\le -\frac{1}{2\varepsilon}\left\|\mathrm{curl}^2\tilde{\mathbf{A}} + \frac{i}{2}\left(\overline{\psi}D_{\tilde{\mathbf{A}}}\psi - \psi\overline{D_{\tilde{\mathbf{A}}}\psi}\right) + \mathrm{curl}\mathbf{H}_{ext}\right\|^2_{\mathbf{X}_0'}\\
&\quad + C_3\left(\|\psi_{1t}\|^2+\|\psi_{2t}\|^2+\|\nabla\psi_{1t}\|^2+\|\nabla\psi_{2t}\|^2+\|\tilde{\mathbf{A}}_t\|^2+\|\nabla\Phi\|^2\right).
\end{aligned} \tag{5.27}$$

Proof. By direct calculation and (1.2), we get

$$\begin{aligned}
&\frac{d}{dt}\mathcal{G}\left[(\psi_1,\psi_2),\tilde{\mathbf{A}},\Phi\right]\\
&= \left\langle\left(\mathrm{curl}^2\tilde{\mathbf{A}} + \frac{i}{2}\left(\overline{\psi}D_{\tilde{\mathbf{A}}}\psi - \psi\overline{D_{\tilde{\mathbf{A}}}\psi}\right) + \mathrm{curl}\mathbf{H}_{ext}\right)_t,\ \tilde{\mathbf{A}}_t - \nabla\Phi\right\rangle_{\mathbf{X}_0'}\\
&\quad -\frac{1}{\varepsilon}\left\|\mathrm{curl}^2\tilde{\mathbf{A}} + \frac{i}{2}\left(\overline{\psi}D_{\tilde{\mathbf{A}}}\psi - \psi\overline{D_{\tilde{\mathbf{A}}}\psi}\right) + \mathrm{curl}\mathbf{H}_{ext}\right\|^2_{\mathbf{X}_0'}\\
&\quad -\frac{\sigma}{\varepsilon}\left\langle \mathrm{curl}^2\tilde{\mathbf{A}} + \frac{i}{2}\left(\overline{\psi}D_{\tilde{\mathbf{A}}}\psi - \psi\overline{D_{\tilde{\mathbf{A}}}\psi}\right) + \mathrm{curl}\mathbf{H}_{ext},\ \tilde{\mathbf{A}}_t - \nabla\Phi\right\rangle_{\mathbf{X}_0'}\\
&:= I_1 + I_2 + I_3.
\end{aligned} \tag{5.28}$$

Let us estimate the right-hand side term by term. Observe first that

$$\left|\left\langle \mathrm{curl}^2\tilde{\mathbf{A}}_t, \tilde{\mathbf{A}}_t - \nabla\Phi\right\rangle_{\mathbf{X}_0'}\right| = \left|\left\langle\tilde{\mathbf{A}}_t,\tilde{\mathbf{A}}_t\right\rangle - \left\langle\tilde{\mathbf{A}}_t,\nabla\Phi\right\rangle\right| \le \frac{3}{2}\|\tilde{\mathbf{A}}_t\|^2 + \frac{1}{2}\|\nabla\Phi\|^2. \tag{5.29}$$

Recalling that we are in dimension two, it follows from a well-known Sobolev embedding theorem that

$$\mathbf{X}_0 \subset \mathbf{H}^1(\Omega) \hookrightarrow \mathbf{L}^p(\Omega) \quad \forall\ p>1, \tag{5.30}$$

$$\mathbf{L}^q(\Omega) \hookrightarrow (\mathbf{H}^1(\Omega))' \subset \mathbf{X}_0' \quad \forall\ q>1. \tag{5.31}$$

Therefore, using (2.12) and taking $q = \frac{4}{3}$ in (5.31), we have

$$\begin{aligned}
&\left| \left\langle \frac{i}{2} \left(\overline{\psi} D_{\tilde{\mathbf{A}}} \psi - \psi \overline{D_{\tilde{\mathbf{A}}} \psi} \right)_t, \tilde{\mathbf{A}}_t - \nabla \Phi \right\rangle_{\mathbf{X}_0'} \right| \\
&\leq \left\| \left[-(\psi_1 \nabla \psi_2 - \psi_2 \nabla \psi_1) + \tilde{\mathbf{A}} |\psi|^2 \right]_t \right\|_{\mathbf{X}_0'} \left\| \tilde{\mathbf{A}}_t - \nabla \Phi \right\|_{\mathbf{X}_0'} \\
&\leq \Big(\|\psi_{1t} \nabla \psi_2\|_{\mathbf{L}^{\frac{4}{3}}(\Omega)} + \|\psi_1 \nabla \psi_{2t}\| + \|\psi_{2t} \nabla \psi_1\|_{\mathbf{L}^{\frac{4}{3}}(\Omega)} + \|\psi_2 \nabla \psi_{1t}\| \\
&\quad + \left\| \tilde{\mathbf{A}}_t |\psi|^2 \right\| + 2 \left\| \tilde{\mathbf{A}} (\psi_1 \psi_{1t} + \psi_2 \psi_{2t}) \right\| \Big) \left(\left\| \tilde{\mathbf{A}}_t \right\| + \|\nabla \Phi\| \right) \\
&\leq \Big(\|\nabla \psi_2\|_{\mathbf{L}^4(\Omega)} \|\psi_{1t}\| + \|\nabla \psi_1\|_{\mathbf{L}^4(\Omega)} \|\psi_{2t}\| + C \|\nabla \psi_{1t}\| + C \|\nabla \psi_{2t}\| \\
&\quad + C \left\| \tilde{\mathbf{A}}_t \right\| + C \|\psi_{1t}\| + C \|\psi_{2t}\| \Big) \left(\left\| \tilde{\mathbf{A}}_t \right\| + \|\nabla \Phi\| \right) \\
&\leq C \left(\|\psi_{1t}\|^2 + \|\psi_{2t}\|^2 + \|\nabla \psi_{1t}\|^2 + \|\nabla \psi_{2t}\|^2 + \left\| \tilde{\mathbf{A}}_t \right\|^2 + \|\nabla \Phi\|^2 \right). \qquad (5.32)
\end{aligned}$$

As a result, we get

$$I_1 \leq C \left(\|\psi_{1t}\|^2 + \|\psi_{2t}\|^2 + \|\nabla \psi_{1t}\|^2 + \|\nabla \psi_{2t}\|^2 + \left\| \tilde{\mathbf{A}}_t \right\|^2 + \|\nabla \Phi\|^2 \right). \qquad (5.33)$$

On the other hand, thanks to the Cauchy–Schwarz inequality, we obtain

(5.34)

$$I_3 \leq \frac{1}{2\varepsilon} \left\| \mathrm{curl}^2 \tilde{\mathbf{A}} + \frac{i}{2} \left(\overline{\psi} D_{\tilde{\mathbf{A}}} \psi - \psi \overline{D_{\tilde{\mathbf{A}}} \psi} \right) + \mathrm{curl} \mathbf{H}_{ext} \right\|_{\mathbf{X}_0'}^2 + C \left(\left\| \tilde{\mathbf{A}}_t \right\|^2 + \|\nabla \Phi\|^2 \right).$$

Thus, (5.27) follows from (5.28), (5.33), and (5.34). ∎

Let

$$\mathcal{G}_1[(\psi_1, \psi_2), \Phi] = \frac{1}{2} \|\psi_{1t}\|^2 + \frac{1}{2} \|\psi_{2t}\|^2 + \frac{\varepsilon}{2} \|\Delta \Phi\|^2 + \frac{\sigma}{2} \|\nabla \Phi\|^2. \qquad (5.35)$$

Then we have the following.

LEMMA 5.3. *The following inequality holds for all $t \geq 0$:*

$$\begin{aligned}
\frac{d}{dt} \mathcal{G}_1[(\psi_1, \psi_2), \Phi] &\leq -\frac{1}{2} \|\nabla \psi_{1t}\|^2 - \frac{1}{2} \|\nabla \psi_{2t}\|^2 - \frac{\sigma}{2} \|\Delta \Phi\|^2 - \frac{\varepsilon}{2} \|\nabla \Phi_t\|^2 \\
&\quad + C_4 \left(\|\psi_{1t}\|^2 + \|\psi_{2t}\|^2 + \left\| \tilde{\mathbf{A}}_t \right\|^2 + \|\nabla \Phi\|^2 \right). \qquad (5.36)
\end{aligned}$$

Proof. Differentiating (1.14) with respect to time, adding the resulting equation multiplied by $\overline{\psi}_t$ to its conjugate multiplied by ψ_t, then integrating on Ω yield

$$\begin{aligned}
\frac{1}{2} \frac{d}{dt} \|\psi_t\|^2 &= -\|\nabla \psi_t\|^2 + \frac{i}{2} \left\langle \Phi_t, \psi \overline{\psi}_t - \overline{\psi} \psi_t \right\rangle + \lambda^2 \left\langle \left(1 - 2|\psi|^2 \right), |\psi_t|^2 \right\rangle \\
&\quad - \frac{\lambda^2}{2} \left(\left\langle \psi^2, \overline{\psi}_t^2 \right\rangle + \left\langle \overline{\psi}^2, \psi_t^2 \right\rangle \right) - \left\langle \left| \tilde{\mathbf{A}} \right|^2, |\psi_t|^2 \right\rangle + i \left\langle \mathbf{A}_t, \psi \nabla \overline{\psi}_t - \overline{\psi} \nabla \psi_t \right\rangle \\
&\quad - \left\langle \tilde{\mathbf{A}} \cdot \mathbf{A}_t, \psi \overline{\psi}_t + \overline{\psi} \psi_t \right\rangle - i \left\langle \tilde{\mathbf{A}}, \overline{\psi}_t \nabla \psi_t - \psi_t \nabla \overline{\psi}_t \right\rangle. \qquad (5.37)
\end{aligned}$$

It follows from (2.12), standard Sobolev embeddings, and the Poincaré inequality that the terms on the right-hand side of (5.37) can be estimated as follows:

$$\begin{aligned}
&\frac{i}{2} \left\langle \Phi_t, \psi \overline{\psi}_t - \overline{\psi} \psi_t \right\rangle + \lambda^2 \left\langle \left(1 - 2|\psi|^2 \right), |\psi_t|^2 \right\rangle - \frac{\lambda^2}{2} \left(\left\langle \psi^2, \overline{\psi}_t^2 \right\rangle + \left\langle \overline{\psi}^2, \psi_t^2 \right\rangle \right) \qquad (5.38) \\
&\leq \frac{\varepsilon}{4} \|\nabla \Phi_t\|^2 + C \|\psi_t\|^2
\end{aligned}$$

$$
\begin{aligned}
&-\left\langle \left|\tilde{\mathbf{A}}\right|^2, |\psi_t|^2\right\rangle + i\left\langle \mathbf{A}_t, \psi\nabla\overline{\psi}_t - \overline{\psi}\nabla\psi_t\right\rangle - \left\langle \tilde{\mathbf{A}}\cdot\mathbf{A}_t, \psi\overline{\psi}_t + \overline{\psi}\psi_t\right\rangle \\
&-i\left\langle \tilde{\mathbf{A}}, \overline{\psi}_t\nabla\psi_t - \psi_t\nabla\overline{\psi}_t\right\rangle \\
&\le \frac{1}{2}\|\nabla\psi_t\|^2 + C\|\psi_t\|^2 + C\left\|\tilde{\mathbf{A}}_t\right\|^2.
\end{aligned} \tag{5.39}
$$

Multiplying now (1.16) by $-\Delta\Phi + \Phi_t$ and integrating on Ω yield

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\left(\varepsilon\|\Delta\Phi\|^2 + \sigma\|\nabla\Phi\|^2\right) \\
&= -\sigma\|\Delta\Phi\|^2 - \varepsilon\|\nabla\Phi_t\|^2 - \frac{i}{2}\left\langle \overline{\psi}\psi_t - \psi\overline{\psi}_t, -\Delta\Phi\right\rangle - \left\langle |\psi|^2\Phi, -\Delta\Phi\right\rangle \\
&-\frac{i}{2}\left\langle \overline{\psi}\psi_t - \psi\overline{\psi}_t, \Phi_t\right\rangle - \left\langle |\psi|^2\Phi, \Phi_t\right\rangle \\
&\le -\sigma\|\Delta\Phi\|^2 - \varepsilon\|\nabla\Phi_t\|^2 + \frac{\sigma}{2}\|\Delta\Phi\|^2 + \frac{\varepsilon}{4}\|\nabla\Phi_t\|^2 + C\|\psi_t\|^2 + C\|\nabla\Phi\|^2.
\end{aligned} \tag{5.40}
$$

Then from (5.37)–(5.40), it follows that

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\left(\|\psi_t\|^2 + \varepsilon\|\Delta\Phi\|^2 + \sigma\|\nabla\Phi\|^2\right) \\
&\le -\frac{1}{2}\|\nabla\psi_t\|^2 - \frac{\sigma}{2}\|\Delta\Phi\|^2 - \frac{\varepsilon}{2}\|\nabla\Phi_t\|^2 + C\left(\|\psi_t\|^2 + \left\|\tilde{\mathbf{A}}_t\right\|^2 + \|\nabla\Phi\|^2\right),
\end{aligned} \tag{5.41}
$$

which gives (5.36). ∎

We are now able to prove that ψ_t vanishes at infinity.

LEMMA 5.4. *There holds*

$$
\lim_{t\to+\infty}\|\psi_t(t)\| = 0. \tag{5.42}
$$

Proof. It follows from (5.41) that

$$
\frac{1}{2}\frac{d}{dt}\left(\|\psi_t\|^2 + \varepsilon\|\Delta\Phi\|^2 + \sigma\|\nabla\Phi\|^2\right) + \frac{\sigma}{2}\|\Delta\Phi\|^2 \le C\left(\|\psi_t\|^2 + \left\|\tilde{\mathbf{A}}_t\right\|^2 + \|\nabla\Phi\|^2\right). \tag{5.43}
$$

Since $z_0 = (\psi_0, \mathbf{A}_0, \mathbf{A}_1, \Phi_0) \in \mathbb{X}^0_\infty$, from (1.1) and the Sobolev embedding theorem we infer that $\|\psi_t|_{t=0}\| \le C$, where C is a constant depending on $\|z_0\|_0$. Therefore, (5.4), (5.15), and (5.43) yield

$$
\|\psi_t(t)\| \le C \quad \forall\ t \ge 0, \tag{5.44}
$$

$$
\int_0^\infty \|\Delta\Phi(\tau)\|^2 d\tau < \infty. \tag{5.45}
$$

Denote $h_2(t) = \|\psi_t\|^2 + \varepsilon\|\Delta\Phi\|^2 + \sigma\|\nabla\Phi\|^2$. Then (5.43), Lemma 5.1, and (5.44) imply

$$
\frac{d}{dt}h_2 \le C \quad \forall\ t \ge 0. \tag{5.46}
$$

On the other hand, we deduce $h_2(t) \in L^1(0,\infty)$ from (5.4), (5.15), and (5.45). As a consequence of [36, Lemma 6.2.1],

$$
\lim_{t\to+\infty} h_2(t) = 0. \tag{5.47}
$$

The proof is complete. ∎

As we mentioned in the Introduction, due to the hyperbolic nature of the equation for $\tilde{\mathbf{A}}$, a key point in the proof of convergence result is to construct an appropriate auxiliary functional in order to implement the Łojasiewicz–Simon approach. For this purpose, we introduce the following functional $\mathfrak{F}:\mathbb{X}^0\to\mathbb{R}$:

$$\mathfrak{F}\left[(\psi_1,\psi_2),\tilde{\mathbf{A}},\Phi\right]:=E\left[(\psi_1,\psi_2),\tilde{\mathbf{A}},\Phi\right]+\eta\mathcal{G}\left[(\psi_1,\psi_2),\tilde{\mathbf{A}},\Phi\right]+\eta_1\mathcal{G}_1[(\psi_1,\psi_2),\Phi], \tag{5.48}$$

where η and η_1 are positive constants to be specified later.

Furthermore, it follows from (5.25), Lemma 5.2, and Lemma 5.3 that

$$\begin{aligned}
&\frac{d}{dt}\mathfrak{F}\left[(\psi_1,\psi_2),\tilde{\mathbf{A}},\Phi\right]\\
&\leq -(C_2-C_3\eta-C_4\eta_1)\left(\|\psi_{1t}\|^2+\|\psi_{2t}\|^2\right)\\
&\quad-\left(\frac{\sigma}{2}-C_3\eta-C_4\eta_1\right)\left(\left\|\tilde{\mathbf{A}}_t\right\|^2+\|\nabla\Phi\|^2\right)\\
&\quad-\left(\frac{\eta_1}{2}-C_3\eta\right)\left(\|\nabla\psi_{1t}\|^2+\|\nabla\psi_{2t}\|^2\right)-\frac{\eta_1\sigma}{2}\|\Delta\Phi\|^2-\frac{\eta_1\varepsilon}{2}\|\nabla\Phi_t\|^2\\
&\quad-\frac{\eta}{2\varepsilon}\left\|\mathrm{curl}^2\tilde{\mathbf{A}}+\frac{i}{2}\left(\overline{\psi}D_{\tilde{\mathbf{A}}}\psi-\psi\overline{D_{\tilde{\mathbf{A}}}\psi}\right)+\mathrm{curl}\mathbf{H}_{ext}\right\|^2_{\mathbf{X}_0'}.
\end{aligned} \tag{5.49}$$

Taking

$$\eta=\frac{\min\{C_2,\sigma\}}{4C_3(1+4C_4)},\quad \eta_1=4C_3\eta=\frac{\min\{C_2,\sigma\}}{(1+4C_4)}, \tag{5.50}$$

we can deduce that

$$\begin{aligned}
&\frac{d}{dt}\mathfrak{F}\left[(\psi_1,\psi_2),\tilde{\mathbf{A}},\Phi\right]\\
&\leq -C\Bigg(\|\psi_{1t}\|^2_{H^1}+\|\psi_{2t}\|^2_{H^1}+\left\|\tilde{\mathbf{A}}_t\right\|^2+\|\nabla\Phi\|^2+\|\Delta\Phi\|^2+\|\nabla\Phi_t\|^2\\
&\qquad+\left\|\mathrm{curl}^2\tilde{\mathbf{A}}+\frac{i}{2}\left(\overline{\psi}D_{\tilde{\mathbf{A}}}\psi-\psi\overline{D_{\tilde{\mathbf{A}}}\psi}\right)+\mathrm{curl}\mathbf{H}_{ext}\right\|^2_{\mathbf{X}_0'}\Bigg)\\
&\leq 0
\end{aligned} \tag{5.51}$$

for some $C>0$.

Thanks to Lemma 3.1, Lemma 3.2, and Lemma 3.4, we can find an increasing unbounded sequence $\{t_n\}_{n=1}^{\infty}$ and a pair $(\psi_\infty,\tilde{\mathbf{A}}_\infty)$ satisfying (2.19)–(2.22) such that

$$\lim_{t_n\to+\infty}\left\|\tilde{\mathbf{A}}(t_n)-\tilde{\mathbf{A}}_\infty\right\|_{\mathbf{H}^2(\Omega)}=0, \tag{5.52}$$

$$\lim_{t_n\to+\infty}\|\psi(t_n)-\psi_\infty\|_{H^2_{\mathbb{C}}(\Omega)}=0. \tag{5.53}$$

Thus, it follows from (5.52), (5.53), Lemma 5.1, and Lemma 5.4 that

$$\lim_{t_n\to+\infty}\mathfrak{F}(t_n)=E_0\left[(\psi_{1\infty},\psi_{2\infty}),\tilde{\mathbf{A}}_\infty\right]. \tag{5.54}$$

Since $\mathfrak{F}$ is decreasing in time (see (5.51)), we have

$$\mathfrak{F}(t)\geq E_0\left[(\psi_{1\infty},\psi_{2\infty}),\tilde{\mathbf{A}}_\infty\right]\quad\forall\, t\geq 0. \tag{5.55}$$

After these preparations, we proceed to prove the convergence result following a simple argument introduced in [20]. The key observation is that, after a certain time t_0, the solution $[(\psi_1(t),\psi_2(t)),\tilde{\mathbf{A}}(t)]$ always satisfies the condition of Lemma 4.2, that is, it falls in a neighborhood of $[(\psi_{1\infty},\psi_{2\infty}),\tilde{\mathbf{A}}_\infty]$, where (4.9) holds.

We now consider all the possible cases.

(1) If there is a $t_0>0$ such that at this time $\mathfrak{F}[(\psi_1,\psi_2),\tilde{\mathbf{A}},\Phi]=E_0[(\psi_{1\infty},\psi_{2\infty}),\tilde{\mathbf{A}}_\infty]$, then $\forall\ t>t_0$, we deduce from (5.51) that $[(\psi_1(t),\psi_2(t)),\tilde{\mathbf{A}}(t)]$ is independent of t. We obtain the convergence for ψ and $\tilde{\mathbf{A}}$ from (5.52) and (5.53).

(2) If $\mathfrak{F}[(\psi_1,\psi_2),\tilde{\mathbf{A}},\Phi]>E_0[(\psi_{1\infty},\psi_{2\infty}),\tilde{\mathbf{A}}_\infty]\ \forall\ t\geq 0$ and there is $t_0>0$ such that $\forall\ t\geq t_0$, $[(\psi_1,\psi_2),\tilde{\mathbf{A}}]$ satisfies the condition of Lemma 4.2, i.e.,

$$\left\|\left[(\psi_1,\psi_2),\tilde{\mathbf{A}}\right]-\left[(\psi_{1\infty},\psi_{2\infty}),\tilde{\mathbf{A}}_\infty\right]\right\|_{\mathcal{X}}<\beta, \tag{5.56}$$

then, for the Łojasiewicz exponent $\theta\in(0,\frac{1}{2})$ (see Lemma 4.2), we calculate

$$\begin{aligned}&\frac{d}{dt}\left(\mathfrak{F}\left[(\psi_1,\psi_2),\tilde{\mathbf{A}},\Phi\right]-E_0\left[(\psi_{1\infty},\psi_{2\infty}),\tilde{\mathbf{A}}_\infty\right]\right)^{\theta}\\ &=\theta\left(\mathfrak{F}\left[(\psi_1,\psi_2),\tilde{\mathbf{A}},\Phi\right]-E_0\left[(\psi_{1\infty},\psi_{2\infty}),\tilde{\mathbf{A}}_\infty\right]\right)^{\theta-1}\frac{d}{dt}\mathfrak{F}\left[(\psi_1,\psi_2),\tilde{\mathbf{A}},\Phi\right].\end{aligned} \tag{5.57}$$

Recalling (5.20) and the definitions of $\mathcal{G}[(\psi_1,\psi_2),\tilde{\mathbf{A}},\Phi]$ and $\mathcal{G}_1[(\psi_1,\psi_2),\Phi]$, we have

$$\begin{aligned}&\left(\mathfrak{F}\left[(\psi_1,\psi_2),\tilde{\mathbf{A}},\Phi\right]-E_0\left[(\psi_{1\infty},\psi_{2\infty}),\tilde{\mathbf{A}}_\infty\right]\right)^{1-\theta}\\ &\leq C\left(\left|E_0\left[(\psi_1,\psi_2),\tilde{\mathbf{A}}\right]-E_0\left[(\psi_{1\infty},\psi_{2\infty}),\tilde{\mathbf{A}}_\infty\right]\right|^{1-\theta}+\left\|\tilde{\mathbf{A}}_t\right\|^{2(1-\theta)}\right.\\ &\quad+\|\nabla\Phi\|^{2(1-\theta)}+\|\Delta\Phi\|^{2(1-\theta)}+\|\psi_{1t}\|^{2(1-\theta)}+\|\psi_{2t}\|^{2(1-\theta)}\\ &\quad\left.+\left\|\mathrm{curl}^2\tilde{\mathbf{A}}+\frac{i}{2}\left(\overline{\psi}D_{\tilde{\mathbf{A}}}\psi-\psi\overline{D_{\tilde{\mathbf{A}}}\psi}\right)+\mathrm{curl}\mathbf{H}_{ext}\right\|_{\mathbf{X}_0'}^{2(1-\theta)}\right).\end{aligned} \tag{5.58}$$

On the other hand, on account of (4.9), Lemma 4.1 (cf. the explicit form of DE_0), Proposition 2.1, and the Poincaré inequality, we deduce that

$$\begin{aligned}&\left|E_0\left[(\psi_1,\psi_2),\tilde{\mathbf{A}}\right]-E_0\left[(\psi_{1\infty},\psi_{2\infty}),\tilde{\mathbf{A}}_\infty\right]\right|^{1-\theta}\leq\left\|DE_0\left[(\psi_1,\psi_2),\tilde{\mathbf{A}}\right]\right\|_{\mathcal{X}'}\\ &\leq C\left(\|\psi_{1t}+\Phi\psi_2\|+\|\psi_{2t}-\Phi\psi_1\|+\left\|\mathrm{curl}^2\tilde{\mathbf{A}}+\frac{i}{2}\left(\overline{\psi}D_{\tilde{\mathbf{A}}}\psi-\psi\overline{D_{\tilde{\mathbf{A}}}\psi}\right)+\mathrm{curl}\mathbf{H}_{ext}\right\|_{\mathbf{X}_0'}\right)\\ &\leq C\left(\|\psi_{1t}\|+\|\psi_{2t}\|+\|\nabla\Phi\|+\left\|\mathrm{curl}^2\tilde{\mathbf{A}}+\frac{i}{2}\left(\overline{\psi}D_{\tilde{\mathbf{A}}}\psi-\psi\overline{D_{\tilde{\mathbf{A}}}\psi}\right)+\mathrm{curl}\mathbf{H}_{ext}\right\|_{\mathbf{X}_0'}\right).\end{aligned} \tag{5.59}$$

Since $\theta\in(0,\frac{1}{2})$, thanks to (5.1) and (5.2), it follows that

$$\begin{aligned}&\left(\mathfrak{F}\left[(\psi_1,\psi_2),\tilde{\mathbf{A}},\Phi\right]-E_0\left[(\psi_{1\infty},\psi_{2\infty}),\tilde{\mathbf{A}}_\infty\right]\right)^{1-\theta}\\ &\leq C\left(\|\psi_{1t}\|+\|\psi_{2t}\|+\left\|\tilde{\mathbf{A}}_t\right\|+\|\nabla\Phi\|+\|\Delta\Phi\|\right.\\ &\quad\left.+\left\|\mathrm{curl}^2\tilde{\mathbf{A}}+\frac{i}{2}\left(\overline{\psi}D_{\tilde{\mathbf{A}}}\psi-\psi\overline{D_{\tilde{\mathbf{A}}}\psi}\right)+\mathrm{curl}\mathbf{H}_{ext}\right\|_{\mathbf{X}_0'}\right).\end{aligned} \tag{5.60}$$

Hence, combining (5.51), (5.57), and (5.60), we infer that, for all $t \geq t_0$,

$$\frac{d}{dt}\left(\mathfrak{F}\left[(\psi_1,\psi_2),\tilde{\mathbf{A}},\Phi\right]-E_0\left[(\psi_{1\infty},\psi_{2\infty}),\tilde{\mathbf{A}}_\infty\right]\right)^\theta + C_5\Bigg(\|\psi_{1t}\|+\|\psi_{2t}\|+\left\|\tilde{\mathbf{A}}_t\right\| + \|\nabla\Phi\|+\|\Delta\Phi\|+\|\nabla\Phi_t\|+\left\|\mathrm{curl}^2\tilde{\mathbf{A}}+\frac{i}{2}\left(\overline{\psi}D_{\tilde{\mathbf{A}}}\psi-\psi\overline{D_{\tilde{\mathbf{A}}}\psi}\right)+\mathrm{curl}\mathbf{H}_{ext}\right\|_{\mathbf{X}_0'}\Bigg) \leq 0. \tag{5.61}$$

As a result, we can deduce that

$$\int_{t_0}^{t}\left\|\tilde{\mathbf{A}}_t(\tau)\right\|d\tau<\infty, \tag{5.62}$$

$$\int_{t_0}^{t}\left(\|\psi_{1t}(\tau)\|+\|\psi_{2t}(\tau)\|\right)\,d\tau<\infty \tag{5.63}$$

$\forall\ t\geq t_0$. This yields

$$\lim_{t\to+\infty}\left\|\tilde{\mathbf{A}}(t)-\tilde{\mathbf{A}}_\infty\right\|=0,\qquad \lim_{t\to+\infty}\|\psi(t)-\psi_\infty\|=0.$$

From Lemmas 3.1 and 3.2, owing to the uniqueness of the limit, we infer that

$$\lim_{t\to+\infty}\left\|\tilde{\mathbf{A}}(t)-\tilde{\mathbf{A}}_\infty\right\|_{\mathbf{H}^2(\Omega)}=0,\qquad \lim_{t\to+\infty}\|\psi(t)-\psi_\infty\|_{H^2_{\tilde{\mathbb{C}}}(\Omega)}=0.$$

Recalling Lemma 5.1, we conclude that $(\psi(t),\tilde{\mathbf{A}}(t),\tilde{\mathbf{A}}_t(t),\Phi(t))$ converges to $(\psi_\infty,\tilde{\mathbf{A}}_\infty,\mathbf{0},0)$ in $\mathbb{X}^0$ as t goes to infinity.

(3) Combining (5.52), (5.53), and Lemma 5.1, we obtain that, for any $\rho\in(0,\beta)$, there exists an integer $N=N(\rho)$ such that, when $n\geq N$,

$$\|\psi_1(t_n)-\psi_{1\infty}\|_{H^2(\Omega)}<\frac{\rho}{4},\quad \|\psi_2(t_n)-\psi_{2\infty}\|_{H^2(\Omega)}<\frac{\rho}{4}, \tag{5.64}$$

$$\left\|\tilde{\mathbf{A}}(t_n)-\tilde{\mathbf{A}}_\infty\right\|_{\mathbf{H}^2(\Omega)}<\frac{\rho}{4}, \tag{5.65}$$

$$\frac{1}{C_5}\left(\mathfrak{F}\left[(\psi_1(t_n),\psi_2(t_n)),\tilde{\mathbf{A}}(t_n),\Phi(t_n)\right]-E_0\left[(\psi_{1\infty},\psi_{2\infty}),\tilde{\mathbf{A}}_\infty\right]\right)^\theta<\frac{\rho}{4}. \tag{5.66}$$

We now define

$$\bar{t}_n=\sup_{t>t_n}\Big\{\ |\ \|\psi_1(s)-\psi_{1\infty}\|_{H^2(\Omega)}+\|\psi_2(s)-\psi_{2\infty}\|_{H^2(\Omega)} +\left\|\tilde{\mathbf{A}}(s)-\tilde{\mathbf{A}}_\infty\right\|_{\mathbf{H}^2(\Omega)}<\beta,\quad \forall s\in[t_n,t]\Big\}, \tag{5.67}$$

and we note that it follows from (5.64), (5.65), and the continuity of the orbit in $\mathbb{X}^0$ that $\bar{t}_n>t_n\forall\ n\geq N$. Then there are two possibilities, namely,

(i) There exists $n_0\geq N$ such that $\bar{t}_{n_0}=+\infty$, then arguing as in (1) and (2), the convergence result can be obtained.

(ii) Otherwise, for all $n\geq N$, we have $t_n<\bar{t}_n<+\infty$, and, for all $t\in[t_n,\bar{t}_n]$, we have

$$\mathfrak{F}\left[(\psi_1(t),\psi_2(t)),\tilde{\mathbf{A}}(t),\Phi(t)\right]>E_0\left[(\psi_{1\infty},\psi_{2\infty}),\tilde{\mathbf{A}}_\infty\right].$$

Hence, we deduce from (5.61) that

$$\begin{aligned}
&\int_{t_n}^{\bar{t}_n} \left(\|\psi_{1t}(\tau)\| + \|\psi_{2t}(\tau)\| + \|\tilde{\mathbf{A}}_t(\tau)\| \right) d\tau \\
&\leq \frac{1}{C_5} \left(\mathfrak{F}\left[(\psi_1(t_n), \psi_2(t_n)), \tilde{\mathbf{A}}(t_n), \Phi(t_n) \right] - E_0\left[(\psi_{1\infty}, \psi_{2\infty}), \tilde{\mathbf{A}}_\infty \right] \right)^\theta \\
&< \frac{\rho}{4}. \qquad (5.68)
\end{aligned}$$

Noticing that $\rho > 0$ is arbitrary in the above argument, we are able to conclude that for any $\rho > 0$, there exist $N = N(\rho) \in \mathbb{N}$ such that, for all $n \geq N$, there holds

$$\begin{aligned}
&\|\psi_1(\bar{t}_n) - \psi_{1\infty}\| + \|\psi_2(\bar{t}_n) - \psi_{2\infty}\| + \|\tilde{\mathbf{A}}(\bar{t}_n) - \tilde{\mathbf{A}}_\infty\| \\
&\leq \|\psi_1(t_n) - \psi_{1\infty}\| + \|\psi_2(t_n) - \psi_{2\infty}\| + \|\tilde{\mathbf{A}}(t_n) - \tilde{\mathbf{A}}_\infty\| \\
&\quad + \int_{t_n}^{\bar{t}_n} \left(\|\psi_{1t}(\tau)\| + \|\psi_{2t}(\tau)\| + \|\tilde{\mathbf{A}}_t(\tau)\| \right) d\tau \\
&< \rho. \qquad (5.69)
\end{aligned}$$

This implies

$$\lim_{n\to+\infty} \left(\|\psi_1(\bar{t}_n) - \psi_{1\infty}\| + \|\psi_2(\bar{t}_n) - \psi_{2\infty}\| + \|\tilde{\mathbf{A}}(\bar{t}_n) - \tilde{\mathbf{A}}_\infty\| \right) = 0. \qquad (5.70)$$

Therefore, due to Lemmas 3.1 and 3.2, there exists a subsequence of $\{\bar{t}_n\}$, denoted by $\{\bar{t}_{n_k}\}$ such that

$$\lim_{n_k\to+\infty} \left(\|\psi_1(\bar{t}_{n_k}) - \psi_{1\infty}\|_{H^2(\Omega)} + \|\psi_2(\bar{t}_{n_k}) - \psi_{2\infty}\|_{H^2(\Omega)} + \|\tilde{\mathbf{A}}(\bar{t}_{n_k}) - \tilde{\mathbf{A}}_\infty\|_{\mathbf{H}^2(\Omega)} \right) = 0, \qquad (5.71)$$

which contradicts the definition of $\bar{t}_n$.

Summing up, we conclude that

$$\lim_{t\to+\infty} \|z(t) - z_\infty\|_0 = 0. \qquad (5.72)$$

6. Convergence rate. In this section we demonstrate estimate (2.25) on convergence rate. Let us begin by noting that a combination of (5.60) and (5.61) implies, for all $t \geq t_0$, that

$$\begin{aligned}
&\frac{d}{dt} \left(\mathfrak{F}\left[(\psi_1, \psi_2), \tilde{\mathbf{A}}, \Phi \right] - E_0\left[(\psi_{1\infty}, \psi_{2\infty}), \tilde{\mathbf{A}}_\infty \right] \right) \\
&+ C \left(\mathfrak{F}\left[(\psi_1, \psi_2), \tilde{\mathbf{A}}, \Phi \right] - E_0\left[(\psi_{1\infty}, \psi_{2\infty}), \tilde{\mathbf{A}}_\infty \right] \right)^{2(1-\theta)} \leq 0. \qquad (6.1)
\end{aligned}$$

Keeping in mind that $\theta \in (0, \frac{1}{2})$, we deduce

$$\mathfrak{F}\left[(\psi_1(t), \psi_2(t)), \tilde{\mathbf{A}}(t), \Phi(t) \right] - E_0\left[(\psi_{1\infty}, \psi_{2\infty}), \tilde{\mathbf{A}}_\infty \right] \leq C(1+t)^{-1/(1-2\theta)}, \qquad (6.2)$$

for all $t \geq 0$ and for a suitable $C > 0$, because for $t \in [0, t_0]$, the term on the left-hand side is bounded by a constant depending only on initial data (cf. (2.12)). Hence, it

follows from (5.61) that

$$\begin{aligned}&\int_t^\infty \left(\|\psi_{1t}\| + \|\psi_{2t}\| + \|\nabla\Phi\| + \|\nabla\Phi_t\| + \left\|\tilde{\mathbf{A}}_t\right\|\right) d\tau\\ &+ \int_t^\infty \left(\left\|\mathrm{curl}^2\tilde{\mathbf{A}} + \frac{i}{2}\left(\overline{\psi}D_{\tilde{\mathbf{A}}}\psi - \psi\overline{D_{\tilde{\mathbf{A}}}\psi}\right) + \mathrm{curl}\mathbf{H}_{ext}\right\|_{\mathbf{X}_0'}\right) d\tau\\ &\leq C(1+t)^{-\theta/(1-2\theta)}, \qquad \forall\ t \geq 0,\end{aligned} \tag{6.3}$$

and, as a consequence (cf. (1.2)),

$$\int_t^\infty \left\|\tilde{\mathbf{A}}_{tt}(\tau)\right\|_{\mathbf{X}_0'} d\tau \leq C(1+t)^{-\theta/(1-2\theta)} \qquad \forall\ t \geq 0. \tag{6.4}$$

Therefore, from (6.3) and (6.4), we deduce the following preliminary convergence rate estimate:

$$\begin{aligned}&\|\psi(t) - \psi_\infty\| + \left\|\tilde{\mathbf{A}}(t) - \tilde{\mathbf{A}}_\infty\right\| + \left\|\tilde{\mathbf{A}}_t(t)\right\|_{\mathbf{X}_0'} + \|\nabla\Phi\|\\ &\leq \int_t^\infty \left(\|\psi_t(\tau)\| + \left\|\tilde{\mathbf{A}}_t(\tau)\right\| + \left\|\tilde{\mathbf{A}}_{tt}(\tau)\right\|_{\mathbf{X}_0'} + \|\nabla\Phi_t(\tau)\|\right) d\tau\\ &\leq C(1+t)^{-\theta/(1-2\theta)} \qquad \forall\ t \geq 0.\end{aligned} \tag{6.5}$$

In order to get higher order estimates, we first subtract the stationary equations from the time-dependent equations. This gives

$$\begin{aligned}&\psi_t - i\Phi\psi - \Delta(\psi - \psi_\infty) + 2i\left(\tilde{\mathbf{A}}\cdot\nabla\psi - \tilde{\mathbf{A}}_\infty\cdot\nabla\psi_\infty\right) + \tilde{\mathbf{A}}^2\psi - \tilde{\mathbf{A}}_\infty^2\psi_\infty\\ &-\lambda^2\left(1 - |\psi|^2\right)\psi + \lambda^2\left(1 - |\psi_\infty|^2\right)\psi_\infty = 0,\end{aligned} \tag{6.6}$$

$$\begin{aligned}&\varepsilon\left(\tilde{\mathbf{A}}_t - \nabla\Phi\right)_t + \sigma\left(\tilde{\mathbf{A}}_t - \nabla\Phi\right) + \mathrm{curl}^2\tilde{\mathbf{A}} - \mathrm{curl}^2\tilde{\mathbf{A}}_\infty + |\psi|^2\tilde{\mathbf{A}} - |\psi_\infty|^2\tilde{\mathbf{A}}_\infty\\ &\qquad + \frac{i}{2}\left(\overline{\psi}\nabla\psi - \psi\nabla\overline{\psi}\right) - \frac{i}{2}\left(\overline{\psi}_\infty\nabla\psi_\infty - \psi_\infty\nabla\overline{\psi}_\infty\right) = 0,\end{aligned} \tag{6.7}$$

$$-\varepsilon\Delta\Phi_t - \sigma\Delta\Phi + \frac{i}{2}\left(\overline{\psi}\psi_t - \psi\overline{\psi}_t\right) + |\psi|^2\Phi = 0, \tag{6.8}$$

in $\Omega\times(0,\infty)$, with boundary conditions

$$\partial_{\mathbf{n}}(\psi - \psi_\infty) = 0, \quad \left(\tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty\right)\cdot\mathbf{n} = 0, \quad \left(\mathrm{curl}\tilde{\mathbf{A}} - \mathrm{curl}\tilde{\mathbf{A}}_\infty\right)\times\mathbf{n} = 0, \ \ \partial_{\mathbf{n}}\Phi = 0, \tag{6.9}$$

on $\Gamma\times(0,\infty)$.

Multiplying (6.6) by $\overline{\psi}_t$ and its conjugate by ψ_t, respectively, then integrating on Ω and adding the results together, we have

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\|\nabla\psi - \nabla\psi_\infty\|^2 + \|\psi_t\|^2 + \frac{i}{2}\left\langle\Phi, \overline{\psi}\psi_t - \psi\overline{\psi}_t\right\rangle\\ &+ i\left\langle\tilde{\mathbf{A}}\cdot\nabla\psi - \tilde{\mathbf{A}}_\infty\cdot\nabla\psi_\infty,\ \overline{\psi}_t\right\rangle - i\left\langle\tilde{\mathbf{A}}\cdot\nabla\overline{\psi} - \tilde{\mathbf{A}}_\infty\cdot\nabla\overline{\psi}_\infty,\ \psi_t\right\rangle\\ &+ \frac{1}{2}\left\langle\tilde{\mathbf{A}}^2\psi - \tilde{\mathbf{A}}_\infty^2\psi_\infty,\ \overline{\psi}_t\right\rangle + \frac{1}{2}\left\langle\tilde{\mathbf{A}}^2\overline{\psi} - \tilde{\mathbf{A}}_\infty^2\overline{\psi}_\infty,\ \psi_t\right\rangle\\ &- \frac{\lambda^2}{2}\int_\Omega\left[\left(1 - |\psi|^2\right)\psi - \left(1 - |\psi_\infty|^2\right)\psi_\infty\right]\overline{\psi}_t dx\\ &- \frac{\lambda^2}{2}\int_\Omega\left[\left(1 - |\psi|^2\right)\overline{\psi} - \left(1 - |\psi_\infty|^2\right)\overline{\psi}_\infty\right]\psi_t dx = 0.\end{aligned} \tag{6.10}$$

It follows from inequalities (2.7), (2.8), and bound (2.12) that

$$
\begin{aligned}
I_4 &:= \left| i \left\langle \tilde{\mathbf{A}} \cdot \nabla \psi - \tilde{\mathbf{A}}_\infty \cdot \nabla \psi_\infty,\ \overline{\psi}_t \right\rangle \right| \\
&\leq \left| \left\langle \tilde{\mathbf{A}} \cdot \nabla \psi - \tilde{\mathbf{A}}_\infty \cdot \nabla \psi,\ \overline{\psi}_t \right\rangle \right| + \left| \left\langle \tilde{\mathbf{A}}_\infty \cdot \nabla \psi - \tilde{\mathbf{A}}_\infty \cdot \nabla \psi_\infty,\ \overline{\psi}_t \right\rangle \right| \\
&\leq \|\nabla \psi\|_{\mathbf{L}^4} \left\| \tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty \right\|_{\mathbf{L}^4} \|\psi_t\| + \left\| \tilde{\mathbf{A}}_\infty \right\|_{\mathbf{L}^\infty} \|\nabla \psi - \nabla \psi_\infty\| \|\psi_t\| \\
&\leq C \left\| \tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty \right\|^{\frac{1}{2}} \left\| \mathrm{curl} \tilde{\mathbf{A}} - \mathrm{curl} \tilde{\mathbf{A}}_\infty \right\|^{\frac{1}{2}} \|\psi_t\| + C \|\nabla \psi - \nabla \psi_\infty\| \|\psi_t\| \\
&\leq \frac{1}{16} \|\psi_t\|^2 + \frac{\alpha}{8} \left\| \mathrm{curl} \tilde{\mathbf{A}} - \mathrm{curl} \tilde{\mathbf{A}}_\infty \right\|^2 + C \|\tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty\|^2 \\
&\quad + C_6 \|\nabla \psi - \nabla \psi_\infty\|^2. \qquad (6.11)
\end{aligned}
$$

Similarly, we deduce

$$
\begin{aligned}
I_5 &:= \left| -i \left\langle \tilde{\mathbf{A}} \cdot \nabla \overline{\psi} - \tilde{\mathbf{A}}_\infty \cdot \nabla \overline{\psi}_\infty,\ \psi_t \right\rangle \right| \\
&\leq \frac{1}{16} \|\psi_t\|^2 + \frac{\alpha}{8} \left\| \mathrm{curl} \tilde{\mathbf{A}} - \mathrm{curl} \tilde{\mathbf{A}}_\infty \right\|^2 + C \left\| \tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty \right\|^2 \\
&\quad + C_6 \|\nabla \psi - \nabla \psi_\infty\|^2. \qquad (6.12)
\end{aligned}
$$

Next, we have

$$
\begin{aligned}
I_6 &:= \frac{1}{2} \left| \left\langle \tilde{\mathbf{A}}^2 \psi - \tilde{\mathbf{A}}_\infty^2 \psi_\infty,\ \overline{\psi}_t \right\rangle \right| \\
&\leq \frac{1}{2} \left| \left\langle \tilde{\mathbf{A}}^2 \psi - \tilde{\mathbf{A}}^2 \psi_\infty,\ \overline{\psi}_t \right\rangle \right| + \frac{1}{2} \left| \left\langle \tilde{\mathbf{A}}^2 \psi_\infty - \tilde{\mathbf{A}}_\infty^2 \psi_\infty,\ \overline{\psi}_t \right\rangle \right| \\
&\leq \frac{1}{16} \|\psi_t\|^2 + C \left(\left\| \tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty \right\|^2 + \|\psi - \psi_\infty\|^2 \right). \qquad (6.13)
\end{aligned}
$$

In the same manner, we get

$$
I_7 := \frac{1}{2} \left| \left\langle \tilde{\mathbf{A}}^2 \overline{\psi} - \tilde{\mathbf{A}}_\infty^2 \overline{\psi}_\infty,\ \psi_t \right\rangle \right| \leq \frac{1}{16} \|\psi_t\|^2 + C \left(\left\| \tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty \right\|^2 + \|\psi - \psi_\infty\|^2 \right). \qquad (6.14)
$$

Moreover, we have

$$
\begin{aligned}
I_8 &:= \frac{\lambda^2}{2} \left| \int_\Omega \left[\left(1 - |\psi|^2\right) \psi - \left(1 - |\psi_\infty|^2\right) \psi_\infty \right] \overline{\psi}_t dx \right| \\
&\leq \frac{\lambda^2}{2} \left| \int_\Omega |\psi|^2 (\psi - \psi_\infty) \overline{\psi}_t dx \right| + \frac{\lambda^2}{2} \left| \int_\Omega \psi_\infty \left(|\psi|^2 - |\psi_\infty|^2 \right) \overline{\psi}_t dx \right| \\
&\quad + \frac{\lambda^2}{2} \left| \int_\Omega (\psi - \psi_\infty) \overline{\psi}_t dx \right| \\
&\leq \frac{1}{32} \left\| \overline{\psi}_t \right\|^2 + C \|\psi - \psi_\infty\|^2 + C \left\| \psi_1^2 + \psi_2^2 - \psi_{1\infty}^2 - \psi_{2\infty}^2 \right\| \left\| \overline{\psi}_t \right\| \\
&\leq \frac{1}{32} \left\| \overline{\psi}_t \right\|^2 + C \|\psi - \psi_\infty\|^2 + C \left\| \overline{\psi}_t \right\| (\|\psi_1 - \psi_{1\infty}\| + \|\psi_2 - \psi_{2\infty}\|) \\
&\leq \frac{1}{32} \left\| \overline{\psi}_t \right\|^2 + C \|\psi - \psi_\infty\|^2 + C \left\| \overline{\psi}_t \right\| \|\psi - \psi_\infty\| \\
&\leq \frac{1}{16} \|\psi_t\|^2 + C \|\psi - \psi_\infty\|^2, \qquad (6.15)
\end{aligned}
$$

(6.16)
$$I_9 := \frac{\lambda^2}{2}\left|\int_\Omega \left[\left(1-|\psi|^2\right)\overline{\psi}-\left(1-|\psi_\infty|^2\right)\overline{\psi}_\infty\right]\psi_t dx\right| \le \frac{1}{16}\|\psi_t\|^2 + C\|\psi-\psi_\infty\|^2.$$

Multiplying (6.6) by $\overline{\psi}-\overline{\psi}_\infty$ and its conjugate by $\psi-\psi_\infty$, respectively, then integrating on Ω and adding the results together, we obtain

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\psi-\psi_\infty\|^2 + \|\nabla\psi-\nabla\psi_\infty\|^2 + \frac{i}{2}\left\langle \Phi,\ \psi\overline{\psi}_\infty - \overline{\psi}\psi_\infty\right\rangle \\
&\quad + i\left\langle \tilde{\mathbf{A}}\cdot\nabla\psi - \tilde{\mathbf{A}}_\infty\cdot\nabla\psi_\infty,\ \overline{\psi}-\overline{\psi}_\infty\right\rangle - i\left\langle \tilde{\mathbf{A}}\cdot\nabla\overline{\psi} - \tilde{\mathbf{A}}_\infty\cdot\nabla\overline{\psi}_\infty,\ \psi-\psi_\infty\right\rangle \\
&\quad + \frac{1}{2}\left\langle \tilde{\mathbf{A}}^2\psi - \tilde{\mathbf{A}}_\infty^2\psi_\infty,\ \overline{\psi}-\overline{\psi}_\infty\right\rangle + \frac{1}{2}\left\langle \tilde{\mathbf{A}}^2\overline{\psi} - \tilde{\mathbf{A}}_\infty^2\overline{\psi}_\infty,\ \psi-\psi_\infty\right\rangle \\
&\quad + \frac{1}{2}\left\langle -\lambda^2\left(1-|\psi|^2\right)\psi + \lambda^2\left(1-|\psi_\infty|^2\right)\psi_\infty,\ \overline{\psi}-\overline{\psi}_\infty\right\rangle \\
&\quad + \frac{1}{2}\left\langle -\lambda^2\left(1-|\psi|^2\right)\overline{\psi} + \lambda^2\left(1-|\psi_\infty|^2\right)\overline{\psi}_\infty,\ \psi-\psi_\infty\right\rangle = 0. \qquad (6.17)
\end{aligned}$$

Then, arguing as before, we infer the following estimates:

$$\begin{aligned}
I_{10} &:= \frac{1}{2}\left|i\left\langle \Phi,\ \psi\overline{\psi}_\infty - \overline{\psi}\psi_\infty\right\rangle\right| \\
&\le \frac{1}{2}\left|\left\langle \Phi,\ \psi\overline{\psi}_\infty - \psi_\infty\overline{\psi}_\infty\right\rangle\right| + \frac{1}{2}\left|\left\langle \Phi,\ \psi_\infty\overline{\psi}_\infty - \psi_\infty\overline{\psi}\right\rangle\right| \\
&\le \|\Phi\|_{L^4}\left\|\overline{\psi}_\infty\right\|_{L^4}\|\psi-\psi_\infty\| + \|\Phi\|_{L^4}\|\psi_\infty\|_{L^4}\left\|\overline{\psi}-\overline{\psi}_\infty\right\| \\
&\le \|\nabla\Phi\|^2 + C\|\psi-\psi_\infty\|^2, \qquad (6.18)
\end{aligned}$$

$$\begin{aligned}
I_{11} &:= \left|i\left\langle \tilde{\mathbf{A}}\cdot\nabla\psi - \tilde{\mathbf{A}}_\infty\cdot\nabla\psi_\infty,\ \overline{\psi}-\overline{\psi}_\infty\right\rangle\right| \\
&\le \left|\left\langle \tilde{\mathbf{A}}\cdot\nabla\psi - \tilde{\mathbf{A}}_\infty\cdot\nabla\psi,\ \overline{\psi}-\overline{\psi}_\infty\right\rangle\right| + \left|\left\langle \tilde{\mathbf{A}}_\infty\cdot\nabla\psi - \tilde{\mathbf{A}}_\infty\cdot\nabla\psi_\infty,\ \overline{\psi}-\overline{\psi}_\infty\right\rangle\right| \\
&\le \|\nabla\psi\|_{\mathbf{L}^4}\left\|\tilde{\mathbf{A}}-\tilde{\mathbf{A}}_\infty\right\|_{\mathbf{L}^4}\|\psi-\psi_\infty\| + \left\|\tilde{\mathbf{A}}_\infty\right\|_{\mathbf{L}^\infty}\|\nabla\psi-\nabla\psi_\infty\|\|\psi-\psi_\infty\| \\
&\le \frac{\epsilon\alpha}{16}\left\|\mathrm{curl}\tilde{\mathbf{A}}-\mathrm{curl}\tilde{\mathbf{A}}_\infty\right\|^2 + \frac{1}{16}\|\nabla\psi-\nabla\psi_\infty\|^2 + C\|\psi-\psi_\infty\|^2, \qquad (6.19)
\end{aligned}$$

$$\begin{aligned}
I_{12} &:= \left|i\left\langle \tilde{\mathbf{A}}\cdot\nabla\overline{\psi} - \tilde{\mathbf{A}}_\infty\cdot\nabla\overline{\psi}_\infty,\ \psi-\psi_\infty\right\rangle\right| \\
&\le \frac{\epsilon\alpha}{16}\left\|\mathrm{curl}\tilde{\mathbf{A}}-\mathrm{curl}\tilde{\mathbf{A}}_\infty\right\|^2 + \frac{1}{16}\|\nabla\psi-\nabla\psi_\infty\|^2 + C\|\psi-\psi_\infty\|^2, \qquad (6.20)
\end{aligned}$$

$$\begin{aligned}
I_{13} &:= \left|\frac{1}{2}\left\langle \tilde{\mathbf{A}}^2\psi - \tilde{\mathbf{A}}_\infty^2\psi_\infty,\ \overline{\psi}-\overline{\psi}_\infty\right\rangle\right| \\
&\le \left|\frac{1}{2}\left\langle \tilde{\mathbf{A}}^2\psi - \tilde{\mathbf{A}}_\infty^2\psi,\ \overline{\psi}-\overline{\psi}_\infty\right\rangle\right| + \left|\frac{1}{2}\left\langle \tilde{\mathbf{A}}_\infty^2\psi - \tilde{\mathbf{A}}_\infty^2\psi_\infty,\ \overline{\psi}-\overline{\psi}_\infty\right\rangle\right| \\
&\le C\|\psi\|_{L^\infty}\left\|\tilde{\mathbf{A}}+\tilde{\mathbf{A}}_\infty\right\|_{\mathbf{L}^\infty}\left\|\tilde{\mathbf{A}}-\tilde{\mathbf{A}}_\infty\right\|\|\psi-\psi_\infty\| + C\left\|\tilde{\mathbf{A}}_\infty^2\right\|_{\mathbf{L}^\infty}\|\psi-\psi_\infty\|^2 \\
&\le C\left(\left\|\tilde{\mathbf{A}}-\tilde{\mathbf{A}}_\infty\right\|^2 + \|\psi-\psi_\infty\|^2\right), \qquad (6.21)
\end{aligned}$$

$$I_{14} := \left|\frac{1}{2}\left\langle \tilde{\mathbf{A}}^2\overline{\psi} - \tilde{\mathbf{A}}_\infty^2\overline{\psi}_\infty,\ \psi-\psi_\infty\right\rangle\right| \le C\left(\left\|\tilde{\mathbf{A}}-\tilde{\mathbf{A}}_\infty\right\|^2 + \|\psi-\psi_\infty\|^2\right), \qquad (6.22)$$

$$I_{15} := \left|\frac{1}{2}\left\langle -\lambda^2\left(1-|\psi|^2\right)\psi + \lambda^2\left(1-|\psi_\infty|^2\right)\psi_\infty,\ \overline{\psi}-\overline{\psi}_\infty\right\rangle\right| \le C\|\psi-\psi_\infty\|^2, \qquad (6.23)$$

$$I_{16} := \left| \frac{1}{2} \left\langle -\lambda^2 \left(1 - |\psi|^2\right) \overline{\psi} + \lambda^2 \left(1 - |\psi_\infty|^2\right) \overline{\psi}_\infty,\ \psi - \psi_\infty \right\rangle \right| \le C \|\psi - \psi_\infty\|^2. \tag{6.24}$$

In the above estimates, $\alpha > 0$ and $\epsilon > 0$ are constants to be chosen later.

We now consider (6.7) and take the inner product with $\tilde{\mathbf{A}}_t + \alpha(\tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty)$ in $\mathbf{L}^2(\Omega)$. We thus have

$$\begin{aligned}
&\frac{d}{dt}\left(\frac{1}{2} \|\mathrm{curl}\tilde{\mathbf{A}} - \mathrm{curl}\tilde{\mathbf{A}}_\infty\|^2 + \frac{\varepsilon}{2} \|\tilde{\mathbf{A}}_t\|^2 + \alpha\varepsilon \left\langle \tilde{\mathbf{A}}_t, \tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty \right\rangle \right. \\
&\qquad \left. + \frac{\alpha\sigma}{2} \|\tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty\|^2 \right) \\
&+ \alpha \|\mathrm{curl}\tilde{\mathbf{A}} - \mathrm{curl}\tilde{\mathbf{A}}_\infty\|^2 + (\sigma - \alpha\varepsilon) \|\tilde{\mathbf{A}}_t\|^2 \\
&+ \left\langle \tilde{\mathbf{A}} \cdot \tilde{\mathbf{A}}_t, |\psi|^2 \right\rangle - \left\langle \tilde{\mathbf{A}}_\infty \cdot \tilde{\mathbf{A}}_t, |\psi_\infty|^2 \right\rangle \\
&+ \frac{i}{2} \left\langle \overline{\psi}\nabla\psi - \psi\nabla\overline{\psi}, \tilde{\mathbf{A}}_t \right\rangle - \frac{i}{2} \left\langle \overline{\psi}_\infty \nabla\psi_\infty - \psi_\infty \nabla\overline{\psi}_\infty, \tilde{\mathbf{A}}_t \right\rangle \\
&+ \alpha \left\langle |\psi|^2 \tilde{\mathbf{A}} - |\psi_\infty|^2 \tilde{\mathbf{A}}_\infty,\ \tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty \right\rangle \\
&+ \frac{\alpha i}{2} \left\langle \overline{\psi}\nabla\psi - \psi\nabla\overline{\psi}, \tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty \right\rangle - \frac{\alpha i}{2} \left\langle \overline{\psi}_\infty \nabla\psi_\infty - \psi_\infty \nabla\overline{\psi}_\infty, \tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty \right\rangle = 0.
\end{aligned} \tag{6.25}$$

Let us estimate some of the terms in (6.25). We have

$$\begin{aligned}
I_{17} &:= \left| \left\langle \tilde{\mathbf{A}} \cdot \tilde{\mathbf{A}}_t, |\psi|^2 \right\rangle - \left\langle \tilde{\mathbf{A}}_\infty \cdot \tilde{\mathbf{A}}_t, |\psi_\infty|^2 \right\rangle \right| \\
&\le \left| \left\langle \tilde{\mathbf{A}} \cdot \tilde{\mathbf{A}}_t, |\psi|^2 \right\rangle - \left\langle \tilde{\mathbf{A}}_\infty \cdot \tilde{\mathbf{A}}_t, |\psi|^2 \right\rangle \right| \\
&\quad + \left| \left\langle \tilde{\mathbf{A}}_\infty \cdot \tilde{\mathbf{A}}_t, |\psi|^2 \right\rangle - \left\langle \tilde{\mathbf{A}}_\infty \cdot \tilde{\mathbf{A}}_t, |\psi_\infty|^2 \right\rangle \right| \\
&\le \frac{\sigma}{16} \|\tilde{\mathbf{A}}_t\|^2 + C \left(\|\tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty\|^2 + \|\psi - \psi_\infty\|^2 \right),
\end{aligned} \tag{6.26}$$

$$\begin{aligned}
I_{18} &:= \left| \frac{i}{2} \left\langle \overline{\psi}\nabla\psi - \psi\nabla\overline{\psi}, \tilde{\mathbf{A}}_t \right\rangle - \frac{i}{2} \left\langle \overline{\psi}_\infty \nabla\psi_\infty - \psi_\infty \nabla\overline{\psi}_\infty, \tilde{\mathbf{A}}_t \right\rangle \right| \\
&\le \left| \left\langle \overline{\psi}\nabla\psi - \overline{\psi}_\infty \nabla\psi, \tilde{\mathbf{A}}_t \right\rangle \right| + \left| \left\langle \overline{\psi}_\infty \nabla\psi - \overline{\psi}_\infty \nabla\psi_\infty, \tilde{\mathbf{A}}_t \right\rangle \right| \\
&\quad + \left| \left\langle \psi\nabla\overline{\psi} - \psi_\infty \nabla\overline{\psi}, \tilde{\mathbf{A}}_t \right\rangle \right| + \left| \left\langle \psi_\infty \nabla\overline{\psi} - \psi_\infty \nabla\overline{\psi}_\infty, \tilde{\mathbf{A}}_t \right\rangle \right| \\
&\le \frac{\sigma}{16} \|\tilde{\mathbf{A}}_t\|^2 + C_7 \|\nabla\psi - \nabla\psi_\infty\|^2 + C\|\psi - \psi_\infty\|^2,
\end{aligned} \tag{6.27}$$

$$\begin{aligned}
I_{19} &:= \left| \alpha \left\langle |\psi|^2 \tilde{\mathbf{A}} - |\psi_\infty|^2 \tilde{\mathbf{A}}_\infty,\ \tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty \right\rangle \right| \\
&\le \alpha \left| \left\langle |\psi|^2 \tilde{\mathbf{A}} - |\psi|^2 \tilde{\mathbf{A}}_\infty,\ \tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty \right\rangle \right| + \alpha \left| \left\langle |\psi|^2 \tilde{\mathbf{A}}_\infty - |\psi_\infty|^2 \tilde{\mathbf{A}}_\infty,\ \tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty \right\rangle \right| \\
&\le C\alpha \left(\|\tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty\|^2 + \|\psi - \psi_\infty\|^2 \right),
\end{aligned} \tag{6.28}$$

$$\begin{aligned}
I_{20} &:= \left| \frac{\alpha i}{2} \left\langle \overline{\psi}\nabla\psi - \psi\nabla\overline{\psi}, \tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty \right\rangle - \frac{\alpha i}{2} \left\langle \overline{\psi}_\infty \nabla\psi_\infty - \psi_\infty \nabla\overline{\psi}_\infty, \tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty \right\rangle \right| \\
&\le \alpha \left| \left\langle \overline{\psi}\nabla\psi - \overline{\psi}_\infty \nabla\psi, \tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty \right\rangle \right| + \alpha \left| \left\langle \overline{\psi}_\infty \nabla\psi - \overline{\psi}_\infty \nabla\psi_\infty, \tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty \right\rangle \right| \\
&\quad + \alpha \left| \left\langle \psi\nabla\overline{\psi} - \psi_\infty \nabla\overline{\psi}, \tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty \right\rangle \right| + \alpha \left| \left\langle \psi_\infty \nabla\overline{\psi} - \psi_\infty \nabla\overline{\psi}_\infty, \tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty \right\rangle \right| \\
&\le \frac{\alpha}{16} \|\nabla\psi - \nabla\psi_\infty\|^2 + C\alpha \left(\|\psi - \psi_\infty\|^2 + \|\tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty\|^2 \right).
\end{aligned} \tag{6.29}$$

On the other hand, recalling (3.14) and (6.10), we observe that

$$I_{21} := \left| i \left\langle \Phi, \overline{\psi}\psi_t - \psi\overline{\psi}_t \right\rangle \right| \leq \frac{1}{16}\|\psi_t\|^2 + C\|\nabla\Phi\|^2. \tag{6.30}$$

Thus, adding (6.10), the product of (6.17) by constant $M > 0$, (6.25) and (3.14) together, using the above estimates of I_j $(j = 4, \ldots, 21)$, and taking

$$\alpha \in \left(0, \frac{\sigma}{2\varepsilon}\right), \quad M \geq \frac{\sigma}{\varepsilon} + 4C_6 + 2C_7, \quad \epsilon \in \left(0, \frac{3}{M}\right), \tag{6.31}$$

we can find two positive constants C and γ such that

$$\begin{aligned} &\frac{d}{dt}\left(\frac{1}{2}\|\nabla\psi - \nabla\psi_\infty\|^2 + \frac{M}{2}\|\psi - \psi_\infty\|^2 + \frac{1}{2}\left\|\mathrm{curl}\tilde{\mathbf{A}} - \mathrm{curl}\tilde{\mathbf{A}}_\infty\right\|^2\right. \\ &\left. + \frac{\varepsilon}{2}\left\|\tilde{\mathbf{A}}_t\right\|^2 + \alpha\varepsilon\left\langle \tilde{\mathbf{A}}_t, \tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty \right\rangle + \frac{\alpha\sigma}{2}\left\|\tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty\right\|^2 + \frac{\varepsilon}{2}\|\nabla\Phi\|^2\right) \\ &+ \gamma\left(\|\psi_t\|^2 + \|\nabla\psi - \nabla\psi_\infty\|^2 + \left\|\mathrm{curl}\tilde{\mathbf{A}} - \mathrm{curl}\tilde{\mathbf{A}}_\infty\right\|^2 + \left\|\tilde{\mathbf{A}}_t\right\|^2 + \|\nabla\Phi\|^2\right) \\ &\leq C\left(\left\|\tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty\right\|^2 + \|\psi - \psi_\infty\|^2 + \|\nabla\Phi\|^2\right). \end{aligned} \tag{6.32}$$

Let us now set

$$\begin{aligned} y := &\frac{1}{2}\|\nabla\psi - \nabla\psi_\infty\|^2 + \frac{M}{2}\|\psi - \psi_\infty\|^2 + \frac{1}{2}\left\|\mathrm{curl}\tilde{\mathbf{A}} - \mathrm{curl}\tilde{\mathbf{A}}_\infty\right\|^2 \\ &+ \frac{\varepsilon}{2}\left\|\tilde{\mathbf{A}}_t\right\|^2 + \alpha\varepsilon\left\langle \tilde{\mathbf{A}}_t, \tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty \right\rangle + \frac{\alpha\sigma}{2}\left\|\tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty\right\|^2 + \frac{\varepsilon}{2}\|\nabla\Phi\|^2. \end{aligned} \tag{6.33}$$

Choosing α small enough, by the Cauchy–Schwarz inequality, we get

$$c_1 y \leq \|\psi_t\|^2 + \|\nabla\psi - \nabla\psi_\infty\|^2 + \left\|\mathrm{curl}\tilde{\mathbf{A}} - \mathrm{curl}\tilde{\mathbf{A}}_\infty\right\|^2 + \left\|\tilde{\mathbf{A}}_t\right\|^2 + \|\nabla\Phi\|^2 \leq c_2 y. \tag{6.34}$$

Therefore, recalling (6.5), we have

$$\frac{d}{dt}y(t) + \hat{\gamma}y(t) \leq C(1+t)^{-2\theta/(1-2\theta)} \qquad \forall\, t \geq 0 \tag{6.35}$$

for some constant $\hat{\gamma} > 0$. Thus, it follows that (see also [14, 34, 35])

$$y(t) \leq C(1+t)^{-2\theta/(1-2\theta)} \qquad \forall\, t \geq 0. \tag{6.36}$$

Furthermore, on account of (6.34), we obtain the estimate of convergence rate in $\mathbb{X}^{-1}$-norm (cf. (2.5)), namely,

$$\|z(t) - z_\infty\|_{-1} \leq C(1+t)^{-\theta/(1-2\theta)} \qquad \forall\, t \geq 0. \tag{6.37}$$

To get (2.25), a further step is needed. We adapt a higher order estimate derived in [3, section 3]. Let us set

$$\begin{aligned} y_1 := &\alpha_1\|\psi_t\|^2 + \left\|\mathrm{curl}\tilde{\mathbf{A}}_t\right\|^2 + \varepsilon\left\|\tilde{\mathbf{A}}_{tt}\right\|^2 + 2\alpha_2\varepsilon\left\langle \tilde{\mathbf{A}}_t,\ \tilde{\mathbf{A}}_{tt} \right\rangle + \alpha_2\sigma\left\|\tilde{\mathbf{A}}_t\right\|^2 \\ &+ \varepsilon\|\Delta\Phi\|^2 + \sigma\|\nabla\Phi\|^2 \end{aligned} \tag{6.38}$$

and choose $\alpha_1, \alpha_2 > 0$ such that (cf. also (2.9))

$$\alpha_1 \geq \max\left\{\frac{\sigma}{\varepsilon}, \frac{1}{\sigma\kappa}\right\} + \frac{24}{\sigma}, \qquad 0 < \alpha_2 \leq \min\left\{\frac{\sigma}{4\varepsilon}, \frac{1}{4\sigma\kappa}\right\}. \tag{6.39}$$

Then we observe that (see [3, equations (3.7) and (3.8)])

$$\frac{1}{2}\left(\alpha_1\|\psi_t\|^2 + \left\|\mathrm{curl}\tilde{\mathbf{A}}_t\right\|^2 + \varepsilon\left\|\tilde{\mathbf{A}}_{tt}\right\|^2 + \varepsilon\|\Delta\Phi\|^2 + \sigma\|\nabla\Phi\|^2\right) \leq y_1, \tag{6.40}$$

$$y_1 \leq \frac{3}{2}\left(\alpha_1\|\psi_t\|^2 + \left\|\mathrm{curl}\tilde{\mathbf{A}}_t\right\|^2 + \varepsilon\left\|\tilde{\mathbf{A}}_{tt}\right\|^2 + \varepsilon\|\Delta\Phi\|^2 + \sigma\|\nabla\Phi\|^2\right). \tag{6.41}$$

Furthermore, a refinement of the argument used in [3, section 3, p. 628] yields

$$\frac{d}{dt}y_1 + \gamma_1 y_1 + \frac{\alpha_1}{2}\|\nabla\psi_t\|^2 + \varepsilon\|\nabla\Phi_t\|^2 \leq C\left(\|\psi_t\|^2 + \|\nabla\Phi\|^2 + \left\|\tilde{\mathbf{A}}_t\right\|^2\right), \tag{6.42}$$

for some $\gamma_1 > 0$, provided that α_1 and α_2 satisfy (6.39).

Hence, on account of (6.37), it follows that

$$\frac{d}{dt}y_1(t) + \gamma_1 y_1(t) \leq C(1+t)^{-2\theta/(1-2\theta)}, \tag{6.43}$$

which gives, using (6.40),

$$\|\psi_t(t)\| + \left\|\mathrm{curl}\tilde{\mathbf{A}}_t(t)\right\| + \left\|\tilde{\mathbf{A}}_{tt}(t)\right\| + \|\Delta\Phi(t)\| \leq C(1+t)^{-\theta/(1-2\theta)} \qquad \forall\, t \geq 0. \tag{6.44}$$

On the other hand, we infer from (6.6) that

$$\begin{aligned}\|\Delta(\psi-\psi_\infty)\| &\leq \|\psi_t\| + C\|\Phi\| + C\left\|\tilde{\mathbf{A}}\cdot\nabla\psi - \tilde{\mathbf{A}}_\infty\cdot\nabla\psi_\infty\right\| + \left\|\tilde{\mathbf{A}}^2\psi - \tilde{\mathbf{A}}_\infty^2\psi_\infty\right\| \\ &\quad + \left\|-\lambda^2\left(1-|\psi|^2\right)\psi + \lambda^2\left(1-|\psi_\infty|^2\right)\psi_\infty\right\|, \end{aligned} \tag{6.45}$$

$$\begin{aligned}\left\|\tilde{\mathbf{A}}\cdot\nabla\psi - \tilde{\mathbf{A}}_\infty\cdot\nabla\psi_\infty\right\| &\leq \left\|\tilde{\mathbf{A}}\cdot\nabla\psi - \tilde{\mathbf{A}}_\infty\cdot\nabla\psi\right\| + \left\|\tilde{\mathbf{A}}_\infty\cdot\nabla\psi - \tilde{\mathbf{A}}_\infty\cdot\nabla\psi_\infty\right\| \\ &\leq \left\|\tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty\right\|_{\mathbf{L}^4}\|\nabla\psi\|_{\mathbf{L}^4} + \left\|\tilde{\mathbf{A}}_\infty\right\|_{\mathbf{L}^\infty}\|\nabla\psi - \nabla\psi_\infty\| \\ &\leq C\left(\left\|\mathrm{curl}\tilde{\mathbf{A}} - \mathrm{curl}\tilde{\mathbf{A}}_\infty\right\| + \|\nabla\psi - \nabla\psi_\infty\|\right), \end{aligned} \tag{6.46}$$

$$\begin{aligned}\left\|\tilde{\mathbf{A}}^2\psi - \tilde{\mathbf{A}}_\infty^2\psi_\infty\right\| &\leq \left\|\tilde{\mathbf{A}}^2\psi - \tilde{\mathbf{A}}^2\psi_\infty\right\| + \left\|\tilde{\mathbf{A}}^2\psi_\infty - \tilde{\mathbf{A}}_\infty^2\psi_\infty\right\| \\ &\leq \left\|\tilde{\mathbf{A}}^2\right\|_{\mathbf{L}^\infty}\|\psi - \psi_\infty\| + \left\|\tilde{\mathbf{A}} + \tilde{\mathbf{A}}_\infty\right\|_{\mathbf{L}^\infty}\left\|\tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty\right\|\|\psi_\infty\|_{L^\infty_{\mathbb{C}}} \\ &\leq C\left(\left\|\tilde{\mathbf{A}} - \tilde{\mathbf{A}}_\infty\right\| + \|\psi - \psi_\infty\|\right), \end{aligned} \tag{6.47}$$

$$\left\|-\lambda^2\left(1-|\psi|^2\right)\psi + \lambda^2\left(1-|\psi_\infty|^2\right)\psi_\infty\right\| \leq C\|\psi - \psi_\infty\|. \tag{6.48}$$

Therefore, due to (6.37), we deduce

$$\|\Delta(\psi(t) - \psi_\infty)\| \leq C(1+t)^{-\theta/(1-2\theta)} \qquad \forall\, t \geq 0. \tag{6.49}$$

Similarly, from (6.7), we obtain

$$\begin{aligned}&\left\|\mathrm{curl}^2\tilde{\mathbf{A}} - \mathrm{curl}^2\tilde{\mathbf{A}}_\infty\right\| \\ &\leq \varepsilon\left\|\tilde{\mathbf{A}}_{tt}\right\| + \varepsilon\|\nabla\Phi_t\| + \sigma\left\|\tilde{\mathbf{A}}_t\right\| + \sigma\|\nabla\Phi\| + \left\||\psi|^2\tilde{\mathbf{A}} - |\psi_\infty|^2\tilde{\mathbf{A}}_\infty\right\| \\ &\quad + \left\|\frac{i}{2}\left(\overline{\psi}\nabla\psi - \psi\nabla\overline{\psi}\right) - \frac{i}{2}\left(\overline{\psi}_\infty\nabla\psi_\infty - \psi_\infty\nabla\overline{\psi}_\infty\right)\right\|. \end{aligned} \tag{6.50}$$

Besides, the following estimates hold:

$$\left\||\psi|^2\tilde{\mathbf{A}} - |\psi_\infty|^2\tilde{\mathbf{A}}_\infty\right\| \le C\left(\|\psi-\psi_\infty\| + \left\|\tilde{\mathbf{A}}-\tilde{\mathbf{A}}_\infty\right\|\right), \tag{6.51}$$

$$\left\|\frac{i}{2}\left(\overline{\psi}\nabla\psi - \psi\nabla\overline{\psi}\right) - \frac{i}{2}\left(\overline{\psi}_\infty\nabla\psi_\infty - \psi_\infty\nabla\overline{\psi}_\infty\right)\right\| \le C\left(\|\psi-\psi_\infty\| + \|\nabla\psi-\nabla\psi_\infty\|\right). \tag{6.52}$$

As a result, owing to (6.37), we infer from (5.7) and (6.50)–(6.52) that

$$\left\|\mathrm{curl}^2\tilde{\mathbf{A}}(t) - \mathrm{curl}^2\tilde{\mathbf{A}}_\infty\right\| \le C(1+t)^{-\theta/(1-2\theta)} \qquad \forall\, t \ge 0. \tag{6.53}$$

Finally, (2.25) follows from (6.37), (6.44), (6.49), and (6.53).

The proof of Theorem 2.1 is now complete.

Acknowledgments. This paper was conceived on occasion of a first author's visit to the Institute of Mathematics in Fudan University, whose hospitality is gratefully acknowledged.

REFERENCES

[1] T. Akiyama and Y. Shibata, *On an L^p approach to the stationary and nonstationary problems of the Ginzburg–Landau–Maxwell equations*, J. Differential Equations, 243 (2007), pp. 1–23.

[2] J. Berger, *Time-dependent Ginzburg–Landau equations with charged boundaries*, J. Math. Phys., 46 (2005), p. 095106.

[3] V. Berti and S. Gatti, *Parabolic-hyperbolic time-dependent Ginzburg–Landau–Maxwell equations*, Quart. Appl. Math., 64 (2006), pp. 617–639.

[4] T. Cazenave and A. Haraux, *An Introduction to Semilinear Evolution Equations*, Oxford University Press, New York, 1998.

[5] S.J. Chapman, Q. Du, M.D. Gunzburger, and J.S. Peterson, *Simplified G–L models for superconductivity valid for high kappa and high fields*, Adv. Math. Sci. Appl., 5 (1995), pp. 193–218.

[6] K. Deimling, *Nonlinear functional analysis*, Springer, New York, 1985.

[7] Q. Du, *Global existence and uniqueness of solutions of the time-dependent Ginzburg-Landau model for superconductivity*, Appl. Anal., 53 (1994), pp. 1–18.

[8] M. Fabrizio and A. Morro, *Electromagnetism of Continuous Media. Mathematical Modelling and Applications*, Oxford University Press, Oxford, 2003.

[9] E. Feireisl and P. Takáč, *Long-time stabilization of solutions to the Ginzburg–Landau equations of superconductivity*, Monatsh. Math., 133 (2001), pp. 197–221.

[10] J. Fleckinger-Pellé and H.G. Kaper, *Gauges for the Ginzburg-Landau equations of superconductivity*, Z. Angew. Math. Mech., 96 (1996), pp. 345–348.

[11] J. Fleckinger-Pellé, H.G. Kaper, and P. Takáč, *Dynamics of the Ginzburg-Landau equations of superconductivity*, Nonlinear Anal., 32 (1998), pp. 647–665.

[12] V. Girault and P.A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer, New York, 1986.

[13] L.P. Gor'kov and G.M. Èliashberg, *Generalizations of Ginzburg-Landau equations for nonstationary problems in the case of alloys with paramagnetic impurities*, Soviet. Phys. JETP, 27 (1968), pp. 328–334.

[14] M. Grasselli, H. Wu, and S. Zheng, *Asymptotic behavior of a non-isothermal Ginzburg–Landau model*, Quart. Appl. Math., to appear.

[15] A. Haraux and M.A. Jendoubi, *Convergence of bounded weak solutions of the wave equation with dissipation and analytic nonlinearity*, Calc. Var. Partial Differential Equations., 9 (1999), pp. 95–124.

[16] A. Haraux, *Systèmes Dynamiques Dissipatifs et Applications*, Masson, Paris, 1991.

[17] A. Haraux and M.A. Jendoubi, *Decay estimates to equilibrium for some evolution equations with an analytic nonlinearity*, Asymptot. Anal., 26 (2001), pp. 21–36.

[18] K.-H. Hoffmann and Q. Tang, *Ginzburg-Landau phase-transition theory and superconductivity*, Internat. Ser. Numer. Math. 134, Birkhäuser, Basel, 2001.

[19] S.-Z. Huang and P. Takáč, *Convergence in gradient-like systems which are asymptotically autonomous and analytic*, Nonlinear Anal., 46 (2001), pp. 675–698.
[20] M.A. Jendoubi, *A simple unified approach to some convergence theorem of L. Simon*, J. Funct. Anal., 153 (1998), pp. 187–202.
[21] H.G. Kaper and P. Takáč, *An equivalence relation for the Ginzburg–Landau equations of superconductivity*, Z. Angew. Math. Phys., 48 (1997), pp. 665–675.
[22] H.G. Kaper and P. Takáč, *Ginzburg-Landau dynamics with a time-dependent magnetic field*, Nonlinearity, 11 (1998), pp. 291–305.
[23] J. Liang and Q. Tang, *Asymptotic behavior of the solutions of an evolutionary Ginzburg–Landau superconductivity model*, J. Math. Anal. Appl., 195 (1995), pp. 92–107.
[24] F.-H. Lin and Q. Du, *Ginzburg–Landau vortices: Dynamics, pinning, and hysteresis*, SIAM J. Math. Anal., 28 (1997), pp. 1265–1293.
[25] P. Monk, *Finite Element Methods for Maxwell's Equation*, Clarendon Press, Oxford, 2003.
[26] A. Rodriguez-Bernal, B. Wang, and R. Willie, *Asymptotic behavior of the time-dependent Ginzburg-Landau equations of superconductivity*, Math. Methods Appl. Sci., 22 (1999), pp. 1647–1669.
[27] A. Schmid, *A time dependent Ginzburg-Landau equation and its application to the problem of resistivity in the mixed state*, Phys. Kondens. Mater., 5 (1966), pp. 302–317.
[28] L. Simon, *Asymptotics for a class of nonlinear evolution equation with applications to geometric problems*, Ann. of Math., 118 (1983), pp. 525–571.
[29] Q. Tang, *On an evolutionary system of Ginzburg-Landau equations with fixed total magnetic flux*, Comm. Partial Differential Equations, 20 (1995), pp. 1–36.
[30] Q. Tang and S. Wang, *Time dependent Ginzburg-Landau equations of superconductivity*, Phys. D, 88 (1995), pp. 139–166.
[31] R. Temam, *Infinite-dimensional dynamical systems in mechanics and physics*, Appl. Math. Sci. 68, Springer, New York, 1988.
[32] M. Tsutsumi and H. Kasai, *The time-dependent Ginzburg-Landau-Maxwell equations*, Nonlinear Anal., 37 (1999), pp. 187–216.
[33] S. Wang and M.Q. Zhan, *L^p solutions to the time-dependent Ginzburg-Landau equations of superconductivity*, Nonlinear Anal., 36 (1999), pp. 661–677.
[34] H. Wu, *Convergence to equilibrium for a Cahn-Hilliard model with the Wentzell boundary condition*, Asymptot. Anal., 54 (2007), pp. 71–92.
[35] H. Wu, M. Grasselli, and S. Zheng, *Convergence to equilibrium for a parabolic–hyperbolic phase–field system with Neumann boundary conditions*, Math. Models Methods Appl. Sci., 17 (2007), pp. 1–29.
[36] S. Zheng, *Nonlinear evolution equations*, Chapman & Hall-CRC Monogr. Surv. Pure Appl. Math. 133, Chapman & Hall/CRC, Boca Raton, FL, 2004.

SIAM J. MATH. ANAL.
Vol. 40, No. 5, pp. 2034–2054

ON A FOURTH ORDER NONLINEAR ELLIPTIC EQUATION WITH NEGATIVE EXPONENT*

ZONGMING GUO† AND JUNCHENG WEI‡

Abstract. We consider the following nonlinear fourth order equation: $T\Delta u - D\Delta^2 u = \frac{\lambda}{(L+u)^2}$, $-L < u < 0$, in Ω, $u = 0$, $\Delta u = 0$ on $\partial\Omega$, where $\lambda > 0$ is a parameter. This nonlinear equation models the deflection of charged plates in electrostatic actuators under the pinned boundary condition (Lin and Yang [*Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 463 (2007), pp. 1323–1337]). Lin and Yang proved that there exists a $\lambda_c > 0$ such that for $\lambda > \lambda_c$ there is no solution, while for $\lambda < \lambda_c$ there is a branch of maximal solutions. In this paper, we show that in the physical domains (two or three dimensions) the maximal solution is unique and regular at $\lambda = \lambda_c$. In a two-dimensional (2D) convex smooth domain, we also establish the existence of a second mountain-pass solution for $\lambda \in (0, \lambda_c)$. The asymptotic behavior of the second solution is also studied. The main difficulty is the analysis of the touch-down behavior.

Key words. fourth order, electrostatic actuation, touch-down, pull-in threshold

AMS subject classifications. 35B45, 35J40

DOI. 10.1137/070703375

1. Introduction.

We consider the structure of solutions to the problem

$$(P_\lambda) \qquad \begin{cases} T\Delta u - D\Delta^2 u = \frac{\lambda}{(L+u)^2} & \text{in } \Omega, \\ -L < u \le 0 & \text{in } \Omega, \\ u = 0, \quad \Delta u = 0 & \text{on } \partial\Omega, \end{cases}$$

where $\lambda > 0$ is a parameter, $T > 0$, $D > 0$, and $L > 0$ are fixed constants, and $\Omega \subset \boldsymbol{R}^N$ $(N \ge 2)$ is a bounded smooth domain.

When $D = 0$, problem (P_λ) becomes

$$(Q_\lambda) \qquad \begin{cases} T\Delta u = \frac{\lambda}{(L+u)^2} & \text{in } \Omega, \\ -L < u \le 0 & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

which models a simple electrostatic microelectromechanical system (MEMS) device consisting of a thin dielectric elastic membrane with boundary supported at 0 above a rigid plate located at $-L$. Here $L + u$ represents the distance from the membrane to the plate. Recently there have been many studies on (Q_λ). See, for example, [9], [14], [15], [16], [10], [11], [12], [7], [8], [17], [25], [24], and the references therein. These papers deal only with second order semilinear elliptic equations with singular nonlinearities. Equation (Q_λ) also appears in the study of thin film; see, for example, [2], [3], [5], [19], [20], [21], [18], and the references therein.

*Received by the editors September 21, 2007; accepted for publication (in revised form) October 2, 2008; published electronically January 21, 2009.

http://www.siam.org/journals/sima/40-5/70337.html

†Department of Mathematics, Henan Normal University, Xinxiang, 453007, People's Republic of China (gzm@henannu.edu.cn). The research of this author was supported by grants of NSFC (10571022 and 10871060).

‡Department of Mathematics, The Chinese University of Hong Kong, Shatin, Hong Kong (wei@math.cuhk.edu.hk). The research of this author was partially supported by earmarked grants from RGC of Hong Kong and a direct grant of CUHK.

In a recent paper [22], Lin and Yang derived the fourth order equation (P_λ) in the study of the deflection of charged plates in electrostatic actuators. Here $\lambda = aV^2$, where V is the electric voltage and a is positive constant. Associated with (P_λ) is the energy functional

$$E(u) = \int_\Omega \left\{ \frac{T}{2}|\nabla u|^2 + \frac{D}{2}|\Delta u|^2 - \frac{\lambda}{L+u} \right\}, \tag{1.1}$$

where $P = \int_\Omega \frac{T}{2}|\nabla u|^2 dx$ is the stretching energy, $Q = \int_\Omega \frac{D}{2}|\Delta u|^2 dx$ corresponds to the bending energy, and $W = -\int_\Omega \frac{\lambda}{L+u(x)} dx$ is the electric potential energy.

Lin and Yang [22] considered two kinds of boundary conditions: the pinned boundary condition

$$u = \Delta u = 0 \text{ on } \partial\Omega$$

and the clamped boundary condition

$$u = \frac{\partial u}{\partial n} = 0 \text{ on } \partial\Omega.$$

For the pinned boundary condition problem (P_λ), they found that there exists $0 < \lambda_c < \infty$ such that for $\lambda \in (0, \lambda_c)$, (P_λ) has a maximal regular solution u_λ, which can be obtained from an iterative scheme. (By a regular solution u_λ of (P_λ), we mean that $u_\lambda \in C^4(\Omega) \cap C^3(\overline{\Omega})$ satisfies (P_λ).) For $\lambda > \lambda_c$, (P_λ) does not have any regular solution. Moreover, if λ', $\lambda'' \in (0, \lambda_c)$ and $\lambda' < \lambda''$, then the corresponding maximal solutions $u_{\lambda'}$ and $u_{\lambda''}$ satisfy

$$u_{\lambda'} > u_{\lambda''} \quad \text{in } \Omega.$$

Physically, this is a natural relation because a higher supply voltage results in a greater elastic deformation or deflection.

The number λ_c, which determines the pull-in voltage, is called the pull-in threshold. It is known from [22] that, for $\lambda \in (0, \lambda_c)$, $\min_\Omega(L + u_\lambda) > 0$. Let $\Sigma_\lambda = \{x \in \Omega : L + u_\lambda(x) = 0\}$ be the singular set of (P_λ). An interesting question is to study the limit of u_λ as $\lambda \nearrow \lambda_c$. The monotonicity of u_λ with respect to λ implies that there is a well-defined function U so that

$$U(x) = \lim_{\lambda \to \lambda_c^-} u_\lambda(x); \quad -L \le U(x) < 0, \quad x \in \Omega.$$

However, $U(x)$ may touch down to $-L$ and cease to be a regular solution to (P_{λ_c}). (By [22], $U \in W^{2,2}_{loc}(\Omega)$.) For the one-dimensional case, Lin and Yang showed that U is a regular solution; that is, the set $\Sigma_{\lambda_c} = \emptyset$.

In this paper, we will show that *for two dimensions and three dimensions, U is a regular solution*. Moreover, we also show that there is a *unique solution for* (P_λ) *at* $\lambda = \lambda_c$. To obtain our results, we first prove that the solutions u_λ for $\lambda \in (0, \lambda_c)$ obtained in [22] are stable in some sense. Furthermore, we also obtain the structure of solutions of (P_λ) in the two-dimensional (2D) case. Our main results of this paper are as follows.

THEOREM 1.1. *For dimension $N = 2$ or 3, there exists a constant $0 < C := C(N, L)$ independent of λ such that for any $0 < \lambda < \lambda_c$, the maximal solution u_λ of (P_λ) satisfies* $\min_\Omega(L + u_\lambda) \ge C$.

Consequently, $u_{\lambda_c} = \lim_{\lambda \nearrow \lambda_c} u_\lambda$ *exists in the topology of* $C^4(\Omega)$. *It is the unique regular solution to* (P_{λ_c}).

THEOREM 1.2. *Let* $N = 2$ *and* Ω *be a bounded, smooth, and convex domain in* $\boldsymbol{R}^2$. *For* $\lambda \in (0, \lambda_c]$, *any solution of the problem* (P_λ) *is regular and the following hold.*

(i) *For* $0 < \lambda < \lambda_c$, *problem* (P_λ) *admits two solutions: the maximal solution and a mountain-pass solution.*

(ii) *For* $\lambda = \lambda_c$, *problem* (P_λ) *admits a unique regular solution.*

(iii) *For* $\lambda > \lambda_c$, *problem* (P_λ) *admits no regular solution.*

Furthermore, the mountain-pass solution V_λ *has the following asymptotic behavior as* $\lambda \to 0$*:*

$$\max_\Omega V_\lambda \to L \ \text{ as } \lambda \to 0, \quad \lim_{\lambda \to 0^+} \frac{[\min_\Omega (L - V_\lambda)]^3}{\lambda} = 0. \tag{1.2}$$

Remark. Theorem 1.2 shows that the bifurcation diagram of (P_λ) changes drastically when $D > 0$. In a nice 2D domain (see [16]), it has been proved in [16] that for λ small, the maximal solution is *unique*, and there exists $0 < \lambda_* < \lambda_c$ such that the solutions to (Q_λ) undergo infinitely many turning points. An interesting question is the asymptotic behavior as $D \to 0$. When $\Omega = B_1 \subset \boldsymbol{R}^2$, the complete bifurcation picture as well as the asymptotic behavior when $D \to 0$ has been considered in [23].

The organization of the paper is as follows: in section 2, we present some preliminary results on the first eigenvalue and the corresponding eigenfunction of the problem

$$-T\Delta\phi + D\Delta^2\phi = \sigma\phi \ \text{ in } \Omega, \quad \phi = \Delta\phi = 0 \ \text{ on } \partial\Omega.$$

In section 3, we derive a key L^1 bound for $\frac{1}{(L+u)^2}$. In section 4, we show the stability of the maximal solutions of (P_λ). In section 5, we show that the solution at the pull-in threshold is regular for $N = 2$ or 3. In section 6, we show that any weak solution at the pull-in threshold is unique. In section 7, we present the structure of the solutions of (P_λ) for the 2D case. We show that for $0 < \lambda < \lambda_c$, (P_λ) admits at least two solutions: the maximal solution and a mountain-pass solution. Finally in section 8, we give some asymptotic behaviors of the mountain-pass solution as $\lambda \to 0^+$.

2. The first eigenfunction. In this section, we study the following eigenvalue problem:

$$-T\Delta\phi + D\Delta^2\phi = \sigma\phi \ \text{ in } \Omega, \quad \phi = \Delta\phi = 0 \ \text{ on } \partial\Omega, \tag{2.1}$$

where T, $D > 0$. We will show that (2.1) has the least eigenvalue σ_1 and the corresponding eigenfunction $\phi_1 > 0$ in Ω. Moreover, ϕ_1 is simple; i.e., all the eigenfunctions corresponding to σ_1 assume the forms of $C\phi_1$ with $C \in \boldsymbol{R}$.

PROPOSITION 2.1. *Problem* (2.1) *has the least eigenvalue* σ_1 *such that all the eigenfunctions corresponding to* σ_1 *assume the forms of* $C\phi_1$, *where* $\phi_1 \in C^\infty(\Omega)$ *and* $\phi_1 > 0$ *in* Ω.

Proof. This proposition may be known, but we cannot find the reference. We give a proof here for completeness.

Consider the following minimization problem:

$$\sigma_1 := \inf \left\{ \int_\Omega [T|\nabla\phi|^2 + D|\Delta\phi|^2]dx : \ \phi \in \mathcal{H}, \ \ \|\phi\|_{L^2(\Omega)} = 1 \right\}, \tag{2.2}$$

where $\mathcal{H} = H^2(\Omega) \cap H_0^1(\Omega)$ is the function space obtained by taking the completion under the norm of $H^2(\Omega) \cap H_0^1(\Omega)$ (i.e., $\|\psi\| = (\int_\Omega [T|\nabla\psi|^2 + D|\Delta\psi|^2]dx)^{1/2}$) for the set of smooth functions that satisfy the boundary condition $\phi = \Delta\phi = 0$ on $\partial\Omega$. Since the Sobolev embedding $\mathcal{H} \hookrightarrow L^2(\Omega)$ is compact, by the standard direct method of calculus of variations, we have at least one minimizer ϕ_1 for the problem (2.2), where $\phi_1 \in \mathcal{H}$, $\|\phi_1\|_{L^2(\Omega)} = 1$. Furthermore, ϕ_1 is a weak solution to (2.1); namely,

$$\int_\Omega [T\nabla\phi_1\nabla\phi + D\Delta\phi_1\Delta\phi] = \sigma_1 \int_\Omega \phi_1\phi dx \quad \forall\phi \in \mathcal{H}. \tag{2.3}$$

Using the L^p-estimates due to Agmon, Douglis, and Nirenberg [1], we conclude that

$$\|\phi_1\|_{W^{4,p}(\Omega)} \le C\|\phi_1\|_{L^p(\Omega)}$$

for any $p > 1$. Thus we have $\phi_1 \in C^4(\Omega) \cap C^3(\bar\Omega)$, and hence $\Delta\phi_1 = 0$ on $\partial\Omega$, and ϕ_1 satisfies (2.1). (See a similar argument in Lemma B.3 of [27].)

It is clear that

$$\sigma_1 = \frac{\int_\Omega [T|\nabla\phi_1|^2 + D|\Delta\phi_1|^2]dx}{\int_\Omega \phi_1^2 dx} = \inf_{\phi\in\mathcal{H}\setminus\{0\}} \frac{\int_\Omega [T|\nabla\phi|^2 + D|\Delta\phi|^2]dx}{\int_\Omega \phi^2 dx}.$$

In order to show that ϕ_1 is of fixed sign, we consider the following new problem:

$$-T\Delta\psi_1 + D\Delta^2\psi_1 = \sigma_1|\phi_1| \ \text{ in } \Omega, \quad \psi_1 = \Delta\psi_1 = 0 \ \text{ on } \partial\Omega. \tag{2.4}$$

By the maximum principle, $\psi_1 > 0$, $-D\Delta\psi_1 + T\psi_1 > 0$ in Ω. Furthermore, we have $\psi_1 \ge \phi_1, \psi_1 \ge -\phi_1$, and hence $\psi_1 \ge |\phi_1|$ in Ω.

On the other hand, from (2.4) we obtain

$$\int_\Omega [T|\nabla\psi_1|^2 + D|\Delta\psi_1|^2]dx = \sigma_1 \int_\Omega \psi_1|\phi_1|dx \le \sigma_1 \int_\Omega |\psi_1|^2 dx. \tag{2.5}$$

By the minimality of σ_1, we have

$$\sigma_1 = \frac{\int_\Omega [T|\nabla\psi_1|^2 + D|\Delta\psi_1|^2]dx}{\int_\Omega |\psi_1|^2 dx}. \tag{2.6}$$

Thus ψ_1 also attains σ_1, and hence the inequality of (2.5) is actually an equality. This implies that $\psi_1 = |\phi_1|$ in Ω. Since $\psi_1 > 0$ in Ω, we conclude that ϕ_1 is of fixed sign in Ω.

The above argument actually proves that any nonzero eigenfunction corresponding to σ_1 must be of fixed sign in Ω. So if ϕ_1 and ϕ_2 are two eigenfunctions corresponding to σ_1, we may choose $\phi_1 > 0, \phi_2 > 0$. Let $x_0 \in \Omega$ and $C = \frac{\phi_1(x_0)}{\phi_2(x_0)}$. Then the function $\phi_1 - C\phi_2$ is again an eigenfunction corresponding to σ_1. By the previous argument, we see that $\phi_1 \equiv C\phi_2$ in Ω. This completes the proof. □

3. A uniform L^1 bound. In this section, we establish a key uniform L^1 bound for $\frac{1}{(L-v)^2}$, where v satisfies

$$(T_\lambda) \qquad \begin{cases} -T\Delta v + D\Delta^2 v = \frac{\lambda}{(L-v)^2} & \text{in } \Omega, \\ 0 < v < L & \text{in } \Omega, \\ v = 0, \quad \Delta v = 0 & \text{on } \partial\Omega \end{cases}$$

(which is equivalent to (P_λ) by taking $u = -v$). Note that $v \in C^4(\Omega)\cap C^2(\overline{\Omega})$ provided that v satisfies (T_λ).

THEOREM 3.1. *Let Ω be a bounded, smooth, and convex domain. Then there exists a constant C (independent of λ) such that for any solution v to (T_λ) we have*

$$\int_\Omega \frac{1}{(L-v)^2} \le \frac{C}{\lambda}. \tag{3.1}$$

As a consequence, we have

$$\int_\Omega (D|\Delta v|^2 + T|\nabla v|^2) \le C. \tag{3.2}$$

Proof. Let ϕ_1 be as given in Proposition 2.1. Multiplying (T_λ) by ϕ_1 and integrating over Ω, we obtain

$$\lambda \int_\Omega \frac{1}{(L-v)^2}\phi_1 = \sigma_1 \int_\Omega v\phi_1 \le C, \tag{3.3}$$

which implies that

$$\int_{\Omega'} \frac{1}{(L-v)^2} \le \frac{C_{\Omega'}}{\lambda} \tag{3.4}$$

for any $\Omega' \subset\subset \Omega$, where $C_{\Omega'}$ is independent of λ.

We write (T_λ) as

$$\begin{cases} \Delta v + \frac{1}{D}w - \frac{T}{D}v = 0 & \text{in } \Omega, \\ \Delta w + \frac{\lambda}{(L-v)^2} = 0 & \text{in } \Omega, \\ v = w = 0 & \text{on } \partial\Omega. \end{cases}$$

If we denote $f_1(v,w) = -\frac{T}{D}v + \frac{1}{D}w$, $f_2(v,w) = \frac{\lambda}{(L-v)^2}$, we see that $\frac{\partial f_1}{\partial w} = \frac{1}{D} > 0$ and $\frac{\partial f_2}{\partial v} = \frac{2\lambda}{(L-v)^3} > 0$. Therefore, the convexity of Ω, Lemma 5.1 of [26], and the moving plane method near $\partial\Omega$ as in the appendix of [13] imply that there exist $t_0 > 0$ and $\alpha > 0$ depending only on the domain Ω, such that *$v(x - t\nu)$ and $w(x - t\nu)$ are nondecreasing for $t \in [0, t_0]$, $\nu \in R^N$ satisfying $|\nu| = 1$ and $(\nu, n(x)) \ge \alpha$ and $x \in \partial\Omega$.* Therefore, we can find $\gamma, \delta > 0$ such that for any $x \in \Omega_\delta := \{z \in \bar{\Omega} : d(z, \partial\Omega) < \delta\}$ there exists a fixed-sized cone Γ_x (with x as its vertex) with

(i) $meas(\Gamma_x) \ge \gamma$,

(ii) $\Gamma_x \subset \{z \in \Omega : d(z, \partial\Omega) < \delta\}$, and

(iii) $v(y) \ge v(x)$ for any $y \in \Gamma_x$.

Then, for any $x \in \Omega_\delta$, we have

$$\frac{1}{(L-v(x))^2} \le \frac{1}{meas(\Gamma_x)} \int_{\Gamma_x} \frac{1}{(L-v)^2} \le \frac{1}{\gamma} \int_{\Omega_\delta} \frac{1}{(L-v)^2} \le \frac{C}{\lambda}.$$

This implies that $\frac{1}{(L-v)^2} \in L^\infty(\Omega_\delta)$ and there is $C > 0$ independent of λ such that

$$\sup_{x\in\Omega_\delta} v < L - C\sqrt{\lambda}. \tag{3.5}$$

Next, we derive estimates for w near $\partial\Omega$. Multiplying the second equation of the equivalent system of (T_λ) by φ_0, the first eigenfunction of $-\Delta$, and integrating over Ω, we obtain

$$\lambda \int_\Omega \frac{1}{(L-v)^2}\varphi_0 = \lambda_1 \int_\Omega \varphi_0 w, \tag{3.6}$$

where λ_1 is the first eigenvalue of $-\Delta$. By (3.3), we have that

$$\lambda_1 \int_\Omega \varphi_0 w = \lambda \int_\Omega \frac{1}{(L-v)^2}\varphi_0 \le \lambda C \int_\Omega \frac{1}{(L-v)^2}\phi_1 \le C, \tag{3.7}$$

and hence

$$\int_{\Omega'} w \le C_{\Omega'}$$

for any $\Omega' \subset\subset \Omega$. To see the second inequality of (3.7), we notice that there exist $\ell_i > 0$ $(i = 1, 2, 3, 4)$ such that

$$\ell_1 d(x) \le \varphi_0(x) \le \ell_2 d(x), \quad \ell_3 d(x) \le \phi_1(x) \le \ell_4 d(x),$$

where $d(x) = \operatorname{dist}(x, \partial\Omega)$. Hence $\varphi_0(x) \le C\phi_1(x)$. The same reason as above shows that $w \le C(\Omega_\delta)$.

By elliptic regularity applied to the system (T_λ) (noting that $v, w, and\ \frac{1}{(L-v)^2}$ are all bounded in Ω_δ), we have $v \in C^3(\Omega_\delta)$, and hence

$$\lambda \int_\Omega \frac{1}{(L-v)^2} = D \int_{\partial\Omega} \frac{\partial \Delta v}{\partial n} - T \int_\Omega \frac{\partial v}{\partial n} \le C.$$

To prove the inequality (3.2), we multiply (T_λ) by v and integrate over Ω to obtain

$$T \int_\Omega |\nabla v|^2 + D \int_\Omega |\Delta v|^2 = \lambda \int_\Omega \frac{Lv}{(L-v)^2} \le C. \qquad \square$$

4. Stability of the maximal solutions of (P_λ). In this section, we show that the maximal solutions u_λ to (P_λ) obtained in [22] for $\lambda \in (0, \lambda_c)$ are stable in some sense. Let $v_\lambda = -u_\lambda$. Then, from [22], for each $\lambda \in (0, \lambda_c)$, (T_λ) has a minimal positive solution v_λ.

We call v_λ stable if the first eigenvalue $\sigma_{1,\lambda}(v_\lambda)$ of the problem

$$-T\Delta h + D\Delta^2 h = \frac{2\lambda}{(L-v_\lambda)^3}h + \sigma h \ \text{ in } \Omega, \quad h = \Delta h = 0 \ \text{ on } \partial\Omega \tag{4.1}$$

is nonnegative. By arguments similar to those in the proof of Proposition 2.1, we see that the first eigenvalue $\sigma_{1,\lambda}(v_\lambda)$ exists and every eigenfunction corresponding to $\sigma_{1,\lambda}(v_\lambda)$ is of fixed sign if $\sigma_{1,\lambda}(v_\lambda) \ge 0$.

LEMMA 4.1. *Suppose that v is a regular solution of (T_λ), and u is a regular supersolution of (T_λ); that is,*

$$\begin{cases} -T\Delta u + D\Delta^2 u \ge \frac{\lambda}{(L-u)^2} & \text{in } \Omega, \\ 0 < u < L & \text{in } \Omega, \\ u = 0, \ \ \Delta u = 0 & \text{on } \partial\Omega. \end{cases}$$

If $\sigma_{1,\lambda}(v) > 0$, then $u \geq v$ in Ω, and if $\sigma_{1,\lambda}(v) = 0$, then $u = v$ in Ω.

Proof. For a given λ and $x \in \Omega$, by the fact that $s \to (L-s)^{-2}$ is convex on $(0, L)$, we see that

$$-T\Delta(v+\tau(u-v)) + D\Delta^2(v+\tau(u-v)) - \frac{\lambda}{[L-(v+\tau(u-v))]^2} \geq 0 \text{ in } \Omega \tag{4.2}$$

for $\tau \in [0,1]$. Note that (4.2) is an identity at $\tau = 0$, which means that the first derivative of the left-hand side of (4.2) with respect to τ is nonnegative at $\tau = 0$, i.e.,

$$\begin{cases} -T\Delta(u-v) + D\Delta^2(u-v) - \frac{2\lambda}{(L-v)^3}(u-v) \geq 0 & \text{in } \Omega, \\ u - v = 0, \ \ \Delta(u-v) = 0 & \text{on } \partial\Omega. \end{cases} \tag{4.3}$$

Thus, the fact $\sigma_{1,\lambda}(v) > 0$ implies that $u \geq v$ in Ω. Indeed, on the contrary, we see that $0 \not\equiv (u-v)^- \in H^2(\Omega) \cap H_0^1(\Omega)$. Multiplying $(u-v)^-$ on both the sides of (4.3) and integrating it on Ω, we see that

$$\begin{aligned} &\sigma_{1,\lambda}(v) \int_\Omega [(u-v)^-]^2 dx \\ &\quad \leq T \int_\Omega |\nabla(u-v)^-|^2 dx + D \int_\Omega |\Delta(u-v)^-|^2 dx - \int_\Omega \frac{2\lambda}{(L-v)^3}[(u-v)^-]^2 dx \\ &\quad \leq 0. \end{aligned}$$

This contradicts $\sigma_{1,\lambda}(v) > 0$.

If $\sigma_{1,\lambda}(v) = 0$, we have

$$-T\Delta(u-v) + D\Delta^2(u-v) - \frac{2\lambda}{(L-v)^3}(u-v) = 0 \text{ in } \Omega.$$

Moreover, the second derivative of the left-hand side of (4.2) with respect to τ at $\tau = 0$ is

$$-6\lambda(L-v)^{-4}(u-v)^2 \geq 0,$$

which implies that $u \equiv v$ in Ω. This completes the proof. □

PROPOSITION 4.2. *For each $\lambda \in (0, \lambda_c)$, the minimal positive solution v_λ of (T_λ) is stable.*

Proof. Since $\sigma_1 > 0$, we easily see that the first eigenvalue $\sigma_{1,\lambda}(v_\lambda)$ of problem (4.1) is positive provided that λ is sufficiently small. Now we prove that $\sigma_{1,\lambda}(v_\lambda) > 0$ for $\lambda \in (0, \lambda_c)$.

We define

$$\lambda^* = \sup\{\rho : \ v_\lambda \text{ is a stable solution for } \lambda \in (0, \rho)\}.$$

It is clear that $\lambda^* \leq \lambda_c$. To show $\lambda^* = \lambda_c$, it suffices to prove that there is no regular minimal solution for (T_λ) with $\lambda > \lambda^*$. For that, suppose w is a regular minimal solution of $(T_{\lambda^*+\delta})$ with $\delta > 0$; then we would have for $\lambda \leq \lambda^*$

$$-T\Delta w + D\Delta^2 w = \frac{\lambda^* + \delta}{(L-w)^2} \geq \frac{\lambda}{(L-w)^2} \text{ in } \Omega.$$

Since for $0 < \lambda < \lambda^*$ the minimal solution v_λ is stable, it follows from Lemma 4.1 that $L > w \geq v_\lambda$. Consequently, $\overline{v} = \lim_{\lambda \nearrow \lambda^*} v_\lambda$ exists in $C^4(\Omega)$, and it is a regular solution to (T_{λ^*}). Now, from the definition of λ^* and the implicit function theorem, we necessarily have $\sigma_{1,\lambda^*}(\overline{v}) = 0$. By Lemma 4.1 again, we obtain that $w \equiv \overline{v}$ in Ω and hence $\delta = 0$. This is a contradiction. Therefore, $\lambda^* = \lambda_c$. This completes the proof. □

5. The regularity of the minimal solution of (T_λ) at $\lambda = \lambda_c$. In this section, we are concerned with the regularity of the minimal solutions of (T_λ) at $\lambda = \lambda_c$. Normally, the minimal solution v_λ at $\lambda = \lambda_c$ may have a singular set in Ω; i.e., there exists a set $\Sigma_{\lambda_c} \subset \Omega$ such that $v_{\lambda_c}(x) = L$ for $x \in \Sigma_{\lambda_c}$. But we will see that for the lower dimensional case, $v_{\lambda_c} < L$ in Ω.

By a weak solution $v \in \mathcal{H}$ of (T_λ) we mean $0 < v \le L$ in Ω and $(L-v)^{-2} \in L^1(\Omega)$ such that for any $\varphi \in \mathcal{H}$,

$$\int_\Omega [T\nabla v \cdot \nabla\varphi + D\Delta v\Delta\varphi]dx = \lambda \int_\Omega (L-v)^{-2}\varphi dx.$$

LEMMA 5.1. *If $v \in \mathcal{H}$ is a weak solution to (T_λ), then there exists $C := C(\lambda) > 0$ such that*

$$\int_\Omega \frac{dx}{(L-v)^2} \le C.$$

For $N \ge 2$, any solution v satisfying $(L-v)^{-2} \in L^p(\Omega)$ with $p = N/2$ is a classical solution.

Proof. For the first conclusion, we see that, since $v \in \mathcal{H}$ is a solution of (T_λ),

$$\int_\Omega \frac{v}{(L-v)^2}dx = \frac{1}{\lambda}\left[\int_\Omega (|\nabla v|^2 + |\Delta v|^2)dx\right] \le C. \tag{5.1}$$

On the other hand, we see that

$$\frac{v}{(L-v)^2} = \frac{L}{(L-v)^2} - \frac{1}{(L-v)}.$$

Thus, (5.1) implies that

$$\int_\Omega \frac{L}{(L-v)^2}dx = \int_\Omega \frac{1}{(L-v)}dx + \int_\Omega \frac{v}{(L-v)^2}dx. \tag{5.2}$$

By Young's inequality, we have that

$$\int_\Omega \frac{1}{(L-v)}dx \le \epsilon L \int_\Omega \frac{1}{(L-v)^2}dx + C(\epsilon, L)|\Omega|, \tag{5.3}$$

where $0 < \epsilon < 1/4$ and $C(\epsilon, L) > 0$ is a constant. Our first conclusion can be obtained from (5.1), (5.2), and (5.3).

For $N = 2$, suppose that v is a weak solution such that $\frac{1}{(L-v)^2} \in L^1(\Omega)$. Thus,

$$-D\Delta^2 v = \lambda(L-v)^{-2} - T\Delta v \in L^1(\Omega).$$

This and the Sobolev embedding imply that $\nabla^3 v \in L^q(\Omega)$ for any $1 < q < 2$. In particular, $\nabla v \in C^{2-\frac{2}{q}}(\Omega)$ for any $1 < q < 2$. This and the fact that $(L-v)^{-2} \in L^1(\Omega)$ clearly imply that $v < L$ in Ω. In fact, on the contrary, suppose that there exists $x_0 \in \Omega$ such that $v(x_0) = \max_\Omega v = L$. Then, $\nabla v(x_0) = 0$ and

$$v(x) - v(x_0) = \nabla v(\xi) \cdot (x - x_0) \text{ for } x \in \Omega \text{ near } x_0,$$

where $\xi = tx_0 + (1-t)x$ with $t \in (0,1)$. Moreover, since $\nabla v \in C^{2-\frac{2}{q}}(\Omega)$, we see that

$$|\nabla v(\xi) - \nabla v(x_0)| \le M|\xi - x_0|^{2-\frac{2}{q}} \le M|x - x_0|^{2-\frac{2}{q}},$$

and thus

$$|v(x)-v(x_0)| \le M|x-x_0|^{3-\frac{2}{q}} \text{ for } x\in\Omega \text{ near } x_0.$$

This inequality shows that

$$\infty > \int_\Omega \frac{1}{(L-v)^2}dx \ge M^{-2}\int_\Omega |x-x_0|^{-(6-\frac{4}{q})}dx = \infty,$$

which is a contradiction, which implies that we must have $\|v\|_{C(\overline{\Omega})} < L$.

For $N \ge 3$, suppose that v is a weak solution such that $\frac{1}{(L-v)^2} \in L^p(\Omega)$ with $p = \frac{N}{2}$. By the regularity of Δ^2, we see that $v \in W^{4,p}(\Omega)$. The Sobolev embedding theorem then implies that $v \in C^{1,\alpha}(\Omega)$ with $\alpha < 1$ since $4 - \frac{N}{p} = 2$. To show that v is a classical solution, it suffices to show that $v < L$ in Ω. Indeed, on the contrary, there exists $x_0 \in \Omega$ such that $v(x_0) = \max_\Omega v = L$. Then, $\nabla v(x_0) = 0$ and

$$v(x)-v(x_0) = \nabla v(\xi)\cdot(x-x_0) \text{ for } x\in\Omega \text{ near } x_0, \tag{5.4}$$

where $\xi = tx_0 + (1-t)x$ with $t \in (0,1)$. Moreover, since $v \in C^{1,1}(\Omega)$, we see that $|\nabla v(\xi) - \nabla v(x_0)| \le M|\xi - x_0| \le M|x-x_0|$. This and (5.4) imply that

$$|v(x)-v(x_0)| \le M|x-x_0|^{1+\alpha} \text{ for } x\in\Omega. \tag{5.5}$$

This inequality shows that

$$\infty > \int_\Omega \left(\frac{1}{(L-v)^2}\right)^p dx \ge M^{-2p}\int_\Omega |x-x_0|^{-2(1+\alpha)p}dx = \infty,$$

which is a contradiction, which implies that we must have $\|v\|_{C(\overline{\Omega})} < L$. This completes the proof. □

PROPOSITION 5.2. *There exists a constant $C := C(L,\lambda) > 0$ such that for each $\lambda \in (0,\lambda_c)$, the minimal solution v_λ satisfies $\|(L-v_\lambda)^{-2}\|_{L^{3/2}(\Omega)} \le C$.*

Proof. Since the minimal solutions v_λ are stable, we have

$$\int_\Omega \frac{2\lambda}{(L-v_\lambda)^3}w^2dx \le \int_\Omega [T|\nabla w|^2 + D|\Delta w|^2]dx \tag{5.6}$$

for all $0 < \lambda < \lambda_c$ and nonnegative $w \in \mathcal{H}$.

Let $w = v_\lambda$; we then have

$$\int_\Omega \frac{2\lambda}{(L-v_\lambda)^3}v_\lambda^2dx \le \int_\Omega [T|\nabla v_\lambda|^2 + D|\Delta v_\lambda|^2]dx = \int_\Omega \frac{\lambda v_\lambda}{(L-v_\lambda)^2}. \tag{5.7}$$

Since $v_\lambda < L$, this implies that

$$\int_\Omega \frac{v_\lambda^2}{(L-v_\lambda)^3}dx \le C \tag{5.8}$$

and

$$\int_\Omega \frac{L^2}{(L-v_\lambda)^3}dx \le \int_\Omega \frac{v_\lambda^2}{(L-v_\lambda)^3}dx + \int_\Omega \frac{(L-v_\lambda)^2}{(L-v_\lambda)^3}dx \le C + \int_\Omega \frac{1}{L-v_\lambda}dx. \tag{5.9}$$

Hence

$$\int_\Omega \frac{1}{(L-v_\lambda)^3}dx \le C. \tag{5.10}$$

This completes the proof. □

Now we obtain the following theorem, from which our Theorem 1.1 can be obtained.

THEOREM 5.3. *For dimension $N = 2$ or 3, there exists a constant $0 < C := C(N, L) < L$ independent of λ such that for any $0 < \lambda < \lambda_c$, the minimal solution v_λ of (T_λ) satisfies $\|v_\lambda\|_{C(\Omega)} \le C$.*

Consequently, $v_{\lambda_c} = \lim_{\lambda \nearrow \lambda_c} v_\lambda$ exists in the topology of $C^4(\Omega)$. It is the unique classical solution to (T_{λ_c}).

Proof. By Proposition 5.2 and (3.2), we see that there is $C > 0$ independent of λ such that

$$\|v_\lambda\|_{H^2(\Omega)} \le C.$$

Since the mapping $\lambda \mapsto v_\lambda$ is increasing for $\lambda \in (0, \lambda_c)$, we see that there is a function $v_{\lambda_c} \in H^2(\Omega)$ such that

$$\lim_{\lambda \nearrow \lambda_c} v_\lambda = v_{\lambda_c} \text{ weakly in } H^2(\Omega).$$

Consequently, v_{λ_c} is a weak solution of the equation (T_λ) at the critical parameter λ_c,

$$-T\Delta v_{\lambda_c} + D\Delta^2 v_{\lambda_c} = \frac{\lambda_c}{(L-v_{\lambda_c})^2} \text{ in } \Omega,$$

and in the sense of weak solutions, the critical value λ_c is attainable.

Now we show that v_{λ_c} is a classical solution. The implicit function theorem implies that the mapping $\lambda \mapsto v_\lambda$ from $(0, \lambda_c)$ to $C(\overline{\Omega})$ is continuous. Thus, we see that $\sigma_{1,\lambda_c} = 0$. (Otherwise, the implicit function theorem implies that v_λ will exist for $\lambda > \lambda_c$.) By arguments similar to those in the proof of Proposition 5.2, we see that

$$\|(L-v_{\lambda_c})^{-2}\|_{L^{3/2}(\Omega)} \le C(L).$$

Note that (5.6) holds with the inequality replaced by an equality. Then Lemma 5.1 implies that for $N = 2$ and 3, v_{λ_c} is a classical solution. Thus, there exists $C < L$ such that $\|v_{\lambda_c}\|_{C(\overline{\Omega})} \le C$. Note that $\|v_\lambda\|_{C(\overline{\Omega})} \le \|v_{\lambda_c}\|_{C(\overline{\Omega})} \le C < L$ for $\lambda \in (0, \lambda_c)$. The uniqueness of v_{λ_c} of (T_λ) at $\lambda = \lambda_c$ follows from Lemma 4.1. This completes the proof. □

6. Uniqueness of the solution of (T_λ) at $\lambda = \lambda_c$. We first note that the monotonicity with respect to λ and the uniform boundedness of the branch of the minimal solutions imply that the extremal function defined by $v_{\lambda_c} = \lim_{\lambda \nearrow \lambda_c} v_\lambda$ always exists and can always be considered as a solution for (T_{λ_c}) in a weak sense. On the other hand, if there is a $0 < C < L$ such that $\|v_\lambda\|_{C(\overline{\Omega})} \le C$ for each $\lambda < \lambda_c$, just as in the case $N = 2$ or 3, then we see from Theorem 5.3 that v_{λ_c} is the unique classical solution.

In the following, we consider only the case that v_{λ_c} is a weak solution (i.e., $v_{\lambda_c} \in W^{2,2}_{loc}(\Omega)$; note that we can obtain $v_{\lambda_c} \in \mathcal{H}$ provided that $v_{\lambda_c} \in W^{2,2}_{loc}(\Omega)$ by the moving plane argument) but with the possibility that $\|v_{\lambda_c}\|_{L^\infty(\Omega)} = L$.

THEOREM 6.1. *For $\lambda > 0$, assume that $v \in \mathcal{H}$ is a weak solution to (T_λ) such that $\|v\|_{L^\infty(\Omega)} = L$. The following assertions are equivalent:*

(i) $\sigma_{1,\lambda}(v) \geq 0$*; that is, v satisfies*

$$2\lambda \int_\Omega (L-v)^{-3}\phi^2 \leq \int_\Omega [T|\nabla\phi|^2 + D|\Delta\phi|^2]dx \quad \forall \phi \in \mathcal{H}.$$

(ii) $\lambda = \lambda_c$ *and* $v \equiv v_{\lambda_c}$ *in* Ω.

Theorem 6.1 can be easily obtained from the following proposition.

PROPOSITION 6.2. *Let v_1, v_2 be two $\mathcal{H}$-weak solutions of (T_λ) so that $\sigma_{1,\lambda}(v_i) \geq 0$ for $i = 1, 2$. Then $v_1 = v_2$ a.e. in Ω.*

Proof. For any $\theta \in [0,1]$ and $\phi \in \mathcal{H}$, $\phi \geq 0$, we have that

$$\begin{aligned} I_{\theta,\phi} :&= T\int_\Omega \nabla(\theta v_1 + (1-\theta)v_2)\nabla\phi dx + D\int_\Omega \Delta(\theta v_1 + (1-\theta)v_2)\Delta\phi dx \\ &\quad - \lambda\int_\Omega [L - (\theta v_1 + (1-\theta)v_2)]^{-2}\phi dx \\ &= \lambda\int_\Omega \Big[[\theta(L-v_1)^{-2} + (1-\theta)(L-v_2)^{-2}] - [(L-(\theta v_1 + (1-\theta)v_2))^{-2}]\Big]dx \\ &\geq 0 \end{aligned}$$

due to the convexity of $(L-s)^{-2}$ with respect to $s \in (0, L)$. Since $I_{0,\phi} = I_{1,\phi} = 0$, the derivative of $I_{\theta,\phi}$ at $\theta = 0, 1$ provides

$$\begin{aligned} \int_\Omega [T\nabla(v_1-v_2)\nabla\phi + D\Delta(v_1-v_2)\Delta\phi] - 2\lambda\int_\Omega (L-v_2)^{-3}(v_1-v_2)\phi \geq 0, \\ \int_\Omega [T\nabla(v_1-v_2)\nabla\phi + D\Delta(v_1-v_2)\Delta\phi] - 2\lambda\int_\Omega (L-v_1)^{-3}(v_1-v_2)\phi \leq 0 \end{aligned}$$

for any $\phi \in \mathcal{H}$ with $\phi \geq 0$. Testing the first inequality on $\phi = (v_1-v_2)^-$ and the second one on $(v_1-v_2)^+$, we obtain that

$$\begin{aligned} \int_\Omega \Big[T|\nabla(v_1-v_2)^-|^2 + D|\Delta(v_1-v_2)^-|^2\Big] - 2\lambda\int_\Omega (L-v_2)^{-3}((v_1-v_2)^-)^2 \leq 0, \\ \int_\Omega \Big[T|\nabla(v_1-v_2)^+|^2 + D|\Delta(v_1-v_2)^+|^2\Big] - 2\lambda\int_\Omega (L-v_1)^{-3}((v_1-v_2)^+)^2 \leq 0. \end{aligned}$$

Since $\sigma_{1,\lambda}(v_1) \geq 0$, we have the following:

(1) If $\sigma_{1,\lambda}(v_1) > 0$, then $v_1 \leq v_2$ a.e. in Ω.

(2) If $\sigma_{1,\lambda}(v_1) = 0$, then

$$\int_\Omega \Big[T\nabla(v_1-v_2)\nabla\overline{\varphi} + D\Delta(v_1-v_2)\Delta\overline{\varphi}\Big] - 2\lambda\int_\Omega (L-v_1)^{-3}(v_1-v_2)\overline{\varphi} = 0,$$

where $\overline{\varphi} = (v_1-v_2)^+$. Since $I_{\theta,\overline{\varphi}} \geq 0$ for any $\theta \in [0,1]$ and $I_{1,\overline{\varphi}} = \partial I_{1,\overline{\varphi}} = 0$, we get that $\partial^2_{\theta\theta} I_{1,\overline{\varphi}} = -\int_\Omega \frac{6\lambda}{(L-v_1)^4}((v_1-v_2)^+)^3 \geq 0$. Thus, $(v_1-v_2)^+ = 0$ a.e. in Ω. Hence, $v_1 \leq v_2$ a.e. in Ω. The same argument applies to prove the reversed inequality, and the proof of the proposition is complete. ∎

7. Structure of solutions of (T_λ) in the 2D case. In this section we obtain the structure of positive solutions of (T_λ) in the 2D case. The main theorem of this section is the following theorem. Our Theorem 1.2 can be obtained from this theorem and Theorem 8.2.

THEOREM 7.1. *Let Ω be a convex smooth domain in $\boldsymbol{R}^2$. For $\lambda \in (0, \lambda_c]$, any solution of the problem (T_λ) is regular and the following hold.*

(i) *For $0 < \lambda < \lambda_c$, problem (T_λ) admits two solutions: the minimal solution and a mountain-pass solution.*

(ii) *For $\lambda = \lambda_c$, problem (T_λ) admits a unique regular solution.*

(iii) *For $\lambda > \lambda_c$, problem (T_λ) admits no regular solution.*

To prove this theorem, we first show the following lemma.

LEMMA 7.2. *For any fixed $\lambda > 0$, if $v_\lambda \in \mathcal{H}$ is a positive solution of (T_λ), then there exists $0 < \tau_\lambda < L$ such that $v_\lambda \le L - \tau_\lambda$ in Ω. This also implies that v_λ is regular.*

Proof. The embedding theorem implies that $v_\lambda \in C^\alpha(\overline{\Omega})$ for any $0 < \alpha < 1$, and thus the moving plane arguments as in the proof of Theorem 3.1 imply that if $v_\lambda(x_\lambda) = \max_\Omega v_\lambda$, then $x_\lambda \in \Omega_0$, where $\Omega_0 \subset\subset \Omega$. Moreover, by Theorem 3.1, we have

$$\int_\Omega (L - v_\lambda)^{-2} \le \frac{C}{\lambda} \tag{7.1}$$

and

$$\int_\Omega [T|\nabla v_\lambda|^2 + D(\Delta v_\lambda)^2]dx \le C. \tag{7.2}$$

Suppose that there is $\lambda_0 > 0$ and sequences $\{\lambda_i\}$ and $\{v_i\}$ with $\max_\Omega v_i = L - \epsilon_i$ such that $\lambda_i \to \lambda_0$, $\epsilon_i \to 0$ as $i \to \infty$. Making the transformation $w_i = L - v_i$, we see that w_i with $\min_\Omega w_i = \epsilon_i$ satisfies the problem

$$T\Delta w_i - D\Delta^2 w_i = \lambda_i w_i^{-2} \text{ in } \Omega, \quad w_i = L, \Delta w_i = 0 \text{ on } \partial\Omega.$$

Define $z_i = \Delta w_i$; then

$$-D\Delta z_i + T z_i = \lambda_i w_i^{-2} \text{ in } \Omega, \ \ z_i = 0 \text{ on } \partial\Omega. \tag{7.3}$$

It is known from (7.3) that $z_i(x) = \lambda_i \int_\Omega G_{T,D}(x,y) w_i^{-2}(y)dy$, where $G_{T,D}(x,y)$ is the Green's function of the operator $-D\Delta + TId$. Let $w_i(x_i) = \min_\Omega w_i$. Then $x_i \in \Omega_0 \subset\subset \Omega$. Setting $\tilde{w}_i(y) = \frac{w_i}{\epsilon_i}$ and $y = \lambda_i^{1/4}\epsilon_i^{-3/4}(x - x_i)$, we see that $\tilde{w}_i$ with $\tilde{w}_i(0) = \min_{\Omega_i} \tilde{w}_i = 1$ and $\tilde{w}_i$ satisfies the problem

$$\lambda_i^{-1/2}\epsilon_i^{3/2} T\Delta_y \tilde{w}_i - D\Delta_y^2 \tilde{w}_i = \tilde{w}_i^{-2} \text{ in } \Omega_i, \quad \tilde{w}_i = \frac{L_i}{\epsilon_i}, \ \ \Delta_y \tilde{w}_i = 0 \text{ on } \partial\Omega_i, \tag{7.4}$$

where $\Omega_i = \{y = \lambda_i^{1/4}\epsilon_i^{-3/4}(x - x_i) : \ x \in \Omega\}$. On the other hand,

$$\Delta_y \tilde{w}_i = \lambda_i^{-1/2}\epsilon_i^{1/2}\Delta_x w_i = \lambda_i^{1/2}\epsilon_i^{1/2}\int_\Omega G_{T,D}(x,\xi) w_i^{-2}(\xi)d\xi.$$

Note that $N = 2$ and $w_i \geq \epsilon_i$ in Ω. The Hölder inequality implies that

$$\begin{aligned}|\Delta_y \tilde{w}_i| &\leq C\epsilon_i^{1/2}\left(\int_\Omega [G_{T,D}(x,\xi)]^p d\xi\right)^{1/p}\left(\int_\Omega w_i^{-2} w_i^{2-2q}(\xi)d\xi\right)^{1/q}\\ &\leq C\epsilon_i^{1/2}\epsilon_i^{-2/p}\left(\int_\Omega [G_{T,D}(x,\xi)]^p d\xi\right)^{1/p}\left(\int_\Omega w_i^{-2}(\xi)d\xi\right)^{1/q}\\ &\leq C\epsilon_i^{\frac{1}{2}-\frac{2}{p}},\end{aligned}$$

where we have applied (7.1). Choosing p sufficiently large, we see that

$$|\Delta_y \tilde{w}_i(y)| \to 0 \ \text{ for } y \in \Omega_i \text{ a.e. as } i \to \infty. \tag{7.5}$$

On the other hand, it follows from (7.4), (7.5), and the regularity of the operator $T\Delta - D\Delta^2$ that $\tilde{w}_i \to W$ in $C^4_{loc}(\boldsymbol{R}^2)$ as $i \to \infty$, where W with $W(0) = 1$ and $W \geq 1$ in $\boldsymbol{R}^2$ satisfies the equation

$$-D\Delta^2 W = W^{-2} \ \text{ in } \boldsymbol{R}^2, \quad W(0) = 1. \tag{7.6}$$

Meanwhile, (7.5) implies that $\Delta W = 0$ in $\boldsymbol{R}^2$. This contradicts (7.6) and completes the proof of this lemma. □

In the remainder of this section, we establish the existence of the second solution. Note that in the energy functional (1.1), the integral $\int_\Omega \frac{1}{L+u(x)}dx$ is not well defined for $u \in H^2(\Omega)$. Therefore, we do not have a good energy functional to work with. Our idea is to modify the nonlinearity so that the mountain-pass lemma works, and then show that the resulting solution has no singularity.

We first modify the nonlinearity. Since the nonlinearity $g(v) = \frac{1}{(L-v)^2}$ is singular at $v = L$, we need to consider a regularized C^1 nonlinearity $g_\epsilon(v)$, $0 < \epsilon < L$, of the following form:

$$g_\epsilon(v) = \begin{cases} \frac{1}{(L-v)^2}, & v \leq L-\epsilon,\\ \frac{1}{\epsilon^2} - \frac{(L-\epsilon)}{\epsilon^3} + \frac{1}{\epsilon^3(L-\epsilon)}v^2, & v > L-\epsilon.\end{cases}$$

For $\lambda \in (0, \lambda_c)$, we study the regularized semilinear elliptic problem:

$$-T\Delta v + D\Delta^2 v = \lambda g_\epsilon(v) \ \text{ in } \Omega, \quad v = \Delta v = 0 \ \text{ on } \partial\Omega. \tag{7.7}$$

From a variational viewpoint, the action functional associated to (7.7) is

$$J_{\epsilon,\lambda}(v) = \frac{1}{2}\int_\Omega [T|\nabla v|^2 + D(\Delta v)^2]dx - \lambda\int_\Omega G_\epsilon(v)dx, \quad v \in \mathcal{H},$$

where $G_\epsilon(v) = \int_{-\infty}^v g_\epsilon(s)ds$.

Fix now $0 < \epsilon < \tau_\lambda/4$, where τ_λ is as given in Lemma 7.2. The minimal solution $\underline{v}_\lambda$ of (T_λ) is still a solution of (7.7) so that $\sigma_{1,\lambda}(\underline{v}_\lambda) > 0$. In order to motivate the choice of $g_\epsilon(v)$, we briefly sketch the proof of Theorem 7.1. First, we prove that $\underline{v}_\lambda$ is a local minimum for $J_{\epsilon,\lambda}(v)$. Then, by the well-known mountain-pass theorem, we show the existence of a second solution $V_{\epsilon,\lambda}$ for (7.7). (A similar idea has been used in [6].) The subcritical growth

$$0 \leq g_\epsilon(v) \leq C_\epsilon(1 + |v|^2) \tag{7.8}$$

and the inequality

$$3G_\epsilon(v) \le v g_\epsilon(v) \ \text{ for } v \ge L-\theta, \tag{7.9}$$

for some sufficiently small $\theta > 10\epsilon$ independent of ϵ, $C_\epsilon > 0$, will yield that $J_{\epsilon,\lambda}$ satisfies the Palais–Smale condition.

In order to complete the details of the proof of Theorem 7.1, we first need to show the following lemma.

LEMMA 7.3. *The minimal solution $\underline{v}_\lambda$ of (T_λ) is a local minimum of $J_{\epsilon,\lambda}$ on $\mathcal{H}$.*

Proof. Since $\mathcal{H} \hookrightarrow C^\alpha(\overline{\Omega})$ for any $0 < \alpha < 1$, we need only to show that $\underline{v}_\lambda$ is a local minimum of $J_{\epsilon,\lambda}$ in $C^\alpha(\overline{\Omega})$ for some $0 < \alpha < 1$. Indeed, since $\sigma_{1,\lambda}(\underline{v}_\lambda) > 0$, we have the following inequality:

$$\int_\Omega [T|\nabla\varphi|^2 + D(\Delta\varphi)^2]dx - 2\lambda \int_\Omega \frac{1}{(L-\underline{v}_\lambda)^3}\varphi^2 dx \ge \sigma_{1,\lambda}\int_\Omega \varphi^2 dx \tag{7.10}$$

for any $\varphi \in \mathcal{H}$, since $\underline{v}_\lambda \le L - \tau_\lambda < L - \epsilon$ (see Lemma 7.2). Now, take any $\varphi \in \mathcal{H} \cap C^\alpha(\overline{\Omega})$ such that $\|\varphi\|_{C^\alpha} \le \delta_\lambda$. Since $\underline{v}_\lambda \le L - \tau_\lambda$, if $\delta_\lambda \le \epsilon$, then $\underline{v}_\lambda + \varphi \le L - \epsilon$, and we have that

$$\begin{aligned}
&J_{\epsilon,\lambda}(\underline{v}_\lambda + \varphi) - J_{\epsilon,\lambda}(\underline{v}_\lambda) \\
&\quad = \frac{1}{2}\int_\Omega [T|\nabla\varphi|^2 + D(\Delta\varphi)^2]dx + \int_\Omega [T\nabla\underline{v}_\lambda\cdot\nabla\varphi + D\Delta\underline{v}_\lambda\Delta\varphi]dx \\
&\qquad\quad - \lambda\int_\Omega \left(\frac{1}{L-\underline{v}_\lambda-\varphi} - \frac{1}{L-\underline{v}_\lambda}\right) \\
&\quad \ge \frac{\sigma_{1,\lambda}}{2}\int_\Omega \varphi^2 - \lambda\int_\Omega \left(\frac{1}{L-\underline{v}_\lambda-\varphi} - \frac{1}{L-\underline{v}_\lambda} - \frac{\varphi}{(L-\underline{v}_\lambda)^2} - \frac{\varphi^2}{(L-\underline{v}_\lambda)^3}\right),
\end{aligned}$$

where we have used (7.10). Since now

$$\left|\frac{1}{L-\underline{v}_\lambda-\varphi} - \frac{1}{L-\underline{v}_\lambda} - \frac{\varphi}{(L-\underline{v}_\lambda)^2} - \frac{\varphi^2}{(L-\underline{v}_\lambda)^3}\right| \le C|\varphi|^3$$

for some $C > 0$, we have that

$$J_{\epsilon,\lambda}(\underline{v}_\lambda + \varphi) - J_{\epsilon,\lambda}(\underline{v}_\lambda) \ge \left(\frac{\sigma_{1,\lambda}}{2} - C\lambda\delta_\lambda\right)\int_\Omega \varphi^2 > 0$$

provided δ_λ is small enough. This proves that $\underline{v}_\lambda$ is a local minimum of $J_{\epsilon,\lambda}$ in the C^α topology and completes the proof of this lemma. ∎

Fix some ball $B_{2r} \subset \Omega$ of radius $2r$, $r > 0$. Take a cut-off function χ so that $\chi = 1$ on B_r and $\chi = 0$ outside B_{2r}. Let $w_\epsilon = (L-\epsilon)\chi \in \mathcal{H}$. We have that

$$J_{\epsilon,\lambda}(w_\epsilon) \le \frac{(L-\epsilon)^2}{2}\int_\Omega [T|\nabla\chi|^2 + D(\Delta\chi)^2]dx - \frac{\lambda}{\epsilon^2}|B_r| \to -\infty$$

as $\epsilon \to 0$. Moreover, we can find for $\epsilon > 0$ small the inequality

$$J_{\epsilon,\lambda}(w_\epsilon) < J_{\epsilon,\lambda}(\underline{v}_\lambda). \tag{7.11}$$

Now fix $\epsilon > 0$ small enough in order that (7.11) holds, and define

$$c_{\epsilon,\lambda} = \inf_{\gamma\in\Gamma}\max_{v\in\gamma} J_{\epsilon,\lambda}(v),$$

where $\Gamma = \{\gamma : [0,1] \to \mathcal{H}; \gamma \text{ continuous and } \gamma(0) = \underline{v}_\lambda,\ \gamma(1) = w_\epsilon\}$. We can then apply the mountain-pass theorem to get a solution $V_{\epsilon,\lambda}$ of (7.7), provided the Palais–Smale condition holds at level $c_{\epsilon,\lambda}$. The embedding theorem and the maximum principle imply that $V_{\epsilon,\lambda} > 0$ in Ω.

LEMMA 7.4. *Assume that $\{v_n\} \subset \mathcal{H}$ satisfies*

$$J_{\epsilon,\lambda_n}(v_n) \le C, \quad J'_{\epsilon,\lambda_n}(v_n) \to 0 \ \text{ in } \mathcal{H}' \tag{7.12}$$

for $\lambda_n \to \lambda > 0$. Then the sequence $(v_n)_n$ is uniformly bounded in $\mathcal{H}$ and therefore admits a convergent subsequence in $\mathcal{H}$.

Proof. By (7.12) we have that

$$\int_\Omega [T|\nabla v_n|^2 + D(\Delta v_n)^2]dx = \lambda_n \int_\Omega g_\epsilon(v_n)v_n dx + o(\|v_n\|_{\mathcal{H}})$$

as $n \to +\infty$, and then

$$\begin{aligned}
C &\ge \frac{1}{2}\int_\Omega [T|\nabla v_n|^2 + D(\Delta v_n)^2]dx - \lambda_n \int_\Omega G_\epsilon(v_n)dx \\
&= \frac{1}{6}\int_\Omega [T|\nabla v_n|^2 + D(\Delta v_n)^2]dx + \lambda_n \int_\Omega \left(\frac{1}{3}v_n g_\epsilon(v_n) - G_\epsilon(v_n)\right)dx + o(\|v_n\|_{\mathcal{H}}) \\
&\ge \frac{1}{6}\int_\Omega [T|\nabla v_n|^2 + D(\Delta v_n)^2]dx + \lambda_n \int_{\{v_n \ge L-\theta\}} \left(\frac{1}{3}v_n g_\epsilon(v_n) - G_\epsilon(v_n)\right)dx \\
&\quad + o(\|v_n\|_{\mathcal{H}}) - C(\theta) \\
&\ge \frac{1}{6}\int_\Omega [T|\nabla v_n|^2 + D(\Delta v_n)^2]dx + o(\|v_n\|_{\mathcal{H}}) - C(\theta)
\end{aligned}$$

in view of (7.9), where $C(\theta) > 0$ depends on θ but is independent of ϵ. Hence, $\sup_{n\in\mathbf{N}} \|v_n\|_{\mathcal{H}} < +\infty$.

The compactness of the embedding $\mathcal{H} \hookrightarrow C^\alpha(\overline{\Omega})$ for any $0 < \alpha < 1$ provides that, up to a subsequence, $v_n \to v$ weakly in $\mathcal{H}$ and strongly in $C^\alpha(\overline{\Omega})$ for some $0 < \alpha < 1$ and some $v \in \mathcal{H}$. By (7.12) we get that $\int_\Omega [T|\nabla v|^2 + D(\Delta v)^2]dx = \lambda \int_\Omega g_\epsilon(v)v$, and then

$$\begin{aligned}
&\int_\Omega [T|\nabla(v_n - v)|^2 + D(\Delta(v_n - v))^2]dx \\
&\quad = T\left[\int_\Omega |\nabla v_n|^2 - \int_\Omega |\nabla v|^2\right] + D\left[\int_\Omega (\Delta v_n)^2 - \int_\Omega (\Delta v)^2\right] + o(1) \\
&\quad = \lambda_n \int_\Omega g_\epsilon(v_n)v_n - \lambda \int_\Omega g_\epsilon(v)v + o(1) \to 0
\end{aligned}$$

as $n \to +\infty$. This completes the proof. □

Proof of Theorem 7.1. We need only to show (i), and it is enough to show that for any fixed $\lambda > 0$, the mountain-pass solution $V_{\epsilon,\lambda}$ satisfies $V_{\epsilon,\lambda} \le L - \epsilon$ in Ω.

Since $V_{\epsilon,\lambda} \in \mathcal{H}$, by the same argument as in Theorem 3.1, we easily see that

$$\int_\Omega g_\epsilon(V_{\epsilon,\lambda})dx \le C/\lambda, \tag{7.13}$$

where C is independent of ϵ. In fact, we see that

$$J_{\epsilon,\lambda}(V_{\epsilon,\lambda}) \le \max_{v\in\gamma_0} J_{\epsilon,\lambda}(v),$$

where $\gamma_0 : [0,1] \to \mathcal{H}$; $\gamma_0(v) = t\underline{v}_\lambda + (1-t)w_\epsilon$ for $t \in [0,1]$. Thus,

$$J_{\epsilon,\lambda}(V_{\epsilon,\lambda}) \le C,$$

where $C > 0$ is independent of ϵ. On the other hand, we see that

$$\begin{aligned} C &\ge \frac{1}{2}\int_\Omega [T|\nabla V_{\epsilon,\lambda}|^2 + D(\Delta V_{\epsilon,\lambda})^2]dx - \lambda_n \int_\Omega G_\epsilon(V_{\epsilon,\lambda})dx \\ &= \frac{1}{6}\|V_{\epsilon,\lambda}\|^2_{\mathcal{H}} + \lambda\int_\Omega \left(\frac{1}{3}V_{\epsilon,\lambda}g_\epsilon(V_{\epsilon,\lambda}) - G_\epsilon(V_{\epsilon,\lambda})\right)dx \\ &\ge \frac{1}{6}\int_\Omega \|V_{\epsilon,\lambda}\|^2_{\mathcal{H}} + \lambda\int_{\{V_{\epsilon,\lambda}\ge L-\theta\}} \left(\frac{1}{3}V_{\epsilon,\lambda}g_\epsilon(V_{\epsilon,\lambda}) - G_\epsilon(V_{\epsilon,\lambda})\right)dx - C(\theta) \\ &\ge \frac{1}{6}\|V_{\epsilon,\lambda}\|^2_{\mathcal{H}}dx - C(\theta). \end{aligned}$$

Thus,

$$\|V_{\epsilon,\lambda}\|_{\mathcal{H}} \le C,$$

where $C > 0$ is independent of ϵ. The embedding $\mathcal{H} \hookrightarrow C^0(\overline{\Omega})$ implies $V_{\epsilon,\lambda} \le C$ in Ω. By the moving plane argument, as in the proof of Theorem 3.1, we have that $V_{\epsilon,\lambda} \le L - \theta$ in Ω_δ, where Ω_δ is as given in the proof of Theorem 3.1 and θ is as given by (7.9). This implies that (7.13) holds.

Let $W_\epsilon = \Delta V_{\epsilon,\lambda}$. Then W_ϵ satisfies the equation

$$-TW_\epsilon + D\Delta W_\epsilon = \lambda g_\epsilon(V_{\epsilon,\lambda}) \in L^1(\Omega).$$

Since $N = 2$, the Brezis–Merle inequality [4] implies that

$$\int_\Omega |W_\epsilon|^q dx \le C \quad \forall q > 1,$$

where C is independent of ϵ. This also yields that

$$\|V_{\epsilon,\lambda}\|_{W^{2,q}(\Omega)} \le C.$$

By choosing $q > 3$ sufficiently large, we see from the embedding $W^{2,q}_0(\Omega) \hookrightarrow C^{1+\frac{1}{2}}(\Omega)$ that $V_{\epsilon,\lambda} \le C$ in Ω.

Now we show that $V_{\epsilon,\lambda} < L$ in Ω for ϵ sufficiently small. On the contrary, we suppose that there is a sequence $\{\epsilon_i\}$ with $\epsilon_i \to 0$ as $i \to \infty$ such that $\max_\Omega V_{\epsilon_i,\lambda} \ge L$. Denote $V_{\epsilon_i,\lambda}(x_i) = \max_\Omega V_{\epsilon_i,\lambda}$. By arguments similar to those in the proof of Lemma 5.1, we see that

$$V_{\epsilon_i,\lambda}(x_i) - V_{\epsilon_i,\lambda}(x) \le C|x - x_i|^{3/2}.$$

Thus,

$$V_{\epsilon_i,\lambda}(x) \ge V_{\epsilon_i,\lambda}(x_i) - C|x - x_i|^{3/2} > L - \epsilon_i$$

provided that $|x - x_i| < (\epsilon_i/C)^{2/3}$. But

$$C \ge \int_\Omega g_{\epsilon_i}(V_{\epsilon_i,\lambda})dx \ge \epsilon_i^{-2}\int_{\{|x-x_i|\le(\epsilon_i/C)^{2/3}\}} dx = C\epsilon_i^{-2/3} \to \infty$$

as $i \to \infty$. This is a contradiction.

Now we claim that there exists $\delta > 0$ independent of ϵ such that

$$V_{\epsilon,\lambda} \le L - \delta \ \text{ in } \Omega$$

for ϵ sufficiently small. On the contrary, there are sequences $\{\epsilon_i\}$ and $\{V_i\} \equiv \{V_{\epsilon_i,\lambda}\}$ with $\epsilon_i \to 0$ as $i \to \infty$ such that $\max_\Omega V_i = L - \xi_i$ and $\xi_i \to 0$ as $i \to \infty$. Set $Z_i = L - V_i$. Then $Z_i(x_i) := \min_\Omega Z_i = \xi_i$ and Z_i satisfies

$$T\Delta Z_i - D\Delta^2 Z_i = \lambda h_i(Z_i) \ \text{ in } \Omega, \quad Z_i = T, \ \ \Delta Z_i = 0 \ \text{ on } \partial\Omega,$$

where

$$h_i(Z_i) = \begin{cases} \frac{1}{Z_i^2}, & Z_i \ge \epsilon_i, \\ \frac{1}{\epsilon_i^2} + \frac{2(\epsilon_i - Z_i)}{\epsilon_i^3} + \frac{(\epsilon_i - Z_i)^2}{\epsilon_i^3(L-\epsilon_i)}, & Z_i < \epsilon_i. \end{cases}$$

Making the transformations $\tilde{Z}_i(y) = Z_i/\xi_i$ and $y = \xi_i^{-3/4}(x - x_i)$, we see that $\tilde{Z}_i(0) = \min_\Omega \tilde{Z}_i = 1$ and $\tilde{Z}_i$ satisfies the problem

$$\xi_i^{3/2}\Delta_y \tilde{Z}_i - D\Delta_y^2 \tilde{Z}_i = \lambda \tilde{h}_i(\tilde{Z}_i) \ \text{ in } \tilde{\Omega}_i, \quad \tilde{Z}_i = T/\xi_i, \ \ \Delta_y \tilde{Z}_i = 0 \ \text{ on } \partial\tilde{\Omega}_i, \tag{7.14}$$

where $\tilde{\Omega}_i = \{y = \xi_i^{-3/4}(x - x_i) : \ x \in \Omega\}$ and

$$\tilde{h}_i(\tilde{Z}_i) = \begin{cases} \frac{1}{\tilde{Z}_i^2}, & \tilde{Z}_i \ge \frac{\epsilon_i}{\xi_i}, \\ 3\Big(\frac{\xi_i}{\epsilon_i}\Big)^2 - 2\Big(\frac{\xi_i}{\epsilon_i}\Big)^3 \tilde{Z}_i + \frac{\xi_i^2}{\epsilon_i(L-\epsilon_i)} - 2\Big(\frac{\xi_i^3}{\epsilon_i^2(L-\epsilon_i)}\Big)\tilde{Z}_i + \Big(\frac{\xi_i^4}{\epsilon_i^3(L-\epsilon_i)}\Big)\tilde{Z}_i^2, & \tilde{Z}_i < \frac{\epsilon_i}{\xi_i}. \end{cases}$$

We consider two cases for $\{\frac{\epsilon_i}{\xi_i}\}$ (we can choose subsequences if necessary):

(i) There is $0 < A < \infty$ such that $\frac{\epsilon_i}{\xi_i} \le A$ for all i,

(ii) $\frac{\epsilon_i}{\xi_i} \to \infty$ as $i \to \infty$.

For the first case, we have that there is $0 \le A_1 \le A$ such that $\lim_{i\to\infty} \frac{\epsilon_i}{\xi_i} = A_1$. If $A_1 < 1$, since $\tilde{Z}_i \ge 1$, we have that $\tilde{h}_i = \tilde{Z}_i^{-2} \le 1$ in $\tilde{\Omega}_i$ for i sufficiently large. If $1 \le A_1 \le A$, we also have that $\tilde{h}_i \le C$ in $\tilde{\Omega}_i$ for i sufficiently large, where $0 < C < \infty$ is independent of i. Moreover, for $q > 3$,

$$\begin{aligned} \int_{\tilde{\Omega}_i} |\Delta_y \tilde{Z}_i|^q dy &= \xi_i^{q/2} \int_{\tilde{\Omega}_i} |\Delta_x Z_i|^q dy \\ &= \xi^{\frac{q-3}{2}} \int_\Omega |\Delta_x Z_i|^q dx \to 0 \end{aligned}$$

as $i \to \infty$. Thus, the regularity of Δ^2 implies that $\tilde{Z}_i \to \tilde{Z}$ in $C^3_{loc}(\boldsymbol{R}^2)$ with $\tilde{Z}(0) = \min_{\mathbf{R}^2} \tilde{Z} = 1$, and $\tilde{Z}$ satisfies the equation

$$-D\Delta^2 \tilde{Z} = \lambda \tilde{Z}^{-2} \ \text{ in } \boldsymbol{R}^2$$

provided $A_1 \le 1$ and the equation

$$D\Delta^2 \tilde{Z} = \lambda \tilde{h}(\tilde{Z}) \ \text{ in } \boldsymbol{R}^2$$

provided $1 < A_1 \le A$, where

$$\tilde{h}(\tilde{Z}) = \begin{cases} \tilde{Z}^2, & \tilde{Z} \ge A_1, \\ \frac{3}{A_1^2} - \frac{2}{A_1^3}\tilde{Z}, & \tilde{Z} < A_1. \end{cases}$$

Moreover, for any large ball B_R of $\boldsymbol{R}^2$, $\int_{B_R} |\Delta \tilde{Z}|^q(y)dy = 0$. This is impossible.

For the second case, we see that $\xi_i = o(\epsilon_i)$ for i sufficiently large. Thus, $Z_i(x_i) = \xi_i = o(\epsilon_i)$. Noting that $\int_\Omega |\Delta Z_i|^q dx \le C$, we see that $Z_i \in W_0^{2,q}(\Omega)$. The embedding $W_0^{2,q}(\Omega) \hookrightarrow C^{1+\frac{1}{2}}(\Omega)$ gives

$$|Z_i(x)| \le Z_i(x_i) + C|x - x_i|^{3/2} < \epsilon_i$$

provided $|x - x_i| \le (\frac{\epsilon_i}{2C})^{2/3}$. Thus

$$C \ge \int_\Omega h_i(Z_i)dx \ge \frac{1}{\epsilon_i^2} \int_{Z_i < \epsilon_i} dx \ge C\epsilon_i^{-2/3} \to \infty$$

as $i \to \infty$. This is also a contradiction. Therefore,

$$V_{\epsilon,\lambda} \le L - \delta \ \text{ in } \Omega,$$

where $\delta > 0$ is independent of ϵ. This also implies that $V_{\epsilon,\lambda}$ is a solution of (T_λ). This completes the proof of (i) of Theorem 7.1. □

8. The asymptotic behavior of the mountain-pass solution as $\lambda \to 0$. In this section we will study the asymptotic behavior of the mountain-pass solution V_λ obtained in Theorem 7.1 as $\lambda \to 0$.

LEMMA 8.1.

$$\sigma_{1,\lambda}(V_\lambda) < 0 \ \textit{for } 0 < \lambda < \lambda_c.$$

Proof. Let $\underline{v}_\lambda$ be the minimal solution of (T_λ) so that $V_\lambda \ge \underline{v}_\lambda$. If the linearization around V_λ had nonnegative first eigenvalue, then Lemma 4.1 would also yield $V_\lambda \le \underline{v}_\lambda$ so that $\overline{v}_\lambda$ and V_λ would necessarily coincide, which would be a contradiction. □

THEOREM 8.2.

$$\max_\Omega V_\lambda \to L \ \textit{ as } \lambda \to 0. \tag{8.1}$$

Moreover,

$$\lim_{\lambda \to 0^+} \frac{[\min_\Omega (L - V_\lambda)]^3}{\lambda} = 0. \tag{8.2}$$

Proof. Suppose that there are sequences $\{\lambda_i\}$ and $\{V_i\} \equiv \{V_{\lambda_i}\}$ such that $\lambda_i \to 0$ as $i \to \infty$ and $\max_\Omega V_i \le L - \delta$, where $0 < \delta < L$ is independent of i. Then it follows from the equation of V_i that $V_i \to 0$ in $C^0(\overline{\Omega})$ as $i \to \infty$ (we can choose subsequences if necessary). This contradicts the fact that $\sigma_{1,\lambda_i}(V_i) < 0$ for all i. Thus, (8.1) holds.

By Theorem 3.1, we see that

$$\lambda \int_\Omega (L - V_\lambda)^{-2} dx + \int_\Omega |\Delta V_\lambda|^2 \le C$$

for any λ sufficiently small, where C is independent of λ. Since ΔV_λ satisfies

$$-D\Delta(\Delta V_\lambda) = D\Delta V_\lambda + \frac{\lambda}{(L - V_\lambda)^2} \in L^1(\Omega), \quad \Delta V_\lambda = 0 \text{ on } \partial\Omega,$$

by the Brezis–Merle inequality [4], we have, for any $q > 1$,

$$\int_\Omega |\Delta V_\lambda|^q dx \le C \tag{8.3}$$

for any λ sufficiently small.

Let $V_\lambda(x_\lambda) = \max_\Omega V_\lambda$. Setting $W_\lambda = L - V_\lambda$, we see that $\xi_\lambda := W_\lambda(x_\lambda) = \min_\Omega W_\lambda$ and $\xi_\lambda \to 0$ as $\lambda \to 0$. Now we claim that

$$\lim_{\lambda\to 0} \frac{\xi_\lambda^3}{\lambda} = 0. \tag{8.4}$$

Suppose not; there are sequences $\{\lambda_i\}$ and $\{\xi_i\}$ with $\lambda_i \to 0$ as $i \to \infty$ such that $\frac{\xi_i^3}{\lambda_i} \to C > 0$ or $\frac{\xi_i^3}{\lambda_i} \to \infty$ as $i \to \infty$.

We first consider the case that $\frac{\xi_i^3}{\lambda_i} \to \infty$ as $i \to \infty$. Then, defining $\hat{W}_i = W_i/\xi_i$, we see that $\hat{W}_i$ satisfies the problem

$$T\Delta\hat{W}_i - D\Delta^2\hat{W}_i = \frac{\lambda_i}{\xi_i^3}\hat{W}_i^{-2} \text{ in } \Omega, \quad \hat{W}_i = L/\xi_i, \ \Delta W_i = 0 \text{ on } \partial\Omega.$$

Since $\hat{W}_i \geq 1$, we see that $\hat{W}_i \to \hat{W}$ in $C^3_{loc}(\Omega)$ as $i \to \infty$ and $\hat{W}$ with $\hat{W}(0) = \min_\Omega \hat{W} = 1$ satisfies the equation

$$T\Delta\hat{W} - D\Delta^2\hat{W} = 0 \text{ in } \Omega, \quad \hat{W} = \infty, \ \Delta\hat{W} = 0 \text{ on } \partial\Omega. \tag{8.5}$$

Setting $Z = \Delta\hat{W}$, we see from (8.5) that

$$TZ - D\Delta Z = 0 \text{ in } \Omega, \quad Z = 0 \text{ on } \partial\Omega.$$

The strong maximum principle then implies that $Z \equiv 0$ in Ω and hence $\Delta\hat{W} \equiv 0$ in Ω. The maximum principle then implies that $\hat{W} \equiv 1$ in Ω, which is a contradiction.

Now we consider the case that $\lim_{i\to\infty} \frac{\xi_i^3}{\lambda_i} \to C > 0$. Defining $\hat{W}_i = W_i/\xi_i$ again, we see that $\hat{W}_i$ satisfies the problem

$$T\Delta\hat{W}_i - D\Delta^2\hat{W}_i = \frac{\lambda_i}{\xi_i^3}\hat{W}_i^{-2} \text{ in } \Omega, \quad \hat{W}_i = L/\xi_i, \ \Delta\hat{W}_i = 0 \text{ on } \partial\Omega. \tag{8.6}$$

Setting $\hat{Z}_i = \Delta\hat{W}_i$, we see that $\hat{Z}_i$ satisfies the problem

$$T\hat{Z}_i - D\Delta\hat{Z}_i = \frac{\lambda_i}{\xi_i^3}\hat{W}_i^{-2} \text{ in } \Omega, \quad \hat{Z}_i = 0 \text{ on } \partial\Omega. \tag{8.7}$$

Therefore,

$$\hat{Z}_i = \frac{\lambda_i}{\xi_i^3}\int_\Omega G_{T,D}(x,y)\hat{W}_i^{-2}(y)dy,$$

and hence $|\hat{Z}_i| \leq C$, where $C > 0$ is independent of i. We now obtain from the regularity of Δ^2 and (8.6) that $\hat{W}_i \to \hat{W}$ in $C^3_{loc}(\Omega)$ and $\hat{W}$ satisfies the equation

$$T\Delta\hat{W} - D\Delta^2\hat{W} = \frac{1}{C}\hat{W}^{-2} \text{ in } \Omega, \quad \hat{W} = \infty, \ \Delta\hat{W} = 0 \text{ on } \partial\Omega.$$

On the other hand, we see from (8.7) that $\hat{Z}_i \to \hat{Z}$ in $C^1(\overline{\Omega})$ as $i \to \infty$ and $\hat{Z} \equiv \Delta\hat{W}$ satisfies the problem

$$T\hat{Z} - D\Delta\hat{Z} = \frac{1}{C}\hat{W}^{-2} \text{ in } \Omega, \quad \hat{Z} = 0 \text{ on } \partial\Omega.$$

Since we easily know that $\Delta\hat{W} \leq C$ on $\overline{\Omega}$ and hence $\Delta(\hat{W} - C\rho) \leq 0$ in Ω, where $-\Delta\rho = 1$ in Ω and $\rho = 0$ on $\partial\Omega$, the maximum principle implies that $\hat{W}$ cannot be ∞ on $\partial\Omega$. Thus, (8.2) holds. This completes the proof of Theorem 8.2. ∎

Acknowledgments. This paper was done while the first author was visiting the Department of Mathematics at the Chinese University of Hong Kong. He would like to thank the Department for its hospitality. We thank Professor Michael Ward and Professor Dong Ye for useful discussions. We also thank the referees for their valuable suggestions.

REFERENCES

[1] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations, satisfying general boundary conditions*, I, Comm. Pure Appl. Math., 12 (1959), pp. 623–727.

[2] A. L. Bertozzi and M. C. Pugh, *Long-wave instabilities and saturation in thin film equations*, Comm. Pure Appl. Math., 51 (1998), pp. 625–661.

[3] A. L. Bertozzi and M. C. Pugh, *Finite-time blow-up of solutions of some long-wave unstable thin film equations*, Indiana Univ. Math. J., 49 (2000), pp. 1323–1366.

[4] H. Brezis and F. Merle, *Uniform estimates and blow-up behavior for solutions of* $-\Delta u = V(x)e^u$ *in two dimensions*, Comm. Partial Differential Equations, 16 (1991), pp. 1223–1254.

[5] J. P. Burelbach, S. G. Bankoff, and S. H. Davis, *Nonlinear stability of evaporating /condensing liquid films*, J. Fluid Mech., 195 (1998), pp. 463–494.

[6] M. G. Crandall and P. H. Rabinowitz, *Some continuation and variational methods for positive solutions of nonlinear elliptic eigenvalue problems*, Arch. Ration. Mech. Anal., 58 (1975), pp. 207–218.

[7] P. Esposito, N. Ghoussoub, and Y. Guo, *Compactness along the branch of semi-stable and unstable solutions for an elliptic problem with a singular nonlinearity*, Comm. Pure Appl. Math., 60 (2008), pp. 1731–1768.

[8] G. Flores, G. A. Mercado, and J. A. Pelesko, *Dynamics and touchdown in electrostatic MEMS*, in Proceedings of ICMENS 2003, IEEE Computer Society, Washington, DC, 2003, pp. 182–187.

[9] F. Gazzola and H.-Ch. Grunau, *Critical dimensions and higher order Sobolev inequalities with remainder terms*, NoDEA Nonlinear Differential Equations Appl., 8 (2001), pp. 35–44.

[10] N. Ghoussoub and Y. Guo, *On the partial differential equations of electrostatic MEMS devices: Stationary case*, SIAM J. Math. Anal., 38 (2007), pp. 1423–1449.

[11] N. Ghoussoub and Y. Guo, *On the partial differential equations of electrostatic MEMS devices* II: *Dynamic case*, NoDEA Nonlinear Differential Equations Appl., 15 (2008), pp. 115–145.

[12] Y. Guo, *On the partial differential equations of electrostatic MEMS devices* III: *Refined touchdown behavior*, J. Differential Equations, 244 (2008), pp. 2277–2309.

[13] Z. M. Guo and J. R. L. Webb, *Large and small solutions of a class of quasilinear elliptic eigenvalue problems*, J. Differential Equations, 180 (2002), pp. 1–50.

[14] Z. M. Guo and J. C. Wei, *Hausdorff dimension of ruptures for solutions of a semilinear elliptic equation with singular nonlinearity*, Manuscripta Math., 120 (2006), pp. 193–209.

[15] Z. M. Guo and J. C. Wei, *Symmetry of nonnegative solutions of a semilinear elliptic equation with singular nonlinearity*, Proc. Roy. Soc. Edinburgh Sect. A, 137 (2007), pp. 963–994.

[16] Z. M. Guo and J. C. Wei, *Infinitely many turning points for an elliptic problem with a singular nonlinearity*, J. London Math. Soc., 78 (2008), pp. 21–35.

[17] Y. Guo, Z. Pan, and M. J. Ward, *Touchdown and pull-in voltage behavior of a MEMS device with varying dielectric properties*, SIAM J. Appl. Math., 66 (2005), pp. 309–338.

[18] C. C. Hwang, C. K. Lin, and W. Y. Uen, *A nonlinear three-dimensional rupture theory of thin liquid films*, J. Colloid Interf. Sci., 190 (1997), pp. 250–252.

[19] R. S. Laugesen and M. C. Pugh, *Properties of steady states for thin film equations*, European J. Appl. Math., 11 (2000), pp. 293–351.

[20] R. S. Laugesen and M. C. Pugh, *Energy levels of steady-states for thin-film-type equations*, J. Differential Equations, 182 (2002), pp. 377–415.

[21] R. S. Laugesen and M. C. Pugh, *Linear stability of steady states for thin film and Cahn–Hilliard type equations*, Arch. Ration. Mech. Anal., 154 (2000), pp. 3–51.

[22] F. H. Lin and Y. S. Yang, *Nonlinear non-local elliptic equation modelling electrostatic actuation*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 463 (2007), pp. 1323–1337.

[23] A. Lindsay and M. Ward, *Asymptotics of some nonlinear eigenvalue problems for a MEMS capacitor. Part I: Fold point asymptotics*, Methods Anal. Appl., to appear.

[24] J. A. Pelesko, *Mathematical modeling of electrostatic MEMS with tailored dielectric properties*, SIAM J. Appl. Math., 62 (2002), pp. 888–908.
[25] J. A. Pelesko and D. H. Bernstein, *Modeling MEMS and NEMS*, Chapman & Hall/CRC Press, Boca Raton, FL, 2002.
[26] W. C. Troy, *Symmetry properties in systems of semilinear elliptic equations*, J. Differential Equations, 42 (1981), pp. 400–413.
[27] R. C. A. M. van der Vorst, *Best constant for the embedding of the space* $H^2 \cap H_0^1(\Omega)$ *into* $L^{2N/(N-4)}(\Omega)$, Differential Integral Equations, 6 (1993), pp. 259–276.

SIAM J. MATH. ANAL.
Vol. 40, No. 5, pp. 2055–2076

STABILITY AND SYMMETRY IN THE NAVIER PROBLEM FOR THE ONE-DIMENSIONAL WILLMORE EQUATION*

KLAUS DECKELNICK† AND HANS-CHRISTOPH GRUNAU†

Abstract. We consider the one-dimensional Willmore equation subject to Navier boundary conditions; i.e., the position and the curvature are prescribed on the boundary. In a previous work, explicit symmetric solutions to symmetric data have been constructed. Within a certain range of boundary curvatures one has precisely two symmetric solutions, while for boundary curvatures outside the closure of this range there are none. The solutions are ordered; one is "small," and the other is "large." In the first part of this paper we address the stability problem and show that the small solution is (linearized) stable in the whole open range of admissible boundary curvatures, while the large one is unstable and has Morse index 1. A second goal is to investigate whether the small solution is minimal for the corresponding Willmore functional. It turns out that for a certain subrange of admissible boundary curvatures the small solution is the unique minimum, while for curvatures outside that range the minimum is not attained. As a byproduct of our argument we show that for any admissible function there exists a symmetric function with smaller Willmore energy.

Key words. Willmore equation, Navier boundary conditions, stability, Morse index, symmetry

AMS subject classifications. 53C21, 34B15, 35J65, 35B35

DOI. 10.1137/07069033X

1. Introduction.

Recently, Willmore surfaces (see [19]) and the related flow have attracted quite some attention; see, e.g., [1, 8, 9, 10, 13, 17, 18], [3] for numerical studies, and [15, 5] for elastic curves, which are the one-dimensional analogues. The mentioned work is concerned with closed surfaces and curves, while only very few results concerning boundary value problems are available. Quite recently, Schätzle [16] considered Willmore surfaces with a boundary, which are subject to the constraint to be submanifolds of $\mathbb{S}^n$ and which satisfy Dirichlet-type boundary conditions.

In order to gain some more insight into general boundary conditions for the "free" Willmore equation, in [4] we had a look at the one-dimensional case, where, in some situations, almost explicit solutions can be found for suitable boundary value problems. For further background information and references, see [4] and also [14]. In [4], we were interested in Willmore graphs and studied among others the *Navier boundary value problem* with symmetric data $\alpha \in \mathbb{R}$ for the *one-dimensional Willmore equation*:

$$\begin{cases} \frac{1}{\sqrt{1+u'(x)^2}} \frac{d}{dx} \left(\frac{\kappa'(x)}{\sqrt{1+u'(x)^2}} \right) + \frac{1}{2}\kappa^3(x) = 0, \quad x \in (0,1), \\ u(0) = u(1) = 0, \quad \kappa(0) = \kappa(1) = -\alpha. \end{cases} \tag{1.1}$$

Here

$$\kappa(x) = \frac{d}{dx} \left(\frac{u'(x)}{\sqrt{1+u'(x)^2}} \right) = \frac{u''(x)}{(1+u'(x)^2)^{3/2}} \tag{1.2}$$

*Received by the editors May 2, 2007; accepted for publication (in revised form) October 2, 2008; published electronically January 21, 2009.
http://www.siam.org/journals/sima/40-5/69033.html
†Fakultät für Mathematik, Otto-von-Guericke-Universität Magdeburg, Postfach 4120, D-39016 Magdeburg, Germany (Klaus.Deckelnick@mathematik.uni-magdeburg.de, Hans-Christoph.Grunau@mathematik.uni-magdeburg.de).

denotes the curvature of the graph of u at the point $(x, u(x))$. Solutions of (1.1) are critical points of the modified one-dimensional Willmore functional

(1.3)
$$\tilde{W}_\alpha(u) = \int_{\mathrm{graph}(u)} \big(\kappa(x)^2 + 2\alpha\kappa(x)\big)\, ds(x) = \int_0^1 \big(\kappa(x)^2 + 2\alpha\kappa(x)\big)\, \sqrt{1+u'(x)^2}\, dx,$$

with $u \in H^2(0,1) \cap H_0^1(0,1)$; see [4, section 2]. The boundary conditions $u(0) = u(1) = 0$ are formulated by working in the space H_0^1, while the curvature boundary conditions $\kappa(0) = \kappa(1) = -\alpha$ arise as natural boundary conditions since also the admissible testing functions have only to be in $H^2 \cap H_0^1$. By reflection it is sufficient to consider

$$\alpha \geq 0.$$

Our first results are concerned with solutions of (1.1) which are symmetric about $x = 1/2$. This class is particularly important for two reasons: on the one hand, it is possible to associate with each function in $H^2 \cap H_0^1$ a symmetric function with the same or smaller Willmore energy. The corresponding construction is nontrivial and will be described in the proof of Theorem 1.5. Moreover, exploiting an observation due to Euler [6, p. 234, line 13], we can give symmetric solutions more or less in closed form. This issue was discussed in detail in [4], where, among other things, the following result was proved.

PROPOSITION 1.1 (see [4, Theorem 1]). *There exists* $\alpha_{\max} = 1.343799725\ldots$ *such that for* $0 < \alpha < \alpha_{\max}$, *the Navier boundary value problem* (1.1) *has precisely two smooth (graph) solutions* u *in the class of smooth functions that are symmetric around* $x = \frac{1}{2}$. *If* $\alpha = \alpha_{\max}$, *one has precisely one such solution, for* $\alpha = 0$ *one has only the trivial solution, and for* $\alpha > \alpha_{\max}$ *no such solutions exist.*

Both solutions are positive, and one of these solutions is larger than the other. The small solutions are ordered with respect to α, while the large ones become smaller for increasing α; see Figure 1.1. For the bifurcation diagram, see Figure 1.2.

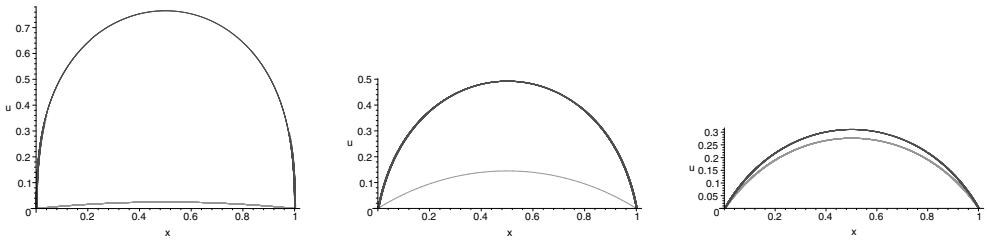


FIG. 1.1. *Solutions of the Navier boundary value problem* (1.1) *for* $\alpha = 0.2$, $\alpha = 1$, *and* $\alpha = 1.34$ *(left to right). See* [4, *Figure* 1].

It is an obvious conjecture that for $0 \leq \alpha < \alpha_{\max}$ the small solutions are (linearized) stable. This property was left open in [4], and to prove it is the first goal of this paper.

THEOREM 1.2. *Assume that* $0 \leq \alpha < \alpha_{\max}$ *and that* u *is the symmetric* small solution *of the Navier boundary value problem* (1.1). *Then, this solution is linearized stable; i.e., the spectrum of the (self-adjoint) linearization of* (1.1) *around* u *is contained in* $(0, \infty)$.

That these linearizations are the second variation of the functional $\tilde{W}_\alpha$ proves that the small solution is a local minimum of the functional $\tilde{W}_\alpha$ in $H^2 \cap H_0^1(0,1)$.

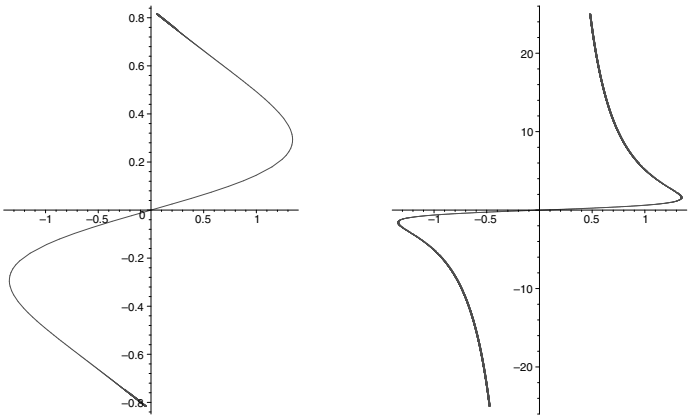


FIG. 1.2. *Bifurcation diagram for* (1.1)*: The extremals value of the solution* $u(1/2)$ *(left) and of the derivative* $u'(0)$ *(right) plotted over* α*. See* [4, *Figure* 2].

Furthermore, we will show that on $0 < \alpha < \alpha_{\max}$, the large solutions are unstable. More precisely, we prove that the following holds.

THEOREM 1.3. *Assume that* $0 < \alpha < \alpha_{\max}$ *and that* u *is the symmetric* large solution *of the Navier boundary value problem* (1.1)*. Then, this solution is unstable and has Morse index* 1*; i.e., one eigenvalue of the (self-adjoint) linearization of* (1.1) *is negative, while the remaining spectrum is contained in* $(0, \infty)$.

We emphasize that no symmetry assumptions are made in the discussion of the linearizations of (1.1).

A further important question is whether the small solutions are not only a local but also a global minimum of the functional $\tilde{W}_\alpha$.

THEOREM 1.4. *There exists* $\alpha^* = 1.132372323\ldots \in (0, \alpha_{\max})$ *such that for* $0 \le \alpha \le \alpha^*$ *the small solution* u *is the unique global minimum of the functional* $\tilde{W}_\alpha$ *in the class* $H^2 \cap H_0^1(0,1)$*. If* $\alpha^* < \alpha \le \alpha_{\max}$*, the infimum of* $\tilde{W}_\alpha$ *in* $H^2 \cap H_0^1(0,1)$ *is not attained, and in that case*

$$\inf_{v \in H^2 \cap H_0^1(0,1)} \tilde{W}_\alpha(v) = \left(\int_{\mathbb{R}} \frac{1}{(1+\tau^2)^{5/4}}\, d\tau \right)^2 - 2\alpha\pi.$$

The main idea of proving Theorem 1.4 consists in reducing the minimization of $\tilde{W}_\alpha$ over $H^2 \cap H_0^1(0,1)$ to the minimization of a function of two variables. As a byproduct of this approach we shall see that the infimum of the Willmore energy in $H^2 \cap H_0^1(0,1)$ coincides with the infimum in the subspace M of functions that are symmetric about $x = 1/2$; i.e., for every function in $H^2 \cap H_0^1(0,1)$, there exists a symmetric function with the same or smaller Willmore energy. This is remarkable since we deal with a fourth order problem and the well-known symmetrization procedures do not apply.

THEOREM 1.5. *Let* M *be the class of functions in* $H^2 \cap H_0^1(0,1)$*, which are symmetric about* $x = 1/2$*. Then we have*

$$\inf_{v \in H^2 \cap H_0^1(0,1)} \tilde{W}_\alpha(v) = \inf_{v \in M} \tilde{W}_\alpha(v).$$

2. Linearized stability. To prove Theorem 1.2 we describe in more detail how the symmetric solutions to (1.1) were obtained in [4].

In what follows, the function

$$G : \mathbb{R} \to \left(-\frac{c_0}{2}, \frac{c_0}{2}\right), \qquad G(s) := \int_0^s \frac{1}{(1+\tau^2)^{5/4}}\, d\tau, \tag{2.1}$$

$$c_0 = \int_{\mathbb{R}} \frac{1}{(1+\tau^2)^{5/4}}\, d\tau = \mathcal{B}\left(\frac{1}{2}, \frac{3}{4}\right) = 2.396280469\ldots,$$

plays a crucial role. It is straightforward to see that G is strictly increasing and bijective with $G'(s) > 0$. So, also the inverse function

$$G^{-1} : \left(-\frac{c_0}{2}, \frac{c_0}{2}\right) \to \mathbb{R} \tag{2.2}$$

is strictly increasing, bijective, and smooth with $G^{-1}(0) = 0$.

LEMMA 2.1 (see [4, Lemma 4]). *Let $u \in C^4([0,1])$ be a function symmetric about $x = 1/2$. Then u solves the Willmore equation in (1.1) iff there exists $c \in (-c_0, c_0)$ such that*

$$\forall x \in [0,1] : \qquad u'(x) = G^{-1}\left(\frac{c}{2} - cx\right). \tag{2.3}$$

For the curvature, one has that

$$\kappa(x) = -\frac{c}{\sqrt[4]{1 + G^{-1}\left(\frac{c}{2} - cx\right)^2}}. \tag{2.4}$$

Moreover, if we additionally assume that $u(0) = u(1) = 0$, then one has

$$u(x) = \frac{2}{c\sqrt[4]{1 + G^{-1}\left(\frac{c}{2} - cx\right)^2}} - \frac{2}{c\sqrt[4]{1 + G^{-1}\left(\frac{c}{2}\right)^2}} \quad (c \neq 0). \tag{2.5}$$

In order to solve the Navier boundary value problem (1.1), in [4], we had to study the function

$$h : (-c_0, c_0) \to \mathbb{R}, \qquad h(c) = \frac{c}{\sqrt[4]{1 + G^{-1}\left(\frac{c}{2}\right)^2}} \tag{2.6}$$

and the equation $h(c) = \alpha$. See Figure 2.1. The range of h is precisely the set of α, for which the Navier boundary value problem (1.1) has a smooth symmetric graph solution. The number of solutions c of the equation $\alpha = h(c)$ is the number of such solutions of the boundary value problem.

LEMMA 2.2 (see [4, Lemma 6]). *We have $h > 0$ in $(0, c_0)$, $h < 0$ in $(-c_0, 0)$, $\lim_{c \nearrow c_0} h(c) = \lim_{c \searrow -c_0} h(c) = 0$. The function h is odd and has precisely one local maximum in $c_{\max} = 1.840428142\ldots$ and one local minimum in $c_{\min} = -c_{\max}$. The corresponding value is $\alpha_{\max} = h(c_{\max}) = 1.343799725\ldots$.*

The small solutions correspond precisely to $c \in (0, c_{\max})$ and the large ones to $c \in (c_{\max}, c_0)$. Let us fix $c \in (0, c_0)$ with corresponding $\alpha = h(c)$ and solution u given by (2.5). First we have to calculate the linearization of (1.1) around u, i.e., the second variation of the modified Willmore functional $\tilde{W}_\alpha$ in u.

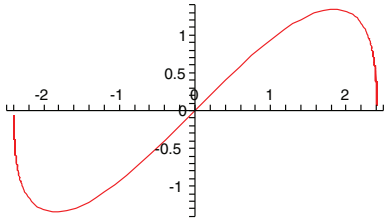


FIG. 2.1. *The function $c \mapsto h(c)$. According to* [4, *Lemma* 4], *solutions to* (1.1) *are given by solving $h(c) = \alpha$.*

LEMMA 2.3. *We have*

$$D^2\tilde{W}_\alpha(u)(\varphi,\eta) = 2\int_0^1 \frac{\varphi''(x)\eta''(x)}{(1+u'(x)^2)^{5/2}}\,dx + 5\int_0^1 \frac{1-u'(x)^2}{(1+u'(x)^2)^{3/2}}\kappa(x)^2\varphi'(x)\eta'(x)\,dx + 6\alpha\left[\frac{u'(x)\varphi'(x)\eta'(x)}{(1+u'(x)^2)^2}\right]_0^1, \quad \varphi,\eta \in H^2 \cap H_0^1(0,1).$$

Proof. According to [4, Lemma 2 and Corollary 1], the first variation of $\tilde{W}_\alpha(u)$ is given by

$$D\tilde{W}_\alpha(u)(\varphi) = 2\int_0^1 \frac{u''(x)\varphi''(x)}{(1+u'(x)^2)^{5/2}}\,dx - 5\int_0^1 \frac{u'(x)u''(x)^2\varphi'(x)}{(1+u'(x)^2)^{7/2}}\,dx + 2\alpha\left[\frac{\varphi'(x)}{1+u'(x)^2}\right]_0^1, \qquad \varphi \in H^2 \cap H_0^1(0,1).$$

In order to obtain the second derivative, we also consider $\eta \in H^2 \cap H_0^1(0,1)$ and differentiate the previous expression with respect to this direction:

$$\begin{aligned}
D^2\tilde{W}_\alpha(u)(\varphi,\eta) &= \frac{d}{dt}D\tilde{W}_\alpha(u+t\eta)(\varphi)|_{t=0} \\
&= 2\int_0^1 \frac{\varphi''(x)\eta''(x)}{(1+u'(x)^2)^{5/2}}\,dx - 10\int_0^1 \frac{u'(x)u''(x)\varphi''(x)\eta'(x)}{(1+u'(x)^2)^{7/2}}\,dx \\
&\quad - 10\int_0^1 \frac{u'(x)u''(x)\varphi'(x)\eta''(x)}{(1+u'(x)^2)^{7/2}}\,dx - 5\int_0^1 \frac{u''(x)^2\varphi'(x)\eta'(x)}{(1+u'(x)^2)^{7/2}}\,dx \\
&\quad + 35\int_0^1 \frac{u'(x)^2u''(x)^2\varphi'(x)\eta'(x)}{(1+u'(x)^2)^{9/2}}\,dx - 4\alpha\left[\frac{u'(x)\varphi'(x)\eta'(x)}{(1+u'(x)^2)^2}\right]_0^1 \\
&= 2\int_0^1 \frac{\varphi''(x)\eta''(x)}{(1+u'(x)^2)^{5/2}}\,dx - 5\int_0^1 \frac{\kappa(x)^2\varphi'(x)\eta'(x)}{\sqrt{1+u'(x)^2}}\,dx \\
&\quad - 10\int_0^1 \kappa(x)\cdot\frac{u'(x)}{\sqrt{1+u'(x)^2}}\cdot\frac{1}{(1+u'(x)^2)^{3/2}}\cdot\frac{d}{dx}\left(\varphi'(x)\eta'(x)\right)dx \\
&\quad + 35\int_0^1 \frac{u'(x)^2\kappa(x)^2\varphi'(x)\eta'(x)}{(1+u'(x)^2)^{3/2}}\,dx - 4\alpha\left[\frac{u'(x)\varphi'(x)\eta'(x)}{(1+u'(x)^2)^2}\right]_0^1.
\end{aligned}$$

To proceed further we would like to integrate the third term by parts. Here we will exploit that u is a solution to (1.1). In particular, u is smooth and satisfies the Navier

boundary data $\kappa(x) = -\alpha$, $x \in \{0,1\}$.

$$\begin{aligned}
D^2\tilde{W}_\alpha(u)(\varphi,\eta) &= 2\int_0^1 \frac{\varphi''(x)\eta''(x)}{(1+u'(x)^2)^{5/2}}\,dx - 5\int_0^1 \frac{\kappa(x)^2\varphi'(x)\eta'(x)}{\sqrt{1+u'(x)^2}}\,dx \\
&\quad + 35\int_0^1 \frac{u'(x)^2\kappa(x)^2\varphi'(x)\eta'(x)}{(1+u'(x)^2)^{3/2}}\,dx - 4\alpha\left[\frac{u'(x)\varphi'(x)\eta'(x)}{(1+u'(x)^2)^2}\right]_0^1 \\
&\quad - 10\left[\kappa(x)\frac{u'(x)\varphi'(x)\eta'(x)}{(1+u'(x)^2)^2}\right]_0^1 + 10\int_0^1 \frac{\kappa'(x)u'(x)\varphi'(x)\eta'(x)}{(1+u'(x)^2)^2}\,dx \\
&\quad + 10\int_0^1 \frac{\kappa(x)^2\varphi'(x)\eta'(x)}{(1+u'(x)^2)^{3/2}}\,dx - 30\int_0^1 \frac{\kappa(x)u'(x)^2u''(x)\varphi'(x)\eta'(x)}{(1+u'(x)^2)^3}\,dx \\
&= 2\int_0^1 \frac{\varphi''(x)\eta''(x)}{(1+u'(x)^2)^{5/2}}\,dx - 5\int_0^1 \frac{\kappa(x)^2\varphi'(x)\eta'(x)}{\sqrt{1+u'(x)^2}}\,dx \\
&\quad + 5\int_0^1 \frac{u'(x)^2\kappa(x)^2\varphi'(x)\eta'(x)}{(1+u'(x)^2)^{3/2}}\,dx + 6\alpha\left[\frac{u'(x)\varphi'(x)\eta'(x)}{(1+u'(x)^2)^2}\right]_0^1 \\
&\quad + 10\int_0^1 \frac{\kappa'(x)u'(x)\varphi'(x)\eta'(x)}{(1+u'(x)^2)^2}\,dx + 10\int_0^1 \frac{\kappa(x)^2\varphi'(x)\eta'(x)}{(1+u'(x)^2)^{3/2}}\,dx.
\end{aligned}$$

We infer from (2.3) and (2.4) that

$$\forall x \in [0,1], \qquad \kappa(x)\left(1+u'(x)^2\right)^{1/4} = -c,$$

and hence

$$\forall x \in [0,1], \quad \kappa'(x)\left(1+u'(x)^2\right)^{1/4} + \frac{1}{2}u'(x)\kappa(x)^2\left(1+u'(x)^2\right)^{3/4} = 0.$$

Consequently,

$$\begin{aligned}
D^2\tilde{W}_\alpha(u)(\varphi,\eta) &= 2\int_0^1 \frac{\varphi''(x)\eta''(x)}{(1+u'(x)^2)^{5/2}}\,dx - 5\int_0^1 \frac{\kappa(x)^2\varphi'(x)\eta'(x)}{\sqrt{1+u'(x)^2}}\,dx \\
&\quad + 6\alpha\left[\frac{u'(x)\varphi'(x)\eta'(x)}{(1+u'(x)^2)^2}\right]_0^1 + 10\int_0^1 \frac{\kappa(x)^2\varphi'(x)\eta'(x)}{(1+u'(x)^2)^{3/2}}\,dx.
\end{aligned}$$

This proves our claim. □

Looking at $\eta \in H^2(0,1) \cap H_0^1(0,1)$ as a test function, we now obtain the linearization from the second variation with the help of integration by parts. Note that the first integral gives rise to a further boundary term $2\left[\frac{\varphi''(x)\eta'(x)}{(1+u'(x)^2)^{5/2}}\right]_0^1$. Expressing u and α in terms of c according to Lemmas 2.1 and 2.2, the linearization of (1.1) around u reads as follows:

(2.7)

$$\begin{cases}
\left(\frac{\varphi''(x)}{\left(1+G^{-1}\left(\frac{c}{2}-cx\right)^2\right)^{5/2}}\right)'' + \frac{5}{2}c^2\left(\frac{G^{-1}\left(\frac{c}{2}-cx\right)^2-1}{\left(1+G^{-1}\left(\frac{c}{2}-cx\right)^2\right)^2}\varphi'(x)\right)' = 0, \quad x \in (0,1), \\
\varphi(0) = \varphi(1) = 0, \\
\frac{\varphi''(0)}{\left(1+G^{-1}\left(\frac{c}{2}\right)^2\right)^{5/2}} + 3\frac{c\,G^{-1}\left(\frac{c}{2}\right)\varphi'(0)}{\left(1+G^{-1}\left(\frac{c}{2}\right)^2\right)^{9/4}} = 0, \quad \frac{\varphi''(1)}{\left(1+G^{-1}\left(\frac{c}{2}\right)^2\right)^{5/2}} - 3\frac{c\,G^{-1}\left(\frac{c}{2}\right)\varphi'(1)}{\left(1+G^{-1}\left(\frac{c}{2}\right)^2\right)^{9/4}} = 0.
\end{cases}$$

For $c = 0$, the small solution of (1.1) is $u(x) \equiv 0$, and $D^2\tilde{W}_0(u)(\varphi,\varphi) = \int_0^1 \varphi''(x)^2\,dx$ is positive definite in $H^2 \cap H_0^1(0,1)$ with respect to the $L^2(0,1)$-norm. The spectrum of the linearization as a regular elliptic operator on a bounded interval together with suitable boundary conditions consists only of eigenvalues. Since these eigenvalues depend smoothly on u and u depends smoothly on c, $D^2\tilde{W}_\alpha(u)(\varphi,\varphi)$ remains positive definite for c increasing from 0 as long as (2.7) has only the trivial solution $\varphi(x) \equiv 0$.

We assume that (2.7) has a solution φ and put

$$\chi(x) := \varphi'(x).$$

Then, there exists a constant $A \in \mathbb{R}$ such that χ solves the second order differential equation

$$\left(\frac{\chi'(x)}{\left(1+G^{-1}\left(\frac{c}{2}-cx\right)^2\right)^{5/2}}\right)' + \frac{5}{2}c^2\left(\frac{G^{-1}\left(\frac{c}{2}-cx\right)^2-1}{\left(1+G^{-1}\left(\frac{c}{2}-cx\right)^2\right)^2}\chi\right) = c^2A.$$

We introduce more suitable variables,

$$\begin{aligned} y &= G^{-1}\left(\frac{c}{2}-cx\right) \in \left[-G^{-1}\left(\frac{c}{2}\right), G^{-1}\left(\frac{c}{2}\right)\right], \quad x = \frac{1}{2} - \frac{G(y)}{c}, \\ \psi(y) &:= \chi(x) = \chi\left(\frac{1}{2}-\frac{G(y)}{c}\right), \qquad \chi(x) = \psi\left(G^{-1}\left(\frac{c}{2}-cx\right)\right), \\ \chi'(x) &= -c\left(1+G^{-1}\left(\frac{c}{2}-cx\right)^2\right)^{5/4}\psi'\left(G^{-1}\left(\frac{c}{2}-cx\right)\right), \end{aligned}$$

and conclude that ψ solves the following boundary value problem:

$$\text{(2.8)} \qquad \begin{cases} \psi''(y) - \frac{5y}{2(1+y^2)}\psi'(y) + \frac{5(y^2-1)}{2(1+y^2)^2}\psi(y) = A, \quad y \in (-y_0, y_0), \\ \psi'(-y_0) + \frac{3y_0}{1+y_0^2}\psi(-y_0) = 0, \qquad \psi'(y_0) - \frac{3y_0}{1+y_0^2}\psi(y_0) = 0. \end{cases}$$

Here, we denote

$$\text{(2.9)} \qquad y_0 := G^{-1}\left(\frac{c}{2}\right).$$

To simplify the boundary conditions we make a last change of variables and put

$$\text{(2.10)} \qquad \Phi(y) := \frac{\psi(y)}{(1+y^2)^{3/2}}, \qquad y \in [-y_0, y_0],$$

and finally consider the following boundary value problem:

$$\text{(2.11)} \qquad \begin{cases} (1+y^2)^{3/2}\Phi''(y) + \frac{7}{2}y(1+y^2)^{1/2}\Phi'(y) \\ \qquad\qquad + \left(y^2+\frac{1}{2}\right)(1+y^2)^{-1/2}\Phi(y) = A, \quad y \in (-y_0, y_0), \\ \Phi'(-y_0) = \Phi'(y_0) = 0. \end{cases}$$

We recall the definition of $G(y) := \int_0^y \frac{1}{(1+\tau^2)^{5/4}}\,d\tau$ and put

$$\text{(2.12)} \qquad \Phi_0(y) := -2\frac{1}{\sqrt{1+y^2}}, \qquad \Phi_1(y) := \frac{1}{\sqrt[4]{1+y^2}}, \qquad \Phi_2(y) := \frac{G(y)}{\sqrt[4]{1+y^2}}.$$

Then, one directly verifies that the general solution of the differential equation in (2.11) is given by

$$\Phi(y) := A \cdot \Phi_0(y) + \gamma_1 \cdot \Phi_1(y) + \gamma_2 \cdot \Phi_2(y) \tag{2.13}$$

with $\gamma_1, \gamma_2 \in \mathbb{R}$. Since $A \cdot \Phi_0(y) + \gamma_1 \cdot \Phi_1(y)$ is even and $\gamma_2 \cdot \Phi_2(y)$ is odd, the boundary conditions in (2.11) are equivalent to

$$A \cdot \Phi_0'(y_0) + \gamma_1 \cdot \Phi_1'(y_0) = 0 \text{ and } \gamma_2 \cdot \Phi_2'(y_0) = 0, \tag{2.14}$$

which in turn are equivalent to

$$\gamma_1 = \frac{4A}{\sqrt[4]{1+y_0^2}} \text{ and } (\gamma_2 = 0 \text{ or } \Phi_2'(y_0) = 0)\,. \tag{2.15}$$

A beautiful coincidence between these solutions and the functions involved in the proof of Theorem 1.1 can be observed, namely,

$$\Phi_2(y) = \frac{1}{2} h(2G(y)), \qquad \Phi_2'(y) = \frac{h'(2G(y))}{(1+y^2)^{5/4}}. \tag{2.16}$$

With the help of these observations we are now ready to conclude the following lemma.

LEMMA 2.4. *For $c \in [0, c_0) \setminus \{c_{\max}\}$, the boundary value problem (2.7) has only the trivial solution $\varphi(x) \equiv 0$. For $c = c_{\max}$, it has a one-dimensional null space which is spanned by*

$$\varphi(x) = \frac{1}{c} \int_{G^{-1}(\frac{c}{2} - cx)}^{G^{-1}(\frac{c}{2})} G(\eta)\, d\eta.$$

If $c = c_{\max}$, $\alpha = \alpha_{\max}$, instabilities will occur first from the corresponding solution u in direction of this function φ; see Figure 2.2.

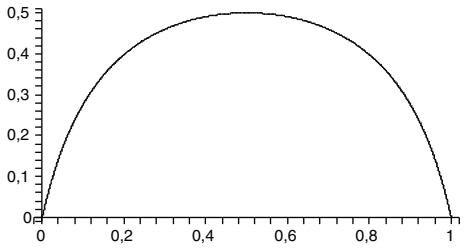


FIG. 2.2. *Profile of the unstable direction in $c = c_{\max}$.*

Proof. The case $c = 0$ is obvious, and we consider only $c \in (0, c_0)$. We denote

$$\tilde{\Phi}_1(y) := \Phi_0(y) + \frac{4}{\sqrt[4]{1+y_0^2}} \cdot \Phi_1(y) = \frac{2}{\sqrt[4]{1+y^2} \cdot \sqrt[4]{1+y_0^2}} \left(2 - \frac{\sqrt[4]{1+y_0^2}}{\sqrt[4]{1+y^2}} \right).$$

According to (2.13), we have to study

$$\Phi(y) = A\tilde{\Phi}_1(y) + \gamma_2 \Phi_2(y)$$

with some suitable $A, \gamma_2 \in \mathbb{R}$. Let φ be the corresponding solution of (2.7) which is obtained from Φ by tracing back the changes of variables and integrating χ. We want to show first that necessarily $A = 0$ for any $c \in [0, c_0)$:

$$\begin{aligned}
0 = \varphi(1) - \varphi(0) &= \int_0^1 \chi(x)\,dx = \int_0^1 \psi\left(G^{-1}\left(\frac{c}{2} - cx\right)\right) dx \\
&= \frac{1}{c}\int_{-G^{-1}\left(\frac{c}{2}\right)}^{G^{-1}\left(\frac{c}{2}\right)} \psi(y)(1+y^2)^{-5/4}\,dy = \frac{1}{c}\int_{-G^{-1}\left(\frac{c}{2}\right)}^{G^{-1}\left(\frac{c}{2}\right)} \Phi(y)(1+y^2)^{1/4}\,dy \\
&= \frac{A}{c}\int_{-G^{-1}\left(\frac{c}{2}\right)}^{G^{-1}\left(\frac{c}{2}\right)} \tilde{\Phi}_1(y)(1+y^2)^{1/4}\,dy + \frac{\gamma_2}{c}\int_{-G^{-1}\left(\frac{c}{2}\right)}^{G^{-1}\left(\frac{c}{2}\right)} \Phi_2(y)(1+y^2)^{1/4}\,dy \\
&= \frac{A}{c}\int_{-G^{-1}\left(\frac{c}{2}\right)}^{G^{-1}\left(\frac{c}{2}\right)} \tilde{\Phi}_1(y)(1+y^2)^{1/4}\,dy
\end{aligned}$$

since Φ_2 is odd. Hence we may conclude that

$$\begin{aligned}
0 &= \frac{A}{c}\int_{-G^{-1}\left(\frac{c}{2}\right)}^{G^{-1}\left(\frac{c}{2}\right)} \tilde{\Phi}_1(y)(1+y^2)^{1/4}\,dy \\
&= \frac{2A}{c}\int_{-G^{-1}\left(\frac{c}{2}\right)}^{G^{-1}\left(\frac{c}{2}\right)} \left(\frac{2}{\left(1+G^{-1}\left(\frac{c}{2}\right)^2\right)^{1/4}} - \frac{1}{(1+y^2)^{1/4}}\right) dy \\
&= \frac{4A}{c} F\left(G^{-1}\left(\frac{c}{2}\right)\right),
\end{aligned}$$

where F is defined by

$$F(\eta) := \frac{2\eta}{(1+\eta^2)^{1/4}} - \int_0^\eta \frac{1}{(1+s^2)^{1/4}}\,ds.$$

Since $F(0) = 0$ and

$$\begin{aligned}
F'(\eta) &= \frac{2}{(1+\eta^2)^{1/4}} - \frac{\eta^2}{(1+\eta^2)^{5/4}} - \frac{1}{(1+\eta^2)^{1/4}} \\
&= \frac{1}{(1+\eta^2)^{1/4}} - \frac{\eta^2}{(1+\eta^2)^{5/4}} = \frac{1}{(1+\eta^2)^{5/4}} > 0,
\end{aligned}$$

we have

$$F\left(G^{-1}\left(\frac{c}{2}\right)\right) > 0.$$

As a consequence, $A = 0$, and hence $\gamma_1 = 0$ by (2.15), and we are left with considering $\gamma_2\Phi_2$. We have that $h'(c) > 0$ for $c \in (0, c_{\max})$ and $h'(c) < 0$ for $c \in (c_{\max}, c_0)$. By making use of

$$\Phi_2'(y) = \frac{h'(2G(y))}{(1+y^2)^{5/4}}$$

and the boundary condition $\gamma_2\Phi_2'(G^{-1}(c/2)) = 0$, we conclude that $\gamma_2 = 0$ provided $c \in (0, c_0) \setminus \{c_{\max}\}$. If $c = c_{\max}$, then Φ_2 is a nontrivial solution of (2.11). For the

corresponding nontrivial solution φ of (2.7) we derive

$$\begin{aligned}\varphi(x) &= \gamma_2 \frac{c_{\max}}{2} \int_0^x \left(1 + G^{-1}\left(\frac{c_{\max}}{2} - c_{\max}\xi\right)^2\right)^{5/4} (1-2\xi)\, d\xi \\ &= \frac{\gamma_2}{c_{\max}} \int_{G^{-1}(\frac{c_{\max}}{2} - c_{\max}x)}^{G^{-1}(\frac{c_{\max}}{2})} G(\eta)\, d\eta,\end{aligned}$$

where we made use of the boundary conditions $\varphi(0) = \varphi(1) = 0$. □

Proof of Theorem 1.2. The proof is now immediate. By the preceding lemma we have that on $[0, c_{\max})$, 0 is not an eigenvalue of (2.7). Since $D^2\tilde{W}_0(u)(\varphi,\varphi)$ is positive definite in $H^2 \cap H_0^1(0,1)$ with respect to the $L^2(0,1)$-norm, by continuity, the same holds true for $D^2\tilde{W}_\alpha(u)(\varphi,\varphi)$ for $c \in [0, c_{\max})$, which is the stated linearized stability of the corresponding small solutions of (1.1). □

As an immediate consequence of Theorem 1.2 we obtain a global existence result for the geometric flow associated with (1.1), namely,

$$V = -\kappa_{ss} - \frac{1}{2}\kappa^3 \quad \text{on } \Gamma(t).$$

Here, V denotes the upward normal velocity of the evolving graphs

$$\Gamma(t) = \{(x, v(x,t)) \mid x \in [0,1]\}.$$

The above evolution law then leads to the parabolic initial-boundary value problem (2.18) below. The principle of linearized stability as it was proved in great generality by Latushkin, Prüss, and Schnaubelt [11, Proposition 16] can be applied to our situation and allows us to obtain global existence and asymptotic stability for initial data close to a small solution to (1.1).

COROLLARY 2.5. *Assume that* $c \in [0, c_{\max})$, *and let* $\alpha = h(c) = \frac{c}{\sqrt[4]{1+G^{-1}(\frac{c}{2})^2}}$ *and*

$$u(x) = \frac{2}{c\sqrt[4]{1 + G^{-1}\left(\frac{c}{2} - cx\right)^2}} - \frac{2}{c\sqrt[4]{1 + G^{-1}\left(\frac{c}{2}\right)^2}} \tag{2.17}$$

be the corresponding small solution of (1.1). *We fix some* $p > 5$. *Then, there exist* $\delta, \rho, C > 0$ *such that for* $v_0 \in W^{4,p}(0,1)$ *with* $v_0(0) = v_0(1) = 0$, $\kappa_{v_0}(0) = \kappa_{v_0}(1) = -\alpha$, *and*

$$\|v_0 - u\|_{W^{4,p}(0,1)} \le \delta,$$

there exists a global solution $v \in L^p(0,\infty, W^{4,p}(0,1)) \cap W^{1,p}(0,\infty, L^p(0,1))$ *of the initial Navier boundary value problem*

(2.18)
$$\begin{cases} \frac{v_t(t,x)}{\sqrt{1+v_x(t,x)^2}} + \frac{1}{\sqrt{1+v_x(t,x)^2}} \frac{d}{dx}\left(\frac{\kappa_{v,x}(t,x)}{\sqrt{1+v_x(t,x)^2}}\right) + \frac{1}{2}\kappa_v^3(t,x) = 0, & (t,x) \in [0,\infty) \times [0,1], \\ v(t,0) = v(t,1) = 0, \quad \kappa_v(t,0) = \kappa_v(t,1) = -\alpha, & t \in [0,\infty), \\ v(0,x) = v_0(x), & x \in [0,1]. \end{cases}$$

One has exponential convergence toward the steady state u:

$$\|v(t,\,.\,) - u(\,.\,)\|_{W^{4,p}(0,1)} \le C\exp(-\rho t) \qquad (t \ge 1). \tag{2.19}$$

With similar but simpler techniques and calculations one finds that the unique solution (cf. [4, Theorem 2]), being symmetric about $x = 1/2$ of the Dirichlet problem

$$\begin{cases} \frac{1}{\sqrt{1+u_x(x)^2}} \frac{d}{dx} \left(\frac{\kappa_x(x)}{\sqrt{1+u_x(x)^2}} \right) + \frac{1}{2}\kappa^3(t,x) = 0, & x \in [0,1], \\ u(0) = u(1) = 0, \quad u_x(0) = -u_x(1) = \beta, \end{cases} \tag{2.20}$$

$\beta \in \mathbb{R}$, is (linearized) stable. Analogously, a global existence result follows for the initial Dirichlet boundary value problem

$$\begin{cases} \frac{v_t(t,x)}{\sqrt{1+v_x(t,x)^2}} + \frac{1}{\sqrt{1+v_x(t,x)^2}} \frac{d}{dx} \left(\frac{\kappa_{v,x}(t,x)}{\sqrt{1+v_x(t,x)^2}} \right) + \frac{1}{2}\kappa_v^3(t,x) = 0, & (t,x) \in [0,\infty) \times [0,1], \\ v(t,0) = v(t,1) = 0, \quad v_x(t,0) = -v_x(t,1) = \beta, & t \in [0,\infty), \\ v(0,x) = v_0(x), & x \in [0,1], \end{cases} \tag{2.21}$$

provided the initial datum v_0 obeys the same boundary data and is sufficiently close to the stationary solution u of (2.20) with respect to the $W^{4,p}$-norm ($p > 5$).

3. Morse index of the large solution. For $c \in (0, c_0)$ we consider as in (2.5)

$$u_c(x) = \frac{2}{c\sqrt[4]{1 + G^{-1}\left(\frac{c}{2} - cx\right)^2}} - \frac{2}{c\sqrt[4]{1 + G^{-1}\left(\frac{c}{2}\right)^2}}.$$

In order to prove Theorem 1.3, we have to show that exactly one eigenvalue of the quadratic form

$$\varphi \mapsto D^2\tilde{W}_\alpha(u_c)(\varphi,\varphi), \qquad \alpha = h(c),$$

passes through 0 when c passes through $c_{\max}$ and that for $c \in (c_{\max}, c_0)$, 0 is not an eigenvalue of $D^2\tilde{W}_\alpha(u_c)$, i.e., of (2.7). The latter was already done in Lemma 2.4. Moreover, its proof yields that there is at most one eigenvalue, which crosses 0 in $c = c_{\max}$. It remains to show that for $c > c_{\max}$ and suitable $\varphi \in H^2 \cap H_0^1(0,1)$, one indeed has $D^2\tilde{W}_\alpha(u_c)(\varphi,\varphi) < 0$. Making use of the same transformations and notations of section 2 and restricting ourselves to symmetric φ, we find

$$\begin{aligned} D^2\tilde{W}_\alpha(u_c)(\varphi,\varphi) &= 2\int_0^1 \frac{\chi'(x)^2}{\left(1 + G^{-1}\left(\frac{c}{2} - cx\right)^2\right)^{5/2}}\,dx \\ &\quad - 5c^2 \int_0^1 \frac{G^{-1}\left(\frac{c}{2} - cx\right)^2 - 1}{\left(1 + G^{-1}\left(\frac{c}{2} - cx\right)^2\right)^2}\,\chi(x)^2\,dx - 12h(c)\frac{G^{-1}\left(\frac{c}{2}\right)}{\left(1 + G^{-1}\left(\frac{c}{2}\right)^2\right)^2}\chi(1)^2 \\ &= 2c\int_{-G^{-1}\left(\frac{c}{2}\right)}^{G^{-1}\left(\frac{c}{2}\right)} \frac{\psi'(y)^2}{(1+y^2)^{5/4}}\,dy - 5c\int_{-G^{-1}\left(\frac{c}{2}\right)}^{G^{-1}\left(\frac{c}{2}\right)} \frac{(y^2-1)}{(1+y^2)^{13/4}}\psi(y)^2\,dy \\ &\quad - 12\frac{c\,G^{-1}\left(\frac{c}{2}\right)}{\left(1 + G^{-1}\left(\frac{c}{2}\right)^2\right)^{9/4}}\psi\left(G^{-1}\left(\frac{c}{2}\right)\right)^2. \end{aligned}$$

We choose

$$\psi_c(y) := (1+y^2)^{3/2}\Phi_2(y) = (1+y^2)^{5/4}G(y)$$

and obtain for the corresponding $\varphi_c \in H^2 \cap H_0^1(0,1)$

$$\begin{aligned}&\frac{1}{4c} D^2\tilde{W}_\alpha(u_c)(\varphi_c,\varphi_c)\\ &= \int_0^{G^{-1}(\frac{c}{2})} \left(\left((1+y^2)^{-1/4} + \frac{5}{2} y G(y) \right)^2 - \frac{5}{2}(y^2-1)G(y)^2 \right) \frac{dy}{(1+y^2)^{3/4}}\\ &\quad - \frac{3}{4} c^2 G^{-1}\left(\frac{c}{2}\right) \left(1 + G^{-1}\left(\frac{c}{2}\right)^2 \right)^{1/4}. \end{aligned} \tag{3.1}$$

According to Theorem 1.2, we know that this expression is equal to 0 for $c = c_{\max}$. Writing $c = 2G(d)$, we see that the asymptotic behavior of the right-hand side is dominated by

$$\frac{c_0^2}{4}\left(\frac{25}{4}\cdot\frac{2}{3} - \frac{5}{2}\cdot\frac{2}{3} - 3 \right) d^{3/2} = -\frac{c_0^2}{8} d^{3/2} \to -\infty$$

for $d \to \infty$, i.e., $c \nearrow c_0$. This shows, together with Lemma 2.4, that $D^2\tilde{W}_\alpha(u_c)(\varphi_c,\varphi_c) < 0$ for $c \in (c_{\max}, c_0)$ and concludes the proof of Theorem 1.3.

The right-hand side of (3.1) is plotted in Figure 3.1. Since $\varphi_c \to 0$ for $c \searrow 0$, the curve starts in $(0,0)$, although there, $D^2\tilde{W}_\alpha(u_0)$ is positive definite.

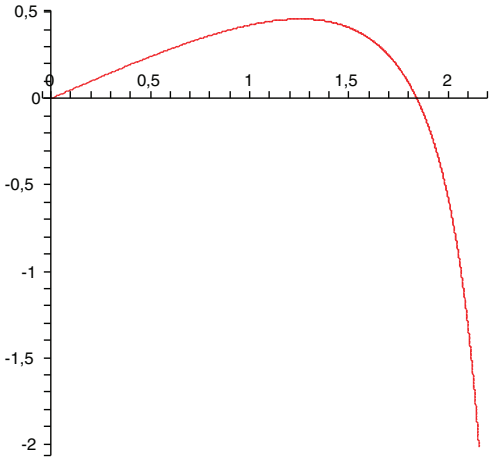


FIG. 3.1. $c \mapsto \frac{1}{4c} D^2\tilde{W}_\alpha(u_c)(\varphi_c,\varphi_c)$.

4. Global minima and symmetry. The aim of this section is to examine whether the small solutions which were found to be local minima in section 2 are also global minima for the functional $\tilde{W}_\alpha$. In what follows it will be convenient to write

$$\begin{aligned}\tilde{W}_\alpha(v) &= \int_0^1 \left(\kappa(x)^2 + 2\alpha\kappa(x)\right)\sqrt{1+v'(x)^2}\,dx,\\ &= \int_0^1 \kappa(x)^2\sqrt{1+v'(x)^2}\,dx + 2\alpha\left[\arctan(v'(x))\right]_0^1 =: W(v) + BC_\alpha(v).\end{aligned}$$

We remark that all quantities are geometric and so are invariant under rotation. Moreover, when stretching a curve by a factor k, W is multiplied by a factor $1/k$ while BC_α remains unchanged.

We shall see that the task of minimizing $\tilde{W}_\alpha$ can be reduced to a minimization problem for a function of two variables. As a byproduct of the analysis of this function we find that in order to determine $\inf_{v\in H^2\cap H_0^1}\tilde{W}_\alpha(v)$ it is sufficient to minimize over all symmetric functions. The reduction to a two-dimensional problem is achieved in two steps. We begin by showing that it is enough to consider concave functions.

LEMMA 4.1. *Suppose that $u\in H^2\cap H_0^1(0,1)$ is not concave. Then there exists a concave function $v\in H^2\cap H_0^1(0,1)$ with $\tilde{W}_\alpha(v)<\tilde{W}_\alpha(u)$.*

Proof. It is natural to think of v as the concave envelope of u, so that we are led to consider the following obstacle problem: find $v\in K$ such that

$$\forall \eta\in K,\qquad \int_0^1 v'(\eta'-v')\geq 0, \tag{4.1}$$

where $K=\{\eta\in H_0^1(0,1)\mid \eta\geq u \text{ a.e. in } (0,1)\}$. It is shown in Chapter IV of [7] that v can be obtained as the limit of a sequence $(v_\varepsilon)_{\varepsilon>0}$, where $v_\varepsilon\in H^2\cap H_0^1(0,1)$ solves

$$-v_\varepsilon''=(-u'')^+\,\vartheta_\varepsilon(v_\varepsilon-u)\quad \text{in } (0,1). \tag{4.2}$$

Here, $\vartheta_\varepsilon:\mathbb{R}\to\mathbb{R}$ satisfies

$$\vartheta_\varepsilon(t)=\begin{cases}1, & t<0,\\ 1-\frac{t}{\varepsilon}, & 0\leq t\leq\varepsilon,\\ 0, & t>\varepsilon.\end{cases}$$

It follows from the analysis in [7] that $v_\varepsilon\to v$ in $H^1(0,1)$ and $v_\varepsilon''\rightharpoonup v''$ in $L^2(0,1)$ as $\varepsilon\to 0$, so that $v\in H^2\cap H_0^1(0,1)$ and $v''\leq 0$ a.e. in $(0,1)$; in particular, v is concave. Denoting by $I=\{x\in[0,1]\mid v(x)=u(x)\}$ the coincidence set, we have that $v''=0$ a.e. in $[0,1]\setminus I$. Furthermore, using (4.2)

$$\begin{aligned}W(v)&=\int_0^1\frac{|v''|^2}{(1+(v')^2)^{\frac{5}{2}}}=\int_I\frac{|v''|^2}{(1+(u')^2)^{\frac{5}{2}}}\leq\liminf_{\varepsilon\to 0}\int_I\frac{|v_\varepsilon''|^2}{(1+(u')^2)^{\frac{5}{2}}}\\&\leq\int_I\frac{|(-u'')^+|^2}{(1+(u')^2)^{\frac{5}{2}}}\leq\int_0^1\frac{|(-u'')^+|^2}{(1+(u')^2)^{\frac{5}{2}}}\leq\int_0^1\frac{|u''|^2}{(1+(u')^2)^{\frac{5}{2}}}=W(u).\end{aligned}$$

If we had $W(v)=W(u)$, then the above argument would imply that $(-u'')^-=0$ a.e. in $(0,1)$ and therefore $u''\leq 0$ a.e. in $(0,1)$, contradicting our assumption that u is not concave. Hence $W(v)<W(u)$; since $v\geq u$ we have that $u'(0)\leq v'(0)$ and $u'(1)\geq v'(1)$, and therefore $\tilde{W}_\alpha(v)<\tilde{W}_\alpha(u)$. ∎

In what follows we shall make use of the prototype solution

$$U_0(x)=\frac{2}{c_0\sqrt[4]{1+G^{-1}\left(\frac{c_0}{2}-c_0x\right)^2}}. \tag{4.3}$$

Formally, it is the large solution of the Navier boundary value problem (1.1) for $\alpha=0$. However, one should observe that this solution is no longer smooth as a graph near $x=0$ and $x=1$, and for this reason, it was not included in Proposition 1.1.

Suppose that $0\leq x_0<x_1\leq 1$ are two points with $x_1-x_0<1$. Then $U_{0|[x_0,x_1]}$ can be written as a graph over the segment connecting $(x_0,U_0(x_0))$ and $(x_1,U_0(x_1))$. We denote by $u_{x_0,x_1}:[0,1]\to\mathbb{R}$ the strictly concave function which is obtained by translating, rotating, and rescaling the above graph to the unit interval $[0,1]$. Note

that $u_{x_0,x_1} \in H^2 \cap H_0^1(0,1)$. Our next lemma essentially reduces the minimization of $\tilde{W}_\alpha$ to a two-dimensional minimization problem.

LEMMA 4.2. *Suppose that $u \in H^2 \cap H_0^1(0,1) \setminus \{0\}$ is concave. Then there exist $0 \le x_0 < x_1 \le 1$, $x_1 - x_0 < 1$ such that $v = u_{x_0,x_1}$ satisfies $BC_\alpha(u) = BC_\alpha(v)$ and either $W(v) \le W(u)$, $u'(0) = v'(0)$, or $W(v) < W(u)$, $u'(0) \ne v'(0)$.*

Proof. Let us denote by β_ℓ and β_r the boundary angles of graph(u) on the left and on the right, respectively. Since u is assumed to be concave and nontrivial we have $\beta_\ell, \beta_r \in (0, \frac{\pi}{2})$. Consider

$$\mathcal{K} := \text{graph}(U_0) \cup \{(0,y) : y \le 0\} \cup \{(1,y) : y \le 0\}.$$

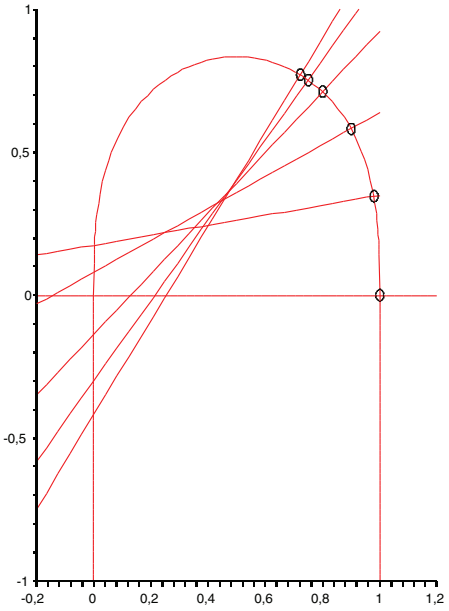


FIG. 4.1. *Left angle β_ℓ; right angle $\pi/2$.*

This is neither a graph nor a solution of the Willmore equation. However, it is a regular H^2-curve, locally an H^2-graph over the x- or the y-axis, respectively, and it has minimal Willmore energy c_0^2 among all concave curves connecting any point from $\{(0,y) : y \le 0\}$ with any point from $\{(1,y) : y \le 0\}$ with tangential directions $(0,1)$ and $(0,-1)$, respectively. This minimality follows similarly as in [4, end of section 5].

Claim. There exist two points $P = (x_P, y_P)$, $Q = (x_Q, y_Q) \in \mathcal{K}$, $P \ne Q$, such that the segment $[P,Q]$ intersects $\mathcal{K}$ under the angles β_ℓ at P and β_r at Q.

To see this, we start with the point $(x_1, y_1) = (1,0)$ and the orthogonal straight line through this point. This line intersects the left part of $\mathcal{K}$ in (x_0, y_0) under a right angle. Now we move the point (x_1, y_1) and the corresponding orthogonal straight line counterclockwise. The corresponding (x_0, y_0) finally moves down, and the intersection angle (at least finally) decreases and becomes arbitrarily small. In particular, the left angle β_ℓ is attained. See Figure 4.1. Now we keep this angle fixed and move the point (x_0, y_0) clockwise. We consider (x_1, y_1) on the right part of $\mathcal{K}$ as an intersection point with the straight line building the angle β_ℓ with $\mathcal{K}$ in (x_0, y_0). At the beginning this right angle is $\pi/2$, while it becomes arbitrarily small when (x_0, y_0) moves clockwise. In particular, β_r is attained as the angle on the right and the claim is proved.

In view of the above-mentioned minimality property of $\mathcal{K}$, $\mathcal{K}'$ enjoys a similar minimality among those arcs with boundary angles β_ℓ, β_r. We infer that

$$W(\mathcal{K}') \le \frac{1}{|P-Q|} W(u), \tag{4.4}$$

where $\mathcal{K}'$ denotes the subarc of $\mathcal{K}$ between P and Q. Observing that by construction y_P and y_Q cannot both be negative, we may distinguish two cases.

Case 1. $y_P \ge 0$ and $y_Q \ge 0$. Setting $x_0 = x_P$, $x_1 = x_Q$, we have $x_1 - x_0 < 1$ since $\beta_\ell, \beta_r \in (0, \frac{\pi}{2})$. The function $v = u_{x_0,x_1}$ then satisfies

$$W(v) = |P-Q| W(\mathcal{K}') \le W(u)$$

as well as $v'(0) = u'(0)$ and $v'(1) = u'(1)$.

Case 2. Either $y_P < 0$ or $y_Q < 0$. If $y_P < 0$, then $y_Q > 0$ since $\beta_r < \frac{\pi}{2}$, and we let $x_0 = 0$, $x_1 = x_Q$, and $v = u_{x_0,x_1}$. Denoting by $L(x_0, x_1)$ the length of the segment connecting $(x_0, U_0(x_0))$ and $(x_1, U_0(x_1))$, we have

$$W(v) = L(x_0,x_1) W(\mathcal{K}') \le \frac{L(x_0,x_1)}{|P-Q|} W(u) < W(u)$$

since $u \neq 0$ and by construction any point on $\mathrm{graph}(U_0)$ is strictly closer to $(0,0)$ than to any other point on $\{(0,y) \mid y < 0\}$. A similar argument applies if $y_Q < 0$. Finally note that while $BC_\alpha(u) = BC_\alpha(v)$, we have $u'(0) \neq v'(0)$ in this case. ∎

We deduce from Lemmas 4.1 and 4.2 that when determining $\inf \tilde{W}_\alpha$ over $H^2 \cap H_0^1(0,1)$ it is sufficient to calculate the Willmore energy for functions $v = u_{x_0,x_1}$ with $0 \le x_0 < x_1 \le 1$ and $x_1 - x_0 < 1$. The integrand for W on $[x_0, x_1]$ is c_0^2, so the integral is $c_0^2 \cdot (x_1 - x_0)$. The length of the base line is $\big((x_1-x_0)^2 + (U_0(x_1) - U_0(x_0))^2\big)^{1/2}$. As for BC_α, we have $2\alpha(\arctan(U_0'(x_1)) - \arctan(U_0'(x_0)))$. After rotation and rescaling we come up with

$$\begin{aligned}
\tilde{W}_\alpha(u_{x_0,x_1}) &= c_0^2 \cdot (x_1 - x_0) \left((x_1-x_0)^2 + (U_0(x_1) - U_0(x_0))^2\right)^{1/2} \\
&\quad + 2\alpha(\arctan(U_0'(x_1)) - \arctan(U_0'(x_0))) \\
&= c_0^2 \cdot (x_1 - x_0) \\
&\quad \cdot \left((x_1-x_0)^2 + \frac{4}{c_0^2} \left(\frac{1}{\sqrt[4]{1 + G^{-1}(c_0/2 - c_0 x_1)^2}} - \frac{1}{\sqrt[4]{1 + G^{-1}(c_0/2 - c_0 x_0)^2}} \right)^2 \right)^{1/2} \\
&\quad + 2\alpha(\arctan(G^{-1}(c_0/2 - c_0 x_1)) - \arctan(G^{-1}(c_0/2 - c_0 x_0))).
\end{aligned}$$

We now introduce the new variables

$$d_0 := G^{-1}(c_0/2 - c_0 x_0), \qquad d_1 := -G^{-1}(c_0/2 - c_0 x_1), \qquad d_1 > -d_0, \tag{4.5}$$

so that

$$x_0 = \frac{1}{2} - \frac{1}{c_0} G(d_0), \qquad x_1 = \frac{1}{2} + \frac{1}{c_0} G(d_1).$$

Defining

$$\hat{W}_\alpha(d_0, d_1) := \tilde{W}_\alpha(u_{x_0,x_1})$$

(see Figure 4.2), we end up with

$$\begin{aligned}&\hat W_\alpha(d_0,d_1)\\&= (G(d_0)+G(d_1))\left((G(d_0)+G(d_1))^2+4((1+d_0^2)^{-1/4}-(1+d_1^2)^{-1/4})^2\right)^{1/2}\\&\qquad -2\alpha(\arctan(d_0)+\arctan(d_1)).\end{aligned}$$

The following result summarizes what we have achieved so far.

THEOREM 4.3. *Let $\alpha \ge 0$. Then*

$$\inf_{v\in H^2\cap H_0^1(0,1)} \tilde W_\alpha(v) = \inf_{(d_0,d_1)\in\mathbb{R}^2,\, d_1\ge -d_0} \hat W_\alpha(d_0,d_1).$$

Let us remark that ideas similar to those employed to obtain Theorem 4.3 were used in [2] in order to prove an existence result for axially symmetric Willmore surfaces satisfying Dirichlet boundary conditions. The corresponding Navier problem, however, is still open.

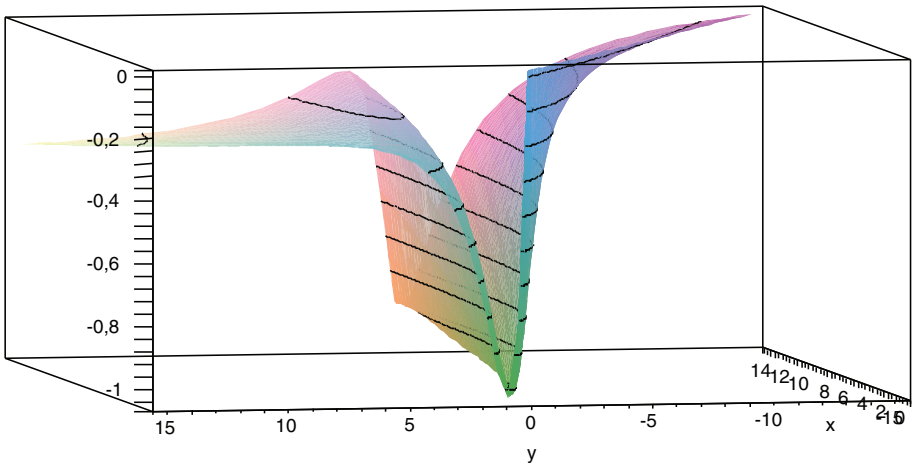


FIG. 4.2. *Cross section of the graph of $\hat W_1$ along the axis $d_0 = d_1$.*

It remains to discuss the two-dimensional function $\hat W_\alpha(d_0,d_1)$, $(d_1 \ge -d_0)$. Here, the key step is proving positivity for the following expression.

LEMMA 4.4. *For $d_1 > -d_0$ we have that*

$$\begin{aligned}&(G(d_0)+G(d_1))\cdot\left(\frac{d_0}{\sqrt[4]{1+d_0^2}}+\frac{d_1}{\sqrt[4]{1+d_1^2}}\right)-(G(d_0)+G(d_1))^2\\&\qquad -2\left(\frac{1}{\sqrt[4]{1+d_0^2}}-\frac{1}{\sqrt[4]{1+d_1^2}}\right)^2>0.\end{aligned}$$

Proof. By the fundamental theorem of calculus and since G is odd, we have

$$G(d_0)+G(d_1) = G(d_0)-G(-d_1) = \int_{-d_1}^{d_0} \frac{1}{(1+\tau^2)^{5/4}}\,d\tau,$$

$$\frac{d_0}{\sqrt[4]{1+d_0^2}}+\frac{d_1}{\sqrt[4]{1+d_1^2}} = \left[\frac{\tau}{(1+\tau^2)^{1/4}}\right]_{-d_1}^{d_0} = \int_{-d_1}^{d_0}\frac{1+\frac12\tau^2}{(1+\tau^2)^{5/4}}\,d\tau,$$

$$\frac{1}{\sqrt[4]{1+d_0^2}} - \frac{1}{\sqrt[4]{1+d_1^2}} = \left[\frac{1}{(1+\tau^2)^{1/4}}\right]_{-d_1}^{d_0} = -\frac{1}{2}\int_{-d_1}^{d_0} \frac{\tau}{(1+\tau^2)^{5/4}}\,d\tau.$$

One may observe that $d_1 > -d_0$ is equivalent to $-d_1 < d_0$. The first two terms in the expression under consideration combine as follows:

$$\begin{aligned}&(G(d_0)+G(d_1))\cdot\left(\frac{d_0}{\sqrt[4]{1+d_0^2}} + \frac{d_1}{\sqrt[4]{1+d_1^2}}\right) - (G(d_0)+G(d_1))^2\\ &= \frac{1}{2}\left(\int_{-d_1}^{d_0} \frac{1}{(1+\tau^2)^{5/4}}\,d\tau\right)\cdot\left(\int_{-d_1}^{d_0} \frac{\tau^2}{(1+\tau^2)^{5/4}}\,d\tau\right).\end{aligned}$$

We now apply the Cauchy–Schwarz inequality and make use of $\tau \mapsto \frac{1}{(1+\tau^2)^{5/8}}$ and $\tau \mapsto \frac{\tau}{(1+\tau^2)^{5/8}}$ being linearly independent to obtain

$$\begin{aligned}2\left(\frac{1}{\sqrt[4]{1+d_0^2}} - \frac{1}{\sqrt[4]{1+d_1^2}}\right)^2 &= \frac{1}{2}\left(\int_{-d_1}^{d_0} \frac{\tau}{(1+\tau^2)^{5/4}}\,d\tau\right)^2\\ &< \frac{1}{2}\left(\int_{-d_1}^{d_0} \frac{1}{(1+\tau^2)^{5/4}}\,d\tau\right)\cdot\left(\int_{-d_1}^{d_0} \frac{\tau^2}{(1+\tau^2)^{5/4}}\,d\tau\right)\\ &= (G(d_0)+G(d_1))\cdot\left(\frac{d_0}{\sqrt[4]{1+d_0^2}} + \frac{d_1}{\sqrt[4]{1+d_1^2}}\right) - (G(d_0)+G(d_1))^2,\end{aligned}$$

thereby proving the claim. □

Next we show that in the open interior of the domain of definition of the two-dimensional energy function $\hat{W}_\alpha$, critical points may occur at most on the diagonal, i.e., on symmetric graphs in the original context.

LEMMA 4.5. *Let $\alpha \geq 0$ and assume that*

$$\begin{aligned}&(d_0,d_1) \mapsto \hat{W}_\alpha(d_0,d_1)\\ &= (G(d_0)+G(d_1))\cdot\left((G(d_0)+G(d_1))^2 + 4\left((1+d_0^2)^{-1/4} - (1+d_1^2)^{-1/4}\right)^2\right)^{1/2}\\ &\quad - 2\alpha\left(\arctan(d_0)+\arctan(d_1)\right)\end{aligned}$$

has a critical point (d_0,d_1) with $d_1 > -d_0$. Then

$$d_0 = d_1.$$

Proof. In a critical point of $\hat{W}_\alpha$, we have that

$$\begin{aligned}0 &= \frac{\partial}{\partial d_0}\hat{W}_\alpha(d_0,d_1)\\ &= \frac{1}{2}\left(G(d_0)+G(d_1)\right)\cdot\left((G(d_0)+G(d_1))^2 + 4\left((1+d_0^2)^{-1/4} - (1+d_1^2)^{-1/4}\right)^2\right)^{-1/2}\\ &\quad\cdot\left(2(G(d_0)+G(d_1))(1+d_0^2)^{-5/4} - 4d_0(1+d_0^2)^{-5/4}\left((1+d_0^2)^{-1/4} - (1+d_1^2)^{-1/4}\right)\right)\\ &\quad + (1+d_0^2)^{-5/4}\left((G(d_0)+G(d_1))^2 + 4\left((1+d_0^2)^{-1/4} - (1+d_1^2)^{-1/4}\right)^2\right)^{1/2} - \frac{2\alpha}{1+d_0^2};\end{aligned}$$

$$
\begin{aligned}
0 &= \frac{\partial}{\partial d_1}\hat{W}_\alpha(d_0,d_1)\\
&= \frac{1}{2}\,(G(d_0)+G(d_1))\cdot\left((G(d_0)+G(d_1))^2+4\left((1+d_0^2)^{-1/4}-(1+d_1^2)^{-1/4}\right)^2\right)^{-1/2}\\
&\quad\cdot\left(2(G(d_0)+G(d_1))(1+d_1^2)^{-5/4}+4d_1(1+d_1^2)^{-5/4}\left((1+d_0^2)^{-1/4}-(1+d_1^2)^{-1/4}\right)\right)\\
&\quad+(1+d_1^2)^{-5/4}\left((G(d_0)+G(d_1))^2+4\left((1+d_0^2)^{-1/4}-(1+d_1^2)^{-1/4}\right)^2\right)^{1/2}\\
&\quad-2\alpha\frac{1}{1+d_1^2}.
\end{aligned}
\tag{4.6}
$$

Equivalently,

$$
\begin{aligned}
0 &= (G(d_0)+G(d_1))\cdot\left((G(d_0)+G(d_1))^2+4\left((1+d_0^2)^{-1/4}-(1+d_1^2)^{-1/4}\right)^2\right)^{-1/2}\\
&\quad\cdot\left((G(d_0)+G(d_1))(1+d_0^2)^{-1/4}-2d_0(1+d_0^2)^{-1/4}\left((1+d_0^2)^{-1/4}-(1+d_1^2)^{-1/4}\right)\right)\\
&\quad+(1+d_0^2)^{-1/4}\left((G(d_0)+G(d_1))^2+4\left((1+d_0^2)^{-1/4}-(1+d_1^2)^{-1/4}\right)^2\right)^{1/2}-2\alpha;\\
0 &= (G(d_0)+G(d_1))\cdot\left((G(d_0)+G(d_1))^2+4\left((1+d_0^2)^{-1/4}-(1+d_1^2)^{-1/4}\right)^2\right)^{-1/2}\\
&\quad\cdot\left((G(d_0)+G(d_1))(1+d_1^2)^{-1/4}+2d_1(1+d_1^2)^{-1/4}\left((1+d_0^2)^{-1/4}-(1+d_1^2)^{-1/4}\right)\right)\\
&\quad+(1+d_1^2)^{-1/4}\left((G(d_0)+G(d_1))^2+4\left((1+d_0^2)^{-1/4}-(1+d_1^2)^{-1/4}\right)^2\right)^{1/2}-2\alpha.
\end{aligned}
$$

Subtracting both equations yields

$$
\begin{aligned}
0 &= \left((1+d_0^2)^{-1/4}-(1+d_1^2)^{-1/4}\right)\\
&\quad\cdot\left((G(d_0)+G(d_1))^2+4\left((1+d_0^2)^{-1/4}-(1+d_1^2)^{-1/4}\right)^2\right)^{-1/2}\\
&\quad\cdot\left\{(G(d_0)+G(d_1))^2-2(G(d_0)+G(d_1))\left(\frac{d_0}{\sqrt[4]{1+d_0^2}}+\frac{d_1}{\sqrt[4]{1+d_1^2}}\right)\right.\\
&\qquad\left.+\left((G(d_0)+G(d_1))^2+4\left((1+d_0^2)^{-1/4}-(1+d_1^2)^{-1/4}\right)^2\right)\right\}.
\end{aligned}
$$

By Lemma 4.4, the curly bracket is strictly negative, since we assume that $d_1 > -d_0$. We conclude that

$$
0 = \left((1+d_0^2)^{-1/4}-(1+d_1^2)^{-1/4}\right),
$$

which yields that $d_0 = d_1$. □

We are now in position to solve the two-dimensional minimization problem.

PROPOSITION 4.6. *Let $0 < \alpha \le \alpha_{\max}$. There exists $\alpha^* = 1.132372323\ldots \in (0,\alpha_{\max})$ such that*

$$\inf_{(d_0,d_1)\in\mathbb{R}^2,\, d_1\geq -d_0} \hat{W}_\alpha(d_0,d_1) = \begin{cases} \hat{W}_\alpha(G^{-1}(\frac{c}{2}), G^{-1}(\frac{c}{2})), & 0<\alpha\leq\alpha^*, \\ c_0^2 - 2\alpha\pi, & \alpha^*<\alpha\leq\alpha_{\max}, \end{cases}$$

where $c\in(0,c_{\max})$ solves $h(c)=\alpha$. In the first case $d_0=d_1=G^{-1}(\frac{c}{2})$ is the only point for which the minimum is attained, while it is not attained for $\alpha^*<\alpha\leq\alpha_{\max}$.

Proof. In view of Lemma 4.5 and the symmetry of $\hat{W}_\alpha$,

$$\inf_{(d_0,d_1)\in\mathbb{R}^2,\, d_1\geq -d_0} \hat{W}_\alpha(d_0,d_1)$$

is the minimum between

$$\inf_{d\in(0,\infty)} \hat{W}_\alpha(d,d), \tag{4.7}$$

$$\inf_{d\in\mathbb{R}} \hat{W}_\alpha(d,-d) = 0, \tag{4.8}$$

and

$$\inf_{d\in\mathbb{R}} \hat{W}_\alpha(d,\infty). \tag{4.9}$$

Since

$$\hat{W}_\alpha(d,d) = 4G(d)^2 - 4\alpha\arctan(d)$$

is certainly negative for $d>0$ close to 0, we see that $\inf_{d\in(0,\infty)}\hat{W}_\alpha(d,d)<0$, so we need not consider (4.8). As for (4.9) we have

$$\hat{W}_\alpha(d,\infty) = \Big(G(d)+\frac{c_0}{2}\Big)\cdot\left(\Big(G(d)+\frac{c_0}{2}\Big)^2 + 4(1+d^2)^{-1/2}\right)^{1/2} - 2\alpha\Big(\arctan(d)+\frac{\pi}{2}\Big).$$

It is sufficient to discuss local mimima, since $\hat{W}_\alpha(\infty,\infty)$ is already covered by (4.7) and $\hat{W}_\alpha(-\infty,\infty)=0$ by (4.8). Passing to the $c=2G(d)$-variable, we see that $\hat{W}_\alpha$ attains its minimum on $\{(d_0,d_1): d_0\in[-\infty,\infty],\, d_1\in[-d_0,\infty]\}$. For fixed $d_0\in\mathbb{R}$, we infer from (4.6) that for d_1 large enough, $\frac{\partial\hat{W}_\alpha}{\partial d_1}>0$. This follows since the slowest term $4d_1(1+d_1^2)^{-5/4}(1+d_0^2)^{-1/4}$ decays of order $-3/2$ and has a positive coefficient. Hence, the minimum is not attained on $\mathbb{R}\times\{\infty\}$ but either in (∞,∞) or in the interior of our domain. This proves that

$$\inf_{(d_0,d_1)\in\mathbb{R}^2,\, d_1\geq -d_0} \hat{W}_\alpha(d_0,d_1) = \inf_{d\in(0,\infty)} \hat{W}_\alpha(d,d). \tag{4.10}$$

It remains to evaluate the right-hand side of (4.10). Let

$$\phi(d) := \hat{W}_\alpha(d,d) = 4G(d)^2 - 4\alpha\arctan(d).$$

We have

$$\phi'(d) = \frac{8G(d)}{(1+d^2)^{\frac{5}{4}}} - \frac{4\alpha}{1+d^2} = \frac{4}{1+d^2}\big(h(2G(d))-\alpha\big),$$

with h defined in (2.6). Thus, $\phi'(d)=0$ iff $d=G^{-1}(\frac{c}{2})$, where c is one of the solutions of $h(c)=\alpha$. Only the solution $c\in(0,c_{\max})$ is a local minimum so that

$$\inf_{d\in(0,\infty)}\phi(d) = \min\left(c^2 - 4\alpha\arctan\left(G^{-1}\left(\frac{c}{2}\right)\right), c_0^2 - 2\alpha\pi\right);$$

we take into account that $\phi(0) = 0$ and $\phi(d) < 0$ for small $d > 0$. In order to calculate the last minimum we introduce the following auxiliary function $f : [0, c_{\max}] \to \mathbb{R}$:

$$f(c) := c_0^2 - 2h(c)\pi - c^2 + 4h(c) \arctan G^{-1}\left(\frac{c}{2}\right).$$

We find that $f(0) = c_0^2 > 0$ and $f(c_{\max}) = -0.6674542140\ldots < 0$, and a short calculation shows that

$$f'(c) = \left(4 \arctan G^{-1}\left(\frac{c}{2}\right) - 2\pi\right) h'(c) < 0, \quad c \in (0, c_{\max}),$$

so that f has a unique zero

$$c^* = 1.274998908\ldots \in [0, c_{\max}] \text{ with } \alpha^* := h(c^*) = 1.132372323\ldots.$$

This proves the formula for $\inf_{(d_0,d_1)\in\mathbb{R}^2,\, d_1 \geq -d_0} \hat{W}_\alpha(d_0, d_1)$. The uniqueness of the minimum for $0 \leq \alpha \leq \alpha^*$ follows from Lemma 2.2. □

We are now in position to prove Theorems 1.4 and 1.5. The second result is an immediate consequence of (4.10) and Theorem 4.3. As for Theorem 1.4, we focus on the case $0 < \alpha \leq \alpha^*$. Let $c \in (0, c_{\max})$ be the unique solution of $h(c) = \alpha$ with corresponding small solution u_c. Clearly,

$$\begin{aligned}\tilde{W}_\alpha(u_c) = \hat{W}_\alpha\left(G^{-1}\left(\frac{c}{2}\right), G^{-1}\left(\frac{c}{2}\right)\right) &= \inf_{(d_0,d_1)\in\mathbb{R}^2,\, d_1 \geq -d_0} \hat{W}_\alpha(d_0, d_1) \\ &= \inf_{v \in H^2 \cap H_0^1(0,1)} \tilde{W}_\alpha(v)\end{aligned}$$

by Proposition 4.6 and Lemma 4.2. It remains to show that u_c is the only function in $H^2 \cap H_0^1(0, 1)$ for which the minimum is attained. Suppose that $u \in H^2 \cap H_0^1(0, 1)$ satisfies $\tilde{W}_\alpha(u) = \inf_{v\in H^2\cap H_0^1(0,1)} \tilde{W}_\alpha(v)$. In view of Lemma 4.1, u is necessarily concave. Let $v = u_{x_0,x_1} \in H^2 \cap H_0^1(0, 1)$ be the function appearing in Lemma 4.2 with d_0, d_1 given by (4.5). Using the minimality of u, Proposition 4.6, and Lemma 4.2, we obtain

$$\tilde{W}_\alpha(u) \leq \tilde{W}_\alpha(u_c) = \hat{W}_\alpha\left(G^{-1}\left(\frac{c}{2}\right), G^{-1}\left(\frac{c}{2}\right)\right) \leq \hat{W}_\alpha(d_0, d_1) = \tilde{W}_\alpha(v) \leq \tilde{W}_\alpha(u).$$

This implies that $\hat{W}_\alpha(G^{-1}(\frac{c}{2}), G^{-1}(\frac{c}{2})) = \hat{W}_\alpha(d_0, d_1)$ and hence by Proposition 4.6 that $d_0 = d_1 = G^{-1}(\frac{c}{2})$ so that $v = u_c$. In particular, we infer with the help of Lemma 4.2 that $u'(0) = v'(0) = u_c'(0)$ and $u'(1) = v'(1) = u_c'(1)$. As a consequence we have $BC_\alpha(u) = BC_\alpha(u_c)$, and therefore $W(u) = W(u_c)$. However, in view of Theorem 2 in [4], u_c is the *unique* minimum of W in the class $M_\beta = \{w \in H^2 \cap H_0^1(0, 1) \mid w'(0) = -w'(1) = \beta\}$ ($\beta = u_c'(0)$) so that we must have $u = u_c$. This completes the proof of Theorem 1.4. □

For selected values of α, Figure 4.3 shows plots of the function $c \mapsto \tilde{W}_\alpha(u_c)$ on the interval $[0, c_0)$.

The uniqueness part of Theorem 1.4 guarantees that a global minimizer of $\tilde{W}_\alpha$ is symmetric. A more difficult question is whether any solution of (1.1) is necessarily symmetric. The following example (see also [12, p. 461]) shows that such a result certainly does not hold if one extends the class of admissible functions to include graphs with singularities in their first derivatives. Let

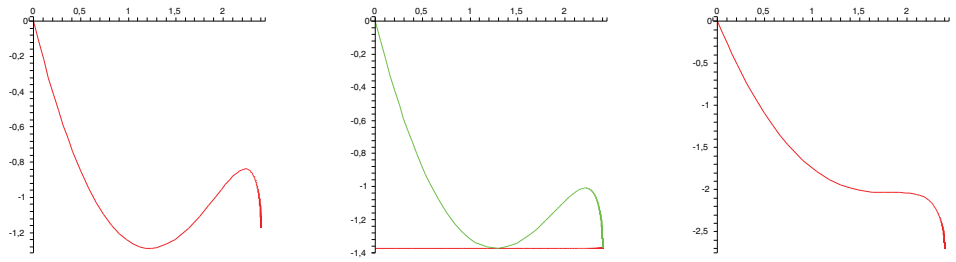


FIG. 4.3. *Graphs of the function* $c \mapsto \tilde{W}_\alpha(u_c)$ *for* $\alpha = 1.1$, $\alpha = \alpha^*$, *and* $\alpha = 1.34$ *(left to right).*

$$u(x) := \begin{cases} \frac{1}{2}U_0(2x), & 0 \le x \le \frac{1}{2}, \\ -\frac{1}{2}U_0(2-2x), & \frac{1}{2} \le x \le 1, \end{cases}$$

where the function U_0 was defined in (4.3). Clearly, as a curve in $\mathbb{R}^2$, the graph of u is a nonsymmetric solution of (1.1) for $\alpha = 0$. Note, however, that $u \notin H^2(0,1) \cap H_0^1(0,1)$ so that u is not a smooth solution of (1.1).

Acknowledgment. The authors thank N. Masel (Minsk) for pointing out an error in an earlier version of Lemma 2.3.

REFERENCES

[1] M. Bauer and E. Kuwert, *Existence of minimizing Willmore surface of prescribed genus*, Int. Math. Res. Not., 2003, No. 10 (2003), pp. 553–576.
[2] A. Dall'Acqua, K. Deckelnick, and H.-Ch. Grunau, *Classical solutions to the Dirichlet problem for Willmore surfaces of revolution*, Adv. Calc. Var., to appear.
[3] K. Deckelnick and G. Dziuk, *Error analysis of a finite element method for the Willmore flow of graphs*, Interfaces Free Bound., 8 (2006), pp. 21–46.
[4] K. Deckelnick and H.-Ch. Grunau, *Boundary value problems for the one-dimensional Willmore equation*, Calc. Var. Partial Differential Equations, 30 (2007), pp. 293–314.
[5] G. Dziuk, E. Kuwert, and R. Schätzle, *Evolution of elastic curves in* $\mathbb{R}^n$*: Existence and computation*, SIAM J. Math. Anal., 33 (2002), pp. 1228–1245.
[6] L. Euler, *Opera Omnia*, Ser. 1, 24, Orell Füssli, Zürich, 1952.
[7] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Classics in Appl. Math. 31, SIAM, Philadelphia, 2000.
[8] E. Kuwert and R. Schätzle, *The Willmore flow with small initial energy*, J. Differential Geom., 57 (2001), pp. 409–441.
[9] E. Kuwert and R. Schätzle, *Gradient flow for the Willmore functional*, Comm. Anal. Geom., 10 (2002), pp. 307–339.
[10] E. Kuwert and R. Schätzle, *Removability of point singularities of Willmore surfaces*, Ann. of Math. (2), 160 (2004), pp. 315–357.
[11] Yu. Latushkin, J. Prüss, and R. Schnaubelt, *Stable and unstable manifolds for quasilinear parabolic systems with fully nonlinear boundary conditions*, J. Evol. Equ., 6 (2006), pp. 537–576.
[12] A. Linnér, *Explicit elastic curves*, Ann. Global Anal. Geom., 16 (1998), pp. 445–475.
[13] U. F. Mayer and G. Simonett, *A numerical scheme for axisymmetric solutions of curvature-driven free boundary problems, with applications to the Willmore flow*, Interfaces Free Bound., 4 (2002), pp. 89–109.
[14] J. C. C. Nitsche, *Boundary value problems for variational integrals involving surface curvatures*, Quart. Appl. Math., 51 (1993), pp. 363–387.
[15] A. Polden, *Curves and Surfaces of Least Total Curvature and Fourth-Order Flows*, Ph.D. thesis, University of Tübingen, Tübingen, Germany, 1996.
[16] R. Schätzle, *The Willmore Boundary Value Problem*, preprint, University of Tübingen, Tübingen, Germany, 2006.

[17] L. Simon, *Existence of surfaces minimizing the Willmore functional*, Comm. Anal. Geom., 1 (1993), pp. 281–326.
[18] G. Simonett, *The Willmore flow near spheres*, Differential Integral Equations, 14 (2001), pp. 1005–1014.
[19] T. J. Willmore, *Total Curvature in Riemannian Geometry*, Ellis Horwood Ser. Math. Appl., Ellis Horwood Limited, Chichester, Halsted Press, New York, 1982.

SIAM J. MATH. ANAL.
Vol. 40, No. 5, pp. 2077–2104

ON THE ENERGY OF SUPERCONDUCTORS IN LARGE AND SMALL DOMAINS*

MATTHIAS KURZKE† AND DANIEL SPIRN‡

Abstract. We study the Ginzburg–Landau energy functional for superconductors in an applied magnetic field. We focus on asymptotically large or small domains and establish the asymptotic behavior of the energy as a function of the Ginzburg–Landau parameter, applied magnetic field, and domain size. For a large class of domain sizes, we calculate the critical field strength where vortex nucleation becomes energetically favorable, and describe the vorticity of minimizers. For supercritical magnetic field strengths, we recover the energy of a classical Abrikosov vortex lattice. Our findings generalize several known results of Sandier and Serfaty for domains of fixed size.

Key words. Ginzburg–Landau, vortices, asymptotics, critical field

AMS subject classifications. 82D55, 35B25, 35J60

DOI. 10.1137/070687992

1. Introduction.

In the widely studied Ginzburg–Landau model, superconducting materials are modeled via the free energy functional

(1.1)
$$G_{phys}(\psi, \hat{A}) = G_0 + \int_\Omega \frac{|\operatorname{curl} \hat{A} - H_{ex}|^2}{8\pi} + \frac{1}{2m_*} \left| \left(\hbar \nabla - \frac{ie_* \hat{A}}{c} \right) \psi \right|^2 + \alpha |\psi|^2 + \frac{\beta}{2} |\psi|^4,$$

where Ω is the region occupied by the superconductor, ψ a complex-valued order parameter, $\hat{A}$ the vector potential of the magnetic field, H_{ex} an applied magnetic field, e_* and m_* physical constants corresponding to the charge and mass of superconducting carriers, c the speed of light, and $\hbar$ the Planck constant. G_0 is the energy of the normal state and independent of $(\psi, \hat{A})$. The quantities α and β are temperature-dependent constants. We will assume subcritical temperatures, so $\alpha < 0$ and $\beta > 0$. An introduction to the physics of (1.1) explaining the meaning of these constants can be found in [26].

We will work with only two-dimensional domains (corresponding to cylindrical symmetry in the x_3-direction) and will assume that the domain size is s. As we want to study regimes in which the domain size becomes small or large, we renormalize by assuming that $U = \frac{\Omega}{s}$ is a fixed domain of size $O(1)$.

Setting $u(x) = \sqrt{\frac{\beta}{|\alpha|}} \psi(sx)$, $A(x) = \frac{e_* s}{\hbar c} \hat{A}(sx)$, and $h_{ex} = \frac{e_*}{\hbar c} H_{ex}$, we rewrite the energy as

$$G_0 + K \left(\frac{1}{2} \int_U \frac{1}{\ell^2} |\operatorname{curl} A - h_{ex}|^2 + \frac{1}{2\varepsilon^2} (1 - |u|^2)^2 + |(\nabla - iA)u|^2 \right), \tag{1.2}$$

where $K = \frac{\hbar^2 |\alpha|}{2m_* \beta}$, $\ell^2 = \frac{2\pi s^2 e_*^2 |\alpha|}{m_* \beta c^2}$, and $\varepsilon^2 = \frac{\hbar^2}{4m_* |\alpha| s^2}$.

*Received by the editors April 11, 2007; accepted for publication (in revised form) September 8, 2008; published electronically January 23, 2009.
http://www.siam.org/journals/sima/40-5/68799.html

†Institut für angewandte Mathematik, Universität Bonn, Germany D-53115 (kurzke@iam.uni-bonn.de).

‡School of Mathematics, University of Minnesota, Minneapolis, MN 55455 (spirn@math.umn.edu). This author was supported in part by NSF grant DMS-0510121.

Dropping K and G_0 by an affine transformation of the energy, we arrive at the functional

$$G_\varepsilon(u,A) = \frac{1}{2}\int_U |\nabla_A u|^2 + \frac{1}{\ell^2}\,|\operatorname{curl} A - h_{ex}|^2 + \frac{1}{2\varepsilon^2}\left(1-|u|^2\right)^2 dx. \tag{1.3}$$

Using the commonly used comparison length scales of *penetration depth*,

$$\lambda = \sqrt{\frac{m_*\beta c^2}{4\pi|\alpha|e_*^2}}, \tag{1.4}$$

and *coherence length*,

$$\xi = \frac{\hbar}{\sqrt{2m_*|\alpha|}}, \tag{1.5}$$

we find that our dimensionless parameters ε and ℓ can be expressed as

$$\varepsilon = \frac{\xi}{\sqrt{2}s}, \quad \ell = \frac{s}{\sqrt{2}\lambda}. \tag{1.6}$$

We will consider ε small (i.e., domains of size $s \gg \xi$) and various scalings of ℓ.

We will study *type II* superconductors, those where the Ginzburg–Landau parameter $\kappa = \frac{\lambda}{\xi}$ is large, so the coherence length will be the smallest of the three physical length scales. This corresponds to $\ell \ll \frac{1}{\varepsilon}$, which will be implied by other assumptions throughout this article. Under this limit minimizers will start to energetically favor the formation of vortices once the applied magnetic field grows large enough.

The asymptotics of (1.3) with $\ell \equiv 1$, i.e.,

$$\frac{1}{2}\int_U |\nabla_A u|^2 + |\operatorname{curl} A - h_{ex}|^2 + \frac{1}{2\varepsilon^2}\left(1-|u|^2\right)^2 dx \tag{1.7}$$

under the asymptotic limit $\varepsilon \to 0$, has been widely studied in the past decade and a half.

1.1. Background results. In the last 15 years, there has been considerable progress made in the mathematical understanding of the Ginzburg–Landau model. A major step has been the groundbreaking work of Bethuel, Brezis, and Hélein [5] on the related functional without gauge term, the so-called BBH functional

$$E_\varepsilon(u) = \frac{1}{2}\int_U |\nabla u|^2 + \frac{1}{2\varepsilon^2}(1-|u|^2)^2, \tag{1.8}$$

and much of the analysis for the full gauge-invariant functional G_ε is based on analysis of E_ε. We can only sketch some of the developments for the static Ginzburg–Landau model with magnetic field, and refer the reader to the recent monograph [18] by Sandier and Serfaty, which contains a thorough treatment of vortex solutions and critical fields for vortex nucleation for (1.7).

We mention some works that are of particular relevance to the topic of this article. The first rigorous treatment of (1.7) in the $\varepsilon \to 0$ limit can be found in Bethuel and Riviere [6], who discovered many important features of the standard Ginzburg–Landau functional. Serfaty [22, 23] built on this work and gave the first rigorous treatment of the critical field question by a study of local minimizers close to the critical field.

The technique used assumptions on the BBH energy (1.8) to obtain an a priori bound on the number of vortices. Using the "vortex ball construction" of Sandier [17] and Jerrard [10], a key ingredient in most of the later research, Sandier and Serfaty [21] were then able to show that the global minimizer below the critical field is indeed vortex-free and $|u_\varepsilon|$ is bounded away from zero.

The structure of global minimizers with an unbounded number of vortices and with external field of order $h_{ex} = O(|\log \varepsilon|)$ was analyzed by Sandier and Serfaty in [20]. This result, combined with a Jacobian compactness theorem [11], was rephrased by Jerrard and Soner [12] in the framework of Γ-convergence. The limit problem is equivalent to a certain obstacle problem, and the limiting vorticity (after rescaling with $|\log \varepsilon|$) is constant in the set where the obstacle is active.

For asymptotically larger applied magnetic fields ($|\log \varepsilon| \ll h_{ex} \ll \frac{1}{\varepsilon^2}$), the vortices fill the whole domain as an Abrikosov-type lattice with uniform limiting density of vortices. This was established by Sandier and Serfaty in [19].

There are few results as of yet on the influence of the domain size on the behavior of the functional. The asymptotic expansion results of Chapman et al. [8] contain a scaling argument similar to ours above, and then study domains of size $s = O(\xi)$ with $\kappa \gg 1$, which corresponds to very small domains.

Aftalion and Dancer [1] studied critical points of the Ginzburg–Landau energy. For small domains ($\ell < C \min(1, \frac{C}{\kappa})$), they showed that any solution that is not the normal solution (where $u \equiv 0$) will be bounded away from zero, regardless of the external field. For the special case where the domain is a ball, $U_\ell = B_\ell(0)$, they showed that solutions in small domains are necessarily radially symmetric, and there exists a critical field of order $O(\frac{1}{\ell})$ such that, above this field, only the normal solution exists, while a unique superconducting solution exists below this threshold.

A numerical study was performed by Aftalion and Du [2], who studied the response of a superconductor to the raising and lowering of the external field depending on Ginzburg–Landau parameter κ and domain size. They found bifurcation diagrams in several distinct regimes, including a critical line separating type I and type II behavior.

Recently, there have been results by Aydi [3] and Aydi and Sandier [4] on minimizing sequences of the Ginzburg–Landau energy in periodic boxes. In these papers the authors establish a detailed description of the critical field strength for vortex nucleation; vortices are located at minimizers of an explicit renormalized energy. Furthermore, the limiting induced magnetic field satisfies a periodic London equation with point sources at the vortex locations. These results can be understood as a study of the interior behavior of an Abrikosov lattice in the large domain limit.

There have also been a few results that study (1.7) with applied magnetic fields and domain-dependence between h_{c_2} and h_{c_3}, the regime associated with surface superconductivity. However, we restrict ourselves to field strengths asymptotically below h_{c_2}; hence we do not attempt to review results within this class of field strengths.

There are similarities between the Ginzburg–Landau energy (1.3) and the Chern–Simons–Higgs energy

$$G_{csh}(u, A) = \frac{1}{2} \int_U |\nabla_A u|^2 + \frac{\mu^2}{4} \frac{|\operatorname{curl} A - h_{ex}|^2}{|u|^2} + \frac{1}{\varepsilon^2} |u|^2 \left(1 - |u|^2\right)^2 dx \tag{1.9}$$

for an applied magnetic field, h_{ex}, and a bounded simply connected domain, $U \subset \mathbb{R}^2$. The Chern–Simons–Higgs model is an anyon theory that is of interest in connection with high-temperature superconductors and the quantum Hall effect. For an overview

of the study in the self-dual case $\mu = \varepsilon$, see Yang [27].

In [14, 15] the authors proved several results of a similar nature to those found here: For $h_{ex} = H|\log\varepsilon|$ and $G_{csh}(u_\varepsilon, A_\varepsilon) = O(|\log\varepsilon|^2)$, we were able to show Γ-convergence results for the cases $\mu = \mu_\varepsilon \to \mu_0 \in (0, \infty]$. These enabled us to calculate the critical field for vortex nucleation. The main ingredient in these results is a compactness proof that relates the Jacobian of u, $J(u) = \det \nabla u$, to the energy

$$E_{csh}(u) = \frac{1}{2}\int_U |\nabla u|^2 + \frac{1}{\varepsilon^2}|u|^2(1-|u|^2).$$

Using this compactness result from [14] and an energy decomposition, we showed Γ-convergence for finite μ in [14] and for $\mu \to \infty$ in [15]. For $\mu \to 0$, we gave an explicit counterexample that illustrates why this method, using a decomposition and bounds for E_{csh}, fails in this case. However, we were later able to show that for h_{ex} much larger than the critical field, and under certain restrictions on μ, the energy of minimizers scales in the same fashion as the energy of an Abrikosov-type lattice just as for the Ginzburg–Landau energy (1.7); see [16].

All of our results here carry over from the Chern–Simons–Higgs energy (1.9) under the assumptions above. In particular, we can extend the results of [14, 15, 16, 25] and understand vortices in non–self-dual Chern–Simons–Higgs for a wider range of parameters and in more detail. Results for (1.3) in the next subsection can be related to results for (1.9) by simply setting $\ell = \frac{2}{\mu}$.

1.2. Main results. In this subsection, we list our main theorems on the behavior of minimizers for various parameter regimes. These results, most of which are generalizations of known results from the last section, provide a partial solution to Open Problem 1 of [18].

Our first result is the calculation of the first critical field where minimizers of the Ginzburg–Landau energy start to have vortices. This field is $O(|\log\varepsilon|)$ if the domains stay bounded and $O(\ell^2|\log\varepsilon|)$ if the domains are unbounded and ℓ is bounded by a power of $|\log\varepsilon|$.

THEOREM 1.1. *There exists a sequence of critical fields $h_{c_1}(\varepsilon)$ such that any minimizer of the Ginzburg–Landau functional with $h_{ex} < h_{c_1}(\varepsilon) - o(|\log\varepsilon|)$ is vortex-free, while any minimizer with $h_{ex} > h_{c_1}(\varepsilon) + o(|\log\varepsilon|)$ has vortices.*

As $\varepsilon \to 0$, the critical field $h_{c_1}(\varepsilon)$ satisfies the following expansion: If $\ell_\varepsilon \to \ell_0$ with $0 \le \ell_0 < \infty$, then

$$\frac{h_{c_1}(\varepsilon)}{|\log\varepsilon|} \to H_1(\ell_0), \tag{1.10}$$

where

$$H_1(\ell_0) = \frac{1}{2\max_{\overline{U}} |y_{\ell_0}|}, \tag{1.11}$$

where y_{ℓ_0} is the solution of

$$-\Delta y_{\ell_0} + \ell_0^2 y_{\ell_0} + 1 = 0$$

with Dirichlet boundary conditions $y_{\ell_0} = 0$ on ∂U.

Finally, if $\ell_\varepsilon \to \infty$ and $\ell_\varepsilon \le |\log\varepsilon|^\gamma$ for any fixed $\gamma > 0$, then

$$\frac{h_{c_1}(\varepsilon)}{\ell_\varepsilon^2|\log\varepsilon|} \to \frac{1}{2} \tag{1.12}$$

as $\varepsilon \to 0$. Therefore, the critical field scales as $h_{c_1} = \frac{\ell_\varepsilon^2 |\log \varepsilon|}{2}$ in this regime of domain sizes.

For small or bounded domains Theorem 1.1 follows from adapting the proof of Sandier and Serfaty [21], where $\ell \equiv 1$. Formally examining the resulting critical field (1.11), one finds $H_1(\ell_0) \to \frac{\ell_0^2}{2}$ as $\ell_0 \to \infty$, so we expect that $h_{c_1} = \frac{\ell_\varepsilon^2 |\log \varepsilon|}{2}$ for any $\ell_\varepsilon \to +\infty$ and $\ell_\varepsilon \ll \frac{1}{\sqrt{\varepsilon |\log \varepsilon|}}$; see the discussion before Lemma 2.7. We give a proof for the case of large (but not too large) domains in section 2; see Proposition 2.1 for details.

The following results can be used to characterize the minimizers of $(u_\varepsilon, A_\varepsilon)$ for external fields of order $O(|\log \varepsilon|)$ and small or bounded domains. The first step is a Γ-convergence result that relates $G_\varepsilon(u_\varepsilon, A_\varepsilon)$ to a simpler functional that no longer involves ε. We skip some of the detailed convergence statements for ease of presentation. The full statement is given in Theorem 3.1.

THEOREM 1.2. *As $\varepsilon \to 0$, the functional $\frac{1}{|\log \varepsilon|^2} G_\varepsilon$ is Γ-convergent to $G(v, a)$, where the limit functional G is given by*

(1.13)
$$G(v,a) := \begin{cases} \frac{1}{2}\int_U |v-a|^2 + \frac{1}{\ell_0^2} |\operatorname{curl} a - H|^2 + \frac{1}{2} \|\operatorname{curl} v\|_{\mathscr{M}} & \text{if } \ell_0 > 0, \\ \frac{1}{2}\int_U |v-a|^2 + \frac{1}{2} \|\operatorname{curl} v\|_{\mathscr{M}} & \text{if } \ell_0 = 0 \text{ and } \operatorname{curl} a = H, \\ +\infty & \text{otherwise,} \end{cases}$$

under a convergence that includes $\left(\frac{1}{|\log \varepsilon|}(iu_\varepsilon, \nabla u_\varepsilon) - \frac{1}{|\log \varepsilon|}|u|^2 A_\varepsilon\right) \rightharpoonup (v - a)$ and $\frac{1}{|\log \varepsilon|} A_\varepsilon \rightharpoonup \operatorname{curl} a$ in L^2.

Since Γ-convergence and the compactness we have here imply that minimizers of G_ε and of G approximate each other, we study minimizers of G to gain insight into the structure of minimizers of G_ε.

THEOREM 1.3. *If (v_0, a_0) is a minimizer of (1.13) and $\ell_0 > 0$, then $z_0 = \ell_0^{-2}(\operatorname{curl} a_0 - H)$ is the unique minimizer of the following obstacle problem: Minimize*

$$F_{\ell_0, H}(z) = \frac{1}{2}\int_U |\nabla z|^2 + \ell_0^2 z^2 + 2zH \tag{1.14}$$

in the admissible class

$$\mathscr{K} = \left\{ z \in H_0^1(U) : z \geq -\frac{1}{2} \text{ a.e. in } U \right\}. \tag{1.15}$$

The limit (v_0, a_0) satisfies the following additional properties:

$$\begin{aligned} \ell_0^{-2} \operatorname{curl}(\operatorname{curl} a_0 - H) + a_0 &= v_0 \quad \text{in } U, \\ \operatorname{curl} a_0 - H &= 0 \quad \text{on } \partial U, \\ -\frac{1}{2} \leq z_0 &\leq 0, \\ \operatorname{curl} v_0 &\geq 0, \\ \operatorname{spt}(\operatorname{curl} v_0) &\subset \left\{ z_0 = -\frac{1}{2} \right\}. \end{aligned}$$

In the case where $\ell_0 = 0$, we have $\operatorname{curl} a_0 = H$ and obtain a slightly different obstacle problem: Let y_0 be the solution of $-\Delta y_0 = \operatorname{curl} v_0 - H$ with zero boundary conditions. Then y_0 is the unique minimizer of

$$F_{0,H}(y) = \frac{1}{2}\int_U |\nabla y|^2 + 2yH \tag{1.16}$$

in the admissible class

$$\mathscr{K} = \left\{ y \in H^1_0(U) : y \geq -\frac{1}{2} \text{ a.e. in } U \right\}. \tag{1.17}$$

Moreover, $\operatorname{curl} v_0 \geq 0$ *and* $\operatorname{spt}(\operatorname{curl} v_0) \subset \{y_0 = -\frac{1}{2}\}$.

This theorem, proved later in section 3, implies again the results on the first critical field: When the obstacle is not active, the minimizer satisfies $\operatorname{curl} v = 0$. This happens if and only if $H < H_1(\ell_0)$ with the same fields as above. However, since we rescaled the vorticity to obtain convergence, this only shows that an approximating sequence $(u_\varepsilon, A_\varepsilon)$ has at most $o(|\log \varepsilon|)$ vortices, a result that is weaker than the "no vortices below the critical field" obtained in Theorem 1.1.

Finally, we study minimizers of the Ginzburg–Landau functional with a very large (supercritical) applied external magnetic field and obtain energy asymptotics of a uniform vortex lattice, as follows.

THEOREM 1.4. *Assume that*

$$\max\{1, \ell^2\} |\log \varepsilon| \ll h_{ex} \ll \frac{1}{\varepsilon^2}; \tag{1.18}$$

then minimizing sequence $\{u_\varepsilon, A_\varepsilon\}$ *satisfies*

$$G_\varepsilon(u_\varepsilon, A_\varepsilon) = \frac{1}{2} |U| \, h_{ex} \log \frac{1}{\varepsilon \sqrt{h_{ex}}} \left(1 - o_\varepsilon(1)\right). \tag{1.19}$$

Furthermore, the vortex density is uniform in the limit; see Proposition 5.2 *for a precise statement.*

The proof of Theorem 1.4 is given in Proposition 5.2 and relies on the upper bound Proposition 4.1. For the lower bound, we have followed the simple proof of Sandier and Serfaty [18] and reduced the theorem by a well-chosen rescaling to the lower bound part of the Γ-convergence theorem in the $|\log \varepsilon|^2$ scaling.

For the upper bound, we use a simple approach with Fourier series (Proposition 4.1). A refined version of the upper bound, (4.2), motivates our conjecture on the behavior close to the critical field for large domains.

Remark 1.5. For $\ell_\varepsilon \to \infty$ and $h_{ex} = O(\ell^2 |\log \varepsilon|)$, we do not yet have a rigorous result on the structure of minimizers. However, we expect from formal calculations that a uniform lattice, such as those constructed in section 4, should be minimizing. This is further supported by the analysis of Aydi [3], who showed such a result in the periodic setting. Aydi's upper bound and ours are essentially equivalent.

1.3. Discussion. We conclude the introduction with several unresolved questions regarding asymptotics of (1.3). There is still work to do to complete the answer to Problem 1 in [18]. In particular, a complete phase diagram for the minimizing behavior depending on κ, ℓ, and h_{ex} should be given, including the cross-over between type I and type II behavior that happens for $\kappa\ell = O(1)$, and the results of Aftalion and Du [2] should be made fully rigorous. For such a study, it would also be necessary to understand local minimizers and hysteresis phenomena for slowly changing fields.

It is an interesting problem to further study beginning vortex nucleation close to the critical field in the large domain limit, $\ell \to +\infty$. Based on the construction of Proposition 4.1 and the structure of the Meissner state, we expect that minimizers exhibit a uniform vortex lattice that fills the whole domain. However, vortices will be far apart and interact only weakly, making this a subtle problem. Finally, it would

be interesting to study (1.3) with applied fields in the "intermediate range," recently undertaken for (1.7) in [18]. For states with few vortices in large domains, we similarly expect very slow motion for the gradient flow, as vortices will move in an almost flat potential.

2. Critical field calculation. In this section we establish Theorem 1.1. The proof of Theorem 1.1 for $\ell \to [0, +\infty)$ follows from a direct insertion of ℓ^{-2} into the magnetic field term of (1.7) and following the proof found in [22, 21]. However, when $\ell \to +\infty$, a simple scaling argument fails, and we need to be more careful. In the following we show that for a substantial class of large-domain asymptotics, the critical field strength is indeed

$$h_{c_1} = \frac{\ell^2}{2}|\log \varepsilon| + o(|\log \varepsilon|),$$

as suggested by the formal analysis of the scaled renormalized energy.

PROPOSITION 2.1. *Let $\ell \to +\infty$ with $\ell \le C|\log \varepsilon|^{\gamma}$ for any fixed $\gamma \in \mathbb{R}^+$, and suppose that (u, A) is a minimizer of the Ginzburg–Landau energy* (1.3). *Then the first critical field for vortex nucleation is $h_{c_1} = \frac{\ell^2}{2}|\log \varepsilon| = \frac{1}{2}|\log \varepsilon|^{2\gamma+1}$. In particular for $h_{ex} < \frac{\ell^2}{2}|\log \varepsilon|$, any minimizer will satisfy $|u| \ge \frac{3}{4}$ for all ε sufficiently small, and for $h_{ex} > \frac{\ell^2}{2}|\log \varepsilon|$ any minimizer must have a vortex.*

Remark 2.2. Although we establish the conjectured critical field for $\ell = |\log \varepsilon|^{\gamma}$, we believe the critical field should be true over length scales up to $\ell \le \frac{C}{\sqrt{\varepsilon}|\log \varepsilon|}$. In particular, the more refined vortex ball estimates found in [13, 18] should be powerful enough to handle larger domains, but in the interest of brevity we consider only ℓ's that satisfy $1 \ll \ell \le C|\log \varepsilon|^{\gamma}$. We establish Proposition 2.1 by using the explicit vortex structure that exists for these intermediate-sized domains.

The Euler–Lagrange equations of (1.3),

$$\begin{aligned} 0 &= \nabla_A^2 u + \frac{1}{\varepsilon^2} u\left(1 - |u|^2\right), \\ 0 &= \operatorname{curl}\operatorname{curl} A + \ell^2 j_A(u), \end{aligned} \tag{2.1}$$

in U and $n \cdot \nabla_A u = 0$ and $\operatorname{curl} A = h_{ex}$ on ∂U. Setting the Coulomb gauge, we see that

$$\operatorname{div} A = 0 \text{ in } U, \qquad n \cdot A = 0 \text{ on } \partial U.$$

Solutions to (2.1) satisfy the maximum principle

$$\|u\|_{L^\infty(U)} \le 1, \tag{2.2}$$

the proof of which can be found in [6, 19].

The key to establishing Proposition 2.1 is a good energy decomposition. In order to establish this decomposition we use the following result of Sandier and Serfaty that supplies the vortex structure for our range of ℓ's. Their result, based on the method of Jerrard [10] and Jerrard and Soner [11] is the following.

PROPOSITION 2.3 (see Sandier and Serfaty[21]). *Let $u : U \to \mathbb{C}$ be such that $|\nabla u| \le \frac{C}{\varepsilon}$ and that $E_\varepsilon(u) \le C|\log \varepsilon|^M$ for $M \ge 2$ a fixed number. Then, for any $\alpha > 0$ there exist disjoint balls $\{B_{r_i}\}_{i \in I}$ of radii r_i such that for sufficiently small ε,*

1. $\{|u| < \frac{3}{4}\} \subset \cup_{i \in I} B_{r_i}$,

2. $\operatorname{card} I \leq C|\log\varepsilon|^M$,
3. $r_i \leq \frac{C}{|\log\varepsilon|^\alpha}$,
4. *if* $\overline{B_{r_i}} \subset U$ *and* $d_i = \deg(u, \partial B_{r_i})$, *then*

$$\int_{B_{r_i}} e_\varepsilon(u)dx \geq \pi\,|d_i|\,(|\log\varepsilon| - O(\log|\log\varepsilon|))\,. \tag{2.3}$$

Remark 2.4. The result in [21] is restricted to energies of the size $E_\varepsilon(u) \leq K|\log\varepsilon|^2$; however, the same proof holds for the higher energies in the assumptions found in Proposition 2.1.

We now state our energy decomposition, in the spirit of Bethuel and Riviere [6], Serfaty [22], and Sandier and Serfaty [19].

PROPOSITION 2.5. *Let* (u, A) *be a minimizer, where* A *satisfies the Coulomb gauge and* $1 \ll \ell \leq C|\log\varepsilon|^\gamma$. *Let* $A = h_{ex}\nabla^\perp\xi_\ell + \nabla^\perp\zeta$, *where* ξ_ℓ *satisfies*

$$\begin{aligned} -\frac{1}{\ell^2}\Delta^2\xi_\ell + \Delta\xi_\ell &= 0 \text{ in } U,\\ \Delta\xi_\ell &= 1 \text{ on } \partial U,\\ \xi_\ell &= 0 \text{ on } \partial U. \end{aligned} \tag{2.4}$$

Then

$$\begin{aligned} G_\varepsilon(u, A) \geq \sum_{i\in I}\int_{B_{r_i}} e_\varepsilon(u)dx + \frac{1}{\ell^2}\int_U |\Delta\zeta|^2 + G^0\\ + 2\pi h_{ex}\sum_{i\in I} d_i\xi_\ell(a_i) - \frac{C}{|\log\varepsilon|}, \end{aligned} \tag{2.5}$$

where $G^0 = G_\varepsilon(1, h_{ex}\nabla^\perp\xi_\ell)$. *Here the vortex balls* B_{r_i} *and degrees* d_i *are defined via Proposition* 2.3

We prove several intermediate lemmas before attempting the proof of Proposition 2.5. The first facts we establish are on the scaled London equation. This limiting equation for the stream function of the magnetic field potential is the expected Meissner solution.

LEMMA 2.6. *Let* ξ_ℓ *be the solution of* (2.4) *with* $\ell \gg 1$; *then*

$$-\frac{1}{\ell^2} \leq \xi_\ell \leq 0 \tag{2.6}$$

and

$$\sup_{x\in\overline{U}} |\xi_\ell| = \frac{1}{\ell^2}\,(1 - o_\ell(1))\,. \tag{2.7}$$

Further,

$$\|\xi_\ell\|_{H^2} \leq C \quad \text{and} \quad \|\nabla\xi_\ell\|_{L^\infty} \leq C, \tag{2.8}$$

where C *depends only on* U.

Proof. These results are similar to results found in [6, 22, 24] for $\ell \equiv 1$. If $\Delta\xi_\ell = h_\ell$ in U and $\xi_\ell = 0$ on ∂U, then h_ℓ satisfies

$$\begin{aligned} -\frac{1}{\ell^2}\Delta h_\ell + h_\ell &= 0 \text{ in } U,\\ h_\ell &= 1 \text{ on } \partial U. \end{aligned} \tag{2.9}$$

If we let $\chi = \xi_\ell - \frac{1}{\ell^2}(h_\ell - 1)$, then $\Delta\chi = \Delta\xi_\ell - \frac{1}{\ell^2}\Delta(h_\ell - 1) = h_\ell - \frac{1}{\ell^2}\Delta h_\ell = 0$ in U and $\chi = 0$ on ∂U. Therefore, $\chi \equiv 0$, and thus

$$\xi_\ell = \frac{1}{\ell^2}(h_\ell - 1). \tag{2.10}$$

Applying the maximum principle to (2.9) yields $0 < h_\ell < 1$. In particular, if a minimum occurs at a point x_m in the interior of U, then $0 < \frac{1}{\ell^2}\Delta h_\ell(x_m) = h_\ell(x_m)$, and by the boundary condition we see $h_\ell \geq 0$. On the other hand, if the maximum occurs at a point x_M in the interior of U, then $0 > \frac{1}{\ell^2}\Delta h_\ell(x_M) = h_\ell(x_M)$, and by the boundary condition $h_\ell \leq 1$. Applying this to (2.10) yields (2.6).

Next, using the boundary conditions on ξ_ℓ,

$$\begin{aligned} 0 &= \int_U \xi_\ell \left[-\frac{1}{\ell^2}\Delta^2\xi_\ell + \Delta\xi_\ell\right] = \int_U \frac{1}{\ell^2}\nabla\xi_\ell \cdot \nabla\Delta\xi_\ell - |\nabla\xi_\ell|^2 \\ &= \int_{\partial U} \frac{1}{\ell^2}\partial_n\xi - \int_U \left[\frac{1}{\ell^2}|\Delta\xi_\ell|^2 + |\nabla\xi_\ell|^2\right]. \end{aligned}$$

Thus by (2.9), (2.10), and the bound on h_ℓ,

$$\int_U \frac{1}{\ell^2}|\Delta\xi_\ell|^2 + |\nabla\xi_\ell|^2 = \frac{1}{\ell^2}\int_U \Delta\xi_\ell = \frac{1}{\ell^2}\int_U h_\ell \leq \frac{1}{\ell^2}|U|.$$

This implies $\|\xi_\ell\|_{H^2(U)} \leq 2\sqrt{|U|}$. Since $0 < \Delta\xi_\ell < 1$, then $\|\xi_\ell\|_{W^{2,p}} \leq C_p$ for any chosen $p > 2$. Therefore, by Sobolev embedding, we have (2.8).

Set $z_\ell = \partial_\ell h_\ell$; then z_ℓ satisfies

$$\Delta z_\ell - \ell^2 z_\ell = 2\ell h_\ell \geq 0$$

in U and $z_\ell = 0$ on ∂U. By the maximum principle $z_\ell \leq 0$; hence, h_ℓ is monotonically decreasing in ℓ for all $x \in U$. Since h_ℓ is bounded below by -1, then $h_\ell(x) = -1 + o_\ell(1)$ for all $x \in U' \Subset U$. Thus $\max|\xi_\ell| = \frac{1}{\ell^2}(1 - o_\ell(1))$ by (2.10). ∎

In order to use Proposition 2.3 we need to establish a bound on $E_\varepsilon(u)$; see (1.8). As we see below, the BBH energy can be much larger than the Ginzburg–Landau energy $G_\varepsilon(u, A)$, since the magnetic field term in the energy can absorb large induced fields generated by a large number of vortices. We have the following result.

LEMMA 2.7. *Let (u, A) be a minimizer of the Ginzburg–Landau energy. Suppose $h_{ex} \leq C\ell^2|\log\varepsilon| \leq \frac{C}{\varepsilon}$ and $G_\varepsilon(u, A) \leq K\ell^2|\log\varepsilon|^2$; then*

$$E_\varepsilon(u) \leq C\ell^4|\log\varepsilon|^2 \tag{2.11}$$

and

$$\|A\|_{H^1(U)} \leq C\ell^2|\log\varepsilon| \quad \text{and} \quad \|A\|_{H^2(U)} \leq C\ell^3|\log\varepsilon|. \tag{2.12}$$

Proof. We first establish a uniform H^1 estimate on A. From the assumption on the energy,

$$\int |h - h_{ex}|^2 \leq K\ell^4|\log\varepsilon|^2,$$

and hence from the bound on h_{ex} we see that

$$\|h\|_{L^2(U)} \leq C\ell^2|\log\varepsilon|.$$

Since $\operatorname{div} A = 0$ and $n \cdot A = 0$ on ∂U, there exists ξ such that $\nabla^\perp \xi = A$ and $\xi = 0$ on ∂U. From standard elliptic estimates we get $\|\xi\|_{H^2(U)} \le C\ell^2 |\log\varepsilon|$. Thus

$$\|A\|_{H^1(U)} \le C\ell^2 |\log\varepsilon|.$$

Decomposing $\frac{1}{2}|\nabla_A u|^2 = \frac{1}{2}|\nabla u|^2 - A \cdot j(u) + \frac{1}{2}A^2 |u|^2$, we control the cross term via

$$\begin{aligned} A \cdot j(u) &\le \frac{1}{4}\left|\frac{j(u)}{|u|}\right|^2 + A^2 |u|^2 \\ &\le \frac{1}{4}\left|\frac{j(u)}{|u|}\right|^2 + A^2 + 2\varepsilon^2 A^4 + \frac{1}{8\varepsilon^2}\left(1-|u|^2\right)^2. \end{aligned}$$

Therefore, from the algebraic bounds, the estimate on A, and Sobolev embedding,

$$\begin{aligned} G_\varepsilon(u,A) &\ge E_\varepsilon + \frac{1}{2\ell^2}\|h - h_{ex}\|^2 - \left[\frac{1}{2}E_\varepsilon(u) + \int_U A^2 + 2\varepsilon^2 A^4\right] \\ &\ge \frac{1}{2}E_\varepsilon(u) - C\left[\|A\|^2_{H^1(U)} + \varepsilon^2 \|A\|^4_{H^1(U)}\right] \\ &\ge \frac{1}{2}E_\varepsilon(u) - C\ell^4 |\log\varepsilon|^2 - C\varepsilon^2\ell^8 |\log\varepsilon|^4. \end{aligned}$$

The upper bound on $E_\varepsilon(u)$ follows.

In order to establish higher bounds on A we use the Euler–Lagrange equation $-\nabla^\perp h = \ell^2 j_A(u)$. Therefore,

$$\|\nabla h\|_{L^2(U)} \le \ell^2 \|\nabla_A u\|_{L^2(U)} \|u\|_{L^\infty(U)} \le C\ell^3 |\log\varepsilon|,$$

and hence (2.8). ∎

The fact that $E_\varepsilon(u)$ can have a much larger energy than $G_\varepsilon(u,A)$ is an essential difference in the large-ℓ asymptotics. It implies a more complicated global vortex structure. Given the energy bound on $E_\varepsilon(u)$, we can split apart the full Ginzburg–Landau energy into its chief components. We start with an initial energy splitting.

LEMMA 2.8. *We can decompose* $G_\varepsilon(u,A) = G_\varepsilon(u, \nabla^\perp\xi) = G_\varepsilon(u, h_{ex}\nabla^\perp\xi_\ell + \nabla^\perp\zeta)$ *as*

$$\begin{aligned} G_\varepsilon(u,A) \ge{}& \frac{1}{2}\int_U \left|\nabla u - iu\nabla^\perp\xi\right|^2 + \frac{1}{2\varepsilon^2}\left(1-|u|^2\right)^2 + \frac{1}{2\ell^2}\int_U |\Delta\xi|^2 \\ &+ \int_U \left(\nabla u, -ih_{ex}\nabla^\perp\xi_\ell u\right) + G^0 - \frac{C}{|\log\varepsilon|}. \end{aligned} \tag{2.13}$$

Proof. We decompose the energy in a series of steps.

Our first step is to compute the approximate energy of the Meissner state via the method of Serfaty [22]. Since $\operatorname{div} A = 0$ and $n \cdot A = 0$ in ∂U, we can write $A = \nabla^\perp\xi$ with $\xi = 0$ on ∂U and so $\Delta\xi = h$. We further decompose $\nabla^\perp\xi$ as $\nabla^\perp\xi = h_{ex}\nabla^\perp\xi_\ell + \nabla^\perp\zeta$, where $\zeta = \Delta\zeta = 0$ on ∂U and where ξ_ℓ satisfies (2.4). Consider now the Meissner energy associated with $G^0 = G_\varepsilon(1, h_{ex}\nabla^\perp\xi_\ell)$. We compute the form of the Meissner energy, setting $(u_0, A_0) = (1, h_{ex}\nabla^\perp\xi_\ell)$. Then

$$\begin{aligned} G_\varepsilon(u_0,A_0) &= \frac{1}{2}\int_U |\nabla_{A_0}u_0|^2 + \frac{1}{\ell^2}|\operatorname{curl} A_0 - h_{ex}|^2 + \frac{1}{2\varepsilon^2}\left(1-|u_0|^2\right)^2 \\ &= \frac{h_{ex}^2}{2}\int_U |\nabla\xi_\ell|^2 + \frac{1}{\ell^2}|\Delta\xi_\ell - 1|^2. \end{aligned}$$

Multiplying ξ_ℓ against $-\frac{1}{\ell^2}\Delta^2\xi_\ell + \Delta\xi_\ell$ and integrating over U yields

$$\int_U |\nabla\xi_\ell|^2 + \frac{1}{\ell^2}|\Delta\xi_\ell|^2\,dx = \frac{1}{\ell^2}\int_U \Delta\xi_\ell.$$

We use the above identity to rewrite the Meissner energy as

$$\begin{aligned} G^0 &= \frac{1}{2}\int_U h_{ex}^2|\nabla\xi_\ell|^2 + \frac{h_{ex}^2}{\ell^2}|\Delta\xi_\ell|^2 + \frac{h_{ex}^2}{\ell^2} - 2\frac{h_{ex}^2}{\ell^2}\Delta\xi_\ell \\ &= -\frac{h_{ex}^2}{2}\int_U \left[|\nabla\xi_\ell|^2 + \frac{1}{\ell^2}|\Delta\xi_\ell|^2\right] + \frac{h_{ex}^2}{2\ell^2}|U|. \end{aligned} \tag{2.14}$$

Therefore the Meissner energy is of order $O(\frac{h_{ex}^2}{\ell^2})$.

Next we write

$$\begin{aligned} \frac{1}{2}\int_U |\nabla_A u|^2 &= \frac{1}{2}\int_U \left|\nabla u - ih_{ex}\nabla^\perp\xi_\ell u - i\nabla^\perp\zeta u\right|^2 \\ &= \frac{1}{2}\int_U \left|\nabla u - i\nabla^\perp\zeta u\right|^2 + \frac{h_{ex}^2}{2}\int_U |u|^2\left|\nabla^\perp\xi_\ell\right|^2 \\ &\quad + \int_U \left(\nabla u - i\nabla^\perp\zeta u, -ih_{ex}\nabla^\perp\xi_\ell u\right) \end{aligned}$$

and

$$\begin{aligned} &\int_U \left(\nabla u - i\nabla^\perp\zeta u, -ih_{ex}\nabla^\perp\xi_\ell u\right) \\ &\quad = \int_U \left(\nabla u, -ih_{ex}\nabla^\perp\xi_\ell u\right) + h_{ex}\int_U |u|^2\nabla\xi_\ell\cdot\nabla\zeta. \end{aligned}$$

Therefore, we can write the Ginzburg–Landau energy as

$$\begin{aligned} G_\varepsilon(u,A) &= \frac{1}{2}\int_U \left|\nabla u - i\nabla^\perp\zeta u\right|^2 + \frac{1}{2\varepsilon^2}\left(1-|u|^2\right)^2 \\ &\quad + \int_U \left(\nabla u, -ih_{ex}\nabla^\perp\xi_\ell u\right) \\ &\quad + \frac{h_{ex}^2}{2}\int_U \left(|u|^2-1\right)\left|\nabla^\perp\xi_\ell\right|^2 + h_{ex}\int_U \left(|u|^2-1\right)\nabla\xi_\ell\cdot\nabla\zeta \\ &\quad + \frac{1}{2\ell^2}\int_U |h-h_{ex}|^2 + \frac{h_{ex}^2}{2}\int_U \left|\nabla^\perp\xi_\ell\right|^2 + h_{ex}\int_U \nabla\xi_\ell\cdot\nabla\zeta. \end{aligned} \tag{2.15}$$

The terms in the third line of (2.15) are small since

$$\begin{aligned} h_{ex}^2\int_U \left(|u|^2-1\right)|\nabla\xi_\ell|^2 &\le Ch_{ex}^2\|\nabla\xi_\ell\|_{L^\infty}^2\left\|1-|u|^2\right\|_{L^2} \le C\varepsilon h_{ex}^2 E_\varepsilon^{\frac{1}{2}}(u) \\ &\le C\varepsilon\ell^6|\log\varepsilon|^3 \le C\varepsilon|\log\varepsilon|^{6\gamma+3} \le \frac{C}{|\log\varepsilon|} \end{aligned}$$

and

$$\begin{aligned} h_{ex}\int_U \left(|u|^2-1\right)\nabla\xi_\ell\cdot\nabla\zeta &\le Ch_{ex}\|\nabla\xi_\ell\|_{L^\infty}\|\nabla\zeta\|_{L^2}\left\|1-|u|^2\right\|_{L^2} \\ &\le C\varepsilon h_{ex}\left\|A - h_{ex}\nabla^\perp\xi_\ell\right\|_{L^2} E_\varepsilon^{\frac{1}{2}}(u) \\ &\le C\varepsilon\ell^2|\log\varepsilon|\left(\ell^2|\log\varepsilon| + \ell^2|\log\varepsilon|\right)\ell^2|\log\varepsilon| \\ &\le C\varepsilon|\log\varepsilon|^{6\gamma+3} \le \frac{C}{|\log\varepsilon|}. \end{aligned}$$

For the fourth line of (2.15) we have

$$\begin{aligned}
&\frac{1}{2\ell^2}\int_U |h-h_{ex}|^2 + \frac{h_{ex}^2}{2}\int_U \left|\nabla^\perp \xi_\ell\right|^2 + h_{ex}\int_U \nabla\xi_\ell\cdot\nabla\zeta \\
&\quad = \frac{1}{2\ell^2}\int_U |h_{ex}\Delta\xi_\ell - h_{ex} + \Delta\zeta|^2 + \frac{h_{ex}^2}{2}\int_U |\nabla\xi_\ell|^2 + h_{ex}\int_U \nabla\xi_\ell\cdot\nabla\zeta \\
&\quad = \frac{h_{ex}^2}{2\ell^2}\int |\Delta\xi_\ell - 1|^2 + \frac{h_{ex}^2}{2}\int |\nabla\xi_\ell|^2 + \frac{1}{2\ell^2}\int |\Delta\zeta|^2 \\
&\qquad + \frac{h_{ex}}{\ell^2}\int (\Delta\xi_\ell - 1)\,\Delta\zeta + h_{ex}\int \nabla\xi_\ell\cdot\nabla\zeta.
\end{aligned}$$

Multiplying ζ against (2.4) and integrating over U, we have

$$0 = -\frac{1}{\ell^2}\int_U \Delta\zeta\,(\Delta\xi_\ell - 1) - \int_U \nabla\xi_\ell\cdot\nabla\zeta,$$

and then

$$\begin{aligned}
&\frac{1}{2\ell^2}\int_U |h-h_{ex}|^2 + \frac{h_{ex}^2}{2}\int_U \left|\nabla^\perp \xi_\ell\right|^2 + h_{ex}\int_U \nabla\xi_\ell\cdot\nabla\zeta \\
&\qquad = G^0 + \frac{1}{2\ell^2}\int_U |\Delta\zeta|^2 .
\end{aligned} \tag{2.16}$$

Combining (2.15), the bounds on the third line, and (2.16) yields (2.13). ∎

We can now prove Proposition 2.5 by carefully extracting the concentration of the Ginzburg–Landau energy against the magnetic field potential ξ_ℓ. Note that there are potentially an unbounded number of vortices, so we need to extract good decay on each vortex ball.

Proof of Proposition 2.5. We follow the approach in [21] for $\ell \equiv 1$. The first step is to establish the concentration in the cross term $\int \nabla^\perp\xi_\ell \cdot j(u)$. In particular, we claim

$$\left|\int_U \left(\nabla u, -ih_{ex}\nabla^\perp\xi_\ell u\right) - 2\pi h_{ex}\sum_{i\in I} d_i\xi_\ell(a_i)\right| \le \frac{C}{|\log\varepsilon|}, \tag{2.17}$$

where a_j is the center of the vortex ball B_{r_i} and I is the vortex ball collection.

Step 1. Since $E_\varepsilon(u) \le C\ell^4|\log\varepsilon|^2 \le C|\log\varepsilon|^{4\gamma+2}$ and $h_{ex} = C\ell^2|\log\varepsilon| = C|\log\varepsilon|^{2\gamma+1}$, then by Proposition 2.3 we have balls $\{B_{r_i}\}_{i\in I}$ such that

$$\left\{|u| < \frac{3}{4}\right\} \subset \cup_{i\in I} B_{r_i}, \qquad \operatorname{card} I \le C|\log\varepsilon|^{4\gamma+2}, \qquad r_i \le \frac{C}{|\log\varepsilon|^{10\gamma+6}},$$

$$\text{if } \overline{B_{r_i}} \subset U \text{ and } d_i = \deg(u,\partial B_{r_i}), \text{ then } \int_{B_{r_i}} e_\varepsilon(u)dx \ge \pi\,|d_i|\,(|\log\varepsilon| - O(\log|\log\varepsilon|)),$$

where we chose $\alpha = 10\gamma + 6$ in Proposition 2.3. Therefore,

$$\begin{aligned}
\left|\int_{\cup_I B_{r_i}} \left(\nabla u, -ih_{ex}\nabla^\perp\xi_\ell u\right)\right| &\le (\operatorname{card} I)\,h_{ex}\,\|\nabla u\|_{L^2}\max_{i\in I} r_i \\
&\le C|\log\varepsilon|^{8\gamma+4-\alpha} \le \frac{C}{|\log\varepsilon|}.
\end{aligned}$$

Setting $\widetilde{U} = U \setminus \cup_{i\in I} B_{r_i}$, for $v = \frac{u}{|u|}$ we have

$$\int_{\widetilde{U}} \left(\nabla u, -ih_{ex}\nabla^{\perp}\xi_\ell u\right) = h_{ex}\int_{\widetilde{U}} \nabla\xi_\ell \times j(v) + h_{ex}\int_{\widetilde{U}} \nabla\xi_\ell \times (j(u) - j(v)).$$

The second term is small, using (2.8) and (2.11), since

$$\begin{aligned} h_{ex}\int_{\widetilde{U}} \nabla\xi_\ell \times (j(u) - j(v)) &\le h_{ex}\left\|\nabla^{\perp}\xi_\ell\right\|_{L^\infty}\int_{\widetilde{U}}\left|j(u) - \frac{j(u)}{|u|^2}\right| \\ &\le Ch_{ex}\int_{\widetilde{U}} \frac{|j(u)|}{|u|}\frac{\left||u|^2-1\right|}{|u|} \le C\varepsilon h_{ex}E_\varepsilon(u) \\ &\le C\varepsilon|\log\varepsilon|^{6\gamma+3} = \frac{C}{|\log\varepsilon|}, \end{aligned}$$

where we used $|u| \ge \frac{1}{2}$ in $\widetilde{U}$ in the second line. Therefore,

$$\left|\int_{\widetilde{U}} \left(\nabla u, -ih_{ex}\nabla^{\perp}\xi_\ell u\right) - h_{ex}\int_{\widetilde{U}} \nabla\xi_\ell \times j(v)\right| \le \frac{C}{|\log\varepsilon|}. \tag{2.18}$$

Next for $J = \{i \text{ such that } \overline{B_{r_i}} \subset U\}$ we claim we can extract the following bound:

$$\left|h_{ex}\int_{B_{r_i}} \nabla\xi_\ell \times j(v) - 2\pi h_{ex}d_i\xi_\ell(a_i)\right| \le \frac{C}{|\log\varepsilon|^{4\gamma+3}}. \tag{2.19}$$

For $\Omega_i = B_{r_i} \cap \{x \in U \text{ such that } |u| \le \frac{1}{2}\}$, $\Omega_i \cap \partial B_{r_i} = \varnothing$. Since $|u| \ge \frac{1}{2}$ in Ω_i, by Stokes' theorem,

$$\begin{aligned} &h_{ex}\left|\int_{\partial B_{r_i}} (\xi_\ell - \xi_\ell(a_i))\, j(v)\cdot\tau - \int_{\partial U_i} (\xi_\ell - \xi_\ell(a_i))\, j(v)\cdot\tau\right| \\ &\quad = h_{ex}\left|\int_{B_i\setminus U_i} \nabla\xi_\ell \times j(u)\right| \le Ch_{ex}\left\|\nabla\xi_\ell\right\|_{L^\infty}\left\|\nabla u\right\|_{L^2} r_i \\ &\quad \le C|\log\varepsilon|^{4\gamma+2-\alpha} \le \frac{C}{|\log\varepsilon|^{4\gamma+3}}; \end{aligned}$$

thus

$$\begin{aligned} h_{ex}\left|\int_{\partial B_{r_i}} (\xi_\ell - \xi_\ell(a_i))\, j(v)\cdot\tau - \int_{\partial\Omega_i} (\xi_\ell - \xi_\ell(a_i))\, j(v)\cdot\tau\right| \\ \le \frac{C}{|\log\varepsilon|^{4\gamma+3}}. \end{aligned} \tag{2.20}$$

On the other hand, since $|u| = \frac{1}{2}$ on $\partial\Omega_i$ we find

$$\begin{aligned}
&h_{ex}\left|\int_{\partial\Omega_i}(\xi_\ell - \xi_\ell(a_i))\, j(v)\cdot\tau\right| \\
&\quad= h_{ex}\left|\int_{\partial\Omega_i}(\xi_\ell - \xi_\ell(a_i))\,\frac{j(u)}{|u|^2}\cdot\tau\right| = 4h_{ex}\left|\int_{\partial\Omega_i}(\xi_\ell - \xi_\ell(a_i))\, j(u)\cdot\tau\right| \\
&\quad= 4h_{ex}\left|\int_{\Omega_i}\operatorname{curl}\left[(\xi_\ell - \xi_\ell(a_i))\, j(u)\right]\right| \\
&\quad\le 4h_{ex}\left|\int_{\Omega_i}\nabla\xi_\ell\times j(u)\right| + 8h_{ex}\left|\int_{\Omega_i}(\xi_\ell - \xi_\ell(a_i))\, J(u)\right| \\
&\quad\le Ch_{ex}\left\|\nabla\xi_\ell\right\|_{L^\infty}\left\|\nabla u\right\|_{L^2} r_i + Ch_{ex}\left\|\nabla\xi_\ell\right\|_{L^\infty}\left\|\nabla u\right\|_{L^2}^2 r_i \\
&\quad\le C|\log\varepsilon|^{4\gamma+2-\alpha} + C|\log\varepsilon|^{6\gamma+3-\alpha} \le \frac{C}{|\log\varepsilon|^{4\gamma+3}},
\end{aligned}$$

and consequently, since $\operatorname{card} I \le C|\log\varepsilon|^{4\gamma+2}$ and $\int_{\partial B_{r_i}}\xi_\ell(a_i)j(v)\cdot\tau = 2\pi d_i$,

$$\sum_{i\in J}\left|h_{ex}\int_{\partial B_{r_i}}\xi_\ell j(v)\cdot\tau - h_{ex}\int_{\partial B_{r_i}}\xi_\ell(a_i)j(v)\cdot\tau\right| \le \frac{C}{|\log\varepsilon|}. \tag{2.21}$$

Finally, for the balls that intersect ∂U, $I\backslash J$. Since $\xi_\ell = 0$ on ∂U, then for $\Omega_i = B_i\cap\{x\in U \text{ such that } |u|\le\frac{1}{2}\}$ we follow the above argument and see

$$\begin{aligned}
\left|h_{ex}\int_{\partial B_{r_i}\cap U}\xi_\ell j(v)\cdot\tau\right| &\le \left|h_{ex}\int_{\partial(\Omega_i\cap U)}\xi_\ell j(v)\cdot\tau\right| + \frac{C}{|\log\varepsilon|^{4\gamma+3}} \\
&\le 4h_{ex}\left|\int_{\Omega_i\cap U}\nabla\xi_\ell\times j(v) + 2\xi_\ell J(v)\right| + \frac{C}{|\log\varepsilon|^{4\gamma+3}} \\
&\le \frac{C}{|\log\varepsilon|^{4\gamma+3}}.
\end{aligned}$$

Combining this estimate along with $\operatorname{card} I \le C|\log\varepsilon|^{4\gamma+2}$ and (2.21) yields estimate (2.17).

Step 2. We bound $\int_U\left|\nabla u - i\nabla^\perp\zeta u\right|^2 \ge \int_{\cup_{i\in I}B_{r_i}}|\nabla u|^2 - \frac{C}{|\log\varepsilon|}$. In particular,

$$\begin{aligned}
\int_U\left|\nabla u - i\nabla^\perp\zeta u\right|^2 &\ge \int_{\cup_{i\in I}B_{r_i}}\left|\nabla u - i\nabla^\perp\zeta u\right|^2 \\
&= \int_{\cup_{i\in I}B_{r_i}}|\nabla u|^2 - 2\nabla^\perp\zeta\cdot j(u) + |\nabla\zeta|^2\,|u|^2.
\end{aligned}$$

From (2.12) we see $\|A\|_{L^\infty(U)} \le C\|A\|_{H^2(U)} \le C|\log\varepsilon|^{3\gamma+2}$, and thus

$$\begin{aligned}
\left|\int_{\cup_{i\in I}B_{r_i}}\nabla^\perp\zeta\cdot j(u)\right| &\le (\operatorname{card} I)\left\|A - h_{ex}\nabla^\perp\xi_\ell\right\|_{L^\infty}\left\|\nabla u\right\|_{L^2}\max_{i\in I} r_i \\
&\le C|\log\varepsilon|^{4\gamma+2}\left(|\log\varepsilon|^{3\gamma+2} + |\log\varepsilon|^{2\gamma+1}\right)|\log\varepsilon|^{2\gamma+1}|\log\varepsilon|^{-\alpha} \\
&\le C|\log\varepsilon|^{9\gamma+5-\alpha} \le \frac{C}{|\log\varepsilon|},
\end{aligned}$$

and so

$$\int_U \left|\nabla u - i\nabla^\perp \zeta u\right|^2 \geq \int_{\cup_{i\in I} B_{r_i}} |\nabla u|^2 . \tag{2.22}$$

Combining (2.13) with (2.17) and (2.22) yields (2.5). ∎

We are finally in the position to establish the next claim.

Proof of Proposition 2.1. The first part of the proof establishes that a minimizing sequence must be in the Meissner state when $h_{ex} < \frac{\ell^2 |\log\varepsilon|}{2}$.

Step 1. From Proposition 2.5 and the minimality of (u, A)

$$\begin{aligned} G^0 &\geq G_\varepsilon(u, A) \\ &\geq \sum_{i\in I} \int_{B_{r_i}} e_\varepsilon(u)dx + \frac{1}{2\ell^2}\int_U |\Delta\zeta|^2 + G^0 + 2\pi h_{ex}\sum_{i\in I} d_i \xi_\ell(a_i) - \frac{C}{|\log\varepsilon|}. \end{aligned}$$

Therefore, since $\xi_\ell \leq 0$, we use (2.7) and lower bound in Proposition 2.3.4 to get

$$\begin{aligned} \pi \sum_{i\in I} |d_i| \left(|\log\varepsilon| + O(\log|\log\varepsilon|)\right) &\leq 2\pi h_{ex} \sum_{i\in I} d_i \left|\xi_\ell(a_i)\right| \\ &\leq 2\pi h_{ex} \left(\sum_{i\in I} |d_i|\right) \max|\xi_\ell| \\ &\leq \left(\sum_{i\in I} |d_i|\right) 2\pi \frac{h_{ex}}{\ell^2} \left(1 - o_\ell\right). \end{aligned}$$

So if $\sum_{i\in I} |d_i| \neq 0$, then

$$h_{ex} \geq \frac{\ell^2}{2}\left(|\log\varepsilon| + O(\log|\log\varepsilon|)\right).$$

Hence, for $h_{ex} < \frac{\ell^2}{2}|\log\varepsilon|$, either $\deg(u, \partial B_{r_i}) = 0$ or $B_{r_i} \cap U \neq \varnothing$. It is straightforward to show from this point that $|u| \geq \frac{3}{4}$ in U; see [5, 6].

Step 2. We now complete the proof of the critical field strength. In particular, we show that if $h_{ex} > \frac{\ell^2|\log\varepsilon|}{2}$, then there must be a vortex. We prove this by contradiction. Let $(u_\varepsilon, A_\varepsilon)$ be a minimizing sequence with $\sum_{j\in J} |d_j| = 0$; then we claim $G_\varepsilon(u, A) \geq G_\varepsilon(1, h_{ex}\nabla^\perp\xi_\ell) - \frac{C}{|\log\varepsilon|}$.

In order to get better bounds on $\nabla^\perp\zeta = A - h_{ex}\nabla^\perp\xi_\ell$, we replace lower bound (2.22) with

$$\frac{1}{2}\int_U \left|\nabla u - i\nabla^\perp\zeta u\right|^2 \geq \frac{1}{2}\int_U |\nabla u|^2 - 2j(u)\cdot\nabla^\perp\zeta + |\nabla\zeta|^2 - \frac{C}{|\log\varepsilon|},$$

where we used the argument for the estimate of the third line of (2.15) in the proof of Lemma 2.8. By (2.8) and (2.12) we see that ζ is continuous. Since there are no nontrivial-degree vortex balls, then by an argument identical to the proof of (2.17) we have the lower bound

$$G_\varepsilon(u, A) \geq E_\varepsilon(u) + G^0 + \frac{1}{2}\int_U |\nabla\zeta|^2 + \frac{1}{\ell^2}|\Delta\zeta|^2 - \frac{C}{|\log\varepsilon|}.$$

Since (u,A) is an energy minimizer, $G^0 \geq G_\varepsilon(u,A)$, and so $E_\varepsilon(u)+\frac{1}{2}\int_U |\nabla\zeta|^2 \leq \frac{C}{|\log\varepsilon|}$. Even more so, the boundary condition $\zeta = 0$ implies $\zeta \to 0$ and $E_\varepsilon(u) \to 0$ as $\varepsilon \to 0$. We see that

$$G_\varepsilon(u,A) \geq G^0 - \frac{C}{|\log\varepsilon|} \tag{2.23}$$

when $\sum_{j\in J}|d_j| = 0$.

To prove that $G_\varepsilon(u,A)$ is no longer the Meissner state, we construct a sequence of functions $(u_\varepsilon, A_\varepsilon)$ which have lower energy than the Meissner energy when $h_{ex} > \frac{\ell^2|\log\varepsilon|}{2}$. Set $A_\varepsilon = h_{ex}\nabla^\perp\xi_\ell + \nabla^\perp\zeta$, where ξ_ℓ is defined in (2.4) and

$$-\frac{1}{\ell^2}\Delta^2\zeta + \Delta\zeta = 2\pi\delta_a \text{ in } U, \qquad \Delta\zeta = \xi_\ell = 0 \text{ on } \partial U. \tag{2.24}$$

To define $u_\varepsilon = \rho_\varepsilon e^{i\varphi_\varepsilon}$ we set $\nabla\varphi_\varepsilon = A_\varepsilon + \frac{1}{\ell^2}\nabla^\perp \operatorname{curl} A_\varepsilon$ and

$$\rho_\varepsilon = \begin{cases} 0 & |x-a| \leq \frac{\varepsilon}{2}, \\ 1 & |x-a| \geq \varepsilon. \end{cases}$$

Then for any $B_R \supset \{a\}$, $\int_{\partial B_R}\partial_\tau\varphi_\varepsilon = \int_{B_R} h_\varepsilon - \frac{1}{\ell^2}\Delta h_\varepsilon = 2\pi$, which correctly quantizes the phase. A straightforward calculation shows that $E_\varepsilon(u_\varepsilon) \leq \pi\log\frac{\operatorname{diam} U}{\varepsilon} + C \leq \pi|\log\varepsilon| + C$, where C is a fixed constant. The arguments in section 4 contain more refined upper bound calculations; however, they are similar in spirit.

Following Step 1 of the proof of Proposition 2.5 yields

$$\begin{aligned}
&\frac{1}{2}\int_U A_\varepsilon^2 |u_\varepsilon|^2 + \frac{1}{\ell^2}|\operatorname{curl} A_\varepsilon - h_{ex}|^2 \\
&\leq \frac{h_{ex}^2}{2}\int_U |\nabla\xi_\ell|^2 + \frac{1}{\ell^2}|\Delta\xi_\ell - h_{ex}|^2 + \frac{1}{2}\int_U |\nabla\zeta|^2 + \frac{1}{\ell^2}|\Delta\zeta|^2 + \frac{C}{|\log\varepsilon|}.
\end{aligned}$$

Again we decompose $\|\nabla u - i\nabla^\perp\zeta u\|^2_{L^2(U)} = \|\nabla u\|^2_{L^2(U)} - 2\int j(u)\cdot\nabla^\perp\zeta + \|\nabla^\perp\zeta u\|^2_{L^2(U)}$, and a similar calculation as in Step 1 shows

$$\begin{aligned}
G_\varepsilon(u_\varepsilon, A_\varepsilon) &\leq E_\varepsilon(u_\varepsilon) - \int_U j(u_\varepsilon)\cdot\nabla^\perp(h_{ex}\xi_\ell + \zeta) + G^0 \\
&\quad + \frac{1}{2}\int_U |\nabla\zeta|^2 + \frac{1}{\ell^2}|\Delta\zeta|^2 + \frac{C}{|\log\varepsilon|} \\
&\leq G^0 + \pi|\log\varepsilon| - \frac{2\pi h_{ex}}{\ell^2}(1 - o_\ell(1)) + C \\
&\quad + \frac{1}{2}\int_U |\nabla\zeta|^2 + \frac{1}{\ell^2}|\Delta\zeta|^2 + 2\pi\zeta(a),
\end{aligned}$$

where we used (2.7) in the last inequality. Multiplying (2.24) by ζ and integrating over U shows $2\pi\zeta(a) = -\int_U |\nabla\zeta|^2 + \frac{1}{\ell^2}|\Delta\zeta|^2 < 0$; hence $\frac{1}{2}\int_U |\nabla\zeta|^2 + \frac{1}{\ell^2}|\Delta\zeta|^2 + 2\pi\zeta(a) < 0$. Therefore,

$$G_\varepsilon(u_\varepsilon, A_\varepsilon) \leq G^0 + \pi|\log\varepsilon| - \frac{2\pi h_{ex}}{\ell^2}(1 - o_\ell(1)) + C.$$

Since $h_{ex} > \frac{\ell^2|\log\varepsilon|}{2}$, there exists $\delta > 0$, bounded away from zero, such that $\pi|\log\varepsilon| - \frac{2\pi h_{ex}}{\ell^2}(1 - o_\ell(1)) + C < C - \delta|\log\varepsilon| < -\frac{|C|}{2}$ for ε small enough; thus

$$G_\varepsilon(u_\varepsilon, A_\varepsilon) < G^0 - \frac{|C|}{2} < G_\varepsilon(u_\varepsilon, A_\varepsilon).$$

Therefore, a vortex-less configuration *cannot* be minimizing in the $h_{ex} > \frac{\ell^2|\log\varepsilon|}{2}$ regime. □

Remark 2.9. For values of h_{ex} well above the critical field, we expect the minimizers to be similar to the functions constructed in the proof of (4.2) in section 4.

Remark 2.10. The proof of the critical field for $\ell_0 \in [0,+\infty)$ proceeds in the same way as for the proof of Proposition 2.1 and can be done by a suitable modification of the method in [21]. Since we handled the more difficult case $\ell \to +\infty$ such that $\ell_\varepsilon = |\log\varepsilon|^\gamma$ for some $\gamma \in \mathbb{R}^+$, we leave out the proof for the case $\ell_\varepsilon \to \ell_0 \in [0,+\infty)$.

3. Obstacle problem for small and bounded domains. In this section, we study the functional (1.3) where $\ell_\varepsilon \to \ell_0 \in [0,\infty)$, i.e., for domain sizes s that are smaller than or comparable with the penetration depth ($s \ll \lambda$ or $s = O(\lambda)$), in the critical scaling of energy and magnetic field.

The following result is a generalization of Theorem 1.3 in [12] (where it is proved for $\ell = 1$). Closely related results in the context of the Chern–Simons–Higgs energy were shown by the authors in [14, Theorem 1.3] and [15, Theorem 3]. We state the theorem in its gauge-invariant form.

THEOREM 3.1. *Let $(u_\varepsilon, A_\varepsilon)$ be a sequence with $G_\varepsilon(u_\varepsilon, A_\varepsilon) \le K|\log\varepsilon|^2$ and assume that h_{ex} satisfies $\frac{h_{ex}}{|\log\varepsilon|} \to H$ for some $H \ge 0$ and $\ell_\varepsilon \to \ell_0 \in [0,\infty)$. Define the following rescaled quantities:*

$$a_\varepsilon := \frac{1}{|\log\varepsilon|}A_\varepsilon,$$
$$v_\varepsilon := \frac{1}{|\log\varepsilon|}(iu_\varepsilon, \nabla u_\varepsilon),$$
$$w_\varepsilon := v_\varepsilon - |u_\varepsilon|^2 a_\varepsilon.$$

Then $\operatorname{curl} a_\varepsilon$ is weakly compact in $L^2(U)$, and w_ε is weakly compact in L^p for $1 \le p < 2$. Furthermore, $\frac{w_\varepsilon}{|u_\varepsilon|}$ converges weakly in L^2 if and only if w_ε converges weakly, and the weak limits are equal.

Any weak limit of $(w_\varepsilon, \operatorname{curl} a_\varepsilon)$ can be expressed in the form $(v-a, \operatorname{curl} a)$ for some $(v,a) \in L^2(U;\mathbb{R}^2) \times H^1(U;\mathbb{R}^2)$ such that $\operatorname{curl} v$ is a Radon measure. In addition, we have the following $\Gamma - \liminf$ inequality:

$$\liminf_{\varepsilon\to 0} \frac{1}{|\log\varepsilon|^2} G_\varepsilon(u_\varepsilon, A_\varepsilon) \ge G(v,a),$$

where the limit functional G is given by

(3.1)
$$G(v,a) := \begin{cases} \frac{1}{2}\int_U |v-a|^2 + \frac{1}{\ell_0^2}|\operatorname{curl} a - H|^2 + \frac{1}{2}\|\operatorname{curl} v\|_{\mathscr{M}} & \text{if } \ell_0 > 0, \\ \frac{1}{2}\int_U |v-a|^2 + \frac{1}{2}\|\operatorname{curl} v\|_{\mathscr{M}} & \text{if } \ell_0 = 0 \text{ and } \operatorname{curl} a = H, \\ +\infty & \text{otherwise.} \end{cases}$$

Conversely, for every $(v,a) \in L^2(U;\mathbb{R}^2) \times H^1(U;\mathbb{R}^2)$ such that $\operatorname{curl} v$ is a Radon measure there exist approximating sequences $(\tilde{u}_\varepsilon, \tilde{A}_\varepsilon)$ such that the convergences above hold and such that

$$\lim_{\varepsilon\to 0} \frac{1}{|\log\varepsilon|^2} G_\varepsilon(\tilde{u}_\varepsilon, \tilde{A}_\varepsilon) = G(v,a). \tag{3.2}$$

Proof. It suffices to check the theorem for sequences $(u_\varepsilon, A_\varepsilon)$ that satisfy the Coulomb gauge condition $\operatorname{div} A_\varepsilon = 0$ in U, $A \cdot \nu = 0$ on ∂U, since $G(u_\varepsilon, A_\varepsilon) =$

$G(u_\varepsilon e^{i\chi}, A_\varepsilon + \nabla\chi)$ and the quantities w_ε and $\operatorname{curl} a_\varepsilon$ are invariant under this gauge transformation. The limit functional $G(v,a)$ also has the gauge invariance $G(v + \nabla\chi, a + \nabla\chi) = G(v,a)$.

From the energy bound $G_\varepsilon(u_\varepsilon, A_\varepsilon) \le K|\log\varepsilon|^2$ we infer that

$$\int_U |\operatorname{curl} a_\varepsilon - H|^2 \le 2K\ell_\varepsilon^2 \le C,$$

since ℓ_ε is bounded, and together with $\operatorname{div} a_\varepsilon = 0$ this implies $\|a_\varepsilon\|_{H^1(U)} \le C$, and via Sobolev embedding $\|A_\varepsilon\|_{L^p(U)} \le C_p|\log\varepsilon|$ for $p \ge 1$.

We can now establish that the BBH energy $E_\varepsilon(u_\varepsilon)$ is bounded, using the following decomposition:

$$|\nabla_A u|^2 = |\nabla u|^2 - 2j(u)\cdot A + |u|^2|A|^2,$$

which implies that

$$E_\varepsilon(u_\varepsilon) \le G_\varepsilon(u_\varepsilon, A_\varepsilon) + \int_U |j(u_\varepsilon)\cdot A_\varepsilon|.$$

As in [12], we can estimate the cross term via

$$\begin{aligned}
|j(u_\varepsilon)\cdot A_\varepsilon| &\le \frac{1}{4}|\nabla u_\varepsilon|^2 + |u_\varepsilon|^2|A_\varepsilon|^2 \\
&\le \frac{1}{4}|\nabla u_\varepsilon|^2 + \left(|u_\varepsilon|^2 - 1\right)|A_\varepsilon|^2 + |A_\varepsilon|^2 \\
&\le \frac{1}{4}|\nabla u_\varepsilon|^2 + \frac{1}{8\varepsilon^2}\left(1 - |u_\varepsilon|^2\right)^2 + 2\varepsilon^2|A_\varepsilon|^4 + |A_\varepsilon|^2 \\
&\le \frac{1}{2}E_\varepsilon(u_\varepsilon) + C\varepsilon^2|\log\varepsilon|^4 + C|\log\varepsilon|^2,
\end{aligned}$$

and it follows that $E_\varepsilon(u_\varepsilon) \le C|\log\varepsilon|^2$. We are therefore able to use the compactness results of [12] that show compactness for v_ε and the estimate

$$\liminf_{\varepsilon\to 0} E_\varepsilon(u_\varepsilon) \ge \frac{1}{2}\int_U |v|^2 + \frac{1}{2}\|\operatorname{curl} v\|_{\mathscr{M}}.$$

It is then not difficult to show the lower bound for the full energy using the same decomposition as above and the weak convergence of a_ε implied by the bounds.

The Γ-limsup property (3.2) can be shown as follows: Given a limit (v,a) with $\operatorname{div} a = 0$, we set $\tilde{A}_\varepsilon = a|\log\varepsilon|$ and construct $\tilde{u}_\varepsilon$ as in [12, section 7]. It is then easy to see that the claimed convergence holds, using the Γ-convergence result for E_ε from [12] and the same decomposition as above. □

Remark 3.2. Note that compactness for v_ε only holds due to our choice of gauge. The representative $\tilde{u}_\varepsilon = u_\varepsilon e^{i\chi_\varepsilon}$ corresponds to $\tilde{v}_\varepsilon = v_\varepsilon + \frac{1}{|\log\varepsilon|}\nabla\chi_\varepsilon$, and so v_ε and $\tilde{v}_\varepsilon$ need not have the same compactness properties. The limit functional $G(v,a)$ also has the gauge invariance $G(v + \nabla\chi, a + \nabla\chi) = G(v,a)$. If $\ell_\varepsilon \to \infty$, the compactness argument for a_ε fails, since we only know that $\int_U |\operatorname{curl} a_\varepsilon - H|^2 \le K\ell_\varepsilon^2$, so this sequence need not be bounded. The example given (for $H = 0$) in [15, Theorem 5], which can be used for (1.3), shows that v_ε also need not be compact in this case, even if $\operatorname{div} A_\varepsilon = 0$. In fact we construct a sequence of $(v_\varepsilon, a_\varepsilon)$ with bounded energy but $\|v_\varepsilon\|_{L^2(U)} \gtrsim \log(\ell_\varepsilon \wedge \frac{1}{\varepsilon|\log\varepsilon|^{1/2}}) \to +\infty$ by constructing a set of vortices that

concentrate about a single point. Therefore, the energy splitting approach of [12] is insufficient to treat the case of large domains.

As in [20], we can characterize the minimizers of the limit functional. We obtain, following the presentation of [12], Theorem 1.3.

Proof of Theorem 1.3. We prove only the part for $\ell_0 = 0$; the first half can be shown by a completely straightforward insertion of ℓ^{-2} into the argument of [12]. Our proof of the second half also follows the structure of their argument.

For a Radon measure $\mu \in H^{-1}$, define the vector field $v^\mu \in L^2(U;\mathbb{R}^2)$ by $\operatorname{curl} v^\mu = \mu$ and $\operatorname{div} v^\mu = 0$. We decompose μ as $\mu^{ac} + \mu^{sing}$ into an absolutely continuous and a singular part. Setting $g(t,\mu) = G(v_0 + tv_\mu, a)$, we calculate

$$0 \le \lim_{t\to 0^+} \frac{g(t,\mu) - g(0,\mu)}{t} = \int_U (v_0 - a_0, v^\mu) + \frac{1}{2}\int_U \operatorname{sgn}(\mu_0) d\mu^{ac} + \frac{1}{2}\left\|\mu^{sing}\right\|(U).$$

Integrating by parts and using the definition of y_0, we see that

$$\int_U (v_0 - a_0, v^\mu) = \int_U y_0 \mu,$$

so we obtain

$$0 \le \int_U \left(y_0 + \frac{1}{2}\operatorname{sgn}(\mu_0)\right) d\mu^{ac} + \int_U \left(y_0 + \frac{1}{2}\operatorname{sgn}(\mu)\right) d\mu^{sing} \tag{3.3}$$

and similarly by one-sided differentiation in the opposite direction,

$$0 \ge \int_U \left(y_0 + \frac{1}{2}\operatorname{sgn}(\mu_0)\right) d\mu^{ac} + \int_U \left(y_0 - \frac{1}{2}\operatorname{sgn}(\mu)\right) d\mu^{sing}. \tag{3.4}$$

Together, (3.3) and (3.4) imply, due to the arbitrariness of μ^{ac} and μ^{sing}, that $|y_0| \le \frac{1}{2}$ everywhere and $y_0 = -\frac{1}{2}\operatorname{sgn}(\mu_0)$ in $\operatorname{spt}\mu_0$. It follows that for any smooth function φ with $\varphi(z) = 0$ for $z \le 0$ and $\varphi'(z) \ge 0$ there holds

$$\int_U \varphi(y_0) d\mu_0 = -\varphi\left(\frac{1}{2}\right)\mu_0^-(U),$$

where μ_0^- denotes the negative part in the Hahn decomposition of μ_0. Since $\mu_0 = -\Delta y_0 + H$, we can integrate by parts and obtain

$$\int_U \varphi(y_0) d\mu_0 = \int_U \varphi'(y_0)|\nabla y_0|^2 + \varphi(y_0)H \ge 0,$$

and we conclude that $\mu_0^-(U) = 0$ and so $\mu_0 \ge 0$.

To see that y_0 is a solution of the obstacle problem, we take any $y \in \mathscr{K}$ and compare using $|v|^2 - |w|^2 \ge 2(v-w)\cdot w$ and integration by parts:

$$F_{0,H}(y) - F_{0,H}(y_0) \ge \int_U \nabla(y - y_0)\cdot\nabla y_0 + (y - y_0)H = \int_U (y - y_0) d\mu_0.$$

Now $y_0 = -\frac{1}{2}$ on $\operatorname{spt}(\mu_0)$, so $(y - y_0) \ge 0$ on $\operatorname{spt}(\mu_0)$ for all $y \in \mathscr{K}$. It follows that y_0 is a minimizer of the obstacle problem. Standard theory [9] can now be used to show uniqueness of y_0. □

Remark 3.3. Another proof of Theorem 1.3 can be given in the framework of Brezis and Serfaty [7], who examine the obstacle problem arising from (1.7).

COROLLARY 3.4. *Let (v_0, a_0) be a minimizer of $G(v,a)$. Then* $\operatorname{curl} v_0 = 0$ *for* $H < H_1(\ell_0)$ *and* $\operatorname{curl} v_0 \neq 0$ *for* $H > H_1(\ell_0)$*, where $H_1(\ell_0)$ is given by*

$$H_1(\ell_0) = \frac{1}{2\max_{\overline{U}} |y_{\ell_0}|}, \tag{3.5}$$

where y_{ℓ_0} is the solution of

$$-\Delta y_{\ell_0} + \ell_0^2 y_{\ell_0} + 1 = 0$$

with Dirichlet boundary conditions $y_{\ell_0} = 0$ on ∂U.

Remark 3.5. We reiterate that the function H_1 in (3.5) satisfies $\frac{2H_1(\ell_0)}{\ell_0^2} \to 1$ as $\ell_0 \to \infty$.

Remark 3.6. In the case where $U = B_1(0)$ is a ball, the function H_1 can be written down explicitly since the solutions of $-\Delta y + \alpha y + 1 = 0$ with Dirichlet boundary conditions are given by known special functions. Denoting by I_0 the modified Bessel function of zeroth order, we have that

$$H_1(\ell_0) = \frac{\ell_0^2 I_0(\ell_0)}{2(I_0(\ell_0) - 1)}.$$

Since $I_0(x) \sim \frac{e^x}{\sqrt{2\pi x}}$ as $x \to \infty$ and $I_0(x) = 1 + \frac{x^2}{4} + O(x^4)$ as $x \to 0$, it is easy to see that this matches the claimed behavior at zero and infinity.

4. Upper bound for vortex lattices. In this section, we construct good comparison sequences that correspond to vortex lattices and calculate their energy.

PROPOSITION 4.1. *Assume $\varepsilon < \frac{C}{\sqrt{h_{ex}}}$ and $\varepsilon \to 0$. There exists a sequence of functions $(u_\varepsilon, A_\varepsilon)$ such that the Ginzburg–Landau energy satisfies*

$$G_\varepsilon(u_\varepsilon, A_\varepsilon) \leq h_{ex}\frac{|U|}{2}\left(\log\frac{1}{\sqrt{h_{ex}}\varepsilon} + C\right). \tag{4.1}$$

If $h_{ex} - \frac{\ell^2|\log\varepsilon|}{2} = S \gg 1$ and $h_{ex} \leq \frac{1}{\varepsilon^2}$, then there exists a sequence of functions with Ginzburg–Landau energy

$$G_\varepsilon(u_\varepsilon, A_\varepsilon) \leq \frac{|U|}{2}\left(S\left(\left|\log\left(\varepsilon\max(\sqrt{S}, \ell)\right)\right| + C\right) + \frac{\ell^2}{4}|\log\varepsilon|^2\right). \tag{4.2}$$

Remark 4.2. The bound given in (4.2) is better than that of (4.1). It is essentially equivalent to the one given for $\ell = 1$ in Proposition 5.8 of Aydi's thesis [3]. Aydi's upper bound, adapted to our setting, reads as

$$G_\varepsilon(u_\varepsilon, A_\varepsilon) \leq \frac{|U|}{2}\left(\left(\frac{1}{2}|\log\varepsilon| + S\right)^2 - \frac{1}{2}S^2\right),$$

and, expanding the square, this is essentially (4.2).

Remark 4.3. If $\ell^2 \geq Kh_{ex}$, then

$$G_\varepsilon(1,0) \leq h_{ex}\frac{|U|}{2K}.$$

In particular, there is a constant $K > 0$ such that our vortex lattice construction is not minimizing for $\ell^2 \geq Kh_{ex}$.

Remark 4.4. Under the assumptions for the upper bound (4.2), the trivial Meissner-like state $(u, A) = (1, 0)$ has the energy

$$\frac{|U|}{2}\left(\frac{S^2}{\ell^2} + S|\log\varepsilon| + \frac{\ell^2}{4}|\log\varepsilon|^2\right).$$

Since $|\log(\varepsilon\max(\ell, \sqrt{S}))| \leq |\log\varepsilon| - C$, the vortex lattice state with $O(S)$ vortices is energetically favorable compared to $(u, A) = (1, 0)$. Consult Proposition 2.1 for a more detailed statement regarding the first critical field for vortex nucleation.

We now turn to the proof of Proposition 4.1. We present a new elementary approach using Fourier series to estimate the energy of a vortex lattice. This approach shows the cross-over that happens at $L = \ell^{-1}$; this is related to the decay properties of the Bessel function–type solutions. On the unit cell of our lattice, we investigate solutions of

$$-\alpha\Delta h + h = 2\pi\delta_\varepsilon \quad \text{in } K_L \tag{4.3}$$

with homogeneous Neumann boundary conditions. Here $K_L = \left(-\frac{L}{2}, \frac{L}{2}\right)^2$ for some $L > 0$. This is equivalent to looking at $L\mathbb{Z}^2$-periodic solutions in $\mathbb{R}^2$. For δ_ε we use the Dirac sequence

$$\delta_\varepsilon(x) = \frac{1}{4\varepsilon^2}\chi_{(-\varepsilon,\varepsilon)}(x_1)\chi_{(-\varepsilon,\varepsilon)}(x_2),$$

where χ_A is the characteristic function of a set $A \subset \mathbb{R}$. We assume $2\varepsilon < L$. We obtain the following results on the lattice.

PROPOSITION 4.5. *There exists a $C > 0$ such that for any L, ε with $\varepsilon < \frac{L}{2}$ there exists a periodic function h such that $-\ell^{-2}\Delta h + h = 2\pi\delta_\varepsilon$ and*

$$\int_{K_L} \ell^{-4}|\nabla h|^2 + \ell^{-2}|h - h_{ex}|^2 \leq 2\pi\log\frac{1}{\max(\ell, L^{-1})\varepsilon} + C + L^2\left(h_{ex} - \frac{2\pi}{L^2}\right)^2. \tag{4.4}$$

Proof. We calculate the energy

$$\int_{K_L} \alpha^2|\nabla h|^2 + \ell^{-2}|h - h_{ex}|^2.$$

It will become apparent later that we should use $\alpha = \ell^{-2}$.

We use double Fourier series as follows. For $f \in L^2(K_L)$ and $k \in \mathbb{Z}^2$, set

$$a_k = \frac{1}{L^2}\int_{K_L} f(x)e^{-i\gamma k\cdot x},$$

where $\gamma = \frac{2\pi}{L}$. Then f can be reconstructed as

$$f(x) = \sum_{k\in\mathbb{Z}^2} a_k e^{i\gamma k\cdot x}.$$

By Plancherel's theorem we have

$$\int_{K_L} |f|^2 = L^2\sum_{k\in\mathbb{Z}^2} |a_k|^2.$$

It is standard that ∇f corresponds to the series $(i\gamma k a_k)$, and Δf to the series $(-\gamma^2|k|^2 a_k)$. Solving (4.3) therefore corresponds to

$$(\alpha\gamma^2|k|^2+1)a_k = b_k,$$

where b_k are the Fourier coefficients for δ_ε.

We calculate these coefficients. Set $k=(k_1,k_2)$. If $k_1k_2\neq 0$, then

$$b_k = \frac{2\pi}{4\varepsilon^2L^2}\frac{4\sin(\gamma k_1\varepsilon)\sin(\gamma k_2\varepsilon)}{\gamma^2 k_1k_2}.$$

In other cases we have

$$b_{(m,0)} = b_{(0,m)} = \frac{2\pi}{4\varepsilon^2L^2}\frac{4\varepsilon\sin(\gamma m\varepsilon)}{\gamma m}$$

and finally $b_0 = \frac{2\pi}{L^2}$ corresponding to $\int_{K_L} h = 2\pi$.

To simplify notation we write this using $\operatorname{sinc}(x)$, the continuous continuation of $\frac{\sin x}{x}$, which yields

$$b_k = \frac{2\pi}{L^2}\operatorname{sinc}(\gamma k_1\varepsilon)\operatorname{sinc}(\gamma k_2\varepsilon).$$

Since $a_0 = b_0 = \frac{2\pi}{L^2}$, we have that

$$\int |h-h_{ex}|^2 = L^2\sum_{k\neq 0}|a_k|^2 + L^2\left(\frac{2\pi}{L^2}-h_{ex}\right)^2.$$

We want to calculate

$$E = L^2\sum_{k\in\mathbb{Z}^2\setminus 0}(\alpha^2\gamma^2|k|^2+\alpha)|a_k|^2 + L^2\left(\frac{2\pi}{L^2}-h_{ex}\right)^2.$$

Using the expressions obtained for b_k above, it follows that we have

$$E = \frac{4\pi^2}{L^2}\sum_{k\in\mathbb{Z}^2\setminus 0}\frac{\alpha}{\alpha|\gamma|^2|k|^2+1}\operatorname{sinc}^2(\gamma\varepsilon k_1)\operatorname{sinc}^2(\gamma\varepsilon k_2) + L^2\left(\frac{2\pi}{L^2}-h_{ex}\right)^2.$$

We split up the double sum as follows. First, consider $k=(k_1,k_2)$ with $1\leq|k|\leq\frac{1}{\gamma\varepsilon}$. For these terms we estimate $|\operatorname{sinc}|\leq 1$. We label this part of the energy E_1, and so

$$E_1 \leq \frac{4\pi^2}{L^2}\sum_{K_1}\frac{\alpha}{\alpha\gamma^2|k|^2+1}.$$

Now we compare the sum with an integral. For any decreasing function f, we have

$$\sum_{1\leq|k|\leq A} f(|k|) \leq \int_{1-\frac{1}{\sqrt{2}}}^{A+\frac{1}{\sqrt{2}}} f(r)2\pi r\,dr$$

and so

$$E_1 \leq \frac{4\pi^2}{L^2}\int_{1-c}^{\frac{1}{\gamma\varepsilon}+c}\frac{2\pi\alpha r}{1+\alpha\gamma^2r^2}dr = \frac{4\pi^2}{L^2\gamma^2}\left(\frac{2\pi}{2}\log(\alpha\gamma^2x^2+1)\right)\Bigg|_{x=1-c}^{x=\frac{1}{\gamma\varepsilon}+c},$$

where $c = \frac{1}{\sqrt{2}}$. As $\frac{4\pi^2}{L^2\gamma^2} = 1$, we obtain

$$E_1 \le \frac{2\pi}{2} \log \frac{\frac{\alpha}{\varepsilon^2} + \frac{2c\alpha\gamma}{\varepsilon} + c^2 + 1}{\alpha\gamma^2(1-c)^2 + 1}.$$

We distinguish two cases. If $\alpha\gamma^2 \le 1$, we estimate the denominator as ≥ 1 and obtain

$$E_1 \le \frac{2\pi}{2} \log \frac{\alpha}{\varepsilon^2} + C \le 2\pi \log \frac{\sqrt{\alpha}}{\varepsilon} + C.$$

In the case where $\alpha\gamma^2 \ge 1$, we estimate the denominator as $\ge C\alpha\gamma^2$ and obtain

$$E_1 \le \frac{2\pi}{2} \log \frac{1}{\gamma^2\varepsilon^2} + C \le 2\pi \log \frac{1}{\gamma\varepsilon} + C$$

if $\alpha\gamma > \varepsilon$.

We still need to deal with the frequencies k with $|k| \ge \frac{1}{\gamma\varepsilon}$. For this we use that $\operatorname{sinc}^2(x)\operatorname{sinc}^2(y) \le \frac{2}{r^2}$, which can be seen as follows. Assume without loss of generality that $|x| \le |y|$. Then $r^2 = x^2 + y^2 \le 2y^2$. Estimating $|\operatorname{sinc}(x)| \le 1$ and $|\operatorname{sinc}(y)| \le \frac{1}{y}$, we see that $\operatorname{sinc}^2(x)\operatorname{sinc}^2(y) \le \frac{1}{y^2} \le \frac{2}{r^2}$, as claimed.

To calculate the energy contribution E_2 of those k with $|k| \ge \frac{1}{\gamma\varepsilon}$, we again replace the sum by an integral. Using the sinc bound, we see that

$$E_2 \le \frac{4\pi^2}{L^2} \int_{\frac{1}{\gamma\varepsilon}-c}^{\infty} \frac{4\pi\alpha r}{(\alpha r^2\gamma^2 + 1)\gamma^2 r^2 \varepsilon^2}.$$

We estimate this using $4\pi^2 L^{-2}\gamma^{-2} = 1$ as

$$E_2 \le C \int_{\frac{1}{\gamma\varepsilon}-c}^{\infty} \frac{1}{\gamma^2\varepsilon^2 r^3} dr \le \frac{C}{\gamma^2\varepsilon^2} \cdot \frac{1}{(\frac{1}{\gamma\varepsilon} - c)^2},$$

and for $\frac{1}{\gamma\varepsilon} > 2c$ we obtain that $E_2 \le C$. The claim then follows using the definitions. ∎

Choosing $L = \frac{2\pi}{h_{ex}}$ in (4.4) implies the following upper bound.

COROLLARY 4.6. *If $\varepsilon < \frac{1}{2\sqrt{h_{ex}}}$, then there exists a periodic function h with period $L = \sqrt{\frac{2\pi}{h_{ex}}}$ such that $-\ell^{-2}\Delta h + h = \delta_\varepsilon$ and*

$$\int_{K_L} \ell^{-4}|\nabla h|^2 + \ell^{-2}|h - h_{ex}|^2 \le 2\pi \log \frac{1}{\max(\ell, \sqrt{h_{ex}})\varepsilon} + C.$$

Furthermore, for any $L > 2\varepsilon$ there exists h with $-\ell^{-2}\Delta h + h = \delta_\varepsilon$ and

$$\int_{K_L} \ell^{-4}|\nabla h|^2 + \ell^{-2}|h - h_{ex}|^2 \le 2\pi \log \frac{1}{\max(\ell, \sqrt{h_{ex}})\varepsilon} + C + L^2\ell^{-2}\left(\frac{2\pi}{L^2} - h_{ex}\right)^2.$$

Remark 4.7. This can be easily extended to $\varepsilon \le \frac{C}{\sqrt{h_{ex}}}$ for any C that is bounded independently of ε, ℓ, and h_{ex} by choosing $\tilde{\varepsilon} = \frac{2\varepsilon}{C}$ and constructing with $\tilde{\varepsilon}$ instead of ε.

To construct a pair (u, A) from h, we do the following. To define the modulus ρ, we set

$$\rho(r) = \begin{cases} 0, & r < \varepsilon\sqrt{2}, \\ \frac{r-\varepsilon\sqrt{2}}{\varepsilon}, & \varepsilon\sqrt{2} < r < \varepsilon(1+\sqrt{2}), \\ 1, & r > \varepsilon(1+\sqrt{2}). \end{cases}$$

We take any A with $\operatorname{curl} A = h$. Outside $B_{\varepsilon\sqrt{2}}$, we define u as $\rho e^{i\varphi}$, where $\nabla\varphi - A = \alpha\nabla^\perp h$. This is possible since for any simple closed curve $\Gamma \subset K_L \setminus B_{\varepsilon\sqrt{2}}$ with $\Gamma = \partial G$ we have

$$\int_\Gamma \frac{\partial\varphi}{\partial\tau} = \int_\Gamma A\cdot\tau - \alpha\frac{\partial h}{\partial\nu} = \int_G(-\Delta h + h) = \begin{cases} 2\pi & \text{if } G \supset B_{\varepsilon\sqrt{2}}, \\ 0 & \text{otherwise.} \end{cases} \tag{4.5}$$

On K_L, we can therefore estimate

$$\begin{aligned} \frac{1}{2}\int_{K_L} |(\nabla - iA)u|^2 + \ell^{-2}|h - h_{ex}|^2 + \frac{1}{2\varepsilon^2}(1-\rho^2)^2 \\ \le \frac{1}{2}\int_{K_L} \rho^2\ell^{-4}|\nabla h|^2 + \ell^{-2}|h-h_{ex}|^2 + |\nabla\rho|^2 + \frac{1}{2\varepsilon^2}(1-\rho^2)^2 \\ \le C + \pi\log\frac{1}{\max(\ell,\sqrt{h_{ex}})\varepsilon}. \end{aligned} \tag{4.6}$$

We are now in the position to establish the following argument.

Proof of Proposition 4.1. This will be done in two steps.

Step 1. We use the above construction to build an h in $\mathbb{R}^2$ and to define a periodic ρ_ε corresponding to the lattice. As the equivalent of (4.5) holds in all of $\mathbb{R}^2$, we can define (u, A) in all of $\mathbb{R}^2$ such that (4.6) holds on every cell of the lattice. All we need to do is choose a proper origin for our lattice: For any $a \in K_L$ we can set $(u^a(z), A^a(z)) = (u(z-a), A(z-a))$, which has energy density $gl^a(z) = gl(z-a)$, where $gl(z) = \frac{1}{2}|(\nabla - iA)u|^2(z) + \ell^{-2}|\operatorname{curl} A(z) - h_{ex}|^2 + \frac{1}{2\varepsilon^2}(1-|u(z)|^2)^2$. Integrating over the unit cell, we see that

$$\int_{K_L} G_\varepsilon(u_\varepsilon^a, A_\varepsilon^a; U)da = \int_{K_L}\int_U gl^a(z)dzda = |U|G_\varepsilon(u, A; K_L).$$

The mean value theorem shows that there exists some a such that $G(u^a, A^a; U) \le \frac{|U|}{|K_L|}G(u, A; K_L)$, and since $|K_L| = \frac{h_{ex}}{2\pi}$, this finishes the proof of (4.1).

Step 2. We follow the argument in Step 1; however, we choose a lattice of size $L = \sqrt{\frac{2\pi}{S}}$, which is optimal up to logarithmic terms. Since $h_{ex} \le \varepsilon^{-2}$ and $S \gg 1$, we have $2\varepsilon \le L \ll 1$, and we can follow the same construction as above and obtain for the energy after choosing a suitable origin

$$\begin{aligned} G(u_\varepsilon, A_\varepsilon; U) &\le \frac{|U|}{2|K_L|}\left(2\pi\log\frac{1}{\max(\sqrt{S},\ell)\varepsilon} + C + L^2\ell^{-2}\left|h_{ex} - \frac{2\pi}{L^2}\right|\right) \\ &= \frac{|U|}{2}\left(\frac{S}{2\pi}\left(2\pi\log\frac{1}{\max(\sqrt{S},\ell)\varepsilon} + C\right) + \ell^{-2}\left(\frac{1}{2}\ell^2|\log\varepsilon|\right)^2\right) \\ &= \frac{|U|}{2}\left(S\log\frac{1}{\max(\sqrt{S},\ell)\varepsilon} + C + \frac{1}{4}\ell^2|\log\varepsilon|^2\right) \end{aligned}$$

since $h_{ex} - \frac{2\pi}{L^2} = h_{ex} - S = \frac{1}{2}\ell^2|\log\varepsilon|$. This completes the proof of (4.2). ∎

5. Lower bound for vortex lattices. The lower bound counterpart to Proposition 4.1 can be obtained similarly to the calculation in [18] (see also [19]). We can actually get a uniform vortex density estimate.

The heart of the lower bound argument is the following blow-up estimate.

PROPOSITION 5.1. *Assume that* $\ell^2|\log\varepsilon| \ll h_{ex} \ll \frac{1}{\varepsilon^2}$, *and let* $(u_\varepsilon, A_\varepsilon)$ *be a minimizer of the Ginzburg–Landau energy* (1.3). *Then for every* $H > 0$ *we can find* λ *with* $\ell \ll \lambda \ll \frac{1}{\varepsilon}$ *such that for every* x *such that the ball* $B(x, \frac{1}{\lambda}) \subset U$ *there holds*

$$G_\varepsilon\left(u_\varepsilon, A_\varepsilon; B\left(x, \frac{1}{\lambda}\right)\right) \geq \frac{\alpha_H |B(x, \frac{1}{\lambda})|}{2} h_{ex} \log \frac{1}{\varepsilon\sqrt{h_{ex}}}(1 - o_\varepsilon(1)), \tag{5.1}$$

where α_H *satisfies* $\alpha_H \to 1$ *for* $H \to \infty$.

Proof. We follow the proof of Proposition 8.2 in [18] and rescale by λ. Setting $u(x) = u'(\lambda x)$, $A'(x) = \lambda A(\lambda x)$, $\varepsilon' = \lambda\varepsilon$, and $h'_{ex} = \frac{h_{ex}}{\lambda^2}$, the claim is equivalent to proving for a ball B_1 of radius 1 that

$$\frac{1}{2}\int_{B_1} |(\nabla - iA')u'|^2 + \frac{\lambda^2}{\ell^2}|\operatorname{curl} A' - h'_{ex}|^2 + \frac{1}{2\varepsilon'^2}(1 - |u'|^2)^2 \geq \frac{\alpha_H |B_1|}{2} h'_{ex} \log \frac{1}{\varepsilon\sqrt{h_{ex}}}(1 - o_\varepsilon(1)). \tag{5.2}$$

Since $\varepsilon^2\ell^2|\log\varepsilon| \ll \varepsilon^2 h_{ex} \ll 1$ and using the asymptotic behavior of the function $x \mapsto Hx^2 \log\frac{1}{x}$ near 0, we can find λ such that $\varepsilon^2 h_{ex} = H(\varepsilon\lambda)^2 \log\frac{1}{\varepsilon\lambda}$. It also follows that λ satisfies $\ell \ll \lambda \ll \frac{1}{\varepsilon}$.

With this choice of λ, we have $\varepsilon' \to 0$ and $h'_{ex} = H|\log\varepsilon'|$. We also have $\log\frac{1}{\varepsilon\sqrt{h_{ex}}} = |\log\varepsilon'|(1 - o_\varepsilon(1))$, and so it suffices to prove that (dropping some ε's)

$$G'_{\varepsilon'}(u', A'; x) = \frac{1}{2}\int_{B_1(x)} |(\nabla - iA')u'|^2 + \frac{\lambda^2}{\ell^2}|\operatorname{curl} A' - h'_{ex}|^2 + \frac{1}{2\varepsilon'^2}(1 - |u'|^2)^2 \geq \frac{H\alpha_H |B_1|}{2}|\log\varepsilon'|^2(1 - o_\varepsilon(1)) \tag{5.3}$$

for some α_H with $\alpha_H \to 1$ as $H \to \infty$.

Depending on the blow-up origin x, we distinguish two cases. Either $G'_{\varepsilon'}(u', A'; x) \gg |\log\varepsilon'|^2$ (in which case (5.3) is true trivially) or $G'_{\varepsilon'}(u', A'; x) \leq K|\log\varepsilon'|^2$. In the latter case we can use the Gamma-convergence result Theorem 3.1 to see that $\frac{1}{|\log\varepsilon'|}(iu'_\varepsilon, \nabla u'_\varepsilon) \rightharpoonup v$ and $\frac{1}{|\log\varepsilon'|}A'_\varepsilon \rightharpoonup a$. As $\lambda \gg \ell$, we are in the case $\ell_0 = 0$.

Letting (v_0, a_0) denote the minimizer of the functional G defined in (3.1), we now use the characterization of the limit in Theorem 1.3. It follows that $\operatorname{curl} a_0 = H$. Using Lemma 5.3 below, $\operatorname{curl} v_0 = H$ on $B_{r_0(H)}$ with $r_0(H) \to 1$ for $H \to \infty$. As $\operatorname{curl} v_0 \geq 0$ by Theorem 1.3, it follows that

$$\liminf_{\varepsilon\to 0} \frac{1}{|\log\varepsilon'|^2} G'_{\varepsilon'}(u', A'; x) \geq G(v, a) \geq G(v_0, a_0) \geq \frac{1}{2}\|\operatorname{curl} v_0\|_{\mathscr{M}} \geq H\frac{|B_{r_0(H)}|}{2} = H\alpha_H\frac{|B_1|}{2}, \tag{5.4}$$

and $\alpha_H \to 1$ by Lemma 5.3. ∎

Combining this with the result of the previous section, we obtain the following claim.

PROPOSITION 5.2. *Assume that* $\ell^2|\log\varepsilon| \ll h_{ex} \ll \frac{1}{\varepsilon^2}$, *and let* $(u_\varepsilon, A_\varepsilon)$ *be a minimizer of the Ginzburg–Landau energy* (1.3). *Then the energy density*

$$g_\varepsilon = \frac{1}{2}\left(|\nabla_{A_\varepsilon} u_\varepsilon|^2 + \frac{1}{\ell^2}|\operatorname{curl} A_\varepsilon - h_{ex}|^2 + \frac{1}{2\varepsilon^2}(1-|u_\varepsilon|^2)^2\right)$$

satisfies

$$\frac{g_\varepsilon}{h_{ex}\log\frac{1}{\varepsilon\sqrt{h_{ex}}}} \rightharpoonup \mathscr{L}^2$$

in the sense of measures, and the energy satisfies

$$G_\varepsilon(u_\varepsilon, A_\varepsilon) = \frac{1}{2}\,|U|\,h_{ex}\log\frac{1}{\varepsilon\sqrt{h_{ex}}}\,(1-o_\varepsilon(1))\,. \tag{5.5}$$

Proof. As in [18], this follows by integrating (5.1) over the domain. If $W \subset\subset U$ is a compactly contained subdomain, then we can use Fubini's theorem and estimate

$$\begin{aligned}\int_W G_\varepsilon\left(u_\varepsilon, A_\varepsilon; B\left(x,\frac{1}{\lambda}\right)\cap W\right)dx &= \int_{x\in W}\int_{y\in B(x,\frac{1}{\lambda})\cap W} g_\varepsilon\,dy\,dx \\ = \int_{y\in B(x,\frac{1}{\lambda})\cap W}\int_{x\in W} g_\varepsilon\,dx\,dy &= \int_{y\in W}\left|B\left(y,\frac{1}{\lambda}\right)\cap W\right| g_\varepsilon\,dy \le \frac{\pi}{\lambda^2}G_\varepsilon(u_\varepsilon, A_\varepsilon; W).\end{aligned} \tag{5.6}$$

Using now (5.1) and Fatou's lemma, we continue to estimate

$$\begin{aligned}\liminf_{\varepsilon\to0}\frac{G_\varepsilon(u_\varepsilon, A_\varepsilon; W)}{h_{ex}\log\frac{1}{\varepsilon\sqrt{h_{ex}}}} &\ge \liminf_{\varepsilon\to0}\int_{x\in W}\frac{\lambda^2 G_\varepsilon(u_\varepsilon, A_\varepsilon; B(x,\frac{1}{\lambda}\cap W))}{\pi h_{ex}\log\frac{1}{\varepsilon\sqrt{h_{ex}}}} \\ &\ge \int_{x\in W}\liminf_{\varepsilon\to0}\left(\chi_{\{x:B(x,\frac{1}{\lambda})\subset W\}}\frac{G_\varepsilon(u_\varepsilon, A_\varepsilon; B(x,\frac{1}{\lambda}))}{|B(x,\frac{1}{\lambda})|h_{ex}\log\frac{1}{\varepsilon\sqrt{h_{ex}}}}\right) \ge \frac{1}{2}|W|.\end{aligned} \tag{5.7}$$

Using the upper bound of the previous section, we can deduce the existence of a weak limit g of $\frac{g_\varepsilon}{h_{ex}\log\frac{1}{\varepsilon\sqrt{h_{ex}}}}$. Using an approximation result, we see from (5.7) that $g \ge \frac{1}{2}\mathscr{L}^2$, and by the upper bound, equality follows. ∎

LEMMA 5.3. *For the domain* $U = B_1(0)$, *the unique minimizer* y_0 *of*

$$F_{0,H}(y) = \frac{1}{2}\int_{B_1(0)} |\nabla y|^2 + 2yH$$

in the admissible class

$$\mathscr{K} = \left\{y \in H^1_0(B_1(0)) : y \ge -\frac{1}{2} \text{ a.e. in } B_1(0)\right\}$$

satisfies $y_0 \equiv -\frac{1}{2}$ *in* $B_{r_0(H)}$ *with* $r_0(H) \to 1$ *as* $H\to\infty$. *Furthermore, the minimizer* (v_0, a_0) *of the functional* G *given in* (3.1) *satisfies* $\operatorname{curl} v_0 = H$ *on* $B_{r_0}(0)$, *and* $\alpha_H = \frac{\|\operatorname{curl} v_0\|_{\mathscr{M}}}{H|B_1(0)|} \to 1$ *as* $H\to\infty$.

Proof. Using the uniqueness and regularity properties [9], we can calculate the solution of the obstacle problem explicitly: It is given by

$$y_0(r) = \max\left\{H\left(\frac{r^2}{4} - \frac{1}{4} + \frac{r_0^2}{2}\log r\right), -\frac{1}{2}\right\},$$

where r_0 is chosen such that

$$H\left(\frac{r_0^2}{4}-\frac{1}{4}+\frac{r_0^2}{2}\log r_0\right)=-\frac{1}{2},$$

which we rewrite as

$$1-\frac{2}{H}=r_0^2(1+2\log r_0). \tag{5.8}$$

From (5.8), it is easy to see that $r_0\to 1$ as $H\to\infty$.

In B_{r_0}, $y_0\equiv -\frac{1}{2}$, so $\Delta y_0=0$, and from Theorem 1.3 we infer $\operatorname{curl} v_0=H$ in B_{r_0}, and $\operatorname{curl} v_0\geq 0$ elsewhere, which shows the claim. □

Acknowledgments. The authors wish to thank the referees for many helpful suggestions, including the approach to Proposition 5.2, that significantly improved the paper.

REFERENCES

[1] A. Aftalion and E.N. Dancer, *On the symmetry and uniqueness of solutions of the Ginzburg-Landau equations for small domains*, Commun. Contemp. Math., 3 (2001), pp. 1–14.

[2] A. Aftalion and Q. Du, *The bifurcation diagrams for the Ginzburg-Landau system of superconductivity*, Phys. D, 163 (2002), pp. 94–105.

[3] H. Aydi, *Vorticité dans le Modèle de Ginzburg-Landau de la Supraconductivité*, Ph.D. thesis, Department of Mathematics, Paris XII, Paris, France, 2004.

[4] H. Aydi and E. Sandier, *Vortex analysis of the periodic Ginzburg–Landau model*, Ann. Inst. H. Poincare Anal. Non Lineaire, to appear.

[5] F. Bethuel, H. Brezis, and F. Hélein, *Ginzburg-Landau Vortices*, Progr. Nonlinear Differential Equations Appl. 13, Birkhäuser Boston, Cambridge, MA, 1994.

[6] F. Bethuel and T. Riviere, *Vortices for a variational problem related to superconductivity*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 12 (1995), pp. 243–303.

[7] H. Brezis and S. Serfaty, *A variational formulation for the two-sided obstacle problem with measure data*, Comm. Contemp. Math., 4 (2002), pp. 357–374.

[8] G. Chapman, Q. Du, M. Gunzburger, and J. Peterson, *Simplified Ginzburg-Landau models for superconductivity valid for high kappa and high fields*, Adv. Math. Sci. Appl., 5 (1995), pp. 193–218.

[9] A. Friedman, *Variational Principles and Free-boundary Problems*, John Wiley & Sons, New York, 1982.

[10] R.L. Jerrard, *Lower bounds for generalized Ginzburg–Landau functionals*, SIAM J. Math. Anal., 30 (1999), pp. 721–746.

[11] R.L. Jerrard and H.M. Soner, *The Jacobian and the Ginzburg-Landau energy*, Calc. Var. Partial Differential Equations, 14 (2002), pp. 151–191.

[12] R.L. Jerrard and H.M. Soner, *Limiting behavior of the Ginzburg-Landau functional*, J. Funct. Anal., 192 (2002), pp. 524–561.

[13] R.L. Jerrard and D. Spirn, *Refined Jacobian estimates and the Ginzburg-Landau energy*, Indiana Univ. Math. J., 56 (2007), pp. 135–186.

[14] M. Kurzke and D. Spirn, *Gamma limit of the nonself-dual Chern-Simons-Higgs energy*, J. Funct. Anal., 244 (2008), pp. 535–588.

[15] M. Kurzke and D. Spirn, *Scaling limits of the Chern-Simons-Higgs energy*, Commun. Contemp. Math., 10 (2008), pp. 1–16.

[16] M. Kurzke and D. Spirn, *On the energy of a Chern-Simons-Higgs vortex lattice*, CRM Proc. Lecture Notes, 44 (2008), pp. 127–152.

[17] E. Sandier, *Lower bounds for the energy of unit vector fields and applications*, J. Funct. Anal., 152 (1998), pp. 379–403.

[18] E. Sandier and S. Serfaty, *Vortices in the Magnetic Ginzburg-Landau Model*, Progr. Nonlinear Differential Equations Appl. 70, Birkhäuser Boston, Cambridge, MA, 2007.

[19] E. Sandier and S. Serfaty, *On the energy of type-*II *superconductors in the mixed phase*, Rev. Math. Phys., 12 (2000), pp. 1219–1257.

[20] E. Sandier and S. Serfaty, *A rigorous derivation of a free boundary problem arising in superconductivity*, Ann. Sci. École Norm. Sup. (4), 33 (2000), pp. 561–592.

[21] E. Sandier and S. Serfaty, *Global Minimizers for the Ginzburg-Landau functional below the first critical field*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 17 (2000), pp. 119–145.
[22] S. Serfaty, *Local minimizers for the Ginzburg-Landau energy near critical magnetic field*. I., Commun. Contemp. Math., 1 (1999), pp. 213–254.
[23] S. Serfaty, *Local minimizers for the Ginzburg-Landau energy near critical magnetic field*. II., Commun. Contemp. Math., 1 (1999), pp. 295–333.
[24] D. Spirn, *Vortex motion law for the Schrödinger–Ginzburg–Landau equations*, SIAM J. Math. Anal., 34 (2003), pp. 1435–1476.
[25] D. Spirn and X. Yan, *Minimizers near the first critical field for the non self-dual Chern-Simons-Higgs energy*, Calc. Var. Partial Differential Equations, to appear.
[26] M. Tinkham, *Introduction to Superconductivity*, Dover Publications, New York, 2004.
[27] Y. Yang, *Solitons in Field Theory and Nonlinear Analysis*, Springer Monogr. Math., Springer-Verlag, New York, 2001.

SIAM J. MATH. ANAL.
Vol. 40, No. 5, pp. 2105–2131

INVERSE SPECTRAL AND SCATTERING THEORY FOR THE HALF-LINE LEFT-DEFINITE STURM–LIOUVILLE PROBLEM*

C. BENNEWITZ†, B. M. BROWN‡, AND R. WEIKARD§

Abstract. The problem of integrating the Camassa–Holm equation leads to the scattering and inverse scattering problem for the Sturm–Liouville equation $-u'' + \frac{1}{4}u = \lambda w u$, where w is a weight function which may change sign but where the left-hand side gives rise to a positive quadratic form so that one is led to a left-definite spectral problem. In this paper the spectral theory and a generalized Fourier transform associated with the equation $-u'' + \frac{1}{4}u = \lambda w u$ posed on a half-line are investigated. An inverse spectral theorem and an inverse scattering theorem are established. A crucial ingredient of the proofs of these results is a theorem of Paley–Wiener type which is shown to hold true. Additionally, the accumulation properties of eigenvalues are investigated.

Key words. inverse scattering problems, inverse spectral problems, left-definite problems, Sturm–Liouville, Camassa–Holm equation

AMS subject classifications. 37K15, 34A55, 34B24, 34L25, 35Q53

DOI. 10.1137/080724575

1. Introduction.

Standard Sturm–Liouville theory deals with the eigenvalue problem

$$-(pu')' + qu = \lambda w u, \tag{1.1}$$

together with appropriate boundary conditions, in the space L^2_w of functions square integrable with respect to the weight w, *i.e.*, the norm-square of the space is $\|u\|^2 = \int |u|^2 w$. A basic assumption for this to be possible is that $w \geq 0$. In some situations of interest this is not the case, but instead one has $p > 0$, $q \geq 0$. One may then use as a norm-square the integral $\int (p|u'|^2 + q|u|^2)$, and a problem of this type is usually called *left-definite*. A left-definite problem of current interest is the spectral problem associated with the Camassa–Holm equation, which is of the form

$$-u'' + \tfrac{1}{4}u = \lambda w u. \tag{1.2}$$

The Camassa–Holm equation is an integrable system in a similar sense as the Korteweg–de Vries (KdV) equation. It was first derived as an abstract bi-Hamiltonian system by Fuchssteiner and Fokas [22]. Subsequently, it was shown by Camassa and Holm [11] that it may serve as an integrable model for shallow water waves. In that paper Camassa and Holm also showed that the solitons are peaked and called them peakons (see also Fokas and Liu [21] and Johnson [23]). In contrast to the KdV equation the Camassa–Holm equation may model breaking waves, *i.e.*, smooth

*Received by the editors May 19, 2008; accepted for publication (in revised form) August 19, 2008; published electronically January 23, 2009. This paper was written with partial support from the Mittag-Leffler Institute in Stockholm, Sweden, the Newton Institute in Cambridge, UK, and the National Science Foundation under grant DMS-0304280.

http://www.siam.org/journals/sima/40-5/72457.html

†Department of Mathematics, Lund University, Box 118, SE-221 00 Lund, Sweden (christer.bennewitz@math.lu.se).

‡School of Computer Science, Cardiff University, Cardiff, P.O. Box 916, Cardiff CF2 3XF, UK (Malcolm.Brown@cs.cardiff.ac.uk).

§Department of Mathematics, University of Alabama at Birmingham, Birmingham, AL 35226-1170 (rudi@math.uab.edu).

initial data may develop singularities in finite time; cf. Constantin and Escher [15] and Constantin [13] (see also Bressan and Constantin [10] for a way to resolve the singularities due to wave breaking). This, however, happens only when w changes sign and it is this fact which motivates us to consider (1.2) without the assumption that w is positive. The well developed theory of scattering and inverse scattering for the Schrödinger equation is of crucial importance to the theory of the KdV equation. In the same way scattering/inverse scattering theory for (1.2) is important for dealing with the Camassa–Holm equation. Unfortunately, no such theory is available unless $w \geq 0$, and even then current theory requires more smoothness of w than is convenient to assume, in view of the lack of smoothness for the corresponding peakons.

The problem of inverse scattering for (1.2) is considerably more difficult than for the Schrödinger equation, which may be viewed as a rather mild perturbation of the equation $-u'' = \lambda u$. In case of (1.2) the perturbation is of the equation $-u''+\frac{1}{4}u = \lambda u$, and thus changes the coefficient containing the eigenvalue parameter λ. It appears that the methods used so far for dealing with the Schrödinger equation are no longer applicable.

In this paper we will prove some uniqueness results for inverse spectral theory and inverse scattering for the left-definite case which apply to (1.2) posed on a half-line. One would also like to have results for the full-line, but this appears to be more difficult. One exception is the case of odd initial data for the Camassa–Holm equation on the full-line because the problem can be reduced to one on a half-line. We mention here that the half-line case was also investigated by Boutet de Monvel and Shepelsky [8], [9], who employ Riemann–Hilbert techniques but assume that w is positive. Our approach is via the inverse spectral theory for the left-definite problem, which also is not very well developed. Even the spectral theory for left-definite problems is not widely known (but see for example [1]), in the level of detail necessary for dealing with the inverse problem. We will therefore start by presenting a reasonably comprehensive spectral theory, then prove some uniqueness theorems for the inverse spectral problem, and finally a uniqueness theorem for inverse half-line scattering.

Spectral theory for left-definite Sturm–Liouville problems seems to have been initiated by Weyl [28], who called such problems *polar*. Later many authors have dealt with more or less general left-definite problems. In particular we mention a series of papers by Niessen, Schneider, and their collaborators on singular left-definite so-called S-hermitian systems; see, *e.g.*, [26]. See also [1] and the references cited there. For a more recent contribution, see Kong, Wu, and Zettl [24]. However, papers in inverse spectral theory for left-definite problems are much more scarce; one example is Binding, Browne, and Watson [7].

Because of the connection with the Camassa–Holm equation the inverse scattering problem for (1.2) has attracted some attention. From the physical point of view the full-line case where w decays at infinity and the periodic case are most interesting. The former was treated by Fokas [20] and Constantin and various co-authors, for example in [14], [16], and [17]. The latter was addressed by Constantin and McKean [18], Constantin [12], and Vaninsky [27]. The full-line case with odd initial data reduces to a half-line case, but the half-line case is also of interest independently.

It will be convenient to deal only with the equation

$$-u'' + qu = \lambda wu. \tag{1.3}$$

There is no loss of generality in doing this, since the change of variable $t = \int_0^x 1/p$ will, as is readily seen, turn (1.1) into an equation of this form.

The plan of the paper is as follows. In section 2 we give a general spectral theory for left-definite problems on intervals with at least one regular endpoint, modelled on standard Titchmarsh–Weyl theory. One may extend this to intervals with two singular endpoints, in the same way as one can extend the right-definite theory, but since we will have no use of it here we have abstained from this.

In section 3 we deal with the generalized Fourier transform associated with a left-definite problem. To simplify the discussion we have restricted ourselves to one case, when so-called finite functions are dense in the Hilbert space associated with the equation. There are no fundamental difficulties involved in dealing with the general situation, but again we have no need of it in the applications we are thinking of.

Section 4 discusses uniqueness of the inverse spectral problem. Unfortunately we have neither a characterization nor a reconstruction algorithm, but the fundamental uniqueness theorem is quite general.

In section 5 we prove a theorem of Paley–Wiener type which is crucial for our approach to the inverse spectral theory, and section 6 deals with the uniqueness theorem for the half-line inverse scattering of a left-definite problem. Section 7 is devoted to some results about the number of eigenvalues for a left-definite problem under scattering conditions. Some elementary, but rather lengthy, calculations needed in section 4 have been relegated to the appendix.

2. Spectral theory. We shall consider (1.3) on an interval $[0, b)$ and assume that q and w are real-valued and integrable on compact subsets of $[0, b)$, that $q \geq 0$, and that neither q nor w vanish a.e. Let $\mathcal{H}_1$ be the set of locally absolutely continuous functions u defined in $[0, b)$ such that $u' \in L^2(0, b)$ and $q|u|^2 \in L^1(0, b)$. As we shall see presently $\mathcal{H}_1$ is a Hilbert space with scalar product

$$\langle u, v\rangle = \int_0^b (u'\overline{v'} + qu\overline{v})$$

and norm $\|u\| = \sqrt{\langle u, u\rangle}$. In order to show completeness of $\mathcal{H}_1$ and discuss how to find self-adjoint realizations corresponding to (1.3) we first note the following simple result.

LEMMA 2.1. *For any $a \in [0, b)$ there exists a constant C_a such that*

$$|u(x)| \leq C_a\|u\| \tag{2.1}$$

for any $x \in [0, a]$ and any $u \in \mathcal{H}_1$.

Proof. By the fundamental theorem of calculus and the Cauchy–Schwarz inequality $|u(x)| \leq |u(y)| + |y - x|^{1/2}\big(\int_0^b |u'|^2\big)^{1/2}$. If $c \in [a, b)$ is such that $\int_0^c q > 0$, multiplication by $q(y)$ and integrating with respect to y gives

$$|u(x)| \int_0^c q \leq \int_0^c q|u| + c^{1/2} \int_0^c q \left(\int_0^b |u'|^2\right)^{1/2}.$$

Using Cauchy–Schwarz again we obtain (2.1) with $C_a = \big(c + 1/\int_0^c q\big)^{1/2}$. □

PROPOSITION 2.2. *The space $\mathcal{H}_1$ is complete.*

Proof. By (2.1) a Cauchy sequence $u_1, u_2, \ldots$ in $\mathcal{H}_1$ converges locally uniformly to a continuous function u. Furthermore, $\sqrt{q}u_j$ and u_j' converge in $L^2[0, b)$ to $\sqrt{q}u$ and, say, v, respectively. Now

$$u_j(x) - u_j(0) = \int_0^x u_j'.$$

Letting $j \to \infty$ we obtain $u(x) = u(0) + \int_0^x v$. Thus u is absolutely continuous with derivative v and u_j converges to u in $\mathcal{H}_1$. □

Denote the set of integrable functions with compact support in $(0,b)$ by L_0. Then, if $u \in \mathcal{H}_1$ and $v \in L_0$, it follows that $|\int u\overline{v}| \le C_a \int |v|\,\|u\|$ if $\operatorname{supp} v \subset [0,a]$, so that the linear form $\mathcal{H}_1 \ni u \mapsto \int u\overline{v}$ is bounded. By Riesz's representation theorem we may therefore find a unique $v^* \in \mathcal{H}_1$ so that $\int u\overline{v} = \langle u, v^*\rangle$. Clearly v^* depends linearly on v, so we obtain a (bounded) operator $G_0 : L_0 \to \mathcal{H}_1$ such that

$$\langle u, G_0 v\rangle = \int_0^b u\overline{v} \quad \text{for} \quad u \in \mathcal{H}_1,\ v \in L_0.$$

The operator G_0 is central for the left-definite spectral theory of (1.3).

PROPOSITION 2.3. *The operator G_0 is an integral operator $G_0u(x) = \int u\, g_0(x,\cdot)$, it is injective, and its restriction to $L_0 \cap \mathcal{H}_1$ is symmetric with range dense in $\mathcal{H}_1$.*

Proof. By (2.1) the map $\mathcal{H}_1 \ni u \mapsto u(x)$ is for each fixed $x \in [0,b)$ a bounded linear form, so there exists an element $g_0(x,\cdot) \in \mathcal{H}_1$ so that $u(x) = \langle u, \overline{g_0(x,\cdot)}\rangle$ for $u \in \mathcal{H}_1$, and therefore $G_0v(x) = \langle G_0v, \overline{g_0(x,\cdot)}\rangle = \int_0^b v\, g_0(x,\cdot)$ for any $v \in L_0$. Thus G_0 is an integral operator with kernel $g_0(x,y)$ (actually, as we shall see in Proposition 2.7, g_0 is real-valued). If u and $v \in L_0 \cap \mathcal{H}_1$, then

$$\langle G_0u, v\rangle = \overline{\langle v, G_0u\rangle} = \int_0^b u\overline{v} = \langle u, G_0v\rangle,$$

so the restriction of G_0 to $L_0 \cap \mathcal{H}_1$ is symmetric.

Let $[c,d] \subset (0,b)$ and $u_j(x) = \min(1, j(x-c), j(d-x))$ for $x \in [c,d]$ and $u_j(x) = 0$ otherwise. Then $u_j \in L_0 \cap \mathcal{H}_1$ and tends boundedly to the characteristic function of $[c,d]$ as $j \to \infty$, so if $G_0v = 0$, it follows from $0 = \langle G_0v, u_j\rangle = \int v\overline{u_j}$ that $\int_c^d v = 0$ for all $[c,d] \subset (0,b)$. Thus $v = 0$ a.e. so that G_0 is injective. On the other hand, if $u \in \mathcal{H}_1$ is orthogonal to G_0v for all $v \in L_0 \cap \mathcal{H}_1$, we may put $v = u_j$, so that $0 = \langle u, G_0u_j\rangle \to \int_c^d u$. It follows that $u = 0$ so the range of G_0 restricted to $L_0 \cap \mathcal{H}_1$ is dense and the proof is complete. □

We shall have to briefly use the theory of symmetric relations as presented in [1, section 1], and define maximal and minimal relations corresponding to (1.3). We start by setting

$$T_c = \{(G_0(wv), v) \mid v \in L_0 \cap \mathcal{H}_1\}.$$

Then, since w is real-valued, T_c is a symmetric relation in $\mathcal{H}_1$ for

$$\langle G_0(wu), v\rangle = \overline{\langle v, G_0(wu)\rangle} = \int_0^b wu\overline{v} = \langle u, G_0(wv)\rangle.$$

Proposition 2.3 implies that T_c is the graph of a densely defined symmetric operator in $\mathcal{H}_1$ if $\operatorname{supp} w = [0,b)$, but at this point we do not want to exclude the possibility of w vanishing on an open set. We define the *minimal relation* T_0 as the closure (in $\mathcal{H}_1 \oplus \mathcal{H}_1$) of T_c, and the *maximal relation* T_1 as the adjoint of this, *i.e.*,

$$T_1 = \{(u,f) \in \mathcal{H}_1 \oplus \mathcal{H}_1 \mid \langle u, v\rangle = \langle f, G_0(wv)\rangle \text{ for all } v \in L_0 \cap \mathcal{H}_1\}.$$

We must show that T_1 is a differential relation.

PROPOSITION 2.4. *We have $(u,f) \in T_1$ if and only if u and $f \in \mathcal{H}_1$, u' is locally absolutely continuous, and $-u'' + qu = wf$.*

Proof. First note that if u and $f \in \mathcal{H}_1$, then the definition of G_0 shows that

$$\langle u, v\rangle - \langle f, G_0(wv)\rangle = \int_0^b (u'\overline{v'} + qu\overline{v} - wf\overline{v}) \tag{2.2}$$

for any $v \in L_0 \cap \mathcal{H}_1$. If in addition u' is locally absolutely continuous and satisfies $-u'' + qu = wf$, integrating by parts gives

$$\langle u, v\rangle - \langle f, G_0(wv)\rangle = \int_0^b (-u'' + qu - wf)\overline{v} = 0.$$

This proves one direction of the proposition.

In proving the other direction the assumption is that the quantity (2.2) is zero. But since $C_0^\infty(0,b) \subset \mathcal{H}_1$ this means that the distributional derivative of u' is $qu - wf$ so that u' is locally absolutely continuous and u satisfies the differential equation.

To give a proof without the use of distribution theory we prove a variant of the classical du Bois-Reymond lemma. If $v \in L_0 \cap \mathcal{H}_1$, integration by parts in (2.2) gives

$$\int_0^b \left\{ u' - \int_0^x (qu - wf) - C \right\} \overline{v'} = 0 \tag{2.3}$$

for any constant C. Now let $[c,d] \subset (0,b)$ and choose $C = \frac{1}{d-c}\int_c^d \{u' - \int_0^x (qu - wf)\}$. Put $v(y) = 0$ for $y \notin [c,d]$ and

$$v(y) = \int_c^y \left\{ u'(x) - \int_0^x (qu - wf) - C \right\} dx$$

for $y \in [c,d]$. Then $v \in L_0 \cap \mathcal{H}_1$ and (2.3) gives

$$\int_c^d \left| u' - \int_0^x (qu - wf) - C \right|^2 = 0$$

so that $u' - \int_0^x (qu - wf)$ is constant in $[c,d]$. Thus u' is locally absolutely continuous, and differentiation gives $-u'' + qu = wf$. ∎

Let $\mathcal{D}_\lambda = \{(u, \lambda u) \in T_1\}$ and let D_λ be the projection of $\mathcal{D}_\lambda$ onto its first components, *i.e.*, $u \in D_\lambda$ means that $u \in \mathcal{H}_1$ and u satisfies $-u'' + qu = \lambda wu$. We then have

$$T_1 = T_0 \dotplus \mathcal{D}_\lambda \dotplus \mathcal{D}_{\overline{\lambda}}$$

as a direct sum for any nonreal λ. Here $\dim \mathcal{D}_\lambda = \dim D_\lambda$ is constant in each of the upper and lower half-planes, and these dimensions will be called the *deficiency indices* of T_1. See [1, Theorem 1.4] for this simple generalization of the von Neumann formula for symmetric operators and its consequences. It is clear that $\dim D_\lambda \le 2$, and that $\dim D_{\overline{\lambda}} = \dim D_\lambda$, since $\overline{u} \in D_{\overline{\lambda}}$ if and only if $u \in D_\lambda$. Thus deficiency indices are always equal, and there are always self-adjoint extensions of T_0, which will at the same time be restrictions of T_1, and therefore realizations of (1.3). It is of course of interest to have criteria in terms of the coefficients q and w for different values of the deficiency indices $\dim D_\lambda$. In surprising contrast to the right-definite case, we have the following simple and explicit criteria.

THEOREM 2.5. *Suppose* $\operatorname{Im}\lambda \neq 0$ *and let* W *be an antiderivative of* w. *Then* $\dim D_\lambda = 2$ *if* $b < \infty$ *and* $q + W^2 \in L^1[0,b)$. *Otherwise* $\dim D_\lambda = 1$ *for* $\operatorname{Im}\lambda \neq 0$.

The theorem is a special case of [2, Theorem 2.3]. See also [5]. In the right-definite case a simple variation of constants argument shows that if $\dim D_\lambda = 2$ for one real or nonreal value of λ, then this holds for all $\lambda \in \mathbb{C}$. A similar argument shows that this remains true in the left-definite case, with the exception that it is possible that $\dim D_0 = 2$ even if $\dim D_\lambda < 2$ for all $\lambda \neq 0$. This is to be expected, since D_0 does not depend on the choice of w. We characterize $\dim D_0$ completely in the following theorem, which also brings out the significance of the space D_0. We use the expression *finite function* in $\mathcal{H}_1$ to denote a function which vanishes near b.

THEOREM 2.6.

(1) *The set D_0 is the orthogonal complement in $\mathcal{H}_1$ of $L_0 \cap \mathcal{H}_1$ and has dimension 1 or 2.*
(2) $\dim D_0 = 2$ *if and only if* $b < \infty$ *and* $q \in L^1[0, b)$.
(3) *If $b < \infty$ and $q \in L^1[0, b)$, then v and v' have finite limits at b for all $v \in D_0$, and these limits uniquely determine v.*
(4) *If $b < \infty$ and $q \in L^1[0, b)$, then every $u \in \mathcal{H}_1$ has a limit at b which is a bounded linear form on $\mathcal{H}_1$.*
(5) *If $\dim D_0 = 1$ and $D_0 \ni v \not\equiv 0$, then $v(0)\overline{v'(0)} < 0$ and $u(x)\overline{v'(x)} \to 0$ as $x \to b$ for any $u \in \mathcal{H}_1$.*
(6) *Finite functions are dense in $\mathcal{H}_1$ if and only if* $\dim D_0 = 1$.

Most of this is also a special case of the results of [2] and [5], but we give a simple proof, an elaboration of which can also prove Theorem 2.5.

Proof. We have $u \in D_0$ precisely if $(u, 0) \in T_1$, which holds precisely if $\langle u, v\rangle = \langle u, v\rangle - \langle 0, G_0(wv)\rangle = 0$ for all $v \in L_0 \cap \mathcal{H}_1$, proving the first claim. Since there are elements $v \in \mathcal{H}_1$ with $v(0) \neq 0$, and since $u(0) = 0$ for every $u \in L_0 \cap \mathcal{H}_1$, it follows from (2.1) for $x = 0$ that $\dim D_0 \geq 1$ and we have proved (1).

If b is finite and q integrable, standard existence and uniqueness theorems show that all solutions of $-v'' + qv = 0$ are continuously differentiable with absolutely continuous derivative in $[0, b]$, and thus in $\mathcal{H}_1$, and that they are uniquely determined by the values of v and v' at b. In this case the proof of Lemma 2.1 clearly also works for $a = b$, so we have proved (3), (4), and one direction of (2).

Now let $u \in \mathcal{H}_1$ and $v \in D_0$. Integration by parts gives

$$\int_0^x (u'\overline{v'} + qu\overline{v}) + u(0)\overline{v'(0)} = u(x)\overline{v'(x)}. \tag{2.4}$$

Thus $u(x)\overline{v'(x)}$ has a limit at b. If this is not 0, then $(u(x)\overline{v'(x)})^{-1}$ is bounded close to b. Therefore $u'/u = u'\overline{v'}/(u\overline{v'})$ is integrable near b, so that u has a nonzero limit at b. Since $q|u|^2$ is integrable it follows that $q \in L^1(0, b)$. Similarly, $v''/v' = qv/v' = qv\overline{u}/(v'\overline{u})$ is integrable near b, so v' has a nonzero limit at b. Since $|v'|^2$ is integrable it follows that b is finite.

Now, setting $u = v \not\equiv 0$ in (2.4) the integral is increasing, ≥ 0, and not constant, so if $v(0)\overline{v'(0)} \geq 0$, then $v(x)\overline{v'(x)}$ cannot tend to 0 at b. However, if $\dim D_0 = 2$, we may choose $v \in D_0$ with $v'(0) = 0$, so it follows that $q \in L^1(0, b)$ and b finite, completing the proof of (2).

On the other hand, if $\dim D_0 = 1$, then $u(x)\overline{v'(x)}$ must tend to zero for any $u \in \mathcal{H}_1$. In particular, for $u = v$ one therefore has $v(0)\overline{v'(0)} < 0$ for any nonzero $v \in D_0$ which proves (5).

Finally, if $u \in \mathcal{H}_1$ is finite and $v \in D_0$, integration by parts shows that $\langle u, v\rangle = -u(0)\overline{v'(0)}$, so the orthogonal complement of the finite functions consists of those

$v \in D_0$ for which $v'(0) = 0$. According to (5) this implies $v = 0$ if $\dim D_0 = 1$ and the proof is complete. □

It is now possible to give a detailed description of the kernel g_0.

PROPOSITION 2.7. *The kernel $g_0(x, y)$ is real-valued and symmetric in x, y. As a function of y it satisfies* (1.3) *with $\lambda = 0$ for $y \neq x$, and there are real-valued functions ψ_0 and φ_0 which solve* (1.3) *with $\lambda = 0$, such that if $u \in \mathcal{H}_1$, then*

(1) $\psi_0 \in \mathcal{H}_1$, $\psi_0'(0) = 1$ *and* $\psi_0'(x)u(x) \to 0$ *as* $x \to b$,

(2) $\varphi_0(0) = -1$, $\varphi_0'(0) = 0$,

(3) $g_0(x, y) = \varphi_0(\min(x, y))\psi_0(\max(x, y))$.

Proof. The existence of the solution φ_0 is not in question, and if a solution with the properties of ψ_0 exists, it is easy to verify that the kernel $\varphi_0(\min(x, y))\psi_0(\max(x, y))$ has the properties required of $g_0(x, y)$.

The existence of ψ_0 follows from Theorem 2.6. Indeed, if $\dim D_0 = 2$, the element $v \in D_0$ with $v(b) = 1$, $v'(b) = 0$ is real-valued and must have $v(0)v'(0) < 0$ by (2.4), so $v'(0) \neq 0$, and an appropriate multiple will have the properties required of ψ_0.

On the other hand, if $\dim D_0 = 1$, any nonzero $v \in D_0$ satisfies $v(0)\overline{v'(0)} < 0$ so $v'(0) \neq 0$, and an appropriate multiple will satisfy the requirements for ψ_0. Note that this solution is real-valued, since its real and imaginary parts also are in D_0, and are thus proportional, and the initial condition guarantees that the imaginary part vanishes. □

Now let T be a self-adjoint restriction of T_1 and assume that (u, f) and $(v, g) \in T$. Integrating by parts we then obtain

$$\int_0^x (u'\overline{g'} + qu\overline{g}) - \int_0^x (f'\overline{v'} + qf\overline{v}) = (u'\overline{g} - f\overline{v'})\big|_0^x. \tag{2.5}$$

As $x \to b$ this vanishes, since the left-hand side tends to $\langle u, g\rangle - \langle f, v\rangle$. Thus the condition for symmetry is that

$$(u'\overline{g} - f\overline{v'})\big|_0^b = 0.$$

Comparing this with $(u'\overline{v} - u\overline{v'})\big|_0^b = 0$, which is the similar condition in the right-definite case, we see that only exceptionally would self-adjoint boundary conditions in the left-definite case also be self-adjoint boundary conditions in the right-definite case.

Separated boundary conditions are those that make $u'\overline{g} - f\overline{v'}$ vanish at each endpoint separately, and are thus at 0 of the form

$$f(0)\cos\alpha + u'(0)\sin\alpha = 0, \tag{2.6}$$

for some $\alpha \in [0, \pi)$. Again comparing with the right-definite case, where the condition is $u(0)\cos\alpha + u'(0)\sin\alpha = 0$, the conditions coincide only in the case $\alpha = \pi/2$, the Neumann boundary condition. However, for eigenfunctions, where $f = \lambda u$, it is clear that also $\alpha = 0$, the Dirichlet boundary condition, gives the same spectra outside of $\lambda = 0$.

We shall not need a detailed description of self-adjoint boundary conditions at a singular endpoint. However, one may always impose the condition (2.6) at 0. It is easy to see that the corresponding restriction of T_1 has a symmetric adjoint, which is a strict extension of T_0. If the deficiency indices of T_0 equal 1, this is sufficient to obtain a self-adjoint restriction T of T_1, and all self-adjoint realizations are of this form. Otherwise, a condition needs to be imposed also at b. From (2.5) it follows

immediately that every $(u,f) \in T_1$ satisfying such a condition at b must satisfy $\mathrm{Im}(u'(x)\overline{f(x)}) \to 0$ as $x \to b$.

Assuming now that we have a self-adjoint relation T, the spectral theorem looks as follows (see [1, Theorem 1.15]). Consider the set $\mathcal{H}_\infty = \{u \in \mathcal{H}_1 \mid (0,u) \in T\}$. Then $\mathcal{H}_\infty$ is a subspace of $\mathcal{H}_1$, and setting $\mathcal{H} = \mathcal{H}_1 \ominus \mathcal{H}_\infty$ the *domain* D_T of T (*i.e.*, the set of first components of T) is a dense subset of $\mathcal{H}$, and $T \cap \mathcal{H} \oplus \mathcal{H}$ is the graph of a self-adjoint operator in $\mathcal{H}$. We will denote this operator by T as well, and may now apply the usual spectral theorem to T. If the resolution of the identity for the operator T is $\{E_t\}_{t\in\mathbb{R}}$, we extend the domain of the projection E_t to all of $\mathcal{H}_1$ by setting $E_t\mathcal{H}_\infty = 0$. Clearly one may view $\mathcal{H}_\infty$ as an eigenspace for the relation T belonging to the eigenvalue ∞, so adjoining the orthogonal projection onto $\mathcal{H}_\infty$ to $\{E_t\}_{t\in\mathbb{R}}$ gives a resolution of the identity in $\mathcal{H}_1$ for the relation T. In the present case one may give a rather complete description of $\mathcal{H}_\infty$.

PROPOSITION 2.8. *The space $\mathcal{H}_\infty$ consists of those elements $g \in \mathcal{H}_1$ for which $wg = 0$ a.e., and for which $(0,g)$ satisfies the boundary conditions that define T. In particular, if $wg = 0$ a.e. and $g \in L_0 \cap \mathcal{H}_1$, then $g \in \mathcal{H}_\infty$.*

Proof. Now $g \in \mathcal{H}_\infty$ means that $(0,g) \in T$, which therefore satisfies the boundary conditions defining T. In particular, $0 = \langle g, G_0(wf)\rangle - \langle 0, f\rangle = \langle g, G_0(wf)\rangle = \int g\overline{f}w$ for any $f \in L_0 \cap \mathcal{H}_1$. It follows, as in the proof of Proposition 2.3, that $wg = 0$ a.e.

Conversely, if $(0,g)$ satisfies the boundary conditions and $gw = 0$ a.e., then if $(u,f) \in T$, an integration by parts gives

$$\langle u, g\rangle - \langle f, 0\rangle = \lim_{x\to b}(u'\overline{g} - f\cdot 0)\big|_0^x = 0,$$

i.e., $(0,g) \in T$, so the proof is complete. □

We remark that if an endpoint is regular, then the boundary condition implied by $u \in \mathcal{H}_\infty$ is in most cases the vanishing of u in that endpoint. For separated boundary conditions an exception occurs when the boundary condition is of Neumann type (*i.e.*, when $\alpha = \pi/2$ in (2.6)). If we have Neumann conditions at both ends, or at one end when deficiency indices equal 1, there are *no* boundary conditions for elements of $\mathcal{H}_\infty$.

We will base our derivation of the expansion theorem for the operator T on a detailed description of the resolvent $R_\lambda = (T-\lambda)^{-1}$. Thus R_λ is defined on $\mathcal{H}$, but we extend its domain to $\mathcal{H}_1$ by setting $R_\lambda\mathcal{H}_\infty = 0$. The range of R_λ is of course D_T, which is a dense set in $\mathcal{H}$. Using the kernel g_0 for the evaluation operator on $\mathcal{H}_1$ introduced in the proof of Proposition 2.3, we have $R_\lambda u(x) = \langle R_\lambda u, \overline{g_0(x,\cdot)}\rangle = \langle u, R_{\overline{\lambda}}g_0(x,\cdot)\rangle$, since the adjoint of R_λ is $R_{\overline{\lambda}}$. Thus we may view $G(x,\cdot,\lambda) = \overline{R_{\overline{\lambda}}g_0(x,\cdot)}$ as Green's function for our operator; note, however, that G is not the kernel of a standard integral operator. It will turn out to be convenient to introduce the kernel $g(x,y,\lambda) = G(x,y,\lambda) + g_0(x,y)/\lambda$, so that we obtain

$$R_\lambda u(x) = \langle u, \overline{g(x,\cdot,\lambda)}\rangle - u(x)/\lambda. \tag{2.7}$$

Note that $G(x,\cdot,\lambda) \in \mathcal{H}$ but this is not true of $g(x,\cdot,\lambda)$ unless $\mathcal{H}_\infty = \{0\}$. We shall need a precise description of $g(x,y,\lambda)$. To do this we must introduce solutions of (1.3) satisfying initial conditions at 0, so let $\varphi(x,\lambda)$, $\theta(x,\lambda)$ be solutions of (1.3) for $\lambda \neq 0$ satisfying

$$\begin{cases}\lambda\varphi(0,\lambda) = -\sin\alpha\\ \varphi'(0,\lambda) = \cos\alpha\end{cases}, \quad \begin{cases}\lambda\theta(0,\lambda) = \cos\alpha\\ \theta'(0,\lambda) = \sin\alpha\end{cases}. \tag{2.8}$$

This means that φ satisfies the boundary condition (2.6) and θ another similar boundary condition at 0. We have the following theorem.

THEOREM 2.9. *Suppose T is a self-adjoint realization of* (1.3) *given by* (2.6) *and, if needed, an appropriate condition at b. Then there exists a function $m(\lambda)$ defined for* $\operatorname{Im}\lambda \neq 0$, *the* Titchmarsh–Weyl m-function for T, *depending only on λ and such that $\psi(x,\lambda) = \theta(x,\lambda) + m(\lambda)\varphi(x,\lambda)$, called the* Weyl solution *for T, is in $\mathcal{H}_1$ and satisfies the boundary condition at b, if any. Furthermore*

$$g(x,y,\lambda) = \varphi(\min(x,y),\lambda)\psi(\max(x,y),\lambda).$$

Proof. For nonreal λ neither φ nor θ can be in $\mathcal{H}_1$ and satisfy the boundary condition at b, since that would make λ a nonreal eigenvalue for a self-adjoint problem. Thus there is a solution $\psi(x,\lambda) = \theta(x,\lambda) + m(\lambda)\varphi(x,\lambda)$ in $\mathcal{H}_1$ which also satisfies the boundary condition at b, since if $\dim D(\lambda) = 2$, one linear, homogeneous condition still leaves a one-dimensional space, whereas if $\dim D(\lambda) = 1$, no boundary condition is imposed at b.

Define, for fixed x and $\lambda \notin \mathbb{R}$, the function

$$F(y) = \varphi(\min(x,y),\lambda)\psi(\max(x,y),\lambda) - \lambda^{-1}g_0(x,y).$$

Since $\psi(\cdot,\lambda)$ and ψ_0 are in $\mathcal{H}_1$ so is F. We claim that $F \in D_T$. In fact, one easily checks that F' is locally absolutely continuous and that F satisfies $-F'' + qF = \lambda wF + wg_0(x,\cdot)$. It is also easy to check that F satisfies the boundary condition (2.6).

Finally, for $y > x$ the function F is a linear combination of $\psi(\cdot,\lambda)$ and ψ_0. The former satisfies the boundary condition at b by construction, and ψ_0 satisfies the boundary condition at b by Theorem 2.6(5), since if $(u,f) \in T$, then $\psi_0'\overline{f} - 0\overline{u'} = \psi_0'\overline{f} \to 0$ at b. All this means that $F = R_\lambda g_0(x,\cdot) = \overline{R_{\overline{\lambda}} g_0(x,\cdot)} = G(x,\cdot,\lambda)$ so that $g(x,y,\lambda)$ is as claimed. ∎

THEOREM 2.10. *The function m is analytic outside $\mathbb{R}$, it maps the upper half plane into itself, and it satisfies $\overline{m(\lambda)} = m(\overline{\lambda})$.*

Proof. Since R_λ is analytic outside $\mathbb{R}$ in the strong operator topology $R_\lambda u(x)$ is, by (2.1), pointwise analytic. It follows that $g(x,\cdot,\lambda)$ is weakly analytic for each x, and thus, again by (2.1), $g(x,y,\lambda)$ is analytic outside $\mathbb{R}$ for each x and y. Since $\varphi(x,\lambda)$ and $\theta(x,\lambda)$ also are analytic and since an integration by parts shows that they are nonzero for $x > 0$ and $\lambda \notin \mathbb{R}$, it follows that $m(\lambda)$ is analytic in $\mathbb{C} \setminus \mathbb{R}$.

If (v,g) defines a boundary condition at b, then so does either its real part or its imaginary part, which is easily seen. Therefore, since $\psi(x,\lambda)$ satisfies (1.3) and the boundary condition at b, so does $\overline{\psi(x,\overline{\lambda})}$, and is thus a multiple of $\psi(x,\lambda)$. Now $\overline{\varphi(x,\overline{\lambda})} = \varphi(x,\lambda)$, $\overline{\theta(x,\overline{\lambda})} = \theta(x,\lambda)$ and $\psi(x,\lambda) = \theta(x,\lambda) + m(\lambda)\varphi(x,\lambda)$ so it follows that $m(\overline{\lambda}) = \overline{m(\lambda)}$.

Integrating by parts we have

$$\operatorname{Im}\lambda \int_0^x (|\psi'(\cdot,\lambda)|^2 + q|\psi(\cdot,\lambda|^2) = \operatorname{Im}(\overline{\psi'(\cdot,\lambda)}\lambda\psi(\cdot,\lambda))\big|_0^x.$$

Since ψ satisfies a boundary condition at b, the integrated term vanishes as $x \to b$. At 0 the integrated term evaluates to $-\operatorname{Im} m(\lambda)$, so we obtain

$$\|\psi(\cdot,\lambda)\|^2 = \operatorname{Im} m(\lambda)/\operatorname{Im}\lambda. \tag{2.9}$$

Thus m maps the upper and lower half-planes into themselves. ∎

A function with the properties of m is a so-called Nevanlinna or Herglotz function, and has a unique representation

$$m(\lambda) = A + B\lambda + \int_{\mathbb{R}} \left(\frac{1}{t-\lambda} - \frac{t}{t^2+1} \right) d\rho, \tag{2.10}$$

where $A \in \mathbb{R}$, $B \geq 0$, and $d\rho$ is a positive measure with $\int_{\mathbb{R}} \frac{d\rho(t)}{1+t^2} < \infty$. We will call the measure $d\rho$ the *spectral measure* for T, for reasons that will become clear presently.

We finally note the following proposition.

PROPOSITION 2.11. *Unless $\alpha = \pi/2$ and $0 \notin \operatorname{supp} w$ the functions ψ_0 and $\psi(\cdot, \lambda)$ are in $\mathcal{H}$.*

Proof. Suppose $g \in \mathcal{H}_\infty$. An integration by parts then gives

$$\langle g, \psi \rangle = -g(0)\overline{\psi'(0)},$$

where $\psi = \psi_0$ or $\psi(\cdot, \lambda)$. The boundary condition at 0 requires $g(0) = 0$ unless $\alpha = \pi/2$, and even then $g(0) = 0$ unless $w = 0$ in a neighbourhood of 0. ∎

3. The Fourier transform. We shall call functions that vanish near b *finite* and from now on make the following simplifying assumption.

Assumption 3.1. Assume that finite functions are dense in $\mathcal{H}_1$.

According to Theorem 2.6 this means exactly that either $q \notin L^1(0, b)$ or else $b = \infty$. Note that, according to Theorem 2.5, the assumption implies that the deficiency indices of T_1 equal 1.

The spectral measure introduced in the previous section gives rise to a Hilbert space L^2_ρ with scalar product $\langle \hat{u}, \hat{v} \rangle_\rho = \int_{-\infty}^{\infty} \hat{u}\overline{\hat{v}}\, d\rho$. We shall define a generalized Fourier transform $\mathcal{F} : \mathcal{H}_1 \to L^2_\rho$ with the following properties.

THEOREM 3.2.

(1) *The map $u \mapsto \int_0^b (u'\varphi'(\cdot, t) + qu\varphi(\cdot, t))$, defined for finite $u \in \mathcal{H}_1$, extends by continuity to a map $\mathcal{F} : \mathcal{H}_1 \to L^2_\rho$ called the* generalized Fourier transform. *The image of $u \in \mathcal{H}_1$ is denoted by $\mathcal{F}(u)$ or $\hat{u}$. We write this as $\hat{u}(t) = \langle u, \varphi(\cdot, t) \rangle$ although the integral in general does not converge pointwise.*

(2) *The mapping $\mathcal{F} : \mathcal{H}_1 \to L^2_\rho$ has kernel $\mathcal{H}_\infty$ and is unitary between $\mathcal{H}$ and L^2_ρ so that* Parseval's formula *$\langle u, v \rangle = \langle \hat{u}, \hat{v} \rangle_\rho$ holds if at least one of u and v is in $\mathcal{H}$.*

(3) *If $u \in D_T$, then $\mathcal{F}(Tu)(t) = t\hat{u}(t)$. Conversely, if $\hat{u}$ and $t\hat{u}(t)$ are in L^2_ρ, then $\mathcal{F}^{-1}(\hat{u}) \in D_T$.*

(4) *Suppose $\alpha \neq 0$ in (2.6). Then $\varphi(x, \cdot) \in L^2_\rho$ for each x and $\int_{-\infty}^{\infty} \hat{u}\varphi(x, \cdot)\, d\rho = \langle \hat{u}, \varphi(x, \cdot) \rangle_\rho$ converges in $\mathcal{H}$, and hence locally uniformly in x, for $\hat{u} \in L^2_\rho$. This is the adjoint of $\mathcal{F} : \mathcal{H}_1 \to L^2_\rho$ and thus the inverse of $\mathcal{F}$ restricted to $\mathcal{H}$. If M is a Borel set in $\mathbb{R}$, then*

$$E_M u(x) = \int_M \hat{u}\varphi(x, \cdot)\, d\rho. \tag{3.1}$$

If $\alpha = 0$, the same is true, except that we must replace $\varphi(\cdot, t)$ for $t = 0$ by the function ψ_0 of Proposition 2.7. Note that ψ_0 is the eigenfunction for the eigenvalue 0 in this case.

We first consider the Fourier transform for finite functions $u \in \mathcal{H}_1$, for every $\lambda \in \mathbb{C}$ setting

$$\hat{u}(\lambda) = \langle u, \varphi(\cdot, \overline{\lambda}) \rangle.$$

It is clear that $\hat{u}$ is an entire function, since integration by parts shows that

$$\hat{u}(\lambda) = \langle u, \varphi(\cdot, \overline{\lambda})\rangle = \int_0^b u\lambda\varphi(\cdot,\lambda)w - u(0)\cos\alpha,$$

and by (2.8) $\lambda\varphi(x,\lambda)$ is an entire function of λ, locally uniformly in x.

LEMMA 3.3. *For finite u and $v \in \mathcal{H}_1$ we have $\hat{u}$ and $\hat{v} \in L^2_\rho$. If E_Δ is the spectral projection for T associated with an interval Δ, then $\langle E_\Delta u, v\rangle = \int_\Delta \hat{u}\overline{\hat{v}}\, d\rho$.*

Proof. We have $\langle R_\lambda u, v\rangle = \hat{u}(\lambda)\overline{\hat{v}(\overline{\lambda})}m(\lambda) + g(\lambda)$, where g is entire, as is easily verified by direct calculation. Integrating around a rectangle γ with corners at $c \pm i$ and $d \pm i$ we therefore have $\int_\gamma \langle R_\lambda u, v\rangle\, d\lambda = \int_\gamma \hat{u}(\lambda)\overline{\hat{v}(\overline{\lambda})}m(\lambda)\, d\lambda$ whenever one of the integrals exists. By the spectral theorem the first integral equals $\int_\gamma \int_{\mathbb{R}} \frac{d\langle E_t u, v\rangle}{t-\lambda}\, d\lambda$, so if the integral is absolutely convergent, changing the order of integration gives $-2\pi i\langle E_{(c,d)}u, v\rangle$ if c and d are points of continuity for $\langle E_t u, v\rangle$.

Similarly, using the Nevanlinna representation (2.10), the other integral equals $-2\pi i \int_c^d \hat{u}(t)\overline{\hat{v}(t)}\, d\rho(t)$ if it is absolutely convergent and c, d are points of continuity for ρ.

The absolute convergence of the double integrals is ensured if $\langle E_t u, v\rangle$ and ρ are differentiable at c and d as is easily seen. For more details of the identical calculation carried out for the right-definite case, see [6, Lemmas 14.3, 14.4].

As functions of bounded variation $\langle E_t u, v\rangle$ and ρ are both differentiable a.e., so the second claim of the lemma is true if the endpoints of Δ belong to this dense set of points, and so in general by continuity. In particular, letting $c \to -\infty$, $d \to \infty$ through such points it follows that $\langle E_{\mathbb{R}} u, u\rangle = \langle \hat{u}, \hat{u}\rangle_\rho$, so that $\hat{u}, \hat{v} \in L^2_\rho$. ∎

Since finite functions are dense in $\mathcal{H}_1$, and since $E_{\mathbb{R}}$ has kernel $\mathcal{H}_\infty$, we now obtain Theorem 3.2(1) by continuity and also (2) except for the surjectivity of $\mathcal{F}$. To prove this we need the following lemmas.

LEMMA 3.4. *The transform of $R_\lambda u$ is $\hat{u}(t)/(t-\lambda)$.*

Proof. According to the spectral theorem we have $\langle R_\lambda u, v\rangle = \int_{\mathbb{R}} \frac{d\langle E_t u, v\rangle}{t-\lambda}$ and by Lemma 3.3 we have $\langle E_t u, v\rangle = \int_{-\infty}^t \hat{u}\overline{\hat{v}}\, d\rho$ so that

$$\langle R_\lambda u, v\rangle = \int_{\mathbb{R}} \frac{\hat{u}(t)}{t-\lambda}\,\overline{\hat{v}(t)}\, d\rho(t).$$

We also have $R_\lambda - R_{\overline{\lambda}} = (\lambda - \overline{\lambda})R_{\overline{\lambda}}R_\lambda$ and $\langle R_\lambda u, R_\lambda u\rangle = \langle R_{\overline{\lambda}}R_\lambda u, u\rangle$ so

$$\langle R_\lambda u, R_\lambda u\rangle = \frac{1}{\lambda - \overline{\lambda}}(\langle R_\lambda u, u\rangle - \langle R_{\overline{\lambda}}u, u\rangle) = \left\| \frac{\hat{u}(t)}{t-\lambda}\right\|_\rho^2.$$

Expanding $\|\frac{\hat{u}(t)}{t-\lambda} - \mathcal{F}(R_\lambda u)\|_\rho^2$ and using Parseval's formula and the above yields 0, thus proving the lemma. ∎

LEMMA 3.5. *The operator T has eigenvalue 0 if and only if $\alpha = 0$, in which case the eigenfunction is ψ_0, and for any $u \in \mathcal{H}_1$ we then have $\hat{u}(0) = -u(0)$.*

Furthermore, the measure $d\rho$ has mass at 0 ($\{0\}$ is not a nullset with respect to $d\rho$) precisely if $\alpha = 0$. In this case $\hat{\psi}_0 = \chi_{\{0\}}/\rho\{0\}$, where $\chi_{\{0\}}$ is the characteristic function of the singleton $\{0\}$ and $\rho\{0\}$ the spectral measure of this set.

Proof. According to Theorem 2.6 the only nontrivial solutions of (1.3) for $\lambda = 0$ in $\mathcal{H}_1$ are multiples of a solution u for which $u'(0)\overline{u(0)} < 0$, so that $u'(0) \neq 0$. These solutions satisfy the boundary condition (2.6) precisely if $\alpha = 0$, which proves the first

claim. If u is any finite function, integrating by parts gives $\hat{u}(0) = \langle u, \varphi(\cdot,0)\rangle = -u(0)$. This holds in general by continuity, $u(0)$ being a bounded linear form on $\mathcal{H}_1$ by (2.1), and $\hat{u}(0)$ on L^2_ρ since $d\rho$ has mass at 0, as we shall see presently.

Now $u \in D_T$ and $Tu = 0$ precisely if $u + \lambda R_\lambda u = 0$, and the Fourier transform of $u + \lambda R_\lambda u$ is $(1 + \frac{\lambda}{t-\lambda})\hat{u}(t) = \frac{t\hat{u}(t)}{t-\lambda}$. If this is 0, then $\hat{u} = 0$ a.e. with respect to $d\rho$ except possibly at $t = 0$. Thus, if $\alpha = 0$, then $\{0\}$ cannot be a nullset with respect to $d\rho$. It also follows that $\hat{\psi}_0$ is a multiple of the characteristic function of the set $\{0\}$. On the other hand, since $\dim D_\lambda = 1$, Weyl solutions for different α are proportional so it immediately follows that

$$m_0(\lambda) = \frac{\psi'(0,\lambda)}{\lambda\psi(0,\lambda)} = \frac{\sin\alpha + m_\alpha(\lambda)\cos\alpha}{\cos\alpha - m_\alpha(\lambda)\sin\alpha}, \tag{3.2}$$

where m_α denotes the m-function associated with the boundary condition parameter α. Now $m_0(i\nu) \to \infty$ as $\nu \downarrow 0$, as a consequence of the mass at 0, so that $m_\alpha(i\nu) \to \cot\alpha$ for $\alpha \neq 0$. For $\alpha \neq 0$ the spectral measure therefore has no mass at 0.

It only remains to prove the formula for $\hat{\psi}_0$. By Parseval's formula (note that $\psi_0 \in \mathcal{H}$ by Proposition 2.11) we have $\hat{\psi}_0(0) = -\psi_0(0) = \|\psi_0\|^2 = \|\hat{\psi}_0\|^2_\rho = |\hat{\psi}_0(0)|^2\rho\{0\}$. Hence $-\psi_0(0) = \hat{\psi}_0(0) = 1/\rho\{0\}$. □

It is now easy to prove that $\mathcal{F}$ is surjective.

LEMMA 3.6. *The Fourier transform $\mathcal{H} \to L^2_\rho$ is surjective.*

Proof. Suppose that $\hat{u} \in L^2_\rho$ is orthogonal to all Fourier transforms $\hat{v}$. Since $\hat{v}(t)/(t-\lambda)$ is also a transform, for any nonreal λ, we have $\int \frac{1}{t-\lambda}\hat{u}(t)\overline{\hat{v}(t)}\,d\rho(t) = 0$ for all nonreal λ. Thus the Stieltjes transform of the measure $\hat{u}\overline{\hat{v}}\,d\rho$ is 0, so by the uniqueness of the Stieltjes transform it follows that this measure is the zero measure.

Now, if $\hat{v}$ is the transform of a finite function in $\mathcal{H}_1$, then it is an entire function, so to prove that t is outside the support of $\hat{u}\,d\rho$ it is enough to show that there is such a $\hat{v}$ for which $\hat{v}(t) \neq 0$. If $t \neq 0$ and $\hat{v}(t) = 0$ for all compactly supported $v \in \mathcal{H}_1$, then as in the proof of Proposition 2.4 it follows that $\varphi(\cdot,t)$ satisfies (1.3) both for $\lambda = 0$ and $\lambda = t$, so that $\varphi(\cdot,t)w = 0$ a.e., which is not possible since it implies that $\varphi(\cdot,t) = 0$ in a set of positive Lebesgue measure. It therefore follows that $\hat{u}\,d\rho$ vanishes outside 0. But according to Lemma 3.5 this proves that the measure is zero, unless $\alpha = 0$. However, also in this case $\hat{u} = 0$ since otherwise $\hat{u}$ would be the transform of an eigenfunction. □

We next turn to Theorem 3.2(3).

LEMMA 3.7. *If $u \in D_T$, then $\mathcal{F}(Tu)(t) = t\hat{u}(t)$. Conversely, if $\hat{u}$ and $t\hat{u}(t)$ are in L^2_ρ, then $\mathcal{F}^{-1}(\hat{u}) \in D_T$.*

Proof. We have $u \in D_T$ if and only if for some $v \in \mathcal{H}_1$ we have $u = R_\lambda(v - \lambda u)$, *i.e.*, if and only if $\hat{u}(t) = (\hat{v}(t) - \lambda\hat{u}(t))/(t-\lambda)$ or $t\hat{u}(t) = \hat{v}(t)$ for some $\hat{v} \in L^2_\rho$. □

We obtain the following corollary which will be useful later on.

COROLLARY 3.8. *If $u \in D_T$, then $\hat{u}$ is integrable with respect to $d\rho$.*

Proof. The functions $t\hat{u}(t)$, $\hat{u}$, and $1/(t-i)$ are all in L^2_ρ, so that $\hat{u}(t) = (t\hat{u}(t) - i\hat{u}(t))/(t-i)$ is integrable with respect to $d\rho$. □

To finish the proof of Theorem 3.2 it only remains to consider the inverse transform.

LEMMA 3.9. *If $\alpha \neq 0$, the integral $\langle \hat{u}, \varphi(x,\cdot)\rangle_\rho$ converges in $\mathcal{H}$ and locally uniformly for every $\hat{u} \in L^2_\rho$. If $\hat{u} = \mathcal{F}(u)$ for some $u \in \mathcal{H}_1$, then the integral is the orthogonal projection of u onto $\mathcal{H}$.*

If $\alpha = 0$, the same statement is true if one replaces $\varphi(\cdot,0)$ by ψ_0 in the integral.

Remark 3.10. A simple integration by parts shows that every finite function is orthogonal to φ_0. Now suppose $\alpha = 0$ and let $\theta_0 = \varphi(\cdot, 0)$ so that θ_0 solves (1.3) for $\lambda = 0$ with initial data $\theta_0(0) = 0$, $\theta_0'(0) = 1$. Then, when calculating the Fourier transform at 0 we may replace $\varphi(\cdot, 0)$ by any function $\theta_0 + A\varphi_0$ for A constant, with no change to the Fourier transform.

In particular we may choose $A = -\psi_0(0) = 1/\rho\{0\}$, according to Lemma 3.5, so that $\theta_0 + A\varphi_0 = \psi_0$. This might seem a more natural choice of kernel for the Fourier transform, in view of the fact that it must be used for the inverse transform, and that ψ_0 is an eigenfunction to the eigenvalue 0, but would thus not actually change the Fourier transform.

Proof of Lemma 3.9. We have $u(x) = \langle u, g_0(x, \cdot)\rangle = \langle \hat{u}, e(x, \cdot)\rangle_\rho$ for $u \in \mathcal{H}$, where $e(x,t) = \mathcal{F}(g_0(x,\cdot))(t)$. If $u \in \mathcal{H}_1$, we instead get the projection of u onto $\mathcal{H}$, so that the integral operator $\hat{u} \mapsto \langle \hat{u}, e(x,\cdot)\rangle_\rho$ is the adjoint of $\mathcal{F}$. We must prove that $e(x,t) = \varphi(x,t)$, so suppose $\hat{u}$ has compact support and consider $\tilde{u}(x) = \langle \hat{u}, \varphi(x,\cdot)\rangle_\rho$ which satisfies the equation $-\tilde{u}'' + q\tilde{u} = w(x)\langle \hat{u}, t\varphi(x,\cdot)\rangle_\rho$, differentiating under the integral sign. Since $\hat{u}$ has compact support $u \in D_T$, so that $-u'' + qu = w(x)\langle t\hat{u}(t), e(x,t)\rangle_\rho$. Thus $u_1 = u - \tilde{u}$ satisfies $-u_1'' + qu_1 = w(x)\langle t\hat{u}(t), e(x,t) - \varphi(x,t)\rangle_\rho$.

Now, if v is finite, then

$$\langle \tilde{u}, v\rangle = \iint \hat{u}(t)(\varphi'(\cdot,t)\overline{v'} + q\varphi(\cdot,t)\overline{v})\, d\rho(t) = \langle \hat{u}, \hat{v}\rangle_\rho = \langle u, v\rangle,$$

since the double integral is absolutely convergent. Hence u_1 is orthogonal to all finite v so it satisfies $-u_1'' + qu_1 = 0$. It follows that $w(x)\langle t\hat{u}(t), e(x,t) - \varphi(x,t)\rangle_\rho = 0$ a.e., so that $\langle t\hat{u}(t), e(x,t) - \varphi(x,t)\rangle_\rho = 0$ on a set of positive measure. But this function also satisfies (1.3) for $\lambda = 0$, as is seen by replacing $\hat{u}$ by $t\hat{u}(t)$ in the previous calculations. It follows that $t(e(x,t) - \varphi(x,t)) = 0$ for a.a. t with respect to $d\rho$, so that $e(x,t) = \varphi(x,t)$ except possibly if $t = 0$ and $\alpha = 0$.

However, 0 is an eigenvalue for $\alpha = 0$ and the eigenfunction ψ_0 has transform $\chi_{\{0\}}/\rho\{0\}$ according to Lemma 3.5, so we must choose $e(x,0) = \psi_0(x)$. □

The proof of Theorem 3.2 is now complete if we note that from $\langle E_t u, v\rangle = \int_{-\infty}^t \hat{u}\overline{\hat{v}}$ follows that the transform of $E_t u$ is $\hat{u}$ multiplied by the characteristic function of $(-\infty, t]$. The formula $E_M u(x) = \int_M \hat{u}\varphi(x,\cdot)\, d\rho$ therefore follows from the inversion formula.

In Lemma 3.5 we calculated the Fourier transform of ψ_0 in the case $\alpha = 0$. We shall need to find a few more Fourier transforms.

LEMMA 3.11. *If $\lambda \notin \mathbb{R}$, the Fourier transform of $\psi(\cdot,\lambda)$ is $\hat{\psi}(t,\lambda) = 1/(t-\lambda)$. Furthermore, the Fourier transform of ψ_0 equals $\hat{\psi}_0(t) = \sin\alpha/t$ for $\alpha \neq 0$ and $1/\rho\{0\}$ times the characteristic function of the set $\{0\}$ for $\alpha = 0$.*

Proof. We have already calculated $\hat{\psi}_0$ for $\alpha = 0$ in Lemma 3.5. If $\alpha \neq 0$, we note that $\psi_0(x) = -g_0(0,x)$ so its Fourier transform is $-e(0,t) = -\varphi(0,t) = \sin\alpha/t$.

According to (2.7), Theorem 2.9, and Lemma 3.9, for $u \in \mathcal{H}$ we have

$$-\sin\alpha\langle \hat{u}, \hat{\psi}(\cdot,\lambda)\rangle_\rho = \overline{\lambda}\varphi(0,\overline{\lambda})\langle u, \psi(\cdot,\lambda)\rangle$$

$$= \overline{\lambda}R_{\overline{\lambda}}u(0) + u(0) = \left\langle \left(\frac{\overline{\lambda}}{t-\overline{\lambda}} + 1\right)\hat{u}(t), e(0,t)\right\rangle_\rho$$

$$= \left\langle \hat{u}(t), \frac{t\, e(0,t)}{t-\lambda}\right\rangle_\rho = -\sin\alpha\left\langle \hat{u}(t), \frac{1}{t-\lambda}\right\rangle_\rho$$

so that we have $\hat{\psi}(t,\lambda) = 1/(t-\lambda)$ if $\alpha \neq 0$. If $\alpha = 0$, we assume $\hat{u}$ has compact support so that we may differentiate $u(x) = \langle \hat{u}, e(x,\cdot)\rangle_\rho$ under the integral sign to obtain

$$\langle \hat{u}, \hat{\psi}(\cdot,\lambda)\rangle_\rho = \varphi'(0,\overline{\lambda})\langle u, \psi(\cdot,\lambda)\rangle = (R_{\overline{\lambda}}u)'(0)$$
$$= \left\langle \frac{\hat{u}(t)}{t-\overline{\lambda}}, e'_x(0,t)\right\rangle_\rho = \left\langle \hat{u}(t), \frac{e'_x(0,t)}{t-\lambda}\right\rangle_\rho = \left\langle \hat{u}(t), \frac{1}{t-\lambda}\right\rangle_\rho .$$

Thus, also in this case we obtain $\hat{\psi}(t,\lambda) = 1/(t-\lambda)$. □

COROLLARY 3.12. *Suppose $u \in \mathcal{H}$. Then $\langle u, \psi(\cdot,t\lambda)\rangle \to 0$ as $t \to \infty$, locally uniformly for $\lambda \notin \mathbb{R}$. By (2.1) this means that $\psi(x,t\lambda) \to 0$ as $t \to \infty$, locally uniformly in x and $\lambda \notin \mathbb{R}$.*

In fact, unless $0 \notin \operatorname{supp} w$ and $\alpha = \pi/2$ we have $\psi(\cdot,t\lambda) \to 0$ in $\mathcal{H}$, locally uniformly in $\lambda \notin \mathbb{R}$ as $t \to \infty$.

Proof. We have $\langle u, \psi(\cdot,\lambda)\rangle = \langle \hat{u}, \hat{\psi}(\cdot,\lambda\rangle_\rho$. With the extra assumptions Proposition 2.11 shows that $\psi(\cdot,\lambda) \in \mathcal{H}$ so that $\|\psi(\cdot,\lambda)\| = \|\hat{\psi}(\cdot,\lambda)\|_\rho$.

It follows immediately by dominated convergence from Lemma 3.11 that the claims are true. □

Remark 3.13. All of the theory of sections 2 and 3 extends with no essential change to the case when w is just a measure, or even an element of $H^{-1}_{\mathrm{loc}}(0,b)$.

4. Uniqueness of the inverse problem. We shall here deal with the following question: *To what extent is the operator T,* i.e.*, the interval $[0,b)$, the coefficients q and w, and the boundary condition parameter α, determined by the spectral measure $d\rho$?* To answer this question we introduce the concept of a *Liouville transform* as a map $v \mapsto u$ given by $u(x) = f(x)v(g(x))$, where f and g are fixed functions. We suppose that g is strictly increasing and continuous, and that f is never 0. It is then easy to see that the inverse of a Liouville transform is also a Liouville transform, as is the composition of two Liouville transforms.

Now consider another relation $\breve{T}$ of the same type as T, with Hilbert space $\breve{\mathcal{H}}_1$, interval $[0,\breve{b})$, boundary condition parameter $\breve{\alpha}$, and coefficients $\breve{q}$ and $\breve{w}$. We will assume, as we do for $\mathcal{H}_1$, that finite functions are dense in $\breve{\mathcal{H}}_1$.

THEOREM 4.1. *Suppose that $\alpha = \breve{\alpha}$, or $0 < \alpha = \pi/2 - \breve{\alpha} < \pi/2$, or $\pi/2 < \alpha = 3\pi/2 - \breve{\alpha} < \pi$ and that there is a continuously differentiable bijection g from $[0,b)$ to $[0,\breve{b})$ with the following properties: g, g', and g'' are locally absolutely continuous, $g' > 0$, $g(0) = g''(0) = 0$, $g'(0) = (\sin\breve{\alpha}/\sin\alpha)^2$ if $\alpha \neq 0 \neq \breve{\alpha}$, $g'(0) = 1$ if $\alpha = \breve{\alpha} = 0$, and the coefficients of T and $\breve{T}$ satisfy $\breve{q}(g(x)) = (-f(x)f''(x) + q(x)f(x)^2)/g'(x)$ and $\breve{w}(g(x)) = w(x)/g'(x)^2$, where $f(x) = g'(x)^{-1/2}$.*

Then the spectral measures associated with T and $\breve{T}$ are identical.

Proof. The functions g and f give rise to Liouville transform $\mathcal{L}$ from functions defined on $[0,\breve{b})$ to functions defined on $[0,b)$, in particular to a transform from $\breve{\mathcal{H}}_1$ to $\mathcal{H}_1$. We will first show that this latter transform is unitary. To that end assume that $\breve{u}$ and $\breve{v}$ are in $\breve{\mathcal{H}}_1$ and that at least one of them is a finite function. Obviously $\mathcal{L}\breve{u}$ and $\mathcal{L}\breve{v}$ are locally absolutely continuous. Furthermore we obtain after a partial integration

$$\langle \mathcal{L}\breve{u}, \mathcal{L}\breve{v}\rangle_{\mathcal{H}_1} = \int_0^b (g'(\breve{u}'\overline{\breve{v}'}) \circ g + (-ff'' + qf^2)(\breve{u}\overline{\breve{v}}) \circ g)$$
$$= \int_0^{\breve{b}} (\breve{u}'\overline{\breve{v}'} + \breve{q}\breve{u}\overline{\breve{v}}) = \langle \breve{u}, \breve{v}\rangle_{\breve{\mathcal{H}}_1}.$$

This proves first that $\mathcal{L}\breve{u} \in \mathcal{H}_1$ whenever $\breve{u}$ is a finite function in $\breve{\mathcal{H}}_1$ and second that $\mathcal{L}$ is an isometry from the finite functions in $\breve{\mathcal{H}}_1$ onto the finite functions in $\mathcal{H}_1$. As an isometry $\mathcal{L}$ can be extended to a unitary operator from $\breve{\mathcal{H}}_1$ to $\mathcal{H}_1$.

Next, a straightforward computation, using that $2f'g' + fg'' = 0$, shows that $-u'' + qu = wr$ if $u = \mathcal{L}\breve{u}$, $r = \mathcal{L}\breve{r}$, and $-\breve{u}'' + \breve{q}\breve{u} = \breve{w}\breve{r}$. In particular, $(\breve{u}, \breve{r}) \in \breve{T}$ implies that $(\mathcal{L}\breve{u}, \mathcal{L}\breve{r}) \in T$ and $\mathcal{L}\breve{\psi}(\cdot, \lambda)$ must be a multiple of $\psi(\cdot, \lambda)$.

Also, since $\breve{\varphi}(\cdot, \lambda)$ satisfies the differential equation $-\breve{u}'' + \breve{q}\breve{u} = \lambda\breve{w}\breve{u}$ the function $\mathcal{L}\breve{\varphi}(\cdot, \lambda)$ satisfies $-u'' + qu = \lambda wu$. Our assumptions on α, $\breve{\alpha}$, $g'(0)$, and $g''(0)$ imply that $f(0) = \sin\alpha / \sin\breve{\alpha} = \cos\breve{\alpha} / \cos\alpha$ and that $f'(0) = 0$. Therefore we find $\lambda(\mathcal{L}\breve{\varphi}(\cdot, \lambda))(0) = \lambda f(0)\breve{\varphi}(0, \lambda) = -\sin\alpha$ and $(\mathcal{L}\breve{\varphi}(\cdot, \lambda))'(0) = \breve{\varphi}'(0, \lambda)/f(0) = \cos\alpha$ which shows that $\varphi(\cdot, \lambda) = \mathcal{L}\breve{\varphi}(\cdot, \lambda)$. The situation is a little more complicated for the relationship between θ and $\breve{\theta}$ where one finds that

$$\mathcal{L}\breve{\theta}(\cdot, \lambda) = \theta(\cdot, \lambda) + (\tan\breve{\alpha} - \tan\alpha)\varphi(\cdot, \lambda).$$

By the linearity of $\mathcal{L}$ we have

$$\mathcal{L}\breve{\psi}(\cdot, \lambda) = \theta(\cdot, \lambda) + (\tan\breve{\alpha} - \tan\alpha + \breve{m})\varphi(\cdot, \lambda) = \psi(\cdot, \lambda).$$

This proves that $\breve{m} + \tan\breve{\alpha} = m + \tan\alpha$ and hence that $\breve{\rho} = \rho$. □

In the rest of this section we will make the following additional assumption about (1.3).

Assumption 4.2. The coefficients w and $\breve{w}$ satisfy $\operatorname{supp} w = [0, b)$, $\operatorname{supp}\breve{w} = [0, \breve{b})$.

Note that this does not mean that $w \neq 0$ a.e.; w could vanish on a nowhere dense set of strictly positive measure. However, it does mean that $\mathcal{H}_\infty = \{0\}$, $\mathcal{H} = \mathcal{H}_1$.

Remark 4.3. One may also allow w to be an arbitrary measure. However, then in the definition of the function h below, and in the statement of Lemma 5.1, w should be replaced by the density of the absolutely continuous part of the measure w, and Assumption 4.2 will have to be made on this density. If this is done, the results in the rest of the paper are still true, *mutatis mutandis*, with essentially the same proofs.

Now define the functions $h(x) = \int_0^x \sqrt{|w|}$ on $[0, b)$ and $\breve{h}(x) = \int_0^x \sqrt{|\breve{w}|}$ on $[0, \breve{b})$, respectively. By Assumption 4.2 these are strictly increasing, locally absolutely continuous functions.

Our main theorem is the following.

THEOREM 4.4. *Suppose that T and $\breve{T}$ have the same spectral measure $d\rho$. Then there is a unitary Liouville transform $\mathcal{U}$ taking $\breve{T}$ into T, in the sense that $\mathcal{H} \ni u \mapsto \mathcal{U}u \in \breve{\mathcal{H}}$ through $u(x) = f(x)\mathcal{U}u(g(x))$ and $\mathcal{U}T = \breve{T}\mathcal{U}$. Here $g(x) = \breve{h}^{-1} \circ h(x)$ and $f(x) = (g'(x))^{-1/2}$.*

The functions f and g are continuously differentiable, f is strictly positive, and f' is locally absolutely continuous with $f'(0) = 0$. Also $\alpha = \breve{\alpha}$, in which case $f(0) = 1$, or else $0 < \alpha = \pi/2 - \breve{\alpha} < \pi/2$ or $\pi/2 < \alpha = 3\pi/2 - \breve{\alpha} < \pi$, in which case $f(0) = |\tan\alpha|$.

The relations between the coefficients are $\breve{w}(g(x)) = w(x)/(g'(x))^2$ and $\breve{q}(g(x)) = (-f''(x) + q(x)f(x))/(f(x)(g'(x))^2)$.

It is clear from Theorem 4.1 that Theorem 4.4 is optimal in the sense that it is not possible to deduce more about the relation between T and $\breve{T}$ from the equality of their spectral measures than is done in Theorem 4.4. Sufficient additional information, however, will imply that T and $\breve{T}$ are identical. We give two corollaries of this type.

COROLLARY 4.5. *Suppose T and $\breve{T}$ have the same spectral measure and that $|w| = |\breve{w}|$ in $[0, \min(b, \breve{b}))$. Then $T = \breve{T}$,* i.e., *$b = \breve{b}$, $\alpha = \breve{\alpha}$, $q = \breve{q}$, and $w = \breve{w}$.*

Proof. The assumptions together with Theorem 4.4 show that $g(x) = x$ so that $b = \breve{b}$, and that $f(x) = 1$, so that T and $\breve{T}$ are identical. □

Note that only the absolute value of w need be known, so that all information about sign changes in w is encoded in the spectral measure. Also note that if $|w| = |\breve{w}|$ only in $[0,a)$, where $0 < a < \min(b,\breve{b})$, we still have $\alpha = \breve{\alpha}$ and $q = \breve{q}$, $w = \breve{w}$ in $[0,a)$.

COROLLARY 4.6. *Suppose T and $\breve{T}$ have the same spectral measure, that $q = \breve{q}$ on $[0, \min(b,\breve{b}))$, and that either $b = \breve{b}$ or $\alpha = \breve{\alpha}$. Then $T = \breve{T}$,* i.e., *$b = \breve{b}$, $\alpha = \breve{\alpha}$, $q = \breve{q}$, and $w = \breve{w}$.*

We will postpone the proof and first prove Theorem 4.4. To do this we will use a theorem of Paley–Wiener type. For its statement it will be convenient to introduce a special class of entire functions.

DEFINITION 4.7. *Let $\mathcal{A}$ be the set of entire functions $\hat{u}$ of order $\le 1/2$ which satisfy*

$$\limsup_{t\to\infty} t^{-1} \ln|\hat{u}(t^2\lambda)| \le \int_0^a \operatorname{Re}\sqrt{-\lambda w} \tag{4.1}$$

for some $a \in (0,b)$ and all $\lambda \in \mathbb{C}\setminus\mathbb{R}$. Here the branch of the square root is that with a positive real part.

THEOREM 4.8. *Let $\hat{u}$ be the generalized Fourier transform of $u \in \mathcal{H}$. Then $\hat{u}$ has at most one entire continuation in $\mathcal{A}$, and if* $\sup\operatorname{supp} u = a < b$, *such a continuation is given by*

$$\hat{u}(\lambda) = \int_0^a (u'\varphi'(\cdot,\lambda) + qu\varphi(\cdot,\lambda))$$

in which case (4.1) *holds with equality for all $\lambda \in \mathbb{C}$.*

Conversely, if $\hat{u}$ has an entire continuation of order $\le 1/2$ satisfying (4.1) *for λ on at least two different rays from the origin, then* $\operatorname{supp} u \subset [0,a]$.

We will postpone the proof of Theorem 4.8 to the next section and instead turn to the proof of Theorem 4.4.

LEMMA 4.9. *Let $g : [0,b) \to [0,\breve{b})$ be increasing and $g(0) = 0$. Suppose $\mathcal{U} : \mathcal{H}_1 \to \breve{\mathcal{H}}_1$ is linear with the properties that $(\mathcal{U}u)(0) = 0$ if $u(0) = 0$, that* $\operatorname{supp}\mathcal{U}u \subset [0,g(x)]$ *if* $\operatorname{supp} u \subset [0,x]$, *and that* $\operatorname{supp}\mathcal{U}u \subset [g(x),\breve{b})$ *if* $\operatorname{supp} u \subset [x,b)$. *Then there exists a function f such that $(\mathcal{U}u)(g(x)) = f(x)u(x)$ for all $u \in \mathcal{H}_1$.*

Proof. Fix $x \in [0,b)$. Suppose $u, v \in \mathcal{H}_1$ and that $u(x) = v(x)$. We will first show that $(\mathcal{U}(u-v)(g(x)) = 0$. If $x = 0$, this is by assumption.

For $x > 0$ we define[1] $u_- = \chi_{[0,x]}(u-v)$ and $u_+ = \chi_{[x,b)}(u-v)$. These are elements of $\mathcal{H}$. Thus $\operatorname{supp}\mathcal{U}u_- \subset [0,g(x)]$ and $\operatorname{supp}\mathcal{U}u_+ \subset [g(x),b)$ so that the functions $\mathcal{U}u_\pm$ vanish in $g(x)$. Adding them gives $\mathcal{U}(u-v)(g(x)) = 0$ as desired.

It follows that the value of $\mathcal{U}u$ at $g(x)$ only depends on the value of u at x. Thus, for each fixed $x \in [0,b)$, the map $u(x) \mapsto \mathcal{U}u(g(x))$ is well-defined and linear on $\mathbb{C}$, so we may find $f(x)$ so that $\mathcal{U}u(g(x)) = f(x)u(x)$. □

We will also need the following lemma.

LEMMA 4.10. *Put $m(x,\lambda) = \psi'(x,\lambda)/(\lambda\psi(x,\lambda))$. Then $m(x,\lambda) \to 0$ and $\lambda m(x,\lambda) \to \infty$ for every $x \in [0,b)$ as $\lambda \to \infty$ along any nonreal ray starting from the origin.*

Proof. First note that $m(x,\lambda)$ is the m-function for (1.3) on the interval $[x,b)$, with the Dirichlet boundary condition ($\alpha = 0$) at x. The first claim is then an immediate consequence of [3, Theorem 3.6].

[1] χ_I denotes the characteristic function of an interval I.

To prove the second claim, first assume that q does not have compact support, so that it does not vanish identically on $[x,b)$. Now note that, according to (3.2), $\tilde{m}(\lambda) = -1/m(x,\lambda)$ is the m-function for the Neumann boundary condition ($\alpha = \pi/2$) at x, so we need to show this to be $o(|\lambda|)$. Now, in the Nevanlinna representation (2.10) it is easy to see that the integral is always $o(|\lambda|)$, so we simply need to prove that $B = 0$ in the representation of $\tilde{m}$. Denote the corresponding Weyl solution by $\tilde{\psi}$ and the spectral measure by $d\tilde{\rho}$. Using (2.9) and Lemma 3.11 we obtain

$$\|\tilde{\psi}(\cdot,\lambda)\|^2_{[x,b)} = \frac{\operatorname{Im}\tilde{m}(\lambda)}{\operatorname{Im}\lambda} = B + \int_{-\infty}^{\infty} \frac{d\tilde{\rho}(t)}{|t-\lambda|^2} = B + \|\hat{\tilde{\psi}}(\cdot,\lambda)\|^2_{\tilde{\rho}}.$$

However, by Proposition 2.11, Parseval's formula is correct for $\tilde{\psi}$, so that $B = 0$ and we are done in the case when q does not have compact support.

Now suppose q vanishes identically in $[x,b)$. Consider an auxiliary equation for which q does not have compact support, but which has the same coefficients as (1.3) up to some point c, $x < c < b$. For this equation the above proof of the lemma is valid. Moreover, let $\tilde{\theta}$ and $\tilde{\varphi}$ denote functions analogous to θ and φ for $\alpha = 0$, but with initial data given in the point x. In view of (2.9) both the original $m(x,\lambda)$ and the corresponding function for the auxiliary equation are in the "Weyl disk" defined by

$$\int_x^c |\tilde{\theta}' + m\tilde{\varphi}'|^2 \le \frac{\operatorname{Im} m}{\operatorname{Im}\lambda},$$

so their distance is bounded by the diameter of the disk, which is exponentially small as λ becomes large (see [3, Theorem 6.3] for this result). Since $m(x,\lambda)$ is a nontrivial Nevanlinna function it cannot tend to 0 faster than a multiple of $1/|\lambda|$ for large $|\lambda|$, so that asymptotically $m(x,\lambda)$ is the same as the corresponding function for the auxiliary equation. Thus the lemma is actually valid in all cases. □

Proof of Theorem 4.4. Note first that by Lemma 3.5 we must have either $\alpha = \breve{\alpha} = 0$ or else $\alpha \neq 0 \neq \breve{\alpha}$.

Let $\mathcal{H}$ and $\breve{\mathcal{H}}$ denote the Hilbert spaces and $\mathcal{F}$ and $\breve{\mathcal{F}}$ the generalized Fourier transforms associated with the two equations, and put $\mathcal{U} = \breve{\mathcal{F}}^{-1} \circ \mathcal{F} : \mathcal{H} \to \breve{\mathcal{H}}$, which is unitary since the target space is L^2_ρ for both $\mathcal{F}$ and $\breve{\mathcal{F}}$. By Lemma 3.11 we have $\mathcal{U}\psi_0 = \breve{\psi}_0$ if $\alpha = \breve{\alpha}$, and if $\alpha \neq 0 \neq \breve{\alpha}$, we have $\mathcal{U}\psi_0 = \frac{\sin\alpha}{\sin\breve{\alpha}}\breve{\psi}_0$. Since $\langle u, \psi_0\rangle = -u(0)$ it follows that

$$u(0) = -\langle u, \psi_0\rangle = -\langle \mathcal{U}u, \mathcal{U}\psi_0\rangle = \frac{\sin\alpha}{\sin\breve{\alpha}}\mathcal{U}u(0), \tag{4.2}$$

where the quotient of the sines is to be read as 1 for $\alpha = \breve{\alpha} = 0$. In particular, $\mathcal{U}u(0) = 0$ if and only if $u(0) = 0$.

Now, applying Theorem 4.8 for the rays generated by $\pm i$, it is clear that if $\breve{a} \in (0,\breve{b})$ and $u \in \mathcal{H}$, then $\sup\operatorname{supp} u = a$ if $\sup\operatorname{supp}\mathcal{U}u = \breve{a}$, where $h(a) = \breve{h}(\breve{a})$, provided there is such an $a \in (0,b)$[2] (see [4, p. 29] for more details). This will certainly be the case if $\breve{a}$ is sufficiently close to 0. Suppose for some $\breve{a} \in (0,\breve{b})$ we have $h(b) \le \breve{h}(\breve{a})$. Then, since compactly supported functions are dense in $\mathcal{H}$, the range of $\mathcal{U}$ would be orthogonal to all elements of $\breve{\mathcal{H}}$ with supports in $(\breve{a},\breve{b})$, contradicting the fact that $\mathcal{U}$ is unitary.

A similar reasoning applied to $\mathcal{U}^{-1}$ shows that the mapping

$$g : [0,b) \ni a \mapsto \breve{a} \in [0,\breve{b})$$

[2]Note that $\operatorname{Re}\sqrt{\pm iw} = \sqrt{|w|/2}$.

is bijective, and that $\sup\operatorname{supp}\mathcal{U}u = \breve{a}$ if $\sup\operatorname{supp} u = a$. It follows that $\sup\operatorname{supp} u = a$ if and only if $\sup\operatorname{supp}\mathcal{U}u = g(a)$.

We also have $\inf\operatorname{supp} u = a$ if and only if $\inf\operatorname{supp}\mathcal{U}u = g(a)$. To see this, note that what we have already proved implies that if $\inf\operatorname{supp} u = a > 0$, then $\mathcal{U}u$ is orthogonal to all elements of $\breve{\mathcal{H}}$ with support in $[0, g(a)]$. This means that in this interval $\mathcal{U}u$ is a multiple of $\breve{\varphi}_0$. However, since $u(0) = 0$ we also have $\mathcal{U}u(0) = 0$, so that the multiple is 0, and thus $\inf\operatorname{supp}\mathcal{U}u \geq g(a)$. A similar reasoning applied to $\mathcal{U}^{-1}$ proves the other direction.

We have now verified that $\mathcal{U}$ and $\mathcal{U}^{-1}$ both have the properties required in Lemma 4.9. This implies that there is a nonvanishing function f so that

$$u(x) = f(x)\mathcal{U}u(g(x)). \tag{4.3}$$

We must have f real-valued since $\mathcal{F}$ and $\breve{\mathcal{F}}^{-1}$, and thus $\mathcal{U}$, map real-valued functions to real-valued functions. We note that (4.2) implies that $f(0) = 1$ if $\alpha = \breve{\alpha} = 0$ and $f(0) = \frac{\sin\alpha}{\sin\breve{\alpha}} > 0$ if $\alpha \neq 0 \neq \breve{\alpha}$. Now choose $\mathcal{U}u = 1$ in a neighborhood of $g(x)$. We then have $u = f$ in a neighborhood of x. Since $u \in \mathcal{H}$ is locally absolutely continuous, so is f. This also implies that f is strictly positive, since it cannot change sign and $f(0) > 0$. Similarly, choosing $\mathcal{U}u$ linear in a neighborhood of $g(x)$ it follows that also g is locally absolutely continuous.

According to Lemma 3.11 $\mathcal{U}\psi(\cdot,\lambda) = \breve{\psi}(\cdot,\lambda)$, so we have $\psi(x,\lambda) = f(x)\breve{\psi}(g(x),\lambda)$. Taking the logarithmic derivative we obtain

$$\frac{\psi'(x,\lambda)}{\psi(x,\lambda)} = \frac{f'(x)}{f(x)} + g'(x)\frac{\breve{\psi}'(g(x),\lambda)}{\breve{\psi}(g(x),\lambda)}.$$

Here the left member and the coefficient for $g'(x)$ are locally absolutely continuous, and the coefficient for $g'(x)$ is not independent of λ by Lemma 4.10. It follows that g' and f' are locally absolutely continuous, and differentiating, using the differential equations, we obtain

$$-\frac{f''}{f} + q - (g')^2\breve{q}\circ g - \lambda(w - (g')^2\breve{w}\circ g) = \frac{(f^2g')'}{f^2}\,\frac{\breve{\psi}'(g(\cdot),\lambda)}{\breve{\psi}(g(\cdot),\lambda)}.$$

Here the right member is $o(|\lambda|)$ according to Lemma 4.10 so the coefficient of λ to the left vanishes. On the other hand, the right member is not independent of λ unless $(f^2g')' = 0$, so that we obtain

$$\breve{q}\circ g = \frac{1}{f(g')^2}(-f'' + qf),$$

$$\breve{w}\circ g = (g')^{-2}w,$$

$$f^2g' = C$$

for some constant C. Evaluating (4.3) and its derivative at 0 for $u = \psi(\cdot,\lambda)$ elementary calculations now shows[3] that $C = 1$ and $f'(0) = 0$. One also deduces that either $\alpha = \breve{\alpha}$ or else $0 < \alpha = \pi/2 - \breve{\alpha} < \pi/2$ or $\pi/2 < \alpha = 3\pi/2 - \breve{\alpha} < \pi$. In these calculations one uses that $\breve{m}$ is not a Möbius transform, which is clear since this would give

[3]See the appendix.

a transform space of dimension 1. This can only happen if w, and $d\rho$, is a point mass. □

Finally we have to prove Corollary 4.6.

Proof of Corollary 4.6. The function $\tilde{f} = -\varphi_0$ solves $-\tilde{f}'' + q\tilde{f} = 0$ with initial data $\tilde{f}(0) = 1$, $\tilde{f}'(0) = 0$. Since $q \geq 0$ this solution is strictly positive on $[0, b)$, so we may put $\tilde{g}(x) = \int_0^x 1/\tilde{f}^2$. The pair of functions $\tilde{f}$, $\tilde{g}$ gives us a Liouville transform F_0 mapping $[0, b)$ onto some interval $[0, c)$ and $[0, \breve{b})$ onto $[0, \breve{c})$, and transforming the equations into $-u_0'' = \lambda w_0 u_0$ and $-\breve{u}_0'' = \lambda \breve{w}_0 \breve{u}_0$, respectively. Thus $F_0 F F_0^{-1}$, where F is the Liouville transform of Theorem 4.4, transforms one of these equations into the other.

Being a composition of Liouville transforms this is itself a Liouville transform given, say, by $u_0(x) = f_1(x)\breve{u}_0(g_1(x))$. By construction we obtain $f_1(0) = f(0)$, $f_1'(0) = 0$, and $f_1^2 g_1' \equiv 1$. Since both potentials are identically 0 it follows that $f_1'' = 0$. This means that $f_1 \equiv f(0)$ and $g_1(x) = x/(f(0))^2$.

If $\alpha = \breve{\alpha}$, then by Theorem 4.4 $f(0) = 1$ so that $F_0 F F_0^{-1}$ is the identity, implying that also F is the identity. Similarly, if $b = \breve{b}$, then $c = \breve{c}$ so that $f(0) = 1$, unless $c = \breve{c} = \infty$. We will show that c is always finite, and then it again follows that F is the identity.

Now $c = \int_0^b 1/\tilde{f}^2$, so we need to show that this integral is finite. Put $H = \tilde{f}'\tilde{f}$ which will be strictly positive sufficiently close to b by (2.4).

Differentiating $H' = (\tilde{f}')^2 + \tilde{f}''\tilde{f} = (\tilde{f}')^2 + q\tilde{f}^2 \geq (\tilde{f}')^2$. Thus $1/\tilde{f}^2 = (\tilde{f}')^2/H^2 \leq H'/H^2$ so that $\int_d^b 1/\tilde{f}^2 \leq 1/H(d) < \infty$ if d is sufficiently close to b. This completes the proof. □

5. The Paley–Wiener theorem. The proof of Theorem 4.8 relies on the following lemma, which is taken from [3, Theorem 6.1, Corollary 6.2].

LEMMA 5.1. *The following asymptotic formulas hold, locally uniformly for* $\lambda \in \mathbb{C} \setminus \mathbb{R}$ *and* $x > 0$. *The square root refers to the branch with positive real part:*

$$\lim_{t\to\infty} t^{-1} \ln \varphi(x, t^2\lambda) = \int_0^x \sqrt{-\lambda w},$$

$$\lim_{t\to\infty} t^{-1} \ln \psi(x, t^2\lambda) = -\int_0^x \sqrt{-\lambda w}.$$

The next lemma implies the simple direction of Theorem 4.8.

LEMMA 5.2. *Suppose* $u \in \mathcal{H}$ *and* $\operatorname{supp} u \subset [0, a]$. *Then* $\hat{u}(\lambda)$ *is entire of order* $\leq 1/2$ *and* $\hat{u}(\lambda) = o(|\lambda\varphi(a + \varepsilon, \lambda)|)$ *for every* $\varepsilon > 0$ *as* $\lambda \to \infty$ *along any nonreal ray originating at the origin.*

Proof. For finite u we have $\langle u, \varphi(\cdot, \overline{\lambda})\rangle = -u(0)\cos\alpha + \int_0^b u\lambda\varphi(\cdot, \lambda)w$. Now write

$$\hat{u}(\lambda) = -u(0)\cos\alpha + \lambda\varphi(a + \varepsilon, \lambda)\int_0^a u\varphi(\cdot, \lambda)w/\varphi(a + \varepsilon, \lambda).$$

The function $\varphi(x, \lambda)/\varphi(a + \varepsilon, \lambda)$ tends to zero uniformly for $x \in [0, a]$ and $\lambda\varphi(a + \varepsilon, \lambda) \to \infty$ according to Lemma 5.1 as $\lambda \to \infty$ along a nonreal ray. The lemma follows. □

The hard direction of Theorem 4.8 follows from the next lemma.

LEMMA 5.3. *Suppose* $u \in \mathcal{H}$, *that* $\hat{u}$ *has an entire continuation of order* $\leq 1/2$, *and that* $\hat{u}(\lambda) = \mathcal{O}(1/|\psi(a, \lambda)|)$ *as* $\lambda \to \infty$ *along two different nonreal rays originating at the origin. Then* $\operatorname{supp} u \subset [0, a]$ *and* $\hat{u}(\lambda) = \langle u, \overline{\varphi(\cdot, \lambda)}\rangle$.

Proof. Let $\varepsilon > 0$ and consider $F(\lambda) = \langle R_\lambda u, v\rangle - \hat{u}(\lambda)\langle\psi(\cdot,\lambda), v\rangle$, where $v = G_0(wf)$ and $f \in \mathcal{H}$ has compact support in $(a+\varepsilon, b)$. In particular $v \in D_T$. We shall show that F has an entire continuation of order $\leq 1/2$ which tends to 0 along the given rays. By the Phragmén–Lindelöf principle it follows that F is bounded everywhere and is therefore constant by Liouville's theorem, thus actually identically 0.

Now $F(\lambda) = \int_0^b (R_\lambda u - \hat{u}(\lambda)\psi(\cdot,\lambda))\overline{f}w$ so, arguing like in the proof of Proposition 2.3, it follows that $R_\lambda u - \hat{u}(\lambda)\psi(\cdot,\lambda)$ has support in $[0, a+\varepsilon]$. Applying the differential equation it follows that also u has support in $[0, a+\varepsilon]$. Since $\varepsilon > 0$ is arbitrary, in fact u has support in $[0,a]$. For $x > a$ the formula (2.7) gives $R_\lambda u(x) = \psi(x,\lambda)\langle u, \overline{\varphi(\cdot,\lambda)}\rangle$ so that $\psi(x,\lambda)(\hat{u}(\lambda) - \langle u, \overline{\varphi(\cdot,\lambda)}\rangle) = 0$. The lemma follows from this.

To prove that F is entire, Parseval's formula and Lemma 3.11 show that

$$F(\lambda) = \int_{-\infty}^{\infty} \frac{\hat{u}(t) - \hat{u}(\lambda)}{t-\lambda}\,\overline{\hat{v}(t)}\,d\rho(t).$$

It is obvious that this is an entire function, at least if we can bound the integrand properly. To do this and see that the order is at most $1/2$, note that for $|t-\lambda| \leq 1$ we may estimate the integrand by $\sup_{|z|\leq 1}|\hat{u}'(\lambda+z)||\hat{v}(t)|$. For $|t-\lambda| > 1$ we may estimate the integrand by $|\hat{u}(t)\hat{v}(t)| + |\hat{u}(\lambda)||\hat{v}(t)|$. Hence we have locally uniformly dominated convergence of the integral and

$$|F(\lambda)| \leq \|u\|\|v\| + (\sup_{|z|\leq 1}|\hat{u}'(\lambda+z)| + |\hat{u}(\lambda)|)\int_{-\infty}^{\infty}|\hat{v}|d\rho,$$

which is the required estimate, the integral being finite by Corollary 3.8 and $\hat{u}$ and therefore $\hat{u}'$ being of order $\leq 1/2$.

Finally, to show that F tends to 0 along the rays, we first note that $\psi(x,\lambda)/\psi(a,\lambda)$ converges to 0 uniformly for $x \in [a+\varepsilon, b)$, according to Lemma 5.1. Assuming f has compact support in $[a+\varepsilon, b)$ we obtain $\int_0^b \psi(\cdot,\lambda)\overline{f}w = o(|\psi(a,\lambda)|)$. Since $R_\lambda \to 0$ strongly as $\operatorname{Im}\lambda \to \infty$, it follows that F tends to 0 along the given rays. This finishes the proof. □

Theorem 4.8 is a simple consequence of these lemmas.

Proof of Theorem 4.8. If $\operatorname{supp} u \subset [0,a]$, it follows from Lemmas 5.2 and 5.1 that $\hat{u}(\lambda) = \langle u, \varphi(\cdot,\overline{\lambda})\rangle$ is an entire continuation of $\hat{u}$ of order $\leq 1/2$ such that

$$\limsup_{t\to\infty} t^{-1}\ln|\hat{u}(t^2\lambda)| \leq \lim_{t\to\infty} t^{-1}\ln|\varphi(a+\varepsilon, t^2\lambda)| = \int_0^{a+\varepsilon} \operatorname{Re}\sqrt{-\lambda w}$$

for nonreal λ and all $\varepsilon > 0$.

On the other hand, suppose there is an entire continuation of $\hat{u}$ of order $\leq 1/2$ and such that

$$\limsup_{t\to\infty} t^{-1}\ln|\hat{u}(t^2\lambda)| \leq \int_0^a \operatorname{Re}\sqrt{-\lambda w}$$

for λ on two different rays from the origin. If one or both of these are real, an immediate application of the Phragmén–Lindelöf principle shows this to be true for all other rays as well, so we may assume them nonreal. By Lemma 5.1 this implies that $\hat{u}(\lambda) = \mathcal{O}(|\psi(a+\varepsilon,\lambda)|^{-1})$ for large λ on these rays if $0 < \varepsilon < b-a$. Lemma 5.3 now shows that $\operatorname{supp} u \subset [0, a+\varepsilon]$ for small $\varepsilon > 0$ and thus for $\varepsilon = 0$. The uniqueness

of the continuation also follows from Lemma 5.3. If we have strict inequality on one ray, a simple argument using the Phragmén–Lindelöf principle (see [4, Lemma 3.6]) shows this to hold on all nearby rays as well, so that in fact $\sup \operatorname{supp} u < a$. The proof is now complete. □

6. Inverse scattering on the half-line. In this section we will show that scattering data for the half-line problem determines the coefficient w if q is known. We will of course have to assume that our equation is sufficiently close to a model equation, which, as usual, has constant coefficients.

Thus we consider (1.3) on $[0,\infty)$ with the following additional assumption, which will be in force throughout this section.

Assumption 6.1. There is a constant $q_0 \geq 0$ such that $q(x) - q_0$ and $w(x) - 1$ are both in $L^1(0,\infty)$.

Note that according to Theorems 2.5 and 2.6 finite functions are dense in $\mathcal{H}_1$ and, given the boundary condition (2.6), there is a unique self-adjoint realization T of (1.3) in $\mathcal{H}_1$.

We will need the following standard result.

PROPOSITION 6.2. *For* $\operatorname{Im} k \geq 0$, $k \neq 0$, *there exists a solution* $f(\cdot,k)$ *of* (1.3) *with* $\lambda = k^2 + q_0$ *having the following properties:* (1) $f(x,\cdot)$ *and* $f'(x,\cdot)$ *are analytic for* $\operatorname{Im} k > 0$ *and continuous for* $\operatorname{Im} k \geq 0$, $k \neq 0$; (2) $f(x,k) \sim e^{ikx}$ *and* $f'(x,k) \sim ike^{ikx}$ *as* $x \to \infty$.

This is standard. It is easily proved by first writing the equation for $g(x,k) = f(x,k)e^{-ikx}$ as $g'' + 2ikg' = (q - q_0 - (k^2 + q_0)(w-1))g$ and then solving this equation by successive approximations from its desired initial values $g(\infty) = 1$, $g'(\infty) = 0$ at ∞ using the estimate $|e^{2ik(t-x)} - 1| \leq 2$. See, for instance, Deift and Trubowitz [19].

If $\operatorname{Im} k > 0$, then $f(\cdot,k) \in \mathcal{H}_1$. Thus, if $\lambda \notin \mathbb{R}$ (*i.e.*, also $\operatorname{Re} k \neq 0$), then

$$f(x,k) = F(k)\psi(x,\lambda)$$

for some function F defined in $\operatorname{Im} k > 0$, $\operatorname{Re} k \neq 0$.

Let $[u,v] = u'v - uv'$ denote the Wronskian of the functions u and v and recall that Wronskians of solutions to (1.3) are independent of x. Since

$$[\lambda\varphi(\cdot,\lambda), f(\cdot,k)] = F(k)[\lambda\varphi(\cdot,\lambda), \psi(\cdot,\lambda)] = F(k) \tag{6.1}$$

is analytic for $\operatorname{Im}(k) > 0$ we find that F is analytic and can be extended analytically to the positive imaginary axis. Moreover, since $[\lambda\varphi(\cdot,\lambda), f(\cdot,k)]$ is continuous in $\operatorname{Im}(k) \geq 0$, $k \neq 0$, the function F extends continuously to the positive and negative real line. The zeros of F are located exactly where φ and f are linearly dependent, *i.e.*, when $\lambda = q_0 + k^2$ is an eigenvalue.

Equation (6.1) gives also that $F(-k) = \overline{F(k)}$ for real $k \neq 0$ and that F has no zeros on either the positive or the negative real line since $\varphi(\cdot,\lambda)$ is real for real λ and the real and imaginary parts of $f(x,k) \sim e^{ikx}$ are linearly independent.

For $k > 0$ and thus $\lambda = k^2 + q_0 > q_0$ define

$$\psi_\pm(\cdot,\lambda) = \lim_{\epsilon \to 0} \psi(\cdot, (\pm k + i\epsilon)^2 + q_0)$$

and

$$m_\pm(\lambda) = \lim_{\epsilon \to 0} m((\pm k + i\epsilon)^2 + q_0).$$

Since $\overline{m(\lambda)} = m(\overline{\lambda})$ when λ is not real we find that $m_+(\lambda) = \overline{m_-(\lambda)}$ when λ is real. Therefore

$$\frac{2ik\lambda}{|F(k)|^2} = \lambda[\psi_+(\cdot,\lambda), \psi_-(\cdot,\lambda)] = m_+(\lambda) - m_-(\lambda) = 2i \operatorname{Im} m_+(\lambda)$$

when $k > 0$ so that $\lambda > q_0$. This in turn implies

$$\pi\rho'(\lambda) = \operatorname{Im} m(\lambda + i0) = \frac{k\lambda}{|F(k)|^2}$$

for $\lambda > q_0$. Thus the restriction of F to the positive real line determines the spectral measure on the interval (q_0, ∞). It follows from this that the spectrum of T is absolutely continuous[4] in (q_0, ∞).

In the interval $(-\infty, q_0)$, where λ corresponds to the positive half of the imaginary axis for k, the spectrum is discrete since F is analytic there. There might also be an eigenvalue for $k = 0$, $\lambda = q_0$. Suppose $\lambda \neq 0$ is an eigenvalue. Then $\varphi(\cdot,\lambda)$ is a corresponding eigenfunction, and its Fourier transform $\hat{\varphi}(\lambda)$ is a multiple of the characteristic function of the set $\{\lambda\}$. The inversion formula (3.1) gives $\varphi(x,\lambda) = \hat{\varphi}(\lambda)\varphi(x,\lambda)\rho\{\lambda\}$, where $\rho\{\lambda\}$ is the spectral measure of the set $\{\lambda\}$. Thus $\hat{\varphi}(\lambda) = 1/\rho\{\lambda\}$. Parseval's formula gives $\|\varphi(\cdot,\lambda)\|^2 = |\hat{\varphi}(\lambda)|^2\rho\{\lambda\} = 1/\rho\{\lambda\}$. On the interval $(-\infty, q_0]$ we therefore know the spectral measure if we know all eigenvalues λ and the corresponding *normalization constants* $\|\varphi(\cdot,\lambda)\|^2$. Similarly, if $\alpha = 0$, then by Lemma 3.5 also $\lambda = 0$ is an eigenvalue, and $1/\rho\{0\}$ is the normalization constant for the eigenfunction ψ_0. We obtain the following theorem.

THEOREM 6.3. *Given the absolute value of the coefficient $F(k)$ for positive k, all eigenvalues, the corresponding normalization constants, and either q or $|w|$, the coefficients q and w and the boundary value parameter α are uniquely determined.*

Proof. We have already seen that the given data determine the spectral measure, and may now apply Corollaries 4.5 and 4.6 to draw the desired conclusion. □

7. Eigenvalues. This section is devoted to the proof of the following theorem. Part of the proof is an adaptation of Marchenko [25].

THEOREM 7.1. *Assume that q and w satisfy Assumption* 6.1. *Then we have the following:*

(1) *The eigenvalues of T are isolated and can accumulate only at q_0 or negative infinity.*
(2) *There will be infinitely many negative eigenvalues if and only if w is negative on a set of positive measure.*

If in addition we have $\int_0^\infty t|q(t) - q_0 w(t)|dt < \infty$, we also have the following:

(3) *Eigenvalues will not accumulate at q_0.*
(4) *q_0 is not an eigenvalue unless $q_0 = 0$ and $\alpha = 0$.*

To prove this we need the following strengthening of Proposition 6.2.

PROPOSITION 7.2. *Suppose q and w satisfy Assumption* 6.1 *and the integral $\int_0^\infty t|q(t) - q_0 w(t)|dt$ is finite. Then, for every $x \in [0,\infty)$, the function $f(x,\cdot)$ and its x-derivative, which were previously defined for $\operatorname{Im}(k) \geq 0$, $k \neq 0$, extend continuously to $k = 0$.*

The additional assumption and the improved estimate

$$|e^{2ik(t-x)} - 1| \leq \min(2|k|t, 2)$$

[4] For $q_0 < s < t$ we have $\int_s^t \operatorname{Im} m(\mu + i\varepsilon)\, d\mu \to \pi(\rho(t) - \rho(s))$ as $\varepsilon \downarrow 0$. But the left-hand side converges to $\int_s^t \operatorname{Im} m(\mu + i0)\, d\mu$ so ρ is absolutely continuous.

allow us to perform the successive approximations also near $k = 0$. The proposition follows from this.

Proof of Theorem 7.1. If $\mu = k^2 + q_0 < q_0$ is an eigenvalue of T, then, since F is analytic in the upper half-plane, eigenvalues are isolated and hence cannot accumulate at any point in $(-\infty, q_0)$. This proves (1).

To prove the second statement we make first the assumption that $q_0 > 0$ and $\alpha \neq 0$. By Lemma 3.5 zero is then not in the spectrum of T so that the range of T is $\mathcal{H}$ and we may define a bilinear form Q on $\mathcal{H}$ by setting

$$Q(u, v) = \int_{\mathbb{R}} \frac{1}{t} \hat{u}(t)\overline{\hat{v}(t)} d\rho(t).$$

Note that $Q(u, v) = 0$ if the supports of $\hat{u}$ and $\hat{v}$ do not intersect, which happens, for instance, if u and v are eigenvectors for different eigenvalues. Furthermore, by Lemma 3.7 $Q(u, Tv) = \int_{\mathbb{R}} \hat{u}(t)\overline{\hat{v}(t)} d\rho(t) = \langle u, v\rangle$. An integration by parts gives

$$\int_0^x (u'\overline{v'} + qu\overline{v}) = u(x)\overline{v'(x)} - u(0)\overline{v'(0)} + \int_0^x wu\overline{Tv}$$

for $u \in \mathcal{H}$ and $v \in D_T$. Hence if v is in the range of T and u is finite, or if u and v are exponentially decaying eigenfunctions, then we obtain

$$Q(u, v) = \int_0^\infty wu\overline{v} + \cot(\alpha) u(0)\overline{v(0)} \tag{7.1}$$

taking into account the boundary condition satisfied by $(T^{-1}v, v)$.

Now assume that $w \geq 0$. If $\cot(\alpha) \geq 0$, there can be no negative eigenvalue since $Tv = \lambda v$, $\lambda < 0$, $\|v\| \neq 0$ would imply that

$$0 \leq \int_0^\infty w|v|^2 + \cot\alpha |v(0)|^2 = \frac{1}{\lambda} Q(v, Tv) = \frac{1}{\lambda}\|v\|^2 < 0,$$

giving a contradiction. If $\cot\alpha < 0$, there can be at most one negative eigenvalue as we shall show now. If there were two distinct negative eigenvalues λ_1 and λ_2 with associated eigenvectors v_1 and v_2, we could assume that $v_1(0) = v_2(0)$. This would entail that

$$0 \leq \int_0^\infty w|v_1 - v_2|^2 = Q(v_1 - v_2, v_1 - v_2) = Q(v_1, v_1) + Q(v_2, v_2) < 0$$

since eigenfunctions decay exponentially so that we are allowed to employ (7.1).

Next assume $w < 0$ on a set of positive Lebesgue measure. We shall show that there are infinitely many negative eigenvalues. For any integer n one can choose elements $u_1, \ldots, u_n$ in $\mathcal{H}$, compactly supported in $(0, \infty)$, such that $Q(u_j, u_j) < 0$ and $Q(u_j, u_k) = 0$ if $j \neq k$. To achieve this one may for instance choose first bounded sets $A_1, \ldots, A_n$ of positive measure and positive distances from zero and each other on which w is negative. Then one lets u_j be a suitable mollification of the characteristic function of A_j. Equation (7.1) now guarantees that they have the desired properties.

Thus $Q(u, u) < 0$ whenever u is in the linear span B of $u_1, \ldots, u_n$. Let P be the orthogonal projection of B into the negative spectral subspace of $\mathcal{H}$, *i.e.*, $Pu = \mathcal{F}^{-1}(u\chi)$, where χ is the characteristic function of $(-\infty, 0)$. Suppose now that n is

larger than the number of negative eigenvalues. Then the kernel of P cannot be trivial so that there is a nontrivial $u \in B$ such that $\hat{u}$ is supported in $[0, \infty)$. Hence

$$0 > Q(u,u) = \int_{\mathbb{R}} \frac{1}{t} |\hat{u}(t)|^2 d\rho(t) \geq 0.$$

Since this is impossible the number of negative eigenvalues must be infinite.

If we only have $q_0 \geq 0$, but still $\alpha \neq 0$, then Q remains defined for functions u, v with Fourier transforms bounded near 0, since in this case $1/t \in L^2_\rho$ by Lemma 3.11. But the Fourier transforms of eigenfunctions to nonzero eigenvalues are supported away from 0, and the Fourier transform of a finite function is entire and thus locally bounded. Also, u_j is in the range of T. To see this, solve $-y'' + qy = wu_j$ with 0 initial data at a point to the right of $\operatorname{supp} u_j$ which yields a finite function y. Adding an appropriate multiple of ψ_0 (Proposition 2.7) gives a function in D_T. Thus the proof applies also in this case.

Allowing also $\alpha = 0$ the form Q is still defined if $\hat{u}(t)\overline{\hat{v}(t)}/t$ is continuous at 0. This is the case if u and v are eigenfunctions to negative eigenvalues. Also, if u is a finite function orthogonal to the eigenfunction ψ_0, then $\hat{u}(0) = 0$; so Q is defined for such functions. This last condition is just one linear condition on the space B, so the remainder can still have arbitrarily large dimension. All of the u_j are in the range of T, since the boundary condition now reads $u_j(0) = 0$. Thus the proof applies also in this case, and the proof of (2) is finished.

Now assume that $\int_0^\infty t|q(t) - q_0 w(t)|dt$ is finite, and that, contrary to our claim, there is a sequence $\mu_n = k_n^2 + q_0 < q_0$ of eigenvalues converging to q_0. Since eigenfunctions are orthogonal and satisfy the boundary condition an integration by parts shows

$$\int_0^\infty w f(\cdot, k_n)\overline{f(\cdot, k_m)} = -f(0, k_n)\overline{f(0, k_m)} \cot \alpha \tag{7.2}$$

if $n \neq m$. If $\alpha = 0$, the right-hand side has to be replaced by zero.

Since $\int_0^\infty t|q(t) - q_0 w(t)|dt < \infty$, our construction of f shows that $f(x,k) \sim e^{ikx}$ as $x \to \infty$, uniformly for $k \in i[0,1]$. This shows first that (7.2) is bounded as n and m tend to infinity, and second that we may find a positive c such that $|f(x,k) - e^{ikx}| \leq e^{-|k|x}/4$ if $x \geq c$, $k \in i[0,1]$. Simple estimates then show that

$$\frac{7}{16} e^{-(|k_n| + |k_m|)x} \leq \operatorname{Re}(f(x,k_n)\overline{f(x,k_m)}) \leq \frac{25}{16} e^{-(|k_n| + |k_m|)x}$$

if n and m are large. Since $w - 1$ is integrable this shows that the integral

$$\int_c^\infty \operatorname{Re}(f(x,k_n)\overline{f(x,k_m)}) w \to +\infty$$

as n, m tend to infinity. Now, since $f(x,k)$ is uniformly continuous on $[0,c] \times i[0,1]$ it follows that the integral over $[0,c]$ is bounded, so the integral over $[0,\infty)$ tends to infinity, contradicting the previously established boundedness and proving (3).

Finally, if $q_0 = 0$, we already know q_0 is an eigenvalue if and only if $\alpha = 0$. On the other hand, if $q_0 > 0$, then $f(\cdot, q_0)$ is asymptotic to 1, and any other solution to (1.3) is asymptotically linear, as is easily seen from the well-known reduction of order method. Thus no such solution is in $\mathcal{H}$ and there is no eigenfunction with eigenvalue q_0. This proves (4). □

Remark 7.3. If we allow w to be a general measure, then the negative part of w could be a finite sum of Dirac measures. In this case one may in the same way show that the number of negative eigenvalues is equal to the number of these Dirac measures if $\alpha \neq 0$, $\cot\alpha \geq 0$, and $q_0 > 0$, with suitable modifications in the other cases.

8. Appendix. Here we present some calculations which were omitted from the proof of Theorem 4.4.

For $x = 0$ the relation $\psi(x,\lambda) = f(x)\breve{\psi}(g(x),\lambda)$ gives

$$\cos\alpha - m(\lambda)\sin\alpha = f(0)\{\cos\breve{\alpha} - \breve{m}(\lambda)\sin\breve{\alpha}\}, \tag{8.1}$$

while $\psi'(x,\lambda) = f'(x)\breve{\psi}(g(x),\lambda) + f(x)g'(x)\breve{\psi}'(g(x),\lambda)$ for $x = 0$ gives

$$\sin\alpha + m(\lambda)\cos\alpha = \frac{f'(0)}{\lambda}\{\cos\breve{\alpha} - \breve{m}(\lambda)\sin\breve{\alpha}\} + \frac{C}{f(0)}\{\sin\breve{\alpha} + \breve{m}(\lambda)\cos\breve{\alpha}\}. \tag{8.2}$$

From (8.1), (8.2) we obtain

$$1 = \left\{f(0)\cos\alpha + \frac{f'(0)}{\lambda}\sin\alpha\right\}\{\cos\breve{\alpha} - \breve{m}(\lambda)\sin\breve{\alpha}\} + \frac{C\sin\alpha}{f(0)}\{\sin\breve{\alpha} + \breve{m}(\lambda)\cos\breve{\alpha}\}$$

and

$$m(\lambda) = \left\{-f(0)\sin\alpha + \frac{f'(0)}{\lambda}\cos\alpha\right\}\{\cos\breve{\alpha} - \breve{m}(\lambda)\sin\breve{\alpha}\} + \frac{C\cos\alpha}{f(0)}\{\sin\breve{\alpha} + \breve{m}(\lambda)\cos\breve{\alpha}\},$$

which after rearranging gives

$$\begin{aligned} 1 - \left(f(0)\cos\alpha + \frac{f'(0)}{\lambda}\sin\alpha\right)\cos\breve{\alpha} - \frac{C\sin\alpha\sin\breve{\alpha}}{f(0)} \\ = \breve{m}(\lambda)\left\{-\left(f(0)\cos\alpha + \frac{f'(0)}{\lambda}\sin\alpha\right)\sin\breve{\alpha} + \frac{C\sin\alpha\cos\breve{\alpha}}{f(0)}\right\} \end{aligned} \tag{8.3}$$

and

$$\begin{aligned} \left(f(0)\sin\alpha - \frac{f'(0)}{\lambda}\cos\alpha\right)\cos\breve{\alpha} - \frac{C\cos\alpha\sin\breve{\alpha}}{f(0)} \\ = \breve{m}(\lambda)\left\{\left(f(0)\sin\alpha - \frac{f'(0)}{\lambda}\cos\alpha\right)\sin\breve{\alpha} + \frac{C\cos\alpha\cos\breve{\alpha}}{f(0)}\right\} - m(\lambda). \end{aligned} \tag{8.4}$$

In (8.3) the left member and the coefficient of $\breve{m}$ are linear in $1/\lambda$, while $\breve{m}(\lambda)$ is not constant or a Möbius transform (this would give a one-dimensional transform space). From (8.3) we therefore obtain

$$\left(f(0)\cos\alpha + \frac{f'(0)}{\lambda}\sin\alpha\right)\cos\breve{\alpha} = 1 - \frac{C\sin\alpha\sin\breve{\alpha}}{f(0)},$$

$$\left(f(0)\cos\alpha + \frac{f'(0)}{\lambda}\sin\alpha\right)\sin\breve{\alpha} = \frac{C\sin\alpha\cos\breve{\alpha}}{f(0)},$$

which gives

$$f(0)\cos\alpha + \frac{f'(0)}{\lambda}\sin\alpha = \cos\breve{\alpha},$$

$$\frac{C\sin\alpha}{f(0)} = \sin\breve{\alpha}.$$

From this it is (again) clear that $\sin\alpha = 0$ if and only if $\sin\breve{\alpha} = 0$, so that we have two cases.

- $\alpha = \breve{\alpha} = 0$. We obtain $f(0) = 1$, and insertion in (8.4) shows that $\frac{f'(0)}{\lambda} = m(\lambda) - C\breve{m}(\lambda)$. The right member is $(1-C)m(\lambda)$ since $m(i\nu)$ and $\breve{m}(i\nu) \to 0$ as $\nu \to +\infty$ by Lemma 4.10, and m, $\breve{m}$ have the same spectral measure. Again by Lemma 4.10 it follows that $C = 1$, and thus $f'(0) = 0$.
- $\alpha \neq 0 \neq \breve{\alpha}$. We obtain $f'(0) = 0$, $f(0) = C\sin\alpha/\sin\breve{\alpha}$, and $C\sin(2\alpha) = \sin(2\breve{\alpha})$. But we know that $f(0) = \sin\alpha/\sin\breve{\alpha}$ so that $C = 1$. Insertion in (8.4) gives $m(\lambda) - \breve{m}(\lambda) = \cot\alpha - \cot\breve{\alpha}$.
  Since $\sin(2\alpha) = \sin(2\breve{\alpha})$ we have either $\alpha = \breve{\alpha}$ or $0 < \alpha = \pi/2 - \breve{\alpha} < \pi/2$ or $\pi/2 < \alpha = 3\pi/2 - \breve{\alpha} < \pi$. If $\alpha = \breve{\alpha}$, we obtain $f(0) = 1$ and $m(\lambda) = \breve{m}(\lambda)$. In the other cases we obtain $f(0) = |\tan\alpha|$ and $m(\lambda) - \breve{m}(\lambda) = 2\cot(2\alpha)$.

REFERENCES

[1] C. Bennewitz, *Spectral theory for pairs of differential operators*, Ark. Mat., 15 (1977), pp. 33–61.

[2] C. Bennewitz, *A generalisation of Niessen's limit-circle criterion*, Proc. Roy. Soc. Edinburgh Sect. A, 78 (1977/78), pp. 81–90.

[3] C. Bennewitz, *Spectral asymptotics for Sturm-Liouville equations*, Proc. London Math. Soc. (3), 59 (1989), pp. 294–338.

[4] C. Bennewitz, *A Paley-Wiener theorem with applications to inverse spectral theory*, in Advances in Differential Equations and Mathematical Physics (Birmingham, AL, 2002), Contemp. Math. 327, AMS, Providence, RI, 2003, pp. 21–31.

[5] C. Bennewitz and B. M. Brown, *A limit point criterion with applications to nonselfadjoint equations*, J. Comput. Appl. Math., 148 (2002), pp. 257–265.

[6] C. Bennewitz and W. N. Everitt, *The Titchmarsh-Weyl eigenfunction expansion theorem for Sturm-Liouville differential equations*, in Sturm-Liouville Theory, Birkhäuser, Basel, 2005, pp. 137–171.

[7] P. A. Binding, P. J. Browne, and B. A. Watson, *Inverse spectral problems for left-definite Sturm-Liouville equations with indefinite weight*, J. Math. Anal. Appl., 271 (2002), pp. 383–408.

[8] A. Boutet de Monvel and D. Shepelsky, *The Camassa-Holm equation on the half-line*, C. R. Math. Acad. Sci. Paris, 341 (2005), pp. 611–616.

[9] A. Boutet de Monvel and D. Shepelsky, *The Camassa-Holm equation on the half-line: A Riemann-Hilbert approach*, J. Geom. Anal., 18 (2008), pp. 285–323.

[10] A. Bressan and A. Constantin, *Global conservative solutions of the Camassa-Holm equation*, Arch. Ration. Mech. Anal., 183 (2007), pp. 215–239.

[11] R. Camassa and D. D. Holm, *An integrable shallow water equation with peaked solitons*, Phys. Rev. Lett., 71 (1993), pp. 1661–1664.

[12] A. Constantin, *On the inverse spectral problem for the Camassa-Holm equation*, J. Funct. Anal., 155 (1998), pp. 352–363.

[13] A. Constantin, *Existence of permanent and breaking waves for a shallow water equation: A geometric approach*, Ann. Inst. Fourier (Grenoble), 50 (2000), pp. 321–362.

[14] A. Constantin, *On the scattering problem for the Camassa-Holm equation*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 457 (2001), pp. 953–970.

[15] A. Constantin and J. Escher, *Wave breaking for nonlinear nonlocal shallow water equations*, Acta Math., 181 (1998), pp. 229–243.

[16] A. Constantin, V. S. Gerdjikov, and R. I. Ivanov, *Inverse scattering transform for the Camassa-Holm equation*, Inverse Problems, 22 (2006), pp. 2197–2207.

[17] A. Constantin and J. Lenells, *On the inverse scattering approach to the Camassa-Holm equation*, J. Nonlinear Math. Phys., 10 (2003), pp. 252–255.
[18] A. Constantin and H. P. McKean, *A shallow water equation on the circle*, Comm. Pure Appl. Math., 52 (1999), pp. 949–982.
[19] P. Deift and E. Trubowitz, *Inverse scattering on the line*, Comm. Pure Appl. Math., 32 (1979), pp. 121–251.
[20] A. S. Fokas, *On a class of physically important integrable equations*, Phys. D, 87 (1995), pp. 145–150.
[21] A. S. Fokas and Q. M. Liu, *Asymptotic integrability of water waves*, Phys. Rev. Lett., 77 (1996), pp. 2347–2351.
[22] B. Fuchssteiner and A. S. Fokas, *Symplectic structures, their Bäcklund transformations and hereditary symmetries*, Phys. D, 4 (1981/82), pp. 47–66.
[23] R. S. Johnson, *Camassa-Holm, Korteweg-de Vries and related models for water waves*, J. Fluid Mech., 455 (2002), pp. 63–82.
[24] Q. Kong, H. Wu, and A. Zettl, *Singular left-definite Sturm-Liouville problems*, J. Differential Equations, 206 (2004), pp. 1–29.
[25] V. A. Marchenko, *Sturm-Liouville Operators and Applications*, Oper. Theory Adv. Appl. 22, Birkhäuser Verlag, Basel, 1986.
[26] H. D. Niessen and A. Schneider, *Spectral theory for left-definite singular systems of differential equations*, in Spectral Theory and Asymptotics of Differential Equations (Proc. Conf., Scheveningen, 1973), North–Holland Math. Stud. 13, North–Holland, Amsterdam, 1974, pp. 29–43.
[27] K. L. Vaninsky, *Equations of Camassa-Holm type and Jacobi ellipsoidal coordinates*, Comm. Pure Appl. Math., 58 (2005), pp. 1149–1187.
[28] H. Weyl, *Über gewöhnliche lineare Differentialgleichungen mit singulären Stellen und ihre Eigenfunktionen (2. Note)*, Gött. Nachr., (1910), pp. 442–467.

SIAM J. MATH. ANAL.
Vol. 40, No. 5, pp. 2132–2153

ABSOLUTELY CONTINUOUS LAWS OF JUMP-DIFFUSIONS IN FINITE AND INFINITE DIMENSIONS WITH APPLICATIONS TO MATHEMATICAL FINANCE*

BARBARA FORSTER†, EVA LÜTKEBOHMERT‡, AND JOSEF TEICHMANN†

Abstract. In mathematical Finance calculating the Greeks by Malliavin weights has proved to be a numerically satisfactory procedure for finite-dimensional Itô-diffusions. The existence of Malliavin weights relies on absolute continuity of laws of the projected diffusion process and a sufficiently regular density. In this article we first prove results on absolute continuity for laws of projected jump-diffusion processes in finite and infinite dimensions and a general result on the existence of Malliavin weights in finite dimension. In both cases we assume Hörmander conditions and hypotheses on the invertibility of the so-called linkage operators. The purpose of this article is to show that for the construction of numerical procedures for the calculation of the Greeks in fairly general jump-diffusion cases one can proceed as in a pure diffusion case. We also show how the given results apply to infinite-dimensional questions in mathematical Finance. There we start from the Vasiček model, and add—by pertaining no arbitrage—a jump-diffusion component. We prove that we can obtain in this case an interest rate model, where the law of any projection is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^M$.

Key words. Malliavin calculus, compound Poisson process, Hörmander condition, Greeks, Malliavin weight, stochastic partial differential equation, jump-diffusion, interest rate theory

AMS subject classifications. 60H07, 60H15, 62P05

DOI. 10.1137/070708822

1. Introduction.

We shall consider in this article the question of whether the law of $\mathbf{l}(X_t^x)$, for a finite-dimensional projection $\mathbf{l} : H \to \mathbb{R}^M$, is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^M$, where X_t^x is the solution of the stochastic (partial) differential equation (SPDE)

$$dX_t^x = (AX_{t^-}^x + \alpha(X_{t^-}^x))dt + \sum_{i=1}^{d} V_i(X_{t^-}^x)dB_t^i + \sum_{j=1}^{m} \delta_j(X_{t^-}^x)dL_t^j, \tag{1.1}$$

$$X_0^x = x \in H \tag{1.2}$$

and H is a possibly infinite-dimensional separable Hilbert space. We refer to the previous equation loosely speaking as a jump-diffusion on the Hilbert space H, pointing out that the involved Lévy processes are of finite type. For sake of simplicity we shall always work with the cadlag integrand $t \mapsto X_{t^-}^x$, even though for the dt and dB_t integrals this is superfluous. In the infinite-dimensional setting we are not aware of results on absolute continuity of the projected process in the jump-diffusion case.

*Received by the editors November 20, 2007; accepted for publication (in revised form) September 23, 2008; published electronically January 23, 2009.
http://www.siam.org/journals/sima/40-5/70882.html

†Department of Mathematical Methods in Economics, Vienna University of Technology, Research Group e105 Financial and Actuarial Mathematics, Wiedner Hauptstrasse 8-10, A-1040 Wien, Austria (bforster@fam.tuwien.ac.at, jteichma@fam.tuwien.ac.at). The first author acknowledges the support from the FWF-project P15889 "Utility Maximization in Incomplete Financial Markets" and from the FWF-Wissenschaftskolleg W 8 "Differential Equation Models in Science and Engineering." The third author acknowledges the support from the RTN network HPRN-CT-2002-00281 and from the FWF-grant Y328.

‡Institute of Social Sciences and Economics, University of Bonn, Adenauerallee 24-42, D-53113 Bonn, Germany (eva.luetkebohmert@uni-bonn.de).

Related work has been done in [5] for the construction of first variation processes. In the diffusion case we refer the reader to the work [3] and the references therein and in particular to the recently published inspiring results of Jonathan Mattingly; see, for instance [1]. We point out that we deal here with SPDEs and stochastic differential equations (SDEs) at the same time, where the latter case appears in this setting when the state (Hilbert) space H is finite-dimensional.

Very satisfying results in the finite-dimensional setting with Lévy processes of infinite type have been obtained in [11] through a generalization of the Norris lemma to Lévy processes. These results have been built upon our results presented in this work for the finite activity case (see section 7). Substantial work with respect to absolute continuity and smoothness of the density has already been published in the 1980s, the most prominent being [7] and [6]. Therein several questions of extension of hypoellipticity results (and Malliavin Calculus) to jump-processes are discussed and completely solved; however, the problem of a hypoelliptic diffusion together with a finite-activity jump-structure remained open. It has to be pointed out here that—in contrast to [6]—we do not need any extension of Malliavin Calculus to jump-processes for our results (see also the discussion in Remark 5). This gap was filled by the announced results of [27], but several proofs therein are extremely short. Recently, motivated by questions from financial mathematics (see section 8 for an outline of the problem), there has been increasing interest in those results; see the works [2], [13], [14], and the references therein. This article aims to work out the most general finite activity case under Hörmander conditions on the diffusion part.

From the point of view of existence and uniqueness for jump-diffusions in infinite dimensions, our main reference is [15] and the references therein. Since we consider jump-diffusions as concatenated diffusions on Hilbert space, we mention [12] as the main reference for existence and uniqueness results but also [4] and [10] for many interesting constructions and ideas.

There are two applications added to this work. The first is the Heath–Jarrow–Morton (HJM) equation (as presented in [15]), where we show that the innocent Vasiček model (see, for instance, the seminal work [9]) with a certain jump-structure triggered by a one-dimensional Poisson process yields, under no-arbitrage assumptions, a model where not only do no finite-dimensional realizations exist but also where every projection into a finite-dimensional subspace admits a density (compare with the notion of generic interest rate evolutions from [3]). The second application is concerned with concrete formulas for the calculation of Malliavin weights. There our message is that one can think Poisson-trajectory–wise; i.e., the results from [18] or [17] can be literally applied by replacing the diffusion process by the respective jump-diffusion process.

When we analyze jump-diffusions with values in Hilbert spaces, loosely speaking the following facts hold true:

- Between two consecutive jumps of the jump-diffusion we are given an ordinary diffusion.
- At a jump we add to the left limit the jump size (which usually depends on the left limit, too). In [25, Chapter V.10, Hypothesis (H3)], this operation is formalized by the so-called *linkage operators* $x \mapsto x + \mu\delta^j(x)$, which encode what happens at a jump of size μ at x. We shall apply this notion here, too.

Hence the following picture arises:

- In order to obtain absolute continuity of the projected diffusion process, we need the Hörmander condition to be in force. Otherwise we cannot expect—

conditioned on the event of positive probability that no jump occurs—that the law of $\mathbf{l}(X_t^x)$ is absolutely continuous with respect to Lebesgue measure.
- In order to preserve absolute continuity we need the linkage operators to be invertible in a proper sense.

Remark 1. Both conditions are "sine qua non," since it is easy to imagine counterexamples.

In section 2 we fix the general setting of this article. We shall deal with Lévy processes of finite type as drivers of the SDEs, even though we believe that one should be able to prove similar results in the case of many small jumps, too. We also state the main assumptions of this work in section 2 for later use.

In section 3 we state a "folklore" decomposition theorem, which tells that solving a jump-diffusion S(P)DE is the same as solving associated diffusion S(P)DEs and concatenating the solutions by linkage operators at jumps. In section 4 we show that we can also prove results on first variation processes in the spirit of the decomposition theorem. We prove that under our analytic requirements there is in fact a sufficiently regular first variation process.

In section 5 we show by means of Malliavin calculus for a d-dimensional Brownian motion that the law of a projected jump-diffusion is absolutely continuous with respect to Lebesgue measure. In section 6 we introduce a class of examples from mathematical Finance, where we see very directly the phenomenon of absolute continuity arising from the introduction of jumps and the no-arbitrage condition. This shows again that finite-dimensional realizations, as constructed, for instance, in [8], are a rare case in infinite dimensions. In section 7 we restrict our attention to the finite-dimensional setting to show that the density of the absolutely continuous law is in fact smooth by proving that the inverse of the covariance matrix has pth moments for all $p \geq 1$. In section 8 we apply the invertibility of the covariance matrix to the calculation of Greeks. The appendix shows an important estimate implicitly present in the Norris lemma as presented in Nualart's book [24]. A similar result (which could be directly used in section 7 for the proof of the main theorem) can be found in [21, Corollary 3.25]. The article [21] is most likely the source of the first appearance of the precise polynomial time-dependence in the estimate of the L^p-norm of the inverse of the covariance matrix. We have been choosing here the path via the Norris lemma; we explain the estimate by redoing its proof in the appendix.

2. Setting and assumptions. Let $(\Omega, \mathcal{F}, P, (\mathcal{F}_t)_{t\geq 0})$ be a filtered probability space where the filtration $(\mathcal{F}_t)_{t\geq 0}$ satisfies the usual conditions. Let $(B_t)_{t\geq 0}$ be a d-dimensional Brownian motion and $(L_t^j)_{t\geq 0}$, $j = 1, \ldots, m$, be m independent compound Poisson processes given by

$$L_t^j := \sum_{k=1}^{N_t^j} Z_k^j,$$

where N_t^j denotes a Poisson process with jump intensity $\widetilde{\lambda_j} > 0$ and $Z^j = (Z_k^j)_{k\geq 1}$ is an independently and identically distributed sequence of random variables with distribution μ_j for $j = 1, \ldots, m$ such that each μ_j admits all moments. The compensated compound Poisson process reads as

$$L_t^j - E(L_t^j) = L_t^j - \lambda_j t,$$

where $\lambda_j = E(Z_1^j)\widetilde{\lambda_j}$ is the average jump size times the jump rate.

We could equally take an $\mathbb{R}^m$-valued Lévy process of finite type, i.e., introduce a dependence structure between the jumps of the components, and all theorems would equally hold true with slightly modified proofs, but we believe that this generalization does not bring further insight.

We assume that all sources of randomness are mutually independent and that the filtration $(\mathcal{F}_t)_{t\geq 0}$ is the natural filtration with respect to $(B_t, L_t^1, \ldots, L_t^m)_{t\geq 0}$. Let H be a separable Hilbert space. We fix furthermore a strongly continuous semigroup S on H with generator A. Let $\alpha, V_1, \ldots, V_d$, the diffusion vector fields, and $\delta_1, \ldots, \delta_m$, the jump vector fields, be C^∞-bounded on H; that is, the vector fields are infinitely often differentiable with bounded partial derivatives of all proper orders $n \geq 1$. We consider the mild cadlag solution $(X_t^x)_{t\geq 0}$ of an SDE

$$dX_t^x = (AX_{t^-}^x + \alpha(X_{t^-}^x))dt + \sum_{i=1}^{d} V_i(X_{t^-}^x)dB_t^i + \sum_{j=1}^{m} \delta_j(X_{t^-}^x)dL_t^j, \tag{2.1}$$

$$X_0^x = x \in H. \tag{2.2}$$

See [15] for all necessary details on existence and uniqueness of the previous equation.

The previous conditions are slightly more than standard for existence and uniqueness of mild solutions; i.e., in [15] the authors need Lipschitz conditions on the vector fields, whereas we assume them to be C^∞-bounded. In order to speak about absolute continuity of projections to $\mathbb{R}^M$ we shall need more assumptions; in particular, for conclusions drawn from the geometry of the given vector fields $\alpha, V_1, \ldots, V_d, \delta_1, \ldots, \delta_m$ several quite strong analytic requirements are necessary. We group the assumptions into three groups and indicate in each section which assumptions we shall need.

Let $\mathbf{l} : H \to \mathbb{R}^M$ be a projection; then we want to know whether the law of $\mathbf{l}(X_t^x)$ is absolutely continuous with respect to Lebesgue measure and if the density is smooth. Following the short discussion in the introduction, we need the Hörmander conditions to be in force, and we need to suppose invertibility on linkage operators.

We apply the following notation for Hilbert spaces $\mathrm{dom}(A^k)$:

$$\begin{aligned} \mathrm{dom}(A^k) &:= \left\{h \in H \,|\, h \in \mathrm{dom}(A^{k-1}) \text{ and } A^{k-1}h \in \mathrm{dom}(A)\right\}, \\ ||h||^2_{\mathrm{dom}(A^k)} &:= \sum_{i=0}^{k} ||A^i h||^2, \\ \mathrm{dom}(A^\infty) &= \bigcap_{k\geq 0} \mathrm{dom}(A^k), \end{aligned}$$

which we need in order to specify the analytic conditions.

ASSUMPTION 1. *We assume that the generator A of S generates a* strongly continuous group. *We assume furthermore that $\alpha, V_1, \ldots, V_d$, the diffusion vector fields, and $\delta_1, \ldots, \delta_m$, the jump vector fields, are C^∞-bounded on the Hilbert spaces* $\mathrm{dom}(A^k)$ *for $k \geq 0$; that is, the vector fields are infinitely often differentiable with bounded partial derivatives of all proper orders $n \geq 1$ on the Hilbert space* $\mathrm{dom}(A^k)$ *for $k \geq 0$.*

ASSUMPTION 2. *We take Assumption* 1 *for granted; i.e., we can consider all vector fields on the space* $\mathrm{dom}(A^k)$ *for $k = 0, \ldots, \infty$. For a proper statement of the Hörmander condition we apply the "geometrically relevant" drift*

$$V_0(x) = Ax + \alpha(x) - \frac{1}{2}\sum_{i=1}^{d} \mathrm{T}\, V_i(x) \cdot V_i(x)$$

for $x \in \mathrm{dom}(A)$ and call V_0 the Stratonovich drift of the diffusion. Recall the (tangent) directional derivative operator T *defined through*

$$\mathrm{T}\,V(x)\cdot v = \frac{d}{d\epsilon}\Big|_{\epsilon=0} V(x+\epsilon v).$$

Lie brackets can only be calculated on the Fréchet space $\mathrm{dom}(A^\infty)$ and there we formulate the Hörmander condition. We assume that the distribution $\mathcal{D}(x)$ generated by the vector fields

$$\begin{aligned} &V_1(x),\dots,V_d(x), \quad [V_i(x),V_j(x)]\ (i,j=0,1,\dots,d),\\ &[V_i(x),[V_j(x),V_k(x)]]\ (i,j,k=0,1,\dots,d),\dots \end{aligned} \tag{2.3}$$

is dense in H for one $x \in \mathrm{dom}(A^\infty)$.

ASSUMPTION 3. *We assume that the inverse of $x \mapsto x + z\delta_j(x)$ exists and is C^∞-bounded on each $\mathrm{dom}(A^k)$ for $z \in \mathrm{supp}(\mu_j)$, $j = 1,\dots,m$, and $k \geq 0$ (recall that μ_j was the distribution of the random variable Z_j).*

Remark 2. As far as Assumption 1 is concerned, we do believe that the assertions of this paper also hold true for (most) strongly continuous semigroups. A proof based on an application of the Szőkefalvi–Nagy theorem can be found in the recent preprint [26]; therefore, we could replace the assumption that A generates a strongly continuous group by the assumption that A generates a pseudocontractive, strongly continuous semigroup. This includes most of the second order partial differential operators. However, for this paper we do always assume the group property for the sake of simplicity.

Remark 3. The Hörmander condition could not be formulated without the analytic part of Assumption 1.

Example 1. In order to show examples of vector fields, which are C^∞-bounded on $\mathrm{dom}(A^k)$, consider the following structure. Let H be a separable Hilbert space and A the generator of a strongly continuous semigroup. We know that $\mathrm{dom}(A^\infty)$ is a Fréchet space and an injective limit of the Hilbert spaces $\mathrm{dom}(A^k)$ for $k \geq 0$. Following the analysis as developed in [16] (see also [20] and [19], where the analytic concepts have been originally developed), we can consider vector fields $V : U \subset H \to \mathrm{dom}(A^\infty)$. If V is smooth in the sense explained in [16] and has the property that its derivatives of proper order $n \geq 1$ are bounded on $U \subset H$, then V is obviously a C^∞-bounded vector field, and additionally $V|_{\mathrm{dom}(A^\infty)}$ is a Banach map-vector field in the sense of [16]. Such vector fields constitute a class, where Assumptions 1–3 can be readily checked.

3. Decomposition theorem for jump-diffusions on Hilbert spaces. In order to properly understand how to apply the Malliavin calculus, we state the following rather obvious structure theorem on jump-diffusions, which simply takes into account that stochastic integration with respect to the Poisson process follows the rules of Lebesgue–Stieltjes integration (see, for instance, [25] for a general exposition). Here we need only that the vector fields are C^∞-bounded on H in order to guarantee existence and uniqueness of the respective equations.

THEOREM 1. *Let $(\Omega,\mathcal{F},P,(\mathcal{F}_t)_{t\geq 0})$ be a filtered probability space, $(B_t)_{t\geq 0}$ be a d-dimensional Brownian motion, and $(L_t^j)_{t\geq 0}$ be m independent compound Poisson processes for $j = 1,\dots,m$, such that the filtration is the natural filtration with respect to $(B_t, L_t^1,\dots,L_t^m)_{t\geq 0}$. Let S be a strongly continuous semigroup with generator A on H. Let $\alpha, V_1,\dots,V_d$, the diffusion vector fields, and $\delta_1,\dots,\delta_m$, the jump vector fields,*

be C^∞-bounded on H, and consider the cadlag solution $(X_t^x)_{0\le t\le T}$ of an SPDE

$$dX_t^x = (AX_{t^-}^x + \alpha(X_{t^-}^x))dt + \sum_{i=1}^{d} V_i(X_{t^-}^x)dB_t^i + \sum_{j=1}^{m} \delta_j(X_{t^-}^x)dL_t^j, \tag{3.1}$$

$$X_0^x = x. \tag{3.2}$$

Let η denote a piecewise constant cadlag trajectory $\eta : \mathbb{R}_{\ge 0} \to \mathbb{R}^m$ of the compound Poisson process L with finitely many jumps on compact intervals and starting at 0. We consider

$$\begin{aligned} dY_{s,t}^{x,\eta} = &\Big(AY_{s,t^-}^{x,\eta} + \alpha(Y_{s,t^-}^{x,\eta})\Big)dt + \sum_{i=1}^{d} V_i(Y_{s,t^-}^{x,\eta})dB_t^i \\ &+ \sum_{j=1}^{m} \delta_j(Y_{s,t^-}^{x,\eta})d\eta^j(t), \end{aligned} \tag{3.3}$$

$$Y_{s,s}^{x,\eta} = x. \tag{3.4}$$

Then $(Y_{s,t}^{x,\eta})_{s\ge 0}$ can be given explicitly in terms of the jump times τ_n of η for $n \ge 0$ and the diffusion process between two consecutive jumps:

$$\begin{aligned} Y_t^{x,\eta} &:= Y_{0,t}^{x,\eta} \text{ for } 0 \le t < \tau_1, \\ Y_t^{x,\eta} &:= Y_{\tau_1,t}^{y,\eta}\big|_{y = Y_{0,\tau_1^-}^{x,\eta} + \sum_{j=1}^m \delta_j(Y_{0,\tau_1^-}^{x,\eta})\Delta\eta^j(\tau_1)} \text{ for } \tau_1 \le t < \tau_2 \\ &\vdots \\ Y_t^{x,\eta} &:= Y_{\tau_{n-1},t}^{y,\eta}\big|_{y = Y_{0,\tau_{n-1}^-}^{x,\eta} + \sum_{j=1}^m \delta_j(Y_{0,\tau_{n-1}^-}^{x,\eta})\Delta\eta^j(\tau_{n-1})} \text{ for } \tau_{n-1} \le t < \tau_n. \end{aligned}$$

Here we write $\Delta\eta(t) := \eta(t) - \eta(t^-)$ for $t \ge 0$. We define the process $(Y_t^{x,L})_{t\ge 0}$ by inserting the compound Poisson process L for η into $(Y_t^{x,\eta})_{t\ge 0}$. The resulting process $(Y_t^{x,L})_{t\ge 0}$ is then indistinguishable from $(X_t^x)_{t\ge 0}$.

Proof. For the proof we refer the reader to [25, Chapter V.10, Theorem 57], particularly with respect to the conditioning on the jump part. The proof remains unchanged in the infinite-dimensional setting; see [15] for the existence and uniqueness proof on separable Hilbert spaces. □

Remark 4. For future use we shall always assume that the first jumping time of η is strictly positive, $\tau_1 > 0$, and that each time corresponds to the jump of exactly one coordinate process L^j, which is true for almost all trajectories of the compound Poisson process L. Notice that the dependence of $(Y_t^{x,\eta})_{t\ge 0}$ on the jump times of η is continuous but certainly not smooth since the jump times are inserted instead of the time of a hypoelliptic diffusion process.

Remark 5. Notice that one can also interpret the result in the following way: consider the solution $(X_t^x)_{t\ge 0}$ of (3.1) as an element of $L^2(\Omega_1 \times \Omega_2; H)$, where Ω_1 carries the Brownian motion part (with natural filtration), Ω_2 carries the Poisson part (with natural filtration), and $\Omega_1 \times \Omega_2$ is equipped with the respective product σ-algebra. Then we know by Fubini's theorem that

$$L^2(\Omega_1 \times \Omega_2; H) = L^2(\Omega_2; L^2(\Omega_1; H)).$$

The previous theorem only clarifies the jump-diffusion structure of the dependence on Ω_2. In other words, between jumps we have ordinary diffusions, and at a jump we link by linkage operators.

4. First variation processes. In order to calculate Malliavin derivatives, which is crucial for arguments on absolute continuity, we need precise statements on first variation processes of jump-diffusions. For later purposes, but also in order to see results on the inverse of the first variation process easily, we write our equations in the Stratonovich notation. This is not innocent in infinite dimensions, since mild solutions are in general *not* semimartingales, and therefore the Stratonovich notation fails to be applicable in general. However, by Assumption 1 we are able to determine whether we are given a semimartingale, or not, by analyzing the initial value of the process. Indeed, for fixed $k \geq 0$, if $x \in \mathrm{dom}(A^{k+1})$, then there is a mild solution taking values in $\mathrm{dom}(A^{k+1})$ of the Itô SDE. However, this solution process has to coincide, by uniqueness, with the solution process obtained by considering the same equation on $\mathrm{dom}(A^k)$ with an initial value in $\mathrm{dom}(A^{k+1})$. Therefore, the mild solution in $\mathrm{dom}(A^{k+1})$ is a strong solution in $\mathrm{dom}(A^k)$. Therefore, we assume Assumptions 1 and 3 to be in force in this section.

By the previous arguments for the given SDE (3.1), we can switch to Stratonovich notation for $x \in \mathrm{dom}(A)$ and obtain

$$dX_t^x = V_0(X_{t^-}^x)dt + \sum_{i=1}^{d} V_i(X_{t^-}^x) \circ dB_t^i + \sum_{j=1}^{m} \delta_j(X_{t^-}^x)dL_t^j$$

with the Stratonovich drift given by

$$V_0(x) := Ax + \alpha(x) - \frac{1}{2}\sum_{i=1}^{d} \mathrm{T}\, V_i(x) \cdot V_i(x)$$

for $x \in \mathrm{dom}(A)$. Recall the tangent (derivative) operator T:

$$\mathrm{T}\, V(x) \cdot v = \frac{d}{d\epsilon}\Big|_{\epsilon=0} V(x + \epsilon v).$$

We do also consider the SDE (3.3) with respect to one trajectory η and switch to Stratonovich notation there, too. The following theorem states the result on the first variation process along one trajectory η, which yields in what follows the same result by inserting the compound Poisson process L for η. Notice that the trajectory η is such that the first jumping time is strictly positive and that at each jumping time τ_n for $n \geq 1$ only one coordinate jumps.

THEOREM 2. *Assume Assumptions* 1 *and* 3 *hold. We fix* $k \geq 0$. *The* first variation process $(J_{s\to t}(x,\eta))_{t\geq s}$ *associated with* $(Y_t^{x,\eta})_{t\geq 0}$ *on* $\mathrm{dom}(A^k)$ *is well defined and satisfies the SDE*

$$\begin{aligned} dJ_{s\to t}(x,\eta)\cdot h &= \Big(AJ_{s\to t-}(x,\eta)\cdot h + \mathrm{T}\,\alpha(Y_{s,t^-}^{x,\eta})\cdot J_{s\to t^-}(x,\eta)\cdot h\Big)dt \\ &\quad + \sum_{i=1}^{d}\Big(\mathrm{T}\, V_i(Y_{s,t^-}^{x,\eta})\cdot J_{s\to t^-}(x,\eta)\cdot h\Big)dB_t^i \\ &\quad + \sum_{j=1}^{m}\Big(\mathrm{T}\,\delta_j(Y_{s,t^-}^{x,\eta})\cdot J_{s\to t^-}(x,\eta)\cdot h\Big)d\eta^j(t), \end{aligned}$$

$$J_{s\to s}(x,\eta)\cdot h = h \tag{4.1}$$

for $h, x \in \mathrm{dom}(A^k)$ *and* $t \geq s$. *The Itô equation has a unique global mild solution for* $h, x \in \mathrm{dom}(A^k)$, *and* $J_{s\to t}(x,\eta)$ *defines a continuous linear operator on* $\mathrm{dom}(A^k)$, *which is invertible if* $x \in \mathrm{dom}(A^{k+1})$.

The Stratonovich equation on $\mathrm{dom}(A^k)$ *in turn is well defined only for* $h, x \in \mathrm{dom}(A^{k+1})$. *We apply the (formal) notation here:*

$$\mathrm{T}\, V_0(x)v = Av + \mathrm{T}\, \alpha(x)v - \frac{1}{2}\mathrm{T}\left(x \mapsto \sum_{i=1}^{d} \mathrm{T}\, V_i(x) \cdot V_i(x)\right) v$$

for $x \in \mathrm{dom}(A)$ *and* $v \in \mathrm{dom}(A)$.

$$\begin{aligned} dJ_{s\to t}(x,\eta)\cdot h &= \Big(\mathrm{T}\, V_0(Y^{x,\eta}_{s,t^-})\cdot J_{s\to t^-}(x,\eta)\cdot h\Big)dt \\ &\quad + \sum_{i=1}^{d}\Big(\mathrm{T}\, V_i(Y^{x,\eta}_{s,t^-})\cdot J_{s\to t^-}(x,\eta)\cdot h\Big)\circ dB^i_t \\ &\quad + \sum_{j=1}^{m}\Big(\mathrm{T}\, \delta_j(Y^{x,\eta}_{s,t^-})\cdot J_{s\to t^-}(x,\eta)\cdot h\Big)d\eta^j(t), \\ J_{s\to s}(x,\eta)\cdot h &= h \end{aligned} \tag{4.2}$$

for $h, x \in \mathrm{dom}(A^{k+1})$ *and* $t \geq s$. *The adjoint of the inverse*

$$Z^{x,h}_t := \left(J_{s\to t}(x,\eta)^{-1}\right)^* \cdot h,$$

if it exists, should satisfy the following Stratonovich equation at the point x *in direction* h*:*

$$\begin{aligned} dZ^{x,h}_t &= -\Big(\mathrm{T}\, V_0(Y^{x,\eta}_{s,t^-})^*\cdot Z^{x,h}_{t^-}\Big)dt - \sum_{i=1}^{d}\Big(\mathrm{T}\, V_i(Y^{x,\eta}_{s,t^-})^*\cdot Z^{x,h}_{t^-}\Big)\circ dB^i_t \\ &\quad - \sum_{j=1}^{m}\Big(\mathrm{T}\, \delta_j(Y^{x,\eta}_{s,t^-})^*\cdot Z^{x,h}_{t^-}\Big)d\eta^j(t) \\ &\quad + \sum_{j=1}^{m}\Bigg(\Big(\mathrm{T}\, \delta_j(Y^{x,\eta}_{s,t^-})^2\Big)^*\cdot\Big(\Big(id_H + \Delta\eta^j(t)\,\mathrm{T}\, \delta_j(Y^{x,\eta}_{s,t^-})\Big)^{-1}\Big)^* \\ &\qquad \cdot Z^{x,h}_{t^-}(\Delta\eta^j(t))^2\Bigg) \end{aligned} \tag{4.3}$$

for $h, x \in \mathrm{dom}(A^{k+1})$ *and* $t \geq s \geq 0$ *(here we applied the notions of* [25]*).*

Remark 6. The completely analogous theorem holds when we replace η by a compound Poisson process L. We do not state this theorem again, but we point out that we even have moment estimates for the respective processes, which is the only additional relevant information. To be precise, the first variation process $J_{s\to t}(x)\cdot h$, which equals $J_{s\to t}(x, L)$ by construction, has bounded second moments by [15].

Proof. Under our Assumption 1, the regularity in the initial values is clear by well-known results from [12] and the chain rule on Hilbert spaces (recall that the linkage operators are smooth). We are allowed to pass to the Stratonovich decomposition since we integrate semimartingales by Itô's formula on Hilbert spaces for $x, h \in \mathrm{dom}(A^{k+1})$ due to the arguments of [3]: the core assertion here is that we can replace H by each $\mathrm{dom}(A^k)$ for some $k \geq 0$, which means in turn if we start in $\mathrm{dom}(A^{k+1})$ and obtain a mild solution there, it is indeed a strong solution considered on $\mathrm{dom}(A^k)$ for $k \geq 0$. It remains to show the invertibility results on the respective first variation processes.

Left invertibility of the first variation $J_{s\to t}(y,.)$ follows by Itô's formula since we have cadlag trajectories with finitely many jumps. Calculating the semimartingale decomposition of $(Z_t^{y,\cdot})^* \cdot J_{0\to t}(x,\eta)$ given by (4.2) and (4.3) yields the result

$$(Z_t^{x,\cdot})^* J_{0\to t}(x,\eta) = id_{\mathrm{dom}(A^k)}.$$

Thus, the solution of (4.3) is the left inverse of $J_{s\to t}$.

We prove that the left inverse is also the right inverse by the same reasoning as in the proof of Proposition 2 in [3]. Therefore, we choose an orthonormal basis $(g_i)_{i\geq 1}$ of $\mathrm{dom}(A^k)$ which lies in $\mathrm{dom}(A^{k+1})$. Then we can compute the semimartingale decomposition of

$$\begin{aligned}&\sum_{i=1}^N \langle (Z_t^{x,h_1})^*, g_i\rangle_{\mathrm{dom}(A^k)} \langle g_i, J_{s\to t}(x,\eta)^* \cdot h_2\rangle_{\mathrm{dom}(A^k)}\\ &= \sum_{i=1}^N \langle h_1, Z_t^{x,g_i}\rangle_{\mathrm{dom}(A^k)} \langle J_{s\to t}(x,\eta)\cdot g_i, h_2\rangle_{\mathrm{dom}(A^k)}\end{aligned}$$

for $h_1, h_2 \in \mathrm{dom}(A^{k+1})$ and $N \geq 1$. Applying the Stratonovich decomposition and by adjoining, we can free the g_i's and pass to the limit, which yields the vanishing finite variation and martingale part. Hence

$$\begin{aligned}&\langle J_{s\to t}(x,\eta)(Z_t^{x,h_1})^*, h_2\rangle_{\mathrm{dom}(A^k)}\\ &= \lim_{N\to\infty} \sum_{i=1}^N \langle (Z_t^{x,h_1})^*, g_i\rangle_{\mathrm{dom}(A^k)} \langle g_i, J_{s\to t}(x,\eta)^* \cdot h_2\rangle_{\mathrm{dom}(A^k)}\\ &= \langle h_1, h_2\rangle_{\mathrm{dom}(A^k)},\end{aligned}$$

which is what a right inverse should satisfy. □

5. Absolutely continuous laws in finite and infinite dimensions. In this section we assume Assumptions 1, 2, and 3. We want to determine by means of Malliavin calculus whether the law of $\mathbf{l}(Y_t^{x,\eta})$ is absolutely continuous with respect to Lebesgue measure for $t > 0$.

For details on Malliavin calculus see [22] and [24], where in particular the derivative operator and the Skorohod integral for Malliavin calculus with respect to a d-dimensional Brownian motion are defined. Notice that we do not need a Malliavin calculus with respect to the Poissonian trajectories, since we calculate Poisson-trajectory–wise.

Our first task is the calculation of the Malliavin derivative for a fixed cadlag path η. In a second step, we consider the composed problem, where we replace η by a compound Poisson process L as outlined before. Therefore, we first fix a piecewise constant cadlag trajectory $\eta : \mathbb{R}_{\geq 0} \to \mathbb{R}^m$ of the process $(L_t^1, \dots, L_t^m)_{t\geq 0}$ with finitely many jumps on compact intervals starting at 0.

THEOREM 3. *We take Assumptions* 1, 2, *and* 3 *for granted, where* $x \in \mathrm{dom}(A^\infty)$ *denotes the point where the Hörmander condition* (2.3) *holds true. Let* $(Y_t^x)_{t\geq 0}$ *denote the unique cadlag solution of* (3.3). *Then for projections* $\mathbf{l} : H \to \mathbb{R}^M$ *the law of* $\mathbf{l}(Y_t^x)$ *is absolutely continuous with respect to Lebesgue measure on* $\mathbb{R}^M$ *for* $t > 0$.

Proof. Fix $t > 0$. We are able to write the Malliavin derivative of Y_t^x for each Poissonian trajectory η:

$$D_s^i(\mathbf{l}\circ Y_t^{x,\eta}) = \mathbf{l}\circ J_{0\to t}(x) J_{0\to s}(x)^{-1} V_i(Y_{s^-}^{x,\eta}) 1_{[0,t]}(s).$$

We can calculate the *Malliavin covariance matrix* γ as

$$\langle \gamma(\mathbf{l} \circ Y_t^{x,\eta})\xi, \xi\rangle := \sum_{i=1}^{d} \int_0^t \langle \mathbf{l} \circ J_{0\to t}(x) J_{0\to s}(x)^{-1} V_i(Y_{s^-}^{x,\eta}), \xi\rangle^2 ds.$$

Consequently, the covariance matrix $\gamma(\mathbf{l} \circ Y_t^{x,\eta})$ can be calculated in the usual way via the reduced covariance matrix

$$\langle C_t\xi, \xi\rangle := \sum_{i=1}^{d} \int_0^t \left\langle J_{0\to s}(x)^{-1} V_i(Y_{s^-}^{x,\eta}), \xi\right\rangle^2 ds$$

through the relation

$$\gamma(\mathbf{l} \circ Y_t^{x,\eta}) = (\mathbf{l} \circ J_{0\to t}(x)) C_t (\mathbf{l} \circ J_{0\to t}(x))^*,$$

where $*$ denotes the adjoint operator with respect to the Hilbert space structures on H and $\mathbb{R}^M$. We assume n jumps of η on $[0,t]$, and we denote by $0 = \tau_0 < \tau_1 < \cdots < \tau_n \le t$ the sequence of jump times of η. For convenience, we denote the last point in time t by τ_{n+1}, even if $\tau_{n+1} = \tau_n$, which can in principle happen. Hence we can decompose:

$$\langle C_t\xi, \xi\rangle := \sum_{k=0}^{n} \sum_{i=1}^{d} \int_{\tau_k}^{\tau_{k+1}} \left\langle J_{0\to s}(x)^{-1} V_i(Y_{s^-}^{x,\eta}), \xi\right\rangle^2 ds = \sum_{k=0}^{n} \langle C_t^k \xi, \xi\rangle.$$

Each of the summands determines a symmetric matrix C_t^k and can be interpreted as a reduced covariance matrix coming from a diffusion between τ_k and τ_{k+1} with initial value $Y_{\tau_k^-}^x$ for $k = 0, \ldots, n$. We do not know whether the Hörmander condition is true everywhere. Therefore, we do not know whether C_t^k is a positive definite operator for each $k \ge 0$. From [3], Theorem 1, we do know, however, that C_t^0 is a positive definite operator and there exist null sets N_0 such that on N_0^c the matrix C_t^0 is invertible. Hence the law of $(\mathbf{l} \circ Y_t^{x,\eta})$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^M$, since $J_{0\to t}(x)$ is invertible and therefore $\gamma(\mathbf{l} \circ Y_t^{x,\eta})$ has empty kernel (Theorem 2.1.2 in [24, p. 86]). □

Remark 7. The same conclusions hold for $Y_{t^-}^{x,\eta}$: notice that $Y_t^{x,\eta} = Y_{t^-}^{x,\eta}$ if there is no jump at t. Otherwise, $Y_t^{x,\eta} = Y_{t^-}^{x,\eta} + \sum_{j=1}^{m} \delta_j(Y_{0,t^-}^{x,\eta}) \Delta\eta^j(t)$, but invertible diffeomorphisms transform absolutely continuous laws into absolutely continuous ones.

Now we extend this theorem to the jump-diffusion process $(X_t^x)_{t\ge 0}$, which is easy since, conditioned on one trajectory η, we do have an absolutely continuous law and this property is not perturbed by integration due to Fubini's theorem.

THEOREM 4. *We take Assumptions* 1, 2, *and* 3 *for granted, where* $x \in \mathrm{dom}(A^\infty)$ *denotes the point where the Hörmander condition* (2.3) *holds true. Let* $(X_t^x)_{t\ge 0}$ *denote the unique cadlag solution of* (3.1). *Then for projections* $\mathbf{l} = (l_1, \ldots, l_k) : H \to \mathbb{R}^M$ *the law of* $\mathbf{l}(X_t^x)$ *is absolutely continuous with respect to the Lebesgue measure on* $\mathbb{R}^M$ *for* $t > 0$. *Notice that* $\mathbf{l}(X_t^x)$ *and* $\mathbf{l}(X_{t^-}^x)$ *have the same distribution.*

Proof. The proof applies the following simple corollary of Fubini's theorem on $\mathbb{R}^M$ with Lebesgue measure λ and a probability space $(\Omega, \mathcal{F}, P)$: let ν be a probability measure on $\mathbb{R}^M \times \Omega$ such that there is random density $p : \mathbb{R}^M \times \Omega \to \mathbb{R}_{\ge 0}$ with

$$\int_{\mathbb{R}^M \times \Omega} f(x,\eta) p(x,\eta) (\lambda \otimes P)(dx, d\eta) = \int_{\mathbb{R}^M \times \Omega} f(x,\eta) \nu(dx, d\eta);$$

then the marginal of ν on $\mathbb{R}^M$ is absolutely continuous with respect to Lebesgue measure λ with density

$$p(x) = \int_\Omega p(x,\eta) P(d\eta)$$

for almost all $x \in \mathbb{R}^M$. In our case we know that the law of $(\mathbf{l}(X_t^x)$ is absolutely continuous for almost all trajectories η of the compound Poisson processes L (precisely those where $\tau_1 > 0$, where only one coordinate jumps at each jumping time, and finitely many jumps occur on compact intervals); the probability measure ν corresponds to the distribution of $(\mathbf{l}(X_t^x), L)$, where we choose Ω as the space of cadlag trajectories on $\mathbb{R}_{\geq 0}$ with values in $\mathbb{R}^m$. Finally, we have that the law of $(\mathbf{l}(X_t^x)$ is $p(x)\lambda(dx)$. □

6. Applications of the infinite-dimensional result to interest rate theory. In mathematical Finance, the theory of interest rates deals with the market of interest rate related products like swaps, bonds, bills, etc. If one considers default-free products one can crystallize from the data of real markets the prices of default-free zero-coupon bonds for any maturity. A zero-coupon bond contract with maturity T (a calendar date) can be entered at calendar time $t \leq T$ and (certainly) pays 1 unit of currency at maturity time T. Therefore, bonds reflect the level of interest rate between time t and maturity time T. No coupons are paid between t and T, which explains the notion of "zero-coupon bond." We denote the price of a default-free zero-coupon bond with maturity T at time $t \leq T$ by $P(t,T)$. Commonly one assumes that bond prices are at least C^1 with respect to T, i.e.,

$$P(t,T) = \exp\left(-\int_t^T f(t,r)dr\right)$$

for $0 \leq t \leq T$, where $f(t,T)$ denotes the forward rate. This leads to the concept of the short rate

$$R_t = f(t,t)$$

for $t \geq 0$, which corresponds to the level of interest rate for instantaneous transactions from t to $t + dt$. As usual in mathematical Finance, discounted default-free zero-coupon bonds are modeled by semimartingales, and one assumes the existence of an equivalent martingale measure for discounted price processes. This leads to the following fundamental formula with respect to the martingale measure

$$E\left(\exp\left(-\int_t^T R_s ds\right)|\mathcal{F}_t\right) = P(t,T) = \exp\left(-\int_t^T f(t,r)dr\right).$$

The formula simply expresses the fact that the expected value of discounted value of the payoff $P(T,T) = 1$ conditional on today's information equals today's price with respect to the martingale measure. Assuming a jump-diffusion model for $(f(t,T)_{0\leq t\leq T}$ for $T \geq 0$ for the forward rates together with Musiela's parametrization $r(t,T-t) = f(t,T)$ for $0 \leq t \leq T$ leads to the famous Heath–Jarrow–Morton (HJM) equation of interest rate theory, which is an SDE taking values in a Hilbert space of forward rate curves H (and therefore an SPDE). We quote here as leading reference [15], where the no-arbitrage conditions for the HJM equation are discussed in all necessary detail. A

very readable introduction can also be found in [8], particularly for HJM equations with jumps.

The HJM equation has been analyzed from different points of view:

- The question of which HJM equations driven by finitely many Brownian motions admit finite-dimensional realizations has been treated in detail in [16]. This research was inspired by [9], where the geometric approach was introduced. The satisfying answer is that, under quite natural restrictions, finite-dimensional realizations do exist if and only if the corresponding factor processes are affine processes. This is the case, for instance, for Vasiček's model or for the Cox–Ingersoll–Ross model of interest rate theory. In both cases the finite-dimensional realizations are, in fact, two-dimensional. In [8] finite-dimensional realizations are treated for HJM equations with jumps but under the strong restriction that the vector fields do not depend on the forward rate. In this case one can solve the HJM equation explicitly by variation of constants and read off the respective geometric properties of the solution process.
- The question of whether the solution process of an HJM equation always admits a density with respect to Lebesgue measure when projected to a finite-dimensional subspace has been treated in [3] and could be answered affirmatively under Hörmander-type conditions.

We ask here the question—having the theory of the previous sections in mind—of whether a structure of finite-dimensional realizations, such as for Vasiček's model, can be perturbed so strongly through the introduction of jumps that the resulting HJM evolution is "hypoelliptic," i.e., the assumptions of Theorem 3 are fulfilled. We can answer this question affirmatively in the case of a Vasiček model. In contrast to [8], we allow the vector fields to be state-dependent (therefore, we cannot hope for explicit solutions of the HJM equation, and we have to apply local methods from differential geometry to conclude).

We consider the HJM model with jumps

$$\begin{cases} dr_t &= \left(\dfrac{d}{dx}r_{t^-} + \alpha_{HJM}(r_{t^-}) + \beta_{HJM}(r_{t^-})\right)dt + \sigma(r_{t^-})dB_t + \delta(r_{t^-})dN_t, \\ r_0 &= r^* \in H, \end{cases}$$

on some Hilbert space H of forward rate curves as constructed in [3] and [15]. Here $(B_t)_{t\geq 0}$ is a standard Brownian motion, and $(N_t)_{t\geq 0}$ is a standard Poisson process with intensity $\widetilde{\lambda} > 0$ and jump measure $\mu = \delta_1$ (hence $\widetilde{\lambda} = \lambda$). Define

$$\begin{aligned} \Psi_1(z) &\equiv \ln \mathbb{E}\left[e^{zW_1}\right] = \ln\left(e^{\frac{z^2}{2}}\right) = \frac{z^2}{2} \quad \text{and} \\ \Psi_2(z) &\equiv \ln \mathbb{E}\left[e^{zN_1}\right] = \ln\left(\exp\left(\lambda\left(e^z - 1\right)\right)\right) = \lambda\left(e^z - 1\right). \end{aligned}$$

Then we know from [15, equation (2.4)], that

$$\begin{aligned} \alpha_{HJM}(r)(x) &= -\sigma(r)(x)\Psi_1'\left(-\int_0^x \sigma(r)(y)dy\right) \\ &= \sigma(r)(x)\int_0^x \sigma(r)(y)dy \end{aligned}$$

and

$$\begin{aligned}\beta_{HJM}(r)(x) &= -\delta(r)(x)\Psi_2'\left(-\int_0^x \delta(r)(y)dy\right)\\ &= -\lambda\delta(r)(x)\exp\left(-\int_0^x \delta(r)(y)dy\right).\end{aligned}$$

For an explicit example we choose

$$\sigma(r)(x) = \sigma(x) > 0 \quad \text{and} \quad \delta(r)(x) = -\frac{d}{dx}\ln(B(r)(x)),$$

where the vector field B will be determined later. Then we have

$$\alpha_{HJM}(r)(x) = \sigma(x)\int_0^x \sigma(y)dy$$

and

$$\beta_{HJM}(r)(x) = \lambda\frac{d}{dx}\ln(B(r)(x))\frac{(B(r)(x))}{(B(r)(0))} = \lambda\frac{\frac{d}{dx}(B(r)(x))}{(B(r)(0))}.$$

We choose B such that $(B(r)(x))$ is positive on H for $x \in \mathbb{R}$ and $B(r)(0) = 1$ for all $r \in H$, whence δ is well defined. Thus, for such $r \in U$ we have

$$\beta_{HJM}(r)(x) = \lambda\frac{d}{dx}(B(r)(x))$$

and

$$\delta(r)(x) = -\frac{d}{dx}\ln(B(r)(x)).$$

A particular choice in the spirit of Remark 1 is given through

$$B(r)(x) = \psi(x, l(r)),$$

where the maps $y \mapsto \frac{d}{dx}\psi(.,y)$ and $y \mapsto \frac{1}{\psi(.,y)}$ from $\mathbb{R}$ to $\mathrm{dom}(A^\infty) \subset H$ are supposed to be C^∞-bounded with $\psi(0,y) = 1$ for all $y \in \mathbb{R}$. The map l denotes here a non-vanishing linear functional $l : H \to \mathbb{R}$. Hence δ and β are well defined C^∞-bounded vector fields on the whole Hilbert space, and we have global existence of mild solutions.

The Vasiček model is defined by

$$\sigma(r) = \rho\exp(-ax)$$

for $\rho, a > 0$ without any jump component. By [9] and [16] we know that the Vasiček model admits finite-dimensional realizations, as for

$$V_0(r)(x) = \frac{d}{dx}r(x) + \alpha_{HJM}(r)(x)$$

we have

$$\dim(\{V_0, \sigma\}_{LA}(r)) \leq 2,$$

at any point $r \in \mathrm{dom}((\frac{d}{dx})^\infty)$. Here the index LA stands for the Lie algebra generated by the vector fields V_0, σ on $\mathrm{dom}(A^\infty)$. If we add a jump structure as described above and if we choose ψ generic, the two-dimensional structure (a regular finite-dimensional realization in the sense of [16]) is destroyed, since then the drift changes due to no-arbitrage. We obtain a dense Lie algebra if we choose the vector field B generically.

These results might be of interest for recent works in interest rate theory; see, for instance, [23], where under diffusion assumptions hypoellipticity is tested empirically. If one allows for jumps in an HJM model, the phenomenon of hypoellipticity seems to be more generic.

7. Smooth densities for the law X_T^x on $\mathbb{R}^M$. In the what folllows we consider the cases $\dim H = M$ and $\mathbf{l} = \mathrm{id}$, and we choose a coordinate representation $H = \mathbb{R}^M$. We then want to show that the pth power of the inverse of the Malliavin covariance matrix of X_t^x for $t > 0$ can be integrated even with respect to Poisson trajectory η. We therefore need an extension of the Hörmander condition which is called the uniform Hörmander condition.

Following [24], we define

$$\Sigma_0' := \{V_1, \dots, V_d\},$$
$$\Sigma_n' := \left\{ [V_k, V], k = 1, \dots, d, V \in \Sigma_{n-1}'; [V_0, V] + \frac{1}{2}\sum_{i=1}^d [V_i, [V_i, V]], V \in \Sigma_{n-1}' \right\}$$

for $n \geq 1$. We assume that there exist j_0 and $c > 0$ such that

$$\inf_{\xi \in S^{M-1}} \sum_{j=0}^{j_0} \sum_{V \in \Sigma_j'} \langle V(x), \xi \rangle^2 \geq c \tag{7.1}$$

uniformly in $x \in \mathbb{R}^M$.

THEOREM 5. *Assume that* $\dim H < \infty$. *We take Assumptions* 2 *and* 3 *for granted but assume that the Hörmander condition* (2.3) *holds true uniformly on* $\mathbb{R}^M$ *in the sense of* (7.1). *Let* $(X_t^x)_{t \geq 0}$ *denote the unique cadlag solution of* (3.1), *and fix* $t > 0$. *Then the random variable* X_t^x *admits a smooth density with respect to Lebesgue measure on* $\mathbb{R}^M$. *Furthermore, the covariance matrix of* X_t^x *is invertible with p-integrable inverse for all* $p \geq 1$.

Proof. We write the Malliavin derivative of X_t^x,

$$D_s^i X_t^x = J_{0 \to t}(x) J_{0 \to s}(x)^{-1} V_i(X_{s^-}^x) \mathbf{1}_{[0,t]}(s),$$

and calculate the reduced covariance matrix

$$\langle C_t \xi, \xi \rangle = \sum_{i=1}^d \int_0^t \left\langle J_{0 \to s}(x)^{-1} V_i(X_{s^-}^x), \xi \right\rangle^2 ds.$$

We now apply the result from Theorem 1 and the condition on the trajectories of the compound Poisson process (2)

$$\begin{aligned} &\sup_{\xi \in S^{M-1}} P(\langle C_t \xi, \xi \rangle < \epsilon) \\ &= \sup_{\xi \in S^{M-1}} \sum_{n_1, \dots, n_m \geq 0} \left[\prod_{k=1}^m P(N_t^k = n_k) \right] P(\langle C_t \xi, \xi \rangle < \epsilon | N_t^j = n_j \text{ for } j = 1, \dots, m). \end{aligned} \tag{7.2}$$

As in the proof of Theorem 3, we can decompose $\langle C_t\xi,\xi\rangle$ into

$$\langle C_t\xi,\xi\rangle = \sum_{k=0}^{\infty}\sum_{i=1}^{d}\int_{\tau_k\wedge t}^{\tau_{k+1}\wedge t}\left\langle J_{0\to s}(x)^{-1}V_i(X_{s^-}^x),\xi\right\rangle^2 ds,$$

where $\tau_0 = 0 < \tau_1 < \cdots < \tau_n \le \dots$ denotes the sequence of jump times of $(N_t)_{0\le t\le T}$. Hence, we obtain for $n = n_1 + \cdots + n_m$

$$\begin{aligned}&\sup_{\xi\in S^{M-1}} P(\langle C_t\xi,\xi\rangle < \epsilon | N_t^j = n_j,\ j=1,\dots,m)\\ &\le \sup_{\xi\in S^{M-1}} P\left(\sum_{i=1}^{d}\int_{\tau_k\wedge t}^{\tau_{k+1}\wedge t}\left\langle J_{0\to s}(x)^{-1}V_i(X_{s^-}^x),\xi\right\rangle^2 ds < \epsilon \,\middle|\, N_t^j = n_j,\ j=1,\dots,m\right)\end{aligned}$$

for all $0\le k\le n$. Observing that $\max_{0\le k\le n}(\tau_{k+1}-\tau_k)^{K(p)} \ge (\frac{t}{n})^{K(p)}$ (after all, we have only n jumps, so the maximal distance between two consecutive jumps is bigger than $\frac{t}{n}$), we finally obtain

$$P(\langle C_t\xi,\xi\rangle < \epsilon | N_t^j = n_j \text{ for } j=1,\dots,m) \le \epsilon^p$$

for $0\le\epsilon\le(\frac{t}{n})^{K(p)}\epsilon_0(p)$ due to the calculations outlined in the appendix. Note that we can apply the calculations from the appendix, since $J_{0\to s}(x)^{-1}$ is well defined and bounded due to boundedness of $(id + zd\delta_j)^{-1}$ for $z\in\operatorname{supp}(\mu_j)$ and $j=1,\dots,m$. Hence integration with respect to the measures μ_j is possible and yields finite bounds. Recall also that μ_j has moments of all orders; hence X_t^x is L^p, and so is $J_{0\to s}(x)^{-1}$ (see [25] for all necessary details on SDEs).

Let $\Lambda = \inf_{\xi\in S^{M-1}}\langle C_t\xi,\xi\rangle$ be the smallest eigenvalue of the reduced covariance matrix C_t. Following the steps of [24, Lemma 2.3.1], we know that

$$P(\Lambda < \epsilon \mid N_t^j = n_j \text{ for } j=1,\dots,m) \le \text{const}\cdot\epsilon^p$$

for any $p\ge 2$ and $0\le\epsilon\le\left(\frac{t}{n}\right)^{K(p+2M)}\epsilon_0(p+2M) =: \epsilon_{max}$, where the constant depends on the p-norm of C_t. In what follows we shall denote any constant of this type by D. We denote by ρ the law of Λ conditioned on $N_t^j = n_j$ for $j=1,\dots,m$. Consequently, for $j=1,\dots,m$, we have by Fubini's theorem

$$\begin{aligned}E\left(\frac{1}{\Lambda^{p-1}}\,\middle|\, N_t^j = n_j\right) &= E\left(\frac{1}{\Lambda^{p-1}}\cdot\mathbf{1}_{\{\Lambda>\epsilon_{max}\}}\,\middle|\, N_t^j = n_j\right)\\ &\quad + E\left(\frac{1}{\Lambda^{p-1}}\cdot\mathbf{1}_{\{\Lambda\le\epsilon_{max}\}}\,\middle|\, N_t^j = n_j\right)\\ &\le \frac{1}{\epsilon_{max}^{p-1}} + \int_0^{\epsilon_{max}}\frac{1}{z^{p-1}}\rho(dz)\\ &= \frac{1}{\epsilon_{max}^{p-1}} + \int_0^{\epsilon_{max}}(p-1)\int_z^{\infty}\frac{1}{t^p}dt\rho(dz)\\ &= \frac{1}{\epsilon_{max}^{p-1}} + (p-1)\int_0^{\epsilon_{max}}\frac{1}{z^p}\int_0^z\rho(dt)dz\end{aligned}$$

$$
\begin{aligned}
&+ (p-1)\int_{\epsilon_{max}}^{\infty} \frac{1}{z^p}\int_0^{\epsilon_{max}} \rho(dt)dz \\
&\leq \frac{D}{\epsilon_{max}^{p-1}} + D\underbrace{\int_0^{\epsilon_{max}} \frac{1}{z^p}\, z^p dz}_{=\epsilon_{max}} \\
&\leq D\left(\frac{t}{n}\right)^{K(p+2M)} \epsilon_0(p+2M) \\
&+ \frac{D}{[(\frac{t}{n})^{K(p+2M)}\epsilon_0(p+2M)]^{p-1}}.
\end{aligned}
$$

Here we applied $\int_0^z \rho(dt) \leq D \times z^p$ as previously proved. Hence, through the decomposition (7.2),

$$
\begin{aligned}
E\left(\frac{1}{\Lambda^{p-1}}\right) \leq &\sum_{n_1,\ldots,n_m>0} \prod_{k=1}^{m} P(N_t^k = n_k) \\
&\cdot D \cdot \left[\left(\frac{t}{n}\right)^{K(p+2M)} \epsilon_0(p+2M) + \frac{1}{[(\frac{t}{n})^{K(p+2M)}\epsilon_0(p+2M)]^{p-1}}\right] < \infty;
\end{aligned}
$$

the result follows by $n = n_1 + \cdots + n_m$ and by the following fact for any real number K:

$$
\sum_{n_1,\ldots,n_m>0} \frac{\widetilde{\lambda}_1^{n_1}\cdots\widetilde{\lambda}_m^{n_m}}{n_1!\cdots n_m!} e^{-t\tilde{\lambda}_1 n_1 - \cdots - t\tilde{\lambda}_m n_m} t^n (n_1+\cdots+n_m)^K < \infty. \qquad \square
$$

Remark 8. We could have also applied the beautiful results of [21, Corollary 3.25] to evaluate the L^p-norm of the inverse of the covariance matrix between two jumps. Both methods lead to the same result. We have been choosing our approach since we can root it as much as possible in the standard reference [24].

8. Calculating the Greeks in finite dimension. In what follows we consider the case $\dim H = M$ and $\mathbf{l} = \mathrm{id}$ as in the previous section. Once we are given an invertible Malliavin covariance matrix with p-integrable inverse such as in Theorem 5, we can easily calculate derivatives with respect to initial values and obtain explicit formulas for so-called Malliavin weights (see [17] for successful applications of this method in mathematical Finance). We quickly sum up the main idea: in mathematical Finance the gradient of the function $x \mapsto E(f(X_t^x))$ has the meaning of hedging ratios, which control the hedging portfolios away from jumps. Hence for any hedging portfolio corresponding to prices $E(f(X_t^x))$ of a certain derivative at maturity $t > 0$ it is crucial to know $\nabla E(f(X_t^x))$ to perform hedging off jumps.

Very often pricing results in the applications of a weak approximation scheme for the process X, for instance, the Euler–Maruyama scheme. For the calculation of $\nabla E(f(X_t^x))$ in the direction of some vector $v \in H$, basically three methods can be applied:

- a finite difference method to approximate $\nabla E(f(X_t^x)) \cdot v$, resulting in the calculation of $\frac{E(f(X_t^{x+\epsilon v})) - E(f(X_t^x))}{\epsilon}$ ($v \in H$ denotes some vector) for small $\epsilon > 0$;
- a pathwise method applying the formula

$$
\nabla E(f(X_t^x)) \cdot v = E(df(X_t^x) J_{0\to t}(x) \cdot v),
$$

resulting in the weak numerical approximation of $(X_t^x, J_{0\to t}(x))$; and

- the method of Malliavin weights applying the formula

$$\nabla E(f(X_t^x)) \cdot v = E(f(X_t^x)\pi^v),$$

resulting in the weak numerical approximation of (X_t^x, π^v).

The first method is the most robust in the sense that it can be applied under very weak assumptions both on X_t^x and on the payoff f, but the rate of convergence might be very slow since the errors of Monte Carlo evaluations are amplified. The second method works for all reasonable jump-diffusion processes, but one needs Lipschitz conditions on the payoff f. The third method needs the assumptions of Theorem 5 on X_t^x, but no restrictions on the payoff f, which makes the third method attractive for several problems from mathematical Finance, where measurable, non-Lipschitz payoffs (e.g., digital options) are quite usual and hypoellipticity assumptions as in Theorem 5 are common, too.

The implementation of procedures for all three methods has been outlined in [18] in the pure diffusion case. We shall not work on this issue here, since our main message is that one can implement precisely the same methods in pure diffusion cases as in jump-diffusion cases. The important point is that the formulas have the same structure in both cases, a fact we shall point out in this section on several occasions.

We denote in what follows the Skorohod integral (resp., the divergence operator) by δ and its domain by $\mathrm{dom}(\delta)$.

DEFINITION 1. *Assume that $H = \mathbb{R}^M$ and fix $t > 0$ and a direction $v \in \mathbb{R}^M$. We define a set of Skorohod-integrable processes*

$$\mathbb{A}_{t,x,v} = \left\{ a \in \mathrm{dom}(\delta) \text{ such that } \sum_{i=1}^{d} \int_0^t J_{0\to s}(x)^{-1} V_i(X_{s-}^x) a_s^i ds = v \right\}$$

and call it the set of path-perturbations with target-value v.

Remark 9. In the previous definition, as in the whole section, assertions on Skorohod-integrability are meant Poissonian-trajectory–wise.

PROPOSITION 1. *Assume that $H = \mathbb{R}^M$. We take Assumption 3 for granted. Fix $t > 0$ and a direction $v \in \mathbb{R}^M$. Assume furthermore uniform ellipticity; i.e., $M = d$ and there is $c > 0$ such that*

$$\inf_{\xi \in S^{M-1}} \sum_{k=1}^{M} \langle V_k(x), \xi \rangle^2 \geq c.$$

Then $\mathbb{A}_{t,x,v} \neq \emptyset$ and there exists an integrable, real valued random variable π^v (which depends linearly on v) such that for all bounded random variables f we obtain

$$\frac{d}{d\epsilon}|_{\epsilon=0} E(f(X_t^{x+\epsilon v})) = E(f(X_t^x)\pi^v).$$

Such a random variable π^v is called a Malliavin weight and can be obtained through an Itô integral.

Remark 10. The assertion of this theorem corresponds to Assumption (E) in [18] and to the assumptions of [17]. The assumptions are seen as too restrictive since not every problem in mathematical Finance has an elliptic volatility matrix. The formulas of [18] and [17] correspond precisely to the formulas obtained here, which

leads to the assertion that even in the presence of jumps one can apply the same (numerical) methods for the calculation of Greeks as in the pure diffusion cases.

Proof. Here the proof is particularly simple, since we can take a matrix $\sigma(x) := (V_1(x), \dots, V_M(x))$, which is uniformly invertible with bounded inverse. We define

$$a_s := \frac{1}{t}\sigma(X^x_{s^-})^{-1} \cdot J_{0\to s}(x) \cdot v$$

for $0 \le s \le t$ and obtain that $a \in \mathbb{A}_{t,x,v}$. Furthermore, as in [17] and [13], we obtain

$$\pi^v = \sum_{i=1}^{M} \int_0^t a^i_s dB^i_s,$$

since the Skorohod-integrable process a is in fact adapted, left-continuous, and hence Itô-integrable. □

THEOREM 6. *Assume that $H = \mathbb{R}^M$. We take Assumptions 2 and 3 for granted but assume that the Hörmander condition (2.3) holds true uniformly on $\mathbb{R}^M$ (see section 7). Fix $t > 0$ and a direction $v \in \mathbb{R}^M$. Then $\mathbb{A}_{t,x,v} \neq \emptyset$ and there exists an integrable, real valued random variable π^v (which depends linearly on v) such that for all bounded random variables f we obtain*

$$\frac{d}{d\epsilon}|_{\epsilon=0} E(f(X^{x+\epsilon v}_t)) = E(f(X^x_t)\pi^v).$$

We can choose π^v to be the Skorohod integral of any element $a \in \mathbb{A}_{t,x,v} \neq \emptyset$ and call it a Malliavin weight. Moreover, by the explicit construction of a in the proof, we can assert that π^v is the sum of an Itô integral and an integral with respect to Lebesgue measure; see, for instance, [18].

Remark 11. The assertion of this theorem corresponds to Assumption (E′) in [18]. The assumptions (E) and (E′) are fundamental for the third method in [18]. Again the formulas of [18] correspond to the formulas obtained here.

Proof. We take f bounded with bounded first derivative; then we obtain

$$\frac{d}{d\epsilon}|_{\epsilon=0} E(f(X^{x+\epsilon v}_t)) = E(df(X^x_t)J_{0\to t}(x) \cdot v).$$

If there is $a \in \mathbb{A}_{t,x,v}$, we obtain

$$\begin{aligned}
E(df(X^x_t)J_{0\to t}(x) \cdot v) &= E\left(df(X^x_t)\sum_{i=1}^{d}\int_0^t J_{0\to t}(x)J_{0\to s}(x)^{-1}V_i(X^x_{s^-})a^i_s ds\right)\\
&= E\left(\sum_{i=1}^{d}\int_0^t df(X^x_t)J_{0\to t}(x)J_{0\to s}(x)^{-1}V_i(X^x_{s^-})a^i_s ds\right)\\
&= E\left(\sum_{i=1}^{d}\int_0^t D^i_s f(X^x_t)a^i_s ds\right)\\
&= E(f(X^x_t)\delta(a)).
\end{aligned}$$

Here we cannot assert that the strategy is Itô-integrable, since it will be anticipative in general. In order to see that $\mathbb{A}_{t,x,v} \neq \emptyset$, we construct an element, namely,

$$a^i_s := \left\langle J_{0\to s}(x)^{-1}V_i(X^x_{s^-}), (C_t)^{-1}v\right\rangle,$$

where C_t denotes the reduced covariance matrix from Theorem 5. Indeed,

$$\begin{aligned}
&\sum_{i=1}^{d} \left\langle \int_0^t J_{0\to s}(x)^{-1} V_i(X_{s-}^x) a_s^i ds, \xi \right\rangle \\
&= \sum_{i=1}^{d} \int_0^t \left\langle J_{0\to s}(x)^{-1} V_i(X_{s-}^x), \xi \right\rangle \left\langle J_{0\to s}(x)^{-1} V_i(X_{s-}^x), (C_t)^{-1} v \right\rangle ds \\
&= \left\langle \xi, C_t (C_t)^{-1} v \right\rangle = \langle \xi, v \rangle
\end{aligned}$$

for all $\xi \in \mathbb{R}^M$, since C_t is a symmetric random operator defined via

$$\langle \xi, C_t \xi \rangle = \sum_{i=1}^{d} \int_0^t \left\langle J_{0\to s}(x)^{-1} V_i(X_{s-}^x), \xi \right\rangle^2 ds$$

for $\xi \in \mathbb{R}^M$. ∎

For any other derivative with respect to parameters ϵ, we consider a modified set, namely,

$$\mathbb{B}_{t,x,v} = \left\{ b \in \operatorname{dom}(\delta) \,\middle|\, \sum_{i=1}^{d} \int_0^t J_{0\to s}(x)^{-1} V_i(X_{s-}^x) b_s^i ds = J_{0\to t}(x)^{-1} \frac{d}{d\epsilon}\bigg|_{\epsilon=0} X_t^{x,\epsilon} \right\}.$$

Here we are given a parameter-dependent process $X_t^{x,\epsilon}$, where all derivatives with respect to ϵ can be calculated nicely. Also in this case we can construct—if the reduced covariance matrix is invertible and regular enough—an element, namely,

$$b_s^i := \left\langle J_{0\to s}(x)^{-1} V_i(X_{s-}^x), (C_t)^{-1} J_{0\to t}(x)^{-1} \frac{d}{d\epsilon}|_{\epsilon=0} X_t^{x,\epsilon} \right\rangle.$$

This is a consequence of the reasoning

$$\begin{aligned}
&\sum_{i=1}^{d} \left\langle \int_0^t J_{0\to s}(x)^{-1} V_i(X_{s-}^x) b_s^i ds, \xi \right\rangle = \sum_{i=1}^{d} \int_0^t \left\langle J_{0\to s}(x)^{-1} V_i(X_{s-}^x), \xi \right\rangle \cdot \\
&\quad \cdot \left\langle J_{0\to s}(x)^{-1} V_i(X_{s-}^x), (C_t)^{-1} J_{0\to t}(x)^{-1} \frac{d}{d\epsilon}|_{\epsilon=0} X_t^{x,\epsilon} \right\rangle ds \\
&= \left\langle \xi, C_t (C_t)^{-1} J_{0\to t}(x)^{-1} \frac{d}{d\epsilon}|_{\epsilon=0} X_t^{x,\epsilon} \right\rangle = \left\langle \xi, J_{0\to t}(x)^{-1} \frac{d}{d\epsilon}|_{\epsilon=0} X_t^{x,\epsilon} \right\rangle,
\end{aligned}$$

due to the symmetry of C_t.

Appendix.

THEOREM 7. *Let $(\Omega, \mathcal{F}, P, (\mathcal{F}_t)_{t\geq 0})$ be a filtered probability space, and let $(B_t)_{t\geq 0}$ be a d-dimensional Brownian motion adapted to the filtration (which is not necessarily generated by the Brownian motion). Let $V, V_1, \ldots, V_d$, the diffusion vector fields, be C^∞-bounded on $\mathbb{R}^M$, and consider the continuous solution $(X_t^x)_{0\leq t\leq T}$ of an SDE (in Stratonovich notation). V_0 denotes the Stratonovich corrected drift term,*

$$dX_t^x = V_0(X_t^x)dt + \sum_{i=1}^{d} V_i(X_t^x) \circ dB_t^i, \tag{A.1}$$

$$X_0^x = x. \tag{A.2}$$

Assume that the uniform Hörmander condition holds true (see the proof for the precise statement). Then for any $p \geq 1$ there exist numbers $\epsilon_0(p) > 0$ and an integer $K(p) \geq 1$ such that for each $0 < t < T$

$$\sup_{\xi \in S^{M-1}} P(\langle C_t \xi, \xi \rangle < \epsilon) \leq \epsilon^p$$

holds true for $0 \leq \epsilon \leq t^{K(p)} \epsilon_0(p)$. The result holds uniformly in x.

Remark 12. The time-dependence of the estimate $0 \leq \epsilon \leq t^{K(p)} \epsilon_0(p)$ is best explained by redoing the proof. It is heavily applied in section 7 and the main technical ingredient of the given proof. We could have also directly used the results from [21].

Proof. The proof of the theorem is a careful rereading of the Norris lemma and the classical proof of the Hörmander theorem in probability theory (see [22] or [24]). We shall sketch this path in what follows (see [24, pp. 120–123]).

1. Consider the random quadratic form

$$\langle C_t \xi, \xi \rangle = \sum_{i=1}^{d} \int_0^t \left\langle J_{0 \to s}(x)^{-1} V_i(X_s^x), \xi \right\rangle^2 ds.$$

Following [24], we define

$$\Sigma_0' := \{V_1, \ldots, V_d\},$$
$$\Sigma_n' := \left\{ [V_k, V], k = 1, \ldots, d, V \in \Sigma_{n-1}'; [V_0, V] + \frac{1}{2} \sum_{i=1}^{d} [V_i, [V_i, V]], V \in \Sigma_{n-1}' \right\}$$

for $n \geq 1$. We assume that there exist j_0 and $c > 0$ such that

$$\inf_{\xi \in S^{M-1}} \sum_{j=0}^{j_0} \sum_{V \in \Sigma_j'} \langle V(x), \xi \rangle^2 \geq c$$

uniformly in $x \in \mathbb{R}^M$.

2. We define $m(j) := 2^{-4j}$ for $0 \leq j \leq j_0$ and the sets

$$E_j := \left\{ \sum_{V \in \Sigma_j'} \int_0^t \left\langle J_{0 \to s}(x)^{-1} V(X_s^x), \xi \right\rangle^2 ds \leq \epsilon^{m(j)} \right\}.$$

We consider the decomposition

$$E_0 = \{\langle C_t \xi, \xi \rangle \leq \epsilon\} \subset (E_0 \cap E_1^c) \cup (E_1 \cap E_2^c) \cup \cdots \cup (E_{j_0 - 1} \cap E_{j_0}^c) \cup F,$$
$$F = E_0 \cap \cdots \cap E_{j_0}$$

and proceed with

$$P(F) \leq C \epsilon^{\frac{q\beta}{2}}$$

for $\epsilon \leq \epsilon_1$ and any $q \geq 2$ with a constant C depending on q and the norms of the derivatives of the vector fields $V_0, \ldots, V_d$. Furthermore, $0 < \beta < m(j_0)$. The number ϵ_1 is determined by the following two (!) equations:

$$(j_0 + 1) \epsilon_1^{m(j_0)} < \frac{c \epsilon_1^\beta}{4},$$
$$\epsilon_1^\beta < t.$$

Hence ϵ_1 depends on j_0, c, t, and the choice of β via

$$\epsilon_1 < \min\left(t^{\frac{1}{\beta}}, \left(\frac{c}{4(j_0+1)}\right)^{\frac{1}{m(j_0)-\beta}}\right).$$

This little observation, in addition to the proof in [24], is key for our proof.

3. We obtain furthermore that

$$\begin{aligned}
P(E_j \cap E_{j+1}^c) = P\Bigg(& \sum_{V\in\Sigma_j'} \int_0^t \langle J_{0\to s}(x)^{-1} V(X_s^x), \xi\rangle^2 ds \le \epsilon^{m(j)}, \\
& \sum_{V\in\Sigma_{j+1}'} \int_0^t \langle J_{0\to s}(x)^{-1} V(X_s^x), \xi\rangle^2 ds > \epsilon^{m(j+1)} \Bigg) \\
\le \sum_{V\in\Sigma_j'} P\Bigg(& \int_0^t \left\langle J_{0\to s}(x)^{-1} V(X_s^x), \xi\right\rangle^2 ds \le \epsilon^{m(j)}, \\
& \sum_{k=1}^d \int_0^t \left\langle J_{0\to s}(x)^{-1} [V_k, V](X_s^x), \xi\right\rangle^2 ds \\
& + \int_0^t \left\langle J_{0\to s}(x)^{-1} \left([V_0, V] + \frac{1}{2}\sum_{i=1}^d [V_i, [V_i, V]]\right)(X_s^x), \xi \right\rangle^2 ds > \frac{\epsilon^{m(j+1)}}{n(j)} \Bigg),
\end{aligned}$$

where $n(j) = \#\Sigma_j'$. Since we can find the bounded variation and the quadratic variation parts of the martingale $(\langle J_{0\to s}(x)^{-1} V(X_s^x), \xi\rangle)_{0\le s\le t}$ in the above expression, we are able to apply the Norris lemma (see [24, Lemma 2.3.2]). We observe that $8m(j+1) < m(j)$; hence we can apply it with $q = \frac{m(j)}{m(j+1)}$.

4. We obtain for $p \ge 2$—still by the Norris lemma—the estimate

$$P(E_j \cap E_{j+1}^c) \le d_1 \left(\frac{\epsilon^{m(j+1)}}{n(j)}\right)^{rp} + d_2 \exp\left(-\left(\frac{\epsilon^{m(j+1)}}{n(j)}\right)^{-\nu}\right)$$

for $\epsilon \le \epsilon_2$. Furthermore, $r, \nu > 0$ with $18r + 9\nu < q - 8$, and the numbers d_1, d_2 depend on the vector fields $V_0, \dots, V_d$, and on p, T. The number ϵ_2 can be chosen as $\epsilon_2 = \epsilon_3 t^{k_1}$, where ϵ_3 does not depend on t anymore.

5. Putting all this together, we take the minimum of ϵ_1 and ϵ_2 to obtain the desired dependence on t. □

Acknowledgment. The authors are grateful to two unknown referees who helped to considerably improve the presentation and contents of this paper. In particular, we are grateful that they pointed out Corollary 3.25 of [21].

REFERENCES

[1] Y. Bakhtin and J. Mattingly, *Malliavin calculus for infinite-dimensional systems with additive noise*, J. Funct. Anal., 249 (2007), pp. 307–353.

[2] V. Bally, M.-P. Bavouzet, and M. Messaoud, *Integration by parts formula for locally smooth laws and applications to sensitivity computations*, Ann. Appl. Probab., 17 (2007), pp. 33–66.

[3] F. Baudoin and J. Teichmann, *Hypoellipticity in infinite dimensions and an application in interest rate theory*, Ann. Appl. Probab., 15 (2005), pp. 1765–1777.

[4] Y. I. Belopol'skaya and Y. L. Dalecky, *Stochastic Equations and Differential Geometry*, Mathematics and Its Applications (Soviet Series) 30, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.

[5] Y. I. Belopol'skaya and Y. L. Daletsky, *Smoothness of transition probabilities of Markov processes described by stochastic differential equations in Hilbert space*, Dopov. / Dokl. Akad. Nauk Ukraïni, 9 (1994), pp. 40–45 (in Russian).

[6] K. Bichteler, J.-B. Gravereaux, and J. Jacod, *Malliavin Calculus for Processes with Jumps*, Stochastics Monographs 2, Gordon and Breach Science Publishers, New York, 1987.

[7] J.-M. Bismut, *Calcul des variations stochastique et processus de sauts*, Z. Wahrsch. Verw. Gebiete, 63 (1983), pp. 147–235.

[8] T. Björk and A. Gombani, *Minimal realizations of interest rate models*, Finance Stoch., 3 (1999), pp. 413–432.

[9] T. Björk and L. Svensson, *On the existence of finite-dimensional realizations for nonlinear forward rate models*, Math. Finance, 11 (2001), pp. 205–243.

[10] R. A. Carmona and B. Rozovskii, eds., *Stochastic Partial Differential Equations: Six Perspectives*, Math. Surveys Monogr. 64, AMS, Providence, RI, 1999.

[11] T. Cass, *Smooth densities for stochastic differential equations with jumps*, Stochastic Process Appl., to appear.

[12] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Encyclopedia Math. Appl. 44, Cambridge University Press, Cambridge, UK, 1992.

[13] M. H. A. Davis and M. P. Johansson, *Malliavin Monte Carlo Greeks for jump diffusions*, Stochastic Process. Appl., 116 (2006), pp. 101–129.

[14] Y. El-Khatib and N. Privault, *Computations of Greeks in a market with jumps via the Malliavin calculus*, Finance Stoch., 8 (2004), pp. 161–179.

[15] D. Filipovic and S. Tappe, *Existence of Lévy term structure models*, Finance Stoch., 12 (2008), pp. 83–115.

[16] D. Filipović and J. Teichmann, *Existence of invariant manifolds for stochastic equations in infinite dimension*, J. Funct. Anal., 197 (2003), pp. 398–432.

[17] E. Fournié, J.-M. Lasry, J. Lebuchoux, P.-L. Lions, and N. Touzi, *Applications of Malliavin calculus to Monte Carlo methods in finance*, Finance Stoch., 3 (1999), pp. 391–412.

[18] E. Gobet and R. Munos, *Sensitivity analysis using Itô–Malliavin calculus and martingales, and application to stochastic optimal control*, SIAM J. Control Optim., 43 (2005), pp. 1676–1713.

[19] R. S. Hamilton, *The inverse function theorem of Nash and Moser*, Bull. Amer. Math. Soc. (N.S.), 7 (1982), pp. 65–222.

[20] A. Kriegl and P. W. Michor, *The Convenient Setting of Global Analysis*, Math. Surveys Monogr. 53, AMS, Providence, RI, 1997.

[21] S. Kusuoka and D. Stroock, *Applications of the Malliavin calculus*. II, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 32 (1985), pp. 1–76.

[22] P. Malliavin, *Stochastic Analysis*, Grundlehren Math. Wiss. 313, Springer-Verlag, Berlin, 1997.

[23] P. Malliavin, M. E. Mancino, and M. C. Recchioni, *A non-parametric calibration of the HJM geometry: An application of Itô calculus to financial statistics*, Jpn. J. Math., 2 (2007), pp. 55–77.

[24] D. Nualart, *The Malliavin Calculus and Related Topics*, Probab. Appl. (New York), Springer-Verlag, New York, 1995.

[25] P. Protter, *Stochastic Integration and Differential Equations. A New Approach*, Appl. Math. (New York), Springer-Verlag, Berlin, 1990.

[26] J. Teichmann, *Another approach to some rough and stochastic partial differential equations*, Preprint, 2008.

[27] X. Y. Zhou, *On the existence of solutions with smooth density of stochastic differential equations in plane*, Acta Math. Sinica (N.S.), 8 (1992), pp. 432–446.

SIAM J. MATH. ANAL.
Vol. 40, No. 6, pp. 2155–2178

BILLIARD SCATTERING ON ROUGH SETS: TWO-DIMENSIONAL CASE*

ALEXANDER PLAKHOV†

Abstract. The notion of a rough two-dimensional (convex) body is introduced, and to each rough body there is assigned a measure on $\mathbb{T}^3$ describing billiard scattering on the body. The main result is characterization of the set of measures generated by rough bodies. This result can be used to solve various problems of least aerodynamical resistance.

Key words. billiards, scattering on rough surfaces, Monge–Kantorovich optimal mass transportation, problems of minimal and maximal resistance, shape optimization

AMS subject classifications. 37N05, 49K30, 49Q10

DOI. 10.1137/070709700

1. Definition of a rough set and statement of main theorem.

1.1. Introductory remarks and review of literature. In this paper the notion of a rough two-dimensional (convex) body is given and some properties of rough bodies are established.

Let $B \subset \mathbb{R}^2$ be a convex bounded set with nonempty interior, that is, a bounded convex body. Consider the "set" obtained from B by moving off a set of "very small" area. Such a (heuristically defined) set is called a *rough body*: from the "macroscopic" point of view, it almost coincides with B, and, from the "microscopic" point of view, it contains some "flaws." (One can imagine a detail of a mechanism that, after a period of exploitation, has some defects.) If the removed set adjoins the boundary ∂B, then one can expect that a flow of point particles incident on the rough body is reflected in another way as compared to reflection from B.

The notion of rough body arises naturally when studying Newton-like problems of the body of least resistance. The first problem of such kind was considered by Newton himself [1]. Recently there were several works made concerning the problem of least resistance in various classes of admissible bodies; see, e.g., [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [15]. The solution of a minimization problem for the case of *rotating bodies* can be naturally identified with a rough body [14]; see also the concluding remarks to this paper.

There are many papers on particle scattering by rough bodies (see, e.g., [16], [17], [18]); they describe bodies and flows of particles that occur in nature. On the contrary, we assume that a rough structure can be "manufactured," and our aim is to describe all possible rough structures.

1.2. Definition of a rough body. It is supposed that the "microscopic structure" of the boundary of a rough body can be detected from observations of particle

*Received by the editors November 30, 2007; accepted for publication (in revised form) July 16, 2008; published electronically January 28, 2009. This work was supported by the *Centre for Research on Optimization and Control* (CEOC) from the "*Fundação para a Ciência e a Tecnologia*" (FCT), cofinanced by the European Community Fund FEDER/POCTI, and by FCT research project PTDC/MAT/72840/2006.

http://www.siam.org/journals/sima/40-6/70970.html

†Department of Mathematics, Aveiro University, Aveiro 3810, Portugal. Current address: Institute of Mathematical and Physical Sciences, Aberystwyth University, Aberystwyth SY23 3BZ, UK (axp@aber.ac.uk).

scattering on the body. From this point of view, two rough bodies are considered equal if they scatter flows of particles in an identical manner. Having these observations in mind, we give the definition of a rough body.

Let B be a bounded convex body. Denote by $n(\xi)$ the unit outer normal vector to ∂B at a regular point $\xi \in \partial B$, and denote by $(\partial B \times S^1)_+$ the set of pairs $(\xi, v) \in \partial B \times S^1$ such that $\langle n(\xi), v\rangle \geq 0$. Here and in what follows, $\langle \cdot, \cdot \rangle$ means the standard scalar product in $\mathbb{R}^2$. The set $(\partial B \times S^1)_+$ is equipped with the measure μ which is defined by $d\mu(\xi, v) = \langle n(\xi), v\rangle \, d\xi \, dv$, where $d\xi$ and dv are the one-dimensional Lebesgue measures on ∂B and S^1, respectively.

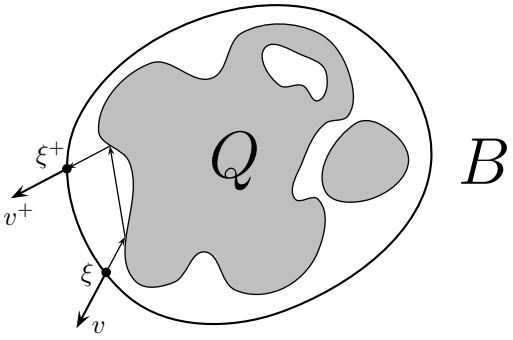


FIG. 1. *A billiard trajectory in* $\mathbb{R}^2 \setminus Q$.

Let Q be a set with piecewise smooth boundary contained in B; consider the billiard in $\mathbb{R}^2 \setminus Q$. Note that Q is not necessarily connected. For $(\xi, v) \in (\partial B \times S^1)_+$, consider a billiard particle starting at the point ξ with the velocity $-v$. After several (maybe none) reflections from $\partial Q \setminus \partial B$, the particle will intersect ∂B again, at a point $\xi^+ = \xi^+_{Q,B}(\xi, v) \in \partial B$; denote by $v^+ = v^+_{Q,B}(\xi, v)$ the velocity at this point (see Figure 1). It may happen that the initial point ξ belongs to ∂Q; in that case we have $\xi^+ = \xi$ and the vector v^+ is symmetric to v with respect to $n(\xi)$. It may also happen that at some moment the particle either gets into a singular point of ∂Q, or touches ∂Q at a regular point, or stays in $B \setminus Q$ forever and does not intersect ∂B again, or makes an infinite number of reflections in finite time. The set of corresponding points (ξ, v) has zero measure, and the corresponding values $\xi^+_{Q,B}(\xi, v)$ and $v^+_{Q,B}(\xi, v)$ are not defined.

Thus, there is defined the one-to-one mapping $T_{Q,B} : (\xi, v) \mapsto (\xi^+_{Q,B}(\xi, v), v^+_{Q,B}(\xi, v))$ of a full measure subset of $(\partial B \times S^1)_+$ onto itself. It has the following properties:

T1. $T_{Q,B}$ preserves the measure μ.
T2. $T^{-1}_{Q,B} = T_{Q,B}$.

The mapping $T_{Q,B}$ induces the measure $\nu_{Q,B}$ on $\mathbb{T}^3 = S^1 \times S^1 \times S^1$ in the following way. Let $A \subset \mathbb{T}^3$ be a Borel set; by definition,

$$\nu_{Q,B}(A) = \mu\left(\{(\xi, v) \in (\partial B \times S^1)_+ : (v, v^+_{Q,B}(\xi, v), n(\xi)) \in A\}\right).$$

In fact, the measure $\nu_{Q,B}$ contains information about particle scattering on Q. Imagine that an observer has no means to track the trajectory of particles inside B. Instead, for each incident particle there is registered the triple of vectors: the initial and final

velocities (measured at the points of first and second intersection with ∂B), and the normal vector to ∂B at the point of first intersection with ∂B. The normal vector at the second point of intersection is not registered; as will be seen later (Lemma 1), if the area of $B \setminus Q$ is small, then the difference between the normal vectors at these two points is also small. The measure $\nu_{Q,B}$ describes the distribution of triples.

DEFINITION 1. *We say that a sequence of sets* $\{Q_m,\ m = 1, 2, \ldots\}$ *represents a* rough body *if it has the following properties:*

M1. $Q_m \subset B$ *and* $\operatorname{Area}(B \setminus Q_m) \to 0$ *as* $m \to \infty$.

M2. *The sequence of measures* $\nu_{Q_m,B}$ *weakly converges.*

Two sequences of such sets are called equivalent if the corresponding limiting measures coincide. An equivalence class is called a body obtained by roughening B*, or simply* rough body*, and denoted by* $\mathcal{B}$*, and the corresponding limiting measure is denoted by* $\nu_{\mathcal{B}}$.

Note that the sets Q_m in this definition are not necessarily connected.

Remark. Since $\mathbb{T}^3$ is compact and the full measure of $\mathbb{T}^3$ satisfies $\nu_{Q,B}(\mathbb{T}^3) \le 2\pi\,|\partial B|$, one concludes that the set of measures $\{\nu_{Q,B}\}$, with fixed B, is weakly precompact. That is, any sequence of measures $\{\nu_{Q_m,B}\}$ contains a weakly converging subsequence. In this sense one can say that a sequence, satisfying only condition M1, can represent more than one rough body.

We would also like to mention that, first, two rough bodies obtained one from another by translation are identified, according to our definition. Second, particle scattering on $\mathcal{B}$ in a small neighborhood of $\xi \in \partial B$ can be detected if ξ is an extreme point of B, and cannot otherwise. Indeed, if ξ is an extreme point of B, then the scattering is described by the restriction of $\nu_{\mathcal{B}}$ on $\mathbb{T}^2 \times \mathcal{N}_{n(\xi)}$, with $\mathcal{N}_{n(\xi)}$ being a small neighborhood of $n(\xi)$ in S^1. If, otherwise, ξ is not an extreme point of B, that is, belongs to an open linear segment contained in ∂B, the scattering can be determined only on the whole segment.

Actually, from the viewpoint of applications to the problems of optimal resistance in *homogeneous* and *rarefied* media (see section 4, containing concluding remarks and applications), these drawbacks are not so serious. Indeed, resistance of a body is invariant under translations (due to homogeneity). Besides, if the boundary of a body contains a linear segment, then one does not need to know scattering at each point of the segment; it suffices to know it on the whole segment (due to homogeneity and rarefaction).

The definition of a rough body could be made in a slightly different way, basing it on measures defined on $S^1 \times S^1 \times \partial B$. In that case the triple (v, v^+, ξ) should be registered, with ξ being the point of first intersection with ∂B. That definition would allow one to register particle scattering at each point of ∂B and to distinguish between bodies obtained by translation one from another. However, we prefer to adopt the former definition, since it seems to us mathematically more transparent and makes the arguments a bit easier.

1.3. Examples. Sometimes it is convenient to use another representation of the measure $\nu_{\mathcal{B}}$. Namely, consider the change of coordinates $(v, v^+, n) \mapsto (\varphi, \varphi^+, n)$, where $\varphi = \operatorname{Arg} v - \operatorname{Arg} n$, $\varphi^+ = \operatorname{Arg} v^+ - \operatorname{Arg} n$. Here $\operatorname{Arg} v$ is the angle between a fixed vector and v measured, say, clockwise from this vector to v. If $(v, v^+, n) \in \operatorname{spt} \nu_{\mathcal{B}}$, then φ and φ^+ belong to $[-\pi/2, \pi/2]$ modulo 2π. Introduce the shorthand notation $\square := [-\pi/2, \pi/2] \times [-\pi/2, \pi/2]$ and define the mapping $\varpi : \square \times S^1 \to \mathbb{T}^3$ by $\varpi(\varphi, \varphi^+, n) = (v, v^+, n)$. One has $\operatorname{spt} \nu_{\mathcal{B}} \subset \varpi(\square \times S^1)$. Denote $\breve{\nu}_{\mathcal{B}} := (\varpi^{-1})^{\#} \nu_{\mathcal{B}}$. Sometimes this measure can be factorized: $\breve{\nu}_{\mathcal{B}} = \eta_{\mathcal{B}} \otimes \tau_B$, where $\eta_{\mathcal{B}}$ is defined on $\square$

and τ_B is the surface measure on B; so to say, the "roughness" is "homogeneous" along the body's boundary. Consider several examples.

Example 1 ("smooth body"). The rough body represented by the sequence $Q_m = B$ is identified with B itself. The corresponding measure is $\breve{\nu}_B = \eta_0 \otimes \tau_B$, where the measure η_0 has the density $\cos\varphi \cdot \delta(\varphi + \varphi^+)$; the support of η_0 is shown in Figure 2(b). In Figure 2(a), B is taken to be an ellipse.

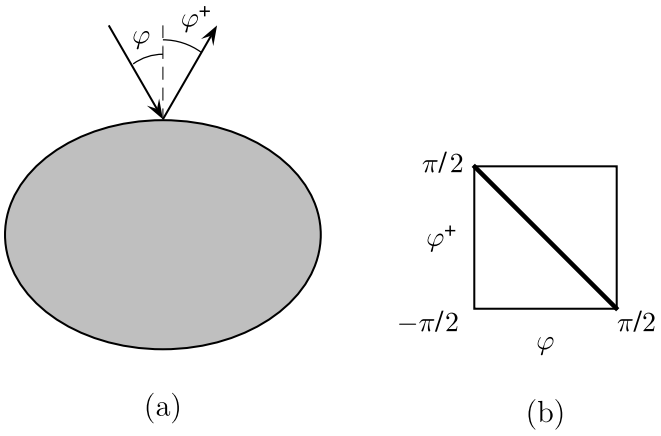


FIG. 2. *A smooth body* (a) *and the support of the corresponding measure* η_0 (b).

Example 2 (roughness formed by triangular hollows). Q_m is a $2m$-polygon; the 270° angles alternate with the angles that are slightly smaller than 90°. All vertices corresponding to the angles smaller than 90° belong to ∂B. Any two sides that form a 270° angle are equal. The largest side length tends to zero as $m \to \infty$. Thus, the set Q_m is obtained by moving off m "hollows" from its convex hull, each of the hollows being an isosceles right triangle.

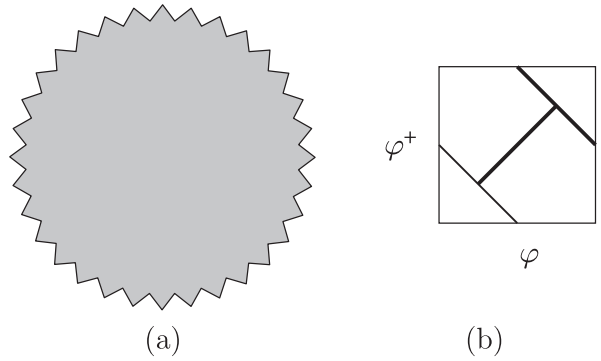
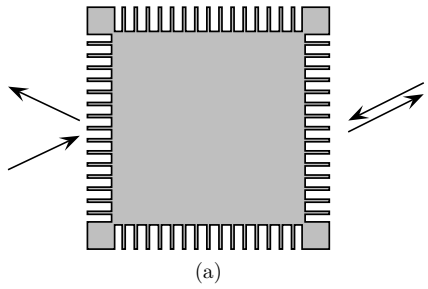


FIG. 3. *A rough body with hollows being isosceles right triangles* (a) *and the support of the corresponding measure* $\eta_\triangledown$ (b).

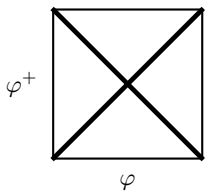
The corresponding measure is $\breve{\nu}_{\mathcal{B}} = \eta_\triangledown \otimes \tau_B$, where the measure $\eta_\triangledown$ has the density $\cos\varphi \cdot [\chi_{[-\pi/2,-\pi/4]}(\varphi)\,\delta(\varphi+\varphi^+ + \frac{\pi}{2}) + \chi_{[-\pi/4,\pi/4]}(\varphi)\,\delta(\varphi-\varphi^+) + \chi_{[\pi/4,\pi/2]}(\varphi)\,\delta(\varphi+\varphi^+ - \frac{\pi}{2})] + |\sin\varphi| \cdot [\chi_{[-\pi/4,0]}(\varphi)\,\delta(\varphi+\varphi^+ + \frac{\pi}{2}) - \chi_{[-\pi/4,\pi/4]}(\varphi)\,\delta(\varphi-\varphi^+) + \chi_{[0,\pi/4]}(\varphi)\,\delta(\varphi+\varphi^+ - \frac{\pi}{2})]$. Thus, the support of $\eta_\triangledown$ is the union of three segments; see Figure 3(b). The

middle segment $\varphi^+ = \varphi$ corresponds to double reflections, and the lateral segments, $\varphi^+ = -\varphi - \pi/2$ and $\varphi^+ = -\varphi + \pi/2$, correspond to single reflections, from the right or from the left side of a triangular hollow. In Figure 3(a), B is a circle.

Example 3 (roughness formed by rectangular hollows). The sets Q_m are obtained by removing a finite number of "rectangular hollows" from B. In other words, one has $Q_m = B \setminus (\cup_n \Omega_{m,n})$, where the removed sets $\Omega_{m,n}$ do not mutually intersect and each set $\partial\Omega_{m,n} \setminus \partial B$ is the union of three sides of a rectangle. The ratio (width)/(depth) of a hollow depends only on m and is denoted by h_m. Denote by $l_m = |\partial B \setminus \cup_n(\partial\Omega_{m,n})|/|\partial B|$ the relative length of the part of boundary ∂B not covered by hollows. We assume that $\lim_{m\to\infty} h_m = 0 = \lim_{m\to\infty} l_m$. In Figure 4(a), B is a square.



(a)



(b)

FIG. 4. *A rough body where hollows are "thin rectangles"* (a)*; the support of the corresponding measure* $\eta_\square$ (b).

The measure $\breve{\nu}_{\mathcal{B}}$ equals $\breve{\nu}_{\mathcal{B}} = \eta_\square \otimes \tau_B$. The density of the measure $\eta_\square$ equals $\frac{1}{2}\cos\varphi \cdot (\delta(\varphi + \varphi^+) + \delta(\varphi - \varphi^+))$, and the support is the union of two diagonals, $\varphi^+ = \varphi$ and $\varphi^+ = -\varphi$; see Figure 4(b). The particles with an even (odd) number of reflections contribute to the first (second) diagonal.

1.4. Main theorem. According to Definition 1, each rough body is identified with a measure on $\mathbb{T}^3$. The question is, What is the set of these measures? The following definition and theorem give the answer.

Let us first introduce some notation: $\pi_{v,n} : \mathbb{T}^3 \to \mathbb{T}^2$, $\pi_n : \mathbb{T}^3 \to S^1$, etc., are projections onto the corresponding subspaces: $\pi_{v,n}(v, v^+, n) = (v, n)$, $\pi_n(v, v^+, n) = n$, etc.; $\pi_d : \mathbb{T}^3 \to \mathbb{T}^3$ is the symmetry with respect to the plane $v = v^+$, that is, $\pi_d(v, v^+, n) = (v^+, v, n)$; $z_+ = \max\{0, z\}$ is the positive part of $z \in \mathbb{R}$; and u means Lebesgue measure on S^1. Recall that τ_B is the surface measure on B and is defined on S^1.

DEFINITION 2. *We denote by* $\mathcal{M}_B$ *the set of measures* ν *on* $\mathbb{T}^3$ *such that the following properties hold:*

A1. *The marginal measures* $\pi^{\#}_{v,n}\nu$ *and* $\pi^{\#}_{v^+,n}\nu$ *are*

$$\pi^{\#}_{v,n}\nu = \langle v, n\rangle_+ \cdot u \otimes \tau_B, \qquad \pi^{\#}_{v^+,n}\nu = \langle v^+, n\rangle_+ \cdot u \otimes \tau_B.$$

A2. $\pi^{\#}_d \nu = \nu$.

Denote also $\mathcal{M} = \cup_B \mathcal{M}_B$*, the union being taken over all bounded convex bodies* B.

Taking into account the Alexandrov theorem on characterization of surface measures, one concludes that $\mathcal{M}$ is the set of measures ν on $\mathbb{T}^3$ such that

(1) the marginal measure $\pi^{\#}_{n}\nu =: \tau$ satisfies the conditions
  1a. $\int_{S^1} n\, d\tau(n) = 0$;
  1b. for any $v \in S^1$ it holds that $\int_{S^1} \langle n, v\rangle^2\, d\tau(n) \neq 0$;
(2) the marginal measures $\pi^{\#}_{v,n}\nu$ and $\pi^{\#}_{v^+,n}\nu$ satisfy the conditions
  2a. $\pi^{\#}_{v,n}\nu = \langle v, n\rangle_+ \cdot u \otimes \tau$;
  2b. $\pi^{\#}_{v^+,n}\nu = \langle v^+, n\rangle_+ \cdot u \otimes \tau$.

Thus, these marginal measures coincide; the only difference is in the notation for the variables: v, n in case 2a and v^+, n in case 2b.

Now we can state the main theorem.

THEOREM. *The set of measures $\{\nu_{\mathcal{B}}\}$, with $\mathcal{B}$ being all possible bodies obtained by roughening B, coincides with $\mathcal{M}_B$. Therefore, $\{\nu_{\mathcal{B}}$, $\mathcal{B}$ is a rough body$\} = \mathcal{M}$.*

In section 2, we formulate two auxiliary lemmas and using them prove the theorem. In section 3, the lemmas are proved. Section 4 contains concluding remarks and applications of the theorem to problems of optimal aerodynamic resistance. Appendices A and B contain proofs of some auxiliary technical results.

2. Statement of auxiliary lemmas and proof of theorem.

2.1. Statement of Lemma 1. Fix a bounded convex body B. Two points $\xi_1, \xi_2 \in \partial B$, $\xi_1 \neq \xi_2$, divide the curve ∂B into two arcs. Denote by $l(\xi_1, \xi_2)$ the length of the smallest arc and denote

$$c = c_B := \inf_{\substack{\xi_1,\xi_2\in\partial B\\ \xi_1\neq\xi_2}} \frac{|\xi_1 - \xi_2|}{l(\xi_1, \xi_2)};$$

one obviously has $0 < c < 1$.

Let $Q \subset B$; denote

$$\overline{|\xi - \xi^+|}_{Q,B} := \iint_{(\partial B\times S^1)_+} |\xi - \xi^+_{Q,B}(\xi, v)|\, d\mu(\xi, v)$$

and

$$\overline{|n - n^+|}_{Q,B} := \iint_{(\partial B\times S^1)_+} |n(\xi) - n(\xi^+_{Q,B}(\xi, v))|\, d\mu(\xi, v).$$

LEMMA 1. (a) *The following holds true:*

$$\overline{|\xi - \xi^+|}_{Q,B} \leq 2\pi \cdot \mathrm{Area}(B \setminus Q).$$

(b) *For sufficiently small* $\mathrm{Area}(B \setminus Q)$,[1] *one has*

$$\overline{|n - n^+|}_{Q,B} \leq \frac{2\pi\sqrt{8\pi}}{\sqrt{c}} \sqrt{\mathrm{Area}(B \setminus Q)}.$$

2.2. Statement of Lemma 2. Let us first introduce the notion of a hollow.

DEFINITION 3. *Let $\Omega \subset \mathbb{R}^2$ be a closed bounded set with piecewise smooth boundary and $I \subset \partial\Omega$, where the following hold:*

(i) *I is an interval contained in a straight line $\langle x, n\rangle = a$.*

(ii) *$\Omega \setminus I$ is contained in the open half-plane $\langle x, n\rangle < a$. Here n is a fixed unit vector.*

Then the pair (Ω, I) is called a hollow oriented by n, or just an n-hollow.

[1] That is, it is smaller than a positive value depending only on B.

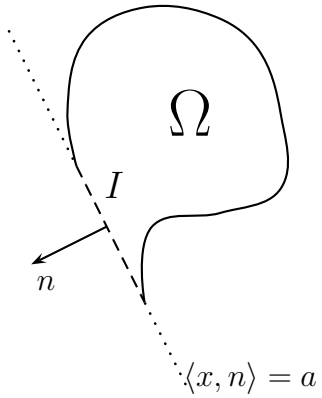


FIG. 5. *A hollow.*

In Figure 5 and in what follows, I is shown by a dashed line, and $\partial\Omega \setminus I$ is shown by a solid line.

Define the measure $\tilde{\mu}_I$ on $I \times S^1$ by $d\tilde{\mu}_I(\xi, v) = \frac{\langle n,v\rangle_+}{|I|}\, d\xi\, dv$, where $|I|$ means the length of I. Obviously, $\tilde{\mu}_I$ is supported on the set $(I \times S^1)_+ := \{(\xi, v) \in I \times S^1 : \langle n, v\rangle \geq 0\}$. Define the one-to-one mapping $(\xi, v) \mapsto (\Xi^+_{\Omega,I}(\xi, v), V^+_{\Omega,I}(\xi, v))$ of a full measure subset of $(I \times S^1)_+$ onto itself. Namely, consider the billiard in Ω. Let $(\xi, v) \in (I \times S^1)_+$; consider the billiard particle starting at the point ξ with the velocity $-v$. It makes several reflections from $\partial\Omega \setminus I$ and then reflects from I again, at a point $\Xi^+ = \Xi^+_{\Omega,I}(\xi, v)$. The velocity immediately before this reflection is denoted by $V^+ = V^+_{\Omega,I}(\xi, v)$. The mapping so defined preserves the measure $\tilde{\mu}_I$ and is an involution, that is, coincides with its inverse.

One can give an equivalent definition based on the mapping $\xi^+_{Q,B}(\xi, v), v^+_{Q,B}(\xi, v)$ just defined in section 1.2. Take a set Q such that Ω is a connected component of $\operatorname{conv} Q \setminus Q$ and I is a connected component of $\partial(\operatorname{conv} Q) \setminus \partial Q$. For $(\xi, v) \in (I \times S^1)_+$, let by definition $(\Xi^+_{\Omega,I}(\xi, v), V^+_{\Omega,I}(\xi, v)) := (\xi^+_{Q,\operatorname{conv} Q}(\xi, v), v^+_{Q,\operatorname{conv} Q}(\xi, v))$. This definition does not depend on the choice of Q.

DEFINITION 4. *Let (Ω, I) be a hollow. The measure $\eta_{\Omega,I}$ on $\mathbb{T}^2 = S^1 \times S^1$ is defined as follows. For a Borel set $A \subset \mathbb{T}^2$, put*

$$\eta_{\Omega,I}(A) := \tilde{\mu}_I(\{(\xi, v) \in (I \times S^1)_+ : (v, V^+_{\Omega,I}(\xi, v)) \in A\}).$$

We shall say that $\eta_{\Omega,I}$ is the measure generated by the hollow (Ω, I).

Here we use the notation $\pi_v, \pi_{v^+} : \mathbb{T}^2 \to S^1$ for the projections onto the subspaces $\{v\}$ and $\{v^+\}$, respectively; $\pi_v(v, v^+) = v$, $\pi_{v^+}(v, v^+) = v^+$. We also denote by π_d the symmetry with respect to the diagonal $v = v^+$; $\pi_d(v, v^+) = (v^+, v)$.

DEFINITION 5. *Denote by Λ_n the set of measures η on $\mathbb{T}^2$ such that*

(i) $d\pi^{\#}_v \eta(v) = \langle v, n\rangle_+\, dv$, $d\pi^{\#}_{v^+}\eta(v^+) = \langle v^+, n\rangle_+\, dv^+$;

(ii) $\pi^{\#}_d \eta = \eta$.

Any measure $\eta_{\Omega,I}$ generated by an n-hollow belongs to Λ_n. Indeed, for any $A \subset S^1$ one has $\pi^{\#}_v \eta_{\Omega,I}(A) = \eta_{\Omega,I}(A \times S^1) = \tilde{\mu}_I(\{(\xi, v) \in (I \times S^1)_+ : v \in A\}) = \frac{1}{|I|} \iint_{I\times A} \langle n, v\rangle_+\, d\xi\, dv = \int_A \langle n, v\rangle_+\, dv$. This proves the first equality in (i).

Similarly, one has $\pi^{\#}_{v^+}\eta_{\Omega,I}(A) = \eta_{\Omega,I}(S^1 \times A) = \tilde{\mu}_I(\{(\xi, v) \in (I \times S^1)_+ : V^+_{\Omega,I}(\xi, v) \in A\})$. Since the mapping $(\xi, v) \mapsto (\Xi^+_{\Omega,I}, V^+_{\Omega,I})$ preserves the measure, one gets the

value $\tilde{\mu}_I(\{(\xi,v)\in(I\times S^1)_+ : v\in A\})$, which in turns equals $\int_A \langle n,v\rangle_+\,dv$. This proves the second equality in (i). Finally, the relation (ii) for $\eta_{\Omega,I}$ is a simple consequence of involutive and measure preserving properties of the mapping $(\xi,v)\mapsto(\Xi^+_{\Omega,I},V^+_{\Omega,I})$.

LEMMA 2. *The set of measures generated by n-hollows is weakly dense in* Λ_n.

2.3. Proof of the direct statement of theorem. Here we prove that for any body $\mathcal{B}$ obtained by roughening B it holds that $\nu_{\mathcal{B}}\in\mathcal{M}_B$.

Let $Q\subset B$; define the measure $\nu'_{Q,B}$ on $\mathbb{T}^3$ by

$$\nu'_{Q,B}(A):=\mu\left(\{(\xi,v)\in(\partial B\times S^1)_+ : (v,v^+_{Q,B}(\xi,v),n(\xi^+_{Q,B}(\xi,v)))\in A\}\right),$$

where A is an arbitrary Borel subset of $\mathbb{T}^3$. Thus, the definition of both $\nu_{Q,B}$ and $\nu'_{Q,B}$ is based on observations of vector triples (v,v^+,n) and (v,v^+,n^+), respectively. Here n and n^+ are the outer normals to ∂B at the points where the particle *gets in* B and *gets out* of B. The measures $\nu_{Q,B}$ and $\nu'_{Q,B}$ have the following properties:

$$\pi^{\#}_{v,n}\nu_{Q,B}=\langle v,n\rangle_+\cdot u\otimes\tau_B, \tag{1}$$

$$\pi^{\#}_{v^+,n^+}\nu'_{Q,B}=\langle v^+,n^+\rangle_+\cdot u\otimes\tau_B, \tag{2}$$

$$\pi^{\#}_{d}\nu_{Q,B}=\nu'_{Q,B}. \tag{3}$$

Consider a sequence $\{Q_m\}$ representing $\mathcal{B}$; let us show that $\nu_{Q_m,B}-\nu'_{Q_m,B}$ weakly converges to zero as $m\to\infty$. It is enough to prove that for any continuous function f on $\mathbb{T}^3$ it holds that

$$\int_{\mathbb{T}^3} f(v,v^+,n)\,d\nu_{Q_m,B}(v,v^+,n)-\int_{\mathbb{T}^3} f(v,v^+,n^+)\,d\nu'_{Q_m,B}(v,v^+,n^+)\to_{m\to\infty}0. \tag{4}$$

Taking into account the formulas for a change of variables

$$\int_{\mathbb{T}^3} f(v,v^+,n)\,d\nu_{Q,B}(v,v^+,n)=\int_{(\partial B\times S^1)_+} f(v,v^+_{Q,B}(\xi,v),n(\xi))\,d\mu(\xi,v)$$

and

$$\int_{\mathbb{T}^3} f(v,v^+,n^+)\,d\nu'_{Q,B}(v,v^+,n^+)=\int_{(\partial B\times S^1)_+} f(v,v^+_{Q,B}(\xi,v),n(\xi^+_{Q,B}(\xi,v)))\,d\mu(\xi,v),$$

formula (4) takes the form

$$\lim_{m\to\infty}\int_{(\partial B\times S^1)_+}\left[f(v,v^+_{Q_m,B}(\xi,v),n(\xi^+_{Q_m,B}(\xi,v)))-f(v,v^+_{Q_m,B}(\xi,v),n(\xi))\right]d\mu(\xi,v)=0. \tag{5}$$

According to Lemma 1, the difference $n(\xi^+_{Q_m,B}(\xi,v))-n(\xi)$ converges to zero in mean, and hence it converges to zero in measure; therefore the difference

$$f(v,v^+_{Q_m,B}(\xi,v),n(\xi^+_{Q_m,B}(\xi,v)))-f(v,v^+_{Q_m,B}(\xi,v),n(\xi))$$

also converges to zero in measure. It follows that formula (5) is true.

Thus, both $\nu_{Q_m,B}$ and $\nu'_{Q_m,B}$ weakly converge to $\nu_{\mathcal{B}}$. Substituting $Q = Q_m$ into formulas (1)–(3) and passing to limit as $m \to \infty$, one gets

$$\pi^{\#}_{v,n}\nu_{\mathcal{B}} = \langle v, n\rangle_+ \cdot u \otimes \tau_B,$$
$$\pi^{\#}_{v^+,n}\nu_{\mathcal{B}} = \langle v^+, n\rangle_+ \cdot u \otimes \tau_B,$$
$$\pi^{\#}_{d}\nu_{\mathcal{B}} = \nu_{\mathcal{B}},$$

that is, $\nu_{\mathcal{B}} \in \mathcal{M}_B$.

2.4. Proof of the inverse statement of theorem. Here it is proved that for any $\nu \in \mathcal{M}_B$ there exists a body $\mathcal{B}$ obtained by roughening B such that $\nu_{\mathcal{B}} = \nu$. The proof is based on two statements.

STATEMENT 1. *Let B be a convex polygon. Then for any measure $\nu \in \mathcal{M}_B$ there exists a body $\mathcal{B}$ obtained by roughening B such that $\nu_{\mathcal{B}} = \nu$.*

Proof. Let us enumerate the sides of the polygon B and denote by c_i the length of the ith side and by n_i the outer unit normal to this side. By δ_n, denote the probabilistic atomic measure on S^1 concentrated at $n \in S^1$, that is, $\delta_n(n) = 1$. The surface measure of B is $\tau_B = \sum c_i\delta_{n_i}$; this implies that any measure $\nu \in \mathcal{M}_B$ has the form $\nu = \sum c_i\eta_i \otimes \delta_{n_i}$, where $\eta_i \in \Lambda_{n_i}$.

According to Lemma 2, any measure η_i is the weak limit as $m \to \infty$ of measures $\eta_{\Omega_i^m, I_i^m}$ generated by a sequence of n_i-hollows (Ω_i^m, I_i^m). Now take a sequence of sets Q_m such that $\operatorname{conv} Q_m = B$ and each connected component of $B\setminus Q_m$ is the image of a set Ω_i^m under the composition of a homothety with positive ratio and a translation; additionally, the image of I_i^m under this transformation belongs to (ith side of B)$\setminus \partial Q_m$. We also require that $\operatorname{Area}(B \setminus Q_m) \to 0$ and $|(i\text{th side of } B) \setminus \partial Q_m| =: c_i^m \to c_i$ as $m \to \infty$. In Appendix A it is shown how to construct such a sequence Q_m. The measure $\nu_{Q_m,B} = \tilde\nu_m + \sum_i \nu_i^m$ is the sum of the measure $\tilde\nu_m$ corresponding to reflections from $\partial B \cap \partial Q_m$ and the measures ν_i^m corresponding to particles getting into the "hollows on the ith side." One has $\tilde\nu_m = \sum_i (c_i - c_i^m)\cdot\eta_0 \otimes \delta_{n_i}$ and $\nu_i^m = c_i^m \cdot \eta_{\Omega_i^m, I_i^m} \otimes \delta_{n_i}$. The norm of $\tilde\nu_m$ goes to zero and ν_i^m weakly converges to $c_i\,\eta_i \otimes \delta_{n_i}$ for any i; it follows that $\nu_{Q_m,B}$ weakly converges to ν as $m \to \infty$. Therefore, the sequence Q_m represents a body $\mathcal{B}$ obtained by roughening B, and $\nu_{\mathcal{B}} = \nu$. □

STATEMENT 2. *For any measure $\nu \in \mathcal{M}_B$ there exist a sequence of convex polygons $B_k \subset B$ with $\operatorname{Area}(B \setminus B_k) \to 0$ and a sequence of measures $\nu_k \in \mathcal{M}_{B_k}$ weakly converging to ν as $k \to \infty$.*

Proof. Consider a partition of the circumference S^1 into a finite number of arcs, $S^1 = \cup_i \mathcal{S}^i$. It induces the partition of ∂B into arcs $\partial B^i = \{\xi \in \partial B : n(\xi) \in \mathcal{S}^i\}$. Consider the polygon $\check B$ inscribed into ∂B whose vertices are separation points of this partition. Denote by n_i the outer normal to the ith side of this polygon. Denote by s_{v_1,v_2} the operator of rotation on S^1 that takes v_1 to v_2, and define the mapping $\Upsilon_i : \mathbb{T}^2 \times \mathcal{S}^i \to \mathbb{T}^2$ by $\Upsilon_i(v, v^+, n) = (s_{n,n_i}v, s_{n,n_i}v^+)$. Finally, consider the measure $\check\nu = \sum_i |b^i|\,\eta_i \otimes \delta_{n_i}$, where $|b^i|$ is the length of the ith side of the polygon, and the measure η_i on $\mathbb{T}^2$ is defined by $\eta_i(A) = \frac{1}{|\partial B^i|}\,\nu(\Upsilon_i^{-1}(A))$ for arbitrary Borel set $A \subset \mathbb{T}^2$. Here $|\partial B^i|$ is the length of the arc ∂B^i. One easily verifies that $\check\nu$ belongs to $\mathcal{M}_{\check B}$.

Now take a sequence of partitions of S^1, $\{\mathcal{S}_k^i\}_i$, $k = 1, 2, \dots$, where the maximum arc length of a partition goes to zero as $k \to \infty$. Denote by $\{\partial B_k^i\}_i$, $k = 1, 2, \dots$, the sequence of induced partitions of ∂B, and take the sequence of polygons B_k generated by these partitions. One clearly has $\operatorname{Area}(B \setminus B_k) \to 0$ and

$$\max_i \frac{|b_k^i|}{|\partial B_k^i|} \to 1 \qquad \text{as} \quad k \to \infty, \tag{6}$$

where $|b_k^i|$ is the length of the ith side of B_k. In the same way as above, one defines the mappings $\Upsilon_{ik} : \mathbb{T}^2 \times \mathcal{S}_k^i \to \mathbb{T}^2$ and the measures $\nu_k = \sum_i |b_k^i|\, \eta_{ik} \otimes \delta_{n_{ik}} \in \mathcal{M}_{B_k}$, where η_{ik} is given by $\eta_{ik}(A) := \frac{1}{|\partial B_k^i|}\, \nu(\Upsilon_{ik}^{-1}(A))$ and n_{ik} is the outer unit normal to the ith side of B_k.

It remains to show that ν_k weakly converges to ν. For any continuous function f on $\mathbb{T}^3$ one has

$$\iiint_{\mathbb{T}^3} f(v, v^+, n)\, d\nu_k(v, v^+, n) = \sum_i |b_k^i| \iint_{\mathbb{T}^2} f(v, v^+, n_{ik})\, d\eta_{ik}(v, v^+)$$

$$= \sum_i \frac{|b_k^i|}{|\partial B_k^i|} \iiint_{\mathbb{T}^2 \times \mathcal{S}_k^i} f(\Upsilon_{ik}(v, v^+, n), n_{ik})\, d\nu(v, v^+, n). \tag{7}$$

For each k define the mapping from $\mathbb{T}^3$ to $\mathbb{T}^3$ by the relations $(v, v^+, n) \mapsto (\Upsilon_{ik}(v, v^+, n), n_{ik})$ if $n \in \mathcal{S}_k^i$. It uniformly converges to the identity mapping as $k \to \infty$; hence the function $\tilde{f}_k$, defined by the relations $\tilde{f}_k(v, v^+, n) := f(\Upsilon_{ik}(v, v^+, n), n_{ik})$ if $n \in \mathcal{S}_k^i$, uniformly converges to f as $k \to \infty$. From here and from (6) it follows that the right-hand side in (7) converges to $\iiint_{\mathbb{T}^3} f(v, v^+, n)\, d\nu(v, v^+, n)$ as $k \to \infty$. Thus, the convergence $\int f\, d\nu_k \to \int f\, d\nu$ is proved. □

The inverse statement of the theorem follows from Statements 1 and 2. Indeed, let $\nu \in \mathcal{M}_B$. Using Statement 2, find a sequence of convex polygons $B_k \subset B$ and a sequence $\nu_k \in \mathcal{M}_{B_k}$ weakly converging to ν. According to Statement 1, each measure ν_k is generated by a rough body. Consider the sequence of sets $Q_{kl} \subset B_k$, $l = 1, 2, \ldots$, representing this body, and then from all of these sequences choose a diagonal sequence $\tilde{Q}_k = Q_{kl_k}$ such that the corresponding sequence of measures $\nu_{\tilde{Q}_k, B}$ weakly converges to ν and $\mathrm{Area}(B \setminus \tilde{Q}_k)$ goes to zero as $k \to \infty$. The sequence $\tilde{Q}_k$ represents a body $\mathcal{B}$ obtained by roughening B and $\nu_{\mathcal{B}} = \nu$.

3. Proof of the lemmas.

3.1. Proof of Lemma 1. Consider the billiard in $\mathbb{R}^2 \setminus Q$. For $(\xi, v) \in (\partial B \times S^1)_+$, denote by $\tau(\xi, v)$ the time the billiard trajectory with the initial data $\xi, -v$ spends in $B \setminus Q$. In particular, if $\xi \in \partial B \cap \partial Q$, then one has $\tau(\xi, v) = 0$.

Denote by D the set of points $(x, w) \in (B \setminus Q) \times S^1$ that are accessible from $(\partial B \times S^1)_+$; that is, there exists $(\xi, v) \in (\partial B \times S^1)_+$ such that the billiard particle with the data $\xi, -v$ at the zero moment of time at some moment $0 \le t \le \tau(\xi, v)$ will pass through x with the velocity w. This description defines the change of coordinates in $D : (\xi, v, t) \mapsto (x, w)$; $(\xi, v) \in (\partial B \times S^1)_+$, $t \in [0, \tau(\xi, v)]$, and the element of phase volume $d^2x\, dw$ in the new coordinates takes the form $d\mu(\xi, v)\, dt$. Hence, the phase volume of D equals $\iiint_D d^2x\, dw = \iint_{(\partial B \times S^1)_+} \tau(\xi, v)\, d\mu(\xi, v)$. Taking into account that $D \subset (B \setminus Q) \times S^1$ and the phase volume of $(B \setminus Q) \times S^1$ equals $2\pi \cdot \mathrm{Area}(B \setminus Q)$, one gets

$$\iint_{(\partial B \times S^1)_+} \tau(\xi, v)\, d\mu(\xi, v) \le 2\pi \cdot \mathrm{Area}(B \setminus Q). \tag{8}$$

This is in fact a simple modification of the well-known *mean free path* formula (see, e.g., [19]).

One has $\tau(\xi, v) \ge |\xi - \xi^+_{Q,B}(\xi, v)|$: the time the particle spends in $B \setminus Q$ exceeds the distance between the initial and final points of the trajectory. This inequality and (8) imply (a).

The points ξ and $\xi^+_{Q,B}(\xi,v)$ divide the curve ∂B into two arcs; denote by $\gamma(\xi,v)$ the shortest one. One has $|\gamma(\xi,v)| = l(\xi,\xi^+_{Q,B}(\xi,v))$, and therefore $|\xi - \xi^+_{Q,B}(\xi,v)| \geq c\,|\gamma(\xi,v)|$. It follows that

(9)
$$c\iint_{(\partial B\times S^1)_+} |\gamma(\xi,v)|\,d\mu(\xi,v) \leq \iint_{(\partial B\times S^1)_+} |\xi-\xi^+_{Q,B}(\xi,v)|\,d\mu(\xi,v) \leq 2\pi\cdot\mathrm{Area}(B\setminus Q).$$

Let $\varrho(y)$ be a natural parametrization of the curve ∂B, $\varrho : [0,|\partial B|] \to \partial B$. By $f(y)$ denote the measure of the values (ξ,v) such that the interval $\gamma(\xi,v)$ contains the point $\varrho(y)$; that is, $f(y) := \iint_{(\partial B\times S^1)_+} \mathbb{I}(\varrho(y)\in\gamma(\xi,v))\,d\mu(\xi,v)$. Making a change of variables in the integral in the left-hand side of (9), one gets

$$\iint_{(\partial B\times S^1)_+} |\gamma(\xi,v)|\,d\mu(\xi,v) = \int_0^{|\partial B|} f(y)\,dy,$$

and therefore

$$\int_0^{|\partial B|} f(y)\,dy \leq \frac{2\pi}{c}\,\mathrm{Area}(B\setminus Q). \tag{10}$$

One easily sees that $|f(y_1)-f(y_2)| \leq 4\,|y_1-y_2|$ for any y_1 and y_2 and $f(y)\geq 0$. From here and from (10) it follows that for sufficiently small $\mathrm{Area}(B\setminus Q)$ (namely, for $\mathrm{Area}(B\setminus Q) \leq c|\partial B|^2/(2\pi)$) it holds that $f(y) \leq \sqrt{8\pi/c}\,\sqrt{\mathrm{Area}(B\setminus Q)}$.

Recall that $\mathrm{Arg}(v)$ is the angle the vector $v\neq 0$ forms with a fixed vector v_0; the angle is measured clockwise from v_0 to v and is defined modulo 2π. Introduce the shorthand notation $\xi^+ := \xi^+_{Q,B}(\xi,v)$, and denote by $\Delta\,\mathrm{Arg}(\xi,v)$ the smallest in modulus of the values $\mathrm{Arg}(n(\xi^+)) - \mathrm{Arg}(n(\xi))$. In other words, $\Delta\,\mathrm{Arg}(\xi,v)$ equals the smallest of the values

$$\int_{\gamma(\xi,v)} |d\,\mathrm{Arg}(n_{\xi'})|, \qquad \int_{\partial B\setminus\gamma(\xi,v)} |d\,\mathrm{Arg}(n_{\xi'})|.$$

Taking into account that $|n(\xi^+)-n(\xi)| \leq |\Delta\,\mathrm{Arg}(\xi,v)|$, one gets that

$$|n(\xi^+)-n(\xi)| \leq \int_{\gamma(\xi,v)} |d\,\mathrm{Arg}(n_{\xi'})|,$$

and therefore

$$\overline{|n-n^+|}_{Q,B} \leq \iint_{(\partial B\times S^1)_+} \left(\int_{\gamma(\xi,v)} |d\,\mathrm{Arg}(n_{\xi'})|\right) d\mu(\xi,v).$$

Making a change of variables in this integral, one obtains

$$\overline{|n-n^+|}_{Q,B} \leq \int_0^{|\partial B|} f(y)\,|d\,\mathrm{Arg}(n_{\varrho(y)})| \leq 2\pi\,\sqrt{8\pi/c}\,\sqrt{\mathrm{Area}(B\setminus Q)}.$$

Thus, (b) is also proved.

3.2. Proof of Lemma 2. Fix $n \in S^1$ and $m \in \mathbb{N}$. Let σ be an involutive permutation of $\{1,\dots,m\}$, that is, $\sigma^2 = \mathrm{id}$. Divide the half-circumference $S^1_n := \{v \in S^1 : \langle v, n\rangle \geq 0\}$ into m arcs $\mathcal{S}^1_{n,m} = \mathcal{S}^1_n$, …, $\mathcal{S}^m_{n,m} = \mathcal{S}^m_n$ numbered clockwise such that, for any i, $\int_{\mathcal{S}^i_n} \langle v, n\rangle\, dv = 2/m$. For the sake of brevity we omit the subscript m when no confusion can arise.

Definition 6. *A measure η is called a (σ, n)-measure if $\eta \in \Lambda_n$ and $\operatorname{spt}\eta \subset \cup_{i=1}^m \left(\mathcal{S}^i_n \times \mathcal{S}^{\sigma(i)}_n\right)$, and therefore, for any i it holds that $\eta\left(\mathcal{S}^i_n \times \mathcal{S}^{\sigma(i)}_n\right) = 2/m$.*

Proposition 1. *For any measure $\eta \in \Lambda_n$ there exists a sequence of involutive permutations σ_k on $\{1,\dots,m_k\}$, $k = 1, 2, \dots$, such that m_k tends to infinity and any sequence of (σ_k, n)-measures weakly converges to η as $k \to \infty$.*

Proposition 2. *Let σ be an involutive permutation on $\{1,\dots,m\}$. Then the distance (in variation) between the set of measures generated by n-hollows and the set of (σ, n)-measures does not exceed $16/m$. In other words, whatever $\varepsilon > 0$, there exist a (σ, n)-measure η and an n-hollow (Ω, I) such that $\|\eta_{\Omega,I} - \eta\| < 16/m + \varepsilon$; here the norm means variation of measure.*

This distance actually equals zero, but we need only the (weaker) claim of Proposition 2.

Lemma 2 follows from Propositions 1 and 2. Indeed, let $\eta \in \Lambda_n$. First, choose the sequence of permutations σ_k, according to Proposition 1, and then, using Proposition 2, for every k choose an an n-hollow (Ω_k, I_k) such that the distance from η_{Ω_k,I_k} to the set of (σ_k, n)-measures does not exceed $17/m_k$. The sequence of chosen measures η_{Ω_k,I_k} weakly converges to η.

3.3. Proof of Proposition 1. Introduce on S^1_n the angular coordinate $\varphi = \operatorname{Arg} v - \operatorname{Arg} n$; that is, φ changes between $-\pi/2$ and $\pi/2$ and increases clockwise. With this notation, to the arcs $\mathcal{S}^i_{n,m}$ correspond the segments $J^i_m = [\arcsin(-1+2(i-1)/m), \arcsin(-1+2i/m)]$. Define the measure λ on $[-\pi/2, \pi/2]$ by $d\lambda(\varphi) = \cos\varphi\, d\varphi$, and denote by Λ the set of measures η on $\square := [-\pi/2, \pi/2] \times [-\pi/2, \pi/2]$ such that (a) $\pi^{\#}_{\varphi}\eta = \lambda = \pi^{\#}_{\varphi^+}\eta$ and (b) $\pi^{\#}_d\eta = \eta$. Here π_φ, π_{φ^+}, and π_d are defined by $\pi_\varphi(\varphi, \varphi^+) = \varphi$, $\pi_{\varphi^+}(\varphi, \varphi^+) = \varphi^+$, and $\pi_d(\varphi, \varphi^+) = (\varphi^+, \varphi)$. Reformulating Definition 6, we shall say that η is a σ-measure if $\eta \in \Lambda$ and $\operatorname{spt}\eta \subset \cup_{i=1}^m \left(J^i_m \times J^{\sigma(i)}_m\right)$. Notice that in the new notation the objects no longer depend on n: we write Λ instead of Λ_n, σ-measure instead of (σ, n)-measure, and hollow instead of n-hollow.

In this notation, Proposition 1 can be reformulated as follows: for any measure $\eta \in \Lambda$ there exists a sequence of involutive permutations σ_k on $\{1,\dots,m_k\}$, $k = 1, 2, \dots$, such that m_k tends to infinity and any sequence of σ_k-measures weakly converges to η as $k \to \infty$.

The idea of the proof is as follows. First, η is approximated by means of a rational matrix, and then this matrix is approximated by means of a larger matrix generated by a permutation.

Consider the partition of $\square$ into smaller rectangles $\square^{ij}_k = J^i_k \times J^j_k$, $i, j = 1, \dots, k$. Choose rational nonnegative numbers c^{ij}_k such that $c^{ij}_k = c^{ji}_k$, $\sum_j c^{ij}_k = 2/k$ for any i, and $\left|\eta(\square^{ij}_k) - c^{ij}_k\right| \leq k^{-3}$ for any i and j. To do so, it suffices to take positive rational values c^{ij}_k such that $\eta(\square^{ij}_k) - k^{-4} \leq c^{ij}_k \leq \eta(\square^{ij}_k)$ for $i > j$ and put $c^{ij}_k = c^{ji}_k$ for $i < j$ and $c^{ii}_k = 2/k - \sum_{j\neq i} c^{ij}_k$ for $i = j$. One has $\eta\left(J^i_k \times [-\pi/2, \pi/2]\right) = \sum_{j=1}^k \eta(\square^{ij}_k) = 2/k$; hence $c^{ii}_k - \eta(\square^{ii}_k) = \sum_{j\neq i}\left(\eta(\square^{ij}_k) - c^{ij}_k\right) \in [0, (k-1)\cdot k^{-4}] \subset [0, k^{-3}]$.

Any sequence of measures η_k satisfying the conditions $\eta_k(\square^{ij}_k) = c^{ij}_k$, $1 \leq i, j \leq k$, weakly converges to η. Indeed, for any continuous function f on $\square$ it holds that

$$\int_{\square} f\,d\eta_k - \int_{\square} f\,d\eta = \sum_{i,j=1}^{k} \int_{\square_k^{ij}} f\,(d\eta_k - d\eta) \le k^{-1} \max f \to 0$$

as $k \to \infty$.

To complete the proof, it suffices to find an integer $m_k > k$ and an involutive permutation σ_k of $\{1, \dots, m_k\}$ such that any σ_k-measure, η_k, satisfies the equalities $\eta_k(\square_k^{ij}) = c_k^{ij}$, $i, j = 1, \dots, k$. Choose a positive integer N such that all of the values $a_{ij} := N \cdot c_k^{ij}$ are integer. The obtained matrix $A = (a_{ij})_{i,j=1}^{k}$ is symmetric, and for any i the value $\sum_{j=1}^{k} a_{ij} = 2N/k$ is a fixed positive integer. In Appendix B it is shown that there exist square matrices $B_{ij} = (b_{ij}^{\mu\nu})_{\mu,\nu}$ of size $2N/k$ such that $B_{ij}^T = B_{ji}$, the sum of elements in any matrix B_{ij} equals a_{ij}, and the block matrix $D = (B_{ij})$ composed of these matrices has exactly one unit in each row and each column, and other elements are zeros.

D is a symmetric square matrix of size $2N$; denote its elements by d_{ij}. Put $m_k = 2N$, and define the mapping σ_k on $\{1, \dots, 2N\}$ in such a way that $d_{i\sigma_k(i)} = 1$ for any i. The so defined mapping σ_k is a permutation; it is involutive since the matrix D is symmetric. Moreover, if η_k is a σ_k-measure, then for any i and j it holds that $\eta_k(\square_k^{ij}) = N^{-1} \sum_{\mu,\nu} b_{ij}^{\mu\nu} = c_k^{ij}$. The proposition is proved.

3.4. Proof of Proposition 2.

1. Whatever the n-hollow (Ω, I), one introduces the reference system (x_1, x_2) in such a way that n coincides with $(0, -1)$, and the interval I belongs to the straight line $x_2 = 0$ and contains the origin $O = (0, 0)$. Like in the proof of Proposition 1, introduce the coordinate $\varphi = \operatorname{Arg} v - \operatorname{Arg} n$ on S_n^1. One has $v = -(\sin\varphi, \cos\varphi)$, $\varphi \in [-\pi/2, \pi/2]$. The definitions of the segments $J_m^i = J^i$, the measure λ, the set of measures Λ, and the σ-measure are seen in the beginning of the previous subsection. The mapping $(\xi, v) \mapsto V_{\Omega,I}^+(\xi, v)$ in the new coordinates ξ, φ is written as $(\xi, \varphi) \mapsto \varphi_{\Omega,I}^+(\xi, \varphi)$. Finally, define the measure μ_I on $I \times [-\pi/2, \pi/2]$ by $d\mu_I(\xi, \varphi) = \frac{\cos\varphi}{|I|}\,d\xi\,d\varphi$.

Denote $\square' = \left(\cup_{i=2}^{m-1} J^i\right) \times \left(\cup_{i=2}^{m-1} J^i\right)$, $\square_1 = J^1 \times [-\pi/2, \pi/2]$, $\square_2 = J^m \times [-\pi/2, \pi/2]$, $\square_3 = \left(\cup_{i=2}^{m-1} J^i\right) \times J^1$, and $\square_4 = \left(\cup_{i=2}^{m-1} J^i\right) \times J^m$. Thus, one has $\square \setminus \square' = \square_1 \cup \square_2 \cup \square_3 \cup \square_4$; see Figure 6.

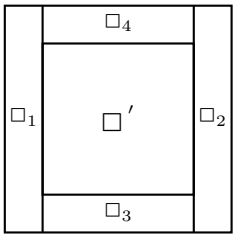


FIG. 6. *Partition of the square into rectangles.*

It suffices to construct a sequence of hollows $(\Omega_\varepsilon, I_\varepsilon)$, $\varepsilon > 0$, such that

(P) *for any* $i \ne 1, m, \sigma(1), \sigma(m)$ *the measure of the set of values* $(\xi, \varphi) \in I_\varepsilon \times J^i$ *such that* $\varphi_{\Omega_\varepsilon, I_\varepsilon}^+(\xi, \varphi) \notin J^{\sigma(i)}$ *goes to zero as* $\varepsilon \to 0$.

Then, speaking of restrictions of measures on the subset $\square'$, one gets that the distance from the restrictions of measures $\eta_{\Omega_\varepsilon, I_\varepsilon}$ to the set of restrictions of σ-measures goes to zero as $\varepsilon \to 0$. On the other hand, for any measure $\eta \in \Lambda$ one has $\eta(\square_1) =$

$\eta(\square_2) = 2/m$, $\eta(\square_3) \le 2/m$, $\eta(\square_4) \le 2/m$, and hence $\eta(\square \setminus \square') \le 8/m$; therefore the distance between the restrictions on $\square \setminus \square'$ of any two measures η_1 and η_2 from Λ does not exceed $16/m$: $\| \eta_1 \rfloor_{\square \setminus \square'} - \eta_2 \rfloor_{\square \setminus \square'} \| \le 16/m$. It follows that the upper limit of distances from $\eta_{\Omega_\varepsilon, I_\varepsilon}$ to the set of σ-measures does not exceed $16/m$, and so Proposition 2 is proved.

2. The rest of this subsection is dedicated to the detailed description of the sequence of hollows $(\Omega_\varepsilon, I_\varepsilon)$ and to the proof of property (P) for them.

First, consider an auxiliary construction (see Figure 7). Take two different points F and F' above the line $l = \{x_2 = 0\}$, with $|OF| = 2 = |OF'|$. Denote by Φ and Φ' the angles the rays OF and OF', respectively, formed with the vector $(0, 1)$. The angles are counted clockwise from $(0, 1)$. Thus, $F = 2(\sin\Phi, \cos\Phi)$ and $F' = 2(\sin\Phi', \cos\Phi')$. Assume, for further convenience, that F is situated on the left of F'; thus, one has $-\pi/2 < \Phi < \Phi' < \pi/2$. (The case where F is situated on the right of F' is completely similar.) Select three positive numbers λ, λ', and δ, and define two ellipses $\mathcal{E}$ and $\mathcal{E}$' and two parabolas $\mathcal{P}$ and $\mathcal{P}$'. The first ellipse has the foci O and F, the length of its large semiaxis is $\sqrt{1+\lambda}$, of the small semiaxis, $\sqrt{\lambda}$, and the focal distance equals 2. The second ellipse has the foci O and F', the lengths of its large and small semiaxes are $\sqrt{1+\lambda'}$ and $\sqrt{\lambda'}$, respectively, and the focal distance is also 2. The parabolas $\mathcal{P}$ and $\mathcal{P}$' have the foci F and F', respectively, the common axis FF', and the same focal distance δ. Thus, the parabolas are symmetric to each other with respect to the bisectrix of the triangle OFF'. The parameter δ is chosen sufficiently small, so that the point O lies in the exterior of both parabolas.

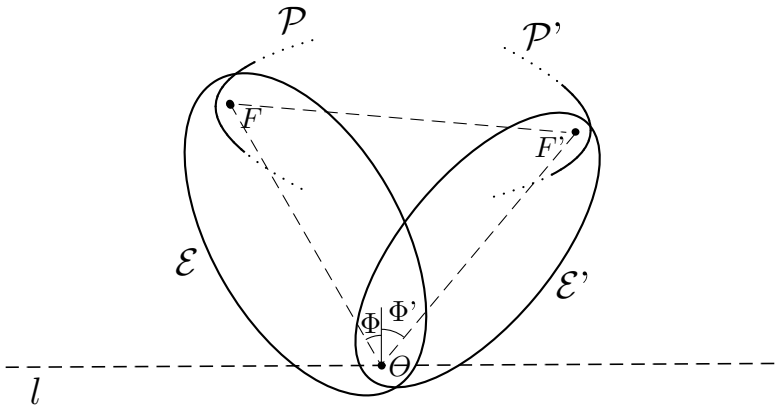


FIG. 7. *Auxiliary construction.*

In what follows, we shall distinguish between the billiard and *pseudobilliard* dynamics. The pseudobilliard dynamics is defined as follows. A particle starts at a point $(\xi, 0) \in l$ and moves with a velocity $(\sin\varphi, \cos\varphi)$ until it reflects from the interior side of $\mathcal{E}$. (Before the reflection it can intersect other curves $\mathcal{E}', \mathcal{P}, \mathcal{P}'$, or even intersect $\mathcal{E}$ from the outer side, without changing the velocity.) Then it moves again with constant velocity until it reflects from the interior side of $\mathcal{P}$. Then, in the same way, it reflects from the interior side of $\mathcal{P}'$, and then from the interior side of $\mathcal{E}'$, and, finally, it intersects l from above to below. Denote by $(\xi', 0)$ the point of intersection and by $-(\sin\varphi', \cos\varphi')$ the velocity at this point.

Consider the admissible set: the set of 7-tuples $(\varphi, \xi, \Phi, \Phi', \lambda, \lambda', \delta)$ such that all of the indicated reflections occur in the prescribed order. This set is open and nonempty.

Indeed, let $\delta(\Phi, \Phi')$ be the least of the values δ such that one of the parabolas (in fact, both of them simultaneously) passes through O. Put $\varphi = \Phi$, $\xi = 0$, and take arbitrary values $\lambda > 0$, $\lambda' > 0$, $-\pi/2 < \Phi < \Phi' < \pi/2$, $0 < \delta < \delta(\Phi, \Phi')$. The particle with initial data $\varphi = \Phi$, $\xi = 0$ first passes along the large semiaxis of $\mathcal{E}$, reflects from $\mathcal{E}$, returns along the same semiaxis, and reflects from $\mathcal{P}$. Then it moves with the velocity parallel to FF', reflects from $\mathcal{P}'$, moves the large semiaxis of the ellipse $\mathcal{E}'$, reflects from it, and returns to O along the same semiaxis. Thus, the admissible set is nonempty. Under a small perturbation of the parameters $\varphi, \xi, \Phi, \Phi', \lambda, \lambda', \delta$, all of the reflections are maintained and the order of reflections remains the same. This implies that the admissible set is open.

This description determines the mapping $\varphi' = \varphi'(\varphi, \xi, \Phi, \Phi', \lambda, \lambda', \delta)$, $\xi' = \xi'(\varphi, \xi, \Phi, \Phi', \lambda, \lambda', \delta)$[2] from the admissible set to $\mathbb{R}^2$. This mapping is infinitely differentiable. For $\varphi = \Phi$ and $\xi = 0$ one has

$$\varphi'(\Phi, 0, \Phi, \Phi', \lambda, \lambda', \delta) = \Phi'. \tag{11}$$

For $\xi = 0$ with arbitrary φ one has

$$\xi'(\varphi, 0, \Phi, \Phi', \lambda, \lambda', \delta) = 0, \tag{12}$$

and

$$\varphi'(\varphi, 0, \Phi, \Phi', \lambda, \lambda', \delta) \quad \text{does not depend on} \quad \delta.$$

Indeed, a particle starting at O, after the reflection from $\mathcal{E}$ passes through F, after reflecting from $\mathcal{P}$ moves in parallel to FF', after the reflection from $\mathcal{P}'$ passes through F', and, finally, after the reflection from $\mathcal{E}'$ returns to O (see Figure 8). The initial and final velocities of the particle are, respectively, $(\sin\varphi, \cos\varphi)$ and $-(\sin\varphi', \cos\varphi')$. Denoting by α and α' the angles the second and fourth segments of the (5-segment) trajectory form, respectively, with OF and OF', one has $\alpha = \alpha'$. The angle α is a function of φ, and φ' is a function of α'; these functions depend only on the parameters of the ellipses $\mathcal{E}$ and $\mathcal{E}'$, respectively, and do not depend on the parameter δ determining the shape of parabolas.

Using properties of ellipses, one derives the formulas connecting φ, α, and $\varphi' = \varphi'(\varphi, 0, \Phi, \Phi', \lambda, \lambda', \delta)$:

$$\sin(\varphi - \Phi) = \frac{\lambda \sin\alpha}{2 + \lambda - 2\cos\alpha\sqrt{1+\lambda}}, \qquad \sin(\varphi' - \Phi') = -\frac{\lambda' \sin\alpha}{2 + \lambda' - 2\cos\alpha\sqrt{1+\lambda'}}. \tag{13}$$

It follows that

$$\left.\frac{\partial\varphi'}{\partial\varphi}\right|_{\substack{\varphi=\Phi \\ \xi=0}} = -\left(\frac{\sqrt{\lambda'}}{1+\sqrt{\lambda'}}\frac{1+\sqrt{\lambda}}{\sqrt{\lambda}}\right)^2. \tag{14}$$

With fixed Φ, Φ', λ, λ', and δ the mapping $\varphi'(\varphi, \xi)$, $\xi'(\varphi, \xi)$ preserves the measure, $\cos\varphi\, d\varphi\, d\xi = \cos\varphi'\, d\varphi'\, d\xi'$, and hence

$$\cos\varphi = \pm\cos\varphi' \begin{vmatrix} \frac{\partial\varphi'}{\partial\varphi} & \frac{\partial\varphi'}{\partial\xi} \\ \frac{\partial\xi'}{\partial\varphi} & \frac{\partial\xi'}{\partial\xi} \end{vmatrix}.$$

[2]Note that throughout this paper the sign ' (prime) never means derivation.

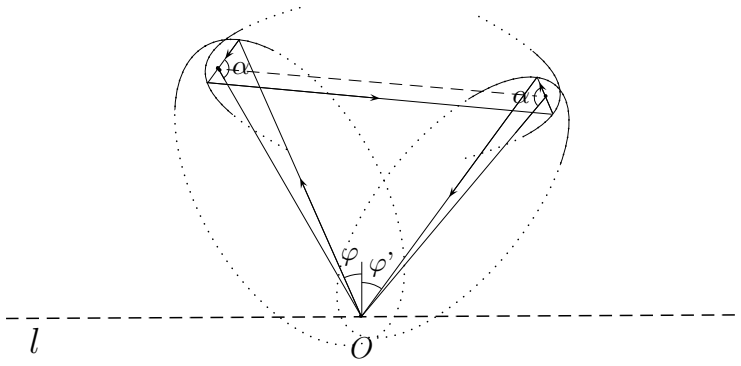


FIG. 8. *Pseudobilliard dynamics.*

Using (12), one gets that $\frac{\partial\xi'}{\partial\varphi}\big\rfloor_{\xi=0} = 0$, and hence

$$\begin{vmatrix} \frac{\partial\varphi'}{\partial\varphi} & \frac{\partial\varphi'}{\partial\xi} \\ \frac{\partial\xi'}{\partial\varphi} & \frac{\partial\xi'}{\partial\xi} \end{vmatrix}_{\xi=0} = \frac{\partial\varphi'}{\partial\varphi}\,\frac{\partial\xi'}{\partial\xi}\Big\rfloor_{\xi=0};$$

therefore

$$\cos\varphi = \pm\cos\varphi'\,\frac{\partial\varphi'}{\partial\varphi}\,\frac{\partial\xi'}{\partial\xi}\Big\rfloor_{\xi=0}. \tag{15}$$

Putting $\varphi = \Phi$ and $\xi = 0$, and taking into account (11), (14), and (15), one gets

$$\cos\Phi = \pm\cos\Phi'\left(\frac{\sqrt{\lambda'}}{1+\sqrt{\lambda'}}\,\frac{1+\sqrt{\lambda}}{\sqrt{\lambda}}\right)^2 \frac{\partial\xi'}{\partial\xi}\Big\rfloor_{\substack{\varphi=\Phi\\ \xi=0}}. \tag{16}$$

Define the positive continuous functions $\lambda(\Phi')$ and $\lambda'(\Phi)$ by the relations

$$\left(\frac{\sqrt{\lambda}}{1+\sqrt{\lambda}}\right)^2 = \frac{1}{2}\cos\Phi', \qquad \left(\frac{\sqrt{\lambda'}}{1+\sqrt{\lambda'}}\right)^2 = \frac{1}{2}\cos\Phi; \tag{17}$$

then one has

$$\left|\frac{\partial\xi'}{\partial\xi}\Big\rfloor_{\substack{\varphi=\Phi;\ \lambda=\lambda(\Phi')\\ \xi=0;\ \lambda'=\lambda'(\Phi)}}\right| = 1. \tag{18}$$

Additionally, taking into account (14) and (17), one gets

$$\frac{\cos\Phi'}{\cos\Phi}\,\frac{\partial\varphi'}{\partial\varphi}\Big\rfloor_{\substack{\varphi=\Phi;\ \lambda=\lambda(\Phi')\\ \xi=0;\ \lambda'=\lambda'(\Phi)}} = -1. \tag{19}$$

Recall that $\varphi' = \varphi'(\varphi, 0, \Phi, \Phi', \lambda, \lambda')$; that is, the restriction of the function φ' to the subspace $\xi = 0$ does not depend on δ. Hence the function $\frac{\partial\varphi'}{\partial\varphi}\big\rfloor_{\xi=0}$ and, by formula (15), the function $\frac{\partial\xi'}{\partial\xi}\big\rfloor_{\xi=0}$ also do not depend on δ. Put $\Phi_0 = \arcsin(1 - 2/m)$,

so that $J^1 = [-\pi/2, -\Phi_0]$ and $J^m = [\Phi_0, \pi/2]$, and put $\Delta\Phi = 2/m$. The set $\{(\Phi, 0, \Phi, \Phi', \lambda(\Phi'), \lambda'(\Phi)) : -\Phi_0 \le \Phi, \Phi' \le \Phi_0,\ \Phi' - \Phi \ge \Delta\Phi\}$ is compact and belongs to the (open) domain of the function φ'. Choose a sufficiently large integer value $k = k(\varepsilon)$, so that for

$$|\sin\varphi - \sin\Phi| < 2/(km), \qquad \xi = 0, \qquad -\Phi_0 \le \Phi, \Phi' \le \Phi_0, \tag{20}$$
$$\Phi' - \Phi \ge \Delta\Phi, \qquad \lambda = \lambda(\Phi'), \qquad \lambda' = \lambda'(\Phi)$$

it holds true that

$$-\frac{\cos\Phi'}{\cos\Phi}\frac{\partial\varphi'}{\partial\varphi} \in [(1+\varepsilon)^{-1}, 1+\varepsilon]. \tag{21}$$

Formulas (21) and (11) mean that, under conditions (20), φ' is also close to Φ'. Increasing k if necessary, ensure (under the same conditions) that

$$\frac{\cos\Phi'}{\cos\Phi}\frac{\cos\varphi}{\cos\varphi'} \in [(1+\varepsilon)^{-1}, 1+\varepsilon]. \tag{22}$$

Taking into account (15), (21), and (22), one obtains that under conditions (20) it holds true that

$$\left|\frac{\partial\xi'}{\partial\xi}\right| \in [(1+\varepsilon)^{-2}, (1+\varepsilon)^2]. \tag{23}$$

3. Now we proceed to the description of the hollow $(\Omega_\varepsilon, I_\varepsilon)$.

(a) If $2 \le i \ne \sigma(i) \le m-1$, then divide the interval J^i into k subintervals $J^{i,j}$ of equal measure λ, going in increasing order: $J^i = \cup_{j=1}^k J^{i,j}$, $\lambda(J^{i,j}) = 2/(km)$ for any $j = 1, \dots, k$. Recall that $d\lambda(\varphi) = \cos\varphi\, d\varphi$ and the value $k = k(\varepsilon)$ is defined above. Without loss of generality assume that $k(\varepsilon) \to \infty$ as $\varepsilon \to 0$.

To each pair of intervals, $J^{i,j}$ and $J^{\sigma(i),j}$, we apply the construction described above; see Figure 9. Namely, draw arcs of ellipses $\mathcal{E}_{i,j} = \overset{\smile}{AB}$, $\mathcal{E}'_{i,j} = \overset{\smile}{A'B'}$ and arcs of parabolas $\mathcal{P}_{i,j}$, $\mathcal{P}'_{i,j}$. Without loss of generality suppose that $i < \sigma(i)$. The angles AOB and $A'OB'$ correspond to the angular intervals $J^{i,j}$ and $J^{\sigma(i),j}$, respectively. The foci $\bar F = F_{i,j}$ and $\bar F' = F'_{i,j}$ belong to the intervals OA and OA', respectively. The endpoints of the arcs $\mathcal{P}_{i,j}$ and $\mathcal{P}'_{i,j}$ also belong to the intervals OA and OA', respectively. The angle corresponding to the ray OA (and therefore to the left endpoint of the interval $J^{i,j}$) will be denoted by $\bar\Phi = \Phi_{i,j}$, and the angle corresponding to the ray OA' (and therefore to the right endpoint of the interval $J^{\sigma(i),j}$) will be denoted by $\bar\Phi' = \Phi'_{i,j}$. Denote $\bar\lambda = \lambda_{i,j} := \lambda(\bar\Phi')$ and $\bar\lambda' = \lambda'_{i,j} := \lambda'(\bar\Phi)$, according to formula (17). Next, select a value $\bar\delta = \delta_{i,j}$ and draw two curves (*lateral reflectors*) in such a way that (i) each of the curves contains an arc of parabola (the first curve contains $\mathcal{P}_{i,j}$ and the second one $\mathcal{P}'_{i,j}$), an arc of circumference centered at O, and three radial segments; (ii) these curves do not intersect the intervals whose endpoints belong to the set $\{F_{\alpha,\beta}, F'_{\gamma,\delta} : (\alpha,\beta) \ne (i,j),\ (\gamma,\delta) \ne (\sigma(i), j)\}$: this will guarantee free passage of particles from one parabola to another; and (iii) the λ-measure of the angular interval occupied by each lateral reflector does not exceed $\varepsilon/(km)$. In Figure 9, the angular reflectors are the curves joining the points A and C, and the points A' and C'.

Notice that $-\Phi_0 \le \bar\Phi, \bar\Phi' \le \Phi_0$, and $\bar\Phi' - \bar\Phi \ge \Delta\Phi$. Indeed, $\bar\Phi$ and $\bar\Phi'$ do not belong to the intervals $J^1 = [-\pi/2, -\Phi_0]$ and $J^m = [\Phi_0, \pi/2]$. On the other hand, one has $\bar\Phi' - \bar\Phi \ge \sin\bar\Phi' - \sin\bar\Phi = \lambda([\bar\Phi, \bar\Phi']) \ge 2/m = \Delta\Phi$.

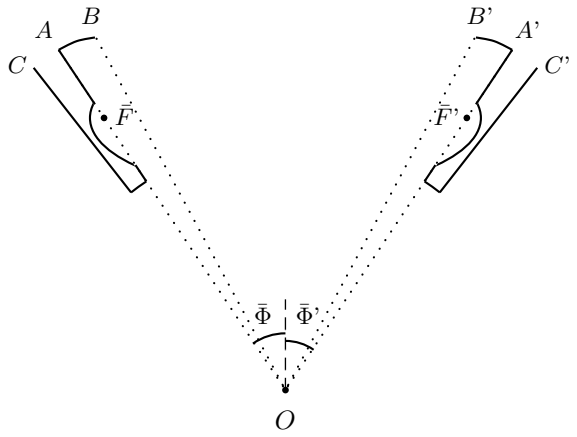


FIG. 9. *Part of the hollow corresponding to the angular intervals $J^{i,j}$ and $J^{\sigma(i),j}$.*

Introduce the shorthand notation $\varphi'(\varphi,\xi)=\varphi'(\varphi,\xi,\Phi_{i,j},\Phi'_{i,j},\lambda_{i,j},\lambda'_{i,j},\delta_{i,j})$. According to (21) and (23), for $\varphi\in J^{i,j}$ it holds true that

$$-\frac{\cos\bar\Phi'}{\cos\bar\Phi}\frac{\partial\varphi'}{\partial\varphi}(\varphi,0)\in[(1+\varepsilon)^{-1},1+\varepsilon] \tag{24}$$

and

$$\left|\frac{\partial\xi'}{\partial\xi}(\varphi,0)\right|\in[(1+\varepsilon)^{-2},(1+\varepsilon)^2]. \tag{25}$$

According to (11), one has $\varphi'(\bar\Phi,0)=\bar\Phi'$; this equality and formula (24) imply that for $\varphi\in J^{i,j}$ and $\varphi'=\varphi'(\varphi,0)$ one has

$$-\frac{\cos\bar\Phi'}{\cos\bar\Phi}\frac{\varphi'-\bar\Phi'}{\varphi-\bar\Phi}\in[(1+\varepsilon)^{-1},1+\varepsilon]. \tag{26}$$

On the other hand, one has

$$\cos\bar\Phi\,|J^{i,j}|=\frac{2}{km}\,(1+o(1)), \tag{27}$$

$$\cos\bar\Phi'\,|J^{\sigma(i),j}|=\frac{2}{km}\,(1+o(1)), \tag{28}$$

with $o(1)$ being uniformly small over all i,j as $\varepsilon\to0$, and $|J|$ being the Lebesgue measure of J. (Recall that the parameters $\bar\Phi,\bar\Phi',k$ and the intervals $J^{i,j}$ implicitly depend on ε.)

Choose closed intervals $\tilde J^{i,j}\subset J^{i,j}$ and $\tilde J^{\sigma(i),j}\subset J^{\sigma(i),j}$ satisfying the following conditions: (i) $\varphi'(\tilde J^{i,j}\times\{0\})=\tilde J^{\sigma(i),j}$; (ii) some neighborhoods of $\tilde J^{i,j}$ and $\tilde J^{\sigma(i),j}$ belong to $J^{i,j}$ and $J^{\sigma(i),j}$, respectively; and (iii) the pseudobilliard trajectory with the initial data $(\varphi,0)$, $\varphi\in\tilde J^{i,j}$ does not intersect the neighbor lateral reflectors (that is, the lateral reflectors corresponding to the intervals $\tilde J^{i,j+1}$ and $\tilde J^{\sigma(i),j-1}$ if $j\ne1,k$; if $j=1$, then $\tilde J^{\sigma(i),j-1}$ should be replaced with $\tilde J^{\sigma(i)-1,k}$, and if $j=k$, then $\tilde J^{i,j+1}$ should be replaced with $\tilde J^{i+1,1}$). Note in this regard that the neighbor lateral reflectors occupy a small part of the angular intervals $J^{i,j}$ and $J^{\sigma(i),j}$ (represented in the figure

by the arcs AB and $B'A'$). Other lateral reflectors will not be intersected by the choice of lateral reflectors.

By virtue of (26), (27), (28) and because of smallness of the angular intervals occupied by the lateral reflectors, $\tilde{J}^{i,j}$ and $\tilde{J}^{\sigma(i),j}$ may be chosen in such a way that the ratios $\lambda(\tilde{J}^{i,j})/\lambda(J^{i,j})$ and $\lambda(\tilde{J}^{\sigma(i),j})/\lambda(J^{\sigma(i),j})$ uniformly (with respect to i, j) tend to 1 as $\varepsilon \to 0$. Thus, a billiard particle going from O in a direction $\varphi \in \tilde{J}^{i,j}$ makes the same reflections and in the same order as under the pseudobilliard dynamics: first, reflection from $\mathcal{E}_{i,j}$, from $\mathcal{P}_{i,j}$, from $\mathcal{P}'_{i,j}$, and from $\mathcal{E}'_{i,j}$; finally, the particle goes back to O in the direction $\varphi'(\varphi, 0) \in \tilde{J}^{\sigma(i),j}$.

Choose $a_{i,j}$ in such a way that the following conditions are fulfilled: if $(\xi, \varphi) \in [-a_{i,j}, a_{i,j}] \times \tilde{J}^{i,j}$, then (i) the corresponding billiard trajectory does not intersect the lateral reflectors and the indicated order of reflections is preserved; (ii) $\varphi'(\varphi, \xi) \in J^{\sigma(i),j}$; (iii) $\left|\frac{\partial \xi'}{\partial \xi}(\varphi, \xi)\right| \in [(1+\varepsilon)^{-3}, (1+\varepsilon)^3]$. Analogously, choose $a_{\sigma(i),j}$ in such a way that the conditions are fulfilled: if $(\xi, \varphi) \in [-a_{\sigma(i),j}, a_{\sigma(i),j}] \times \tilde{J}^{\sigma(i),j}$, then (i) the billiard trajectory does not intersect the lateral reflectors and the order of its reflections is reversed; (ii) $\varphi'(\varphi, \xi) \in J^{i,j}$; (iii) $\left|\frac{\partial \xi'}{\partial \xi}(\varphi, \xi)\right| \in [(1+\varepsilon)^{-3}, (1+\varepsilon)^3]$. Note that the values $a_{i,j}$ and $a_{\sigma(i),j}$ implicitly depend on ε.

Select $a_\varepsilon \le \min_{i,j} a_{ij}$ in such a way that $a_\varepsilon \to 0$ as $\varepsilon \to 0$, and denote $I_\varepsilon = (-a_\varepsilon, a_\varepsilon) \times \{0\}$, $\tilde{I}_\varepsilon = (-a_\varepsilon(1+\varepsilon)^{-3}, a_\varepsilon(1+\varepsilon)^{-3}) \times \{0\}$, and $\tilde{J}^i_\varepsilon = \tilde{J}^i := \cup_j \tilde{J}^{i,j}_\varepsilon$. The part of the boundary of Ω_ε related to the angular intervals $J^{i,j}$ and $J^{\sigma(i),j}$ under consideration is formed by the arcs of ellipses $\mathcal{E}_{i,j}$, $\mathcal{E}'_{i,j}$ and the corresponding lateral reflectors. Then a billiard particle with initial conditions $(\xi, \varphi) \in \tilde{I}_\varepsilon \times \tilde{J}^{i,j}$ after making four reflections will intersect l at a point $(\xi', 0) \in I_\varepsilon$, and the angle at the point of intersection will be $\varphi^+_{\Omega_\varepsilon, I_\varepsilon}(\xi, \varphi) = \varphi'(\varphi, \xi) \in J^{\sigma(i),j} \subset J^{\sigma(i)}$. Thus, the set of values $(\xi, \varphi) \in I_\varepsilon \times J^i$ such that $\varphi^+_{\Omega_\varepsilon, I_\varepsilon}(\xi, \varphi) \notin J^{\sigma(i)}$ is contained in the set $\big(I_\varepsilon \times J^i\big) \setminus \big(\tilde{I}_\varepsilon \times \tilde{J}^i_\varepsilon\big)$, whose measure is vanishing as $\varepsilon \to 0$.

(b) If $2 \le i = \sigma(i) \le m-1$, then the corresponding part of the boundary is the arc of circumference of radius 2 with the center at O occupying the angular interval J^i, that is, the set $\{2(\sin\varphi, \cos\varphi),\ \varphi \in J^i\}$. Next, we will show that for all values $(\xi, 0) \in I_\varepsilon$, $\varphi \in J^i$, except for a portion of order $o(1)$, the corresponding billiard particle makes one reflection from the arc and then goes back to I_ε in the direction $\varphi' \in J^i$.

For all values $\varphi \in J^i$, except for the union of two intervals of vanishing length (each of the intervals is contained in J^i, has the length $2\arctan(a_\varepsilon/4)$, and contains an endpoint of J^i), the particle starting at $(\xi, 0) \in I_\varepsilon$ in the direction φ will reflect from the indicated arc of circumference. Let $\psi \in J^i$ be the angular coordinate of the reflection point. By $(\xi', 0)$ denote the point at which the reflected particle intersects the straight line l. One easily verifies that

$$\frac{1}{\xi} + \frac{1}{\xi'} = \cos\psi. \tag{29}$$

One has

$$|\xi| < a_\varepsilon, \tag{30}$$

and hence

$$\frac{1}{|\xi'|} = \left|\cos\psi - \frac{1}{\xi}\right| > \frac{1}{a_\varepsilon} - 1. \tag{31}$$

From (29) it follows that $|\xi + \xi'|/|\xi\xi'| = |\cos\psi| \le 1$, and, taking into account (30) and (31), one finds that $|\xi + \xi'| < a_\varepsilon^2/(1 - a_\varepsilon)$. This implies that for all values $(\xi, 0) \in I_\varepsilon$, except for a set of measure $O(a_\varepsilon^2)$, the second point of intersection of the billiard trajectory belongs to I_ε; moreover, the velocity at this point, $\varphi^+_{\Omega_\varepsilon, I_\varepsilon}(\xi, \varphi)$, belongs to $\mathcal{N}_{2\arctan(a_\varepsilon/4)}(J^i)$, the neighborhood of J^i of radius $2\arctan(a_\varepsilon/4)$. This finally implies that for all $(\xi, \varphi) \in I_\varepsilon \times J^i$, except for a portion of order $O(a_\varepsilon)$, it holds that $\varphi^+_{\Omega_\varepsilon, I_\varepsilon}(\xi, \varphi) \in J^i$.

(c) The parts of the hollow's boundary, corresponding to J^1 and J^m, are formed by smooth curves joining the corresponding endpoints of I_ε and the points $2(\sin\Phi_0, -\cos\Phi_0)$ and $2(\sin\Phi_0, \cos\Phi_0)$, respectively. The unique condition on these curves is that they can be parametrized by the monotonically increasing angular coordinate. For those values $\sigma(1), \sigma(m)$ that coincide with neither 1 nor m take just the arcs of circumference of radius 2 corresponding to the angular intervals $J^{\sigma(1)}, J^{\sigma(m)}$.

Consider the union of all of the elliptic arcs $\mathcal{E}_{i,j}, \mathcal{E}'_{i,j}$ introduced in item (a), all of the arcs of circumference defined in items (a) and (b), and the two curves introduced in this item (c). Let us call this union the *main element*. Each lateral reflector is a curve; select it in such a way that both its endpoints belong to the main element. Finally, the curve $\partial\Omega_\varepsilon \setminus I_\varepsilon$ is the union of all of the lateral reflectors and the part of the main element visible from O (that is, which is not shielded by the adjacent lateral reflectors). Thus, the definition of the hollow $(\Omega_\varepsilon, I_\varepsilon)$ is complete.

In Figure 10, there is shown a particular hollow $(\Omega_\varepsilon, I_\varepsilon)$ corresponding to the permutation $\sigma = \left(\begin{smallmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 4 & 3 & 2 & 1 \end{smallmatrix}\right)$. The angular intervals $J^1, \dots, J^5$ are separated by dotted lines. The family of hollows $(\Omega_\varepsilon, I_\varepsilon)$, with vanishingly small ε, has the following property: for almost all particles with the initial direction from J^2 (resp. J^3, J^4), the final direction will belong to J^4 (resp. J^3, J^2). In the figure, there is shown the trajectory of a particle with the initial direction $\varphi \in J^2$ and the final direction $\varphi^+ \in J^4$. The particle makes a reflection from an elliptic arc, then two reflections from (very small) parabolic arcs, and, finally, again from an elliptic arc. According to our notation, these arcs are $\mathcal{E}_{2,2}$, $\mathcal{P}_{2,2}$, $\mathcal{P}'_{2,2}$, and $\mathcal{E}'_{2,2}$.

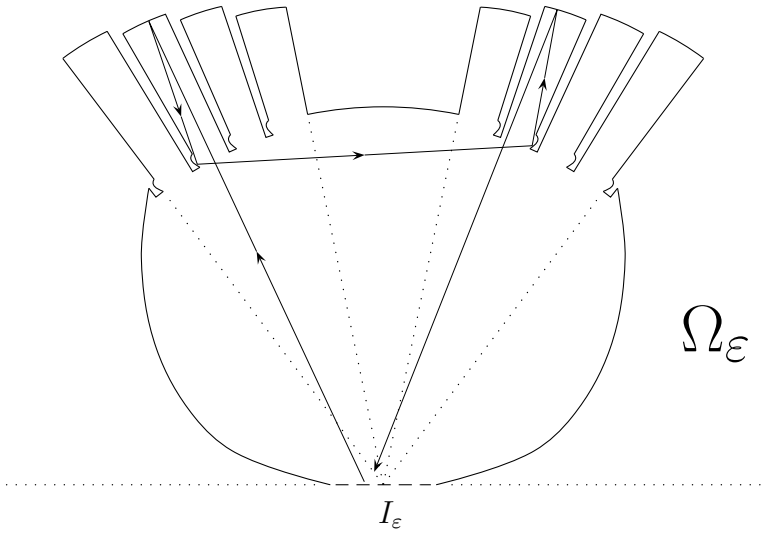


FIG. 10. *A hollow* $(\Omega_\varepsilon, I_\varepsilon)$ *approximating a* σ*-measure.*

4. Concluding remarks and applications. Physical bodies in the real world have atomic structure and therefore are disconnected. This is a reason for using (generally) disconnected sets Q_m in the definition of a rough body. In the future we intend to turn to propose and study the notion of a three-dimensional rough body, where the connectivity assumption is absolutely useless; this is another reason. By removing this assumption, the consideration in two dimensions (namely, the proof of Lemma 1) is made somewhat more difficult, but at the same time prerequisites for passing to the three-dimensional case are created.

In fact, the notions of "disconnected" (as everywhere in this paper) and "connected" rough bodies are equivalent. There is a natural one-to-one correspondence between the equivalence classes in the connected and disconnected cases,[3] the former classes being subclasses of the latter ones under this correspondence.

Let us now consider applications of the main theorem to problems of the body of minimal or maximal aerodynamic resistance. A two-dimensional convex body B moves, at constant velocity, through a rarefied homogeneous medium in $\mathbb{R}^2$, and at the same time it slowly rotates. The rotation is generally nonuniform; we assume that, during a sufficiently long observation period, in a reference system connected with the body the body's velocity is distributed in S^1 according to a given density function ρ, with $\int_{S^1} \rho(v)\,dv = 1$. The medium particles do not mutually interact, and collisions of the particles with the body are absolutely elastic. The resistance of the medium to the motion of the body is a vector-valued function of time. After averaging it over a sufficiently long period of time, one gets a vector. We are interested in the projection of this vector onto the direction of motion; for the sake of brevity, it will be called mean resistance, or just resistance. The problem is as follows: given B, determine the roughness on it in such a way that the main resistance of the resulting rough body is minimal or maximal.

A prototype of such a mechanical system is an artificial satellite of the Earth on relatively low altitudes ($100 \div 200$ km), with restricted capacity of rotation angle control. The satellite's motion is slowing down by the rest of the atmosphere; the problem is to minimize or maximize the effect of slowing down. The problems of resistance *maximization* may also arise when considering solar sail: a spacecraft driven by the pressure of solar photons.

The initial velocity of an incident particle (in the reference system connected with the body) is $-v$, and the final velocity is v^+; therefore, the momentum transmitted by the particle to the body is $v + v^+$. The projection of the transmitted momentum onto the direction of motion of the body equals $1 + \langle v, v^+\rangle$. Averaging this value over all particles incident on the body within a sufficiently long time interval, one gets the mean resistance. The averaging amounts to integration over $\rho(v)\,d\nu_{\mathcal{B}}(v, v^+, n)$; that is, the mean resistance of the rough body equals

$$R(\nu_{\mathcal{B}}) = \iiint_{\mathbb{T}^3} (1 + \langle v, v^+\rangle)\,\rho(v)\,d\nu_{\mathcal{B}}(v, v^+, n).$$

Using the main theorem and Fubini's theorem, one rewrites this formula in the form

$$R(\nu_{\mathcal{B}}) = \int_{S^1} d\tau_B(n) \iint_{\mathbb{T}^2} (1 + \langle v, v^+\rangle)\,\rho(v)\,d\eta_{\mathcal{B},n}(v, v^+), \tag{32}$$

[3]More precisely, we mean equivalence classes formed by sequences of connected/disconnected sets.

where $\eta_{\mathcal{B},n} \in \Lambda_n$. Thus, the minimization problem for $R(\nu_{\mathcal{B}})$ reduces to minimization, for any n, of the functional $\iint_{\mathbb{T}^2} (1+\langle v, v^+\rangle)\,\rho(v)\,d\eta(v,v^+)$ over all $\eta \in \Lambda_n$. Using the notation introduced in section 3.3, one comes to the problem:

$$\inf_{\eta\in\Lambda} \iint_{\square} (1+\cos(\varphi-\varphi^+))\,\varrho(\varphi)\,d\eta(\varphi,\varphi^+), \tag{33}$$

where $\varrho(\varphi) = \rho(v)$ for $\varphi = \operatorname{Arg} v - \operatorname{Arg} n$. This problem, in turn, by symmetrization of the cost function reduces to a particular Monge–Kantorovich problem:

$$\inf_{\eta\in\Lambda_{\lambda,\lambda}} \mathcal{F}(\eta), \quad \text{where} \quad \mathcal{F}(\eta) = \iint_{\square} c(\varphi,\varphi^+)\,d\eta(\varphi,\varphi^+), \tag{34}$$

where $c(\varphi,\varphi^+) = (1+\cos(\varphi-\varphi^+))\,\frac{\varrho(\varphi)+\varrho(\varphi^+)}{2}$ and $\Lambda_{\lambda,\lambda}$ is the set of measures η on $\square$ having both marginal measures equal to λ: $\pi^{\#}_{\varphi}\eta = \lambda = \pi^{\#}_{\varphi^+}\eta$. Recall that λ is defined by $d\lambda(v) = \cos\varphi\,d\varphi$.

Problem (34) can be exactly solved in several particular cases. Consider the case of uniform motion, where the function ρ, and therefore ϱ, is constant, and thus one can take $c(\varphi,\varphi^+) = \frac{3}{8}\,(1+\cos(\varphi-\varphi^+))$.[4] Note that $\mathcal{F}(\eta_0) = 1$ and therefore resistance of the smooth body is equal to its perimeter: $R(\nu_B) = \int_{S^1} d\tau_B(n)\,\mathcal{F}(\eta_0) = |\partial B|$. (Recall that the measure η_0 belongs to Λ and is supported on the diagonal $\varphi^+ = -\varphi$.) The minimization problem (34) for constant ϱ was solved in [14]: one has $\inf_{\mathcal{B}} R(\nu_{\mathcal{B}}) = 0.9878\ldots\cdot|\partial B|$, the infimum being taken over all roughenings of B.

Note that the corresponding *maximization* problem for (34) has the trivial solution, which does not depend on the function ϱ: $\eta = \eta_\star$, the measure $\eta_\star \in \Lambda$ being supported on the diagonal $\varphi^+ = \varphi$. One has $\sup_{\mathcal{B}} R(\nu_{\mathcal{B}}) = \kappa\,|\partial B|$, where $\kappa = \bigl(\int_{-\pi/2}^{\pi/2} \varrho(\varphi)\cos\varphi\,d\varphi\bigr)\big/\bigl(\int_{-\pi/2}^{\pi/2} \varrho(\varphi)\,\cos^3\varphi\,d\varphi\bigr) > 1$; in the case of uniform rotation one has $\kappa = 1.5$. The maximization problem was studied in more detail in [20].

Appendix A. The construction below is simple (see Figure 11), but its description is a bit cumbersome.

Take a point in the interior of B and connect it by segments with all vertices. The polygon is thus divided into several triangles; fix i and m and consider the triangle with the base $\mathfrak{b}_i$, the ith side of B. Denote by $d(\Omega_i^m)$ the diameter of the orthogonal projection of Ω_i^m onto the straight line containing I_i^m; one obviously has $d(\Omega_i^m) \geq |I_i^m|$. Fix a positive number $\kappa < |I_i^m|/d(\Omega_i^m)$.

Take a rectangle Π^1 contained in the triangle and such that one side of Π^1 belongs to $\mathfrak{b}_i$. By δ_1 denote the total length of the part of $\mathfrak{b}_i$ which is not occupied by this side.

For the sake of brevity, the image of a set under the composition of a homothety with positive ratio and a translation will be called a copy of this set. Take several copies of Ω_i^m (copies of first order) that do not mutually interact, that belong to Π^1, whose corresponding copies of I_i^m belong to $\mathfrak{b}_i$, and whose portion of the side of Π^1 occupied by them is more than κ.

Next, take several rectangles that do not mutually intersect and do not intersect with the chosen copies of Ω_i^m, belong to Π^1, and have one side contained in $\mathfrak{b}_i$. Denote by Π^2 the union of these rectangles and by δ_2 the total length of the part of the side of Π^1 which is not occupied by the rectangles from Π^2 and by the copies of I_i^m. Next,

[4] The normalization constant 3/8 is taken for further convenience.

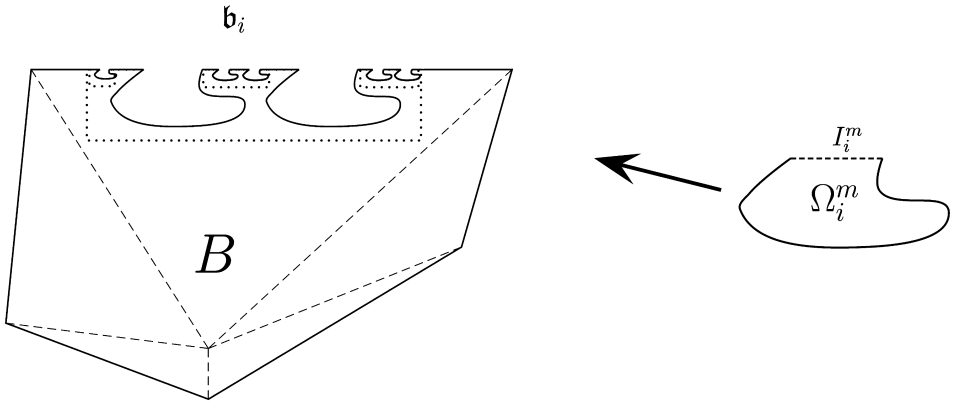


FIG. 11. *Making hollows on the ith side of a polygon.*

for each rectangle from Π^2 choose several copies of Ω_i^m (copies of second order) in the way completely similar to the described above (see Figure 11).

Continuing this process, one obtains a sequence $\Pi^1, \Pi^2, \dots$ of unions of rectangles and collections of copies of Ω_i^m of first, second, ... order. Choose the rectangles in such a way that $\delta_1 + \delta_2 + \dots < 1/m$ and $\mathrm{Area}(\Pi^1) < 1/m$. Finally, choose k such that the total length of sides of rectangles from Π^{k+1} contained in $\mathfrak{b}_i$ is less than $1/m$, and take the collection of copies of Ω_i^m of order $1, 2, \dots, k$ (we shall call it full collection). The total length of the part of $\mathfrak{b}_i$ not occupied by the corresponding copies of I_i^m is less than $2/m$, and therefore it goes to zero as $m \to \infty$.

By definition, the desired set Q_m is B minus the union of full collections of copies of Ω_i^m over all i.

Appendix B. We prove here slightly more than needed.

STATEMENT 3. *Let $A = (a_{ij})_{i,j=1}^k$ be a symmetric matrix, with a_{ij} being non-negative integers. Denote $n_i = \sum_{j=1}^k a_{ij}$. Then there exist matrices $B_{ij} = (b_{ij}^{\mu\nu})_{\mu,\nu}$ of size $n_i \times n_j$ such that $b_{ij}^{\mu\nu} \in \{0, 1\}$, $B_{ij}^T = B_{ji}$, the sum of elements in B_{ij} equals a_{ij}, and the block matrix $D = (B_{ij})$ contains exactly one unit in each row and each column.*

Note that for some values $i = i_1, i_2, \dots$ it may happen that $n_i = 0$, that is, $a_{ij} = 0$ for all $j = 1, \dots, k$. Then the corresponding matrices B_{ij} have the size $0 \times n_j$, that is, are empty. In this case D coincides with the block matrix $D' = (B_{ij})$ having the rows $i_1, i_2, \dots$ and columns $i_1, i_2, \dots$ crossed out.

Proof. The proof is by induction on k. Let the statement be true for $k-1$; prove it for k. Take the matrix $\tilde{A} = (a_{ij})_{i,j=2}^k$; there exists a block matrix $\tilde{B} = (\tilde{B}_{ij})_{i,j=2}^k$ satisfying the statement. Note that the order of $\tilde{B}_{ij}$ is $\tilde{n}_i \times \tilde{n}_j$, where $\tilde{n}_i = \sum_{j=2}^k a_{ij} = n_i - a_{i1}$. Define the matrices B_{ij} as follows.

(a) Put $B_{11} = \mathrm{diag}\{\underbrace{1, \dots, 1}_{a_{11}}, 0, \dots, 0\}$.

(b) Put $b_{12}^{a_{11}+1,1} = \dots = b_{12}^{a_{11}+a_{12},a_{12}} = 1$; $b_{13}^{a_{11}+a_{12}+1,1} = \dots = b_{13}^{a_{11}+a_{12}+a_{13},a_{13}} = 1$; $\dots$; $b_{1k}^{a_{11}+\dots+a_{1,k-1}+1,1} = \dots = b_{1k}^{a_{11}+\dots+a_{1k},a_{1k}} = 1$; the other elements of the matrices B_{1j}, $j = 2, \dots, k$, are zeros. Thus, on the diagonal of B_{1j} starting from the element at the first column and the $(a_{11} + a_{12} + \dots + a_{1,j-1} + 1)$th row, the first a_{1j} elements equal 1, and the remaining elements on this diagonal and all of the elements

off the diagonal are zeros. This defines the matrices B_{1j}, $j = 2, \dots, k$. The matrices B_{i1}, $i = 2, \dots, k$, are determined by the condition $B_{i1} = B_{1i}^T$.

(c) For $i \geq 2$, $j \geq 2$ define the matrix B_{ij} as follows. For $\mu \leq a_{1i}$ or $\nu \leq a_{1j}$, put $b_{ij}^{\mu\nu} = 0$, and for $\mu \geq a_{1i} + 1$, $\nu \geq a_{1j} + 1$, put $b_{ij}^{\mu\nu} = \tilde{b}_{ij}^{\mu - a_{1i}, \nu - a_{1j}}$. Thus, in the obtained matrix B_{ij}, the right lower corner coincides with the matrix $\tilde{B}_{ij}$, and all of the remaining elements are equal to zero. The number of rows of this matrix equals $a_{1i} + \tilde{n}_i = n_i$, and the number of columns equals $a_{1j} + \tilde{n}_j = n_j$. One obviously has $B_{ij}^T = B_{ji}$.

One easily verifies that $\sum_{\mu\nu} b_{ij}^{\mu\nu} = a_{ij}$ and that each row and each column of the obtained block matrix $D = (B_{ij})_{i,j=1}^k$ contains precisely one unit. □

REFERENCES

[1] I. Newton, *Philosophiae Naturalis Principia Mathematica*, William Dawson & Sons, London, 1686.

[2] G. Buttazzo and B. Kawohl, *On Newton's problem of minimal resistance*, Math. Intelligencer, 15 (1993), pp. 7–12.

[3] G. Buttazzo, V. Ferone, and B. Kawohl, *Minimum problems over sets of concave functions and related questions*, Math. Nachr., 173 (1995), pp. 71–89.

[4] F. Brock, V. Ferone, and B. Kawohl, *A symmetry problem in the calculus of variations*, Calc. Var. Partial Differential Equations, 4 (1996), pp. 593–599.

[5] G. Buttazzo and P. Guasoni, *Shape optimization problems over classes of convex domains*, J. Convex Anal., 4 (1997), pp. 343–351.

[6] M. Belloni and B. Kawohl, *A paper of Legendre revisited*, Forum Math., 9 (1997), pp. 655–668.

[7] T. Lachand-Robert and M. A. Peletier, *Newton's problem of the body of minimal resistance in the class of convex developable functions*, Math. Nachr., 226 (2001), pp. 153–176.

[8] M. Comte and T. Lachand-Robert, *Newton's problem of the body of minimal resistance under a single-impact assumption*, Calc. Var. Partial Differential Equations, 12 (2001), pp. 173–211.

[9] M. Comte and T. Lachand-Robert, *Existence of minimizers for Newton's problem of the body of minimal resistance under a single-impact assumption*, J. Anal. Math., 83 (2001), pp. 313–335.

[10] D. Horstmann, B. Kawohl, and P. Villaggio, *Newton's aerodynamic problem in the presence of friction*, NoDEA Nonlinear Differential Equations Appl., 9 (2002), pp. 295–307.

[11] M. Belloni and A. Wagner, *Newton's problem of minimal resistance in the class of bodies with prescribed volume*, J. Convex Anal., 10 (2003), pp. 491–500.

[12] A. Yu. Plakhov, *Newton's problem of a body of minimal aerodynamic resistance*, Dokl. Akad. Nauk, 390 (2003), pp. 314–317.

[13] A. Yu. Plakhov, *Newton's problem of the body of minimal resistance with a bounded number of collisions*, Russian Math. Surveys, 58 (2003), pp. 191–192.

[14] A. Yu. Plakhov, *Newton's problem of the body of minimum mean resistance*, Sb. Math., 195 (2004), pp. 1017–1037.

[15] A. Yu. Plakhov and D. F. M. Torres, *Newton's aerodynamic problem in media of chaotically moving particles*, Sb. Math., 196 (2005), pp. 885–933.

[16] R. G. Barantsev, *Interaction of Rarefied Gases with Streamline Surfaces*, Mir, Moscow, 1975 (in Russian).

[17] D. Blackmore and J. G. Zhou, *A general fractal distribution function for rough surface profiles*, SIAM J. Appl. Math., 56 (1996), pp. 1694–1719.

[18] O. A. Aksenova and I. A. Khalidov, *Fractal and statistical models of rough surface interacting with rarefied gas flow*, in AIP Conference Proceedings, RAREFIED GAS DYNAMICS: 24th International Symposium on Rarefied Gas Dynamics, Vol. 762, American Institute of Physics, College Park, MD, 2005, pp. 993–998.

[19] N. Chernov, *Entropy, Lyapunov exponents, and mean free path for billiards*, J. Statist. Phys., 88 (1997), pp. 1–29.

[20] A. Plakhov and P. Gouveia, *Problems of maximal mean resistance on the plane*, Nonlinearity, 20 (2007), pp. 2271–2287.

SIAM J. MATH. ANAL.
Vol. 40, No. 6, pp. 2179–2216

PERMANENCE AND ASYMPTOTICALLY STABLE COMPLETE TRAJECTORIES FOR NONAUTONOMOUS LOTKA–VOLTERRA MODELS WITH DIFFUSION*

JOSÉ A. LANGA†, JAMES C. ROBINSON‡, ANÍBAL RODRÍGUEZ-BERNAL§, AND ANTONIO SUÁREZ¶

Abstract. Lotka–Volterra systems are the canonical ecological models used to analyze population dynamics of competition, symbiosis, or prey-predator behavior involving different interacting species in a fixed habitat. Much of the work on these models has been within the framework of infinite-dimensional dynamical systems, but this has frequently been extended to allow explicit time dependence, generally in a periodic, quasiperiodic, or almost periodic fashion. The presence of more general nonautonomous terms in the equations leads to nontrivial difficulties which have stalled the development of the theory in this direction. However, the theory of nonautonomous dynamical systems has received much attention in the last decade, and this has opened new possibilities in the analysis of classical models with general nonautonomous terms. In this paper we use the recent theory of attractors for nonautonomous PDEs to obtain new results on the permanence and the existence of forwards and pullback asymptotically stable global solutions associated to nonautonomous Lotka–Volterra systems describing competition, symbiosis, or prey-predator phenomena. We note in particular that our results are valid for prey-predator models, which are not order-preserving: even in the "simple" autonomous case the uniqueness and global attractivity of the positive equilibrium (which follows from the more general results here) is new.

Key words. Lotka–Volterra competition, symbiosis and prey-predator systems, nonautonomous dynamical systems, permanence, attracting complete trajectories

AMS subject classifications. 35B40, 35K55, 92D25, 37L05

DOI. 10.1137/080721790

1. Introduction.

Partial differential equations have proved a very useful tool in the modelling of many ecological phenomena related to the dynamics between species interacting in a given habitat. Many authors have allowed explicit dependence on both space and time in the parameters of the equation, a natural way to take into account the spatial and temporal variations that influence real species interactions.

In this paper we consider a nonautonomous model for two species (u and v), evolving within a habitat Ω that is a bounded domain in $\mathbf{R}^N$, $N \geq 1$, with a smooth

*Received by the editors April 21, 2008; accepted for publication September 13, 2008; published electronically January 28, 2009.
http://www.siam.org/journals/sima/40-6/72179.html

†Dpto. Ecuaciones Diferenciales y Análisis Numérico. C/ Tarfia s/n, 41012. Sevilla. Spain (langa@us.es). Partly supported by Ministerio de Educación y Ciencia (Spain) under grants MTM2005-01412, Consejería de Innovación, Ciencia y Empresa (Junta de Andalucía, Spain) under the Proyecto de Excelencia FQM-02468.

‡Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK (j.c.robinson@warwick.ac.uk). Partially supported by a Royal Society University Research Fellowship.

§Dpto. de Matemática Aplicada, Universidad Complutense de Madrid, Madrid 28040, Spain (arober@mat.ucm.es). Partially suported by Project MTM2006–08262, DGES, Spain and Grupo de Investigación UCM-CAM 920894, CADEDIF.

¶Dpto. Ecuaciones Diferenciales y Análisis Numérico. C/ Tarfia s/n, 41012, Sevilla, Spain (suarez@us.es). Partly supported by Ministerio de Educación y Ciencia (Spain) under grants MTM2006-07932, Consejería de Innovación, Ciencia y Empresa (Junta de Andalucía, Spain) under the Proyecto de Excelencia FQM-520.

boundary $\partial\Omega$, of the following type:

$$\begin{cases} u_t - d_1\Delta u = uf(t,x,u,v) & x\in\Omega,\ t>s \\ v_t - d_2\Delta v = vg(t,x,u,v) & x\in\Omega,\ t>s \\ \mathcal{B}_1 u = 0,\ \mathcal{B}_2 v = 0 & x\in\partial\Omega,\ t>s \\ u(s)=u_s,\ v(s)=v_s, \end{cases} \tag{1.1}$$

where f and g are regular functions, d_1, d_2 are positive constants, and $\mathcal{B}_i$ denotes one of the boundary operators

$$\mathcal{B}u = u, \quad \text{or} \quad \mathcal{B}u = \frac{\partial u}{\partial \vec{n}}, \quad \text{or} \quad \mathcal{B}u = d\frac{\partial u}{\partial \vec{n}} + \sigma(x)u, \tag{1.2}$$

for the Dirichlet, Neumann, or Robin case, respectively, $\vec{n}$ is the outward normal vector-field to $\partial\Omega$, and $\sigma(x)$ a C^1 function. Note that we take diffusion coefficient d_i and boundary potential $\sigma_i(x)$ for the case of Robin boundary condition $\mathcal{B}_i$. Also note that we allow all of the nine possible combinations of boundary conditions in (1.1).

A particularly interesting class of models of the form (1.1) are the nonautonomous Lotka–Volterra models:

$$\begin{cases} u_t - d_1\Delta u = u(\lambda(t,x) - a(t,x)u - b(t,x)v) & x\in\Omega,\ t>s \\ v_t - d_2\Delta v = v(\mu(t,x) - c(t,x)u - d(t,x)v) & x\in\Omega,\ t>s \\ \mathcal{B}_1 u = 0,\ \mathcal{B}_2 v = 0 & x\in\partial\Omega,\ t>s \\ u(s)=u_s,\ v(s)=v_s. \end{cases} \tag{1.3}$$

We refer, for example, to [6] for the biological meaning of the parameters d_1, d_2, λ, μ, a, b, c, d involved in (1.3).

In line with the ecological interpretation of these models, we will consider only positive solutions, and in the light of this we note here that $u_s, v_s \geq 0$ implies that the solution of (1.1) satisfies $u, v \geq 0$.

Note that our hypotheses on b and c allow different models of population dynamics: competition if $b, c > 0$, symbiosis if $b, c < 0$, and prey-predator if $b > 0$ and $c < 0$, although we do not allow sign-changing coefficients.

Of course, it is an important problem to determine the asymptotic behavior of solutions of the system (1.1). Since in general this is a very complicated task, one may try to solve simpler problems; e.g., one can try to determine whether or not the two species will survive in the long term or if, on the contrary, one of them will be driven to extinction. Survival of the species has been formalized in the notion of *permanence*; see Hale and Waltman [15] or Hutson and Schmitt [20]. Loosely speaking, the system (1.1) is said to be *permanent* if for any positive initial data u_s and v_s, within a finite time the values of the solution $(u(t,s,x;u_s,v_s), v(t,s,x;u_s,v_s))$, for $x\in\Omega$, enter and remain within a compact set in $\mathbf{R}^2$ that is strictly bounded away from zero in each component. Note, however, that this is an imprecise statement in the presence of Dirichlet boundary conditions.

Note that permanence is a form of *coexistence* of the species, since none is extinguished at any part of the habitat domain at any time.

A related situation, which implies that the system is permanent but gives more detail since it also indicates the expected final state of the system, is when there exists a solution, bounded away from zero, to which all other solutions tend asymptotically.

These two are the main topics with which we are concerned in this paper.

Before going further observe that both (1.1) and (1.3) always possess the *trivial* solution $(0,0)$ and *semitrivial* solutions of the form $(u,0)$ and $(0,v)$. In the latter

case the nontrivial component satisfies a scalar parabolic problem, of logistic type in the case of (1.3). The dynamics of these solutions have a deep impact on the global dynamics of general solutions. Indeed, if the system is permanent, this implies that semitrivial solutions must be unstable in some sense. On the other hand, if semitrivial solutions are stable, then it can be expected that some solutions of the system exhibit *extinction*, that is, one of the species (or both) approaches asymptotically the value zero.

Some results are already known along these lines. For example, in the autonomous case, assume that all the coefficients in (1.3) are constants and consider, for example, the problem with Dirichlet boundary conditions. In this case results about permanence for problem (1.3) depend on the values of λ and μ with respect to the first eigenvalue of certain associated linear elliptic problems, which we now describe. Given $d \in \mathbf{R}$, $d > 0$, and $f \in L^\infty(\Omega)$, we denote by $\Lambda(d, f)$ (we write $\Lambda_0 := \Lambda(d, 0)$) the first eigenvalue of the problem

$$\begin{cases} -d\Delta w = \sigma w + f(x)w & \text{in } \Omega, \\ w = 0 & \text{on } \partial\Omega, \end{cases}$$

and given $\gamma, \alpha \in \mathbf{R}$ with $\alpha > 0$, we denote by $\omega_{[d,\gamma,\alpha]}$ the unique positive solution of

$$\begin{cases} -d\Delta w = \gamma w - \alpha w^2 & \text{in } \Omega, \\ w = 0 & \text{on } \partial\Omega. \end{cases}$$

If λ and μ satisfy

$$\lambda > \Lambda(d_1, -b\omega_{[d_2,\mu,d]}) \quad \text{and} \quad \mu > \Lambda(d_2, -c\omega_{[d_1,\lambda,a]}), \tag{1.4}$$

then the autonomous version of the competition or prey-predator cases of (1.3), with Dirichlet boundary conditions in both components, are permanent and moreover there exists a positive equilibrium solution (Cantrell et al. [4], [6], [7], [8] and López-Gómez [27]).

Although the case of symbiosis, $b, c < 0$, is not treated in these papers, a similar result holds provided that

$$bc < ad,$$

a condition which is used to obtain *a priori bounds* for the solutions (see, for instance, Pao [30] or Theorem 9.8 in Delgado et al. [12], where moreover the coefficients a, b, c, and d depend on x).

Note that (1.4) is a condition that expresses the instability of semitrivial solutions.

However, in the competition case it is well known that if $\lambda \le \Lambda_0$ or $\mu \le \Lambda_0$, then one of the two species (or both of them) will be driven to extinction (see López-Gómez and Sabina [29] for an improvement of this result). Similar results can be obtained in the other cases; see [6] and [30]. Note that, in contrast with (1.4), the condition above expresses the stability of either one of the semitrivial solutions.

When nonautonomous terms are allowed in the equations, this is usually done under the assumption of periodicity, quasiperiodicity, or almost periodicity, and in this case similar results can be obtained to those for autonomous equations (see Hess [17], Hess and Lazer [18], Hetzer and Shen [19], and references therein). For the case of periodic coefficients, the use of the Poincaré map implies that the system resembles an autonomous one in many respects.

Cantrell and Cosner [5] assume general nonautonomous terms that are bounded by periodic functions, and using a comparison method give conditions on λ and μ that guarantee that (1.3) is permanent.

Note that most of the references cited in the papers above are concerned (besides periodicity or almost periodicity) with some particular choice of boundary conditions (typically Neumann, or even Dirichlet, in both components) and one of the competition, symbiosis or prey-predator cases. In the first two cases a common tool in the references is the use of order-preserving properties of the Lotka–Volterra system.

For example, in the case of almost periodic time dependence, Hetzer and Shen [19] proved similar results for the competition case, assuming that $d_1 = d_2$ and $\lambda = \mu$ are constant and both components of the system satisfy Dirichlet boundary conditions (no such restrictions are required in the case of Neumann BCs). In that paper, the limitation to almost periodic cases is due to the use of skew-product techniques which require, some way or another, some sort of time recurrence in the coefficients of the system.

Note that in [25] Langa et al. studied permanence for the competition case with Dirichlet boundary conditions when only the coefficient a is allowed to depend on time.

In this paper we allow general nonautonomous terms and do not restrict ourselves to (for example) almost periodic time dependence. As said before, we also consider all nine possible choices of boundary conditions and treat competition, symbiosis, and prey-predator models, since we do not rely on monotonicity properties of the system. Note that the only restriction that we impose on the coefficients is that $d_1 = d_2$ in the symbiotic case, a condition that we assume only in order to have explicit upper bounds on the solutions, but not for the permanence results. Also note that as we employ for the solutions of (1.1) or (1.3) the approach of *nonautonomous processes* rather than skew-product techniques, we have to pay attention to both the initial time, s, and the observation time for the solutions, $t > s$. This implies that concepts like permanence, stability, instability, and attractivity can be defined and analyzed in both pullback and forwards senses; see section 2 for further details and also [24]. Observe also that while pullback properties (e.g., permanence, attraction) are usually the most one can expect for general nonautonomous terms, in this case we can also show results on permanence and attractivity forwards in time; see Langa et al. [25, 23] for cases of pullback but not forwards permanence or attraction in nonautonomous reaction-diffusion equations.

In section 3, using results for the scalar nonautonomous logistic equations from, e.g., [25, 34], which we compile in section 3.1, we make use of the theory of attractors for nonautonomous PDEs as developed by Chepyzhov and Vishik [9] (see also Crauel et al. [11] or Kloeden and Schmalfuss [21]). Thus, we prove in section 3.2 that under the assumptions

$$\inf_{\mathbf{R}\times\Omega} a(t,x) > 0 \quad \text{and} \quad \inf_{\mathbf{R}\times\Omega} d(t,x) > 0$$

the system (1.3) has a nonautonomous attractor; see Theorem 3.5. The existence of a nonautonomous attractor in this case implies the presence of bounded complete trajectories, i.e., solutions defined for all time.

From here we derive in section 3.3 some sufficient conditions for the extinction of one (or both) of the species of the system. These conditions are far from optimal but qualitatively describe the stability of semitrivial solutions; see Proposition 3.6.

Then, in section 3.4 we give sufficient conditions reflecting the instability of semitrivial solutions that guarantee that (1.3) is permanent both in a pullback and in

a forwards sense. We want to stress here that these sufficient conditions involve only information about the behavior of the coefficients of the system as either $t \to -\infty$ or $t \to \infty$. Also, they are given in such a way that the result is robust with respect to perturbations of the coefficients.

The rest of the paper is then devoted to a more detailed analysis of the asymptotic behavior of the solutions of (1.3). After some preparatory material in sections 4 and 5, we will prove in section 6 that under appropriate conditions on the parameters all nonsemitrivial solutions of (1.3) have the same asymptotic behavior as $t \to \infty$. In particular all bounded complete trajectories in the nonautonomous attractor have the same asymptotic behavior as $t \to \infty$. For this we make use of the permanence results in section 3.4 and impose a smallness condition on the product of the coupling parameters:

$$\limsup_{t\to\infty} \|b\|_{L^\infty(\Omega)} \limsup_{t\to\infty} \|c\|_{L^\infty(\Omega)} < \rho_0$$

for some suitable constant $\rho_0 > 0$; see Theorem 6.1.

Moreover we show that, under a similar smallness condition on the coupling coefficients, now as $t \to -\infty$, if one of the bounded complete trajectories of (1.3) (which exists from the existence of the nonautonomous attractor) is bounded away from zero at $-\infty$, it is the unique such trajectory, and it also describes the unique pullback asymptotic behavior of all nonsemitrivial solutions of (1.3); see Theorem 6.2. When these two theorems can be applied together, there is a unique bounded complete trajectory $(u^*(t), v^*(t))$ that is both forwards and pullback attracting for (1.3); i.e., (u^*, v^*) is a bounded trajectory such that, for any $s \in \mathbf{R}$ and for any positive solution $(u(t,s), v(t,s))$ of (1.3) defined for $t > s$, one has

$$(u(t,s) - u^*(t), v(t,s) - v^*(t)) \to (0,0) \qquad \text{as} \quad t \to \infty, \text{ or } \quad s \to -\infty. \tag{1.5}$$

To obtain these results we need some nontrivial machinery for the linear scalar case, section 4.1, and some perturbation results about the exponential decay for solutions of linear parabolic nonautonomous systems, section 4.2. In particular, we find conditions guaranteeing that any bounded solution of

$$\begin{cases} u_t - d_1 \Delta u = p(t,x)u \\ v_t - d_2 \Delta v = q(t,x)v \end{cases} \tag{1.6}$$

gives rise to a solution that tends to zero as $t \to \infty$, when (1.6) is perturbed in a certain way; see Theorem 4.6. It is because we are able to study the linear part of the system in detail that we can obtain results for the nonlinear system.

Since we are able to treat the difference of two solutions of problem (1.3) within this framework, as a consequence of this argument we can apply our results to the Lotka–Volterra model in all three standard cases: competition, symbiosis, and prey-predator. It is noteworthy that these different situations are usually studied separately in the literature, but since we do not make any use of monotonicity arguments (which do not apply in the prey-predator case) we are able to give a unified treatment.

We close this paper in section 7 with a discussion of our results and some possibilities for further developments.

In the case in which all the coefficients are autonomous or periodic, our results in section 6 that we described above in (1.5) imply the uniqueness of the asymptotic behavior of all nonsemitrivial solutions.

Hence, in the autonomous case our results agree with all the classical results of uniqueness and stability of the nonsemitrivial steady states of (1.3) for the three cases of competition, symbiosis, and prey-predator (see, for instance, Theorem 4.4 in Furter and López-Gómez [13] and Corollary 4.3 in López-Gómez and Sabina de Lis [29] in the competition case, and Corollary 9.5 in Delgado et al. [12] in the symbiosis case).

Moreover, in the prey-predator case, with (1.5) we are able to conclude the uniqueness and *global stability* of a steady state, solving (for particular ranges of parameter values) one of the most interesting open problems in this field. We emphasize that this result is new even in the autonomous case, where until now only local stability has been proved; see Theorem 4.1 in Leung [26], see also Lakos [22], López-Gómez and Pardo [28], and Yamada [36].

2. Some notations and preliminaries. In this section we introduce some basic notations and terminology that will be used throughout the rest of the paper. In particular, we make precise the way systems (1.1) or (1.3) are said to be permanent.

2.1. Asymptotic behavior and complete trajectories for nonlinear systems. Note that if the solutions of (1.1) are global, then we can define a nonautonomous nonlinear *process* in some Banach space X appropriate for the solutions, i.e., a family of mappings $\{S(t,s)\}_{t\geq s} : X \to X$, $t, s \in \mathbf{R}$ satisfying:

(a) $S(t,s)S(s,\tau)z = S(t,\tau)z$, for all $\tau \leq s \leq t$, $z \in X$,

(b) $S(t,\tau)z$ is continuous in $t > \tau$ and z, and

(c) $S(t,t)$ is the identity in X for all $t \in \mathbf{R}$.

$S(t,\tau)z$ arises as the value of the solution of our nonautonomous system at time t with initial condition z at initial time τ. For an autonomous system the solutions depend only on $t-\tau$, and we can write $S(t,\tau) = S(t-\tau,0)$.

In order to describe the asymptotic behavior of nonautonomous systems like (1.1) and (1.3), we rely on the concept of a nonautonomous pullback attractor (Chepyzhov and Vishik [9], Kloeden and Schmalfuss [21]), which is the sensible generalization of an attractor for nonautonomous systems. For $A, B \subset X$ we denote the Hausdorff semidistance between A and B by $\operatorname{dist}(A,B) = \sup_{a\in A} \inf_{b\in B} d(a,b)$.

DEFINITION 2.1. *We say that a family of compact sets $\{\mathcal{A}(t)\}_{t\in\mathbf{R}} \subset X$ is a pullback attractor associated to S if*

(a) *$S(t,\tau)\mathcal{A}(\tau) = \mathcal{A}(t)$, for all $t \geq \tau$ and*

(b) *for all $t \in \mathbf{R}$ and $D \subset X$ bounded*

$$\lim_{\tau\to-\infty} \mathit{dist}(S(t,\tau)D, \mathcal{A}(t))) = 0.$$

Observe that the attraction in (b) fixes the final time and moves the initial time backwards towards $-\infty$. We are not evolving one trajectory backwards in time, but rather we consider the current state of the system (at the fixed time t) which would result from the same initial condition starting at earlier and earlier times.

To guarantee the existence of such a pullback attractor, one is usually faced with the task of proving the existence of a pullback absorbing family, defined as follows.

DEFINITION 2.2. *Given $t_0 \in \mathbf{R}$, we say that $B(t_0) \subset X$ is pullback absorbing at time t_0 if for every bounded $D \subset X$ there exists a $T = T(t,D) \in \mathbf{R}$ such that*

$$S(t_0,\tau)D \subset B(t_0), \text{ for all } \tau \leq T.$$

A family $\{B(t)\}_{t\in\mathbf{R}}$ is pullback absorbing if $B(t_0)$ is pullback absorbing at time t_0, for all $t_0 \in \mathbf{R}$.

The general result on the existence of nonautonomous pullback attractors is a generalization of the abstract theory for autonomous dynamical systems (Temam [37], Hale [14]).

THEOREM 2.3 (Crauel et al. [11], Schmalfuss [35]). *Assume that there exists a family of compact pullback absorbing sets. Then, there exists a pullback attractor* $\{\mathcal{A}(t)\}_{t\in\mathbf{R}}$ *that is minimal in the sense that if* $\{C(t)\}_{t\in\mathbf{R}}$ *is another family of closed pullback attracting sets, then* $\mathcal{A}(t)\subset C(t)$ *for all* $t\in\mathbf{R}$.

To have a more precise description of the dynamical objects within the pullback attractor, we make the following definition:

DEFINITION 2.4. *Let* S *be a process. We call the continuous map* $w:\mathbf{R}\to X$ *a complete trajectory if, for all* $s\in\mathbf{R}$,

$$S(t,s)w(s)=w(t)\qquad \text{for all}\qquad t\geq s.$$

According to Chepyzhov and Vishik [9], when the family of absorbing sets is uniformly bounded, the pullback attractor can be characterized as

$$\mathcal{A}(t)=\{w(t):\ w(\cdot)\text{ is a bounded complete trajectory for } S\}. \tag{2.1}$$

2.2. Pullback and forwards permanence for nonautonomous systems. Consider the nonlinear system (1.1) and assume that f and g are regular functions. Hence, we can assume that for initial data $(u_s,v_s)\in C_{\mathcal{B}_1}(\overline{\Omega})\times C_{\mathcal{B}_2}(\overline{\Omega})$ there exists a unique (local) smooth solution such that $(u,v)\in C^1_{\mathcal{B}_1}(\overline{\Omega})\times C^1_{\mathcal{B}_2}(\overline{\Omega})$ for $t>s$, where, for $j=0,1$,

$$C^j_{\mathcal{B}}(\overline{\Omega})=\begin{cases} C^j_0(\overline{\Omega}) & \text{for Dirichlet BCs,}\\ C^j(\overline{\Omega}) & \text{for Neumann or Robin BCs,}\end{cases}$$

with $C^j_0(\overline{\Omega})$ denoting functions in $C^j(\overline{\Omega})$ that are zero on $\partial\Omega$ and $C^0(\overline{\Omega})=C(\overline{\Omega})$.

Note that in practice we will be interested only in nonnegative solutions and that if $u_s\geq 0$ and $v_s\geq 0$ in (1.1), then the local solution satisfies $u,v\geq 0$. In fact, the maximum principle implies that if both $u_s\geq 0$ and $v_s\geq 0$ are nontrivial, then u and v are strictly positive in Ω.

Although at this point we assume only local existence of solutions, it still makes sense to consider complete trajectories of (1.1), which roughly speaking are solutions defined for all times. These objects will play a central role in our analysis below, as can be seen from (2.1). More precisely, a restatement of Definition 2.4 gives the following.

DEFINITION 2.5. *A continuous function* $U=\binom{u}{v}:\mathbf{R}\to C_{\mathcal{B}_1}(\overline{\Omega})\times C_{\mathcal{B}_2}(\overline{\Omega})$ *is a complete trajectory of* (1.1)*, if for all* $s<t$ *in* $\mathbf{R}$, $(u(t),v(t))$ *is the solution of* (1.1) *with initial data* $u_s=u(s)$, $v_s=v(s)$.

Now we define several concepts that will help us in making precise the concepts of pullback and forwards permanence for the solutions of (1.1) or (1.3). Note that the concepts below are related to the spaces $C_{\mathcal{B}_i}(\overline{\Omega})$ above. We start with the following.

DEFINITION 2.6. *A set of nonnegative functions* $B\subset C(\overline{\Omega})$ *is bounded away from zero if there exists a nonnegative nontrivial continuous function* $\varphi_0(x)\geq 0$ *in* Ω *(vanishing on* $\partial\Omega$ *in case of Dirichlet boundary conditions) such that*

$$u(x)\geq\varphi_0(x)\quad \text{for all}\quad x\in\Omega,\quad u\in B.$$

The set B *is nondegenerate if the function* $\varphi_0(x)$ *above is in* $C^1(\overline{\Omega})$ *and* $\varphi_0(x)>0$ *in* Ω.

Note that φ_0 above can be a positive constant in the case of Neumann or Robin boundary conditions.

Then we have the following definitions for curves in the space of continuous functions.

DEFINITION 2.7. *A positive function with values in $C(\overline{\Omega})$ is nondegenerate at ∞ (respectively, $-\infty$) if there exists $t_0 \in \mathbf{R}$ such that u is defined in $[t_0, \infty)$ (respectively, $(-\infty, t_0]$) and*

$$\{u(t),\ t \geq t_0\} \quad \textit{is a nondegenerate set}$$

(respectively, for $t \leq t_0$), that is, there exists a $C^1(\overline{\Omega})$ function $\varphi_0(x) > 0$ in Ω, (vanishing on $\partial\Omega$ in case of Dirichlet boundary conditions), such that

$$u(t,x) \geq \varphi_0(x) \qquad \textit{for all} \quad x \in \Omega, \quad t \geq t_0$$

(respectively, for all $t \leq t_0$).

A family of curves in $C(\overline{\Omega})$, denoted $\{u_\sigma(t)\}_{\sigma\in\Sigma}$, is nondegenerate at ∞ if there exists $t_0 \in \mathbf{R}$ such that u_σ is defined in $[t_0, \infty)$ and

$$\{u_\sigma(t),\ t \geq t_0,\ \sigma \in \Sigma\} \quad \textit{is a nondegenerate set.}$$

Finally, a family of curves in $C(\overline{\Omega})$, denoted $\{u_\sigma(t,s)\}_{\sigma\in\Sigma}$, defined in the intervals $[s, \infty)$ is nondegenerate as $s \to -\infty$ if there exists $s_0 \in \mathbf{R}$ such that for all $s \leq s_0$

$$\{u_\sigma(t),\ s \leq t \leq s_0,\ \sigma \in \Sigma\} \quad \textit{is a nondegenerate set.}$$

For systems, analogously to Definition 2.6, a set $B \subset (C(\overline{\Omega}))^2$ is bounded away from zero if each projection of B is bounded away from zero in $C(\overline{\Omega})$. In a similar way, as in Definition 2.7, a family of curves $U_\sigma(x,\cdot) \in (C(\overline{\Omega}))^2$, $\sigma \in \Sigma$, is nondegenerate if both components are nondegenerate in $C(\overline{\Omega})$.

Now we can finally define when the system (1.1) or (1.3) is pullback permanent. Observe that we assume here that solutions are globally defined.

DEFINITION 2.8. *We say that system* (1.1) *is pullback permanent if for any bounded set of intial $B \subset (C(\overline{\Omega}))^2$ bounded away from zero, there exists $t_0 \in \mathbf{R}$ such that for any $t \leq t_0$ the family of solutions*

$$\{\big(u(t,s;u_0,v_0), v(t,s;u_0,v_0)\big),\ s \leq t,\ (u_0,v_0) \in B\} \tag{2.2}$$

is nondegenerate as $s \to -\infty$.

The system (1.1) *is uniformly pullback permanent if it is pullback permament and the functions φ_0 in Definition* 2.7 *are independent of B.*

Note that using the regularizing properties of the solutions of (1.1) or (1.3), if the system is pullback permanent, as defined above, then the set (2.2) is nondegenerate as $s \to -\infty$ for any fixed $t \in \mathbf{R}$.

In an analogous although subtly different way, we can define when system (1.1) or (1.3) is forwards permanent.

DEFINITION 2.9. *We say that system* (1.1) *is forwards permanent if for any bounded set of initial $B \subset (C(\overline{\Omega}))^2$ bounded away from zero, and for any $s \in \mathbf{R}$, the family of solutions*

$$\{\big(u(t,s;u_0,v_0), v(t,s;u_0,v_0)\big),\ s \leq t,\ (u_0,v_0) \in B\} \tag{2.3}$$

is nondegenerate at ∞.

The system (1.1) *is uniformly forwards permanent if it is forwards permanent and the functions* φ_0 *in Definition* 2.7 *are independent of* B.

Note that (1.1) always has the trivial solution $(0,0)$ as well as semitrivial solutions $(u,0)$ and $(0,v)$. Hence, if the system is permanent, as defined above, this implies that trivial and semitrivial solutions are unstable in the pullback or forwards sense; see, e.g., Langa, Robinson, and Suárez [24]. Also, note that permanence implies coexistence of the species, since the values of the solutions eventually remain far from zero in all points of the domain (except at the boundary in the case of Dirichlet boundary conditions).

In the next section we will give conditions on the coefficients of (1.3) for uniform permanence (both forwards and pullback), which will be moreover robust with respect to suitable perturbations on the coefficients.

3. Extinction and permanence for nonautonomous Lotka–Volterra equations: Competition, symbiosis, and prey-predator models. In this section we give results on extinction and pullback and forwards permanence for nonautonomous Lotka–Volterra systems of the type

$$\begin{cases} u_t - d_1\Delta u = u\big(\lambda(t,x) - a(t,x)u - b(t,x)v\big), & x\in\Omega,\ t>s \\ v_t - d_2\Delta v = v\big(\mu(t,x) - c(t,x)u - d(t,x)v\big), & x\in\Omega,\ t>s \\ \mathcal{B}_1 u = 0,\ \mathcal{B}_2 v = 0, & x\in\partial\Omega,\ t>s \\ u(s) = u_s \geq 0,\ v(s) = v_s \geq 0, \end{cases} \tag{3.1}$$

with $d_1, d_2 > 0$; $\lambda, \mu, a, b, c, d \in C^\theta(\overline{Q})$, and $\overline{Q} = \mathbf{R}\times\overline{\Omega}$. Given a function $e \in C^\theta(\overline{Q})$, we define

$$e_L := \inf_{\overline{Q}} e(t,x) \qquad e_M := \sup_{\overline{Q}} e(t,x).$$

We assume from now on that

$$a_L, d_L > 0, \tag{3.2}$$

and consider the three classical cases depending on the signs of b and c:

1. *Competition*: $b_L, c_L > 0$ in $\overline{Q}$.
2. *Symbiosis*: $b_M, c_M < 0$ in $\overline{Q}$.
3. *Prey-predator*: $b_L > 0$, $c_M < 0$ in $\overline{Q}$.

Also, note that we consider all nine possible choices for $\mathcal{B}_i$ as in (1.2).

Using standard techniques, see for instance Pao [30], it can be shown that given $0 \leq u_s \in C(\overline{\Omega})$, $0 \leq v_s \in C(\overline{\Omega})$ there exists, locally in time, a unique solution of (3.1) which is nonnegative, and which we will denote by

$$u = u(t,s,x;u_s,v_s) \geq 0, \quad v = v(t,s,x;u_s,v_s) \geq 0.$$

In fact, due to the strong maximum principle, if $u_s \geq 0$ and $v_s \geq 0$ are both nontrivial, then u and v are strictly positive in Ω. Furthermore, if we denote by $\mathcal{C}_i$ and $\operatorname{int}(\mathcal{C}_i)$ for $i = 1, 2$, respectively, the positive cones in $C^1_{\mathcal{B}_i}(\overline{\Omega})$ and their corresponding interior sets, we have

$$\operatorname{int}(\mathcal{C}_i) := \{u \in \mathcal{C}_i : u > 0 \quad \text{in } \Omega, \text{ and } \frac{\partial u}{\partial \vec{n}} < 0 \quad \text{on } \partial\Omega\} \quad \text{if } \mathcal{B}_i u = u$$

and

$$\operatorname{int}(\mathcal{C}_i) := \{u \in \mathcal{C}_i : u \geq \delta > 0, \quad \text{for some } \delta > 0 \quad \text{in } \overline{\Omega}\},$$

if $\mathcal{B}_i u = \frac{\partial u}{\partial \vec{n}}$ or $\mathcal{B}_i u = d_i \frac{\partial u}{\partial \vec{n}} + \sigma_i(x)u$.

Thus, if $u_s \geq 0$ and $v_s \geq 0$ are both nontrivial, then $(u,v) \in \operatorname{int}(\mathcal{C}_1) \times \operatorname{int}(\mathcal{C}_2)$ for $t > s$.

Note that (3.1) also admits semitrivial solutions of the form $(u,0)$ or $(0,v)$. As indicated in the Introduction, the stability properties of semitrivial solutions play an important role in the global dynamics of (3.1). In fact, extinction requires some semitrivial solution is stable whereas permanence is only possible if semitrivial solutions are somehow unstable.

Thus, we first review some results on the solutions of scalar logistic equations that will be used further below. These results will be used to prove that the local solutions of (3.1) above are, in fact, globally defined. Also, they will be crucially used to prove the existence of a pullback attractor as in section 2.1, and to obtain our results on extinction and permanence as well.

3.1. On the nonautonomous logistic equation. Note that (3.1) always admits semitrivial solutions of the form $(u,0)$ or $(0,v)$. In this case, when one species is not present, the other one satisfies the nonautonomous logistic equation

$$\begin{cases} u_t - d\Delta u = h(t,x)u - g(t,x)u^2 & \text{in } \Omega,\ t > s, \\ \mathcal{B}u = 0 & \text{on } \partial\Omega, \\ u(s) = u_s \geq 0 & \text{in } \Omega, \end{cases} \tag{3.3}$$

where $d > 0$ and $\mathcal{B}$ as in (1.2), $u_s \in C(\overline{\Omega})$, $h, g \in C^\theta(\overline{Q})$, and

$$g_L > 0 \quad \text{in} \quad \overline{Q}.$$

For $m \in L^\infty(\Omega)$ we denote by $\Lambda_{\mathcal{B}}(d,m)$ the first eigenvalue of

$$\begin{cases} -d\Delta u = \lambda u + m(x)u & \text{in } \Omega, \\ \mathcal{B}u = 0 & \text{on } \partial\Omega. \end{cases} \tag{3.4}$$

In particular, we denote by $\Lambda_{0,\mathcal{B}}(d) = \Lambda_{\mathcal{B}}(d,0)$ the first eigenvalue of the operator $-d\Delta$ with boundary conditions $\mathcal{B}$. It is well known that $\Lambda_{\mathcal{B}}(d,m)$ is a simple eigenvalue and a continuous and decreasing function of m. Also note that if m_1 is constant, then

$$\Lambda_{\mathcal{B}}(d, m_1 + m_2) = \Lambda_{\mathcal{B}}(d, m_2) - m_1. \tag{3.5}$$

We write $\varphi_{1,\mathcal{B}}(d,m)$ for the positive eigenfunction associated to $\Lambda_{\mathcal{B}}(d,m)$, normalized such that $\|\varphi_{1,\mathcal{B}}(d,m)\|_{L^\infty(\Omega)} = 1$.

If there is no possible confusion we will suppress the dependence on d and $\mathcal{B}$ in the notations above. When we need to distinguish these quantities with respect to $\mathcal{B}_i$, or d_i, $i = 1,2$, we will employ superscripts as $\Lambda^i(m)$ or Λ_0^i.

Finally, for $h, g \in L^\infty(\Omega)$ with $g_L > 0$ consider the elliptic equation

$$\begin{cases} -d\Delta u = h(x)u - g(x)u^2 & \text{in } \Omega, \\ \mathcal{B}u = 0 & \text{on } \partial\Omega. \end{cases} \tag{3.6}$$

The following result is well known (Cantrell and Cosner [6]).

PROPOSITION 3.1. *If $\Lambda(h) \geq 0$, the unique nonnegative solution of* (3.6) *is the trivial one, i.e., $\omega_{[h,g]}(x) = 0$. On the other hand, if $\Lambda(h) < 0$ there exits a unique positive solution of* (3.6)*, which we denote by $\omega_{[h,g]}(x)$. Moreover, $0 < \omega_{[h,g]}(x) \leq \Psi(x)$ in Ω, where*

$$\Psi(x) = \begin{cases} \dfrac{h_M}{g_L} & \textit{for Dirichlet or Neumann BCs,} \\ -\dfrac{\Lambda(h)}{\varphi_L g_L}\varphi(x) & \textit{for Robin BCs,} \end{cases}$$

with $\varphi = \varphi_{1,\mathcal{B}}(m)$.

The following result will be used in what follows.

LEMMA 3.2. *Assume that* $h_n \in L^\infty(\Omega)$ *and that*

$$h_n \to h_\infty \quad \text{in } L^\infty(\Omega),$$

with $\Lambda(h_\infty) < 0$. *Then, there exist* $n_0 \in \mathbb{N}$, *and* $\varphi \in int(\mathcal{C})$ *such that*

$$\varphi(x) \le \omega_{[h_n,g]}(x) \quad \text{in } \Omega, \quad \text{for all } n \ge n_0,$$

where $\omega_{[h_n,g]}(x)$ *is given by Proposition* 3.1.

Proof. Since $\Lambda(h_\infty) < 0$, we can take $\varepsilon > 0$ such that $0 < \varepsilon < -\Lambda(h_\infty)$. For this $\varepsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that for $n \ge n_0$

$$-\varepsilon < h_n - h_\infty < \varepsilon \quad \text{for all } x \in \Omega.$$

Consider $\varphi_\infty \in \text{int}(\mathcal{C})$ the eigenfunction associated to $\Lambda(h_\infty)$ with $\|\varphi_\infty\|_{L^\infty(\Omega)} = 1$. It is not hard to show that $\delta\varphi_\infty$ is a subsolution of (3.6) with $h = h_n$ provided that

$$\delta \le -\frac{\varepsilon + \Lambda(h_\infty)}{g_M}.$$

So, $\delta\varphi_\infty(x) \le \omega_{[h_n,g]}(x)$ in Ω. This completes the proof. ∎

In [25] and [34] the following properties of (3.3) were proved.

THEOREM 3.3. *Assume that in* (3.3)

$$h_M < \infty \quad \text{and} \quad g_L > 0 \quad \text{in } \overline{Q}.$$

Then

1. *For every nontrivial* $u_s \in C(\overline{\Omega})$, $u_s \ge 0$, *there exists a unique positive solution of* (3.3) *denoted by* $\Theta_{[h,g]}(t,s,u_s)$. *Moreover,*

$$0 \le \Theta_{[h,g]}(t,s,u_s) \le K, \tag{3.7}$$

where

$$K := \begin{cases} \max\left\{(u_s)_M, \dfrac{h_M}{g_L}\right\} & \text{for Dirichlet or Neumann BCs,} \\ \max\left\{\left(\dfrac{u_s}{\varphi}\right)_M, \dfrac{-\Lambda(h_M)}{\varphi_L g_L}\right\} & \text{for Robin BCs,} \end{cases}$$

and φ *is the positive eigenfunction associated to* $\Lambda(h_M)$ *with* $\|\varphi\|_{L^\infty(\Omega)} = 1$.

2. *For fixed* $t > s$, u_s, *the map* $h \mapsto \Theta_{[h,g]}(t,s,u_s)$ *is increasing and* $g \mapsto \Theta_{[h,g]}(t,s,u_s)$ *is decreasing.*
   For fixed $t > s$, h *and* g, *the map* $u_s \mapsto \Theta_{[h,g]}(t,s,u_s)$ *is increasing.*
3. *Define, for* $x \in \Omega$,

$$h_0(x) := \inf_{t\in\mathbf{R}} h(t,x), \qquad H_0(x) := \sup_{t\in\mathbf{R}} h(t,x)$$

and

$$g_0(x) := \inf_{t\in\mathbf{R}} g(t,x), \qquad G_0(x) := \sup_{t\in\mathbf{R}} g(t,x).$$

Then, if $u_s \in int(\mathcal{C})$ and $\Lambda(h_0) < 0$ we have, for any $t > s$,

$$0 < \varepsilon\varphi_1(x) \le \Theta_{[h,g]}(t,s,x;u_s) \quad \text{in } \Omega, \tag{3.8}$$

where φ_1 is the positive eigenfunction associated to $\Lambda(h_0)$ and

$$\varepsilon = \varepsilon(u_s) := \min\left\{\left(\frac{u_s}{\varphi_1}\right)_L, \frac{-\Lambda(h_0)}{g_M}\right\}.$$

4. *If $\Lambda(H_0) > 0$, then for all initial data $u_s \ge 0$, $\Theta_{[h,g]}(t,s,u_s) \to 0$, in $C^1(\overline{\Omega})$, as $t - s \to \infty$. Moreover the convergence is exponential and uniform for bounded sets of initial data u_s.*
5. *If $\Lambda(h_0) < 0$, then there exists a unique bounded, complete, and nondegenerate trajectory at $\pm\infty$ of (3.3), $\varphi_{[h,g]}$, which moreover satisfies that for all s and any bounded set of nontrivial initial data $u_s \ge 0$, bounded away from 0,*

$$\Theta_{[h,g]}(t,s,u_s) - \varphi_{[h,g]}(t) \to 0 \quad \text{as } t \to \infty.$$

That is, $\varphi_{[h,g]}$ describes the forward behavior of all solutions. Also, $\varphi_{[h,g]}$ describes the pullback behavior of all nondegenerate solutions of (3.3), that is, for each t, if $s \mapsto u_s \ge 0$ is bounded and nondegenerate, then

$$\Theta_{[h,g]}(t,s,u_s) - \varphi_{[h,g]}(t) \to 0 \quad \text{as } s \to -\infty.$$

Both limits above are taken in $C^1(\overline{\Omega})$. Furthermore for all $t \in \mathbf{R}$, we have

$$\omega_{[h_0,G_0]}(x) \le \varphi_{[h,g]}(t,x) \le \omega_{[H_0,g_0]}(x) \quad \text{in } \Omega.$$

6. *If h, g are independent of t and are in $L^\infty(\Omega)$ with $g_L > 0$ and $\Lambda(h) < 0$, then $\varphi_{[h,g]}(t,x) = \omega_{[h,g]}(x)$ is the unique positive solution of (3.6) and for all $t > s$ and u_s*

$$\Theta_{[h,g]}(t,s,u_s) = \Theta_{[h,g]}(t-s,u_s) \to \omega_{[h,g]} \quad \text{in } C^1(\overline{\Omega}) \text{ as } t - s \to \infty$$

uniformly for bounded sets of initial data $u_s \ge 0$ bounded away from zero. In particular, there exist $m \le 1 \le M$ such that

$$m\omega_{[h,g]} \le \Theta_{[h,g]}(t,s,u_s) \le M\omega_{[h,g]},$$

for $t - s$ large.

Moreover in statements 4, 5, and 6 above the convergence as $t \to \infty$ is exponentially fast (see [33]).

3.2. Existence of the pullback attractor and complete trajectories for nonautonomous Lotka–Volterra systems. Our first purpose is to prove the existence of a nonautonomous pullback attractor for (3.1). To do this we will derive suitable estimates on the solutions of (3.1). In doing this we will use the following notation for the solutions of (3.3) with diffusion coefficients d_1 and d_2 and boundary conditions $\mathcal{B}_1$ and $\mathcal{B}_2$, respectively,

$$\xi_{[\lambda,a]}(t,s) = \Theta_{[\lambda,a]}(t,s,u_s), \quad \eta_{[\mu,d]}(t,s) = \Theta_{[\mu,d]}(t,s,v_s),$$

where $u_s \ge 0$ and $v_s \ge 0$ in Ω.

THEOREM 3.4. *Provided that $a_L, d_L > 0$, for any solution (u,v) of (3.1), with initial data $u_s \ge 0$, $v_s \ge 0$, the following lower and upper bounds hold:*

1. *Competition,* $b_L > 0, c_L > 0$*:*

$$\xi_{[\lambda - b\eta_{[\mu,d]},a]} \le u \le \xi_{[\lambda,a]}, \qquad \eta_{[\mu - c\xi_{[\lambda,a]},d]} \le v \le \eta_{[\mu,d]}.$$

2. *Symbiosis,* $b_M < 0, c_M < 0$*: Assume that*

$$b_L c_L < a_L d_L. \tag{3.9}$$

Then,

$$\xi_{[\lambda - b\eta_{[\mu,d]},a]} \le u, \quad \eta_{[\mu - c\xi_{[\lambda,a]},d]} \le v.$$

Assume furthermore that $d_1 = d_2$ *and define*

$$\gamma = \max\{\lambda_M, \mu_M\}, \quad M = \frac{a_L - c_L}{d_L - b_L} > 0, \quad K = \frac{a_L d_L - b_L c_L}{d_L - b_L} > 0,$$

and choose w_s *such that* $w_s \ge \max\{u_s, \frac{1}{M} v_s\}$*. Denote by* $\Theta_{[\gamma,K]}(t, s, w_s)$ *the solution of* (3.3) *with* $d = d_1$ *and a certain boundary condition that depends on* $\mathcal{B}_1$ *and* $\mathcal{B}_2$ *and that will be specified in the proof. Then, we have the upper bounds*

$$u \le \Theta_{[\gamma,K]}(t, s, w_s), \quad v \le M\Theta_{[\gamma,K]}(t, s, w_s).$$

3. *Prey-predator,* $b_L > 0, c_M < 0$*:*

$$\xi_{[\lambda - b\eta_{[\mu - c\xi_{[\lambda,a]},d]},a]} \le u \le \xi_{[\lambda - b\eta_{[\mu,d]},a]} \le \xi_{[\lambda,a]}, \qquad \eta_{[\mu,d]} \le v \le \eta_{[\mu - c\xi_{[\lambda,a]},d]}.$$

Proof. 1. Assume that $b_L, c_L > 0$. If we write the equation for u as

$$u_t - d_1 \Delta u = u(\lambda - bv) - au^2,$$

then using Theorem 3.3 we get

$$u = \xi_{[\lambda - bv,a]} \le \xi_{[\lambda,a]},$$

and similarly,

$$v \le \eta_{[\mu,d]}.$$

Hence, again by Theorem 3.3

$$u = \xi_{[\lambda - bv,a]} \ge \xi_{[\lambda - b\eta_{[\mu,d]},a]}.$$

2. Assume now that $b_M, c_M < 0$. To have the lower bounds it is enough to check that in the equation for u one has

$$\xi_{[\lambda - b\eta_{[\mu,d]},a]}(\lambda - a\xi_{[\lambda - b\eta_{[\mu,d]},a]} - b\eta_{[\mu,d]}) \le \xi_{[\lambda - b\eta_{[\mu,d]},a]}(\lambda - a\xi_{[\lambda - b\eta_{[\mu,d]},a]} - b\eta_{[\mu - c\xi_{[\lambda,a]},d]}),$$

or equivalently,

$$\eta_{[\mu - c\xi_{[\lambda,a]},d]} \ge \eta_{[\mu,d]},$$

which is true since $c < 0$. We treat the equation for v similarly.

On the other hand, assuming that $d_1 = d_2$, define

$$\overline{u} = \Theta_{[\gamma,K]}(t,s,w_s), \quad \overline{v} = M\Theta_{[\gamma,K]}(t,s,w_s)$$

with a suitable boundary condition, $\mathcal{B}$, to be described below. Then using the equations we get that $\overline{u}$ and $\overline{v}$ are supersolutions if

$$-K \geq -a - bM, \qquad -K \geq -dM - c,$$

which is satisfied with the choice of M and K. To compare the solutions with the uppersolutions on the boundary, if either u or v satisfies Dirchlet boundary conditions we take $\mathcal{B}$ the boundary condition of the other component. If both u and v satisfy Robin or Neumann (i.e., $\sigma_i = 0$ in the latter case) boundary conditions we define

$$\sigma = \min\{\sigma_1, \sigma_2\},$$

and $\mathcal{B}u = d_1 \frac{\partial u}{\partial \overline{n}} + \sigma(x)u$.

3. Assume finally that $b_L > 0$, $c_M < 0$, then

$$u \leq \xi_{[\lambda,a]} \quad \text{and} \quad \eta_{[\mu,d]} \leq v.$$

Hence

$$v = \eta_{[\mu - cu,d]} \leq \eta_{[\mu - c\xi_{[\lambda,a]},d]},$$

and then,

$$u = \xi_{[\lambda - bv,a]} \geq \xi_{[\lambda - b\eta_{[\mu - c\xi_{[\lambda,a]},d]},a]}. \qquad \square$$

With the upper bounds in Theorem 3.4 and using the results for scalar logistic equations in Theorem 3.3, we get the following result.

THEOREM 3.5. *Under the assumptions in cases* (1)–(3) *of Theorem* 3.4*, all solutions of* (3.1) *are global in time and moreover there exists a pullback attractor* $\mathcal{A}(t)$ *of* (3.1)*, which is bounded for all* $t \in \mathbf{R}$*. More precisely, we have*

$$\limsup_{t-s\to\infty} u(t,s;u_s,v_s) \leq M_\infty, \quad \limsup_{t-s\to\infty} v(t,s;u_s,v_s) \leq N_\infty,$$

uniformly in Ω *and for bounded sets of intial data* $u_s, v_s \geq 0$*, for some constants* $M_\infty \geq 0$ *and* $N_\infty \geq 0$ *that depend on the coefficients of* (3.1)*.*

In particular, there exists at least one complete bounded trajectory $(u^*(t), v^*(t))$*,* $t \in \mathbf{R}$*, for* (3.1)*. Furthermore, all complete bounded trajectories of* (3.1) *are uniformly bounded by* M_∞ *and* N_∞ *and for all* $t \in \mathbf{R}$*.*

Proof. Thanks to the upper bounds in Theorem 3.4, the positive solutions of (3.1) are always bounded by solutions of the logistic equation of the type (3.3). In particular, all solutions of (3.1) are globally defined.

Now we use that

$$0 \leq \Theta_{[\alpha,\beta]}(t,s;z) \leq \Theta_{[\alpha_M,\beta_L]}(t-s;z),$$

statements (4)–(6) in Theorem 3.3, and that $0 \leq \omega_{[\alpha_M,\beta_L]}(x) \leq \Psi_M$, with ω and $\Psi(x)$ as in Proposition 3.1, to get the estimates.

In particular, this implies the existence of bounded pullback absorbing sets for (3.1) in $C(\overline{\Omega}) \times C(\overline{\Omega})$.

Then following the proof of section 6 in Langa et al. [25], we can show the existence of a bounded pullback absorbing set in $C^1(\overline{\Omega}) \times C^1(\overline{\Omega})$, and so compact in $C(\overline{\Omega}) \times C(\overline{\Omega})$. Hence, we conclude using Theorem 2.3 as the existence of a bounded nonautonomous pullback attractor $\mathcal{A}(t)$, and thus the existence of at least one bounded complete trajectory $(u^*(t), v^*(t))$, $t \in \mathbf{R}$, follows. $\square$

3.3. Extinction for nonautonomous Lotka–Volterra systems. Note that with the arguments above there are some cases, when statement 4 in Theorem 3.3 can be used, in which one (or both) constants M_∞ and N_∞ are zero and we have then extinction of one of the species. This implies, in turn, that the semitrivial (or the trivial) solutions are stable in a forwards and pullback senses. More precisely, we have the following result. Observe that these sufficient conditions are far from optimal but qualitatively they describe the global stability of trivial or semitrivial solutions.

PROPOSITION 3.6. *With the notations in Theorems* 3.4 *and* 3.5*, we have*

1. *Competition,* $b_L > 0, c_L > 0$*. If*

$$\lambda_M < \Lambda_0^1, \quad \textit{then} \quad M_\infty = 0,$$

while if

$$\mu_M < \Lambda_0^2, \quad \textit{then} \quad N_\infty = 0.$$

2. *Symbiosis,* $b_M < 0, c_M < 0$*,* $d_1 = d_2$ *and* (3.9)*, that is* $b_L c_L < a_L d_L$*. If*

$$\gamma < \Lambda_0^1, \quad \textit{then} \quad M_\infty = 0,$$

while if

$$\gamma < \Lambda_0^2, \quad \textit{then} \quad N_\infty = 0.$$

3. *Prey-predator,* $b_L > 0, c_M < 0$*. If*

$$\lambda_M < \Lambda_0^1, \quad \textit{then} \quad M_\infty = 0,$$

and in this case, if

$$\mu_M < \Lambda_0^2, \quad \textit{then} \quad M_\infty = 0.$$

On the other hand, if

$$\Lambda_0^1 < \lambda_M, \quad \textit{and} \quad \mu_M - c_L \frac{\lambda_M}{a_L} < \Lambda_0^2, \quad \textit{then} \quad N_\infty = 0.$$

In all the cases, when $M_\infty = 0$ *the* u *component of the solutions of* (3.1) *extinguishes in pullback and forwards senses, while the* v *component of the solutions asymptotically follows the dynamics of the scalar logistic equation* (3.3) *with* $h(t,x) = \mu(t,x)$ *and* $g(t,x) = d(t,x)$ *as described in Theorem* 3.3*.*

The case when $N_\infty = 0$ *is analogous.*

Proof. In fact, in the case of competition we have $0 \le u \le \xi_{[\lambda_M, a_L]}$ and $0 \le v \le \eta_{[\mu_M, d_L]}$. Hence, from statement 4 in Theorem 3.3 and using (3.5), if $\Lambda^1(\lambda_M) = \Lambda_0^1 - \lambda_M > 0$, then $M_\infty = 0$, while $N_\infty = 0$ if $\Lambda^2(\mu_M) = \Lambda_0^2 - \mu_M > 0$.

In the case of symbiosis, assuming $d_1 = d_2$, we have $0 \le u \le \Theta_{[\gamma,K]}(t,s,w_s)$, $0 \le v \le M\Theta_{[\gamma,K]}(t,s,w_s)$. Hence, if $\Lambda^1(\gamma) = \Lambda_0^1 - \gamma > 0$, then $M_\infty = 0$, while $N_\infty = 0$ if $\Lambda^2(\gamma) = \Lambda_0^2 - \gamma > 0$.

Finally, in the case of prey-predator, we have $0 \le u \le \xi_{[\lambda_M, a_L]}$, $0 \le v \le \eta_{[\mu_M - c_L \xi_{[\lambda_M, a_L]}, d_L]}$. Hence, if $\Lambda^1(\lambda_M) = \Lambda_0^1 - \lambda_M > 0$, then $M_\infty = 0$. In this case, $N_\infty = 0$ if $\Lambda^2(\mu_M) = \Lambda_0^2 - \mu_M > 0$.

On the other hand, if $\Lambda_0^1 < \lambda_M$, then for large values of $t - s$ we have $v \le \eta_{[\mu_M - c_L(\omega_{[\lambda_M, a_L]} + \varepsilon), d_L]}$, and then $N_\infty = 0$ if $\Lambda^2(\mu_M - c_L \frac{\lambda_M}{a_L}) = \Lambda_0^2 - \mu_M + c_L \frac{\lambda_M}{a_L} > 0$.

The rest is immediate. ∎

As we are interested in the "permanence" problem for (3.1), we will consider in what follows only the cases in which $M_\infty > 0$ and $N_\infty > 0$. In particular, note that for sufficiently large values of $\lambda_M > 0$ and $\mu_M > 0$ we can take, for the case of Dirichlet or Neumann boundary conditions in either one of the components u or v,

$$M_\infty = \begin{cases} \dfrac{\lambda_M}{a_L} & \text{in the competition case,} \\ \dfrac{\gamma}{K} & \text{in the symbiosis case,} \\ \dfrac{\lambda_M}{a_L} & \text{in the prey-predator case,} \end{cases}$$

$$N_\infty = \begin{cases} \dfrac{\mu_M}{d_L} & \text{in the competition case,} \\ M\dfrac{\gamma}{K} & \text{in the symbiosis case,} \\ \dfrac{\mu_M - c_L \frac{\lambda_M}{a_L}}{d_L} & \text{in the prey-predator case,} \end{cases}$$

while for Robin boundary conditions we have

$$M_\infty = \begin{cases} \dfrac{\lambda_M - \Lambda_0^1}{(\varphi^1)_L a_L} & \text{in the competition case,} \\ \dfrac{\gamma - \Lambda_0^1}{(\varphi^1)_L K} & \text{in the symbiosis case,} \\ \dfrac{\lambda_M - \Lambda_0^1}{(\varphi^1)_L a_L} & \text{in the prey-predator case,} \end{cases}$$

$$N_\infty = \begin{cases} \dfrac{\mu_M - \Lambda_0^2}{(\varphi^2)_L d_L} & \text{in the competition case,} \\ M\dfrac{\gamma - \Lambda_0^2}{(\varphi^2)_L K} & \text{in the symbiosis case,} \\ \dfrac{\mu_M - c_L \frac{\lambda_M - \Lambda_0^1}{(\varphi^1)_L a_L} - \Lambda_0^2}{(\varphi^2)_L d_L} & \text{in the prey-predator case,} \end{cases}$$

where φ^i denotes the positive eigenfunction associated to Λ_0^i with $\|\varphi^i\|_{L^\infty(\Omega)} = 1$. Note that similar expressions can be given in the remaining five cases for the boundary conditions, although their explicit form becomes more cumbersome.

In fact, in the next section we will impose conditions on the coefficients to ensure that the pullback and forwards behavior of the solutions of (3.1), with nontrivial initial data, is far from the semitrivial and the trivial solutions.

3.4. Permanence for nonautonomous Lotka–Volterra systems: Nondegeneracy of solutions. Now, using the lower bounds in Theorem 3.4, we will give sufficient conditions for the system (3.1) to be uniformly permanent in pullback and forwards senses, as in section 2.2. For reasons that will become clear further below, we are interested in obtaining such nondegeneracy in a uniform way with respect to the coefficients λ, μ, a, b, c, d in the system. For this, recall the notations in (3.4) and that we always take nonnegative nontrivial initial data u_s, v_s.

Also note that in the results of this section we will use the quantities $\lambda_I \le \lambda_S$, $\mu_I \le \mu_S$, $a_I \le a_S$, $b_I \le b_S$, $c_I \le c_S$, and $d_I \le d_S$ to control the asymptotic sizes of

the coefficients λ, μ, a, b, c, d as $t \to \pm\infty$. As all the results will be given in terms of such quantities, the statements below show the robustness of the results with respect to perturbations in the coefficients of the system.

Finally, we stress here once again that the results below imply the instability of trivial and semitrivial solutions.

3.4.1. Competition.

PROPOSITION 3.7 (forwards permanence—competitive case). *Assume* (3.2) *and* $b_L, c_L > 0$. *Then:*

(i) *If* $\lambda_I > \Lambda^1(-b_S\omega_{[\mu_S,d_I]})$ *there exists* $\psi_{11} \in int(\mathcal{C}_1)$ *such that whenever*

$$\lambda(t,x) \geq \lambda_I,\ \mu(t,x) \leq \mu_S,\ b(t,x) \leq b_S,\ a(t,x) \leq a_S,\ \text{and}\quad d(t,x) \geq d_I > 0$$

for all $x \in \Omega$ *and* $t \geq t_0$, *for any* $u_s, v_s > 0$, *the solution for* $t > s \geq t_0$ *of* (3.1) *satisfies* $\psi_{11}(x) \leq u(t,s,x;u_s,v_s)$ *for* $t-s$ *large enough.*

(ii) *If* $\mu_I > \Lambda^2(-c_S\omega_{[\lambda_S,a_I]})$ *there exists* $\psi_{22} \in int(\mathcal{C}_2)$ *such that whenever*

$$\lambda(t,x) \leq \lambda_S,\quad \mu(t,x) \geq \mu_I,\quad a(t,x) \geq a_I > 0,\quad d(t,x) \leq d_S,\quad c(t,x) \leq c_S$$

for all $x \in \Omega$ *and* $t \geq t_0$, *for any* $u_s, v_s > 0$, *the solution for* $t > s \geq t_0$ *of* (3.1) *satisfies* $\psi_{22}(x) \leq v(t,s,x;u_s,v_s)$ *for* $t-s$ *large enough.*

Hence, if

$$\lambda_I > \Lambda^1(-b_S\omega_{[\mu_S,d_I]}) \qquad \text{and} \qquad \mu_I > \Lambda^2(-c_S\omega_{[\lambda_S,a_I]}), \tag{3.10}$$

then there exist $\psi_{11} \in int(\mathcal{C}_1)$ *and* $\psi_{22} \in int(\mathcal{C}_2)$ *such that for any choice of coefficients that satisfy*

$$\begin{aligned} &\lambda_I \leq \lambda(t,x) \leq \lambda_S, \qquad \mu_I \leq \mu(t,x) \leq \mu_S, \qquad 0 < a_I \leq a(t,x) \leq a_S,\\ &0 < b_I \leq b(t,x) \leq b_S, \quad 0 < c_I \leq c(t,x) \leq c_S, \quad 0 < d_I \leq d(t,x) \leq d_S, \end{aligned}$$

for all $x \in \Omega$ *and for all* $t \geq t_0$, *and for all nontrivial* $u_s \geq 0$, $v_s \geq 0$ *in a fixed bounded set of* $C(\overline{\Omega})$ *bounded away from* 0, *the solution* (u,v) *of* (3.1) *for* $t > s \geq t_0$ *is nondegenerate at* ∞ *and for all* $t-s$ *large enough,*

$$u(t,s,x;u_s,v_s) \geq \psi_{11}(x) \quad \text{and} \quad v(t,s,x;u_s,v_s) \geq \psi_{22}(x).$$

In particular, (3.1) *is uniformly forwards permanent.*

Proof. Since $\lambda_I > \Lambda^1(-b_S\omega_{[\mu_S,d_I]})$, by the continuity of $\Lambda^1(m)$ with respect to m, there exists $\varepsilon > 0$ such that

$$\lambda_I > \Lambda^1(-b_S(\omega_{[\mu_S,d_I]} + \varepsilon)) \quad \text{or equivalently by (3.5)} \quad \Lambda^1(\lambda_I - b_S(\omega_{[\mu_S,d_I]} + \varepsilon)) < 0.$$

Using Theorems 3.3 and 3.4, we get, for $t > s \geq t_0$,

$$u(t,s,u_s,v_s) \geq \xi_{[\lambda - b\eta_{[\mu,d]},a]}(t,s,u_s) \geq \Theta_{[\lambda_I - b_S\eta_{[\mu_S,d_I]},a_S]}(t-s,u_s).$$

Moreover, $\eta_{[\mu_S,d_I]}(t,s,v_s) \to \omega_{[\mu_S,d_I]}$ in $C^1(\overline{\Omega})$ and uniformly for v_s in bounded sets bounded away from zero, as $t-s \to \infty$, and so

$$u(t,s,u_s,v_s) \geq \Theta_{[\lambda_I - b_S(\omega_{[\mu_S,d_I]}+\varepsilon),a_S]}(t-s,u_s) \to \omega_{[\lambda_I - b_S(\omega_{[\mu_S,d_I]}+\varepsilon),a_S]} \tag{3.11}$$

in $C^1(\overline{\Omega})$ and uniformly for u_s in bounded sets bounded away from zero, as $t-s \to \infty$ by Theorem 3.3 and where we have used (3.10). Hence, the result follows for u.

On the other hand, we have analogously for the v component, for $t > s \geq t_0$,

$$v(t,s,u_s,v_s) \geq \eta_{[\mu - c\xi_{[\lambda,a]},d]}(t,s,v_s) \geq \Theta_{[\mu_I - c_S\xi_{[\lambda_S,a_I]},d_S]}(t-s,v_s).$$

Now, from (3.10), $\xi_{[\lambda_S,a_I]}(t,s,u_s) \to \omega_{[\lambda_S,a_I]}$ in $C^1(\overline{\Omega})$ and uniformly for u_s in bounded sets bounded away from zero, as $t-s \to \infty$, and so

$$v(t,s,u_s,v_s) \geq \Theta_{[\mu_I - c_S(\omega_{[\lambda_S,a_I]}+\varepsilon),d_S]}(t-s,v_s) \to \omega_{[\mu_I - c_S(\omega_{[\lambda_S,a_I]}+\varepsilon),d_S]} \tag{3.12}$$

in $C^1(\overline{\Omega})$ and uniformly for v_s in bounded sets bounded away from zero, as $t-s \to \infty$ by Theorem 3.3. ∎

The same arguments as above, carried out in all pullback sense lead to the following result. Note that, in particular, this proposition guarantees the uniform nondegeneracy at $-\infty$ of complete nondegenerate trajectories with respect to the coefficients in the system.

PROPOSITION 3.8 (pullback permanence—competitive case). *Assume* (3.2) *and* $b_L, c_L > 0$. *Then:*

(i) *If* $\lambda_I > \Lambda^1(-b_S\omega_{[\mu_S,d_I]})$ *there exists* $\psi_{11} \in int(\mathcal{C}_1)$ *such that whenever*

$$\lambda(t,x) \geq \lambda_I, \quad \mu(t,x) \leq \mu_S, \quad b(t,x) \leq b_S, \quad a(t,x) \leq a_S, \quad d(t,x) \geq d_I > 0$$

for all $x \in \Omega$ *and* $t \leq t_0$ *(for some* $t_0 \in \mathbf{R}$*), for any* $u_s, v_s > 0$, *the solution for* $s < t \leq t_0$ *of* (3.1) *satisfies* $\psi_{11}(x) \leq u(t,s,x;u_s,v_s)$ *for* $t-s$ *large enough. In particular, any complete trajectory of* (3.1) *that is nondegenerate at* $-\infty$ *satisfies* $u(t,x) \geq \psi_{11}(x)$ *for all* $x \in \Omega$ *and* $t \leq t_0$.

(ii) *If* $\mu_I > \Lambda^2(-c_S\omega_{[\lambda_S,a_I]})$ *there exists* $\psi_{22} \in int(\mathcal{C}_2)$ *such that whenever*

$$\lambda(t,x) \leq \lambda_S, \quad \mu(t,x) \geq \mu_I, \quad a(t,x) \geq a_I > 0, \quad d(t,x) \leq d_S, \quad c(t,x) \leq c_S$$

for all $x \in \Omega$ *and* $t \leq t_0$ *(some* $t_0 \in \mathbf{R}$*), for any* $u_s, v_s > 0$, *the solution for* $s < t \leq t_0$ *of* (3.1) *satisfies* $\psi_{22}(x) \leq v(t,s,x;u_s,v_s)$ *for* $t-s$ *large enough. In particular, any complete trajectory of* (3.1) *that is nondegenerate at* $-\infty$ *satisfies* $v(t,x) \geq \psi_{22}(x)$ *for all* $x \in \Omega$ *and* $t \leq t_0$.

Hence, if

$$\lambda_I > \Lambda^1(-b_S\omega_{[\mu_S,d_I]}) \qquad \text{and} \qquad \mu_I > \Lambda^2(-c_S\omega_{[\lambda_S,a_I]}) \tag{3.13}$$

there exist functions $\psi_{11} \in int(\mathcal{C}_1)$ *and* $\psi_{22} \in int(\mathcal{C}_2)$ *such that whenever*

$$\begin{aligned} &\lambda_I \leq \lambda(t,x) \leq \lambda_S, \qquad \mu_I \leq \mu(t,x) \leq \mu_S, \qquad 0 < a_I \leq a(t,x) \leq a_S, \\ &0 < b_I \leq b(t,x) \leq b_S, \quad 0 < c_I \leq c(t,x) \leq c_S, \quad 0 < d_I \leq d(t,x) \leq d_S, \end{aligned}$$

for all $x \in \Omega$ *and* $t \leq t_0$ *(for some* $t_0 \in \mathbf{R}$*), and for all nontrivial* $u_s \geq 0$, $v_s \geq 0$ *in a fixed bounded set,* B, *of* $C(\overline{\Omega})$ *bounded away from* 0, *the set of solutions of* (3.1) $\{(u,v),\ s < t \leq t_0,\ (u_s,v_s) \in B\}$ *is nondegenerate as* $s \to -\infty$ *and for all* $t-s$ *large enough*

$$u(t,s,x;u_s,v_s) \geq \psi_{11}(x) \quad \text{and} \quad v(t,s,x;u_s,v_s) \geq \psi_{22}(x).$$

In particular, (3.1) *is uniformly pullback permanent and any bounded complete trajectory that is nondegenerate at* $-\infty$ *satisfies*

$$u(t,x) \geq \psi_{11}(x) \quad \text{and} \quad v(t,x) \geq \psi_{22}(x) \quad \text{for all } x \in \Omega \text{ and } t \leq t_0.$$

Proof. The first part of the statements follow from (3.11) and (3.12), with $t-s \to \infty$ but now $s < t \le t_0$.

For a complete solution, arguing as in Proposition 3.7 we get for any $t_0 \ge t > s$,

$$u(t) \ge \xi_{[\lambda - b\eta_{[\mu,d]},a]}(t,s,u(s)) \ge \Theta_{[\lambda_I - b_S\eta_{[\mu_S,d_I]},a_S]}(t-s,u(s)).$$

As v is nondegenerate at $-\infty$, 5 in Theorem 3.3 implies $\eta_{[\mu_S,d_I]}(t,s,v(s)) \to \omega_{[\mu_S,d_I]}$ in $C^1(\overline{\Omega})$ as $s \to -\infty$. Thus, for sufficiently negative s,

$$u(t) \ge \Theta_{[\lambda_I - b_S(\omega_{[\mu_S,d_I]}+\varepsilon),a_S]}(t-s,u(s)) \to \omega_{[\lambda_I - b_S(\omega_{[\mu_S,d_I]}+\varepsilon),a_S]} \tag{3.14}$$

in $C^1(\overline{\Omega})$ as $s \to -\infty$, because u is nondegenerate at $-\infty$ and 5 in Theorem 3.3 again. Hence the result follows for u.

On the other hand, we have analogously for the v component for any $t_0 \ge t > s$,

$$v(t) \ge \eta_{[\mu - c\xi_{[\lambda,a]},d]}(t,s,v(s)) \ge \Theta_{[\mu_I - c_S\xi_{[\lambda_S,a_I]},d_S]}(t-s,v(s)).$$

Now, $\xi_{[\lambda_S,a_I]}(t,s,u(s)) \to \omega_{[\lambda_S,a_I]}$ in $C^1(\overline{\Omega})$ as $s \to -\infty$, because u is nondegenerate at $-\infty$, and so, for sufficiently negative s,

$$v(t) \ge \Theta_{[\mu_I - c_S(\omega_{[\lambda_S,a_I]}+\varepsilon),d_S]}(t-s,v(s)) \to \omega_{[\mu_I - c_S(\omega_{[\lambda_S,a_I]}+\varepsilon),d_S]} \tag{3.15}$$

in $C^1(\overline{\Omega})$ as $s \to -\infty$ by Theorem 3.3, because v is nondegenerate at $-\infty$. ∎

Results for the other cases can be proved analogously, as we now show.

3.4.2. Symbiosis. First for the case of symbiosis, we have the following result. Note that as we make no use here of the upper bound in Theorem 3.4, we do not assume below that $d_1 = d_2$.

PROPOSITION 3.9 (forwards permanence—symbiotic case). *Assume* (3.2), b_M, $c_M < 0$, *and* (3.9), *that is*

$$b_L c_L < a_L d_L.$$

Then:

(i) *If* $\lambda_I > \Lambda^1(-b_S\omega_{[\mu_I,d_S]})$ *there exists* $\psi_{11} \in int(\mathcal{C}_1)$ *such that whenever*

$$\lambda(t,x) \ge \lambda_I, \quad \mu(t,x) \ge \mu_I, \quad b(t,x) \le b_S < 0, \quad a(t,x) \le a_S, \quad d(t,x) \le d_S$$

for all $x \in \Omega$ *and* $t \ge t_0$ *(some* $t_0 \in \mathbf{R}$*), for any* $u_s, v_s > 0$ *the solution for* $t > s \ge t_0$ *of* (3.1) *satisfies* $\psi_{11}(x) \le u(t,s,x;u_s,v_s)$ *for* $t-s$ *large enough.*

(ii) *If* $\mu_I > \Lambda^2(-c_S\omega_{[\lambda_I,a_S]})$ *there exists* $\psi_{22} \in int(\mathcal{C}_2)$ *such that whenever*

$$\lambda(t,x) \ge \lambda_I, \quad \mu(t,x) \ge \mu_I, \quad a(t,x) \le a_S, \quad d(t,x) \le d_S, \quad c(t,x) \le c_S < 0$$

for all $x \in \Omega$ *and* $t \ge t_0$ *(some* $t_0 \in \mathbf{R}$*), for any* $u_s, v_s > 0$ *the solution for* $t > s \ge t_0$ *of* (3.1) *satisfies* $\psi_{22}(x) \le v(t,s,x;u_s,v_s)$ *for* $t-s$ *large enough.*

Hence, if

$$\lambda_I > \Lambda^1(-b_S\omega_{[\mu_I,d_S]}) \qquad \text{and} \qquad \mu_I > \Lambda^2(-c_S\omega_{[\lambda_I,a_S]}), \tag{3.16}$$

then there are functions $\psi_{11} \in int(\mathcal{C}_1)$ *and* $\psi_{22} \in int(\mathcal{C}_2)$ *such that whenever*

$$\begin{array}{lll} \lambda_I \le \lambda(t,x), & \mu_I \le \mu(t,x), & a(t,x) \le a_S, \\ b(t,x) \le b_S < 0, & c(t,x) \le c_S < 0, & d(t,x) \le d_S \end{array}$$

$x \in \Omega$ and $t \geq t_0$ (some $t_0 \in \mathbf{R}$), and for all $u_s > 0, v_s > 0$ in a fixed bounded set of $C(\overline{\Omega})$ bounded away from 0, the solution (u, v) of (3.1) for $t > s \geq t_0$ is nondegenerate at ∞, and for all $t - s$ large enough

$$u(t, s; u_s, v_s) \geq \psi_{11}(x) \qquad \text{and} \qquad v(t, s; u_s, v_s) \geq \psi_{22}(x).$$

In particular, (3.1) is uniformly forwards permanent.

Proof. We proceed as in the Proposition 3.7 using now that, as $t - s \to \infty$,

$$u \geq \xi_{[\lambda - b\eta_{[\mu,d]},a]} \geq \xi_{[\lambda_I - b_S \eta_{[\mu_I,d_S]},a_S]} \to \omega_{[\lambda_I - b_S \omega_{[\mu_I,d_S]},a_S]}$$

and

$$v \geq \eta_{[\mu - c\xi_{[\lambda,a]},d]} \geq \eta_{[\mu_I - c_S \xi_{[\lambda_I,a_S]},d_S]} \to \omega_{[\mu_I - c_S \omega_{[\lambda_I,a_S]},d_S]}. \qquad \square$$

On the other hand, for pullback permenence and for complete nondegenerate solutions, we have the following proposition along the same lines as that above.

PROPOSITION 3.10 (pullback permanence—symbiotic case). *Assume (3.2), b_M, $c_M < 0$, and (3.9), that is*

$$b_L c_L < a_L d_L.$$

Then

(i) *If $\lambda_I > \Lambda^1(-b_S \omega_{[\mu_I,d_S]})$ there exists $\psi_{11} \in int(\mathcal{C}_1)$ such that whenever*

$$\lambda(t, x) \geq \lambda_I, \quad \mu(t, x) \geq \mu_I, \quad b(t, x) \leq b_S < 0, \quad a(x, t) \leq a_S, \quad d(t, x) \leq d_S.$$

for all $x \in \Omega$ and $t \leq t_0$ (some $t_0 \in \mathbf{R}$), for any $u_s, v_s > 0$, the solution for $s < t \leq t_0$ of (3.1) satisfies $\psi_{11}(x) \leq u(t, s, x; u_s, v_s)$ for $t - s$ large enough.
In particular, any complete trajectory of (3.1) that is nondegenerate at $-\infty$ satisfies $u(t, x) \geq \psi_{11}(x)$ for all $x \in \Omega$ and $t \leq t_0$.

(ii) *If $\mu_I > \Lambda^2(-c_S \omega_{[\lambda_I,a_S]})$ there exists $\psi_{22} \in int(\mathcal{C}_2)$ such that whenever*

$$\lambda(t, x) \geq \lambda_I, \quad \mu(t, x) \geq \mu_I, \quad a(t, x) \leq a_S, \quad d(t, x) \leq d_S, \quad c(t, x) \leq c_S < 0,$$

$x \in \Omega$ and $t \leq t_0$ (some $t_0 \in \mathbf{R}$) for any $u_s, v_s > 0$, the solution for $s < t \leq t_0$ of (3.1) satisfies $\psi_{22}(x) \leq v(t, s, x; u_s, v_s)$ for $t - s$ large enough.
In particular, any complete trajectory of (3.1) that is nondegenerate at $-\infty$ satisfies $v(t, x) \geq \psi_{22}(x)$ for all $x \in \Omega$ and $t \leq t_0$.

Hence, if

$$\lambda_I > \Lambda^1(-b_S \omega_{[\mu_I,d_S]}) \qquad \text{and} \qquad \mu_I > \Lambda^2(-c_S \omega_{[\lambda_I,a_S]}), \tag{3.17}$$

there exist $\psi_{11} \in int(\mathcal{C}_1)$ and $\psi_{22} \in int(\mathcal{C}_2)$ such that whenever

$$\begin{aligned} &\lambda_I \leq \lambda(t, x), \qquad \mu_I \leq \mu(t, x), \qquad a(t, x) \leq a_S, \\ &b(t, x) \leq b_S < 0, \quad c(t, x) \leq c_S < 0, \quad d(t, x) \leq d_S, \end{aligned}$$

for all $x \in \Omega$ and $t \leq t_0$ (some $t_0 \in \mathbf{R}$), and for all nontrivial $u_s \geq 0$, $v_s \geq 0$ in a fixed bounded set, B, of $C(\overline{\Omega})$ bounded away from 0, the set of solutions of (3.1) $\{(u, v),\ s < t \leq t_0,\ (u_s, v_s) \in B\}$ is nondegenerate as $s \to -\infty$ and for all $t - s$ large enough

$$u(t, s, x; u_s, v_s) \geq \psi_{11}(x) \quad \text{and} \quad v(t, s, x; u_s, v_s) \geq \psi_{22}(x).$$

In particular, (3.1) is uniformly pullback permanent and any bounded complete trajectory that is nondegenerate at $-\infty$ satisfies

$$u(t, x) \geq \psi_{11}(x) \quad \text{and} \quad v(t, x) \geq \psi_{22}(x) \quad \text{for all } x \in \Omega \text{ and } t \leq t_0.$$

3.4.3. Prey-predator. We also have for the prey-predator case the following result.

PROPOSITION 3.11 (forwards permanence—prey-predator case). *Assume* (3.2) *and* $b_L > 0$ *and* $c_M < 0$. *Then:*

(i) *If* $\lambda_I > \Lambda^1(-b_S\omega_{[\mu_S - c_I\omega_{[\lambda_S,a_I]},d_I]})$, *there exists* $\psi_{11} \in int(\mathcal{C}_1)$ *such that whenever*

$$\begin{aligned} &\lambda_S \geq \lambda(t,x) \geq \lambda_I, && \mu(t,x) \leq \mu_S, && a_S \geq a(t,x) \geq a_I > 0,\\ &b(t,x) \leq b_S, && c(t,x) \geq c_I, && d(t,x) \geq d_I > 0 \end{aligned}$$

for all $x \in \Omega$ *and* $t \geq t_0$ *(some* $t_0 \in \mathbf{R}$*), for any* $u_s, v_s > 0$ *the solution for* $t > s \geq t_0$ *of* (3.1) *satisfies* $\psi_{11}(x) \leq u(t,s,x;u_s,v_s)$ *for* $t - s$ *large enough.*

(ii) *If* $\mu_I > \Lambda_0^2$, *there exists* $\psi_{22} \in int(\mathcal{C}_2)$ *such that whenever*

$$\mu(t,x) \geq \mu_I, \quad d(x,t) \leq d_S$$

for all $x \in \Omega$ *and* $t \geq t_0$ *(some* $t_0 \in \mathbf{R}$*), for any* $u_s, v_s > 0$ *the solution for* $t > s \geq t_0$ *of* (3.1) *satisfies* $\psi_{22}(x) \leq v(t,s,x;u_s,v_s)$ *for* $t - s$ *large enough.*

Hence, if

$$\lambda_I > \Lambda^1(-b_S\omega_{[\mu_S - c_I\omega_{[\lambda_S,a_I]},d_I]}) \qquad \text{and} \qquad \mu_I > \Lambda_0^2, \tag{3.18}$$

there are functions $\psi_{11} \in int(\mathcal{C}_1)$ *and* $\psi_{22} \in int(\mathcal{C}_2)$ *such that whenever*

$$\begin{aligned} &\lambda_I \leq \lambda(t,x) \leq \lambda_S, && \mu_I \leq \mu(t,x) \leq \mu_S, && a_S \geq a(t,x) \geq a_I > 0,\\ &0 < b_I \leq b(t,x) \leq b_S, && c_I \leq c(t,x) \leq c_S < 0, && d_S \geq d(t,x) \geq d_I > 0 \end{aligned}$$

for all $x \in \Omega$ *and* $t \geq t_0$ *(some* $t_0 \in \mathbf{R}$*), and for all* $u_s > 0, v_s > 0$ *in a fixed bounded set of* $C(\overline{\Omega})$ *bounded away from* 0, *the solution* (u,v) *of* (3.1) *for* $t > s \geq t_0$ *is nondegenerate at* ∞ *and for all* $t - s$ *large enough*

$$u(t,s,x;u_s,v_s) \geq \psi_{11}(x) \qquad \text{and} \qquad v(t,s,x;u_s,v_s) \geq \psi_{22}(x).$$

In particular, (3.1) *is uniformly forwards permanent.*

Proof. As before, we use now that as $t - s \to \infty$,

$$u \geq \xi_{[\lambda - b\eta_{[\mu - c\xi_{[\lambda,a]},d]},a]} \geq \xi_{[\lambda_I - b_S\eta_{[\mu_S - c_I\xi_{[\lambda_S,a_I]},d_I]},a_I]} \to \omega_{[\lambda_I - b_S\eta_{[\mu_S - c_I\omega_{[\lambda_S,a_I]},d_I]},a_I]}$$

and

$$v \geq \eta_{[\mu,d]} \geq \eta_{[\mu_I,d_S]} \to \omega_{[\mu_I,d_S]}. \qquad \square$$

PROPOSITION 3.12 (pullback permanence—prey-predator case). *Assume* (3.2) *and* $b_L > 0$ *and* $c_M < 0$. *Then:*

(i) *If* $\lambda_I > \Lambda^1(-b_S\omega_{[\mu_S - c_I\omega_{[\lambda_S,a_I]},d_I]})$, *there exists* $\psi_{11} \in int(\mathcal{C}_1)$ *such that whenever*

$$\begin{aligned} &\lambda_S \geq \lambda(t,x) \geq \lambda_I, && \mu(t,x) \leq \mu_S, && a_S \geq a(t,x) \geq a_I > 0,\\ &b(t,x) \leq b_S, && c(t,x) \geq c_I, && d(t,x) \geq d_I > 0 \end{aligned}$$

for all $x \in \Omega$ *and* $t \leq t_0$ *(some* $t \in \mathbf{R}$*), for any* $u_s, v_s > 0$, *the solution for* $s < t \leq t_0$ *of* (3.1) *satisfies* $\psi_{11}(x) \leq u(t,s,x;u_s,v_s)$ *for* $t - s$ *large enough. In particular, any complete trajectory of* (3.1) *that is nondegenerate at* $-\infty$ *satisfies* $u(t,x) \geq \psi_{11}(x)$ *for all* $x \in \Omega$ *and* $t \leq t_0$.

(ii) *If $\mu_I > \Lambda_0^2$, there exists $\psi_{22} \in int(\mathcal{C}_2)$ such that whenever*

$$\mu(t,x) \geq \mu_I, \quad d(x,t) \leq d_S$$

for all $x \in \Omega$ and $t \leq t_0$ (some $t_0 \in \mathbf{R}$), for any $u_s, v_s > 0$, the solution for $s < t \leq t_0$ of (3.1) satisfies $\psi_{22}(x) \leq v(t,s,x;u_s,v_s)$ for $t-s$ large enough. In particular, any complete trajectory of (3.1) that is nondegenerate at $-\infty$ satisfies $v(t,x) \geq \psi_{22}(x)$ for all $x \in \Omega$ and $t \leq t_0$.

Hence, if

$$\lambda_I > \Lambda^1(-b_S\omega_{[\mu_S - c_I\omega_{[\lambda_S,a_I]},d_I]}) \qquad \text{and} \qquad \mu_I > \Lambda_0^2, \tag{3.19}$$

there exist functions $\psi_{11} \in int(\mathcal{C}_1)$ and $\psi_{22} \in int(\mathcal{C}_2)$ such that whenever

$$\begin{aligned} &\lambda_I \leq \lambda(t,x) \leq \lambda_S, \qquad \mu_I \leq \mu(t,x) \leq \mu_S, \qquad a_S \geq a(t,x) \geq a_I > 0,\\ &0 < b_I \leq b(t,x) \leq b_S, \quad c_I \leq c(t,x) \leq c_S < 0, \quad d_S \geq d(t,x) \geq d_I > 0, \end{aligned}$$

for all $x \in \Omega$ and $t \leq t_0$ (some $t_0 \in \mathbf{R}$), and for all nontrivial $u_s \geq 0$, $v_s \geq 0$ in a fixed bounded set, B, of $C(\overline{\Omega})$ bounded away from 0, the set of solutions of (3.1) $\{(u,v),\ s < t \leq t_0,\ (u_s,v_s) \in B\}$ is nondegenerate as $s \to -\infty$ and for all $t-s$ large enough

$$u(t,s,x;u_s,v_s) \geq \psi_{11}(x) \quad \text{and} \quad v(t,s,x;u_s,v_s) \geq \psi_{22}(x).$$

In particular, (3.1) is uniformly pullback permanent and any bounded complete trajectory that is nondegenerate at $-\infty$ satisfies

$$u(t,x) \geq \psi_{11}(x) \quad \text{and} \quad v(t,x) \geq \psi_{22}(x) \quad \text{for all } x \in \Omega \text{ and } t \leq t_0.$$

Remark 3.13. Note that in order to apply the previous results one has to check that the assumptions in Propositions 3.7–3.12 are meaningful. Indeed, conditions (3.10), (3.16), and (3.18) must define nonempty sets of coefficients. Here we analyze only Dirichlet or Neumann boundary conditions; Robin ones can be treated in a similar way although the estimates are a little more involved.

In fact, (3.16) includes all coefficients such that

$$\lambda_I > \Lambda_0^1, \qquad \mu_I > \Lambda_0^2$$

since in this case $\lambda_I > \Lambda_0^1 > \Lambda^1(-b_S\omega_{[\mu_I,d_S]})$ and $\mu_I > \Lambda_0^2$ then $\mu_I > \Lambda^2(-c_S\omega_{[\lambda_I,a_S]})$; see also [12].

However, in order to show that (3.10) defines a nonempty set we must impose some conditions on b or c. If, for example, $b_S \to 0$, then $\Lambda^1(-b_S\omega_{[\mu_S,d_I]}) \to \Lambda_0^1$. Also, if $c_S \to 0$, then $\Lambda^2(-c_S\omega_{[\lambda_S,a_I]}) \to \Lambda_0^2$. Hence if b_S or c_S are small, the conditions in (3.10) can be met; see also [27] and [29].

We analyze condition (3.18) for the prey-predator case in more detail. From Proposition 3.1, in the case of Dirichlet or Neumann boundary conditions, we have $\omega_{[h,g]} \leq h_M/g_L$, and so

$$\omega_{[\lambda_S,a_I]} \leq \frac{\lambda_S}{a_I} \quad \text{and then} \quad \omega_{[\mu_S - c_I\omega_{[\lambda_S,a_I]},d_I]} \leq \frac{\mu_S - c_I\left(\frac{\lambda_S}{a_I}\right)}{d_I},$$

and then using the monotonicity of $\Lambda(m)$ with respect to m and (3.5), we get

$$\Lambda^1(-b_S\omega_{[\mu_S - c_I\omega_{[\lambda_S,a_I]},d_I]}) \leq \Lambda^1(-b_S\frac{(a_I\mu_S - c_I\lambda_S)}{a_Id_I}) = \Lambda_0^1 + b_S\frac{(a_I\mu_S - c_I\lambda_S)}{a_Id_I}.$$

Hence, if λ_I and μ_I satisfy

$$\lambda_I > \Lambda_0^1 + \frac{b_S\mu_S}{d_I} + \frac{-b_S c_I}{a_I d_I}\lambda_S, \qquad \mu_I > \Lambda_0^2,$$

then (3.18) defines a nonempty set of parameters.

Observe that the first condition above is a restriction on the oscillation of $\lambda(t,x)$ as $t \to \pm\infty$.

In particular, if $\Lambda_0^1 + \frac{b_S\mu_S}{d_I} > 0$, then a necessary condition is

$$a_I d_I + b_S c_I > 0.$$

In such a case the conditions above can be met.

Now, for reasons that will be apparent in the following sections, we are interested in some uniformity in the previous results with respect to the coefficients $b_I \le b_S$ and $c_I \le c_S$. More precisely, we are going to show that the functions $\psi_{11}(x)$ and $\psi_{22}(x)$ in all the previous propositions can be taken independent of $b(t,x)$ and $c(t,x)$, provided that one of the numbers $b_I \le b_S$ or $c_I \le c_S$ is sufficiently small. In fact we have the following:

THEOREM 3.14. (i) *The competitive case: $b_L, c_L > 0$. Assume either*

1. *$\lambda_I > \Lambda_0^1$, $\mu_I > \Lambda^2(-c_S\omega_{[\lambda_S,a_I]})$, and b_S is sufficiently small, or*
2. *$\lambda_I > \Lambda^1(-b_S\omega_{[\mu_S,d_I]})$, $\mu_I > \Lambda_0^2$, and c_S is sufficiently small, for $t-s$ large enough.*

Then the functions $\psi_{11}(x)$ and $\psi_{22}(x)$ in Propositions 3.7 *and* 3.8 *can be taken also independent of b_S and c_S.*

(ii) *The symbiotic case: $b_M, c_M < 0$ and $b_L c_L < a_L d_L$. Assume either*

1. *$\lambda_I > \Lambda_0^1$, $\mu_I > \Lambda^2(-c_S\omega_{[\lambda_I,a_S]})$, or*
2. *$\lambda_I > \Lambda^1(-b_S\omega_{[\mu_I,d_S]})$, $\mu_I > \Lambda_0^2$.*

Then the functions $\psi_{11}(x)$ and $\psi_{22}(x)$ in Propositions 3.9 *and* 3.10 *can be taken also independent of b_I and c_I.*

(iii) *The prey-predator case: $b_L > 0$, $c_M < 0$. Assume either*

1. *$\lambda_I > \Lambda_0^1$, $\mu_I > \Lambda_0^2$, and b_S is sufficiently small, or*
2. *$\lambda_I > \Lambda^1(-b_S\omega_{[\mu_S,d_I]})$, $\mu_I > \Lambda_0^2$, and $|c_I|$ is sufficiently small.*

Then the functions $\psi_{11}(x)$ and $\psi_{22}(x)$ in Propositions 3.11 *and* 3.12 *can be taken also independent of b_S and c_I.*

Proof. We analyze only the competitive case. By the proof of Proposition 3.7 and Theorem 3.3, statement 6 we get

$$\begin{aligned} u(t,s,u_s,v_s) &\ge \Theta_{[\lambda_I - b_S\eta_{[\mu_S,d_I]},a_S]}(t-s,u_s) \ge \\ &\ge \omega_{[\lambda_I - b_S(\omega_{[\mu_S,d_I]}+\varepsilon),a_S]} \ge m\omega_{[\lambda_I - b_S(\omega_{[\mu_S,d_I]}+\varepsilon),a_S]}. \end{aligned}$$

It suffices to apply Lemma 3.2 as $b_S \to 0$ where $m < 1$.

On the other hand,

$$v(t,s,u_s,v_s) \ge \Theta_{[\mu_I - c_S(\omega_{[\lambda_S,a_I]}+\varepsilon),d_S]}(t-s,v_s) \to \omega_{[\mu_I - c_S(\omega_{[\lambda_S,a_I]}+\varepsilon),d_S]},$$

and so taking ε small, the result follows.

The other cases can be studied in an analogous way by Propositions 3.9, 3.10, 3.11, and 3.12. ∎

4. Exponential decay for nonautonomous linear systems. Once the results on permanence of the previous section have been established, we turn now our

attention to determining ranges of parameters such that there exist some special asymptotically stable trajectories describing the asymptotic behavior of solutions of (3.1), either forwards or in a pullback sense. For this we have to develop some tools for linear systems.

Hence, in this section we give sufficient conditions for certain linear systems to have exponential decay. The results are of a perturbative nature and are based upon results in [33] for scalar equations.

4.1. Preliminary results for the scalar case. We start by recalling some results for the following scalar equation:

$$\begin{cases} u_t - d\Delta u = c(t,x)u & x \in \Omega,\ t > s \\ \mathcal{B}u = 0, & x \in \partial\Omega,\ t > s \\ u(s) = u_s. \end{cases} \tag{4.1}$$

Assume that $d > 0$, $c \in C^\theta(\mathbf{R}, L^p(\Omega))$, with $0 < \theta \le 1$ and some $p > \max(N/2, 1)$. Then for any $u_s \in X$, where $X = L^q(\Omega)$ with $1 \le q < \infty$, or $X = C(\overline{\Omega})$, (4.1) has a unique solution given by $u(t, s; u_s)$, which is a strong solution in $L^r(\Omega)$ for any $1 \le r < p$. This solution can be used to define an order-preserving evolution operator T_c in X via the definition $T_c(t,s)u_s = u(t,s;u_s)$.

Moreover for each q and r with $1 \le q \le r \le \infty$ and $R_0 > 0$ there exist $L_0 = L_0(R_0, r, q) > 0$ and $\delta_0 = \delta(R_0, r, q) > 0$ such that the evolution operator $T_c(t,s)$ satisfies

$$\|T_c(t,s)u_0\|_{L^r(\Omega)} \le L_0 \frac{\mathrm{e}^{\delta_0(t-s)}}{(t-s)^{\frac{N}{2}\left(\frac{1}{q}-\frac{1}{r}\right)}} \|u_0\|_{L^q(\Omega)}, \qquad t > s \tag{4.2}$$

for every $c \in C^\theta(\mathbf{R}, L^p(\Omega))$, with $0 < \theta \le 1$ and some $p > N/2$, such that

$$\|c\|_{L^\infty(\mathbf{R}, L^p(\Omega))} \le R_0.$$

Also, the evolution operator smooths the solutions. More precisely, for every $u_0 \in L^q(\Omega)$ and $t > s$, the map

$$(s, \infty) \ni t \longmapsto u(t,s;u_0) := T_c(t,s)u_0 \in \begin{cases} C^\nu_{\mathcal{B}}(\overline{\Omega}) & \text{if } p > N/2, \\ C^{1,\nu}_{\mathcal{B}}(\overline{\Omega}) & \text{if } p > N, \end{cases}$$

is continuous for some $\nu > 0$. Here

$$C^{j,\nu}_{\mathcal{B}}(\overline{\Omega}) = \begin{cases} C^{j,\nu}_0(\overline{\Omega}) & \text{for Dirichlet BCs,} \\ C^{j,\nu}(\overline{\Omega}) & \text{for Neumann or Robin BCs,} \end{cases}$$

see, e.g., Rodríguez-Bernal [32].

The following proposition is taken from Lemma 4.1 in Robinson et al. [31] and Lemma 2.1 in Rodríguez-Bernal [33]:

PROPOSITION 4.1. *Suppose that for some q with $1 \le q \le \infty$ there exist $M > 0$ and $\beta \in \mathbf{R}$ such that*

$$\|T_c(t,s)\|_{\mathcal{L}(L^q(\Omega))} \le M\mathrm{e}^{\beta(t-s)} \quad \textit{for all} \quad t > s. \tag{4.3}$$

Then for any $1 \le r \le \infty$ there exists a $K \ge 1$ such that

$$\|T_c(t,s)\|_{\mathcal{L}(L^r(\Omega))} \le K\mathrm{e}^{\beta(t-s)} \quad \textit{for all} \quad t > s. \tag{4.4}$$

The constant K can be taken as a continuous function of β, M.

Moreover, for each r with $1 \le r \le q \le \infty$ and for any $\varepsilon > 0$, we have

$$\|T_c(t,s)\|_{\mathcal{L}(L^q(\Omega),L^r(\Omega))} \le M(\beta,\varepsilon)\frac{\mathrm{e}^{(\beta+\varepsilon)(t-s)}}{(t-s)^\delta}, \quad t > s, \tag{4.5}$$

where $\delta = \frac{N}{2}\left(\frac{1}{r} - \frac{1}{q}\right)$,

$$M(\beta,\varepsilon) = \kappa(\beta,M)\begin{cases} \left(\frac{\delta}{\mathrm{e}}\right)^\delta \varepsilon^{-\delta} & \text{if } 0 < \varepsilon < \varepsilon_0 = \frac{\delta}{\mathrm{e}} \\ 1 & \text{if } \varepsilon \ge \varepsilon_0 = \frac{\delta}{\mathrm{e}} \end{cases} \tag{4.6}$$

and

$$\kappa(\beta,M) = L_0\varepsilon^{\delta_0}\max\{1, M\varepsilon^{-\beta}\}.$$

Note that the constants K and κ in the proposition also depend on q and r but we will not pay attention to this dependence.

Our main argument, further below in the paper, will rely on results of the following type. We start with an evolution operator $T_c(t,s)$ that satisfies the estimate

$$\|T_c(t,s)\|_{\mathcal{L}(L^q(\Omega))} \le M_1 \quad \text{for } t \ge s \text{ and } M_1 > 0$$

for either $s \ge s_0$ or for $t \le t_0$. Then, we add to $c(t,x)$ a perturbation $p(t,x)$ in the class $C^\theta(\mathbf{R}, L^p(\Omega))$, with $0 < \theta \le 1$ and some $p > \max(N/2, 1)$, and we want to guarantee that the solutions of the new evolution operator $T_{c+p}(t,s)$ decay exponentially. This means that we want to get estimates of the type

$$\|T_{c+p}(t,s)\|_{\mathcal{L}(L^q(\Omega))} \le M_1'\mathrm{e}^{\beta'(t-s)} \quad \text{for all} \quad t > s \quad \text{and some} \quad \beta' < 0 \tag{4.7}$$

and for either $s \ge s_0$ or for $t \le t_0$. Note also that we can always assume, without loss of generality, that the $L^\infty(\mathbf{R}, L^p(\Omega))$ norms of both $c(t,x)$ and $p(t,x)$ are bounded by R_0, so (4.2) holds for $T_c(t,s)$ and $T_{c+p}(t,s)$.

In this direction, the following important result is a particular case of Corollary 3.3 in Rodríguez-Bernal [33], and it provides sufficient conditions on $p(t,x)$ to ensure that (4.7) holds.

PROPOSITION 4.2. *Assume that*

$$\|T_c(t,s)\|_{\mathcal{L}(L^q(\Omega))} \le M_1 \quad \textit{for} \quad t \ge s \quad \textit{and} \quad M_1 > 0 \tag{4.8}$$

and for either $s \ge s_0$ or for $t \le t_0$.

Let $p \in C^\theta(\mathbf{R}, L^p(\Omega))$, for some $0 < \theta \le 1$ and $p > \max(N/2, 1)$, and assume that for $|t|$ sufficiently large, we have $p(t,x) \le -\varphi(x)$ where

$$\varphi \in C^1(\overline{\Omega}), \quad \varphi \ge 0, \quad \textit{and} \quad \nabla\varphi \ne 0 \quad \textit{at the points at which } \varphi = 0.$$

Then

$$\|T_{c+p}(t,s)\|_{\mathcal{L}(L^q(\Omega))} \le M_1'\mathrm{e}^{\beta'(t-s)} \quad \textit{for all} \quad t > s \quad \textit{and some} \quad \beta' < 0 \tag{4.9}$$

and for either $s \ge s_0$ or for $t \le t_0$, with $M_1' = M_1'(M_1,\varphi)$ and $\beta' = \beta'(M_1,\varphi)$.

The constants $M_1' = M_1'(M_1,\varphi)$ and $\beta' = \beta'(M_1,\varphi)$ depend continuously on M_1 and on $\varphi \in C^1(\overline{\Omega})$.

Note that the condition above holds, in particular if $p(t,x) \le -\delta < 0$ (in which case the constants M_1' and β' can be chosen so that they depend continuously on δ), or if $\varphi \in C_0^1(\overline{\Omega})$ is positive in Ω and $\frac{\partial \varphi}{\partial \vec{n}} < 0$ on $\partial\Omega$. The former is a common situation in the case of Neumann or Robin boundary conditions and the latter in the case of Dirichlet boundary conditions.

In order to apply the above result, we need to show first that (4.8) holds. The next result gives conditions for an evolution operator to have bounds of the type (4.8); see [34], [33]. For this recall the definitions of complete trajectory and of nondegeneracy in section 2.2, which we apply here to solutions of (4.1). Hence, according to [33], we have the following proposition.

PROPOSITION 4.3. (i) *If there exists a positive nondegenerate solution $u(t,s;u_s)$ of (4.1) defined for all $t > s \ge s_0$ such that for some $M > 0$ and some q with $1 \le q \le \infty$*

$$\|u(t,s;u_s)\|_{L^q(\Omega)} \le M,$$

then

$$0 < M_0 \le \|T_c(t,s)\|_{\mathcal{L}(L^q(\Omega))} \le M_1 \quad \text{for} \quad t \ge s \ge s_0, \tag{4.10}$$

where M_0, M_1 are independent of t and s and depend continuously on M and on $\varphi_0 \in C^1(\overline{\Omega})$.

(ii) *If there exists a positive, complete nondegenerate solution $u(t)$ of (4.1) that is bounded as $t \to -\infty$, i.e.,*

$$\|u(t)\|_{L^q(\Omega)} \le M \quad \text{for} \quad t \le t_0,$$

then

$$0 < M_0 \le \|T_c(t,s)\|_{\mathcal{L}(L^q(\Omega))} \le M_1 \quad \text{for} \quad s \le t \le t_0, \tag{4.11}$$

where M_0, M_1 are independent of t and s and depend continuously on M and $\varphi_0 \in C^1(\overline{\Omega})$.

4.2. Perturbation and decay of linear systems. In this section we generalize the perturbation result in the previous section to the case of a system of linear equations. The main theorem in this section will be crucial in the analysis of Lotka–Volterra models in the following sections.

Consider the linear coupled nonautonomous system

$$\begin{cases} u_t - d_1\Delta u = a_{11}(t,x)u + a_{12}(t,x)v, & x \in \Omega,\ t > s \\ v_t - d_2\Delta v = a_{21}(t,x)u + a_{22}(t,x)v, & x \in \Omega,\ t > s \\ \mathcal{B}_1 u = 0,\ \mathcal{B}_2 v = 0 \\ u(s) = u_s,\ v(s) = v_s, \end{cases} \tag{4.12}$$

in $L^q(\Omega, \mathbf{R}^2) \doteq [L^q(\Omega)]^2$. Then define

$$D = \operatorname{diag}(d_1, d_2) \qquad \text{and} \qquad A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

and note that setting $U = \binom{u}{v}$, (4.12) can be written as

$$U_t - D\Delta U = A(t,x)U$$

with boundary conditions $\mathcal{B}U = \binom{\mathcal{B}_1 u}{\mathcal{B}_2 v} = 0$ on the boundary of Ω.

If $A \in C^\theta(\mathbf{R}, L^p(\Omega, \mathbf{R}^4))$, with $0 < \theta \leq 1$, $p > N/2$, and $p > q \geq 1$, the existence of a unique solution $U(t, s; U_s)$ of (4.12), in $L^q(\Omega, \mathbf{R}^2)$, can be obtained from Theorems 11.2, 11.3, and 11.4 in Amann [1]. Thus, the time-dependent operator $-D\Delta - A(t, x)$ generates an evolution operator, $T_A(t, s)$, in $L^q(\Omega, \mathbf{R}^2)$ (Theorem 4.4.1 in Amann [2]) via the definition $T_A(t, s)U_s = U(t, s; U_s)$.

The following result, analogous to (4.2), can be proved along the lines of the scalar arguments in Rodríguez-Bernal [33], [32], and Robinson et al. [31].

PROPOSITION 4.4. *For any $1 \leq q \leq r \leq \infty$, and $R_0 > 0$ there exist $L_0 = L_0(R_0, r, q) > 0$ and $\delta_0 = \delta_0(R_0, r, q) > 0$ such that the evolution operator $T_A(t, s)$ satisfies*

$$\|T_A(t, s)U_s\|_{L^r(\Omega, \mathbf{R}^2)} \leq L_0 \frac{\mathrm{e}^{\delta_0(t-s)}}{(t-s)^{\frac{N}{2}\left(\frac{1}{q}-\frac{1}{r}\right)}} \|U_s\|_{L^q(\Omega, \mathbf{R}^2)}, \tag{4.13}$$

for every $\|A\|_{L^\infty(\mathbf{R}, L^p(\Omega, \mathbf{R}^4))} \leq R_0$. In particular, $T_A(t, s)$ extends to an evolution operator in $L^q(\Omega, \mathbf{R}^2)$ for every $1 \leq q < \infty$.

Furthermore, the results of Proposition 4.1 *for the scalar case remain true for system* (4.12).

Along the same lines as for scalar equations, we consider the linear uncoupled system

$$\begin{cases} u_t - d_1\Delta u = q_{11}(t, x)u, & x \in \Omega,\ t > s \\ v_t - d_2\Delta v = q_{22}(t, x)v, & x \in \Omega,\ t > s \\ \mathcal{B}_1 u = 0,\ \mathcal{B}_2 v = 0 \\ u(s) = u_s,\ v(s) = v_s. \end{cases} \tag{4.14}$$

Observe that with the notations above and setting

$$Q = \operatorname{diag}(q_{11}, q_{22}),$$

then the evolution operator $T_Q(t, s)$ is well defined in $L^q(\Omega, \mathbf{R}^2)$, $1 \leq q < \infty$.

Now we assume that each separate equation in (4.14) satisfies

$$\|T_{q_{ii}}(t, s)\|_{\mathcal{L}(L^q(\Omega))} \leq M_1, \quad t > s,$$

with M_1 independent of t and s and for either $t \leq t_0$ or $s \geq s_0$. Therefore the evolution operator $T_Q(t, s)$ satisfies (4.8).

Our goal is to give conditions on the coupling perturbations such that the solutions of the perturbed system

$$\begin{cases} u_t - d_1\Delta u = q_{11}(t, x)u + p_{11}(t, x)u + p_{12}(t, x)v, & x \in \Omega,\ t > s \\ v_t - d_2\Delta v = q_{22}(t, x)v + p_{21}(t, x)u + p_{22}(t, x)v, & x \in \Omega,\ t > s \\ \mathcal{B}_1 u = 0,\ \mathcal{B}_2 v = 0 \\ u(s) = u_s,\ v(s) = v_s, \end{cases}$$

decay exponentially. Note that the perturbed system can be written as

$$U_t - D\Delta U = Q(t, x)U + P(t, x)U \tag{4.15}$$

with

$$Q = \operatorname{diag}(q_{11}, q_{22}), \quad P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}, \quad \text{and} \quad U = \begin{pmatrix} u \\ v \end{pmatrix},$$

with $Q, P \in C^\theta(\mathbf{R}, L^p(\Omega, \mathbf{R}^4))$, with $0 < \theta \leq 1$, $p > \max(N/2, 1)$.

Hence our goal is to obtain an estimate of the type

$$\|T_{Q+P}(t,s)\|_{\mathcal{L}(L^q(\Omega,\mathbf{R}^2))} \le M_1' \mathrm{e}^{\beta'(t-s)} \quad \text{for all} \quad t > s \text{ and some } \beta' < 0 \tag{4.16}$$

and for either $s \ge s_0$ or for $t \le t_0$.

Note that again we will assume, without loss of generality, that all the evolution operators considered satisfy (4.13) with the same constants L_0 and δ_0.

In what follows we will make use of the following singular Gronwall lemma (see Henry [16]):

LEMMA 4.5 (a singular Gronwall lemma). *Assume that $a \in L^\infty(\tau_0,\infty)$ with $\tau_0 \ge -\infty$ and that $z(t) \ge 0$ is a locally bounded function that for $t \ge s > \tau_0$ satisfies*

$$z(t) \le A + \int_s^t \frac{a(\tau)}{(t-\tau)^\delta} z(\tau)\,\mathrm{d}\tau \tag{4.17}$$

with $\delta < 1$. Then we have for $t \ge s > \tau_0$

$$0 \le z(t) \le A(\delta)\mathrm{e}^{\gamma(t-s)}$$

with $\gamma = \gamma(a,s,\delta) = (\|a\|_{L^\infty(s,\infty)}\Gamma(1-\delta))^{1/(1-\delta)}$ and $A(\delta)$ depends only on the constants A and δ but not on the function $a(\cdot)$ or on s, γ, or τ_0.

Our next result states that if the diagonal perturbing terms $p_{ii}(t,x)$ are sufficiently strong and the coupling terms $p_{ij}(t,x)$, $i \ne j$, are "small" at $\pm\infty$, then (4.16) is achieved.

THEOREM 4.6. *With the notations in* (4.15)*, assume that the scalar evolution operators $T_{q_{ii}}(t,s)$ satisfy*

$$\|T_{q_{ii}}(t,s)\|_{\mathcal{L}(L^q(\Omega))} \le M_1, \quad t > s, \tag{4.18}$$

with M_1 independent of t and s and for either $t \le t_0$ or $s \ge s_0$.

Assume also that $p_{ii}(t,x)$ satisfies $p_{ii}(t,x) \le -\varphi_{ii}(x)$ with $\varphi_{ii}(x)$ as in Proposition 4.2*.*

Then there exists a $\rho = \rho(M_1,\varphi_{11},\varphi_{22}) > 0$ such that if

$$\limsup_{|t|\to\infty} \|p_{12}(t)\|_{L^p(\Omega)} \limsup_{|t|\to\infty} \|p_{21}(t)\|_{L^p(\Omega)} \le \rho^2, \tag{4.19}$$

then

$$\|T_{Q+P}(t,s)\|_{\mathcal{L}(L^q(\Omega,\mathbf{R}^2))} \le M_1'' \mathrm{e}^{\beta''(t-s)} \quad \textit{for all } t > s \textit{ and some } \beta'' < 0 \tag{4.20}$$

and for either $s \ge s_0$ or for $t \le t_0$, where $M_1'' = M_1''(M_1,\varphi_{11},\varphi_{22})$ and $\beta'' = \beta''(M_1,\varphi_{11},\varphi_{22})$.

The constants ρ, M_1'', and β'' depend continuously on M_1 and $\varphi_{11},\varphi_{22}$ as in Proposition 4.2*.*

Proof. Note that, using Proposition 4.4, we just need to prove the result for some suitably chosen $1 \le q < \infty$. We proceed in several steps.

Step 1. If we define

$$P_1 = \begin{pmatrix} p_{11} & 0 \\ 0 & p_{22} \end{pmatrix},$$

then Proposition 4.2 applied to each separate equation gives the estimate

$$\|T_{Q+P_1}(t,s)\|_{\mathcal{L}(L^q(\Omega,\mathbf{R}^2))} \le M_1' \mathrm{e}^{\beta'(t-s)} \quad \text{for all } t > s \text{ and some } \beta' < 0 \tag{4.21}$$

and for either $s \geq s_0$ or for $t \leq t_0$, with $M_1' = M_1'(M_1, \varphi_{11}, \varphi_{22})$ and $\beta' = \beta'(M_1, \varphi_{11}, \varphi_{22})$.

Step 2. We will show that there exists a $\rho = \rho(M_1', \beta')$, which depends continuously on M_1', β', such that if

$$\|p_{12}\|_{L^\infty(\mathbf{R}, L^p(\Omega))} \leq \rho \quad \text{and} \quad \|p_{21}\|_{L^\infty(\mathbf{R}, L^p(\Omega))} \leq \rho,$$

then

$$\|T_{Q+P}(t,s)\|_{\mathcal{L}(L^q(\Omega,\mathbf{R}^2))} \leq M_1'' \mathrm{e}^{\beta''(t-s)} \quad \text{for all } t > s \text{ and some } \beta'' < 0 \tag{4.22}$$

and for either $s \geq s_0$ or for $t \leq t_0$, with

$$P = P_1 + P_2, \qquad P_2 = \begin{pmatrix} 0 & p_{12} \\ p_{21} & 0 \end{pmatrix},$$

where $M_1'' = M_1''(M_1', \beta', \rho)$ and $\beta'' = \beta''(M_1', \beta', \rho)$, depend continuously on M_1', β', ρ.

In fact, we have, by the variation of constants formula, that for every $U_0 \in L^q(\Omega, \mathbf{R}^2)$ the solution $U(t,s;U_0) = T_{Q+P}(t,s)U_0$ of (4.15) satisfies for $t \geq s$,

$$U(t,s;U_0) = T_{Q+P_1}(t,s)U_0 + \int_s^t T_{Q+P_1}(t,\tau)P_2(\tau)U(\tau,s;U_0)\,\mathrm{d}\tau.$$

Now we choose q such that $p \geq q'$, so that $1/p + 1/q \leq 1$. In what follows we will apply (4.5) with $\frac{1}{r} = \frac{1}{p} + \frac{1}{q}$, and so with $\delta = N/2p$. With this choice, we have (4.21) and from (4.5)

$$\|T_{Q+P_1}(t,s)\|_{\mathcal{L}(L^r(\Omega,\mathbf{R}^2), L^q(\Omega,\mathbf{R}^2))} \leq M(\beta', \varepsilon)\, \frac{\mathrm{e}^{(\beta'+\varepsilon)(t-s)}}{(t-s)^{\frac{N}{2p}}},$$

where $M(\beta', \varepsilon)$ is as in (4.6).

Since $P_2(\tau) \in L^p(\Omega, \mathbf{R}^2)$ and $U(\tau, s; U_0) \in L^q(\Omega, \mathbf{R}^2)$, then the term $P_2(\tau)U(\tau, s; U_0)$ can be estimated, using Hölder's inequality, in $L^r(\Omega, \mathbf{R}^2)$ with $\frac{1}{r} = \frac{1}{p} + \frac{1}{q}$. Thus,

$$\begin{aligned} \|U(t,s;U_0)\|_{L^q(\Omega,\mathbf{R}^2)} &\leq M_1' \mathrm{e}^{(\beta'+\varepsilon)(t-s)} \|U_0\|_{L^q(\Omega,\mathbf{R}^2)} \\ &+ M(\beta', \varepsilon) \int_s^t \frac{\mathrm{e}^{(\beta'+\varepsilon)(t-\tau)}}{(t-\tau)^{\frac{N}{2p}}} \|P_2(\tau)\|_{L^p(\Omega,\mathbf{R}^2)} \|U(\tau,s;U_0)\|_{L^q(\Omega,\mathbf{R}^2)}\,\mathrm{d}\tau. \end{aligned}$$

Then, multiplying by $\mathrm{e}^{-(\beta'+\varepsilon)(t-s)}$, and setting $A = M_1' \|U_0\|_{L^q(\Omega,\mathbf{R}^2)}$,

$$z(t) = \mathrm{e}^{-(\beta'+\varepsilon)(t-s)} \|U(t,s;U_0)\|_{L^q(\Omega,\mathbf{R}^2)}, \quad \text{and} \quad a(\tau) = M(\beta', \varepsilon)\|P_2(\tau)\|_{L^p(\Omega,\mathbf{R}^2)}$$

we get, for all $t \geq s$,

$$z(t) \leq A + \int_s^t \frac{a(\tau)}{(\tau - s)^{\frac{N}{2p}}} z(\tau)\,\mathrm{d}\tau.$$

We can apply the singular Gronwall lemma above with $\delta = \frac{N}{2p} < 1$, and we get

$$\|U(t,s;U_0)\|_{L^q(\Omega,\mathbf{R}^2)} \leq M_1'' \mathrm{e}^{(\beta'+\mu(\varepsilon))(t-s)} \|U_0\|_{L^q(\Omega,\mathbf{R}^2)}, \quad t \geq s, \tag{4.23}$$

where

$$\mu(\varepsilon) = \varepsilon + \left(M(\beta', \varepsilon)\Gamma(1-\delta)\|P_2\|_{L^\infty((s,\infty),L^p(\Omega,\mathbf{R}^2))}\right)^{\frac{1}{1-\delta}}.$$

Recalling (4.6), we get that

$$\mu(\varepsilon) = \begin{cases} \varepsilon + \varepsilon^{\frac{-\delta}{1-\delta}} A_0 \|P_2\|^{\frac{1}{1-\delta}}_{L^\infty((s,\infty),L^p(\Omega,\mathbf{R}^2))} & \text{if } 0 < \varepsilon < \varepsilon_0 = \frac{\delta}{\mathrm{e}} \\ \varepsilon + A_1 \|P_2\|^{\frac{1}{1-\delta}}_{L^\infty((s,\infty),L^p(\Omega,\mathbf{R}^2))} & \text{if } \varepsilon \geq \varepsilon_0, \end{cases}$$

where

$$A_1 = \left(L_0 \mathrm{e}^{\delta_0} \max\{1, M_1' \mathrm{e}^{-\beta'}\}\Gamma(1-\delta)\right)^{1/(1-\delta)}, \quad A_0 = A_1 \left(\frac{\delta}{e}\right)^{\delta/(1-\delta)},$$

and L_0 and δ_0 are the constants in (4.2).

Thus $\mu(0) = \mu(\infty) = \infty$. But the function

$$h(\varepsilon) = \varepsilon + \varepsilon^{\frac{-\delta}{1-\delta}} A_0 \|P_2\|^{\frac{1}{1-\delta}}_{L^\infty((s,\infty),L^p(\Omega,\mathbf{R}^2))}$$

has a unique minimum at

$$\varepsilon_1 = \left(A_0 \frac{\delta}{1-\delta}\right)^{1-\delta} \|P_2\|_{L^\infty((s,\infty),L^p(\Omega,\mathbf{R}^2))},$$

and

$$h(\varepsilon_1) = \frac{1}{\delta^\delta}\left(\frac{A_0}{1-\delta}\right)^{1-\delta} \|P_2\|_{L^\infty((s,\infty),L^p(\Omega,\mathbf{R}^2))}.$$

Therefore, comparing ε_0 and ε_1, and minimizing $\mu(\varepsilon)$ leads to

$$\|U(t,s;U_0)\|_{L^q(\Omega,\mathbf{R}^2)} \leq M_1'' \mathrm{e}^{\beta''(t-s)} \|U_0\|_{L^q(\Omega,\mathbf{R}^2)}, \quad t \geq s$$

with

$$\begin{aligned} \beta'' &= \beta' + \min_{\{\varepsilon>0\}} \mu(\varepsilon) \\ &= \beta' + \begin{cases} c_0 \|P_2\|_{L^\infty((s,\infty),L^p(\Omega,\mathbf{R}^2))}, & \text{if } \|P_2\|_{L^\infty((s,\infty),L^p(\Omega,\mathbf{R}^2))} \leq s^*, \\ c_1 + c_2 \|P_2\|^{\frac{1}{1-\delta}}_{L^\infty((s,\infty),L^p(\Omega,\mathbf{R}^2))}, & \text{if } \|P_2\|_{L^\infty((s,\infty),L^p(\Omega,\mathbf{R}^2))} \geq s^*, \end{cases} \end{aligned}$$

where

$$c_0 = \frac{1}{\delta^\delta}\left(\frac{A_0}{1-\delta}\right)^{1-\delta}, \quad c_1 = \frac{\delta}{e}, \quad c_2 = A_1, \quad s^* = \frac{\delta}{e}\left(\frac{1-\delta}{A_0\delta}\right)^{1-\delta}.$$

Thus, it is then clear that (4.22) follows, i.e., $\beta'' < 0$, provided that

$$\|P_2\|_{L^\infty((s,\infty),L^p(\Omega,\mathbf{R}^2))} < \min\left\{s^*, \frac{-\beta'}{c_0}\right\},$$

which reads

$$\|P_2\|_{L^\infty((s,\infty),L^p(\Omega,\mathbf{R}^2))} < \rho := \delta\left(\frac{1-\delta}{\delta A_0}\right)^{1-\delta} \min\left\{-\beta', \frac{1}{\mathrm{e}}\right\}. \tag{4.24}$$

Step 3. Now we show that the result in Step 2 above can be obtained only in terms of $\limsup_{|t|\to\infty} \|P_2(t)\|_{L^p(\Omega,\mathbf{R}^2)}$.

In fact, note that from (4.24), if we take $s \geq s_0$ sufficiently large, the conclusion with $\limsup_{t\to\infty} \|P_2(t)\|_{L^p(\Omega,\mathbf{R}^2)}$ is clear.

On the other hand, observe that we can set $P_2 = 0$ for $t \geq t_0$ and we still have (4.23) for $s \leq t \leq t_0$. Taking then t_0 very negative, (4.24) gives the result for $\limsup_{t\to-\infty} \|P_2(t)\|_{L^p(\Omega,\mathbf{R}^2)}$.

In particular, (4.22) follows, provided that

$$\limsup_{|t|\to\infty} \|P_2(t)\|_{L^p(\Omega,\mathbf{R}^2)} < \rho, \tag{4.25}$$

with ρ as in (4.24).

Step 4. The change of variables

$$U = \begin{pmatrix} u \\ v \end{pmatrix} \to V = \begin{pmatrix} \alpha u \\ \beta v \end{pmatrix}$$

with $\alpha, \beta > 0$, transforms the system (4.15) into

$$V_t - D\Delta V = Q(t,x)V + \tilde{P}(t,x)V$$

with

$$D = \operatorname{diag}(d_1, d_2),\ Q = \operatorname{diag}(q_{11}, q_{22}),\ \tilde{P} = \begin{pmatrix} p_{11} & \frac{\alpha}{\beta}p_{12} \\ \frac{\beta}{\alpha}p_{21} & p_{22} \end{pmatrix}.$$

Hence, we can apply Step 3 provided

$$\frac{\alpha}{\beta}\limsup_{|t|\to\infty} \|p_{12}(t)\|_{L^p(\Omega)} \leq \rho \quad \text{and} \quad \frac{\beta}{\alpha}\limsup_{|t|\to\infty} \|p_{21}(t)\|_{L^p(\Omega)} \leq \rho,$$

with $\rho > 0$ as in (4.24). We can choose α, β such that the above inequalities are satisfied if

$$\limsup_{|t|\to\infty} \|p_{12}(t)\|_{L^p(\Omega)} \limsup_{|t|\to\infty} \|p_{21}(t)\|_{L^p(\Omega)} \leq \rho^2$$

with $\rho > 0$ as in (4.24). □

Remark 4.7. Note that (4.24) gives a quantitative threshold for the size of the perturbation. In fact, from (4.24) and the expression of A_0, it can be deduced that

$$\rho = \rho(M_1', \beta') = \frac{\mathrm{e}^{\delta}(1-\delta)^{1-\delta}}{\Gamma(1-\delta)} \frac{\min\{-\beta', \frac{1}{\mathrm{e}}\}}{L_0 \mathrm{e}^{\delta_0} \max\{1, M_1' \mathrm{e}^{-\beta'}\}},$$

where M_1', β' are from Step 1.

Observe that Step 4 above is the only place where we used the fact that the system has only two components.

5. Attracting trajectories for general nonautonomous nonlinear systems. In this section we sketch out our approach to the existence of asymptotically stable complete trajectories for Lotka–Volterra systems. The key point is to write the equation satisfied by the difference of two solutions as a perturbation of an associated

linear system. Using then the permanence results in section 3, we can apply Theorem 4.6 to conclude that the difference of two solutions converges to zero as $t \to \infty$. A similar convergence result as the initial time $s \to -\infty$ will imply the uniqueness of complete nondegenerate solutions, which moreover describes the pullback behavior of the system.

First we treat the case of general nonautonomous nonlinear systems, before specializing to Lotka–Volterra models. Consider the general nonautonomous nonlinear system

$$\begin{cases} u_t - d_1\Delta u = uf(t,x,u,v) & x \in \Omega,\ t > s \\ v_t - d_2\Delta v = vg(t,x,u,v) & x \in \Omega,\ t > s \\ \mathcal{B}_1 u = 0,\ \mathcal{B}_2 v = 0 & x \in \partial\Omega,\ t > s \\ u(s) = u_s,\ v(s) = v_s. \end{cases} \tag{5.1}$$

We now sketch our strategy for analyzing the asymptotic behavior of solutions to (5.1). Consider two different pairs of nonnegative initial conditions (u_s^1, v_s^1) and (u_s^2, v_s^2) and consider the corresponding solutions of (5.1), $U_1 = \binom{u_1}{v_1}$ and $U_2 = \binom{u_2}{v_2}$, respectively. Write $y = u_2 - u_1$ and $z = v_2 - v_1$. Then, (y, z) satisfies

$$\begin{cases} y_t - d_1\Delta y = q_{11}(t,x)y + p_{11}(t,x)y + p_{12}(t,x)z & x \in \Omega,\ t > s \\ z_t - d_2\Delta z = q_{22}(t,x)z + p_{21}(t,x)y + p_{22}(t,x)z & x \in \Omega,\ t > s \\ \mathcal{B}_1 y = 0,\ \mathcal{B}_2 z = 0 & x \in \partial\Omega,\ t > s \\ y(s) = y_s,\ z(s) = z_s, \end{cases} \tag{5.2}$$

with $y_s = u_s^2 - u_s^1$, $z_s = v_s^2 - v_s^1$, and

$$\begin{aligned} q_{11}(t,x) &= f(t,x,u_2,v_2), \qquad q_{22}(t,x) = g(t,x,u_2,v_2) \\ p_{11}(t,x) &= u_1 \frac{f(t,x,u_2,v_1) - f(t,x,u_1,v_1)}{u_2 - u_1}, \\ p_{12}(t,x) &= u_1 \frac{f(t,x,u_2,v_2) - f(t,x,u_2,v_1)}{v_2 - v_1}, \\ p_{21}(t,x) &= v_1 \frac{g(t,x,u_2,v_1) - g(t,x,u_1,v_1)}{u_2 - u_1}, \\ p_{22}(t,x) &= v_1 \frac{g(t,x,u_2,v_2) - g(t,x,u_2,v_1)}{v_2 - v_1}. \end{aligned}$$

Most of the analysis that follows in the next section will be based on proving that the following results can be applied. The first one gives sufficient conditions to guarantee that two solutions have the same forwards asymptotic behavior, while the second gives a criterion to prove the coincidence of two complete trajectories and also describes the pullback behavior of solutions.

THEOREM 5.1 (forwards behavior). *Assume that both solutions of* (5.1), $U_1 = \binom{u_1}{v_1}$ *and* $U_2 = \binom{u_2}{v_2}$, *are globally defined and bounded in* $L^\infty(\Omega, \mathbf{R}^2)$ *for* $t > s > t_0$. *Moreover, suppose that* u_1, v_1 *are positive in* Ω *and* $U_2(t)$ *is positive, nondegenerate for* $t > t_0$ *and for some* $p > \max(N/2, 1)$ *and* $0 < \theta \le 1$, *the coefficients in* (5.2) *satisfy* $p_{ij}, q_{ii} \in C^\theta(\mathbf{R}, L^p(\Omega))$ *for* $i, j = 1, 2$ *and* $p_{ii}(t,x) \le -\varphi_{ii}(x)$, *for* $t > t_0$, *with* $\varphi_{ii}(x)$ *as in Proposition* 4.2.

Then there exists a $\rho > 0$ *such that if*

$$\limsup_{t\to\infty} \|p_{12}(t)\|_{L^p(\Omega)} \limsup_{t\to\infty} \|p_{21}(t)\|_{L^p(\Omega)} \le \rho^2, \tag{5.3}$$

both solutions have the same forwards asymptotic behavior, i.e.,

$$U_1(t) - U_2(t) \to 0 \quad \textit{exponentially in } C^1_{\mathcal{B}_1}(\overline{\Omega}) \times C^1_{\mathcal{B}_2}(\overline{\Omega}) \quad \textit{as} \quad t \to \infty.$$

In particular, $U_1(t)$ is also nondegenerate at $+\infty$.

Proof. Clearly, (5.2) can be written as

$$W_t - D\Delta W = QW + PW,$$

where

$$D = \operatorname{diag}(d_1, d_2),\ Q = \operatorname{diag}(q_{11}, q_{22}),\ P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}, \quad W = \begin{pmatrix} y \\ z \end{pmatrix}. \tag{5.4}$$

Since $U_2 = \binom{u_2}{v_2}$ is a positive, bounded, and nondegenerate solution, for $t > s > t_0$, of the diagonal system

$$W_t - D\Delta W = QW,$$

it follows from Propositions 4.3 and 4.4 that for any $1 \le q < \infty$,

$$\|T_Q(t,s)\|_{\mathcal{L}(L^q(\Omega,\mathbf{R}^2))} \le M_1, \quad t > s > t_0, \tag{5.5}$$

with M_1 independent of t and s, $t > s > t_0$.

Then, we apply Theorem 4.6 to obtain that there exists $\rho > 0$ such that if (5.3) holds, then

$$\|T_{Q+P}(t,s)\|_{\mathcal{L}(L^q(\Omega,\mathbf{R}^2))} \le M_1'' \mathrm{e}^{\beta''(t-s)} \quad \text{for all} \quad t > s > t_0 \quad \text{and some} \quad \beta'' < 0.$$

Thus, from Proposition 4.4 (see also Proposition 4.1), we have, writing $W_s = (y_s, z_s)$ and for $t > s > t_0$,

$$\|W(t,s;W_s)\|_{L^\infty(\Omega,\mathbf{R}^2)} \le M_2 e^{\beta''(t-s)} \|W_s\|_{L^\infty(\Omega,\mathbf{R}^2)} \to 0, \quad t \to \infty. \tag{5.6}$$

The uniform forwards convergence of trajectories follows. Standard parabolic regularization implies the $C^1_{\mathcal{B}_1}(\overline{\Omega}) \times C^1_{\mathcal{B}_2}(\overline{\Omega})$ convergence. ∎

In the following result we use similar arguments to prove the coincidence of complete nondegenerate trajectories, and show that such a trajectory, when it exists, attracts (in the pullback sense) all bounded positive trajectories. In particular, the following results guarantee the uniqueness of complete nondegenerate solutions.

THEOREM 5.2 (coincidence of complete trajectories and pullback behavior). *Assume that $U_1 = \binom{u_1}{v_1}$ is a complete trajectory that is bounded in $L^\infty(\Omega, \mathbf{R}^2)$ at $-\infty$, and nondegenerate for $t \le t_0$. Suppose further that for some $p > \max(N/2, 1)$ and $0 < \theta \le 1$, the coefficients in (5.2) satisfy $p_{ij}, q_{ii} \in C^\theta(\mathbf{R}, L^p(\Omega))$ for $i, j = 1, 2$ and $p_{ii}(t,x) \le -\varphi_{ii}(x)$, for $t \le t_0$, with $\varphi_{ii}(x)$ as in Proposition 4.2.*

Then there exists a $\rho > 0$ such that if

$$\limsup_{t\to-\infty} \|p_{12}(t)\|_{L^p(\Omega)} \limsup_{t\to-\infty} \|p_{21}(t)\|_{L^p(\Omega)} \le \rho^2, \tag{5.7}$$

then

(i) *$U_1(t)$ is the unique complete trajectory that is bounded in $L^\infty(\Omega,\mathbf{R}^2)$ at $-\infty$, and*

(ii) *if $U_2(s)$ is a family of positive initial data which is bounded in $L^\infty(\Omega,\mathbf{R}^2)$ as $s\to-\infty$, then $U_1(\cdot)$ pullback attracts $S(t,s)U_2(s)$, i.e., for any $t\in\mathbf{R}$*

$$S(t,s)U_2(s)-U_1(t)\to 0 \quad \text{in } C^1_{\mathcal{B}_1}(\overline{\Omega})\times C^1_{\mathcal{B}_2}(\overline{\Omega}) \quad \text{as} \quad s\to-\infty.$$

Proof.

(i) Let $U_2(t)$ be a complete trajectory bounded in $L^\infty(\Omega,\mathbf{R}^2)$ at $-\infty$. We write (5.2) as

$$W_t - D\Delta W = QW + PW, \qquad W(s)=W_s=U_2(s)-U_1(s),$$

where Q, P, and W are defined as in (5.4). Since $U_1=(u_1,v_1)$ is a complete, positive, bounded, and nondegenerate solution of the diagonal system

$$W_t - D\Delta W = QW,$$

it follows from Proposition 4.3 that for any $1\le q<\infty$, and sufficiently negative t_0,

$$\|T_Q(t,s)\|_{\mathcal{L}(L^q(\Omega,\mathbf{R}^2))}\le M_1, \quad s<t\le t_0, \tag{5.8}$$

with M_1 independent of t and s.

Then, we apply Theorem 4.6 to obtain that there exists $\rho>0$ such that if (5.7) holds, then

$$\|T_{Q+P}(t,s)\|_{\mathcal{L}(L^q(\Omega,\mathbf{R}^2))}\le M_1''\mathrm{e}^{\beta''(t-s)} \quad \text{for all } s<t\le t_0 \quad \text{and some} \quad \beta''<0.$$

Thus,

$$\|U_1(t)-U_2(t)\|_{L^q(\Omega,\mathbf{R}^2)}=\|W(t,s;W_s)\|_{L^q(\Omega,\mathbf{R}^2)}\le M_1''\mathrm{e}^{\beta''(t-s)}\|W_s\|_{L^q(\Omega,\mathbf{R}^2)}. \tag{5.9}$$

The right-hand side tends to zero as $s\to-\infty$ since both complete trajectories are bounded, and the result follows.

(ii) Proceeding as above, we obtain

$$\|U_1(t)-S(t,s)U_2(s)\|_{L^q(\Omega,\mathbf{R}^2)}\le M_1''\mathrm{e}^{\beta''(t-s)}\|U_1(s)-U_2(s)\|_{L^q(\Omega,\mathbf{R}^2)}$$

for $s<t\le t_0$ and some $\beta''<0$.

Thus, from Proposition 4.4 (see also Proposition 4.1), we get for $s<t\le t_0$,

$$\|U_1(t)-S(t,s)U_2(s)\|_{L^\infty(\Omega,\mathbf{R}^2)}\le M_2\mathrm{e}^{\beta''(t-s)}\|U_1(s)-U_2(s)\|_{L^q(\Omega,\mathbf{R}^2)}\to 0$$

as $s\to-\infty$. Standard parabolic regularization implies the convergence in $C^1_{\mathcal{B}_1}(\overline{\Omega})\times C^1_{\mathcal{B}_2}(\overline{\Omega})$.

Now for every $\tau\ge t_0$, using the continuity of the nonlinear evolution process, we get, as $s\to-\infty$,

$$U_1(\tau,s)=S(\tau,t)U_1(t,s)\to S(\tau,t)U_2(t). \qquad \square$$

The theorems above may perhaps appear more general than they really are. To verify the assumptions involved one must restrict the nonlinearities of the system and carefully choose the classes of solutions being considered. For example, the conditions $p_{ii}(t,x)\le-\varphi_{ii}(x)$ and the smallness conditions on $p_{ij}(t,x)$, $i\ne j$ depend on the particular solutions considered.

Nevertheless, in the next section we will show that the assumptions required can be verified for our example of a general nonautonomous Lotka–Volterra system.

6. Attracting trajectories for nonautonomous Lotka–Volterra systems. As (5.1) is far too general to apply Theorems 5.1 and 5.2 in a straightforward manner, in this section we apply these results to the solutions of (3.1). Note that we handle the three cases, competition, symbiosis, and prey-predator, in a unified way.

Then, for the difference of two solutions the coefficients in (5.2) are given by

(6.1)
$$q_{11}(t,x) = \lambda(t,x) - a(t,x)u_2 - b(t,x)v_2, \quad q_{22}(t,x) = \mu(t,x) - c(t,x)u_2 - d(t,x)v_2,$$
$$p_{11}(t,x) = -a(t,x)u_1, \qquad p_{12}(t,x) = -b(t,x)u_1,$$
$$p_{21}(t,x) = -c(t,x)v_1, \qquad p_{22}(t,x) = -d(t,x)v_1.$$

Hence, to apply Theorems 5.1 or 5.2, since $a_L, d_L > 0$ and $u_1, v_1 \geq 0$, in order to find positive functions $\varphi_{ii}(x)$ such that $p_{ii}(t,x) \leq -\varphi_{ii}(x)$, $i = 1, 2$ we need positive functions $\psi_{ii}(x)$ such that $\psi_{11}(x) \leq u_1(t,x)$ and $\psi_{22}(x) \leq v_1(t,x)$, that is, we must consider nondegenerate solutions. The results in section 3.4 guarantee then that all solutions are nondegenerate.

On the other hand we must show that the product of the coupling terms

$$p_{12}(t,x)p_{21}(t,x)$$

is small at $\pm\infty$. Having obtained bounds on u_1, v_1 this will be achieved by a smallness condition on the coefficients $b(t,x)$ or $c(t,x)$.

But note that the nondegeneracy of solutions above depends on the functions $b(t,x)$ and $c(t,x)$ themselves. Therefore, we will use the results in section 3.4 which guarantee that solutions of (3.1) are nondegenerate for all sufficiently small "coupling" coefficients $b(t,x)$ or $c(t,x)$ and that the functions $\psi_{ii}(x)$, $i = 1, 2$ do not converge to zero as b or c vanish.

We first start with the forwards behavior in Theorem 5.1. Then we can prove the following theorem.

THEOREM 6.1. *There exists $\rho_0(M_\infty, N_\infty) > 0$, where M_∞ and N_∞ are given in Theorem* 3.5*, such that if*

$$\limsup_{t\to\infty} \|b\|_{L^\infty(\Omega)} \limsup_{t\to\infty} \|c\|_{L^\infty(\Omega)} < \rho_0(M_\infty, N_\infty),$$

and for some t_0 the coefficients of (3.1) *satisfy for $t \geq t_0$ the assumptions of Theorem* 3.14*, then for any bounded set of positive initial data bounded away from zero, all solutions of* (3.1) *that start at a sufficiently large $s > t_0$ have the same asymptotic behavior as $t \to \infty$.*

In particular, all complete positive trajectories in the pullback attractor have the same asymptotic behavior as $t \to \infty$.

Proof. Note that Theorem 3.14 implies that all forward solutions of (3.1) that start at $s \geq t_0$ are uniformly nondegenerate with respect to a bounded set of initial data $u_s > 0$, $v_s > 0$, bounded away from zero, and the coefficients. In particular, from Propositions 4.2 and 4.4 the constant M_1 in (5.5) can be taken independent of such $u_s > 0$, $v_s > 0$ and the coefficients.

Moreover, for such initial data and $t > s \geq t_0$, we have in (6.1)

$$p_{11}(t,x) = -a(t,x)u_1 \leq -a_L\psi_{11}(x) = -\varphi_{11}(x),$$
$$p_{22}(t,x) = -d(t,x)v_1 \leq -d_L\psi_{22}(x) = -\varphi_{22}(x)$$

with $\psi_{11}(x)$ and $\psi_{22}(x)$ independent $u_s > 0$, $v_s > 0$, and of the coefficients. In particular $\varphi_{11}(x)$ and $\varphi_{22}(x)$ satisfy the assumptions in Proposition 4.2.

Hence the threshold value $\rho > 0$ in Theorem 5.1 is also uniform for $u_s > 0$, $v_s > 0$, and of the coefficients as in Theorem 3.14.

Now we have in (6.1) $p_{12}(t,x) = -b(t,x)u_1$, $p_{21}(t,x) = -c(t,x)v_1$, and hence (5.3) is satisfied if

$$\limsup_{t\to\infty} \|b\|_{L^\infty(\Omega)} \limsup_{t\to\infty} \|c\|_{L^\infty(\Omega)} < \rho^2(p,\Omega,M_\infty,N_\infty) = \rho_0,$$

where M_∞ and N_∞ are given in Theorem 3.5.

Therefore, from Theorem 5.1, all solutions have the same forwards behavior. □

Our next result proves that if there is a complete trajectory that is nondegenerate at $-\infty$, then it must be unique and be pullback attracting, as in Theorem 5.2.

THEOREM 6.2. *Assume that there exists a complete, bounded solution of* (3.1) *that is nondegenerate at* $-\infty$, $U^*(t)$, $t \in \mathbf{R}$.

Then there exists $\rho_0(M_\infty, N_\infty) > 0$, *where* M_∞ *and* N_∞ *are given in Theorem* 3.5, *such that if*

$$\limsup_{t\to-\infty} \|b\|_{L^\infty(\Omega)} \limsup_{t\to-\infty} \|c\|_{L^\infty(\Omega)} < \rho_0(M_\infty, N_\infty),$$

and for some t_0 *the coefficients of* (3.1) *satisfy for* $t \le t_0$ *the assumptions of Theorem* 3.14, *then* $U^*(t)$ *is the unique bounded complete solution of* (3.1) *that is nondegenerate at* $-\infty$. *Moreover, for every* $t \in \mathbf{R}$, $U^*(t)$ *pullback attracts solutions* $U_1(t,s)$ *such that* $U_1(s)$ *are positive and bounded as* $s \to -\infty$.

If, in addition,

$$\limsup_{t\to\infty} \|b\|_{L^\infty(\Omega)} \limsup_{t\to\infty} \|c\|_{L^\infty(\Omega)} < \rho_0(M_\infty, N_\infty),$$

and for some t_1 *the coefficients of* (3.1) *satisfy for* $t \ge t_1$ *the assumptions of Theorem* 3.14, *then for any* $s \in \mathbf{R}$ *and for any positive solution* $U(t,s)$ *of* (3.1) *we have*

$$U(t,s) - U^*(t) \to 0 \qquad as \quad t \to \infty.$$

Proof. Assume there exists a complete, bounded nondegenerate solution at $-\infty$. Then Theorem 3.14 implies that all bounded nondegenerate solutions at $-\infty$ are uniformly nondegenerate with respect to the coefficients. In particular, from Propositions 4.2 and 4.4 the constant M_1 in (5.5) can be taken independent of the complete nondegenerate solution under consideration and of the coefficients. Moreover, we have in (6.1)

$$\begin{aligned} p_{11}(t,x) = -a(t,x)u_1 \le -a_L\psi_{11}(x) = -\varphi_{11}(x), \\ p_{22}(t,x) = -d(t,x)v_1 \le -d_L\psi_{22}(x) = -\varphi_{22}(x) \end{aligned}$$

with $\psi_{11}(x)$ and $\psi_{22}(x)$ independent of the complete nondegenerate solution and of the coefficients. In particular $\varphi_{11}(x)$, $\varphi_{22}(x)$ satisfy the assumptions in Proposition 4.2.

Hence the threshold value $\rho > 0$ in Theorem 5.2 is also independent of the complete nondegenerate solution and of the coefficients.

Now we have in (6.1) $p_{12}(t,x) = -b(t,x)u_1$, $p_{21}(t,x) = -c(t,x)v_1$, and hence (5.7) is satisfied if

$$\limsup_{t\to-\infty} \|b\|_{L^\infty(\Omega)} \limsup_{t\to-\infty} \|c\|_{L^\infty(\Omega)} < \rho^2(p,\Omega,M_\infty,N_\infty) = \rho_0.$$

Therefore, from Theorem 5.2, there exists at most a complete nondegenerate solution at $-\infty$.

To show that $U^*(t)$ is pullback attracting, observe that for sufficiently negative t_0 we can proceed as in the proof of Theorem 5.2 to conclude that $U^*(t)$ pullback attracts solutions $U_1(t,s)$ such that $U_1(s)$ is positive and bounded as $s \to -\infty$.

The rest follows from Theorem 6.1. □

7. Conclusions. We have obtained some results on permanence in nonautonomous Lotka–Volterra models without the assumption of any kind of periodicity. In particular we have found conditions under which there exists at least one complete trajectory, and for which all trajectories converge together as $t \to +\infty$. The key argument is a perturbation result for an associated linear system satisfied by the difference between two solutions, and using this we have been able to treat all the different classical cases – competition, symbiosis, and prey-predator – in a unified way. While this unified approach has its advantages, our method requires at least one of the coupling parameters in the system to be sufficiently small. Hence, we hope that a more detailed study of each particular situation could lead to some improvements in the conditions imposed on the nonautonomous terms while still using similar techniques.

It is a very interesting open problem to prove, for this Lotka–Volterra example, the existence of a complete trajectory that is nondegenerate at $-\infty$. Given this nondegeneracy one would get the uniqueness of such a trajectory, and its pullback attracting property. We believe that use of the concepts of sub- and super-trajectories (cf. Arnold and Chueshov [3] and Chueshov [10]), along with the sub- and supersolutions technique (cf. for example Pao [30]) should be able to provide this, and we intend to pursue this direction in a future paper.

However, it is certainly the case that the hypothesis that the time-dependent terms are bounded is important throughout the literature, as this assumption implies the existence of bounded global solutions, and in particular of bounded attracting trajectories. As the analysis in Langa et al. [23] shows, different kinds of forwards asymptotic behavior, such as the nonexistence of asymptotically stable trajectories, is possible if solutions are allowed to be unbounded.

REFERENCES

[1] H. Amann, *Nonhomogeneous linear and quasilinear elliptic and parabolic boundary value problems*, in Function Spaces, Differential Operators and Nonlinear Analysis (Friedrichroda, 1992), Teubner-Texte Math. 133, Teubner, Stuttgart, 1993, pp. 9–126.

[2] H. Amann, *Linear and Quasilinear Parabolic Problems*, Vol. I, Abstract Linear Theory. Monographs in Mathematics, 89, Birkhäuser Boston Inc., Boston, MA, 1995.

[3] L. Arnold and I. Chueshov, *Order-preserving random dynamical systems: Equilibria, attractors, applications*, Dynam. Stability Systems, 13 (1998), pp. 265–280.

[4] R. S. Cantrell and C. Cosner, *Should a park be an island?*, SIAM J. Appl. Math., 53 (1993), pp. 219–252.

[5] R. S. Cantrell and C. Cosner, *Practical persistence in ecological models via comparison methods*, Proc. R. Soc. Edin., 126A (1996), pp. 247–272.

[6] R. S. Cantrell and C. Cosner, *Spatial Ecology via Reaction-Diffusion Equations*, John Wiley & Sons. Ltd., New York, 2003.

[7] R. S. Cantrell, C. Cosner, and V. Hutson, *Permanence in ecological systems with spatial heterogeneity*, Proc. R. Soc. Edin., 123A (1993), pp. 533–559.

[8] R. S. Cantrell, C. Cosner, and V. Hutson, *Ecological models, permanence and spatial heterogeneity*, Rocky Mountain J. Math., 26 (1996), pp. 1–35.

[9] V. V. Chepyzhov and M. I. Vishik, *Attractors for Equations of Mathematical Physics*, AMS Colloquium Publications, Vol. 49, AMS, Providence, RI, 2002.

[10] I. Chueshov, *Monotone Random Systems Theory and Applications*, Lecture Notes in Math. 1779, Springer-Verlag, Berlin, 2002.

[11] H. Crauel, A. Debussche, and F. Flandoli, *Random attractors*, J. Dynam. Differential Equations, 9 (1997), pp. 397–341.

[12] M. Delgado, J. López-Gómez, and A. Suárez, *On the symbiotic Lotka-Volterra model with diffusion and transport effects*, J. Differential Equations, 160 (2000), pp. 175–262.
[13] J. E. Furter and J. López-Gómez, *On the existence and uniqueness of coexistence states for the Lotka-Volterra competition model with diffusion and spatially dependent coefficients*, Nonlinear Anal., 25 (1995), pp. 363–398.
[14] J. Hale, *Asymptotic Behavior of Dissipative Systems*, Math. Surveys and Monographs, AMS, Providence, RI, 1998.
[15] J. Hale and P. Waltman, *Persistence in infinite dimensional systems*, SIAM J. Math. Anal., 9 (1989), pp. 388–395.
[16] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes Math. 840, Springer-Verlag, Berlin, 1981.
[17] P. Hess, *Periodic-Parabolic Boundary Value Problems and Positivity*, Pitman Research Notes in Mathematics 247, Harlow Longman, 1991.
[18] P. Hess and A. C. Lazer, *On an abstract competition model and applications*, Nonlinear Anal., 16 (1991), pp. 917–940.
[19] G. Hetzer and W. Shen, *Uniform persistence, coexistence, and extinction in almost periodic/nonautonomous competition diffusion systems*, SIAM J. Math. Anal., 34 (2002), pp. 204–221.
[20] V. Hutson and K. Schmitt, *Permanence in dynamical systems*, Math. Biosci., 111 (1992), pp. 1–71.
[21] P. E. Kloeden and B. Schmalfuss, *Asymptotic behaviour of non-autonomous difference inclusions*, Systems Control Lett., 33 (1998), pp. 275–280.
[22] N. Lakos, *Existence of steady-state solutions for one-predator-two prey system*, SIAM J. Math. Anal., 21 (1990), pp. 647–659.
[23] J. A. Langa, J. C. Robinson, A. Rodríguez-Bernal, A. Suárez, and A. Vidal-López, *Existence and nonexistence of unbounded forward attractor for a class of non-autonomous reaction diffusion equations*, Discrete Contin. Dyn. Syst., 18 (2007), pp. 483–497.
[24] J. A. Langa, J. C. Robinson, and A. Suárez, *Stability, instability, and bifurcation phenomena in non-autonomous differential equations*, Nonlinearity, 15 (2002), pp. 887–903.
[25] J. A. Langa, J. C. Robinson, and A. Suárez, *Pullback permanence in the non-autonomous Lotka-Volterra competition model*, J. Differential Equations, 190 (2003), pp. 214–238.
[26] A. W. Leung, *Monotone schemes for semilinear elliptic systems related to ecology*, Math. Methods Appl. Sci., 4 (1982), pp. 272–285.
[27] J. López-Gómez, *On the structure of the permanence region for competing species models with general diffusivities and transport effects*, Discrete Contin. Dyn. Syst., 2 (1996), pp. 525–542.
[28] J. López-Gómez and R. Pardo, *Existence and uniqueness of coexistence states for the predator-prey model with diffusion: The scalar case*, Differential Int. Equations, 6 (1993), pp. 1025–1031.
[29] J. López-Gómez and J. C. Sabina de Lis, *Coexistence states and global attractivity for some convective diffusive competing species models*, Trans. Amer. Math. Soc., 347 (1995), pp. 3797–3833.
[30] C. V. Pao, *Nonlinear parabolic and elliptic equations*, Plenum, New York, 1992.
[31] J. C. Robinson, A. Rodríguez-Bernal, and A. Vidal-López, *Pullback attractors and extremal complete trajectories for non-autonomous reaction-diffusion problems*, J. Differential Equations, 238 (2007), pp. 289–337.
[32] A. Rodríguez-Bernal, *On linear and nonlinear non–autonomous parabolic equations*, Departamento de Matemática Aplicada, UCM, Preprint Series MA-UCM-2006-15.
[33] A. Rodríguez-Bernal, *Perturbation of the exponential type of linear nonautonomous parabolic equations and applications to nonlinear equations*, Departamento de Matemática Aplicada, UCM, Preprint Series MA-UCM-2008-08.
[34] A. Rodríguez-Bernal and A. Vidal-López, *Existence, uniqueness and attractivity properties of positive complete trajectories for non-autonomous reaction-diffusion problems*, Discrete Contin. Dyn. Syst., 18 (2007), pp. 537–567.
[35] B. Schmalfuss, *Attractors for the non-autonomous dynamical systems*, Proceedings of Equadiff 99 Berlin, B. Fiedler, K. Gröger, and J. Sprekels, eds., Singapore World Scientific, Singapore, 2000, pp. 684–689.
[36] Y. Yamada, *Stability of steady states for predator-prey diffusion equations with homogeneous Dirichlet conditions*, SIAM J. Math. Anal., 21 (1990), pp. 327–345.
[37] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Springer-Verlag, New York, 1988.

SIAM J. MATH. ANAL.
Vol. 40, No. 6, pp. 2217–2240

AN ENTIRE SOLUTION TO THE LOTKA–VOLTERRA COMPETITION-DIFFUSION EQUATIONS*

YOSHIHISA MORITA† AND KOICHI TACHIBANA‡

Abstract. We deal with a system of Lotka–Volterra competition-diffusion equations on $\mathbb{R}$, which is a competing two species model with diffusion. It is known that the equations allow traveling waves with monotone profile. In this article we prove the existence of an entire solution which behaves as two monotone waves propagating from both sides of the x-axis, where an entire solution is meant by a classical solution defined for all space and time variables. The global dynamics for this entire solution exhibits the extinction of the inferior species by the superior one invading from both sides. The proof is carried out by applying the comparison principle for the competition-diffusion equations, that is, using an appropriate pair of a subsolution and a supersolution.

Key words. Lotka–Volterra, competition-diffusion, entire solution, traveling wave

AMS subject classifications. 35K57, 35B05, 35B40, 92B05

DOI. 10.1137/080723715

1. Introduction.

In the field of population biology Lotka–Volterra competition equations are well accepted as a physiological model describing competing interaction of multiple species. Here we restrict our attention to a two species model. Taking random movement of the species into account, we get to the Lotka–Volterra competition-diffusion equations that are obtained by coupling diffusion terms to the Lotka–Volterra equations. In this article we are dealing with the following Lotka–Volterra competition-diffusion equations on $\mathbb{R}$:

$$\begin{cases} u_t = u_{xx} + (1 - u - k_1 v)u, \\ v_t = dv_{xx} + a(1 - v - k_2 u)v \end{cases} \quad (x \in \mathbb{R}), \tag{1.1}$$

where k_1, k_2, a, and d are positive constants. The variables $u(x,t)$ and $v(x,t)$ stand for the population density of two competing species. Thus we consider nonnegative $u(x,t), v(x,t)$. We note that the above system is normalized so that it has the equilibrium solutions $(u,v) = (1,0), (0,1)$, set as

$$\boldsymbol{e}_u := (1,0), \qquad \boldsymbol{e}_v := (0,1).$$

In the diffusion-free case we can classify the asymptotic behavior of the solution as $t \to \infty$, depending on k_1, k_2:

(i) $(u(t), v(t)) \to \boldsymbol{e}_u$ (u always wins) if $0 < k_1 < 1 < k_2$.
(ii) Depending on the initial condition, $(u,v) \to \boldsymbol{e}_u$ or $(u(t), v(t)) \to \boldsymbol{e}_v$ in general if $1 < k_1, k_2$.
(iii) $(u(t), v(t)) \to \boldsymbol{e}_v$ (v always wins) if $0 < k_2 < 1 < k_1$.

*Received by the editors May 8, 2008; accepted for publication (in revised form) September 23, 2008; published electronically January 28, 2009. This research was supported in part by the Grant-in-Aid for Scientific Research (B) 19340026 and the Grant-in-Aid for Exploratory Research 19654030, Japan Society for the Promotion of Science.
http://www.siam.org/journals/sima/40-6/72371.html
†Department of Applied Mathematics and Informatics, Ryukoku University, Seta Otsu 520-2194, Japan (morita@rins.ryukoku.ac.jp).
‡Daiwa Technique Laboratory Ltd., 915 Mitsushima Kadoma, Osaka 571-0015, Japan (ham_one14@msn.com).

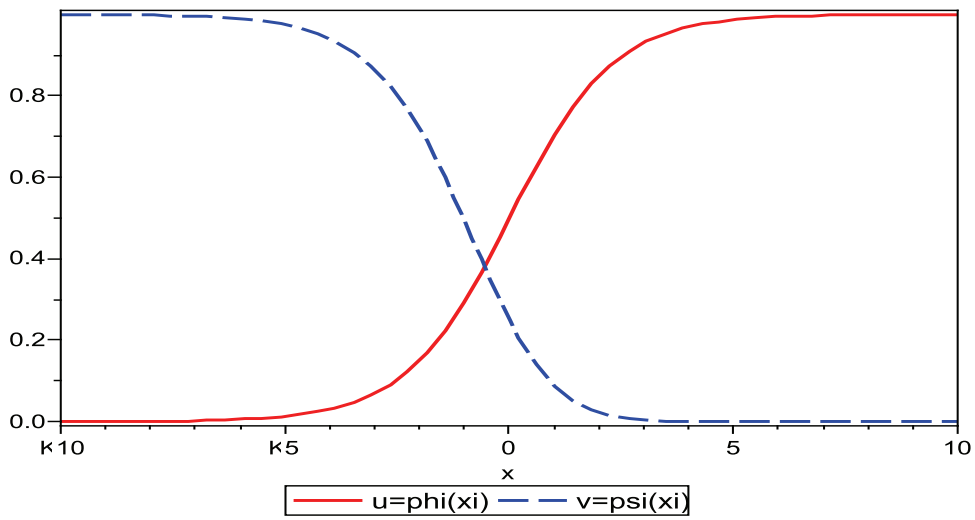


FIG. 1. *The profile of a monotone traveling wave solution of* (1.1). *Monotone increasing and decreasing curves correspond to* $u = \phi(\xi)$ *and* $v = \psi(\xi)$, *respectively. Parameter values are given by* $a = 3/2$, $d = 1/3$, $k_1 = 3/2$, *and* $k_2 = 2$.

(iv) $(u(t), v(t))$ converges to a positive equilibrium (u and v coexist) if $0 < k_1, k_2 < 1$.

Here we do not treat case (iv). Then we may assume $k_2 > 1$ because the third case (iii) is obtained by exchanging the roles of u and v.

The above equations (1.1) are used for describing the invasion of the superior species. As a matter of fact, by virtue of the diffusion, we can observe behavior such as the superior species propagating from one side of the x-axis to the other side, pushing back the inferior species. This dynamics of the propagation is mathematically characterized by a traveling wave solution $(u, v) = (\phi(x + ct), \psi(x + ct))$ satisfying

$$\begin{cases} \phi'' - c\phi' + (1 - \phi - k_1\psi)\phi = 0, \\ d\psi'' - c\psi' + a(1 - \psi - k_2\phi)\psi = 0, \end{cases} \tag{1.2}$$

with

$$\lim_{\xi \to -\infty} (\phi(\xi), \psi(\xi)) = \boldsymbol{e}_v, \quad \lim_{\xi \to \infty} (\phi(\xi), \psi(\xi)) = \boldsymbol{e}_u, \tag{1.3}$$

$$\phi(\xi), \quad \psi(\xi) > 0, \qquad \phi'(\xi) > 0, \quad \psi'(\xi) < 0, \tag{1.4}$$

where $' = d/d\xi$, $'' = d^2/d\xi^2$. A solution to (1.2) with (1.3) and (1.4) gives a profile of the traveling wave. Here we assume the monotonicity of (1.4) which arises in most cases. A traveling wave with the monotonicity is called a monotone wave or a wave front.

We note that for a traveling wave solution $(\phi(\xi), \psi(\xi)) = (\phi(x + ct), \psi(x + ct))$ to (1.2)–(1.4), the reflected $(\phi(-x + ct), \psi(-x + ct))$ also gives a traveling wave with opposite speed. Then $(\tilde{\phi}(\xi), \tilde{\psi}(\xi)) = (\phi(-x + ct), \psi(-x + ct))$ satisfies

$$\lim_{\xi \to -\infty} (\tilde{\phi}(\xi), \tilde{\psi}(\xi)) = \boldsymbol{e}_u, \quad \lim_{\xi \to \infty} (\tilde{\phi}(\xi), \tilde{\psi}(\xi)) = \boldsymbol{e}_v.$$

Thus we obtain traveling wave solutions with opposite speeds simultaneously if we obtain a solution satisfying (1.3)–(1.4).

As for the study of traveling wave solutions, there are various works, including [3], [4], [15], and [7]. In particular, see [8], [9], [10], [11], [12], and the references therein.

Although the study for a traveling wave is a central issue of the Lotka–Volterra competition-diffusion equations, it is not enough for mathematical understanding of the dynamical structure of solutions to (1.1). In fact, as seen in the recent development of the study for entire solutions to a scalar reaction-diffusion equation, some combination of two traveling waves yields characteristic dynamics for the equation (see [6], [16], [2], [5], [1], [13]), where the entire solution is meant by a classical solution which is defined for every x and t. Those results revealed that the equation allows types of entire solutions other than equilibrium solutions and traveling wave solutions. For example, in [6], [16], [2], [5], it is shown that there exists an entire solution which behaves as two wave fronts propagating from both sides of the x-axis and then annihilating, while [13] shows entire solutions which behave as two wave fronts merging.

In this article, motivated by the work of [2] and [5], we consider the entire solution which behaves like two monotone traveling waves propagating from both sides of the x-axis. As t goes forward, this solution converges to the equilibrium $\boldsymbol{e}_u$ when u is superior. Namely, we are interested in the entire solution $(u(x,t), v(x,t))$ which behaves like $(\phi(x+ct),\ \psi(x+ct))$ for $x \in (0,\infty)$ and $(\phi(-x+ct),\ \psi(-x+ct))$ for $x \in (-\infty, 0)$ as $t \to -\infty$ and converges to a uniform state as $t \to \infty$. This solution corresponds to the phenomenon that the superior species invades from both sides and causes the extinction of the inferior species.

We prepare some notation for stating the main result. We assume that u is always superior so that the traveling wave solution of (1.2) with (1.3) and (1.4) has a positive speed $c > 0$. In addition, we allow the equations to have solutions with different speeds (corresponding to case (i)). As a matter of fact, we have traveling waves of the Fisher type for the case (i); that is, there is a minimum speed $c_{min} > 0$ so that the equations allow a family of monotone traveling wave solutions for $c \geq c_{min} > 0$ (see [10]). Thus we set a pair of traveling wave solutions

$$(\phi_j(\xi), \psi_j(\xi)) \quad (j = 1, 2)$$

satisfying

$$\begin{cases} \phi_j'' - c_j\phi_j' + (1 - \phi_j - k_1\psi_j)\phi_j = 0, \\ d\psi_j'' - c_j\psi_j' + a(1 - \psi_j - k_2\phi_j)\psi_j = 0 \end{cases} \qquad (j = 1, 2), \tag{1.5}$$

with

$$\lim_{\xi \to -\infty} (\phi_j(\xi), \psi_j(\xi)) = \boldsymbol{e}_v, \quad \lim_{\xi \to \infty} (\phi_j(\xi), \psi_j(\xi)) = \boldsymbol{e}_u, \tag{1.6}$$

$$\phi_j(\xi), \quad \psi_j(\xi) > 0, \qquad \phi_j'(\xi) > 0, \quad \psi_j'(\xi) < 0, \tag{1.7}$$

where c_j $(j = 1, 2)$ are positive. Note that we consider the case when they coincide, namely, $(\phi_1(\xi), \psi_1(\xi))$ identically equals $(\phi_2(\xi), \psi_2(\xi))$ up to a phase shift.

We assume that there is a positive constant η_0 such that

$$\eta_0 \leq \frac{\phi_j(\xi)}{1 - \psi_j(\xi)} \qquad (\xi \leq 0) \tag{1.8}$$

holds. Then we obtain the following theorem, which is the main result in this article.

THEOREM 1.1. *Assume $k_1 \neq 1$ and $k_2 > 1$ in* (1.1). *Let (ϕ_j, ψ_j) be a solution to* (1.5) *with* (1.6), (1.7), *and $c_j > 0$ $(j = 1, 2)$. If the condition* (1.8) *holds, then there exists a solution $(u(x,t), v(x,t))$ $((x,t) \in \mathbb{R} \times \mathbb{R})$ of* (1.1) *satisfying*

$$\lim_{t\to-\infty} \sup_{x\geq(c_2-c_1)t/2} \{|u(x,t) - \phi_1(x + c_1t)| + |v(x,t) - \psi_1(x + c_1t)|\} = 0,$$
$$\lim_{t\to-\infty} \sup_{x\leq(c_2-c_1)t/2} \{|u(x,t) - \phi_2(-x + c_2t)| + |v(x,t) - \psi_2(-x + c_2t)|\} = 0,$$

and

$$\lim_{t\to\infty} \sup_{x\in\mathbb{R}} \{|u(x,t) - 1| + |v(x,t)|\} = 0.$$

Remark 1.1. If $k_1 > 1$, the monotone traveling wave is unique up to translation (for instance, see [8]). Hence, when $k_1 > 1$, $(\phi_1, \psi_1) = (\phi_2, \psi_2)$ in the theorem.

Considering the fact that the equations (1.1) are invariant under the translation in x and the time shift, we easily see the following result from the above theorem.

COROLLARY 1.2. *Let $(u(x,t), v(x,t))$ be an entire solution to* (1.1) *given in Theorem* 1.1. *Given θ_1, θ_2, there exist ξ and τ such that*

$$\lim_{t\to-\infty} \sup_{x\geq(c_2-c_1)t/2} \{|u(x+\xi,t+\tau) - \phi_1(x + c_1t + \theta_1)| + |v(x+\xi,t+\tau) - \psi_1(x + c_1t + \theta_1)|\} = 0,$$
$$\lim_{t\to-\infty} \sup_{x\leq(c_2-c_1)t/2} \{|u(x+\xi,t+\tau) - \phi_2(-x + c_2t + \theta_2)| + |v(x+\xi,t+\tau) - \psi_2(-x + c_2t + \theta_2)|\} = 0,$$

and

$$\lim_{t\to\infty} \sup_{x\in\mathbb{R}} \{|u(x+\xi,t+\tau) - 1| + |v(x+\xi,t+\tau)|\} = 0.$$

The condition (1.8) is crucial in our argument, though it might be a technical condition. Here we give a sufficient condition for (1.8), which is easier to check. Let $\lambda_3^{(j)}$ be a positive root of

$$d\lambda^2 - c_j\lambda - a = 0. \tag{1.9}$$

For $k_1 > 1$, we have a unique positive root $\lambda_4^{(j)}$ of

$$\lambda^2 - c_j\lambda - (k_1 - 1) = 0, \tag{1.10}$$

while for $0 < k_1 < 1$, there is another positive root $\lambda_5^{(j)}$ $(\leq \lambda_4^{(j)})$ of (1.10) under the condition

$$c_j \geq c_{min} \geq 2\sqrt{1 - k_1}, \tag{1.11}$$

where $\lambda_4^{(j)} = \lambda_5^{(j)}$ if $c_j = 2\sqrt{1 - k_1}$. Those roots $\lambda_k^{(j)}$ $(k = 3, 4, 5)$ are related to the convergence rate of the traveling waves as $(\phi_j(\xi), \psi_j(\xi)) \to (0, 1)$ $(\xi \to -\infty)$. We notice that

$$d\frac{\psi_j''}{1-\psi_j} - c_j\frac{\psi_j'}{1-\psi_j} + a\psi_j\left(1 - k_2\frac{\phi_j}{1-\psi_j}\right) = 0 \tag{1.12}$$

by the second equation of (1.5). We can prove the limit of $\psi_j'(\xi)/(1-\psi_j(\xi))$ as $\xi \to -\infty$ exists. In fact, we have the next lemma.

LEMMA 1.3. *Assume the same assumption of Theorem* 1.1. *Then* (1.8) *holds if*

$$\lim_{\xi \to -\infty} \frac{-\psi_j'(\xi)}{1-\psi_j(\xi)} \neq \lambda_3^{(j)}. \tag{1.13}$$

Moreover, the condition (1.13) *is realized for* $\lambda_4^{(j)} < \lambda_3^{(j)}$ $(j = 1, 2)$.

Remark 1.2. The condition $\lambda_4^{(j)} < \lambda_3^{(j)}$ $(j = 1, 2)$ is clear and easily checked. However, there are cases which break this condition but allow (1.13). For example, (1.13) is verified by the condition $\lambda_5^{(j)} < \lambda_3^{(j)} \leq \lambda_4^{(j)}$ together with a generic condition for the asymptotic behavior of the traveling wave as $t \to -\infty$. We discuss this in section 2.

We briefly sketch an outline of the proof of Theorem 1.1. We will use the following functions to approximate the entire solution as $t \approx -\infty$:

$$\begin{cases} \underline{u} = \max\{\phi_1(x + c_1 t), \phi_2(-x + c_2 t)\}, \\ \overline{v} = \min\{\psi_1(x + c_1 t), \psi_2(-x + c_2 t)\}, \end{cases} \tag{1.14}$$

and

$$\begin{cases} \overline{u} = \phi_1(x + p_1(t)) + \phi_2(-x + p_2(t)) - \phi_1(x + p_1(t))\phi_2(-x + p_2(t)), \\ \underline{v} = \psi_1(x + p_1(t))\psi_2(-x + p_2(t)), \end{cases} \tag{1.15}$$

where

$$p_j(t) \approx c_j t \qquad (t \approx -\infty).$$

Note that $(\underline{u}, \overline{v})$ and $(\overline{u}, \underline{v})$ certainly have the profile we are supposing as $t \to -\infty$. Since the comparison principle does work in the present competition-diffusion system, we prove that (1.14) and (1.15) are a subsolution and a supersolution with suitable $p_j(t)$, respectively, if $t < 0$. This yields a solution sandwiched by the subsolution and the supersolution, defined for every negative t. Then such a solution can be extended for any positive time. In what follows, we obtain the desired entire solution.

Although our idea for the proof of the main result is based on the previous work of [5], [1], and [13], we encounter a difficulty when we prove that the pairing of (1.15) is a supersolution. Fortunately, noticing the identity

$$\begin{aligned} &(u_1 + u_2 - u_1 u_2)v_1 v_2 - (1 - u_2)u_1 v_1 - (1 - u_1)u_2 v_2 \\ &\quad = -u_1(1 - u_2)v_1(1 - v_2) - (1 - u_1)u_2(1 - v_1)v_2 + u_1 u_2 v_1 v_2, \end{aligned} \tag{1.16}$$

we can overcome the difficulty (see section 4 for the details).

In the above theorem we obtain an entire solution characterized by the asymptotic behavior as $t \to -\infty$. Once we obtain such a solution, it is easy to prove the asymptotic behavior as $t \to \infty$ stated in the theorem, since the subsolution is defined globally in time. However, we have only the one side estimated by the subsolution; hence precise transient behavior for positive t is not described yet. Here, using an exact expression for the traveling wave solution, we show the entire behavior of the solution for specific parameter values. In the paper [14], traveling wave solutions are

exactly given in terms of the hyperbolic tangent function under a constraint of the parameters. We use one of them, which is given by

$$\phi(\xi) = \frac{1}{2}\left\{1 + \tanh\left(\frac{\sqrt{k_1}\xi}{2\sqrt{2}}\right)\right\}, \quad \psi(\xi) = \frac{1}{4}\left\{1 - \tanh\left(\frac{\sqrt{k_1}\xi}{2\sqrt{2}}\right)\right\}^2, \tag{1.17}$$

where

$$d = \frac{a}{3k_1}, \quad a = \frac{6 - 3k_1}{3k_2 - 5}, \quad c = \frac{2 - k_1}{\sqrt{2k_1}}, \tag{1.18}$$

with the condition

$$0 < k_1 < 2, \quad 5/3 < k_2. \tag{1.19}$$

It is a lengthy but simple computation to verify that (1.17) with (1.18) and (1.19) is a traveling wave of (1.1) satisfying (1.2)–(1.4). Moreover, this solution satisfies (1.13) as well as (1.8). Indeed, we can compute the corresponding positive roots of (1.9) and (1.10) as

$$\lambda_3 = \frac{1}{2}\sqrt{\frac{k_1}{2}}\{3k_2 - 5 + \sqrt{(3k_2 - 5)^2 + 24}\} > \sqrt{3k_1},$$

$$\lambda_4 = \begin{cases} \sqrt{k_1/2} & (2/3 \le k_1), \\ \sqrt{2}(1 - k_1)/\sqrt{k_1} & (0 < k_1 < 2/3) \end{cases}$$

and

$$\lambda_5 = \begin{cases} \sqrt{2}(1 - k_1)/\sqrt{k_1} & (2/3 < k_1 < 1), \\ \sqrt{k_1/2} & (0 < k_1 \le 2/3), \end{cases}$$

respectively.

We rewrite (1.17) as

$$\phi(\xi) = 1 - \frac{1}{1 + \exp(\xi\sqrt{k_1/2})}, \quad \psi(\xi) = \frac{1}{\{1 + \exp(\xi\sqrt{k_1/2})\}^2}.$$

Modify this expression as

$$u_*(x,t) := 1 - \frac{1}{1 + \Phi(x,t)}, \qquad v^*(x,t) = \frac{1}{(1 + \Phi(x,t))^2} \tag{1.20}$$

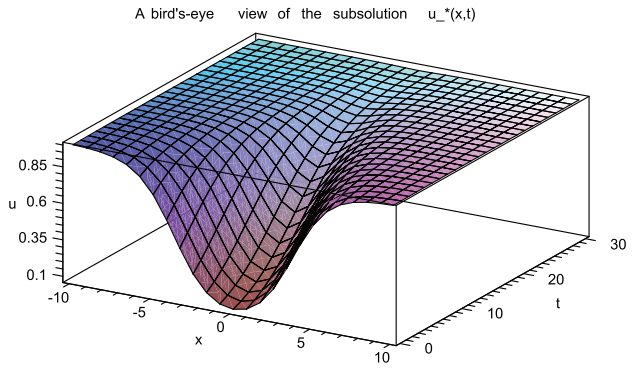
with

$$\Phi = \cosh\left(\frac{x\sqrt{k_1}}{\sqrt{2}}\right)\exp(\omega_0 t), \quad \omega_0 := 1 - \frac{k_1}{2}. \tag{1.21}$$

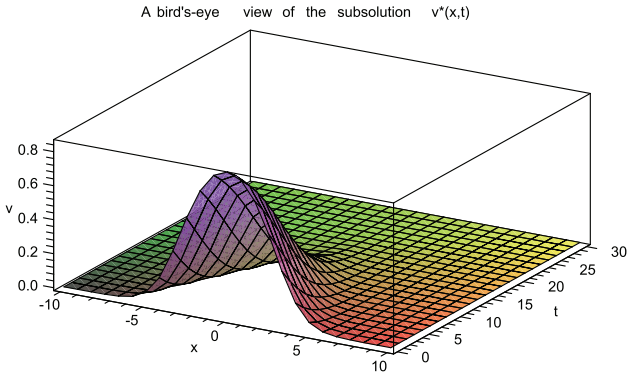
As for a profile of the functions of (1.20), see Figure 2.

We also define

$$u^*(x,t) := 1 - \frac{1}{1 + \Psi(x,t)}, \qquad v_*(x,t) = \frac{1}{(1 + \Psi(x,t))^2} \tag{1.22}$$



(a) A bird's eye view of $u_*(x, t-20)$



(b) A bird's eye view of $v^*(x, t-20)$

FIG. 2. *Bird's eye views of the functions* $u_*(x, t-20)$ *and* $v^*(x, t-20)$ $(0 \le t \le 30, -10 \le x \le 10)$*, respectively, which are defined by* (1.20) *and* (1.21)*. Parameter values are given by* (1.18) *with* $k_1 = 3/2$ *and* $k_2 = 2$*.*

with

$$\Psi = \cosh\left(\frac{x\sqrt{k_1}}{\sqrt{2}}\right)\exp(p(t)), \tag{1.23}$$

where $p(t)$ is given by a solution to

$$\dot{p}(t) = \omega_0 + \frac{L_0\exp(p(t))}{1+\exp(p(t))}, \quad p(0) = p_0, \tag{1.24}$$
$$L_0 := \max\{k_1, a/2\}.$$

Note that the solution $p(t)$ is a monotone increasing function satisfying

$$\lim_{t\to-\infty} p(t) = -\infty, \qquad \lim_{t\to\infty} p(t) = \infty.$$

The next result shows that the above functions give an approximation globally in time together with the existence of an entire solution.

PROPOSITION 1.4. *In addition to* (1.18) *and* (1.19)*, assume that the solution* $p(t)$ *of* (1.24) *satisfies* $p(t_0) = \omega_0 t_0$ *for an arbitrarily given* t_0*. Then a solution*

$(u(x,t),v(x,t))$ to (1.1) with the initial condition

$$(u(x,t_0),v(x,t_0))=(u_*(x,t_0),v^*(x,t_0))=(u^*(x,t_0),v_*(x,t_0))$$

satisfies

$$\begin{cases} u_*(x,t)\le u(x,t)\le u^*(x,t), \\ v_*(x,t)\le v(x,t)\le v^*(x,t) \end{cases} \quad ((t,x)\in(t_0,\infty)\times\mathbb{R}). \tag{1.25}$$

Moreover, if p_0 is a solution of

$$p_0=\frac{L_0}{L_0+\omega_0}\log\left(1+\frac{L_0+\omega_0}{\omega_0}\exp(p_0)\right), \tag{1.26}$$

then the asymptotics

$$\lim_{t\to-\infty}\sup_{x\in\mathbb{R}}[|u^*(x,t)-u_*(x,t)|+|v^*(x,t)-v_*(x,t)|]=0$$

hold; hence there is an entire solution $(u(x,t),v(x,t))$ satisfying (1.25) with $t_0=-\infty$.

We organize the paper as follows: In the next section we classify the asymptotic behavior of the traveling wave solutions as $\xi\to\pm\infty$. This result is given by investigating the linearized equations at the equilibrium points $\boldsymbol{e}_u$ and $\boldsymbol{e}_v$. Then Lemma 1.3 is proved. Although the result is already established, we give a proof in the appendix for the reader's convenience. In section 3 we give crucial estimates for the terms consisting of the traveling wave solutions. By virtue of the estimates we prove Theorem 1.1 together with Proposition 1.4 in section 4. The paper concludes with the appendix.

2. Asymptotic behavior of traveling front waves. All the results in Lemmas 2.1, 2.2, and 2.3 stated below were already obtained in [8] and [12]. For completeness of the present paper we give a proof in the appendix. We let $\lambda_1^{(j)}$ and $\lambda_2^{(j)}$ be negative roots of

$$\lambda^2-c_j\lambda-1=0 \tag{2.1}$$

and

$$d\lambda^2-c_j\lambda-a(k_2-1)=0, \tag{2.2}$$

respectively. The first lemma is on the asymptotic behavior of the traveling wave solutions as $\xi\to\infty$.

LEMMA 2.1. *Let $(\phi_j(\xi),\psi_j(\xi))$ be traveling wave solutions to (1.5)–(1.7) with $c_j>0$ $(j=1,2)$, and let $\lambda_1^{(j)}$ and $\lambda_2^{(j)}$ be negative roots of (2.1) and (2.2), respectively. Then the solution exhibits the asymptotic behavior as $\xi\to\infty$ as follows:*

(i) *If $\lambda_2^{(j)}<\lambda_1^{(j)}$, then*

$$\begin{aligned} 1-\phi_j(\xi)&=\alpha_j\exp(\lambda_1^{(j)}\xi)-\beta_j s_1^{(j)}\exp(\lambda_2^{(j)}\xi)+h.o.t., \\ \psi_j(\xi)&=\beta_j\exp(\lambda_2^{(j)}\xi)+h.o.t. \end{aligned} \tag{2.3}$$

hold for some numbers $\alpha_j\ge0$ and $\beta_j>0$, where

$$s_1^{(j)}:=\frac{k_1}{(\lambda_2^{(j)})^2-c_j\lambda_2^{(j)}-1}=\begin{cases}>0 & (\lambda_2^{(j)}<\lambda_1^{(j)}),\\ <0 & (\lambda_2^{(j)}>\lambda_1^{(j)}).\end{cases} \tag{2.4}$$

(ii) *If* $\lambda_1^{(j)} < \lambda_2^{(j)}$, *then*

$$\begin{aligned} 1-\phi_j(\xi) &= -\beta_j s_1^{(j)} \exp(\lambda_2^{(j)}\xi) + h.o.t., \\ \psi_j(\xi) &= \beta_j \exp(\lambda_2^{(j)}\xi) + h.o.t. \end{aligned} \tag{2.5}$$

for some $\beta_j > 0$, *where* $s_1^{(j)}$ *is as in* (2.4).

(iii) *If* $\lambda_1^{(j)} = \lambda_2^{(j)}$, *then*

$$\begin{aligned} 1-\phi_j(\xi) &= \beta_j \xi \exp(\lambda_1^{(j)}\xi) + h.o.t., \\ \psi_j(\xi) &= -\beta_j \tau_1^{(j)} \exp(\lambda_1^{(j)}\xi) + h.o.t. \end{aligned} \tag{2.6}$$

for some $\beta_j > 0$, *where*

$$\tau_1^{(j)} = \frac{2\lambda_1^{(j)} - c_j}{k_1} < 0.$$

The second lemma is related to the asymptotic behavior of the traveling wave solution as $\xi \to -\infty$ for $k_1 > 1$, namely, the bistable case (ii) stated in the introduction.

LEMMA 2.2. *Let* $(\phi_j(\xi), \psi_j(\xi))$ *be traveling wave solutions to* (1.5)–(1.7) *with* $c_j > 0$ $(j = 1, 2)$ *for* $k_1 > 1$, *and let* $\lambda_3^{(j)}$ *and* $\lambda_4^{(j)}$ *be positive roots of* (1.9) *and* (1.10), *respectively. Then the solution exhibits the asymptotic behavior as* $\xi \to -\infty$ *as follows:*

(i) *If* $\lambda_3^{(j)} < \lambda_4^{(j)}$, *then*

$$\begin{aligned} \phi_j(\xi) &= \beta_j \exp(\lambda_4^{(j)}\xi) + h.o.t., \\ 1-\psi_j(\xi) &= \alpha_j \exp(\lambda_3^{(j)}\xi) - \beta_j s_2^{(j)} \exp(\lambda_4^{(j)}\xi) + h.o.t. \end{aligned} \tag{2.7}$$

hold for some numbers $\alpha_j \geq 0$ *and* $\beta_j > 0$, *where*

$$s_2^{(j)} := \frac{ak_2}{d(\lambda_4^{(j)})^2 - c_j\lambda_4^{(j)} - a} = \begin{cases} > 0 & (\lambda_3^{(j)} < \lambda_4^{(j)}), \\ < 0 & (\lambda_3^{(j)} > \lambda_4^{(j)}). \end{cases} \tag{2.8}$$

(ii) *If* $\lambda_4^{(j)} < \lambda_3^{(j)}$, *then*

$$\begin{aligned} \phi_j(\xi) &= \beta_j \exp(\lambda_4^{(j)}\xi) + h.o.t., \\ 1-\psi_j(\xi) &= -\beta_j s_2^{(j)} \exp(\lambda_4^{(j)}\xi) + h.o.t. \end{aligned} \tag{2.9}$$

for some $\beta_j > 0$, *where* $s_2^{(j)}$ *is defined as in* (2.8).

(iii) *If* $\lambda_3^{(j)} = \lambda_4^{(j)}$, *then*

$$\begin{aligned} \phi_j(\xi) &= \beta_j \tau_2^{(j)} \exp(\lambda_3^{(j)}\xi) + h.o.t., \\ 1-\psi_j(\xi) &= -\beta_j \xi \exp(\lambda_3^{(j)}\xi) + h.o.t. \end{aligned} \tag{2.10}$$

as $\xi \to -\infty$ *for some* β_j, *where*

$$\tau_2^{(j)} := \frac{2d\lambda_3^{(j)} - c_j}{k_1} > 0. \tag{2.11}$$

The next lemma is related to the asymptotic behavior as $\xi \to -\infty$ when $0 < k_1 < 1$, namely, the monostable case (i) stated in the introduction.

LEMMA 2.3. *Let $(\phi_j(\xi), \psi_j(\xi))$ be traveling wave solutions to (1.5)–(1.7) with $c_j > 0$ $(j = 1, 2)$ for $k_1 \in (0, 1)$, and let $\lambda_5^{(j)}$ be a positive root of (1.10) with $\lambda_5^{(j)} \le \lambda_4^{(j)}$, where $c_j \ge 2\sqrt{1-k_1}$. Then the solution behaves as $\xi \to -\infty$ in the following way:*

(i) *$\lambda_5^{(j)} < \lambda_4^{(j)}$, $\lambda_4^{(j)}, \lambda_5^{(j)} \ne \lambda_3^{(j)}$: there are numbers α_j, β_j, and nonnegative γ_j such that*

(2.12)
$$\begin{aligned} \phi_j(\xi) &= \beta_j \exp(\lambda_4^{(j)}\xi) + \gamma_j \exp(\lambda_5^{(j)}\xi) + h.o.t., \\ 1 - \psi_j(\xi) &= \alpha_j \exp(\lambda_3^{(j)}\xi) - \beta_j s_2^{(j)} \exp(\lambda_4^{(j)}\xi) - \gamma_j s_3^{(j)} \exp(\lambda_5^{(j)}\xi) + h.o.t. \end{aligned}$$

with

$$\gamma_j \ge 0, \qquad \beta_j > 0 \quad (\gamma_j = 0), \qquad \alpha_j > 0 \quad (\lambda_3^{(j)} < \lambda_5^{(j)})$$

hold, where

$$s_3^{(j)} = \frac{ak_2}{d(\lambda_5^{(j)})^2 - c_j\lambda_5^{(j)} - a} = \begin{cases} > 0 & (\lambda_5^{(j)} > \lambda_3^{(j)}), \\ < 0 & (\lambda_5^{(j)} < \lambda_3^{(j)}). \end{cases} \tag{2.13}$$

(ii) *$\lambda_5^{(j)} = \lambda_4^{(j)} < \lambda_3^{(j)}$: there are β_j and γ_j such that*

$$\begin{aligned} \phi_j(\xi) &= \beta_j \xi \exp(\lambda_4^{(j)}\xi) + \gamma_j \exp(\lambda_4^{(j)}\xi) + h.o.t., \\ 1 - \psi_j(\xi) &= -\beta_j s_2^{(j)} \xi \exp(\lambda_4^{(j)}\xi) - \gamma_j s_2^{(j)} \exp(\lambda_4^{(j)}\xi) + h.o.t. \end{aligned} \tag{2.14}$$

with $\beta_j \ge 0$ and $\gamma_j > 0$ $(\beta_j = 0)$, where $s_2^{(j)}$ is defined as in (2.8).

(iii) *$\lambda_5^{(j)} < \lambda_3^{(j)} = \lambda_4^{(j)}$: there are β_j and γ_j such that*

$$\begin{aligned} \phi_j(\xi) &= \beta_j \tau_2^{(j)} \exp(\lambda_3^{(j)}\xi) + \gamma_j \exp(\lambda_5^{(j)}\xi) + h.o.t., \\ 1 - \psi_j(\xi) &= -\beta_j \xi \exp(\lambda_3^{(j)}\xi) - \gamma_j s_3^{(j)} \exp(\lambda_5^{(j)}\xi) + h.o.t. \end{aligned} \tag{2.15}$$

with $\gamma_j \ge 0$ and $\beta_j > 0$ $(\gamma_j = 0)$, where $\tau_2^{(j)}$ is defined as in (2.11).

(iv) *$\lambda_5^{(j)} = \lambda_3^{(j)} < \lambda_4^{(j)}$: there are α_j, β_j, and γ_j such that*

(2.16)
$$\begin{aligned} \phi_j(\xi) &= \gamma_j \tau_2^{(j)} \exp(\lambda_3^{(j)}\xi) + \beta_j \exp(\lambda_4^{(j)}\xi) + h.o.t., \\ 1 - \psi_j(\xi) &= \alpha_j \exp(\lambda_3^{(j)}\xi) - \gamma_j \xi \exp(\lambda_3^{(j)}\xi) - \beta_j s_2^{(j)} \exp(\lambda_4^{(j)}\xi) + h.o.t. \end{aligned}$$

with $\gamma_j \ge 0$ and

$$\alpha_j \ge 0, \quad \beta_j > 0 \qquad (\gamma_j = 0).$$

(v) *$\lambda_5^{(j)} = \lambda_4^{(j)} = \lambda_3^{(j)}$: there are β_j and γ_j such that*

$$\begin{aligned} \phi_j(\xi) &= \beta_j \tau_2^{(j)} \exp(\lambda_3^{(j)}\xi) - \gamma_j \tau_2 \xi \exp(\lambda_3^{(j)}\xi) + h.o.t., \\ 1 - \psi_j(\xi) &= -\beta_j \xi \exp(\lambda_3^{(j)}\xi) + \gamma_j (\xi^2/2) \exp(\lambda_3^{(j)}\xi) + h.o.t. \end{aligned} \tag{2.17}$$

with $\gamma_j \ge 0$ and $\beta_j > 0$ $(\gamma_j = 0)$.

The next lemma immediately follows from the above lemmas.

LEMMA 2.4. *There are positive constants r_1, r_2, η_1, η_2, ω_1, ω_2, and M such that*

$$r_1 \le \frac{\phi_j'(\xi)}{1-\phi_j(\xi)} \le r_2 \ (\xi \ge 0), \tag{2.18}$$

$$r_1 \le \frac{1-\phi_j(\xi)}{\psi_j(\xi)} \qquad (\xi \ge 0), \tag{2.19}$$

$$r_1 \le \frac{|\psi_j'(\xi)|}{\psi_j(\xi)} \qquad (\xi \ge 0) \tag{2.20}$$

and

$$0 < 1-\psi_j(\xi) \le M\exp(\omega_1\xi) \ (\xi \le 0), \tag{2.21}$$

$$\eta_1 \le \frac{|\psi_j'(\xi)|}{1-\psi_j(\xi)} \le \eta_2 \qquad (\xi \le 0), \tag{2.22}$$

$$0 \le \phi_j(\xi) \le M\exp(\omega_2\xi) \quad (\xi \le 0), \tag{2.23}$$

$$\eta_1 \le \frac{\phi_j'(\xi)}{\phi_j(\xi)} \le \eta_2 \qquad (\xi \le 0). \tag{2.24}$$

The following lemma is clear from Lemmas 2.2 and 2.3.

LEMMA 2.5. *Inequality* (1.8) *holds provided that one of the following conditions is satisfied:*

(1) $\lambda_4^{(j)} < \lambda_3^{(j)}$,

(2) $\lambda_5^{(j)} < \lambda_3^{(j)} \le \lambda_4^{(j)}, \quad \gamma_j > 0$,

(3) $\lambda_5^{(j)} \le \lambda_3^{(j)} < \lambda_4^{(j)}, \quad \alpha_j = \gamma_j = 0$,

where α_j, β_j, and γ_j are the numbers corresponding to each condition of (i)–(v) *for $\lambda_k^{(j)}$ $(k = 3, 4, 5)$ in Lemma* 2.3.

Proof of Lemma 1.3. By Lemmas 2.2 and 2.3, the limit

$$\lim_{\xi\to-\infty} \frac{-\psi_j'(\xi)}{1-\psi_j(\xi)} = \lim_{\xi\to-\infty} \frac{\psi_j''(\xi)}{\psi_j'(\xi)}$$

exists. Then the condition (1.13) implies that this limit is $\lambda_4^{(j)}$ or $\lambda_5^{(j)}$. Hence $\beta_j > 0$ or $\gamma_j > 0$. We assert that one of the conditions (1), (2), and (3) of Lemma 2.5 holds. This proves the first part of the assertion. The latter part is trivial by Lemma 2.5. □

3. Key estimates. In this section we prove two lemmas which will play a crucial role in the argument for the proof of Theorem 1.1.

LEMMA 3.1. *Let (ϕ_j, ψ_j) be traveling wave solutions of* (1.5)–(1.7) *satisfying* (1.8). *Define*

$$A_1(z,p) := (1-\phi_2(-z+p))\phi_1'(z+p) + (1-\phi_1(z+p))\phi_2'(-z+p), \tag{3.1}$$

and put $\omega = \min\{\omega_1, \omega_2\}$. Then there exist $L_1, L_2 > 0$ such that, given $p < 0$,

$$\frac{\phi_1'(z+p)\phi_2'(-z+p)}{A_1(z,p)} \le L_1\exp(\omega p), \tag{3.2}$$

$$\frac{\phi_1(z+p)(1-\phi_2(-z+p))\psi_1(z+p)(1-\psi_2(-z+p))}{A_1(z,p)} \le L_2\exp(\omega p), \tag{3.3}$$

and

$$\frac{\phi_2(-z+p)(1-\phi_1(z+p))\psi_2(-z+p)(1-\psi_1(z+p))}{A_1(z,p)} \le L_2\exp(\omega p) \tag{3.4}$$

hold.

Proof. We first prove (3.2). With the aid of (2.23) and (2.24), for $z \le 0$,

$$\begin{aligned}\frac{\phi_1'(z+p)\phi_2'(-z+p)}{A_1(z,p)} &\le \frac{\phi_1'(z+p)}{1-\phi_1(z+p)}\\ &\le \frac{\eta_2 M\exp(\omega_2(z+p))}{1-\phi_1(0)}\\ &\le \frac{\eta_2 M\exp(\omega_2 p)}{1-\phi_1(0)},\end{aligned}$$

while, for $z \ge 0$,

$$\begin{aligned}\frac{\phi_1'(z+p)\phi_2'(-z+p)}{A_1(z,p)} &\le \frac{\phi_2'(-z+p)}{1-\phi_2(-z+p)}\\ &\le \frac{\eta_2 M\exp(\omega_2(-z+p))}{1-\phi_2(0)}\\ &\le \frac{\eta_2 M\exp(\omega_2 p)}{1-\phi_2(0)}.\end{aligned}$$

Thus we can take $L_1 = \eta_2 M\max\{1/(1-\phi_1(0)),\, 1/(1-\phi_2(0))\}$ to prove (3.2).

Next we show (3.3) and (3.4). Put

$$B_1(z,p) := \phi_1(z+p)\psi_1(z+p)(1-\phi_2(-z+p))(1-\psi_2(-z+p)), \tag{3.5}$$

$$\tilde{B}_1(z,p) := \phi_2(-z+p)\psi_2(-z+p)(1-\phi_1(z+p))(1-\psi_1(z+p)). \tag{3.6}$$

First we prove (3.3). Let $z \le p$. By (2.18) we have

$$r_1 \le \frac{\phi_2'(-z+p)}{1-\phi_2(-z+p)} \quad (z \le p);$$

hence

$$\begin{aligned}\frac{B_1(z,p)}{A_1(z,p)} &\le \frac{\phi_1(z+p)}{r_1(1-\phi_1(z+p))}\\ &\le \frac{M\exp(\omega_2(z+p))}{r_1(1-\phi_1(0))}\\ &\le \frac{M\exp(\omega_2 p)}{r_1(1-\phi_1(0))}.\end{aligned} \tag{3.7}$$

As for the case $p \le z \le 0$, noticing that

$$\eta_0\eta_1 \le \frac{\phi_2'(-z+p)}{1-\psi_2(-z+p)} \quad (p \le z),$$

from (2.24) and (1.8) we have

$$\begin{aligned}\frac{B_1(z,p)}{A_1(z,p)} &\le \frac{\phi_1(z+p)}{\eta_0\eta_1(1-\phi_1(z+p))}\\ &\le \frac{M\exp(\omega_2(z+p))}{\eta_0\eta_1(1-\phi_1(0))}\\ &\le \frac{M\exp(\omega_2 p)}{\eta_0\eta_1(1-\phi_1(0))}. \end{aligned} \tag{3.8}$$

We let $0 \le z \le -p$. With the aid of

$$\eta_1 \le \frac{\phi_1'(z+p)}{\phi_1(z+p)} \quad (z \le -p),$$

we obtain

$$\begin{aligned}\frac{B_1(z,p)}{A_1(z,p)} &\le \frac{(1-\phi_2(-z+p))(1-\psi_2(-z+p))}{\eta_1(1-\phi_2(-z+p))}\\ &\le \frac{1-\psi_2(-z+p)}{\eta_1}\\ &\le \frac{M\exp(\omega_1(-z+p))}{\eta_1}\\ &\le \frac{M\exp(\omega_1 p)}{\eta_1}. \end{aligned} \tag{3.9}$$

Since

$$r_1^2 \le \frac{\phi_1'(z+p)}{\psi_1(z+p)} \quad (-p \le z),$$

by (2.18) and (2.19), we see that for $-p \le z$,

$$\begin{aligned}\frac{B_1(z,p)}{A_1(z,p)} &\le \frac{(1-\phi_2(-z+p))(1-\psi_2(-z+p))}{r_1^2(1-\phi_2(-z+p))}\\ &\le \frac{M\exp(\omega_1(-z+p))}{r_1^2}\\ &\le \frac{M\exp(\omega_1 p)}{r_1^2}. \end{aligned} \tag{3.10}$$

Combining (3.7), (3.8), (3.9), and (3.10) leads to (3.3).

We go to the case (3.4). In the above argument for (3.5), we replace $\phi_1(z+p)$, $\phi_2(-z+p)$, $\psi_1(z+p)$, and $\psi_2(-z+p)$ by $\phi_2(-z+p)$, $\phi_1(z+p)$, $\psi_2(-z+p)$, and $\psi_1(z+p)$, respectively, and change z to $-z$. Then for (3.6) we similarly obtain the desired estimate. We will leave the details to the reader. □

LEMMA 3.2. *Let*

$$A_2(z,p) := \psi_2(-z+p)\psi_1'(z+p) + \psi_1(z+p)\psi_2'(-z+p) \quad (<0). \tag{3.11}$$

Then there exist $L_3, L_4 > 0$ such that, given $p < 0$,

$$\frac{\psi_1'(z+p)\psi_2'(-z+p)}{|A_2(z,p)|} \le L_3\exp(\omega p) \tag{3.12}$$

and

$$\frac{\psi_1(z+p)\psi_2(-z+p)(1-\psi_1(z+p))(1-\psi_2(-z+p))}{|A_2(z,p)|} \le L_4 \exp(\omega p) \tag{3.13}$$

hold.

Proof. For $z \ge 0$, by (2.21) and (2.22) we have

$$\begin{aligned}\frac{\psi_1'(z+p)\psi_2'(-z+p)}{|A_2(z,p)|} &\le \frac{|\psi_2'(-z+p)|}{\psi_2(-z+p)} \\ &\le \frac{\eta_2 M \exp(\omega_1(-z+p))}{\psi_2(0)} \le \frac{\eta_2 M \exp(\omega_1 p)}{\psi_2(0)}.\end{aligned}$$

Similarly, we obtain that, for $z \le 0$,

$$\frac{\psi_1'(z+p)\psi_2'(-z+p)}{|A_2(z,p)|} \le \frac{\eta_2 M \exp(\omega_1 p)}{\psi_1(0)}.$$

This proves (3.12).

We show (3.13). Put

$$B_2(z,p) := \psi_1(z+p)\psi_2(-z+p)(1-\psi_1(z+p))(1-\psi_2(-z+p)). \tag{3.14}$$

First let $0 \le z \le -p$. By (2.22),

$$\begin{aligned}\frac{B_2(z,p)}{|A_2(z,p)|} &\le \frac{\psi_1(z+p)\psi_2(-z+p)(1-\psi_2(-z+p))}{\eta_1 \psi_2(-z+p)} \\ &\le \frac{1-\psi_2(-z+p)}{\eta_1} \le \frac{M \exp(\omega_1 p)}{\eta_1}.\end{aligned}$$

Next we consider the case $-p \le z$. By (2.20)

$$\begin{aligned}\frac{B_2(z,p)}{|A_2(z,p)|} &\le \frac{\psi_2(-z+p)(1-\psi_1(z+p))(1-\psi_2(-z+p))}{r_1 \psi_2(-z+p)} \\ &\le \frac{1-\psi_2(-z+p)}{r_1} \le \frac{M \exp(\omega_1 p)}{r_1}.\end{aligned}$$

As for $p \le z \le 0$ and $z \le p$, we use the above argument similarly to obtain

$$\frac{B_2(z,p)}{|A_2(z,p)|} \le \frac{M \exp(\omega_1 p)}{\eta_1} \quad (p \le z \le 0),$$

$$\frac{B_2(z,p)}{|A_2(z,p)|} \le \frac{M \exp(\omega_1 p)}{r_1} \quad (z \le p).$$

This concludes the proof of (3.13). ∎

4. Existence of the entire solution. Let $(\phi_j(\xi), \psi_j(\xi))$ $(j = 1, 2)$ be the traveling wave solutions as in Theorem 1.1. To prove the theorem, we consider the combination of the solutions $(\phi_1(x + c_1 t), \psi_1(x + c_1 t))$ and $(\phi_2(-x + c_2 t), \psi_2(-x + c_2 t))$.

By the change of variables

$$(U(z,t), V(z,t)) = (u(x,t), v(x,t)), \quad z = x + \frac{c_1 - c_2}{2} t,$$

(1.1) is transformed into

$$\begin{cases} U_t = U_{zz} - \dfrac{c_1 - c_2}{2} U_z + U(1 - U - k_1 V), \\ V_t = dV_{zz} - \dfrac{c_1 - c_2}{2} V_z + aV(1 - V - k_2 U) \end{cases} \quad (x \in \mathbb{R}). \tag{4.1}$$

Then $(U, V) = (\phi_1(z + c_m t),\ \psi_1(z + c_m t))$ and $(\phi_2(-z + c_m t),\ \psi_2(-z + c_m t))$ give traveling wave solutions of (4.1) with speed $c_m = (c_1 + c_2)/2$ and the opposite speed $-c_m$, respectively. The latter has the profile such that the U component is monotone decreasing while the V component is monotone increasing.

We define supersolutions and subsolutions to (1.1) and (4.1). Put

$$\begin{cases} \mathcal{F}_1(u, v) := u_t - u_{xx} + \beta u_x - f(u, v), \\ \mathcal{F}_2(u, v) := v_t - dv_{xx} + \beta v_x - ag(u, v), \end{cases}$$

and consider the equations

$$\mathcal{F}_1(u, v) = 0, \quad \mathcal{F}_2(u, v) = 0, \tag{4.2}$$

where

$$f(u, v) := (1 - u - k_1 v)u, \qquad g(u, v) := (1 - v - k_2 u)v,$$
$$\beta := \frac{c_1 - c_2}{2}.$$

We call $(\underline{u}(x, t), \overline{v}(x, t))$, $(x, t) \in \mathbb{R} \times [T_1, T_2]$ a subsolution if

$$\mathcal{F}_1(\underline{u}, \overline{v}) \le 0, \qquad \mathcal{F}_2(\underline{u}, \overline{v}) \ge 0 \qquad ((x, t) \in \mathbb{R} \times [T_1, T_2])$$

hold. We say that $(\overline{u}(x, t), \underline{v}(x, t))$, $(x, t) \in \mathbb{R} \times [T_1, T_2]$ is a supersolution to (4.2) if

$$\mathcal{F}_1(\overline{u}, \underline{v}) \ge 0, \qquad \mathcal{F}_2(\overline{u}, \underline{v}) \le 0 \qquad ((x, t) \in \mathbb{R} \times [T_1, T_2]).$$

We note that any solutions $(u(x, t), v(x, t))$ of (4.2) are not only subsolutions but also supersolutions. Let $(u(x, t), v(x, t))$ be solutions to (4.2) in $t \in [T_1, T_2]$. The maximal principle for the competition-diffusion system yields that if

$$\begin{cases} \underline{u}(x, T_1) \le u(x, T_1) \le \overline{u}(x, T_1), \\ \underline{v}(x, T_1) \le v(x, T_1) \le \overline{v}(x, T_1) \end{cases}$$

hold, then these orders are preserved for all $t \in [T_1, T_2]$.

If $(u_1(x, t), v_1(x, t))$ and $(u_1(x, t), v_1(x, t))$ are subsolutions in $t \in (T_1, T_2)$, then the pairing of

$$\underline{u}(x, t) := \max_{x \in \mathbb{R}} \{u_1(x, t), u_2(x, t)\}, \quad \overline{v}(x, t) := \min_{x \in \mathbb{R}} \{v_1(x, t), v_2(x, t)\}$$

is a subsolution in $t \in (T_1, T_2)$. Similarly, if $(u_1(x, t), v_1(x, t))$ and $(u_1(x, t), v_1(x, t))$ are supersolutions, then the pairing of

$$\overline{u}(x, t) := \min_{x \in \mathbb{R}} \{u_1(x, t), u_2(x, t)\}, \quad \underline{v}(x, t) := \max_{x \in \mathbb{R}} \{v_1(x, t), v_2(x, t)\}$$

is a supersolution.

We define a subsolution of (4.1) by the pairing

$$\begin{cases} \underline{U}(z,t) := \max_{z\in\mathbb{R}}\{\phi_1(z+c_mt+q), \phi_2(-z+c_mt+q)\}, \\ \overline{V}(z,t) := \min_{z\in\mathbb{R}}\{\psi_1(z+c_mt+q), \psi_2(-z+c_mt+q)\}, \end{cases} \tag{4.3}$$

where q is an arbitrarily given number.

Next we construct a supersolution. We first notice the following identities:

$$\begin{aligned} &f(u_1+u_2-u_1u_2, v_1v_2) - (1-u_2)f(u_1,v_1) - (1-u_1)f(u_2,v_2) \\ &\quad = -(1-u_1)(1-u_2)u_1u_2 \\ &\qquad - k_1\{(u_1+u_2-u_1u_2)v_1v_2 - (1-u_2)u_1v_1 - (1-u_1)u_2v_2\}, \\ &g(u_1+u_2-u_1u_2, v_1v_2) - v_2g(u_1,v_1) - v_1g(u_2,v_2) \\ &\quad = -v_1v_2(1-v_1)(1-v_2) + k_2u_1u_2v_1v_2. \end{aligned}$$

With the aid of (1.16), we obtain the inequalities

$$\begin{aligned} &f(u_1+u_2-u_1u_2, v_1v_2) - (1-u_2)f(u_1,v_1) - (1-u_1)f(u_2,v_2) \\ &\qquad \le k_1\{u_1(1-u_2)v_1(1-v_2) + (1-u_1)u_2(1-v_1)v_2\}, \end{aligned} \tag{4.4}$$

$$\begin{aligned} &g(u_1+u_2-u_1u_2, v_1v_2) - v_2g(u_1,v_1) - v_1g(u_2,v_2) \\ &\qquad \ge -v_1v_2(1-v_1)(1-v_2) \end{aligned} \tag{4.5}$$

for $0 \le u_1, u_2 \le 1$, and $0 \le v_1, v_2$.

We introduce the following ordinary differential equation:

$$\begin{cases} \dot{p} = c_m + Le^{\omega p} \qquad (t \le 0), \\ p(0) = p_0 < 0. \end{cases} \tag{4.6}$$

A simple computation yields a solution to (4.6) as

$$p(t) = c_mt - \frac{1}{\omega}\log\left\{e^{-\omega p_0} + \frac{L(1-e^{\omega c_m t})}{c_m}\right\} < 0 \quad (t<0) \tag{4.7}$$

having the asymptotics

$$\lim_{t\to-\infty}(p(t) - c_mt) = -\frac{1}{\omega}\log\left(e^{-\omega p_0} + \frac{L}{c_m}\right). \tag{4.8}$$

LEMMA 4.1. *Let* $(\phi_j(x+c_jt), \psi_j(x+c_jt))$ $(j=1,2)$ *be traveling wave solutions to* (1.1) *satisfying* (1.5), (1.6), (1.7), *and* (1.8). *Let* $p(t)$ *be a solution to* (4.6) *with* $L \ge \max\{2L_1 + 2k_1L_2,\ 2dL_3 + aL_4\}$. *Then the pairing of*

$$\begin{cases} \overline{U}(z,t) := \phi_1(z+p(t)) + \phi_2(-z+p(t)) - \phi_1(z+p(t))\phi_2(-z+p(t)), \\ \underline{V}(z,t) := \psi_1(z+p(t))\psi_2(-z+p(t)) \end{cases} \tag{4.9}$$

is a supersolution of (4.1).

Proof. First note that $\phi_1 = \phi_1(\xi)$, $\psi_1 = \psi_1(\xi)$ $(\xi = z + c_mt)$ satisfy

$$\begin{cases} \phi_1'' - c_1\phi_1' + f(\phi_1,\psi_1) = 0, \\ d\psi_1'' - c_1\psi_1' + ag(\phi_1,\psi_1) = 0, \end{cases}$$

while $\phi_2 = \phi_2(\xi)$, $\psi_2 = \psi_2(\xi)$ ($\xi = -z + c_m t$) satisfy

$$\begin{cases} \phi_2'' - c_2\phi_2' + f(\phi_2, \psi_2) = 0, \\ d\psi_2'' - c_2\psi_2' + ag(\phi_2, \psi_2) = 0, \end{cases}$$

where $' = d/d\xi$, $'' = d^2/d\xi^2$. Substituting $(\overline{U}, \underline{V})$ into

$$\begin{cases} \mathcal{F}_1(U,V) = U_t - U_{zz} + \dfrac{c_1 - c_2}{2} U_z - f(U,V), \\ \mathcal{F}_2(U,V) = V_t - dV_{zz} + \dfrac{c_1 - c_2}{2} V_z - ag(U,V) \end{cases}$$

yields

$$\begin{aligned} \mathcal{F}_1(\overline{U}, \underline{V}) &= \dot{p}\{(1-\phi_2)\phi_1' + (1-\phi_1)\phi_2'\} - (\phi_1'' + \phi_2'' - \phi_1''\phi_2 + 2\phi_1'\phi_2' - \phi_1\phi_2'') \\ &\quad + \frac{c_1 - c_2}{2}\{(1-\phi_2)\phi_1' - (1-\phi_1)\phi_2'\} - f(\phi_1 + \phi_2 - \phi_1\phi_2, \psi_1\psi_2) \\ &= \dot{p}\{(1-\phi_2)\phi_1' + (1-\phi_1)\phi_2'\} - (1-\phi_2)\left(\phi_1'' - \frac{c_1 - c_2}{2}\phi_1'\right) \\ &\quad - (1-\phi_1)\left(\phi_2'' + \frac{c_1 - c_2}{2}\phi_2'\right) - 2\phi_1'\phi_2' - f(\phi_1 + \phi_2 - \phi_1\phi_2, \psi_1\psi_2) \\ &= (\dot{p} - c_m)\{(1-\phi_2)\phi_1' + (1-\phi_1)\phi_2'\} - 2\phi_1'\phi_2' \\ &\quad - \{f(\phi_1 + \phi_2 - \phi_1\phi_2, \psi_1\psi_2) - (1-\phi_2)f(\phi_1, \psi_1) - (1-\phi_1)f(\phi_2, \psi_2)\}. \end{aligned}$$

Using (4.4), we obtain

$$\begin{aligned} \mathcal{F}_1(\overline{U}, \underline{V}) &\geq A_1(z,p)(\dot{p} - c_m) - 2\phi_1'\phi_2' - k_1(B_1(z,p) + \tilde{B}_1(z,p)) \\ &= A_1(z,p)\left\{L\exp(\omega p) - \frac{2\phi_1'\phi_2'}{A_1(z,p)} - k_1\frac{B_1(z,p) + \tilde{B}_1(z,p)}{A_1(z,p)}\right\}, \end{aligned} \tag{4.10}$$

where $A_1(z,p)$, $B_1(z,p)$, and $\tilde{B}_1(z,p)$ are defined in (3.1), (3.5), and (3.6), respectively. By Lemma 3.1 we obtain

$$\mathcal{F}_1(\overline{U}, \underline{V}) \geq 0 \qquad ((z,t) \in \mathbb{R} \times (-\infty, 0])$$

for $L \geq 2L_1 + 2k_1L_2$.

Similarly,

$$\begin{aligned} \mathcal{F}_2(\overline{U}, \underline{V}) &= (\psi_1'\psi_2 + \psi_1\psi_2')\dot{p} - d(\psi_1''\psi_2 + \psi_1\psi_2'') + 2d\psi_1'\psi_2' \\ &\quad + \frac{c_1 - c_2}{2}(\psi_1'\psi_2 - \psi_1\psi_2') - ag(\phi_1 + \phi_2 - \phi_1\phi_2, \psi_1\psi_2) \\ &= (\psi_1'\psi_2 + \psi_1\psi_2')\dot{p} + \psi_2\{-c_1\psi_1' + ag(\phi_1, \psi_1)\} \\ &\quad + \psi_1\{-c_2\psi_2' + ag(\phi_2, \psi_2)\} + 2d\psi_1'\psi_2' \\ &\quad + \frac{c_1 - c_2}{2}(\psi_1'\psi_2 - \psi_1\psi_2') - ag(\phi_1 + \phi_2 - \phi_1\phi_2, \psi_1\psi_2) \\ &= (\psi_1'\psi_2 + \psi_1\psi_2')(\dot{p} - c_m) + 2d\psi_1'\psi_2' \\ &\quad - a\{g(\phi_1 + \phi_2 - \phi_1\phi_2, \psi_1\psi_2) - \psi_2 g(\phi_1, \psi_1) - \psi_1 g(\phi_2, \psi_2)\}. \end{aligned}$$

Applying (4.5) to the right-hand side of this equality yields

$$\mathcal{F}_2(\overline{U},\underline{V}) \le A_2(z,p)\left\{L\exp(\omega p) + \frac{2d\psi_1'\psi_2'}{A_2(z,p)} + \frac{aB_2(z,p)}{A_2(z,p)}\right\}, \tag{4.11}$$

where $A_2(z,p)$ and $B_2(z,p)$ are defined as in (3.11) and (3.14), respectively (recall $A_2(z,p) < 0$). By Lemma 3.2 we obtain

$$\mathcal{F}_2(\overline{U},\underline{V}) \le 0 \qquad ((z,t) \in \mathbb{R}\times(-\infty,0])$$

for $L \ge 2dL_3 + aL_4$. This implies that $(\overline{U},\underline{V})$ is a supersolution. □

Proof of Theorem 1.1. We use the subsolution (4.3) with

$$q = -\frac{1}{\omega}\log\left(e^{-\omega p_0} + \frac{L}{c_m}\right)$$

and the supersolution (4.9). For this q we have

$$\lim_{t\to-\infty} |p(t) - (c_m+q)t| = 0,$$

$$\begin{cases} \underline{U}(z,t) < \overline{U}(z,t), \\ \underline{V}(z,t) < \overline{V}(z,t) \end{cases} \qquad (t<0,\ z\in\mathbb{R}),$$

and

$$\lim_{t\to-\infty}[\sup_{z\in\mathbb{R}}(\overline{U}(z,t) - \underline{U}(z,t)) + \sup_{z\in\mathbb{R}}(\overline{V}(z,t) - \underline{V}(z,t))] = 0.$$

Applying the same argument in [2] (see also [1]) with the aid of the comparison theorem yields that there is a solution $(U(z,t),V(z,t))$ satisfying

$$\begin{cases} \underline{U}(z,t) \le U(z,t) \le \overline{U}(z,t), \\ \underline{V}(z,t) \le V(z,t) \le \overline{V}(z,t) \end{cases} \qquad ((z,t)\in\mathbb{R}\times(-\infty,0]). \tag{4.12}$$

Since (1.1) (or (4.1)) has the invariance with time shift, we obtain the asymptotic behavior as $t\to-\infty$ in the statement in Theorem 1.1.

On the other hand, to prove the asymptotic behavior as $t\to\infty$ of the theorem, we recall that $\underline{U}(z,t), \overline{V}(z,t)$ in (4.3) are defined every t. By the asymptotic behavior

$$\lim_{t\to\infty}\sup_{z\in\mathbb{R}}(1-\underline{U}(z,t)) = 0, \quad \lim_{t\to\infty}\sup_{z\in\mathbb{R}}\overline{V}(z,t) = 0$$

and the fact that the solution can be continued for positive t, we obtain the desired behavior as $t\to\infty$. □

Proof of Corollary 1.2. Since the equation is invariant under phase shift and time shift, the desired result follows from Theorem 1.1. For the details of the argument, see [5]. □

Proof of Proposition 1.4. A straightforward computation yields

$$\mathcal{F}_1(u_*,v^*) = -\frac{k_1\exp(2\omega_0 t)}{(1+\Phi)^3} < 0,$$

$$\mathcal{F}_2(u_*,v^*) = \frac{a\exp(2\omega_0 t)}{(1+\Phi)^4} > 0.$$

On the other hand,

$$\begin{aligned}\mathcal{F}_1(u^*, v_*) &= \frac{\Psi}{(1+\Psi)^2}\left\{\dot{p} - \omega_0 - \frac{k_1 \exp(2p)}{\Psi(1+\Psi)}\right\} \\ &\geq \frac{\Psi}{(1+\Psi)^2}\left\{\frac{L_0 \exp(p)}{1+\exp(p)} - \frac{k_1 \exp(p)}{1+\exp(p)}\right\} \geq 0,\end{aligned}$$

$$\begin{aligned}\mathcal{F}_2(u^*, v_*) &= -\frac{2\Psi}{(1+\Psi)^3}\left\{\dot{p} - \omega_0 - \frac{a \exp(2p)}{2\Psi(1+\Psi)}\right\} \\ &\leq -\frac{2\Psi}{(1+\Psi)^3}\left\{\frac{L_0 \exp(p)}{1+\exp(p)} - \frac{a \exp(p)}{2(1+\exp(p))}\right\} \leq 0,\end{aligned}$$

where we used $\Psi(x,t) \geq \Psi(0,t) = \exp(p)$. Hence the first assertion of the proposition is true.

By integration we obtain

$$p(t) - \omega_0 t = \frac{L_0}{L_0 + \omega_0} \log\{\omega_0 + (L_0 + \omega_0)\exp(p(t))\} + q,$$

where

$$q := p_0 - \frac{L_0}{L_0 + \omega_0} \log\{\omega_0 + (L_0 + \omega_0)\exp(p_0)\}.$$

As $t \to -\infty$,

$$p(t) - \omega_0 t \to \frac{L_0}{L_0 + \omega_0} \log \omega_0 + q.$$

Hence, if we take p_0 as a solution of

$$\frac{L_0}{L_0 + \omega_0} \log \omega_0 + q = 0,$$

namely, a solution of (1.26), then

$$\lim_{t \to -\infty} (p(t) - \omega_0 t) = 0.$$

This yields

$$u_*(x,t) < u^*(x,t), \quad v_*(x,t) < v^*(x,t) \qquad ((x,t) \in \mathbb{R} \times \mathbb{R}),$$

and

$$\lim_{t \to -\infty} [\sup_{x \in \mathbb{R}}(u^*(x,t) - u_*(x,t)) + \sup_{x \in \mathbb{R}}(v^*(x,t) - v_*(x,t))] = 0.$$

The existence of the entire solution satisfying (1.25) is proven by the same argument found in [2] (see also [5] or [1]), which is left to the reader. □

Appendix. In this section we prove Lemmas 2.1, 2.2, and 2.3. We investigate the asymptotic behavior as $\xi \to \pm\infty$. For simplicity we will not specify the dependence on j in the notation as long as there is no confusion.

We put

$$(u, w, v, y) = (\phi_j(\xi), \phi_j'(\xi), \psi_j(\xi), \psi_j'(\xi)).$$

Then the equations of (1.5) are written as

$$\begin{pmatrix} u' \\ w' \\ v' \\ y' \end{pmatrix} = \begin{pmatrix} w \\ c_j w - f(u,v) \\ y \\ (c_j/d)y - (a/d)g(u,v) \end{pmatrix}, \tag{A.1}$$

where

$$f(u,v) := u(1-u-k_1 v), \qquad g(u,v) := v(1-v-k_2 u).$$

Proof of Lemma 2.1. We first consider the linearized equations of (A.1) at $(u,w,v,y) = (1,0,0,0)$:

$$\begin{pmatrix} W_1' \\ W_2' \\ Y_1' \\ Y_2' \end{pmatrix} = A_1 \begin{pmatrix} W_1 \\ W_2 \\ Y_1 \\ Y_2 \end{pmatrix}, \tag{A.2}$$

$$A_1 := \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & c_j & k_1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -(a/d)(1-k_2) & c_j/d \end{pmatrix}. \tag{A.3}$$

All the eigenvalues of A_1 consist of roots to (2.1) and (2.2). There is a negative eigenvalue satisfying (2.1) and a corresponding eigenvector, which are given by

$$\lambda_1 = -\frac{\sqrt{c_j^2+4}-c_j}{2}, \qquad \begin{pmatrix} 1 \\ \lambda_1 \\ 0 \\ 0 \end{pmatrix}. \tag{A.4}$$

We recall $k_2 > 1$. Since (2.2) has a negative root, A_1 has another negative eigenvalue and a corresponding eigenvector

$$\lambda_2^{(j)} = -\frac{\sqrt{c_j^2+4ad(k_2-1)}-c_j}{2d}, \qquad \begin{pmatrix} s_1 \\ s_1\lambda_2 \\ 1 \\ \lambda_2 \end{pmatrix}, \quad s_1 := \frac{k_1}{\lambda_2^2 - c_j\lambda_2 - 1}. \tag{A.5}$$

When $\lambda_1 = \lambda_2$, A_1 has a generalized eigenvector

$$\begin{pmatrix} 0 \\ 1 \\ \tau_1 \\ \lambda_1\tau_1 \end{pmatrix}, \quad \tau_1 := \frac{2\lambda_1 - c_j}{k_1}.$$

Thus every solution of (A.2), which converges to zeros as $\xi \to \infty$, is given by

$$\begin{pmatrix} W_1 \\ W_2 \\ Y_1 \\ Y_2 \end{pmatrix} = C_1 \begin{pmatrix} 1 \\ \lambda_1 \\ 0 \\ 0 \end{pmatrix} e^{\lambda_1\xi} + C_2 \begin{pmatrix} s_1 \\ s_1\lambda_2 \\ 1 \\ \lambda_2 \end{pmatrix} e^{\lambda_2\xi} \quad (\lambda_1 \neq \lambda_2)$$

or

$$\begin{pmatrix} W_1 \\ W_2 \\ Y_1 \\ Y_2 \end{pmatrix} = C_1 \begin{pmatrix} 1 \\ \lambda_1 \\ 0 \\ 0 \end{pmatrix} e^{\lambda_1 \xi} + C_2 \left\{ \begin{pmatrix} 1 \\ \lambda_1 \\ 0 \\ 0 \end{pmatrix} \xi + \begin{pmatrix} 0 \\ 1 \\ \tau_1 \\ \lambda_1 \tau_1 \end{pmatrix} \right\} e^{\lambda_1 \xi} \quad (\lambda_1 = \lambda_2),$$

where C_1, C_2 are arbitrarily given real numbers.

Applying the stable manifold theorem to (A.1) yields that as $\xi \to \infty$, there are α and β such that

$$\begin{cases} \phi(\xi) = 1 - \alpha \exp(\lambda_1 \xi) + \beta s_1 \exp(\lambda_2 \xi) + h.o.t., \\ \psi(\xi) = \beta \exp(\lambda_2 \xi) + h.o.t. \end{cases} \tag{A.6}$$

for $\lambda_1 \neq \lambda_2$, while there are $\tilde{\alpha}$ and $\tilde{\beta}$ such that

$$\begin{cases} \phi(\xi) = 1 - \tilde{\alpha} \exp(\lambda_1 \xi) - \tilde{\beta} \xi \exp(\lambda_1 \xi) + h.o.t., \\ \psi(\xi) = -\tilde{\beta} \tau_1 \exp(\lambda_1 \xi) + h.o.t. \end{cases} \tag{A.7}$$

for $\lambda_1 = \lambda_2$.

We show $\beta \neq 0$. Equation (A.1) allows a solution of the form $(u, w, 0, 0)$, which corresponds to a solution to

$$u' = w, \qquad w' = cw - f(u, 0) = cw - u(1 - u). \tag{A.8}$$

There is a stable manifold for the equilibrium $(u, w) = (1, 0)$ of (A.8). The stable manifold for $(u, w, v, y) = (1, 0, 0, 0)$ contains this dynamics that is obtained by putting $\beta = 0$ in (A.6). By this observation we can assert that $\beta \neq 0$ in (A.6) for the present traveling wave.

Since λ_1 is a negative root of (2.1), we notice that

$$s_1 > 0 \quad (\lambda_2 < \lambda_1 < 0), \qquad s_1 < 0 \quad (\lambda_1 < \lambda_2 < 0).$$

Recall that $\phi(\xi)$ and $\psi(\xi)$ are monotone increasing and decreasing, respectively. Thus for $\lambda_2 < \lambda_1$, $\alpha > 0$ and $\beta > 0$ follow from (A.6), while for $\lambda_1 < \lambda_2$, $\beta > 0$ follows.

When $\lambda_1 = \lambda_2$, we can let $\tilde{\beta} \neq 0$ in the same argument above. Since $\tau_1 < 0$ holds, $\tilde{\beta}$ must be positive in (A.7). This concludes the proof of the lemma. □

Next we consider the linearized equation at $(u, w, v, y) = (0, 0, 1, 0)$:

$$\boldsymbol{W}' = A_2 \boldsymbol{W}, \tag{A.9}$$

$$A_2 := \begin{pmatrix} 0 & 1 & 0 & 0 \\ k_1 - 1 & c_j & 0 & 0 \\ 0 & 0 & 0 & 1 \\ ak_2/d & 0 & a/d & c_j/d \end{pmatrix}, \qquad \boldsymbol{W} := \begin{pmatrix} W_1 \\ W_2 \\ Y_1 \\ Y_2 \end{pmatrix}. \tag{A.10}$$

In this case eigenvalues of A_2 are given by the roots of (1.9) and (1.10). We easily see from (1.9) that there is a positive eigenvalue and a corresponding eigenvector

$$\lambda_3 = \frac{c_j + \sqrt{c_j^2 + 4ad}}{2d}, \qquad \begin{pmatrix} 0 \\ 0 \\ 1 \\ \lambda_3 \end{pmatrix}. \tag{A.11}$$

Proof of Lemma 2.2. This corresponds to the bistable case (ii) stated in the introduction. In addition to the positive eigenvalue λ_3, there is a positive eigenvalue

$$\lambda_4 = \frac{c_j + \sqrt{c_j^2 + 4(k_1 - 1)}}{2}, \qquad \begin{pmatrix} 1 \\ \lambda_4 \\ s_2 \\ s_2\lambda_4 \end{pmatrix}, \quad s_2 := \frac{ak_2}{d\lambda_4^2 - c_j\lambda_4 - a} \tag{A.12}$$

unless λ_4 satisfies (1.9). When $\lambda_3 = \lambda_4$, we have a generalized eigenvector

$$\begin{pmatrix} \tau_2 \\ \tau_2\lambda_3 \\ 0 \\ 1 \end{pmatrix}, \quad \tau_2 := \frac{2d\lambda_3 - c_j}{ak_2}.$$

Hence every solution which converges to zeros as $\xi \to -\infty$ is given by

$$\boldsymbol{W}(\xi) = C_1\boldsymbol{p}_1 e^{\lambda_3\xi} + C_2\boldsymbol{p}_4 e^{\lambda_4\xi}$$

or

$$\boldsymbol{W}(\xi) = C_1\boldsymbol{p}_1 e^{\lambda_3\xi} + C_2(\boldsymbol{p}_1\xi + \boldsymbol{p}_2) e^{\lambda_3\xi},$$

where we put

$$\boldsymbol{p}_1 := \begin{pmatrix} 0 \\ 0 \\ 1 \\ \lambda_3 \end{pmatrix}, \quad \boldsymbol{p}_2 := \begin{pmatrix} \tau_2 \\ \tau_2\lambda_3 \\ 0 \\ 1 \end{pmatrix}, \quad \boldsymbol{p}_4 := \begin{pmatrix} 1 \\ \lambda_4 \\ s_2 \\ s_2\lambda_4 \end{pmatrix}.$$

Applying the unstable manifold theorem yields that as $\xi \to -\infty$, there exist α and β such that

$$\begin{cases} \phi_j(\xi) = \beta \exp(\lambda_4\xi) + h.o.t., \\ \psi_j(\xi) = 1 - \alpha \exp(\lambda_3\xi) + \beta s_2 \exp(\lambda_4\xi) + h.o.t. \end{cases} \tag{A.13}$$

for $\lambda_3 \neq \lambda_4$, while there are $\tilde{\alpha}$ and $\tilde{\beta}$ such that

$$\begin{cases} \phi_j(\xi) = \tilde{\beta}\tau_2 \exp(\lambda_3\xi) + h.o.t., \\ \psi_j(\xi) = 1 - \tilde{\alpha} \exp(\lambda_3\xi) + \tilde{\beta}\xi \exp(\lambda_3\xi) + h.o.t. \end{cases} \tag{A.14}$$

for $\lambda_3 = \lambda_4$. In this case we have an unstable manifold for the equilibrium $(v, y) = (1, 0)$ of

$$v' = y, \qquad y' = c_j y - ag(0, v),$$

which corresponds to $\beta = 0$ and $\tilde{\beta} = 0$ in (A.13) and (A.14), respectively. Thus we may assume that $\beta \neq 0$, $\tilde{\beta} \neq 0$. By a similar argument applying to the behavior as $\xi \to \infty$, we obtain the assertion in Lemma 2.2. □

Proof of Lemma 2.3. This corresponds to the monostable case (i) in the introduction. There is a family of traveling waves for $c_j \geq c_{min}^{(j)} \geq 2\sqrt{1 - k_1}$.

In addition to λ_3 and λ_4, we have a positive eigenvalue and a corresponding eigenvector,

$$\lambda_5 = \frac{c_j - \sqrt{c_j^2 - 4(1-k_1)}}{2}, \qquad \begin{pmatrix} 1 \\ \lambda_5 \\ s_3 \\ s_3\lambda_5 \end{pmatrix}, \quad s_3 := \frac{ak_2}{d\lambda_5^2 - c_j\lambda_5 - a}, \tag{A.15}$$

unless λ_5 satisfies (1.9). We note that $\lambda_5 \le \lambda_4$. If $\lambda_5 < \lambda_4$ and $\lambda_4, \lambda_5 \neq \lambda_3$, any solution converging to zeros as $\xi \to -\infty$ is given as

$$\boldsymbol{W}(\xi) = C_1\boldsymbol{p}_1 e^{\lambda_3\xi} + C_2\boldsymbol{p}_4 e^{\lambda_4\xi} + C_3\boldsymbol{p}_5 e^{\lambda_5\xi},$$

where

$$\boldsymbol{p}_5 := \begin{pmatrix} 1 \\ \lambda_5 \\ s_3 \\ s_3\lambda_5 \end{pmatrix}.$$

Applying the unstable manifold theorem, we obtain

$$\begin{cases} \phi_j(\xi) = \beta\exp(\lambda_4\xi) + \gamma\exp(\lambda_5\xi) + h.o.t., \\ \psi_j(\xi) = 1 - \alpha\exp(\lambda_3\xi) + \beta s_2\exp(\lambda_4\xi) + \gamma s_3\exp(\lambda_5\xi) + h.o.t. \end{cases}$$

By the same reason as in the previous case, we can assert $(\beta,\gamma) \neq (0,0)$. This leads to (2.12).

When $\lambda_5 = \lambda_4 < \lambda_3$, we obtain

$$\boldsymbol{W}(\xi) = C_1\boldsymbol{p}_1 e^{\lambda_3\xi} + C_2(\boldsymbol{p}_4\xi + \tilde{\boldsymbol{p}}_4)e^{\lambda_4\xi} + C_3\boldsymbol{p}_4 e^{\lambda_4\xi},$$

where

$$\tilde{\boldsymbol{p}}_4 := \begin{pmatrix} 0 \\ 1 \\ \tau_4 \\ \tau_5 \end{pmatrix}, \quad \begin{pmatrix} \tau_4 \\ \tau_5 \end{pmatrix} := \frac{s_2}{d\lambda_4^2 - c_j\lambda_4 - a}\begin{pmatrix} c_j - 2d\lambda_4 \\ -d\lambda_4^2 - a \end{pmatrix} \quad (\lambda_4 = \lambda_5 = c_j/2).$$

Applying the unstable manifold theorem to this case yields

$$\begin{cases} \phi_j(\xi) = \beta\exp(\lambda_4\xi) + \gamma\exp(\lambda_4\xi) + h.o.t., \\ \psi_j(\xi) = 1 - \alpha\exp(\lambda_3\xi) + \beta(s_2\xi + \tau_4)\exp(\lambda_4\xi) + \gamma s_2\exp(\lambda_4\xi) + h.o.t., \end{cases}$$

where we again see that $(\beta,\gamma) \neq (0,0)$. This leads to (2.14).

For the cases $\lambda_5 < \lambda_3 = \lambda_4$ and $\lambda_5 = \lambda_3 < \lambda_4$ all the solutions converging to zeros as $\xi \to -\infty$ are given by

$$\boldsymbol{W}(\xi) = C_1\boldsymbol{p}_1 e^{\lambda_3\xi} + C_2(\boldsymbol{p}_1\xi + \boldsymbol{p}_2)e^{\lambda_3\xi} + C_3\boldsymbol{p}_5 e^{\lambda_5\xi}$$

and

$$\boldsymbol{W}(\xi) = C_1\boldsymbol{p}_1 e^{\lambda_3\xi} + C_2(\boldsymbol{p}_1\xi + \boldsymbol{p}_2)e^{\lambda_3\xi} + C_3\boldsymbol{p}_4 e^{\lambda_4\xi},$$

respectively. When $\lambda_3 = \lambda_4 = \lambda_5 = c_j/2$, we obtain one more generalized eigenvector so that the solution is given by

$$\boldsymbol{W}(\xi) = C_1\boldsymbol{p}_1 e^{\lambda_3\xi} + C_2(\boldsymbol{p}_1\xi + \boldsymbol{p}_2)e^{\lambda_3\xi} + C_3(\boldsymbol{p}_1\xi^2/2 + \boldsymbol{p}_2\xi + \boldsymbol{p}_3)e^{\lambda_3\xi},$$

where

$$\boldsymbol{p}_3 := \begin{pmatrix} d/ak_2 \\ \lambda_3(3d-2)/ak_2 \\ 0 \\ 0 \end{pmatrix}.$$

By those solutions it is not so difficult to prove the remaining cases (iii), (iv), and (v) of Lemma 2.3. Completing the proof is left to the reader. □

Acknowledgments. The authors would like to express their sincere thanks to Professor Yukio Kan-on for valuable comments on the asymptotic behavior of the monotone traveling wave to the Lotka–Volterra competition-diffusion equations. Their thanks also go to the referees for the useful comments on the revision of the paper.

REFERENCES

[1] X. Chen and J.-S. Guo, *Existence and uniqueness of entire solutions for a reaction-diffusion equation*, J. Differential Equations, 212 (2005), pp. 62–84.
[2] Y. Fukao, Y. Morita, and H. Ninomiya, *Some entire solutions of the Allen-Cahn equation*, Taiwanese J. Math., 8 (2004), pp. 15–32.
[3] R. A. Gardner, *Existence and stability of travelling wave solutions of competition models: A degree theoretic approach*, J. Differential Equations, 44 (1982), pp. 343–364.
[4] R. A. Gardner and C. K. R. T. Jones, *Stability of travelling wave solutions of diffusive predator-prey systems*, Trans. Amer. Math. Soc., 327 (1991), pp. 465–524.
[5] J.-S. Guo and Y. Morita, *Entire solutions of reaction-diffusion equations and an application to discrete diffusive equations*, Discrete Contin. Dynam. Systems, 12 (2005), pp. 193–212.
[6] F. Hamel and N. Nadirashvili, *Entire solutions of the KPP equation*, Comm. Pure Appl. Math., 52 (1999), pp. 1255–1276.
[7] Y. Hosono, *Singular perturbation analysis of travelling waves for diffusive Lotka-Volterra competition models*, in Numerical and Applied Mathematics, Part II (Paris, 1988), IMACS Ann. Comput. Appl. Math. 1.2, Baltzer, Basel, 1989, pp. 687–692.
[8] Y. Kan-on, *Parameter dependence of propagation speed of travelling waves for competition-diffusion equations*, SIAM J. Math. Anal., 26 (1995), pp. 340–363.
[9] Y. Kan-on, *Existence of standing waves for competition-diffusion equations*, Japan J. Indust. Appl. Math., 13 (1996), pp. 117–133.
[10] Y. Kan-on, *Fisher wave fronts for the Lotka-Volterra competition model with diffusion*, Nonlinear Anal., 28 (1997), pp. 145–164.
[11] Y. Kan-on, *Instability of stationary solutions for a Lotka-Volterra competition model with diffusion*, J. Math. Anal. Appl., 208 (1997), pp. 158–170.
[12] Y. Kan-on and Q. Fang, *Stability of monotone travelling waves for competition-diffusion equations*, Japan J. Indust. Appl. Math., 13 (1996), pp. 343–349.
[13] Y. Morita and H. Ninomiya, *Entire solutions with merging fronts to reaction-diffusion equations*, J. Dynam. Differential Equations, 18 (2006), pp. 841–861.
[14] M. Rodrigo and M. Mimura, *Exact solutions of a competition-diffusion system*, Hiroshima Math. J., 30 (2000), pp. 257–270.
[15] M. M. Tang and P. C. Fife, *Propagating fronts for competing species equations with diffusion*, Arch. Ration. Mech. Anal., 73 (1980), pp. 69–77.
[16] H. Yagisita, *Backward global solutions characterizing annihilation dynamics of travelling fronts*, Publ. Res. Inst. Math. Sci., 39 (2003), pp. 117–164.

SIAM J. MATH. ANAL.
Vol. 40, No. 6, pp. 2241–2261

HIGHER ORDER APPROXIMATIONS IN THE HEAT EQUATION AND THE TRUNCATED MOMENT PROBLEM*

YONG JUNG KIM† AND WEI-MING NI‡

Abstract. In this paper, we employ linear combinations of n heat kernels to approximate solutions to the heat equation. We show that such approximations are of order $O(t^{(\frac{1}{2p}-\frac{2n+1}{2})})$ in L^p-norm, $1 \le p \le \infty$, as $t \to \infty$. For positive solutions of the heat equation such approximations are achieved using the theory of truncated moment problems. For general sign-changing solutions these type of approximations are obtained by simply adding an auxiliary heat kernel. Furthermore, inspired by numerical computations, we conjecture that such approximations converge geometrically as $n \to \infty$ for any fixed $t > 0$.

Key words. heat equation, moments, asymptotics convergence rates, approximation of an integral formula, heat kernel

AMS subject classifications. 35K05, 78M05, 41A10

DOI. 10.1137/08071778X

1. Introduction.

It is well known that

$$u(x,t) = \int \frac{u_0(c)}{\sqrt{4\pi t}}\, e^{\frac{-(x-c)^2}{4t}}\, dc \tag{1.1}$$

is the physically meaningful solution to the heat equation

$$u_t = u_{xx}, \quad u(x,0) = u_0(x) \in L^1(\boldsymbol{R}), \qquad x, u \in \boldsymbol{R}, \quad t > 0, \tag{1.2}$$

where, for simplicity, the initial value $u_0(x)$ is assumed to be continuous. In this paper, we shall refer to (1.1) as the solution of the heat equation (1.2) for the sake of brevity. If a general L^1 initial value is considered, no asymptotic convergence order to a fundamental solution is expected in L^1-norm. Hence, the asymptotic convergence order is usually studied under suitable restrictions on its initial value $u_0(x)$ for $|x|$ large.

Since the analysis of this paper is based on the *moments* of the solution, the initial value $u_0(x)$ is required to have finite moments up to certain order, say, $2n$. We set $x^{2n}u_0(x) \in L^1(\boldsymbol{R})$ and the moments of the initial value $u_0(x)$ as

$$\gamma_k := \int x^k u_0(x)dx < \infty, \qquad k = 0, 1, \ldots, 2n. \tag{1.3}$$

For example, if the initial value has an algebraic decay order higher than $2n+1$ for $|x|$ large, i.e., for $\varepsilon > 0$,

$$u_0(x) = O\left(|x|^{-(2n+1+\varepsilon)}\right) \quad \text{as} \quad |x| \to \infty, \tag{1.4}$$

*Received by the editors March 4, 2008; accepted for publication (in revised form) September 23, 2008; published electronically February 20, 2009. This research was supported in part by the National Science Foundation.

http://www.siam.org/journals/sima/40-6/71778.html

†Department of Mathematical Sciences, KAIST, 335 Gwahangno, Yuseong-gu, Daejeon, 305-701, Republic of Korea (yongkim@kaist.edu). This author was supported by the Korea Science and Engineering Foundation (KOSEF) grant R01-2007-000-11307-0 funded by the Korean government (MOST).

‡School of Mathematics, University of Minnesota, 206 Church St. S.E., Minneapolis, MN 55455 (ni@math.umn.edu).

then the moments are well defined up to order $2n$. In the study of asymptotics the initial value is frequently assumed to have the order that a fundamental solution has for $|x|$ large. For the heat equation case the fundamental solution is the Gaussian and the corresponding decay order is $u_0(x) = O(e^{-x^2})$ as $|x| \to \infty$. Hence, the moment γ_k is defined for all order $k \geq 0$.

One may do the integration in the explicit formula (1.1) only approximately, even though the integration gives the exact value of the solution. In numerical computations finding an efficient way to compute such an integration has been an important issue. From this point of view, it seems useful to consider its approximation in a simpler form. Duoandikoetxea and Zuazua [9] showed that the following linear combination of derivatives of the Gaussian

$$\psi_{2n}(x,t) \equiv \sum_{i=0}^{2n-1} \frac{(-1)^i \gamma_i}{(i!)\sqrt{4\pi t}}\, \partial_x^i \left(e^{\frac{-x^2}{4t}} \right) \tag{1.5}$$

approaches to the solution u with a convergence order of

$$\|u(t) - \psi_{2n}(t)\|_p = O\left(t^{\left(\frac{1}{2p} - \frac{2n+1}{2}\right)} \right) \quad \text{as} \quad t \to \infty \ \text{ for } \ 1 \leq p \leq \infty, \tag{1.6}$$

where $\|\cdot\|_p$ denotes the L^p-norm in the whole space $\boldsymbol{R}$ and ∂_x^i the ith order partial differentiation with respect to x. Note that the original multidimensional result is written in a one-dimensional (1D) version for an easier comparison. This asymptotic convergence order indicates that ψ_{2n} is a good approximation of the solution $u(x,t)$ for t large. However, it does not necessarily mean that ψ_{2n} is a good approximation as $n \to \infty$ with a fixed $t > 0$. In fact, Table 7.3 shows that this L^p-norm difference may diverge geometrically as $n \to \infty$ if the fixed time $t > 0$ is not large enough. This is not surprising since the high order derivatives of the Gaussian in (1.5) diverge as their orders increase.

In this article we consider a linear combination of "n" heat kernels

$$\phi_n(x,t) \equiv \sum_{i=1}^{n} \frac{\rho_i}{\sqrt{4\pi t}}\, e^{\frac{-(x-c_i)^2}{4t}} \tag{1.7}$$

as an approximation to the solution $u(x,t)$. One may regard this summation as a discrete version of the integration in (1.1) by considering c_i's as grid points and ρ_i's as approximations of $u_0(c)dc$ in the interval (c_{i-1}, c_i). However, we employ these $2n$ degrees of freedom, ρ_i's and c_i's, to match the first $2n$ initial moments, i.e., to satisfy the following $2n$ moment equations:

$$\lim_{t \to 0} \int x^k \phi_n(x,t)dx = \gamma_k, \qquad k = 0, 1, \ldots, 2n-1. \tag{1.8}$$

If the initial value is positive, the theory of truncated moment problems [3] gives the solvability of this problem. Then ϕ_n and u share identical first $2n$ moments for all $t \geq 0$. Note that ψ_{2n} in (1.5) also has the same property, and Duoandikoetxea and Zuazua obtained the convergence order in (1.6) based on it. Hence, we may obtain the same convergence order for the approximation $\phi_n(x,t)$. In this paper, we actually go a little bit further and obtain the limit of $t^{\frac{2n+1}{2} - \frac{1}{2p}} \|u(t) - \phi_n(t)\|_p$ as $t \to \infty$. This convergence order is then improved in Lemma 2.3 for the case that this limit becomes zero. A multidimensional extension of this approach requires a theory of

multidimensional truncated moment problems. One may find one from a recent work by Curto and Fialkow [4].

From a practical point of view, it is desirable if the solution u can be approximated by ϕ_n as $n \to \infty$ for a fixed $t > 0$. Indeed, our numerical examples in section 7.2 indicate the following geometric convergence order:

$$\frac{\|u(t) - \phi_n(t)\|_\infty}{\|u(t) - \phi_{n+1}(t)\|_\infty} \to 1 + 4\frac{t}{v} \quad \text{as} \quad n \to \infty, \tag{1.9}$$

where the constant $v > 0$ depends on the initial value $u_0(x)$. However, its proof has, thus far, eluded us; nevertheless, we will include a discussion of (1.9) in section 6.

This paper is organized as follows. First, in section 2, we compute the limit of $t^{(\frac{2n+1}{2}-\frac{1}{2p})}\|u(t) - \phi_n(t)\|_p$ as $t \to \infty$ under the assumption (1.8), which gives the convergence order in (1.6). A short introduction to the theory of truncated moment problems is given in section 3, which provides the existence and the uniqueness of ρ_i's and c_i's that solve (1.8). We remark that the theory is applicable for nonnegative initial values only (see [1, 3]). For general sign-changing solutions the existence and the uniqueness of such ρ_i's and c_i's do not hold. In section 4 we discuss this issue in detail for three cases with $n = 1, 2$, and 3. In section 5 we construct approximations for general sign-changing cases by adding an auxiliary heat kernel or by assigning c_i's independently. The conjectured geometric convergence order for large $n > 0$ is discussed in section 6. The asymptotic convergence orders as $t \to \infty$ or $n \to \infty$ are numerically tested in sections 7.1 and 7.2. The convergence of the alternative approach using ψ_{2n} and the conjectured statements in section 6 are numerically tested in sections 7.3 and 7.4.

In the study of nonlinear diffusion or convection, fundamental solutions which have the Dirac measure as their initial value, i.e., $u_0(x) = \delta(x)$, often serve as canonical solutions. The Barenblatt solutions, the diffusion waves, and the N-waves are well-known examples (see [23]). In the study of porous medium equations, the Barenblatt solution is used as an asymptotic profile and the convergence order of general solutions to this special one has been studied in various cases (see [2, 6] and references therein). The diffusion wave and the Gaussian are the asymptotics of convection-diffusion equations for diffusion dominant cases (see [10, 11, 12, 16]). For convection dominant cases (see [14]) and inviscid convection equations (see [5, 8, 13, 20]) or hyperbolic systems (see [7, 17, 18]), N-waves represent the asymptotic behavior, where N-waves can be understood as a special solution with initial value $u_0(x) = \lim_{\varepsilon\to 0}[a\delta(x-\varepsilon) - b\delta(x+\varepsilon)]$ with $a, b > 0$. Placing the Dirac measure at the center of mass, the optimal convergence order of $O(t^{\frac{1}{2p}-\frac{3}{2}})$ in L^p-norm (or of $O(t^{-1})$ in L^1-norm) has been obtained in several cases (see [2, 13, 15]). Therefore, the result of this paper can be viewed as an extreme case that exploits all of the moments of the initial value.

The approach in this paper can be directly employed to approximate the solutions to the Burgers equation via the Cole–Hopf transformation. To obtain the rigorous convergence order for the Burgers case it is required to check the well definedness of the transformed solutions as is done in Lemmas 3.1–3.2 and Theorem 3.3 in [15] for the special case $n = 1$. Considering that the Burgers equation has been used as a tool to study the asymptotic structure of the viscous systems of conservation laws (see, e.g., [19]), we hope the approach in this article may be useful for other general models.

2. Asymptotic convergence order. In this section, we show that the decay rate of a derivative of a solution is naturally transferred to the convergence order

of our approximation. This connection will be made by assigning the moments of a solution to its approximation. Let $\gamma_k(t)$ be the kth order moment of a solution $u(x,t)$ at time $t \geq 0$, i.e.,

$$\gamma_k(t) = \int x^k u(x,t)dx, \qquad k = 0, 1, 2, \ldots, t \geq 0.$$

(Notice that we are slightly abusing the notation γ_k in (1.3) in the following couple of paragraphs.) We can easily show how the moment $\gamma_k(t)$ evolves as $t \to \infty$.

LEMMA 2.1. *Let $u(x,t)$ be the solution to the heat equation and $\gamma_k(t)$ be its kth order moment at time $t \geq 0$. Then*

$$\frac{d}{dt}\gamma_k(t) = \begin{cases} 0\ , & k = 0 \text{ or } 1, \\ k(k-1)\gamma_{k-2}(t), & k \geq 2. \end{cases}$$

Proof. For $k = 0$, the lemma is equivalent to the conservation of mass. For $k = 1$, since $u_t = u_{xx}$, the integration by parts gives

$$\gamma_1'(t) = \int x u_t dx = \int x u_{xx} dx = [x u_x - u]_{-\infty}^{\infty} = 0.$$

Similarly, for $k \geq 2$, we obtain

$$\begin{aligned} \gamma_k'(t) &= \int x^k u_t dx = \int x^k u_{xx} dx \\ &= \left[x^k u_x - k x^{k-1} u\right]_{-\infty}^{\infty} + \int k(k-1)x^{k-2} u dx = k(k-1)\gamma_{k-2}(t). \end{aligned}$$ ∎

This lemma shows that even numbered moments and odd numbered ones evolve independently. One may explicitly write

$$\begin{aligned} \gamma_{2n}(t) &= \sum_{k=0}^{n} \frac{(2n)!}{(n-k)!(2k)!} t^{n-k} \gamma_{2k}(0), \\ \gamma_{2n+1}(t) &= \sum_{k=0}^{n} \frac{(2n+1)!}{(n-k)!(2k+1)!} t^{n-k} \gamma_{2k+1}(0). \end{aligned} \tag{2.1}$$

If $\gamma_k(0) = 0$ for all $0 \leq k \leq n$, then $\gamma_k(t) = 0$ for all $0 \leq k \leq n$, $\gamma_k(t) = \gamma_k(0)$ for $k = n+1, n+2$, $\gamma_k(t)$ is linear for $k = n+3, n+4$, $\gamma_k(t)$ is quadratic for $k = n+5, n+6$, and so on.

Let $v(x,t)$ be an approximation solution of the exact one $u(x,t)$. Since the difference $E(x,t) = v(x,t) - u(x,t)$ is also a solution to the heat equation, the moments of $E(x,t)$ will be always zero up to certain order if they are initially zero. Hence, it is natural to expect a higher convergence order by matching the moments of the approximation solution to those of the exact one. We proceed with our discussion in this respect.

LEMMA 2.2. *If $x^m E_0(x) \in L^1(\boldsymbol{R})$ and*

$$\int_{-\infty}^{\infty} x^k E_0(x)dx = 0 \quad \textit{for all} \quad 0 \leq k < m, \tag{2.2}$$

then there exists $E_m \in W^{m,1}(\boldsymbol{R})$ such that

$$\partial_x^m E_m(x) = E_0(x). \tag{2.3}$$

Proof. The proof was given by Duoandikoetxea and Zuazua [9] for the multidimensional case. Here we provide its 1D counterpart. Consider a sequence of functions defined inductively by

$$E_k(x) = \int_{-\infty}^{x} E_{k-1}(y)dy, \quad 0 < k \le m. \tag{2.4}$$

First, we show that E_k's are well defined,

$$\int_{-\infty}^{\infty} E_k(x)dx = 0 \qquad \text{and} \qquad E_k(x) \to 0 \ \text{ as } |x| \to \infty \tag{2.5}$$

for $k = 0, 1, \dots, m-1$. It suffices to show (2.5) for $k = l < m$ under the assumption that (2.5) holds for all $k = 0, 1, \dots, l-1$. Note that it is clearly satisfied for $k = 0$. Since $\int_{-\infty}^{\infty} E_{l-1}(x)dx = 0$, the integral $E_l(x)$ also decays to zero as $|x| \to \infty$. Using the fact that E_k decays to zero as $|x| \to \infty$ for all $0 \le k \le l$, we obtain

$$\int_{-\infty}^{\infty} E_l(x)dx = (-1)^l \int_{-\infty}^{\infty} \frac{x^l}{l!} E_0(x)dx = 0$$

using the integration by parts and then (2.2). Therefore, (2.5) holds for $k = l$ and, hence, for all $0 \le k \le m-1$. Since $x^m E_0(x) \in L^1(\boldsymbol{R})$, $E_m \in W^{m,1}(\boldsymbol{R})$ and (2.3) is satisfied. ∎

The existence of E_m satisfying (2.3) is the key observation to obtain the asymptotic convergence order. We now continue our discussion under the assumption that the initial value $E_0(x)$ satisfies (2.2) and $E_m(x)$ is its mth order antiderivative given in Lemma 2.2. However, the following discussions about the decay rate of derivatives of a solution can be considered independently. Let $E_m(x,t)$ be the solution to the heat equation with initial value $E_m(x,0) = E_m(x) \in W^{m,1}(\boldsymbol{R})$, i.e.,

$$E_m(x,t) = \frac{1}{\sqrt{4\pi t}} \int e^{-(x-y)^2/(4t)} E_m(y)dy.$$

The dissipation of the solution can be easily shown by introducing similarity variables:

$$\xi = \frac{x}{\sqrt{t}}, \quad \zeta = \frac{y}{\sqrt{t}}, \quad \tilde{E}_m(\xi, t) = E_m(x,t).$$

Then $E_m(x,t)$ is transformed to

$$\tilde{E}_m(\xi,t) = \frac{1}{\sqrt{4\pi}} \int e^{-(\xi-\zeta)^2/4} E_m(\sqrt{t}\zeta)d\zeta,$$

and its mth order derivative is given by

$$\partial_\xi^m \tilde{E}_m(\xi,t) = \partial_x^m E_m(x,t)(\partial_\xi x)^m = \partial_x^m E_m(x,t)\big(\sqrt{t}\big)^m.$$

Now consider the decay order of the mth order derivative of the solution $E_m(x,t)$. First, let $C_m := |\int E_m(y)dy|$ and consider the case $C_m \neq 0$. Then

$$\big(\sqrt{t}\big)^{m+1} |\partial_x^m E_m(x,t)| = \sqrt{t}\left|\partial_\xi^m \tilde{E}_m(\xi,t)\right| = \frac{C_m}{\sqrt{4\pi}}\left|\int f(\zeta) g_t(\xi-\zeta)d\zeta\right|, \tag{2.6}$$

where

$$g_t(\xi) = \sqrt{t}E_m\big(\sqrt{t}\xi\big)/C_m, \ f(\xi) = \partial_\xi^m\left(e^{-\xi^2/4}\right). \tag{2.7}$$

After taking the supremum on both sides of (2.6), one obtains that

$$\left(\sqrt{t}\right)^{m+1}\|\partial_x^m E_m(t)\|_\infty \le \frac{C_m}{\sqrt{4\pi}}\left\|\partial_\xi^m\left(e^{-\xi^2/4}\right)\right\|_\infty.$$

If one takes $t\to\infty$ limit to (2.6), then

$$\lim_{t\to\infty}\left(\sqrt{t}\right)^{m+1}|\partial_x^m E_m(x,t)| = \frac{C_m}{\sqrt{4\pi}}|f(\xi)|.$$

Therefore, after taking the supremum on both sides again, we obtain

$$\lim_{t\to\infty}\left(\sqrt{t}\right)^{m+1}\|\partial_x^m E_m(t)\|_\infty = \frac{C_m}{\sqrt{4\pi}}\left\|\partial_\xi^m\left(e^{-\xi^2/4}\right)\right\|_\infty.$$

On the other hand, if $1\le p<\infty$, then

$$\begin{aligned} t^{\left(\frac{m+1}{2}-\frac{1}{2p}\right)}\ &\|\partial_x^m E_m(t)\|_p \\ &= \left(\sqrt{t}\right)^{m+1}\left(\frac{1}{\sqrt{t}}\right)^{1/p}\left(\int|\partial_x^m E_m(x,t)|^p dx\right)^{1/p} \\ &= \left(\int\left|\left(\sqrt{t}\right)^{m+1}\partial_x^m E_m(x,t)\right|^p d\left(\frac{x}{\sqrt{t}}\right)\right)^{1/p} \\ &= \left(\int\left|\sqrt{t}\partial_\xi^m\tilde{E}_m(\xi,t)\right|^p d\xi\right)^{1/p} \\ &= \frac{C_m}{\sqrt{4\pi}}\left(\int\left|\int f(\zeta)g_t(\xi-\zeta)d\zeta\right|^p d\xi\right)^{1/p} = \frac{C_m}{\sqrt{4\pi}}\|f*g_t\|_p. \end{aligned} \tag{2.8}$$

Standard arguments imply that $\|f*g_t\|_p\to\|f\|_p$ as $t\to\infty$ (see [21, p. 62]). Therefore,

$$\lim_{t\to\infty} t^{\left(\frac{m+1}{2}-\frac{1}{2p}\right)}\|\partial_x^m E_m(t)\|_p = \frac{C_m}{\sqrt{4\pi}}\left\|\partial_\xi^m\left(e^{-\xi^2/4}\right)\right\|_p.$$

Now we consider the case that $C_m=0$. Then one can easily show that this limit is zero. In fact, we will improve the convergence order by working with higher order antiderivatives. Let

$$E_k(x) = \int_{-\infty}^x E_{k-1}(y)dy, \qquad k>m. \tag{2.9}$$

We can easily show that $\int_{-\infty}^\infty E_{k_0}(x)dx(=\lim_{x\to\infty}E_{k_0+1}(x))\neq 0$ for some $k_0>m$. Suppose that $\int_{-\infty}^\infty E_k(x)dx=0$ for all $k>m$. Then $|E_k(x)|$ decays to zero for $|x|$ large, and, therefore, after integrating by parts k times with proper inductive arguments, one obtains

$$(-1)^k k!\int_{-\infty}^\infty E_k(x)dx = \int_{-\infty}^\infty x^k E_0(x)dx = 0.$$

On the other hand, by the Weierstrass approximation theorem, there exists a sequence of polynomials P_n such that

$$P_n(x)\to E_0(x) \qquad \text{as} \qquad n\to\infty$$

uniformly on any bounded domain $[-L, L]$. Therefore, we obtain

$$\|E_0\|_2^2 = \int_{-L}^{L} E_0^2(x)dx = \lim_{n\to\infty} \int_{-L}^{L} P_n(x)E_0(x)dx = 0.$$

Hence, if the initial value E_0 is not a trivial one, there exists $k_0 > m$ such that $\lim_{x\to\infty} E_k(x) = 0$ for all $0 \le k \le k_0$ and $C_{k_0} := |\lim_{x\to\infty} E_{k_0+1}(x)| \neq 0$. If $E_0(x)$ decays with an algebraic order $k_0+1+\varepsilon$, $\varepsilon > 0$, for $|x|$ large, then $C_{k_0} < \infty$. However, C_{k_0} can be unbounded in general.

Let $E_{k_0}(x,t)$ be the solution with $E_{k_0}(x)$ as its initial value. Then, clearly, $\partial_x^m E_m = \partial_x^{k_0} E_{k_0} = E_0$ and, hence,

$$\lim_{t\to\infty} t^{\left(\frac{k_0+1}{2}-\frac{1}{2p}\right)} \|\partial_x^m E_m\|_p = \lim_{t\to\infty} t^{\left(\frac{k_0+1}{2}-\frac{1}{2p}\right)} \left\|\partial_x^{k_0} E_{k_0}\right\|_p = \frac{C_{k_0}}{\sqrt{4\pi}} \left\|\partial_\xi^{k_0} e^{\frac{-\xi^2}{4}}\right\|_p .$$

Therefore, if $C_m := |\int E_m(y)dy| = 0$, one obtains a higher decay order. Summing up, we obtain the following lemma.

LEMMA 2.3. *Let $E_m(x,t)$ be the solution to the heat equation with a nontrivial initial value $E_m(x) \in W^{m,1}(\boldsymbol{R})$ and E_k's be given inductively by* (2.9). *Then there exists $k_0 \ge m$ such that $\lim_{x\to\infty} E_k(x) = 0$ for $0 \le k \le k_0$ and $0 \neq |\lim_{x\to\infty} E_{k_0+1}(x)|$, and, for $m \le k \le k_0$,*

$$\lim_{t\to\infty} t^{\left(\frac{k+1}{2}-\frac{1}{2p}\right)} \|\partial_x^m E_m(t)\|_p = \frac{\left\|\partial_\xi^k e^{-\xi^2/4}\right\|_p}{\sqrt{4\pi}} \left|\int E_k(x)dx\right| , \ 1 \le p \le \infty. \tag{2.10}$$

If $\int E_m(x)dx = 0$, then the limit in (2.10) implies that $\lim_{t\to\infty} t^{\left(\frac{m+1}{2}-\frac{1}{2p}\right)} \|\partial_x^m E_m\|_p = 0$. Hence, we may simply say that

$$\lim_{t\to\infty} t^{\left(\frac{m+1}{2}-\frac{1}{2p}\right)} \|\partial_x^m E_m(t)\|_p = \frac{\left\|\partial_\xi^m \left(e^{-\xi^2/4}\right)\right\|_p}{\sqrt{4\pi}} \left|\int E_m(x)dx\right| , \quad 1 \le p \le \infty, \tag{2.11}$$

which is a weaker statement than (2.10) is. Note that one may obtain the upper bound of the term $t^{\left(\frac{m+1}{2}-\frac{1}{2p}\right)} \|\partial_x^m E_m(t)\|_p$ using Young's inequality. In fact, the corresponding upper bound for the estimate ψ_{2n} was obtained in [9].

In the following, we take the convergence order in (2.11) for simplicity. If an optimal convergence order is concerned and $\int E_m(x)dx = 0$, then one may refer to (2.10). It is well known that an L^1 solution to the heat equation decays to zero with order $O(t^{-1/2})$. Lemma 2.3 says that the decay order of its derivative is increased by $\frac{1}{2}$ after each differentiation. The asymptotic convergence order between two solutions is now obtained as a corollary of previous lemmas.

THEOREM 2.4. *Let $u(x,t)$ and $v(x,t)$ be solutions of the heat equation with initial values $u_0(x)$ and $v_0(x)$, respectively. Suppose that the initial difference $E_0(x) := u_0(x) - v_0(x)$ satisfies the assumptions in Lemma* 2.2. *Then, for $1 \le p \le \infty$,*

$$\lim_{t\to\infty} t^{\left(\frac{m+1}{2}-\frac{1}{2p}\right)} \|u(t) - v(t)\|_p = \frac{\left\|\partial_\xi^m \left(e^{-\frac{1}{4}\xi^2}\right)\right\|_p}{\sqrt{4\pi}} \left|\int E_m(x)dx\right| , \tag{2.12}$$

where $E_m \in W^{m,1}(\boldsymbol{R})$ is the one that satisfies $\partial_x^m E_m(x) = E_0(x)$.

Proof. Let $E_m(x,t)$ be the solution to the heat equation with initial value $E_m(x)$. Then $\partial_x^m E_m(x,t)$ is the solution to the heat equation with initial value $\partial_x^m E_m(x) = E_0(x)$. Hence, $\partial_x^m E_m(x,t)(= E(x,t)) = u(x,t) - v(x,t)$ and (2.12) follows from (2.11). □

Remark 2.5. In this section, we basically considered the convergence order of $\phi_n(x,t)$ ($\cong v(x,t)$) to $u(x,t)$ as $t \to \infty$ with a fixed $n > 0$. However, the relation (2.6), for example, provides certain convergence information as $n \to \infty$ with a fixed $t > 0$, too. To obtain a convergence order as $n \to \infty$ we need to specify $E_m(x)$ corresponding to our approximation $\phi_n(x,t)$, which will be considered in section 6.

3. Positive solutions and truncated moment problems. Consider a linear combination of heat kernels

$$\phi_n(x,t) := \sum_{i=1}^{n} \frac{\rho_i}{\sqrt{4\pi t}}\, e^{-(x-c_i)^2/(4t)}. \tag{3.1}$$

The $2n$ freedom of choices in ρ_i's and c_i's are used to control the first $2n$ moments of the approximation. Remember that γ_k is to denote the initial kth moment, i.e.,

$$\gamma_k := \int x^k u_0(x) dx, \qquad k = 0, 1, \ldots, 2n-1. \tag{3.2}$$

Let $\mathbf{r}_k$ be a column n-vector and $\mathbf{A}$ be the $n \times n$ Hankel matrix given by

$$\begin{array}{ll} \mathbf{r}_k = (\gamma_k, \gamma_{k+1}, \ldots, \gamma_{k+n-1})^t, & k = 0, 1, \ldots, n, \\ \mathbf{A} \equiv (a_{ij}) = (\gamma_{i+j}), & i, j = 0, 1, \ldots, n-1. \end{array} \tag{3.3}$$

Since $\phi_n(x,t) \to \sum_{i=1}^n \rho_i \delta_{c_i}(x)$ as $t \to 0$, the difference between the initial value and its approximation is

$$E_0(x) := u_0(x) - \sum_{i=1}^{n} \rho_i \delta_{c_i}(x),$$

where $\delta_{c_i}(x)$ is the Dirac measure centered at c_i, i.e., $\delta_{c_i}(x) = \delta(x - c_i)$. Hence, the zero moment conditions in (2.2) can be written as

$$\int \sum_{i=1}^{n} x^k \rho_i \delta_{c_i}(x)\, dx = \int x^k u_0(x)\, dx (\equiv \gamma_k), \qquad 0 \le k \le 2n-1, \tag{3.4}$$

or, in a matrix form, as

$$\begin{pmatrix} 1 & \cdots & 1 \\ c_1 & \cdots & c_n \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ c_1^{2n-1} & \cdots & c_n^{2n-1} \end{pmatrix} \begin{pmatrix} \rho_1 \\ \cdot \\ \cdot \\ \rho_n \end{pmatrix} = \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \cdot \\ \cdot \\ \cdot \\ \gamma_{2n-1} \end{pmatrix}. \tag{3.5}$$

After eliminating all ρ_i's (see section 4.3), one may obtain n-equations involving c_i's only:

$$\mathbf{A}\Psi = \mathbf{r}_n, \tag{3.6}$$

where the column vector $\Psi = (\psi_0, \ldots, \psi_{n-1})^t$ is given by

$$\psi_0 = (-1)^{n+1} \prod_{i=1}^{n} c_i, \quad \psi_1 = (-1)^n \sum_{j=1}^{n} \prod_{i \neq j} c_i, \ldots, \psi_{n-1} = \sum_{i=1}^{n} c_i. \tag{3.7}$$

Consequently, we set

$$g_n(x) := x^n - \sum_{j=0}^{n-1} \psi_j x^j = (x - c_1)(x - c_2) \cdots (x - c_n). \tag{3.8}$$

(Note that the coefficient of the leading order term is 1 and, hence, $g_n(x) \to \infty$ as $x \to \infty$.) Hence, if the initial moments in (3.4) are satisfied, then c_i's are zero points of the polynomial $g_n(x)$, where its coefficients are given as a solution of (3.6).

To show the existence and the uniqueness of the approximation we should show that the Hankel matrix in (3.6) is nonsingular. Then there exists a unique column vector $\Psi = (\psi_0, \ldots, \psi_{n-1})^t$ that satisfies (3.6). The next thing to show is that the polynomial $g_n(x)$ in (3.8) has n distinct real zeros $c_1 < \cdots < c_n$. Then ρ_i's are given by solving the Vandermonde given by the first n-equations in (3.5), i.e.,

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ c_1 & c_2 & \cdots & c_n \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ c_1^{n-1} & c_2^{n-1} & \cdots & c_n^{n-1} \end{pmatrix} \begin{pmatrix} \rho_1 \\ \rho_2 \\ \cdot \\ \cdot \\ \cdot \\ \rho_n \end{pmatrix} = \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \cdot \\ \cdot \\ \cdot \\ \gamma_{n-1} \end{pmatrix}. \tag{3.9}$$

It is well known that the Vandermonde matrix is nonsingular if c_i's are all different. Then we can easily check that c_i's and ρ_i's also satisfy the last n-equations in (3.5).

For a general sign-changing initial value $u_0(x)$ the Hankel matrix $\mathbf{A}$ can be singular, and examples are given in section 4. However, if the initial value $u_0(x)$ is nonnegative, then the uniqueness and the existence are resolved by the theory for the moment problem (see [1, 3]). In the following, we assume $u_0(x) \geq 0$ and introduce this technique briefly for the completeness and the later use in this paper. Consider

$$\Psi^t \mathbf{A} \Psi = \sum_{i,j=0}^{n-1} \psi_i \psi_j \gamma_{i+j} = \int \sum_{i,j=0}^{n-1} \psi_i x^i \psi_j x^j u_0(x) dx = \int \left(\sum_{k=0}^{n-1} \psi_k x^k \right)^2 u_0(x) dx.$$

Since the integrand $(\sum_{k=0}^{n-1} \psi_k x^k)^2 u_0(x)$ is nonnegative, we have $\Psi^t \mathbf{A} \Psi \geq 0$. Furthermore, $\Psi^t \mathbf{A} \Psi = 0$ if and only if $(\sum_{k=0}^{n-1} \psi_k x^k)^2 u_0(x) = 0$ for all $x \in \boldsymbol{R}$. For $\Psi \neq 0$, the polynomial $\sum_{k=0}^{n-1} \psi_k x^k$ has at most $n-1$ zeros and, therefore, $\Psi^t \mathbf{A} \Psi > 0$ if the support of the initial value u_0 consists of at least n points. Hence, we may conclude that the Hankel matrix $\mathbf{A} \equiv (\gamma_{i+j})_{i,j=0}^{n-1}$ is nonsingular. (The proof is originally done by Hamburger.)

To show that $g_n(x)$ has n-distinct real zeros, consider a linear functional S on the space of polynomials defined by

$$S(r) := \sum_{i=0}^{l} r_i \gamma_i = r_0 \gamma_0 + \cdots + r_l \gamma_l \quad \text{for} \quad r(x) = \sum_{i=0}^{l} r_i x^i.$$

Then, the same statements used for the positivity of $\Psi^t \mathbf{A} \Psi$ also show that

$$S(r^2) = S\left(\sum_{i,j=0}^{l} r_i r_j x^{i+j} \right) = \sum_{i,j=0}^{l} r_i r_j \gamma_{i+j} > 0.$$

Suppose that $r(x) \geq 0$. Then the degree of the polynomial $r(x)$ is even and there exist two polynomials p, q such that $r(x) = p^2(x) + q^2(x)$ (see [1, p. 2]). So $S(r) = S(p^2) + S(q^2) > 0$.

Since $\mathbf{A}$ is nonsingular, there exists an n-vector $\Psi = (\psi_0, \ldots, \psi_{n-1})$ uniquely so that $\mathbf{A}\Psi = \mathbf{r}_n$, i.e.,

$$\sum_{j=0}^{n-1} \psi_j \mathbf{r}_j = \mathbf{r}_n$$

or

$$\gamma_{n+k} - \sum_{j=0}^{n-1} \psi_j \gamma_{j+k} = 0, \quad k = 0, 1, \ldots, n-1. \tag{3.10}$$

Considering the polynomial $g_n(x)$ and the definition of the functional $S(r)$, we can easily check that (3.10) implies

$$S\left(g_n x^k\right) = 0, \qquad k = 0, 1, \ldots, n-1. \tag{3.11}$$

Suppose that $g_n(x)$ never changes its sign. Then $g_n(x) \geq 0$ and, hence, $S(g_n) > 0$, which contradicts (3.11) with $k = 0$. Suppose that $g_n(x)$ changes its sign at points $c_1, \ldots, c_l$ only. Then $g_n(x)(x - c_1) \cdots (x - c_l) \geq 0$ and $S(g_n(x)(x - c_1) \cdots (x - c_l)) > 0$. On the other hand, if $l < n$, then the linearity of the functional $S(r)$ together with (3.11) implies that $S(g_n(x)(x - c_1) \cdots (x - c_l)) = 0$. Hence, we obtain that $g_n(x)$ has n-distinct real roots, say, $c_1 < \cdots < c_n$.

Now we show that there exist ρ_i's that solve (3.5) in a unique way, i.e.,

$$\sum_{i=1}^{n} \rho_i c_i^l = \gamma_l, \quad l = 0, 1, \ldots, 2n-1. \tag{3.12}$$

Since c_i's are all different, there exists a unique solution for the Vandermonde (3.9); i.e., (3.12) is satisfied for all $0 \leq l < n$. Now we complete the proof using inductive arguments. Let $0 \leq k \leq n-1$. We will show that the identity in (3.12) holds for $l = n + k$ under the assumption that it holds for all $0 \leq l < n + k$. First, observe that, since c_i's are zero points of $x^k g_n(x), k \geq 0$,

$$c_i^{n+k} = \sum_{j=0}^{n-1} \psi_j c_i^{j+k} \qquad \text{for any} \quad 1 \leq i \leq n,\ k \geq 0.$$

Using the relations (3.10) and (3.12) for $l < n + k$, we obtain

$$\gamma_{n+k} = \sum_{j=0}^{n-1} \psi_j \gamma_{j+k} = \sum_{j=0}^{n-1} \psi_j \sum_{i=1}^{n} \rho_i c_i^{j+k} = \sum_{i=1}^{n} \rho_i \sum_{j=0}^{n-1} \psi_j c_i^{j+k} = \sum_{i=1}^{n} \rho_i c_i^{n+k}.$$

Hence, (3.12) holds by the induction.

In summary, the proof of the existence and the uniqueness of the solution to the problem (3.5) consists of three steps. The invertibility of the Hankel matrix $\mathbf{A}$ in (3.6) and the existence of n-distinct real roots c_i's of g_n are the first two. The latter depends on the positive definiteness of the matrix $\mathbf{A}$ which is easily proved for positive initial value $u_0(x)$. On the other hand, after obtaining c_i's, finding ρ_i's that satisfy (3.5) does not require the positivity. It depends only on the recursive structure of the problem. The following theorem is now clear from Theorem 2.4.

THEOREM 3.1. *Let $u(x,t)$ be the solution to the heat equation with initial value $u_0(x)$. If $u_0(x)$ is nonnegative (or nonpositive) and $x^{2n}u_0(x) \in L^1(\boldsymbol{R})$, then there exist $\rho_i, c_i, i = 1, \ldots, n$, such that, for $\phi_n(x,t) \equiv \sum_{i=1}^n \frac{\rho_i}{\sqrt{4\pi t}}\, e^{-(x-c_i)^2/(4t)}$,*

$$\lim_{t\to\infty} t^{\frac{2n+1}{2}-\frac{1}{2p}} \|u(t) - \phi_n(t)\|_p = \frac{\left\|\partial_\xi^{2n}\left(e^{-\frac{1}{4}\xi^2}\right)\right\|_p}{\sqrt{4\pi}} \left|\int E_{2n}(x)dx\right|, \tag{3.13}$$

where $1 \le p \le \infty$ and $E_{2n}(x) \in W^{2n,1}(R)$ is the $2n$th order antiderivative of $E_0(x) = u(x,0) - \phi_n(x,0)$. Furthermore, such a function $\phi_n(x,t)$ is unique.

Remark 3.2. The system (3.5) can be solved by commercial software such as Maple. However, since the problem is highly nonlinear, it takes a very long time even for small n. Therefore, even for the computational purpose, one needs to follow the steps of the proof to construct $\phi_n(x)$.

4. General initial value. In this section, we consider a general initial value which may change its sign. Then the existence and the uniqueness theory of the previous section is not applicable since it is for positive solutions only. In this section, we observe that the existence and uniqueness may fail for a general solution.

4.1. Approximation with a single heat kernel. For the case $n = 1$, the approximation $\phi_1(x,t) = \frac{\rho_1}{\sqrt{4\pi t}}\, e^{-(x-c_1)^2/(4t)}$ is obtained by solving

$$\rho_1 = \gamma_0, \qquad c_1\,\rho_1 = \gamma_1. \tag{4.1}$$

If $\gamma_0 \neq 0$, c_1 is uniquely decided by $c_1 = \gamma_1/\gamma_0$; i.e., c_1 is the *center of the mass* of the initial mass distribution u_0. The convergence order in Theorem 2.4 is written as

$$\lim_{t\to\infty} t^{\left(\frac{3}{2}-\frac{1}{2p}\right)} \|u(t) - \phi_1(t)\|_p = \frac{\left\|\partial_\xi^{2}\left(e^{-\frac{1}{4}\xi^2}\right)\right\|_p}{\sqrt{4\pi}} \left|\int_{-\infty}^{\infty} E_2(x)dx\right|, \quad 1 \le p \le \infty,$$

where $E_2(x)$ is the second order antiderivative of the initial error $E_0(x) := u_0(x) - \rho_1\delta_{c_1}(x)$ given by (2.4), i.e., $E_2(x) = \int_{-\infty}^x \int_{-\infty}^y (u_0(z) - \rho_1\delta_{c_1}(z))dzdy$.

Now consider the singular case $\gamma_0 = 0$. Then the approximation is simply $\phi_1 \equiv 0$. If $\gamma_1 = 0$, then the equation for the first moment is satisfied for any $c_1 \in \boldsymbol{R}$ and we obtain the above convergence order which is equivalent to the decay rate $u(x,t)$. If $\gamma_1 \neq 0$, (4.1) has no solution and we do not obtain a single heat kernel approximation ϕ_1 with the desirable convergence order $O(t^{(\frac{1}{2p}-\frac{3}{2})})$ for t large.

4.2. Approximation with two heat kernels. The double heat kernel solution $\phi_2(x,t) = \sum_{i=1}^2 \frac{\rho_i}{\sqrt{4\pi t}}\, e^{-(x-c_i)^2/(4t)}$ that approximates the solution $u(x,t)$ is obtained by solving

$$\begin{aligned} \rho_1 + \rho_2 &= \gamma_0, & \rho_1 c_1 + \rho_2 c_2 &= \gamma_1, \\ \rho_1 c_1^2 + \rho_2 c_2^2 &= \gamma_2, & \rho_1 c_1^3 + \rho_2 c_2^3 &= \gamma_3. \end{aligned} \tag{4.2}$$

We may simplify the equation by eliminating ρ_i's and obtain two equations of the form $\mathbf{A}\Psi = \mathbf{r}_2$, i.e.,

$$\begin{pmatrix} \gamma_0 & \gamma_1 \\ \gamma_1 & \gamma_2 \end{pmatrix} \begin{pmatrix} \psi_0 \\ \psi_1 \end{pmatrix} = \begin{pmatrix} \gamma_2 \\ \gamma_3 \end{pmatrix}.$$

First, we need to check the invertibility of the Hankel matrix. Its determinant is the variance of the initial value u_0 if it is a probability distribution, i.e.,

$$|\mathbf{A}| = \gamma_0\gamma_2 - \gamma_1^2.$$

If $|\mathbf{A}| \neq 0$, ψ_i's can be solved using Cramer's rule, and c_i's are zeros of a quadratic function

$$g_2(x) = x^2 + \frac{\gamma_1\gamma_2 - \gamma_0\gamma_3}{|\mathbf{A}|}x + \frac{\gamma_1\gamma_3 - \gamma_2^2}{|\mathbf{A}|}.$$

Hence, the centers c_1, c_2 are given by

$$c_{1,2} = \frac{(\gamma_0\gamma_3 - \gamma_1\gamma_2) \pm \sqrt{D}}{2|\mathbf{A}|}, \qquad c_1 < c_2, \tag{4.3}$$

under two assumptions

$$|\mathbf{A}| = \gamma_0\gamma_2 - \gamma_1^2 \neq 0,\ D := (\gamma_1\gamma_2 - \gamma_0\gamma_3)^2 - 4\left(\gamma_0\gamma_2 - \gamma_1^2\right)\left(\gamma_1\gamma_3 - \gamma_2^2\right) > 0. \tag{4.4}$$

After obtaining c_i's, the problem (3.5) is easily solved and gives

$$\rho_1 = \frac{\gamma_0 c_2 - \gamma_1}{c_2 - c_1}, \qquad \rho_2 = \frac{\gamma_0 c_1 - \gamma_1}{c_1 - c_2}. \tag{4.5}$$

From Theorem 2.4 we may conclude that if $D > 0$ and $|\mathbf{A}| \neq 0$, then

$$\lim_{t\to\infty} t^{\left(\frac{5}{2}-\frac{1}{2p}\right)}\|u(t) - \phi_2(t)\|_p \leq \frac{\left\|\partial_\xi^4\left(e^{-\frac{1}{4}\xi^2}\right)\right\|_p}{\sqrt{4\pi}}\left|\int_{-\infty}^{\infty} E_4(x)dx\right|,\ 1 \leq p \leq \infty, \tag{4.6}$$

where $E_4(x)$ is the fourth order antiderivative of the initial error $E_0(x) := u_0(x) - \sum_{i=1}^2 \rho_i\delta_{c_i}(x)$ given by (2.4).

Example 4.1. Consider an initial value

$$U_l(x) = \begin{cases} -1, & -2l - 0.5 < x < -l - 0.5,\ \ l + 0.5 < x < 2l + 0.5, \\ 1, & -l - 0.5 \leq x \leq l + 0.5, \\ 0, & \text{otherwise}, \end{cases} \tag{4.7}$$

where $l > 0$. Let $\gamma_{k,l}$ be the kth moments of the function $U_l(x)$, i.e.,

$$\gamma_{k,l} := \int x^k U_l(x)dx, \quad k = 0, 1, \ldots.$$

Then $\gamma_{0,l} = 1$ for all $l > 0$ and, since U_l is an even function, $\gamma_{k,l} = 0$ for $k = 1, 3, 5, \ldots$. Hence, $|A|$ and D in (4.4) are given by

$$|A| = \gamma_{2,l}, \qquad D = 4(\gamma_{2,l})^3.$$

One may easily check that $\gamma_{2,l} = 0$, if and only if

$$l = l_2 := 0.5\left(\sqrt[3]{2} - 1\right) / \left(2 - \sqrt[3]{2}\right),$$

and $D = 4(\gamma_{2,l})^3 > 0$, if and only if $l < l_2$. Hence, the moment problem (4.2) with the initial value $U_l(x)$ is solvable only for $l < l_2$. This example says that, even if the Hankel matrix is nonsingular, $\phi_2(x,t)$ that satisfies convergence order in (4.6) may not exist.

4.3. Approximation with three heat kernels. The derivation of (3.6) from (3.5) is not clear without some calculations. In the following, such a derivation is given for an example. For the case $n = 3$ the system (3.5) reads

$$\begin{array}{l} \rho_1 \quad + \rho_2 \quad + \rho_3 \quad = \gamma_0, \\ \rho_1 c_1 + \rho_2 c_2 + \rho_3 c_3 = \gamma_1, \\ \rho_1 c_1^2 + \rho_2 c_2^2 + \rho_3 c_3^2 = \gamma_2, \\ \rho_1 c_1^3 + \rho_2 c_2^3 + \rho_3 c_3^3 = \gamma_3, \\ \rho_1 c_1^4 + \rho_2 c_2^4 + \rho_3 c_3^4 = \gamma_4, \\ \rho_1 c_1^5 + \rho_2 c_2^5 + \rho_3 c_3^5 = \gamma_5. \end{array} \tag{4.8}$$

Multiply c_1 to the kth equation and subtract $(k+1)$th one from it for $k = 1, \ldots, 5$ and obtain five equations without ρ_1, i.e.,

$$\begin{array}{l} \rho_2(c_1 - c_2) \quad + \rho_3(c_1 - c_3) \quad = \gamma_0 c_1 - \gamma_1, \\ \rho_2(c_1 - c_2)c_2 + \rho_3(c_1 - c_3)c_3 = \gamma_1 c_1 - \gamma_2, \\ \rho_2(c_1 - c_2)c_2^2 + \rho_3(c_1 - c_3)c_3^2 = \gamma_2 c_1 - \gamma_3, \\ \rho_2(c_1 - c_2)c_2^3 + \rho_3(c_1 - c_3)c_3^3 = \gamma_3 c_1 - \gamma_4, \\ \rho_2(c_1 - c_2)c_2^4 + \rho_3(c_1 - c_3)c_3^4 = \gamma_4 c_1 - \gamma_5. \end{array}$$

Do the similar process two more times and obtain three equations without ρ_i's:

$$\begin{array}{l} 0 = \gamma_0 c_1 c_2 c_3 - \gamma_1(c_1 c_2 + c_2 c_3 + c_3 c_1) + \gamma_2(c_1 + c_2 + c_3) - \gamma_3, \\ 0 = \gamma_1 c_1 c_2 c_3 - \gamma_2(c_1 c_2 + c_2 c_3 + c_3 c_1) + \gamma_3(c_1 + c_2 + c_3) - \gamma_4, \\ 0 = \gamma_2 c_1 c_2 c_3 - \gamma_3(c_1 c_2 + c_2 c_3 + c_3 c_1) + \gamma_4(c_1 + c_2 + c_3) - \gamma_5, \end{array}$$

which are identical to (3.6)–(3.7) with $n = 3$, i.e.,

$$\begin{pmatrix} \gamma_0 & \gamma_1 & \gamma_2 \\ \gamma_1 & \gamma_2 & \gamma_3 \\ \gamma_2 & \gamma_3 & \gamma_4 \end{pmatrix} \begin{pmatrix} \psi_0 \\ \psi_1 \\ \psi_2 \end{pmatrix} = \begin{pmatrix} \gamma_3 \\ \gamma_4 \\ \gamma_5 \end{pmatrix},$$

where $\psi_0 = c_1 c_2 c_3, \psi_1 = -(c_1 c_2 + c_2 c_3 + c_3 c_1)$, and $\psi_2 = c_1 + c_2 + c_3$. The derivation is done for the case $n = 3$.

The determinant of the 3×3 Hankel matrix is given by

$$|\mathbf{A}| = \gamma_0\gamma_2\gamma_4 + 2\gamma_1\gamma_2\gamma_3 - \gamma_2^3 - \gamma_0\gamma_3^2 - \gamma_1^2\gamma_4.$$

If $|\mathbf{A}| \neq 0$, then ψ_i are given by Cramer's rule:

$$\begin{array}{l} \psi_0 = (2\gamma_3\gamma_2\gamma_4 + \gamma_3\gamma_1\gamma_5 - \gamma_3^3 - \gamma_2^2\gamma_5 - \gamma_4^2\gamma_1)/|\mathbf{A}|, \\ \psi_1 = (\gamma_2\gamma_5\gamma_1 + \gamma_0\gamma_4^2 + \gamma_3^2\gamma_2 - \gamma_3\gamma_1\gamma_4 - \gamma_4\gamma_2^2 - \gamma_0\gamma_3\gamma_5)/|\mathbf{A}|, \\ \psi_2 = (\gamma_0\gamma_2\gamma_5 + \gamma_3^2\gamma_1 + \gamma_2\gamma_4\gamma_1 - \gamma_0\gamma_3\gamma_4 - \gamma_3\gamma_2^2 - \gamma_1^2\gamma_5)/|\mathbf{A}|. \end{array}$$

The points c_i's are zeros of third order polynomial

$$g_3(x) = x^3 - \psi_2 x^2 - \psi_1 x - \psi_0. \tag{4.9}$$

Hence, the solvability of the problem (4.8) is equivalent to the existence of three distinct real roots $c_1 < c_2 < c_3$ of (4.9). The convergence order in Theorem 2.4 gives the asymptotic convergence order:

$$\lim_{t\to\infty} t^{\left(\frac{7}{2}-\frac{1}{2p}\right)} \|u(t) - \phi_3(t)\|_p \le \frac{\left\|\partial_\xi^6 \left(e^{-\frac{1}{4}\xi^2}\right)\right\|_p}{\sqrt{4\pi}} \left| \int_{-\infty}^{\infty} E_6(x)dx \right|, \quad 1 \le p \le \infty, \tag{4.10}$$

where $E_6(x)$ is the sixth order antiderivative of the initial error $E_0(x) := u_0(x) - \sum_{i=1}^3 \rho_i \delta_{c_i}(x)$ given by (2.4).

Consider the initial value given in Example 4.1. Since $\gamma_{1,l} = \gamma_{3,l} = \gamma_{5,l} = 0$ and $\gamma_{0,l} = 1$, we obtain

$$|A| = \gamma_{2,l}\left(\gamma_{4,l} - \gamma_{2,l}^2\right), \quad \psi_1 = \gamma_{4,l}/\gamma_{2,l}, \quad \psi_0 = \psi_2 = 0.$$

Hence, if

$$|\mathbf{A}| \neq 0 \quad \text{and} \quad \psi_1 > 0,$$

then $g_3(x)$ has three distinct real roots

$$c_1 = -\sqrt{\psi_1}, \quad c_2 = 0, \quad c_3 = \sqrt{\psi_1}.$$

One may show that $\gamma_{4,l} > 0$ if and only if $0 < l < l_4 := 0.5(\sqrt[5]{2} - 1)/(2 - \sqrt[5]{2})$. Therefore, if $l_4 < l < l_2$, then $\psi_1 < 0$ and the existence of $\phi_3(x,t)$ satisfying (4.10) is not guaranteed. This example shows that the solvability of (3.5) is not obvious for sign-changing initial values.

5. Approximation for sign-changing solutions. Now consider a general sign-changing initial value. First, consider the case that the initial value $u_0(x)$ decays for $|x|$ large with the order that the Gaussian has, i.e.,

$$u_0(x) = O\left(e^{-|x|^2}\right) \quad \text{as} \quad |x| \to \infty. \tag{5.1}$$

Then there exists $M > 0$ such that $v_0(x) := u_0(x) + \frac{M}{\sqrt{4\pi}} e^{-x^2/4} \ge 0$, and we may apply the theory in section 3 to the nonnegative function v_0. Let ρ_i's and c_i's be the solutions of the moment problem with initial value $v_0(x)$. Then the solution $u(x,t)$ can be approximated by

$$u(x,t) \sim \sum_{i=1}^n \frac{\rho_i}{\sqrt{4\pi t}} e^{-(x-c_i)^2/(4t)} - \frac{M}{\sqrt{4\pi(t+1)}} e^{-x^2/4(t+1)}. \tag{5.2}$$

Since the auxiliary part of the approximation is the exact solution with the extra initial value added to $u_0(x)$, the convergence order of this approximation is the same as the one in Theorem 3.1. This example shows that we may obtain the same convergence order for general sign-changing solutions by simply adding an extra term.

On the other hand, if the grid points are preassigned, say, $c_i = \bar{c}_i$, then we have the freedom in choosing the weights ρ_i's only. These ρ_i's are simply obtained by solving

the first n equations in (3.5), where the corresponding matrix is the Vandermonde matrix, i.e.,

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ \bar{c}_1 & \bar{c}_2 & \cdots & \bar{c}_n \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \bar{c}_1^{n-1} & \bar{c}_2^{n-1} & \cdots & \bar{c}_n^{n-1} \end{pmatrix} \begin{pmatrix} \rho_1 \\ \rho_2 \\ \cdot \\ \cdot \\ \cdot \\ \rho_n \end{pmatrix} = \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \cdot \\ \cdot \\ \cdot \\ \gamma_{n-1} \end{pmatrix}. \tag{5.3}$$

The Vandermonde determinant $\prod_{1\le i<j\le n}(\bar{c}_j - \bar{c}_i)$ is not zero if $\bar{c}_i$ are all different and, hence, (5.3) is solvable. Now construct a different kind of approximation:

$$\eta_n(x,t) := \sum_{i=1}^{n} \frac{\rho_i}{\sqrt{4\pi t}} e^{-(x-\bar{c}_i)^2/(4t)}. \tag{5.4}$$

Then $\eta_n(\cdot,t)$ converges to $u(\cdot,t)$ with the order

$$\|u(t) - \eta_n(t)\|_p = O\left(t^{\left(\frac{1}{2p} - \frac{n+1}{2}\right)}\right) \qquad \text{as} \quad t \to \infty,$$

since $\lim_{t\to 0}(\eta_n(x,t) - u_0(x))$ has zero moments up to $(n-1)$th order.

6. Convergence as $n \to \infty$ with fixed $t > 0$. In this section, we discuss the convergence of the approximation $\phi_n(x,t)$ to the solution $u(x,t)$ as $n \to \infty$ with a fixed $t > 0$. An interesting behavior of the approximation $\phi_n(x,t)$ that one may observe numerically is a geometric convergence order such as

$$\beta_n(t) := \frac{\|u(t) - \phi_n(t)\|_\infty}{\|u(t) - \phi_{n+1}(t)\|_\infty} \to 1 + 4\frac{t}{v}\left(\equiv \beta\left(\frac{t}{v}\right)\right) \quad \text{as} \quad n \to \infty, \tag{6.1}$$

where $v > 0$ depends on the initial value $u_0(x)$. (We do not have a proof of it. Hence, the statements here are rather conjectures.) This convergence order implies that the error decays to zero very fast as $n \to \infty$ for any fixed time $t > 0$. This convergence order is somewhat extreme. For example, if $t > v/4$, then the approximation error is reduced into half whenever just a single heat kernel is added.

Set the approximation error as

$$e_n(x,t) = u(x,t) - \phi_n(x,t).$$

Consider a sequence of functions

$$E_k^n(x) = \int_{-\infty}^{x} E_{k-1}^n(y)dy, \qquad k = 1, 2, \ldots, 2n,$$

where

$$E_0^n(x) := e_n(x,0) = u_0(x) - \sum_{i=1}^{n} \rho_i \delta(x - c_i).$$

Notice that the upper index n is to denote that E_k^n is related to the approximation $\phi_n(x,t)$ and the lower index k is to indicate that E_k^n is the kth order antiderivative of the initial approximation error $E_0^n(x)$. Then, from (2.6), one obtains

$$\left(\sqrt{t}\right)^{2n+1} \|e_n(t)\|_\infty = \frac{C_{2n}}{\sqrt{4\pi}} \sup_\xi \left| \int \partial_\xi^{2n}\left(e^{-\zeta^2/4}\right) \sqrt{t} E_{2n}^n \frac{\sqrt{t}(\xi - \zeta)}{C_{2n}} d\zeta \right|, \tag{6.2}$$

where $C_{2n} := \int_{-\infty}^{\infty} E_{2n}^n(x)dx < \infty$.

An interesting observation is that, for n large, $E_{2n}^n(x)$ has a Gaussian-like structure. The following has been observed numerically.

CONJECTURE 6.1. *Suppose that the initial value $u_0(x)$ is nonnegative and has finite moments up to any order. Then there exist $c \in \boldsymbol{R}$ and $v > 0$ such that*

$$\left\| \frac{1}{\sqrt{v\pi}} e^{\frac{-(x-c)^2}{v}} - \frac{1}{C_{2n}} E_{2n}^n(x) \right\|_\infty \to 0 \quad as \quad n \to \infty, \tag{6.3}$$

where $C_{2n} := \int_{-\infty}^{\infty} E_{2n}^n(x)dx < \infty$. Furthermore,

$$\frac{C_{2n}}{C_{2(n+1)}} \frac{\left\| D_x^{2n} e^{\frac{-x^2}{v}} \right\|_\infty}{\left\| D_x^{2n+2} e^{\frac{-x^2}{v}} \right\|_\infty} \to 1 \quad as \quad n \to \infty. \tag{6.4}$$

The $(2n-1)$th order derivative of $E_{2n}^n(x)$ is $E_1^n(x)$ which is, at most, of order $O(1/n)$, which does not make any difference in the geometric convergence order such as (6.1). Hence, we may treat it as of order $O(1)$. Note that C_{2n} is obtained after integrating $E_0^n(x)$ $2n$ times and, hence, its order should be the reciprocal of the order of $\|D_x^{2n}(e^{\frac{-x^2}{v}})\|_\infty$, which is the $2n$th derivative of the Gaussian. Hence, (6.4) is a natural conclusion if (6.3) is assumed. Furthermore, for $\xi = x/\sqrt{v}$,

$$\left\| D_x^{2n} \left(e^{\frac{-x^2}{v}} \right) \right\|_\infty = \left\| D_\xi^{2n} \left(e^{-\xi^2} \right) (\xi_x)^{2n} \right\|_\infty = \frac{1}{v^n} \left\| D_\xi^{2n} \left(e^{-\xi^2} \right) \right\|_\infty .$$

Under Conjecture 6.1, the right-hand side of (6.2) can be approximated using $A(n, v/t)$ given by

$$\begin{aligned} \sup_x \left| \int D_y^{2n} \left(e^{-y^2/4} \right) \sqrt{t} E_{2n}^n \frac{\sqrt{t}(x-y)}{C_{2n}} dy \right| & \\ \cong \frac{\sqrt{t}}{\sqrt{v\pi}} \sup_x \left| \int D_y^{2n} \left(e^{-y^2/4} \right) e^{-\frac{(x-y-c/\sqrt{t})^2}{v/t}} dy \right| & \\ = \frac{\sqrt{t}}{\sqrt{v\pi}} \left| \int D_y^{2n} \left(e^{-y^2/4} \right) e^{-\frac{y^2}{v/t}} dy \right| =: A(n, v/t). & \end{aligned}$$

Notice that due to the symmetry of $D_x^{2n}(e^{-x^2/4})$ and $e^{-\frac{x^2}{v/t}}$ the supremum of the second line is obtained at $x - c/\sqrt{t} = 0$. Then we obtain from the relations (6.2) and (6.4) that

$$t^{-1} \frac{\|e_n(t)\|_\infty}{\|e_{n+1}(t)\|_\infty} \cong \frac{C_{2n}}{C_{2(n+1)}} \frac{A(n, v/t)}{A(n+1, v/t)} \cong \frac{v^n}{v^{n+1}} \frac{\left\| D_x^{2n+2} \left(e^{-x^2} \right) \right\|_\infty}{\left\| D_x^{2n} \left(e^{-x^2} \right) \right\|_\infty} \frac{A(n, v/t)}{A(n+1, v/t)}.$$

One can easily check that

$$\frac{\left\| D_x^{2n+2} \left(e^{-x^2} \right) \right\|_\infty}{\left\| D_x^{2n} \left(e^{-x^2} \right) \right\|_\infty} = 4n + 2, \quad \frac{A(n, v/t)}{A(n+1, v/t)} = \frac{4 + v/t}{4n+2},$$

using a mathematical software such as Maple or by hand. Therefore, we obtain the convergence order in (6.1), i.e.,

$$\frac{\|e_n(t)\|_\infty}{\|e_{n+1}(t)\|_\infty} \cong t \frac{1}{v} (4n+2) \frac{4 + v/t}{4n+2} = 1 + 4\frac{t}{v} \quad \text{for } n \text{ large.}$$

Notice that $c \in \boldsymbol{R}$ in (6.3) does not make any difference in the convergence order. The factor that decides the geometric convergence rate is the variance factor v of the limit function $\frac{1}{\sqrt{v\pi}}e^{-x^2/v}$. It seems that the variance factor v depends on the initial value $u_0(x)$, and another discussion about it will be included in section 7.2.

Remark 6.2. For $n > 0$ small, $E^n_{2n}(x)/C_{2n}$ is not close enough to the Gaussian and the arguments above do not apply. Then it is natural to ask how large n should be. The answer depends on the initial value. Clearly, if $u_0(x)$ itself is like a Gaussian, then such an $n > 0$ can be relatively small. In other cases, the corresponding $n > 0$ could be larger.

7. Numerical examples. In this section, we test the convergence orders numerically for $t > 0$ large and for $n > 0$ large. These tests confirm the convergence orders obtained in the previous sections. This section consists of four subsections. The first two are for $t \to \infty$ and for $n \to \infty$ limits of the approximation $\phi_n(x,t)$. In the third one, we test the behavior of the alternative approach $\psi_{2n}(x,t)$ as $n \to \infty$. In the last one, we do numerical tests for Conjecture 6.1.

There are two difficulties in observing the theoretical convergence order for $t > 0$ large. First, the convergence rate for small time $0 < t \ll 1$ is lower than the theoretical one for $t > 0$ large. So we need to wait a certain amount of time to observe the theoretical convergence order. On the other hand, since the convergence order is so high, the approximation error at the right moment can be as small as of order 10^{-36} or 10^{-64} (see Tables 7.1 and 7.2). So we should employ enough precisions in the computation to obtain meaningful numerical results.

The second difficulty, which is more restrictive, is in computing the solution $u(x,t)$. To compute the decay order of $\|u(x,t) - \phi_n(x,t)\|_\infty$ accurately, we should obtain the exact value $u(x,t)$ or compute it with a smaller error than the actual approximation error. However, it seems impossible to do the integration in (1.1) numerically with such a small error. (In this sense, one may say that the approximation $\phi_n(x,t)$ is more exact than the exact formula in (1.1).) To avoid such a difficulty, we consider the following two examples with explicit solutions. In the following numerical tests we employ these examples.

Example 7.1 (example with a single hump). Consider the solution of

$$u_t = u_{xx}, \quad u(x,0) = K(x,t_0), \quad x \in \boldsymbol{R},\ t > 0, \tag{7.1}$$

where $K(x,t)$ is the heat kernel

$$K(x,t) = \frac{1}{\sqrt{4\pi t}}e^{-x^2/4t}.$$

Then the exact solution is simply $u(x,t) = K(x,t+t_0)$ and the variance of the initial value is $var = 2t_0$. This rather simple example illustrates certain convergence behavior very clearly.

Example 7.2 (example with double humps). Consider the solution of

$$u_t = u_{xx}, \quad u(x,0) = \frac{1}{2}[K(x+1,t_0) + K(x-1,t_0)], \quad x \in \boldsymbol{R},\ t > 0. \tag{7.2}$$

Then the solution is simply $u(x,t) = \frac{1}{2}[K(x+1,t+t_0) + K(x-1,t+t_0)]$ and the variance of the initial value is $var = 1 + 2t_0$.

TABLE 7.1

The error $e_n(x,t) = u(x,t) - \phi_n(x,t)$ and the convergence order α_n in (7.5) have been computed for Examples 7.1 and 7.2 with $n = 4, 8$ and $t_0 = 1$. We observe that $\alpha_n(t) \to -(n + \frac{1}{2})$ as $t \to \infty$. (The norms in this and the following tables are L^∞-norms.)

	Example 7.1				Example 7.2			
t	$\|e_4(t)\|$	$\alpha_4(t)$	$\|e_8(t)\|$	$\alpha_8(t)$	$\|e_4(t)\|$	$\alpha_4(t)$	$\|e_8(t)\|$	$\alpha_8(t)$
0.1	2.17e-01,	0.7	1.13e-01,	1.0	2.24e-01,	0.8	1.30e-01,	0.9
0.2	1.13e-01,	0.9	2.98e-02,	1.9	1.33e-01,	0.8	4.57e-02,	1.5
0.4	3.48e-02,	1.7	3.34e-03,	3.2	5.30e-02,	1.3	7.27e-03,	2.7
0.8	6.24e-03,	2.5	1.38e-04,	4.6	1.21e-02,	2.1	4.27e-04,	4.1
1.6	6.81e-04,	3.2	2.21e-06,	6.0	1.60e-03,	2.9	9.09e-06,	5.6
3.2	5.12e-05,	3.7	1.72e-08,	7.0	1.37e-04,	3.5	8.57e-08,	6.7
6.4	3.03e-06,	4.1	8.40e-11,	7.7	8.79e-06,	4.0	4.68e-10,	7.5
12.8	1.56e-07,	4.3	3.13e-13,	8.1	4.73e-07,	4.2	1.86e-12,	8.0
25.6	7.48e-09,	4.4	1.01e-15,	8.3	2.31e-08,	4.4	6.18e-15,	8.2
51.2	3.44e-10,	4.4	3.01e-18,	8.4	1.08e-09,	4.4	1.88e-17,	8.4
102.4	1.55e-11,	4.5	8.66e-21,	8.4	4.89e-11,	4.5	5.44e-20,	8.4
204.8	6.93e-13,	4.5	2.44e-23,	8.5	2.19e-12,	4.5	1.54e-22,	8.5

TABLE 7.2

The error $e_n(x,t) = u(x,t) - \phi_n(x,t)$ and the geometric convergence rate $\beta_n(t)$ in (6.1) have been computed for Examples 7.1 and 7.2 with $t = 1, 10$ and $t_0 = 1$. The ratio $\beta_n(t)$ converges to the limit in (6.1) quickly with $v = 2$ for Example 1 and slowly for Example 2.

	Example 7.1				Example 7.2			
n	$\|e_n(1)\|$	$\beta_n(1)$	$\|e_n(10)\|$	$\beta_n(10)$	$\|e_n(1)\|$	$\beta_n(1)$	$\|e_n(10)\|$	$\beta_n(10)$
2	2.8e-02	2.91	2.0e-04	20.84	4.28e-02	2.48	3.83e-04	15.83
3	9.6e-03	2.96	9.5e-06	20.92	1.72e-02	2.49	2.32e-05	16.53
4	3.2e-03	2.98	4.5e-07	20.95	6.7e-03	2.57	1.36e-06	17.06
7	1.2e-04	2.99	4.9e-11	20.98	3.67e-04	2.66	2.47e-10	17.89
10	4.5e-06	3.0	5.3e-15	20.99	1.89e-05	2.71	4.09e-14	18.35
13	1.7e-07	3.0	5.8e-19	21.0	9.35e-07	2.74	6.45e-18	18.65
16	6.1e-09	3.0	6.2e-23	21.	4.45e-08	2.76	9.65e-22	18.86
19	2.3e-10	3.0	6.7e-27	21.0	2.08e-09	2.78	1.41e-25	19.02
22	8.4e-12	3.0	7.3e-31	21.0	9.65e-11	2.79	2.02e-29	19.15
25	3.1e-13	3.0	7.9e-35	21.0	4.36e-12	2.8	2.85e-33	19.26
46	2.97e-23	3.0	1.35e-62	21.0	1.38e-21	2.85	2.25e-60	19.70
49	1.1e-24	3.0	1.46e-66	21.0	5.95e-21	2.86	2.94e-64	19.74

7.1. Numerical tests for the long time asymptotics. The approximation

$$\phi_n(x,t) \equiv \sum_{i=1}^{n} \frac{\rho_i}{\sqrt{4\pi t}} e^{\frac{-(x-c_i)^2}{4t}} \tag{7.3}$$

constructed in section 3 converges to the exact solution $u(x,t)$ with order

$$\|u(t) - \phi_n(t)\|_\infty = O\left(t^{-\frac{2n+1}{2}}\right) \quad \text{as} \quad t \to \infty. \tag{7.4}$$

In Table 7.1 the error $e_n(x,t) = u(x,t) - \phi_n(x,t)$ and the convergence order α_n have been computed for $n = 4, 8$ as doubling the time from $t = 0.1$ to $t = 204.8$. The convergence order of the approximation has been measured by computing

$$\alpha_n(t) \sim \frac{\ln(\|e_n(t/2)\|_\infty / \|e_n(t)\|_\infty)}{\ln(1/2)}. \tag{7.5}$$

(Note that we measure the error in L^∞-norm in the following numerical examples and denote it by $\|\cdot\|$ in the tables to get it fitted in the tables.) From Table 7.1, one

clearly observes that the convergence order $\alpha_n(t)$ approaches the optimal convergence order in (7.4) as $t \to \infty$. Notice that these numerical tests for $t \to \infty$ limits show similar patterns of the convergence for both Examples 7.1 and 7.2.

7.2. Numerical tests for $n \to \infty$ limits. Now we are going to check the convergence order for n large with a fixed $t > 0$. Consider the ratio

$$\beta_n(t) := \|e_{n-1}(t)\|_\infty \Big/ \|e_n(t)\|_\infty.$$

(The ratio $r = a_n/a_{n-1}$ is usually considered for a geometric sequence. Here we consider its reciprocal for easier comparison.) In Table 7.2, the error $\|e_n(t)\|_\infty$ and this ratio are computed for Examples 7.1 and 7.2 at two instances $t = 1$ and $t = 10$ with increasing n from $n = 2$ to $n = 25$. One can clearly observe a certain geometric convergence order as in (6.1). In both examples, we can clearly see that $10(\beta_n(1) - 1) \sim (\beta_n(10) - 1)$, which indicates that the corresponding constant $v > 0$ in (6.1) which decides the geometric convergence ratio does not depend on the time $t > 0$.

From the test for Example 7.1, one can clearly see that $\beta_n(1) \to 3$ and $\beta_n(10) \to 21$ as $n \to \infty$. In both cases, the corresponding v is $v = 2$ which is the variance of the initial value. The convergence pattern for Example 7.2 is different. First, the convergence speed of the ratio $\beta_n(t)$ is slow. It seems due to the complexity of the structure of the initial value. At the moment $n = 49$ the index v corresponding to the geometric convergence rate $\beta = 19.74$ is $v = 2.15$ and seems still decreasing. This value is already smaller than the variance of the initial value which is 3 and looks likely to converge to $v = 2$. It seems that the factor that decides the geometric convergence rate is not the variance but the tail of the initial value for $|x|$ large.

7.3. Approximation using derivatives of the Gaussian. We may write $\psi_{2n}(x,t)$ in (1.5) as

$$\psi_{2n}(x,t) = \sum_{i=0}^{2n-1} \frac{-\gamma_i}{2(i!)} \left(\frac{-1}{2\sqrt{t}}\right)^{n+1} H_i\left(\frac{x}{2\sqrt{t}}\right) e^{-x^2/4t}, \tag{7.6}$$

where $H_i(x)$ is the Hermite polynomial of degree i. In Table 7.3, the approximation error and the geometric convergence ratio β_n are given for Example 7.1. To make the relation between the initial value and the convergence ratio more clear, we set

$$\beta_{2n}(t, t_0) := \|e_{2n-2}(t)\|_\infty \Big/ \|e_{2n}(t)\|_\infty, \tag{7.7}$$

where $e_{2n}(x,t) = u(x,t) - \psi_{2n}(x,t)$ and $u_0(x) = K(x,t_0)$ with $t_0 = 10$.

One may observe that $\beta_{2n}(1,10) \to 0.1$, $\beta_{2n}(10,10) \to 1$, and $\beta_{2n}(100,10) \to 10$ as $n \to \infty$. This observation leads us to a conjecture

$$\lim_{n\to\infty} \frac{\|u(t) - \psi_{2n}(t)\|_\infty}{\|u(t) - \psi_{2n+2}(t)\|_\infty} = \frac{t}{t_0}\ , \tag{7.8}$$

which indicates that the approximation error increases geometrically if $t < t_0$ and, hence, $\psi_{2n}(x,t)$ is meaningful for $t > t_0$ only. We may consider t_0 as the age of the initial value since $u_0(x) = K(x,t_0)$. This t_0 seems to be related to the time-shift t_* in [22]. Generally, one may call t_0 the age of a general initial value u_0 of the heat equation if $\lim_{n\to\infty} \frac{\|u(t_0) - \psi_{2n}(t_0)\|_\infty}{\|u(t_0) - \psi_{2n+2}(t_0)\|_\infty} = 1$.

TABLE 7.3

The error $e_n(x,t) = u(x,t) - \psi_{2n}(x,t)$ and the geometric convergence rate $\beta_n(t,t_0)$ in (7.7) have been computed numerically for Example 7.1 with $t_0 = 10$ and $t = 1, 10, 100$. We may observe the convergence rate in (7.8).

$2n$	$\|e_{2n}(1)\|$	$\beta_{2n}(1,10)$	$\|e_{2n}(10)\|$	$\beta_{2n}(10,10)$	$\|e_{2n}(100)\|$	$\beta_{2n}(100,10)$
4	1.21e+00	0.1623821	1.85e-02	1.4142136	9.77e-05	13.4400610
6	9.37e+00	0.1295695	1.50e-02	1.2335625	8.11e-06	12.0475285
8	7.88e+01	0.1188625	1.29e-02	1.1610328	7.08e-07	11.4555359
14	5.74e+04	0.1089029	9.66e-03	1.0825322	5.40e-10	10.7781946
20	4.73e+07	0.1058095	8.05e-03	1.0553099	4.54e-13	10.5307485
26	4.12e+10	0.1043095	7.04e-03	1.0415615	3.99e-16	10.4026370
32	3.69e+13	0.1034247	6.34e-03	1.0332791	3.60e-19	10.3243276
38	3.38e+16	0.1028412	5.81e-03	1.0277462	3.31e-22	10.2715121
44	3.13e+19	0.1024275	5.39e-03	1.0237896	3.07e-25	10.2334861
50	2.93e+22	0.1021190	5.06e-03	1.0208199	2.88e-28	10.2048012

TABLE 7.4

We may observe conjectures in (6.3) and (6.4) numerically. In this table, those conjectures are tested using Examples 7.1 and 7.2 with $t_0 = 1$ and $v = 2t_0$.

	$\left\| \frac{1}{\sqrt{2t_0\pi}} e^{\frac{-x^2}{2t_0}} - \frac{1}{C_{2n}} E_{2n}^n(x) \right\|$		$\frac{C_{2n-2}}{C_{2n}} \frac{\left\| D_x^{2n-2} e^{-x^2/2t_0} \right\|}{\left\| D_x^{2n} e^{-x^2/2t_0} \right\|}$	
n	Example 7.1	Example 7.2	Example 7.1	Example 7.2
2	2.233e-02	4.378e-02	1.0	0.7500000
3	2.070e-02	3.675e-02	1.0	0.7826087
4	1.631e-02	3.547e-02	1.0	0.8070175
5	1.387e-02	3.353e-02	1.0	0.8237512
6	1.207e-02	3.194e-02	1.0	0.8369731
7	1.070e-02	3.054e-02	1.0	0.8474264
8	9.675e-03	2.932e-02	1.0	0.8561101
9	8.743e-03	2.825e-02	1.0	0.8634099
10	8.016e-03	2.729e-02	1.0	0.8696921

7.4. Numerical test for Conjecture 6.1. The geometric convergence rate (6.1) for n large has been obtained under Conjecture 6.1 and observed numerically in section 7.2. The conjecture itself is of an independent interest which has no direct relation with the heat equation. In this section, we test the limits (6.3) and (6.4) in the conjecture.

In Table 7.4, the uniform norm of the difference in (6.3) and the ratio in (6.4) are given for Examples 7.1 and 7.2. In both cases, we have taken $v = 2t_0$. The results for Example 7.1 show the convergence clearly. In particular, the test for (6.4) shows that the ratio is identically one if the initial value $u_0(x)$ is the Gaussian.

The columns for Example 7.2 also show similar convergence behavior. However, the speed of its convergence is a lot slower. In fact, it is not even clear that taking $v = 2t_0$ is the correct one for the case of Example 7.2. Since $2t_0$ is the variance of Example 7.1, one may also try $2t_0 + 1$, which is the variance of Example 7.2. However, one cannot obtain the convergence, and $v = 2t_0$ seems a better choice. The computation is done up to $n = 10$ and that was our best. If computations with higher n are performed, we conjecture that the convergence to the unity will be more clearly observed.

Acknowledgment. Authors would like to thank anonymous reviewers. Their comments and suggestions helped to improve this paper.

REFERENCES

[1] N. I. Akhiezer, *The Classical Moment Problem and Some Related Questions in Analysis*, Hafner, New York, 1965.

[2] J. A. Carrillo and J. L. Vázquez, *Fine asymptotics for fast diffusion equations*, Comm. Partial Differential Equations, 28 (2003), pp. 1023–1056.

[3] R. E. Curto and L. A. Fialkow, *Recursiveness, positivity, and truncated moment problems*, Houston J. Math., 17 (1991), pp. 603–635.

[4] R. E. Curto and L. A. Fialkow, *Truncated K-moment problems in several variables*, J. Operator Theory, 54 (2005), pp. 189–226.

[5] C. M. Dafermos, *Regularity and large time behaviour of solutions of a conservation law without convexity*, Proc. Roy. Soc. Edinburgh Sect. A, 99 (1985), pp. 201–239.

[6] J. Denzler and R. McCann, *Fast diffusion to self-similarity: Complete spectrum, long time asymptotics, and numerology*, Arch. Ration. Mech. Anal., 175 (2005), pp. 301–342.

[7] R. J. DiPerna, *Decay and asymptotic behavior of solutions to nonlinear hyperbolic systems of conservation laws*, Indiana Univ. Math. J., 24 (1974/75), pp. 1047–1071.

[8] J. Dolbeault and M. Escobedo, *L^1 and L^∞ intermediate asymptotics for scalar conservation laws*, Asymptot. Anal., 41 (2005), pp. 189–213.

[9] J. Duoandikoetxea and E. Zuazua, *Moments, masses de Dirac et decomposition de fonctions*, C. R. Acad. Sci. Paris Ser. I Math., 315 (1992), pp. 693–698.

[10] M. Escobedo and E. Zuazua, *Large time behavior for convection-diffusion equations in R^N*, J. Funct. Anal., 100 (1991), pp. 119–161.

[11] M. Escobedo, J. Vazquez, and E. Zuazua, *Asymptotic behaviour and source-type solutions for a diffusion-convection equation*, Arch. Ration. Mech. Anal., 124 (1993), pp. 43–65.

[12] E. Hopf, *The partial differential equation $u_t + uu_x = \mu u_{xx}$*, Comm. Pure Appl. Math., 3 (1950), pp. 201–230.

[13] Y.-J. Kim, *Asymptotic behavior in scalar conservation laws and the optimal convergence order to N-waves*, J. Differential Equations, 192 (2003), pp. 202–224.

[14] Y.-J. Kim, *An Oleinik type estimate for a convection-diffusion equation and the convergence to N-waves*, J. Differential Equations, 199 (2004), pp. 269–289.

[15] Y.-J. Kim and W.-M. Ni, *On the rate of convergence and asymptotic profile of solutions to the viscous Burgers equation*, Indiana Univ. Math. J., 51 (2002), pp. 727–752.

[16] Y.-J. Kim and A. E. Tzavaras, *Diffusive N-waves and metastability in the Burgers equation*, SIAM J. Math. Anal., 33 (2001), pp. 607–633.

[17] P. D. Lax, *Hyperbolic systems of conservation laws.* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[18] T.-P. Liu, *Decay to N-waves of solutions of general systems of nonlinear hyperbolic conservation laws*, Comm. Pure Appl. Math., 30 (1977), pp. 586–611.

[19] T.-P. Liu, *Nonlinear stability of shock waves for viscous conservation laws*, Mem. Amer. Math. Soc., 56 (1985), no. 328.

[20] T.-P. Liu and M. Pierre, *Source-solutions and asymptotic behavior in conservation laws*, J. Differential Equations, 51 (1984), pp. 419–441.

[21] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton Math. Ser. 30, Princeton University Press, Princeton, NJ, 1970.

[22] T. P. Witelski and A. J. Bernoff, *Self-similar asymptotics for linear and nonlinear diffusion equations*, Stud. Appl. Math., 100 (1998), pp. 153–193.

[23] G. Whitham, *Linear and Nonlinear Waves*, Pure Appl. Math., Wiley Interscience, New York, 1974.

SIAM J. MATH. ANAL.
Vol. 40, No. 6, pp. 2262–2285

ON DIVERGENCE FORM SPDES WITH VMO COEFFICIENTS*

N. V. KRYLOV†

Abstract. We present several results on solvability in Sobolev spaces W_p^1 of SPDEs in divergence form in the whole space.

Key words. stochastic partial differential equations, divergence equations, Sobolev spaces

AMS subject classifications. 60H15, 35R60

DOI. 10.1137/080726902

1. Introduction.

The theory of (usual) partial differential equations has two rather different parts depending on whether the equations are written in divergence or nondivergence form. Quite often the starting point is the same: equations with constant coefficients, and then one uses different techniques to treat different types of equations.

By now, one can say that the L_p-theory of evolutional second-order SPDEs is quite well developed. The most advanced results of this theory can be found in the following papers and references therein: [1] (nondivergence type equations), [2] and [3] (divergence type equations). The results of the present paper are close to the corresponding results of [2]. However, unlike [2] we do not assume that the leading coefficients are continuous in the space variable. Instead we assume that the leading coefficients of the "deterministic" part of the equation are in VMO (the space of functions with vanishing mean oscillation), which is a much wider class than C. Still the leading coefficients of the "stochastic" part are assumed to be continuous in x.

The exposition in [2] and [3] is based on the theory of solvability in spaces $H_p^\gamma = (1-\Delta)^{-\gamma/2}L_p$ of SPDEs with coefficients independent of x. Then the method of "freezing" the coefficients is applied as in the general framework set out in [6]. This method does not work if the coefficients are only in VMO so we use a different technique based on recent results from [8] on deterministic parabolic equations with VMO coefficients. In addition, our technique allows us to avoid using the W_2^n theory of SPDEs, which is a starting point in the paper [6] and subsequent articles based on it.

One more difference of our approach from the one in [2] is that we represent the free term in the deterministic part in the form $D_i f^i + f^0$ with $f^j \in L_p$ (see (1.1) below). Of course, this is just a general form of a distribution from H_p^{-1}. However, the spaces H_p^γ are most appropriate for equations in nondivergence form. One general inconvenience of these spaces is that the space or space-time dilations affect the norms in a way which is hard to control. For divergence form equations with low regularity of coefficients the most important space is H_p^1. This space coincides with the Sobolev space W_p^1 and the effect of dilations on the norm or on $D_i f^i + f^0$ can be easily taken into account.

The exposition here is self-contained apart from references to some very basic results of [6], [8], and [12], and is much more elementary than in [2], employing the

*Received by the editors June 11, 2008; accepted for publication (in revised form) October 30, 2008; published electronically February 20, 2009. This work was partially supported by NSF Grant DMS-0653121.

http://www.siam.org/journals/sima/40-6/72690.html

†School of Mathematics, University of Minnesota, 127 Vincent Hall, Minneapolis, MN 55455 (krylov@math.umn.edu).

derivatives instead of the powers of the Laplacian, and yet gives more information. In particular, the author intends to use Corollary 5.5 in order to largely simplify the theory in [2] of divergence form SPDEs in domains. It turns out that to develop this theory one need not first develop the theory of SPDEs in domains with coefficient independent of x, which in itself required quite a bit of work.

The author's interest in divergence-type equations and in simplifying the theory of them appeared after he realized that the corresponding results can be applied to filtering theory of partially observable diffusion processes, given by stochastic Itô equations. It turns out that, under Lipschitz and nondegeneracy conditions only, the filtering density is almost Lipschitz in x and almost Hölder $1/2$ in time. This is proved in [10] on the basis of Theorems 2.2 through 2.6 of the present article. The filtering density satisfies an SPDE usually written in terms of the operators adjoint to operators in nondivergence form with Lipschitz continuous coefficients. Writing these adjoint operators in divergence form makes perfect sense and allows us to obtain the above-mentioned results (see [10]).

Our Theorem 2.2 is very close to Theorem 2.12 of [2]. Apart from weaker conditions on the coefficients, another important difference is the presence of the parameter λ in (2.10). One of differences in the proofs is that we avoid proving the solvability on small consecutive time intervals and then gluing together the results.

Let $(\Omega, \mathcal{F}, P)$ be a complete probability space with an increasing filtration $\{\mathcal{F}_t, t \geq 0\}$ of complete with respect to $(\mathcal{F}, P)$ σ-fields $\mathcal{F}_t \subset \mathcal{F}$. Denote by $\mathcal{P}$ the predictable σ-field in $\Omega \times (0, \infty)$ associated with $\{\mathcal{F}_t\}$. Let w_t^k, $k = 1, 2, \dots$, be independent one-dimensional Wiener processes with respect to $\{\mathcal{F}_t\}$.

We fix a stopping time τ and for $t \leq \tau$ in the Euclidean d-dimensional space R^d of points $x = (x^1, \dots, x^d)$ we consider the following equation

$$du_t = \left(L_t u_t - \lambda u_t + D_i f_t^i + f_t^0\right) dt + \left(\Lambda_t^k u_t + g_t^k\right) dw_t^k, \tag{1.1}$$

where $u_t = u_t(x) = u_t(\omega, x)$ is an unknown function,

$$L_t \psi(x) = D_j \left(a_t^{ij}(x) D_i \psi(x) + a_t^j(x) \psi(x)\right) + b_t^i(x) D_i \psi(x) + c_t(x) \psi(x),$$

$$\Lambda_t^k \psi(x) = \sigma_t^{ik}(x) D_i \psi(x) + \nu_t^k(x) \psi(x),$$

the summation convention with respect to $i, j = 1, \dots, d$ and $k = 1, 2, \dots$ is enforced and detailed assumptions on the coefficients and the free terms will be given later.

One can rewrite (1.1) in the nondivergence form assuming that the coefficients a_t^{ij} and a_t^j are differentiable in x and then one could apply the results from [6]. It turns out that the differentiability of a_t^{ij} and a_t^j is not needed for the corresponding counterparts of the results in [6] to be true and showing this and generalizing the corresponding results of [2] is one of the main purposes of the present article.

The author is sincerely grateful to Kyeong-Hun Kim who kindly pointed out a serious error in the first draft of the article. Doyoon Kim and the referees of the paper made many valuable suggestions and their impact is greatly appreciated.

2. Main results. Fix a number

$$p \geq 2,$$

and denote $L_p = L_p(\mathsf{R}^d)$. We use the same notation L_p for vector- and matrix-valued or else ℓ_2-valued functions such as $g_t = (g_t^k)$ in (1.1). For instance, if $u(x) =$

$(u^1(x),u^2(x),\dots)$ is an ℓ_2-valued measurable function on R^d, then

$$\|u\|_{L_p}^p=\int_{\mathsf{R}^d}|u(x)|_{\ell_2}^p\,dx=\int_{\mathsf{R}^d}\left(\sum_{k=1}^\infty\left|u^k(x)\right|^2\right)^{p/2}dx.$$

Introduce

$$D_i=\frac{\partial}{\partial x^i},\quad i=1,\dots,d,\quad \Delta=D_1^2+\cdots+D_d^2.$$

By Du we mean the gradient with respect to x of a function u on R^d.

As usual,

$$W_p^1=\{u\in L_p:Du\in L_p\},\quad \|u\|_{W_p^1}=\|u\|_{L_p}+\|Du\|_{L_p}.$$

Recall that τ is a stopping time and introduce

$$\mathbf{L}_p(\tau):=L_p(\llbracket 0,\tau\rrbracket,\mathcal{P},L_p),\quad \mathbf{W}_p^1(\tau):=L_p\left(\llbracket 0,\tau\rrbracket,\mathcal{P},W_p^1\right).$$

We also need the space $\mathcal{W}_p^1(\tau)$, which is the space of functions $u_t=u_t(\omega,\cdot)$ on $\{(\omega,t):0\le t\le\tau,t<\infty\}$ with values in the space of generalized functions on R^d and having the following properties:

(i) We have $u_0\in L_p(\Omega,\mathcal{F}_0,L_p)$;

(ii) We have $u\in\mathbf{W}_p^1(\tau)$;

(iii) There exist $f^i\in\mathbf{L}_p(\tau)$, $i=0,\dots,d$, and $g=(g^1,g^2,\dots)\in\mathbf{L}_p(\tau)$ such that for any $\varphi\in C_0^\infty=C_0^\infty(\mathsf{R}^d)$ with probability 1 for all $t\in[0,\infty)$ we have

$$\begin{aligned}(u_{t\wedge\tau},\varphi)=(u_0,\varphi)&+\sum_{k=1}^\infty\int_0^t I_{s\le\tau}\left(g_s^k,\varphi\right)\,dw_s^k\\&+\int_0^t I_{s\le\tau}\left(\left(f_s^0,\varphi\right)-\left(f_s^i,D_i\varphi\right)\right)\,ds.\end{aligned}\tag{2.1}$$

In particular, for any $\phi\in C_0^\infty$, the process $(u_{t\wedge\tau},\phi)$ is $\mathcal{F}_t$-adapted and (a.s.) continuous.

The reader can find in [6] a discussion of (ii) and (iii), in particular, the fact that the series in (2.1) converges uniformly in probability on every finite subinterval of $[0,\tau]$. On the other hand, it is worth saying that the above introduced space $\mathcal{W}_p^1$ is not quite the same as $\mathcal{H}_p^1(\tau)$ in [6] or in [2]. There are three differences. One is that there is an additional restriction on u_0 in [6] and [2]. But in the main part of the article we are going to work with $\mathcal{W}_{p,0}^1(\tau)$ which is the subset of $\mathcal{W}_p^1(\tau)$ consisting of functions with $u_0=0$. Another issue is that in [6] and [2] in place of $D_if^i+f^0$ we have just f such that

$$f\in\mathbf{H}_p^{-1}(\tau)=L_p\left(\llbracket 0,\tau\rrbracket,\mathcal{P},H_p^{-1}\right).$$

Actually, this difference is fictitious because one knows that any $f\in H_p^{-1}$

(a) has the form $D_if^i+f^0$ with $f^j\in L_p$ and

$$\|f\|_{H_p^{-1}}\le N\sum_{j=0}^d\|f^j\|_{L_p},$$

where N is independent of f,f^j, and on the other hand,

(b) for any $f \in H_p^{-1}$ there exist $f^j \in L_p$ such that $f = D_i f^i + f^0$ and

$$\sum_{j=0}^{d} \|f^j\|_{L_p} \leq N \|f\|_{H_p^{-1}},$$

where N is independent of f.

The third difference is that instead of (ii) the condition $D^2 u \in \mathbf{H}_p^{-1}(\tau)$ is required in [6] and [2]. However, as it follows from Theorem 3.7 of [6] and the boundedness of the operator $D : L_p \to H_p^{-1}$, this difference disappears if τ is a bounded stopping time.

To summarize, the spaces $\mathcal{W}^1_{p,0}(\tau)$ introduced above coincide with $\mathcal{H}^1_{p,0}(\tau)$ from [6] if τ is bounded, and we choose a particular representation of the deterministic part of the stochastic differential just for convenience. In the remainder of the article the spaces $\mathcal{H}^1_{p,0}(\tau)$ do not appear and none of their properties is used.

In case that property (iii) holds, we write

$$du_t = \left(D_i f_t^i + f_t^0\right) dt + g_t^k \, dw_t^k \tag{2.2}$$

for $t \leq \tau$ and this explains the sense in which equation (1.1) is understood. Of course, we still need to specify appropriate assumptions on the coefficients and the free terms in (1.1).

Assumption 2.1. (i) The coefficients a_t^{ij}, a_t^i, b_t^i, σ_t^{ik}, c_t, and ν_t^k are measurable with respect to $\mathcal{P} \times \mathcal{B}(\mathsf{R}^d)$, where $\mathcal{B}(\mathsf{R}^d)$ is the Borel σ-field on R^d.

(ii) There is a constant K such that for all values of indices and arguments

$$\left|a_t^i\right| + \left|b_t^i\right| + |c_t| + |\nu|_{\ell_2} \leq K, \quad c_t \leq 0.$$

(iii) There is a constant $\delta > 0$ such that for all values of the arguments and $\xi \in \mathsf{R}^d$

$$a_t^{ij} \xi^i \xi^j \leq \delta^{-1} |\xi|^2, \quad \left(a_t^{ij} - \alpha_t^{ij}\right) \xi^i \xi^j \geq \delta |\xi|^2, \tag{2.3}$$

where $\alpha_t^{ij} = (1/2)(\sigma^{i\cdot}, \sigma^{j\cdot})_{\ell_2}$. Finally, the constant $\lambda \geq 0$.

It is worth emphasizing that we do not require the matrix (a^{ij}) to be symmetric.

Assumption 2.1 guarantees that (1.1) makes perfect sense if $u \in \mathcal{W}^1_p(\tau)$. By the way, adding the term $-\lambda u_t$ with constant $\lambda \geq 0$ is one more technically convenient step. One can always introduce this term, if originally it is absent, by considering $v_t := u_t e^{\lambda t}$.

Let $\mathbf{B}$ denote the set of balls $B \subset \mathsf{R}^d$ and let $\rho(B)$ be the radius of $B \in \mathbf{B}$. For functions $h_t(x)$ on $[0, \infty) \times \mathsf{R}^d$ and $B \in \mathbf{B}$ introduce

$$h_{t(B)} = \frac{1}{|B|} \int_B h_t(x) \, dx,$$

where $|B|$ is the volume of B. Also let $\mathbf{Q}$ denote the set of all cylinders in $[0, \infty) \times \mathsf{R}^d$ of type $Q = (s, t) \times B$, where $B \in \mathbf{B}$ and $t - s = \rho^2(B)$. For such Q set $\rho(Q) = \rho(B)$. For $\rho \geq 0$, $s < t$, a continuous R^d-valued function $x_r, r \in [s, t]$, and a $Q = (s, t) \times B \in \mathbf{Q}$, introduce

$$\operatorname{osc}(h, Q, x.) = \frac{1}{t - s} \int_s^t \left(|h_r - h_{r(B + x_r)}|\right)_{(B + x_r)} dr,$$

$$\operatorname{Osc}(h, Q, \rho) = \sup_{|x.|_C \leq \rho} \operatorname{osc}(h, Q, x.), \quad \operatorname{osc}(h, Q) = \operatorname{osc}(h, Q, 0),$$

where $|x.|_C$ is the sup norm of $|x.|$.

Observe that $\operatorname{osc}(h, Q, x.) = 0$ if $h_t(x)$ is independent of x.

Denote by B_ρ the open ball with radius $\rho > 0$ centered at the origin, define $Q_\rho = (0, \rho^2) \times B_\rho$ and for $t \geq 0$ and $x \in \mathsf{R}^d$ set $B_\rho(x) = B_\rho + x$, $Q_\rho(t, x) = Q_\rho + (t, x)$.

In the remaining two assumptions we use constants $\beta > 0$ and $\beta_1 > 0$, the values of which will be specified later.

Let $t_0 \geq 0$, $x_0 \in \mathsf{R}^d$, and constants $\varepsilon \geq \varepsilon_1 > 0$. We say that the couple (a, σ) is $(\varepsilon, \varepsilon_1)$-regular at point (t_0, x_0) if (for any ω) either

(i) we have $\sigma_t^{nm}(x_0) = 0$ for $t \in (t_0, t_0 + \varepsilon_1^2)$ and all n, m and

$$\operatorname{osc}\left(a^{ij}, Q\right) \leq \beta, \quad \forall i, j, \tag{2.4}$$

for all $Q \in \mathbf{Q}$ such that $Q \subset Q_\varepsilon(t_0, x_0)$, or

(ii) for all $Q \in \mathbf{Q}$ such that $Q \subset Q_\varepsilon(t_0, x_0)$ we have

$$\operatorname{Osc}\left(a^{ij}, Q, \varepsilon\right) \leq \beta, \quad \forall i, j. \tag{2.5}$$

Note that (a, σ) is $(\varepsilon, \varepsilon_1)$-regular at any point (t_0, x_0) for any $\beta > 0$ if, for instance, a^{ij} depend only on x and are of class VMO.

Assumption 2.2. There exist $\varepsilon \geq \varepsilon_1 > 0$ such that (a, σ) is $(\varepsilon, \varepsilon_1)$-regular at any point (t_0, x_0) and

$$\left(a_t^{jk}(x) - \alpha_t^{jk}(y)\right) \xi^j \xi^k \geq \delta |\xi|^2$$

for all t, ξ, x, and y satisfying $|x - y| \leq \varepsilon$.

Assumption 2.3. There exists an $\varepsilon_2 > 0$ such that

$$\left|\sigma_t^{i\cdot}(x) - \sigma_t^{i\cdot}(y)\right|_{\ell_2} \leq \beta_1 \tag{2.6}$$

for all i, t, x, and y satisfying $|x - y| \leq \varepsilon_2$.

Needless to say that Assumptions 2.2 and 2.3 are satisfied with any $\beta, \beta_1 > 0$ and slightly reduced δ if (2.3) holds and $a_t^{ij}(x)$ and $\sigma_t^{i\cdot}(x)$ are uniformly continuous in x uniformly with respect to (ω, t).

Finally, we describe the space of initial data. Recall that for $p \geq 2$ the Slobodetskii space $W_p^{1-2/p} = W_p^{1-2/p}(\mathsf{R}^d)$ of functions $u_0(x)$ can be introduced as the space of traces on $t = 0$ of (deterministic) functions u such that

$$u \in L_p\left(\mathsf{R}_+, H_p^1\right), \quad \partial u / \partial t \in L_p\left(\mathsf{R}_+, H_p^{-1}\right),$$

where $\mathsf{R}_+ = (0, \infty)$. For such functions there is a (unique) modification denoted again u such that u_t is a continuous L_p-valued function on $[0, \infty)$ so that u_0 is well defined. Any such u_t is called an extension of u_0.

The norm in $W_p^{1-2/p}$ can be defined as the infimum of

$$\|u\|_{L_p\left(\mathsf{R}_+, H_p^1\right)} + \|\partial u / \partial t\|_{L_p\left(\mathsf{R}_+, H_p^{-1}\right)}$$

over all extensions u_t of elements u_0. It is also well known that an equivalent norm of u_0 can be introduced as

$$\|u\|_{L_p\left((0,1), W_p^1\right)},$$

where $u = u_t$ is defined as the (unique) solution of the heat equation $\partial u_t(x)/\partial t = \Delta u_t(x)$ with initial condition $u_0(x)$.

For $s \geq 0$ we introduce

$$\mathrm{tr}_s \mathcal{W}^1_p = L_p\left(\Omega, \mathcal{F}_s, W_p^{1-2/p}\right).$$

The following auxiliary result helps understand the role of $\mathrm{tr}_s \mathcal{W}^1_p$. We use spaces $\mathcal{W}^1_p([S,T))$ and $\mathbf{W}^1_p((S,T))$, which are introduced in the same way as $\mathcal{W}^1_p(\tau)$ and $\mathbf{W}^1_p(\tau)$, but the functions are considered only on $[S,T)$ and (S,T), respectively.

LEMMA 2.1. *Let $s \geq 0$ be a fixed number and let u_s be an $\mathcal{F}_s$-measurable function with values in the set of distributions over R^d.*

(i) *We have $u_s \in \mathrm{tr}_s \mathcal{W}^1_p$ if and only if there exists a $v \in \mathcal{W}^1_p([s,\infty))$ satisfying the equation*

$$\partial v/\partial t = \Delta v - v, \quad t \geq s, \tag{2.7}$$

(which is a particular case of (1.1) and is understood in the same sense) with initial data u_s. This v is unique and satisfies

$$\|v\|_{\mathbf{W}^1_p((s,\infty))} \leq N\|u_s\|_{\mathrm{tr}_s \mathcal{W}^1_p}, \quad \|u_s\|_{\mathrm{tr}_s \mathcal{W}^1_p} \leq N\|v\|_{\mathbf{W}^1_p((s,\infty))}, \tag{2.8}$$

where the constants N are independent of s, u_s, and v.

(ii) *We have $u_s \in \mathrm{tr}_s \mathcal{W}^1_p$ if and only if there exists a $v \in \mathcal{W}^1_p([s,s+1))$ such that $v_s = u_s$.*

(iii) *If such a v exists and $dv_t = (D_i f^i_t + f^0_t)\,dt + g^k_t\,dw^k_t$, $t \geq s$, then*

$$\|u_s\|_{\mathrm{tr}_s \mathcal{W}^1_p} \leq N\left(\|v\|_{\mathbf{W}^1_p((s,s+1))} + \sum_{j=0}^{d} \left\|f^j\right\|_{\mathbf{L}_p((s,s+1))} + \|g\|_{\mathbf{L}_p((s,s+1))}\right), \tag{2.9}$$

where the constant N is independent of s, u_s, and v.

(iv) *If $s > 0$ and we have a $u \in \mathcal{W}^1_p(s)$, then $u_s \in \mathrm{tr}_s \mathcal{W}^1_p$ and*

$$\|u_s\|_{\mathrm{tr}_s \mathcal{W}^1_p} \leq N\left(\|u\|_{\mathbf{W}^1_p(s)} + \sum_{j=0}^{d} \left\|f^j\right\|_{\mathbf{L}_p(s)} + \|g\|_{\mathbf{L}_p(s)}\right),$$

where N is independent of u, and f^j and g^k are taken from (2.2).

We prove this lemma in section 5.

Here are our main results concerning (1.1). The following theorem is very close to Theorem 2.12 of [2]. Important differences are the presence of the parameter λ in (2.10) and weaker assumptions on the coefficients of the deterministic part of the equation.

THEOREM 2.2. *Let the above assumptions be satisfied with $\beta = \beta(d,p,\delta) = \beta_0/3$, where β_0 is the constant from Lemma 5.1, and $\beta_1 = \beta_1(d,p,\delta,\varepsilon) > 0$ taken from the proof of Lemma 5.2. Let $\lambda \geq 0$, let $f^j, g \in \mathbf{L}_p(\tau)$, and let $u_0 \in \mathrm{tr}_0 \mathcal{W}^1_p$.*

(i) *Then equation (1.1) for $t \leq \tau \wedge T$ has a unique solution $u \in \mathcal{W}^1_p(\tau \wedge T)$ with initial data u_0 and any $T \in (0,\infty)$. Moreover, if*

$$\lambda \geq \lambda_0(d,p,\delta,K,\varepsilon,\varepsilon_1,\varepsilon_2) \geq 1,$$

then (1.1) for $t \leq \tau$ has a unique solution $u \in \mathcal{W}^1_p(\tau)$ with initial data u_0.

(ii) *Furthermore, if a* $v \in \mathcal{W}^1_p(\infty)$ *is defined by* (2.7) *with initial condition* u_0, *then the above solution* u *satisfies*

$$\lambda^{1/2}\|u\|_{\mathbf{L}_p(\tau)} + \|Du\|_{\mathbf{L}_p(\tau)} \le N\left(\sum_{i=1}^d \|f^i\|_{\mathbf{L}_p(\tau)} + \|g\|_{\mathbf{L}_p(\tau)} + \|Dv\|_{\mathbf{L}_p(\tau)}\right) + N\lambda^{-1/2}\|f^0\|_{\mathbf{L}_p(\tau)} + N\lambda^{1/2}\|v\|_{\mathbf{L}_p(\tau)}, \tag{2.10}$$

provided that $\lambda \ge \lambda_0$, *where the constants* $N, \lambda_0 \ge 1$ *depend only on* d, p, δ, K, ε, ε_1, *and* ε_2.

(iii) *Finally, there exists a set* $\Omega' \subset \Omega$ *of full probability such that* $u_{t\wedge\tau}I_{\Omega'}$ *is a continuous* $\mathcal{F}_t$*-adapted* L_p*-valued function of* $t \in [0,\infty)$.

Observe that estimate (2.10) shows one of good reasons for writing the free term in (1.1) in the form $D_i f^i + f^0$, because f^i, $i = 1, \dots, d$, and f^0 enter (2.10) differently.

Remark 2.3. As it follows from our proofs, if $p = 2$, Assumptions 2.2 and 2.3 are not needed for Theorem 2.2 to be true and mentioning ε, ε_1, and ε_2 can be dropped in the statement. Thus we provide a new way to prove the classical result on Hilbert space solvability of SPDEs (cf., for instance, [13]).

We prove Theorem 2.2 in section 6 after we prepare necessary tools in sections 3–5. In section 3 we prove uniqueness part of Theorem 2.2 on the basis of Itô's formula from [12]. Here Assumptions 2.2 and 2.3 are not used. In section 4 we treat the case of the heat equation with the random right-hand side and present a simplified version of the corresponding result from [6]. In section 5 we prove an auxiliary existence theorem and derive some a priori estimates.

Here is a result about continuous dependence of solutions on the data.

THEOREM 2.4. *Assume that for each* $n = 1, 2, \dots$ *we are given functions* a^{ij}_{nt}, a^i_{nt}, b^i_{nt}, c_{nt}, σ^{ik}_{nt}, ν^k_{nt}, f^j_{nt}, g^k_{nt}, *and* u_{n0} *having the same meaning as the original ones and satisfying the same assumptions as those imposed on the original ones in Theorem* 2.2 *(with the same* $\delta, K, \beta, \dots$*). Assume that for* $i, j = 1, \dots, d$ *and almost all* (ω, t, x) *we have*

$$\left(a^{ij}_{nt}, a^i_{nt}, b^i_{nt}, c_{nt}\right) \to \left(a^{ij}_t, a^i_t, b^i_t, c_t\right),$$

$$\left|\sigma^{i\cdot}_{nt} - \sigma^{i\cdot}_t\right|_{\ell_2} + |\nu_{nt} - \nu_t|_{\ell_2} \to 0,$$

as $n \to \infty$. *Also assume that*

$$\sum_{j=0}^d \left\|f^j_n - f^j\right\|_{\mathbf{L}_p(\tau)} + \|g_n - g\|_{\mathbf{L}_p(\tau)} + \|u_{n0} - u_0\|_{\mathrm{tr}_0\mathcal{W}^1_p} \to 0$$

as $n \to \infty$. *Take* $\lambda \ge \lambda_0$, *take the function* u *from Theorem* 2.2, *and let* $u_n \in \mathcal{W}^1_p(\tau)$ *be the unique solutions of* (1.1) *for* $t \le \tau$ *constructed from* a^{ij}_{nt}, a^i_{nt}, b^i_{nt}, c_{nt}, σ^{ik}_{nt}, ν^k_{nt}, f^j_{nt}, *and* g^k_{nt} *and having initial values* u_{n0}.

Then, as $n \to \infty$, *we have* $\|u_n - u\|_{\mathbf{W}^1_p(\tau)} \to 0$ *and for any finite* $T \in [0,\infty)$

$$E\sup_{t\le\tau\wedge T}\|u_{nt} - u_t\|^p_{L_p} \to 0. \tag{2.11}$$

Proof. Set $v_{nt} = u_{nt} - u_t$. Then

$$dv_{nt} = \left(L_{nt}v_{nt} - \lambda v_{nt} + D_i\tilde{f}^i_{nt} + \tilde{f}^0_{nt}\right)dt + \left(\Lambda^k_{nt}v_{nt} + \tilde{g}^k_{nt}\right)dw^k_t,$$

where L_{nt} and Λ^k_{nt} are the operators constructed from a^{ij}_{nt}, a^i_{nt}, b^i_{nt}, c_{nt}, and σ^{ik}_{nt}, ν^k_{nt}, respectively, and

$$\tilde{f}^i_{nt} = f^i_{nt} - f^i_t + \left(a^{ji}_{nt} - a^{ji}_t\right) D_j u_t + \left(a^i_{nt} - a^i_t\right) u_t,$$
$$\tilde{f}^0_{nt} = f^0_{nt} - f^0_t + \left(b^i_{nt} - b^i_t\right) D_i u_t + (c_{nt} - c_t) u_t,$$
$$\tilde{g}^k_{nt} = g^k_{nt} - g^k_t + \left(\sigma^{ik}_{nt} - \sigma^{ik}_t\right) D_i u_t + \left(\nu^k_{nt} - \nu^k_t\right) u_t.$$

By Theorem 2.2 we know that $u \in \mathbf{W}^1_p(\tau)$. This, along with our assumptions and the dominated convergence theorem, implies that

$$\sum_{j=0}^d \left\| \tilde{f}^j_n \right\|_{\mathbf{L}_p(\tau)} + \|\tilde{g}_n\|_{\mathbf{L}_p(\tau)} \to 0$$

as $n \to \infty$. After that by applying (2.10) to v_{nt} we immediately see that $\|u_n - u\|_{\mathbf{W}^1_p(\tau)} \to 0$.

Assertion (2.11) is, actually, a simple corollary of the above. Indeed, by introducing $\hat{f}^j_n$ and $\hat{g}^k_n$ in an obvious way, we can write

$$dv_{nt} = \left(D_i \hat{f}^i_{nt} + \hat{f}^0_{nt}\right) dt + \hat{g}^k_{nt}\, dw^k_t, \tag{2.12}$$

and

$$\sum_{j=0}^d \left\| \hat{f}^j_n \right\|_{\mathbf{L}_p(\tau)} + \|\hat{g}_n\|_{\mathbf{L}_p(\tau)} \to 0.$$

It is standard (see, for instance, our Theorem 3.1) to derive from here the estimate

$$E \sup_{t \le \tau \wedge T} \|u_{nt} - u_t\|^p_{L_p} \le N \left(\sum_{j=0}^d \left\| \hat{f}^j_n \right\|_{\mathbf{L}_p(\tau \wedge T)} + \|\hat{g}_n\|_{\mathbf{L}_p(\tau \wedge T)} + E\|u_{n0} - u_0\|^p_{L_p} \right),$$

where N is independent of n. It is also well known that $W^{1-2/p}_p \subset L_p$, that is

$$\|u_{n0} - u_0\|_{L_p} \le N \|u_{n0} - u_0\|_{W^{1-2/p}_p}.$$

By combining all this together we obtain (2.11) and the theorem is proved. □

The following result could be proved on the basis of Theorem 2.4 in the same way as Corollary 5.11 of [6], where the solutions are approximated by solutions of equations with smooth coefficients and then a stopping time technique was used. We give here a shorter proof based on a different idea.

THEOREM 2.5. *Let $p_1, p_2 \in [2, \infty)$, $p_1 < p_2$, and let the above assumptions be satisfied with $\beta \le \beta(d, p, \delta)$ for all $p \in [p_1, p_2]$ and $\beta_1 \le \beta_1(d, p, \delta, \varepsilon)$ for all $p \in [p_1, p_2]$. Let $\lambda \ge 0$, and suppose that for $p \in [p_1, p_2]$ we have $f^j, g \in \mathbf{L}_p(\tau)$, and $u_0 \in \mathrm{tr}_0 \mathcal{W}^1_p$.*

Then the solutions corresponding to $p = p_1$ and $p = p_2$ coincide, that is, there is a unique solution $u \in \mathcal{W}^1_{p_1}(\tau) \cap \mathcal{W}^1_{p_2}(\tau)$ of (1.1) with initial data u_0.

Proof. Obviously, it suffices to concentrate on bounded τ. As is explained above in that case we may assume that λ is as large as we like. We take it so large that one could use assertion (ii) of Theorem 2.2 with any $p \in [p_1, p_2]$.

Denote by u the solution corresponding to $p = p_2$ and observe that, owing to uniqueness of solutions in $\mathcal{W}^1_{p_1}(\tau)$, we need only show that $u \in \mathcal{W}^1_{p_1}(\tau)$.

Take a nonnegative $\zeta \in C_0^\infty$ such that $\zeta(0) = 1$, set $\zeta_n(x) = \zeta(x/n)$, and notice that $u^n := u\zeta_n$ satisfies

$$du^n_t = \left(L_t u^n_t - \lambda u^n_t + D_i f^i_{nt} + f^0_{nt}\right) dt + \left(\Lambda^k_t u^n_t + g^k_{nt}\right) dw^k_t,$$

where

$$\begin{aligned}
f^i_{nt} &= f^i_t \zeta_n - u a^{ji}_t D_j \zeta_n, \quad i \geq 1,\\
f^0_{nt} &= f^0_t \zeta_n - f^i_t D_i \zeta_n - \left(a^{ij}_t D_i u_t + a^j_t u\right) D_j \zeta_n - b^i_t u_t D_i \zeta_n,\\
g^k_{nt} &= g^k_t \zeta_n - \sigma^{ik}_t u_t D_i \zeta_n.
\end{aligned}$$

It follows that for $p_1 \leq p \leq p_2$ we have

$$\|u^n\|_{\mathbf{W}^1_p(\tau)} \leq N \left(\sum_{i=0}^d \left\|f^i_n\right\|_{\mathbf{L}_p(\tau)} + \|g_n\|_{\mathbf{L}_p(\tau)} + \|u_0\zeta_n\|_{\mathrm{tr}_0 \mathcal{W}^1_p} \right). \tag{2.13}$$

One knows that with constants N independent of n

$$\|u_0\zeta_n\|_{\mathrm{tr}_0\mathcal{W}^1_p} \leq N \left(\|u_0\zeta_n\|_{\mathrm{tr}_0\mathcal{W}^1_{p_1}} + \|u_0\zeta_n\|_{\mathrm{tr}_0\mathcal{W}^1_{p_2}} \right) \leq N \left(\|u_0\|_{\mathrm{tr}_0\mathcal{W}^1_{p_1}} + \|u_0\|_{\mathrm{tr}_0\mathcal{W}^1_{p_2}} \right).$$

Similarly, and by Hölder's inequality,

$$\left\|f^i_n\right\|_{\mathbf{L}_p(\tau)} \leq N + N\|uD\zeta_n\|_{\mathbf{L}_p(\tau)} \leq N + \|u\|_{\mathbf{L}_{p_2}(\tau)} \|D\zeta_n\|_{\mathbf{L}_q(\tau)},$$

where

$$q = \frac{pp_2}{p_2 - p}.$$

Similar estimates are available for other terms in the right-hand side of (2.13). Since

$$\|D\zeta_n\|_{\mathbf{L}_q(\tau)} = N n^{-1+(p_2-p)d/(p_2 p)} \to 0$$

as $n \to \infty$ if

$$\frac{1}{p} - \frac{1}{p_2} < \frac{1}{d}, \tag{2.14}$$

estimate (2.13) implies that $u \in \mathcal{W}^1_p(\tau)$.

Thus, knowing that $u \in \mathcal{W}^1_{p_2}(\tau)$ allowed us to conclude that $u \in \mathcal{W}^1_p(\tau)$ as long as $p \in [p_1, p_2]$ and (2.14) holds. We can now replace p_2 with a smaller p and keep going in the same way each time increasing $1/p$ by the same amount until p reaches p_1. Then we get that $u \in \mathcal{W}^1_{p_1}(\tau)$. The theorem is proved. □

In many situations the following maximum principle is useful.

THEOREM 2.6. *Let the above assumptions be satisfied with $\beta \leq \beta(d, q, \delta)$ for all $q \in [2, p]$ and $\beta_1 \leq \beta_1(d, q, \delta, \varepsilon)$ for all $q \in [2, p]$. Let $\lambda \geq 0$ and $f^0 \in \mathbf{L}_p(\tau)$, $u_0 \in \mathrm{tr}_0\mathcal{W}^1_p$, $f^i = 0$, $i = 1, \ldots, d$, $g = 0$ be such that $u_0 \geq 0$ and $f^0 \geq 0$. Then for the solution u almost surely we have $u_t \geq 0$ for all finite $t \leq \tau$.*

Proof. If $p = 2$ the result is proved in [9]. For general $p \geq 2$ take the same function ζ_n as in the preceding proof, introduce $f^{ni} = f^i\zeta_n$, $g^k_n = 0$, and call u^n the solution of (1.1) with so modified free terms and the initial data $u_0\zeta_n$. By Theorem 2.5 we have $u^n \in \mathcal{W}^1_p(\tau) \cap \mathcal{W}^1_2(\tau)$. By the above, $u^n \geq 0$ and it only remains to use Theorem 2.4. The theorem is proved. □

3. Itô's formula and uniqueness. The following two "standard" results are taken from [12].

THEOREM 3.1. *Let $u \in \mathcal{W}^1_p(\tau)$, $f^j \in \mathbf{L}_p(\tau)$, $g = (g^k) \in \mathbf{L}_p(\tau)$, and assume that* (2.2) *holds for $t \le \tau$ in the sense of generalized functions. Then there is a set $\Omega' \subset \Omega$ of full probability such that*

(i) *$u_{t\wedge\tau} I_{\Omega'}$ is a continuous L_p-valued $\mathcal{F}_t$-adapted function on $[0,\infty)$;*

(ii) *for all $t \in [0,\infty)$ and $\omega \in \Omega'$ Itô's formula holds:*

$$\begin{aligned}\int_{\mathbf{R}^d} |u_{t\wedge\tau}|^p\,dx &= \int_{\mathbf{R}^d} |u_0|^p\,dx + p\int_0^{t\wedge\tau}\int_{\mathbf{R}^d} |u_s|^{p-2}u_s g^k_s\,dx\,dw^k_s\\ &+ \int_0^{t\wedge\tau}\left(\int_{\mathbf{R}^d}\left[p|u_t|^{p-2}u_t f^0_t - p(p-1)|u_t|^{p-2}f^i_t D_i u_t\right.\right.\\ &\left.\left. + (1/2)p(p-1)|u_t|^{p-2}|g_t|^2_{\ell_2}\right]dx\right)dt.\end{aligned} \tag{3.1}$$

Furthermore, for any $T \in [0,\infty)$

$$\begin{aligned}E\sup_{t\le\tau\wedge T}\|u_t\|^p_{L_p} &\le 2E\|u_0\|^p_{L_p} + NT^{p-1}\left\|f^0\right\|^p_{\mathbf{L}_p(\tau)}\\ &+ NT^{(p-2)/2}\left(\sum_{i=1}^d \left\|f^i\right\|^p_{\mathbf{L}_p(\tau)} + \|g\|^p_{\mathbf{L}_p(\tau)} + \|Du\|^p_{\mathbf{L}_p(\tau)}\right),\end{aligned} \tag{3.2}$$

where $N = N(d,p)$.

Here is an "energy" estimate.

COROLLARY 3.2. *Under the conditions of Theorem* 3.1 *assume that $\tau < \infty$ (a.s.). Then*

$$\begin{aligned}E\int_{\mathbf{R}^d}|u_0|^p\,dx + E\int_0^\tau\left(\int_{\mathbf{R}^d}\left[p|u_t|^{p-2}u_t f^0_t - p(p-1)|u_t|^{p-2}f^i_t D_i u_t\right.\right.\\ \left.\left. + (1/2)p(p-1)|u_t|^{p-2}|g_t|^2_{\ell_2}\right]dx\right)dt \ge EI_{\tau<\infty}\int_{\mathbf{R}^d}|u_\tau|^p\,dx.\end{aligned} \tag{3.3}$$

Furthermore, if τ is bounded, then there is an equality instead of inequality in (3.3).

The next result implies, in particular, uniqueness in Theorem 2.2.

LEMMA 3.3. *Under Assumption* 2.1 *there exist $\lambda_0 \ge 0$ and N depending only on d, p, K, and δ such that, for any strictly positive $\lambda \ge \lambda_0$ and any solution $u \in \mathcal{W}^1_{p,0}(\tau)$ of* (1.1) *for $t \le \tau$, we have*

$$\lambda\|u\|_{\mathbf{L}_p(\tau)} \le N\lambda^{1/2}\left(\sum_{j=1}^d \left\|f^j\right\|_{\mathbf{L}_p(\tau)} + \|g\|_{\mathbf{L}_p(\tau)}\right) + N\left\|f^0\right\|_{\mathbf{L}_p(\tau)}. \tag{3.4}$$

Furthermore, if $a^i = b^i = \nu^k \equiv 0$, then one can take $\lambda_0 = 0$.

Proof. We may assume that $f^j \in \mathbf{L}_p(\tau)$, $g = (g^k) \in \mathbf{L}_p(\tau)$, since otherwise the right-hand side of (3.4) is infinite.

If (3.4) is true for $\tau\wedge T$ in place of τ and any $T \in (0,\infty)$, then it is obviously also true as is. Therefore, we may assume that τ is finite. An advantage of this assumption is that we can use Corollary 3.2. Write (3.3) with $\hat f^i_t$, $\hat f^0_t$, and $\hat g^k_t$ in place of f^i_t, f^0_t, and g^k_t, respectively, where

$$\begin{aligned}&\hat f^i_t = a^{ji}_t D_j u_t + a^i_t u_t + f^i_t,\\ &\hat f^0_t = b^i_t D_i u_t + (c_t - \lambda)u_t + f^0_t, \quad \hat g^k_t = \sigma^{ik}_t D_i u_t + \nu^k_t u_t + g^k_t.\end{aligned}$$

Then observe that inequalities like $(a+b)^2 \leq (1+\varepsilon)a^2 + (1+\varepsilon^{-1})b^2$ show that for any $\varepsilon \in (0,1]$ we have

$$|\hat{g}_t|^2_{\ell_2} \leq (1+\varepsilon)\left|\sum_{i=1}^d \sigma^i_t D_i u_t\right|^2_{\ell_2} + 2\varepsilon^{-1}|\nu_t u_t + g_t|^2_{\ell_2}$$
$$\leq 2(1+\varepsilon)\alpha^{ij}_t (D_i u_t) D_j u_t + N\varepsilon^{-1}\left(|u_t|^2 + |g_t|^2_{\ell_2}\right).$$

Owing to (2.3), for $\varepsilon = \varepsilon(\delta) > 0$ small enough

$$\begin{aligned} I_t &:= (1/2)|u_t|^{p-2}|\hat{g}_t|^2_{\ell_2} - |u_t|^{p-2}\hat{f}^i_t D_i u_t + (p-1)^{-1}|u_t|^{p-2}u_t b^i_t D_i u_t \\ &\leq -(\delta/2)|u_t|^{p-2}|Du_t|^2 \\ &\quad + N|u_t|^{p-2}\left(|u_t|^2 + |g_t|^2_{\ell_2} + |Du_t|\,|u_t| + |Du_t|\sum_{i=1}^d \left|f^i_t\right|\right). \end{aligned} \tag{3.5}$$

Next, we use that for any $\gamma > 0$

$$|u_t|^{p-1}|Du_t| = \left(|u_t|^{(p-2)/2}|Du_t|\right)|u_t|^{p/2} \leq \gamma|u_t|^{p-2}|Du_t|^2 + \gamma^{-1}|u_t|^p,$$
$$|u_t|^{p-2}|Du_t|\,\left|f^i_t\right| \leq \gamma|u_t|^{p-2}|Du_t|^2 + \gamma^{-1}|u_t|^{p-2}\left|f^i_t\right|^2,$$

and by choosing γ appropriately find from (3.5) that

$$I_t \leq N|u_t|^p + N|u_t|^{p-2}\left(\sum_{i=1}^d \left|f^i_t\right|^2 + |g_t|^2_{\ell_2}\right). \tag{3.6}$$

After that Hölder's inequality and (3.3), where the right-hand side is nonnegative, immediately lead to

$$(\lambda - N_1)\|u\|^p_{\mathbf{L}_p(\tau)} \leq N\|u\|^{p-2}_{\mathbf{L}_p(\tau)}\left(\sum_{i=1}^d \|f^i\|^2_{\mathbf{L}_p(\tau)} + \|g\|^2_{\mathbf{L}_p(\tau)}\right) + N\|u\|^{p-1}_{\mathbf{L}_p(\tau)}\left\|f^0\right\|_{\mathbf{L}_p(\tau)}.$$

Furthermore, simple inspection of the above argument shows that, if $a^i = b^i = \nu^k \equiv 0$, then the terms with $|u_t|^2$ and $|u_t|\,|Du_t|$ in (3.5) and the term with $|u_t|^p$ in (3.6) disappear, so that we can take $N_1 = 0$ in this case (recall that $c \leq 0$). Generally, for $\lambda \geq 2N_1$ we have $\lambda - N_1 \geq (1/2)\lambda$ and

$$\bar{U}^p \leq N\bar{U}^{p-2}\bar{G}^2 + N\bar{U}^{p-1}\bar{F},$$

where

$$\bar{U} = \lambda\|u\|_{\mathbf{L}_p(\tau)}, \quad \bar{G} = \lambda^{1/2}\left(\sum_{i=1}^d \left\|f^i\right\|_{\mathbf{L}_p(\tau)} + \|g\|_{\mathbf{L}_p(\tau)}\right), \quad \bar{F} = \left\|f^0\right\|_{\mathbf{L}_p(\tau)}.$$

It follows that $\bar{U} \leq N(\bar{G} + \bar{F})$, which is (3.4) and the lemma is proved. □

4. Case of the heat equation. To move further we need the following analytic fact established in [4] (see also [7] for a complete proof).

LEMMA 4.1. *Denote by T_t the heat semigroup in R^d and let $p \geq 2$, $-\infty \leq a < b \leq \infty$, $g \in L_p((a,b)\times\mathsf{R}^d, \ell_2)$. Then*

$$\int_{\mathsf{R}^d}\int_a^b \left[\int_a^t |DT_{t-s}g_s(x)|^2_{\ell_2}\,ds\right]^{p/2} dtdx \leq N(d,p)\int_{\mathsf{R}^d}\int_a^b |g_t(x)|^p_{\ell_2}\,dtdx.$$

In this section we deal with the following model equation

$$du_t = \Delta u_t\,dt + g_t^k\,dw_t^k. \tag{4.1}$$

LEMMA 4.2. *Assume that $\tau \le T$, where the constant $T \in [0,\infty)$. Then for any $g = (g^1, g^2, \dots) \in \mathbf{L}_p(\tau)$ there exists a unique $u \in \mathcal{W}^1_{p,0}(\tau)$ satisfying (4.1) for $t \le \tau$. Furthermore, for this solution we have*

$$E\sup_{t\le\tau}\|u_t\|_{L_p}^p \le N(d,p)T^{(p-2)/2}\|g\|^p_{\mathbf{L}_p(\tau)}, \tag{4.2}$$

$$\|Du\|_{\mathbf{L}_p(\tau)} \le N(d,p)\|g\|_{\mathbf{L}_p(\tau)}. \tag{4.3}$$

Proof. By replacing the unknown function u_t with $v_t e^{\lambda t}$ we see that v_t satisfies

$$dv_t = (\Delta v_t - \lambda v_t)\,dt + e^{-\lambda t}g_t^k\,dw_t^k.$$

Since τ is bounded, the inclusions $u \in \mathcal{W}^1_{p,0}(\tau)$ and $v \in \mathcal{W}^1_{p,0}(\tau)$ are equivalent and our assertion about uniqueness follows from Lemma 3.3.

In the proof of existence we borrow part of the proof of Theorem 4.2 of [6]. As we have pointed out in the Introduction, the beginning of the theory of divergence and nondivergence type equations is the same. The main difference with that proof is that here we take $f \equiv 0$.

We take an integer $m \ge 1$, some stopping times $\tau_0 \le \tau_1 \le \cdots \le \tau_m \le T$, and some (nonrandom) functions $g^{ij} \in C_0^\infty$, $i,j = 1,\dots,m$. Then we define

$$g_t^k(x) = \sum_{i=1}^m g^{ik}(x) I_{(\tau_{i-1},\tau_i]}(t),$$

$$v_t(x) = \sum_{k=1}^m \int_0^t g_s^k(x)\,dw_s^k = \sum_{i,k=1}^m g^{ik}(x)\left(w^k_{t\wedge\tau_i} - w^k_{t\wedge\tau_{i-1}}\right), \quad t \ge 0.$$

Obviously, for any ω, the function $v_t(x)$ is continuous and bounded in (t,x) along with any derivative in x. Furthermore, the function and its derivative in x are Hölder $1/3$ continuous in t uniformly with respect to x (for almost any ω). Also, $v_t(x)$ has compact support in x.

These properties of $v_t(x)$ imply that for any ω there exists a unique classical solution of the heat equation

$$\frac{\partial}{\partial t}\bar{u}_t = \Delta\bar{u}_t + \Delta v_t, \quad t > 0,$$

with zero initial data. Furthermore,

$$\bar{u}_t(x) = \int_0^t T_{t-s}\Delta v_s(x)\,ds. \tag{4.4}$$

This formula shows, in particular, that $\bar{u}_t(x)$ is $\mathcal{F}_t$-adapted. Adding the fact that $\bar{u}_t$ is continuous in t proves that $\bar{u}_t(x)$ is predictable. The same holds for

$$(\bar{u}_t,\phi) = \int_0^t (T_{t-s}\Delta v_s,\phi)\,ds$$

with any $\phi \in C_0^\infty$. The following corollary of Minkowski's inequality

$$\|\bar{u}_t\|_{L_p} \leq \int_0^t \|\Delta v_s\|_{L_p}\, ds \tag{4.5}$$

shows that $\bar{u}_t$ is L_p-valued. Since $(\bar{u}_t, \phi)$ is predictable for any $\phi \in C_0^\infty$, $\bar{u}_t$ is weakly and, hence, strongly predictable as an L_p-valued process.

One can differentiate (4.4) with respect to x as many times as one wants and get similar statements about the derivatives of $\bar{u}_t$. In particular, (4.5) implies that for any multi-index α

$$E\int_0^T \int_{\mathsf{R}^d} |D^\alpha \bar{u}_t|^p\, dxdt \leq T^p E \int_0^T \int_{\mathsf{R}^d} |D^\alpha \Delta v_t|^p\, dxdt < \infty,$$

so that $\bar{u}_t \in \mathcal{W}^1_{p,0}(T)$.

Now, it is easily seen that

$$u_t(x) := \bar{u}_t(x) + v_t(x)$$

satisfies (4.1) pointwisely and by the above $u_t \in \mathcal{W}^1_{p,0}(T)$. The (deterministic) Fubini's theorem also shows that u_t satisfies (4.1) in the sense of distributions.

Next, we use the same simple transformation as in the proof of Lemma 4.1 of [6] and conclude that for any t and x almost surely

$$Du_t(x) = \sum_{k=1}^m \int_0^t T_{t-s} D g_s^k(x)\, dw_s^k.$$

Hence, by Burkholder–Davis–Gundy inequality

$$E|Du_t(x)|^p \leq NE\left[\int_0^t |T_{t-s} D g_s(x)|_{\ell_2}^2\, ds\right]^{p/2},$$

which along with Lemma 4.1 proves (4.3) for our particular g. Theorem 3.1 shows that (4.2) follows from (4.3) and (4.1).

The rest is trivial since the set of g's like the one above is dense in $\mathbf{L}_p(T)$ by Theorem 3.10 of [6]. The lemma is proved. □

Next we introduce the parameter λ into (4.1).

LEMMA 4.3. *Assume that $\tau \leq T$, where the constant $T \in [0, \infty)$. Let $\lambda > 0$. Then for any $g = (g^1, g^2, \dots) \in \mathbf{L}_p(\tau)$ there exists a unique $u \in \mathcal{W}^1_{p,0}(\tau)$ satisfying*

$$du_t = (\Delta u_t - \lambda u_t)\, dt + g_t^k\, dw_t^k \tag{4.6}$$

for $t \leq \tau$. Furthermore, for this solution we have

$$\lambda^{p/2}\|u\|^p_{\mathbf{L}_p(\tau)} \leq N(d,p)\|g\|^p_{\mathbf{L}_p(\tau)}, \tag{4.7}$$

$$\|Du\|_{\mathbf{L}_p(\tau)} \leq N(d,p)\|g\|_{\mathbf{L}_p(\tau)}. \tag{4.8}$$

Proof. Uniqueness and estimate (4.7) follow from Lemma 3.3. The existence immediately follows from Lemma 4.2 and the result of transformation described in the beginning of its proof. To establish (4.8) consider the heat equation

$$\frac{\partial}{\partial t} v_t = \Delta v_t - \lambda u_t. \tag{4.9}$$

Since $u \in \mathbf{L}_p(\tau)$, for almost any ω we have $u \in L_p((0,\tau)\times \mathsf{R}^d)$ and by a classical result (see, for instance, [11]) for almost any ω equation (4.9) with zero initial data has a unique solution in the class of functions such that along with derivatives in x up to the second order they belong to $L_p((0,\tau)\times \mathsf{R}^d)$. Furthermore, after writing the term $-\lambda u_t$ as $-\lambda v_t - \lambda z_t$, where $z_t = u_t - v_t$, differentiating the equation once with respect to x and using the contraction property of the heat semigroup we find

$$\|Dv\|^p_{L_p((0,\tau)\times\mathsf{R}^d)} \le \|Dz\|^p_{L_p((0,\tau)\times\mathsf{R}^d)}. \tag{4.10}$$

The solution v_t can be given by an integral formula, which implies that v_t is $\mathcal{F}_t$-adapted. It is also continuous as an L_p-valued process, and hence, is a predictable L_p-valued process. Taking expectations of both parts of (4.10) shows that $v \in \mathcal{W}^1_p(\tau)$ if $z \in \mathcal{W}^1_p(\tau)$.

Now observe that

$$dz_t = \Delta z_t\,dt + g^k_t\,dw^k_t,$$

which by Lemma 4.2 implies that $z \in \mathcal{W}^1_p(\tau)$ and

$$\|Dz\|^p_{\mathbf{L}_p(\tau)} \le N\|g\|^p_{\mathbf{L}_p(\tau)}.$$

Upon combining this with (4.10) and using the fact that $u = v + z$, we come to (4.8). The lemma is proved. □

5. A priori estimates in the general case. First we deal with the case when $\sigma = \nu = 0$.

LEMMA 5.1. *Suppose that $\sigma^{ik} \equiv \nu^k \equiv 0$. Also suppose that Assumptions* 2.1 *and* 2.2 *are satisfied with $\beta \le \beta_0$, where the way to estimate the constant $\beta_0(d,p,\delta) > 0$ is described in the proof. Let $f^j \in \mathbf{L}_p(\tau)$ and $g \in \mathbf{L}_p(\tau)$.*

Then there exist constants $\lambda_0 \ge 1$ and N, depending only on d, p, δ, K, and ε, such that for any $\lambda \ge \lambda_0$ there exists a unique $u \in \mathcal{W}^1_{p,0}(\tau)$ satisfying (1.1) *for $t \le \tau$. Furthermore, this solution satisfies the estimate*

$$\lambda^{1/2}\|u\|_{\mathbf{L}_p(\tau)} + \|Du\|_{\mathbf{L}_p(\tau)} \le N\left(\sum_{i=1}^d \left\|f^i\right\|_{\mathbf{L}_p(\tau)} + \|g\|_{\mathbf{L}_p(\tau)}\right) + N\lambda^{-1/2}\left\|f^0\right\|_{\mathbf{L}_p(\tau)}. \tag{5.1}$$

Proof. Uniqueness and part of estimate (5.1) follow from Lemma 3.3. In the rest of the proof we may assume that τ is bounded and split our argument into two parts.

Case $g^k \equiv 0$. First assume that the coefficients and f^j are nonrandom. We extend the coefficients of L following the example $a^{ij}_t(x) = \delta^{ij}$, $t < 0$, and extend f^j_t beyond $(0,\tau)$ arbitrary only requiring $f^j \in L_p(\mathsf{R}^{d+1})$.

Then by Theorem 4.5 and Remark 2.4 of [8] the equation

$$\frac{\partial}{\partial t}u_t = L_t u_t - \lambda u_t + D_i f^i_t + f^0_t \tag{5.2}$$

in R^{d+1} has a unique solution with finite norms

$$\|u\|_{L_p(\mathsf{R}^{d+1})} \quad \text{and} \quad \|Du\|_{L_p(\mathsf{R}^{d+1})}$$

provided that $\lambda \ge \lambda_0$. By Theorem 4.4 of [8]

(5.3)
$$\lambda^{1/2}\|u\|_{L_p(\mathsf{R}^{d+1})}+\|Du\|_{L_p(\mathsf{R}^{d+1})}\leq N\left(\sum_{i=1}^d\left\|f^i\right\|_{L_p(\mathsf{R}^{d+1})}+\lambda^{-1/2}\left\|f^0\right\|_{L_p(\mathsf{R}^{d+1})}\right).$$

By Theorem 3.1 the function u_t is a continuous L_p-valued function.

The proof of Theorem 4.4 of [8] is achieved on the basis of the a priori estimate (5.3) and the method of continuity by considering the family of equations

$$\frac{\partial}{\partial t}u_t=(\theta L_t+(1-\theta)\Delta)u_t-\lambda u_t+D_if_t^i+f_t^0, \tag{5.4}$$

where the parameter θ changes in $[0,1]$. We remind briefly the method of continuity because we want to show that certain properties of (5.4) which we know for $\theta=0$ propagate from $\theta=0$ to $\theta=1$.

We fix a $\theta_0\in[0,1]$ and to solve (5.4) for given f^j define a sequence of $u^n\in L_p(\mathsf{R},W_p^1)$ by solving the equation

$$\begin{aligned}\frac{\partial}{\partial t}u_t^{n+1}&=(\theta_0L_t+(1-\theta_0)\Delta)u_t^{n+1}-\lambda u_t^{n+1}\\&\quad+D_if_t^i+f_t^0+(\theta-\theta_0)(L_t-\Delta)u^n,\quad n\geq1,\quad u^0=0.\end{aligned} \tag{5.5}$$

If we know that (5.4) is uniquely solvable with θ_0 in place of θ for arbitrary $f^j\in L_p(\mathsf{R}^{d+1})$, then the sequence u^n is well defined. Furthermore, estimate (5.3) easily shows that for θ sufficiently close to θ_0 the $L_p(\mathsf{R},W_p^1)$ norm of $u^{n+1}-u^n$ goes to zero geometrically as $n\to\infty$. In this way, passing to the limit in (5.5), we obtain the solution of (5.4) for θ close to θ_0. Then we can repeat the procedure and starting from $\theta=0$ and moving step by step eventually reach $\theta=1$.

For $\theta=0$ we are dealing with solvability of the heat equation which is proved by giving the solution explicitly by means of the heat semigroup. This representation formula has two important implications:

(i) For any constant $T\in\mathsf{R}$, changing f_t^j for $t\geq T$ does not affect u_t for $t\leq T$;

(ii) If f^j are $L_p(\mathsf{R}^{d+1})$-valued measurable functions of a parameter, say ω from a measurable space, say $(\Omega,\mathcal{F}_T)$, then the solution $u\in L_p(\mathsf{R},W_p^1)$, which now depends on ω is also $\mathcal{F}_T$-measurable.

Property (i) is obtained by inspecting the representation formula. Property (ii) is true because the mapping $L_p(\mathsf{R}^{d+1})\ni f^j\to u\in L_p(\mathsf{R},W_p^1)$ is continuous and, hence, Borel measurable.

Obviously, both properties propagate from $\theta=0$ to $\theta=1$ by the above method of continuity. In particular, solutions of (5.2) on the time interval $(-\infty,T]$ depend only on the values of f_t^j for $t\in(-\infty,T]$. It follows that with the same λ and N, for any $T\in\mathsf{R}$,

$$\begin{aligned}&\lambda^{1/2}\|u\|_{L_p((-\infty,T),L_p)}+\|Du\|_{L_p((-\infty,T),L_p)}\\&\qquad\leq N\left(\sum_{i=1}^d\|f^i\|_{L_p((-\infty,T),L_p)}+\lambda^{-1/2}\|f^0\|_{L_p((-\infty,T),L_p)}\right).\end{aligned} \tag{5.6}$$

From now on, we allow the coefficients and f^j to be random, continue f^j as zero for $t<0$, and solve (5.2) for each ω. By (5.6) with $T=0$ we have that $u_t=0$ for $t\leq0$, and it makes sense considering (5.2) on $(0,T)$ for each $T\in(0,\infty)$ with zero initial condition. In such situation properties (i) and (ii) still hold.

In particular, if f^j are measurable $L_p((0,T),L_p)$-valued functions of a parameter, say ω from a measurable space, say $(\Omega,\mathcal{F}_T)$, then the solution $u\in L_p((0,T),W^1_p)$ is also $\mathcal{F}_T$-measurable. Then, from the equation itself, it follows that (u_T,ϕ) is $\mathcal{F}_T$-measurable for any $\phi\in C_0^\infty$. Since u_T takes values in L_p, it is an L_p-valued $\mathcal{F}_T$-measurable function.

If f^i_t are predictable L_p-valued functions, the above conclusions are valid for any $T\in[0,\infty)$. In particular, u_t is $\mathcal{F}_t$-adapted as an L_p-valued function and since it is continuous, u_t is a predictable L_p-valued function.

These properties and the fact that (5.6) holds for any $T\in(0,\infty)$ and ω prove the lemma in the particular case under consideration.

General case. By Lemma 4.3 there is a unique solution $v\in\mathcal{W}^1_{p,0}(\tau)$ of (4.6). Observe that

$$(L_t-\Delta)v_t=D_i\hat{f}^i_t+\hat{f}^0_t,$$

where $\hat{f}^j_t$ are functions of class $\mathbf{L}_p(\tau)$ defined by

$$\hat{f}^j_t=\left(a^{ij}_t-\delta^{ij}\right)D_iv_t+a^j_tv_t,\quad j=1,\dots,d,$$
$$\hat{f}^0_t=b^i_tD_iv_t+c_tv_t.$$

By the above there is a unique solution $u\in\mathcal{W}^1_{p,0}(\tau)$ of

$$\frac{\partial}{\partial t}u_t=L_tu_t-\lambda u_t+(L_t-\Delta)v_t+D_if^i_t+f^0_t.$$

Obviously, v_t+u_t is a solution of class $\mathcal{W}^1_{p,0}(\tau)$ of (1.1). By the particular case

$$\lambda^{1/2}\|u\|_{\mathbf{L}_p(\tau)}+\|Du\|_{\mathbf{L}_p(\tau)}\le N\sum_{i=1}^d\left(\left\|f^i\right\|_{\mathbf{L}_p(\tau)}+\left\|\hat{f}^i\right\|_{\mathbf{L}_p(\tau)}\right)$$
$$+N\lambda^{-1/2}\left(\left\|f^0\right\|_{\mathbf{L}_p(\tau)}+\left\|\hat{f}^0\right\|_{\mathbf{L}_p(\tau)}\right)$$

and to obtain (5.1), it remains only to use the estimates of v_t provided by Lemma 4.3. The lemma is proved. □

Now we allow $\sigma\neq0$.

LEMMA 5.2. (i) *Suppose that Assumption* 2.1 *is satisfied with* $K=0$ *and take* $\varepsilon\ge\varepsilon_1>0$, $\varepsilon_2\in(0,\varepsilon/4]$, $t_0\ge0$, *and* $x_0\in\mathsf{R}^d$.

(ii) *Let* $f^j\in\mathbf{L}_p(\tau)$, $g\in\mathbf{L}_p(\tau)$, *and* $u\in\mathcal{W}^1_{p,0}(\tau)$ *be such that* (1.1) *holds for* $t\le\tau$. *Assume that* $u_t(x)=0$ *if*

$$(t,x)\notin\Gamma:=(t_0,t_0+\varepsilon_1^2)\times B_{\varepsilon_2}(x_0).$$

(iii) *Assume that the couple* (a,σ) *is* $(\varepsilon,\varepsilon_1)$*-regular at* (t_0,x_0) *with* $\beta=\beta_0/3$ *in* (2.4) *and* (2.5), *where* β_0 *is the constant from Lemma* 5.1. *Also assume that*

$$|\sigma^{i\cdot}_t(x)-\sigma^{i\cdot}_t(x_0)|_{\ell_2}\le\beta_1,\quad(a^{jk}_t(y)-\alpha^{jk}_t(x_0))\xi^j\xi^k\ge\delta|\xi|^2$$

for all values of indices and arguments such that $(t,x)\in\Gamma$ *and* $(t,y)\in Q_\varepsilon(t_0,x_0)$, *where* $\beta_1=\beta_1(d,\delta,p,\varepsilon)>0$ *is a constant an estimate from below for which can be obtained from the proof.*

Then there exist constants $\lambda_0 \geq 1$ and N, depending only on d, p, δ, and ε, such that estimate (5.1) *holds provided that $\lambda \geq \lambda_0$.*

Proof. Without loss of generality we may and will assume that $x_0 = 0$. Also we modify, if necessary, a and σ in such a way that $\sigma_t^{ik}(x) = 0$ if $t \notin (t_0, t_0 + \varepsilon_1^2)$, and $a_t^{ij}(x) = \delta^{-1}\delta^{ij}$ if $t \notin (t_0, t_0 + \varepsilon_1^2)$. Obviously, under this modification assumption (iii) is preserved and (1.1) remains unaffected due to assumption (ii). The rest of the proof we split into two cases.

Case $\sigma_t^{ik}(x) = \sigma_t^{ik}(0)$ for $|x| \leq \varepsilon_2$ and $t \geq 0$. We want to apply Lemma 5.1 and for that, even if $\sigma \equiv 0$, we need a^{ij} to satisfy at least the condition $\operatorname{osc}(a^{ij}, Q) \leq \beta$ for *all* $Q \in \mathbf{Q}$ with $\rho(Q) \leq \varepsilon$. To achieve this we modify $a_t^{ij}(x)$ for $|x| \geq \varepsilon/4$ using the fact that such modifications have no effect on (1.1) since $u_t(x) = 0$ for $|x| \geq \varepsilon_2$ and $\varepsilon_2 \leq \varepsilon/4$.

Take a $\xi \in C_0^\infty(\mathsf{R}^d)$ with support lying in the ball of radius $\varepsilon/2$ centered at the origin and such that $\xi(x) = 1$ for $|x| \leq \varepsilon/4$ and $0 \leq \xi \leq 1$. Set

$$\hat{a}_t^{ij} := \xi a_t^{ij} + \delta^{-1}(1 - \xi)\delta^{ij}.$$

We can use $\hat{a}$ in place of a in (1.1). It follows by Lemma 4.7 of [6] (Itô–Wentzell formula) that the function $v_t(x) := u_t(x + x_t)$ satisfies the equation

$$dv_t(x) = \left(\bar{L}_t v_t(x) - \lambda v_t + D_i \bar{f}_t^i + \bar{f}_t^0\right) dt + \bar{g}_t^k(x + x_t)\, dw_t^k, \tag{5.7}$$

where

$$\begin{aligned}
\bar{L}_t\phi &= D_j\left(\bar{a}_t^{ij} D_i \phi\right), \quad \bar{a}_t^{ij}(x) = \hat{a}_t^{ij}(x + x_t) - \alpha_t^{ij}(0), \\
\bar{f}_t^i(x) &:= f_t^i(x + x_t) - \sigma_t^{ik}(0) g_t^k(x + x_t), \quad i = 1, \ldots, d, \\
\bar{f}_t^0(x) &:= f_t^0(x + x_t), \quad \bar{g}_t^k(x) = g_t^k(x + x_t),
\end{aligned}$$

and the process $x_t = (x_t^1, \ldots, x_t^d)$ is defined by

$$x_t^i = -\int_0^t \sigma_s^{ik}(0)\, dw_s^k.$$

This fact shows that the assertion of the present lemma is a direct consequence of Lemma 5.1 in case the latter is applicable to (5.7).

As is easy to see we will be able to apply Lemma 5.1 to (5.7) if we can find $\varepsilon' = \varepsilon'(d, \delta, \varepsilon, p) > 0$ such that

$$\frac{1}{t - s}\int_s^t \left(\left|\bar{a}_r^{ij} - \bar{a}_{r(B)}^{ij}\right|\right)_{(B)} dr \leq \beta_0, \tag{5.8}$$

whenever $(s, t) \times B \in \mathbf{Q}$ and $\rho(B) \leq \varepsilon'$.

Denote by N, with or without subscripts, various (large) constants depending only on d, δ, and ε and observe that $|D\xi| \leq N$. It follows easily that for $B \in \mathbf{B}$ we have

(5.9)

$$\begin{aligned}
\left(\left|\bar{a}_r^{ij} - \bar{a}_{r(B)}^{ij}\right|\right)_{(B)} &\leq \left(\left|\xi a_r^{ij} - (\xi a_r^{ij})_{(B+x_r)}\right|\right)_{(B+x_r)} + \delta^{-1}\delta^{ij}(|\xi - \xi_{(B+x_r)}|)_{(B+x_r)} \\
&\leq \left(\left|\xi a_r^{ij} - (\xi a_r^{ij})_{(B+x_r)}\right|\right)_{(B+x_r)} + N_1\rho =: I_r + N_1\rho,
\end{aligned}$$

where and below $\rho = \rho(B)$.

Let z be the center of B and set

$$y_r = (z + x_r)(\rho + \varepsilon/2)|z + x_r|^{-1}$$

if $|z + x_r| \geq \rho + \varepsilon/2$ and $y_r = z + x_r$ otherwise. Observe that y_r is continuous in r and

$$|y_r| \leq \rho + \varepsilon/2. \tag{5.10}$$

Next we claim that

$$I_r \leq 2\left(\left|a_r^{ij} - a_{r(B_\rho + y_r)}^{ij}\right|\right)_{(B_\rho + y_r)} + N_2\rho. \tag{5.11}$$

If (5.11) is true, then by combining it with (5.9) and using (5.10) we find that the left-hand side of (5.8) is less than

$$(N_1 + N_2)\rho + 2 \sup_{|y| \leq \rho + \varepsilon/2} \operatorname{osc}\left(a^{ij}, Q_\rho + (0, y), 0\right)$$

if $\sigma_t^{nm}(0) = 0$ for all t, n, m or, in general, less than

$$(N_1 + N_2)\rho + 2\operatorname{Osc}\left(a^{ij}, Q_\rho, \rho + \varepsilon/2\right),$$

where $Q_\rho = (s, t) \times B_\rho$. Now (2.4) and (2.5) imply that (5.8) is satisfied for $\rho \leq \varepsilon'$ if we choose $\varepsilon' > 0$ so that

$$(N_1 + N_2)\varepsilon' \leq \beta_0/3, \quad \varepsilon' \leq \varepsilon/4.$$

Therefore, it remains only to prove the claim. Obviously, if $|z + x_r| \geq \rho + \varepsilon/2$, then $I_r = 0$ and (5.11) holds.

In case $|z + x_r| < \rho + \varepsilon/2$ the estimates

$$(|h_r - h_{r(B')}|)_{(B')} \leq \frac{1}{|B'|^2} \int_{B'} \int_{B'} |h_r(y) - h_r(z)|\, dydz \leq 2(|h_r - h_{r(B')}|)_{(B')},$$

$$\left|\xi(y)a_r^{ij}(y) - \xi(z)a_r^{ij}(z)\right| \leq \xi(y)\left|a_r^{ij}(y) - a_r^{ij}(z)\right| + N|\xi(y) - \xi(z)|$$

show that

$$I_r \leq 2\left(\left|a_r^{ij} - a_{r(B + x_r)}^{ij}\right|\right)_{(B + x_r)} + N\rho,$$

which is equivalent to (5.11). This proves the lemma in the particular case under consideration.

General case. We rewrite the term $\Lambda_t^k u_t + g_t^k$ in (1.1) as $\sigma_t^{ik}(0)D_i u_t + \bar{g}_t^k$ with $\bar{g}_t^k = g_t^k + (\sigma_t^{ik} - \sigma_t^{ik}(0))D_i u_t$ and use the above result to conclude that estimate (5.1) holds with $N = N_1 = N_1(d, p, \delta, \varepsilon)$ if we add to its right-hand side

$$N_2(d, p, \delta, \varepsilon)\beta_1 \|Du\|_{\mathbf{L}_p(\tau)}.$$

By choosing $\beta_1 = \beta_1(d, p, \delta, \varepsilon)$ so that $N_2\beta_1 \leq 1/2$, we get (5.1) with $2N_1$ in place of N_1. The lemma is proved. □

Remark 5.3. If Assumption 2.1 is satisfied with $K = 0$ and a_t^{ij} and σ_t^{ik} depend only on ω and t, then the assertion of Lemma 5.2 is true with $\lambda_0 = 0$ and $N = N(d, p, \delta)$ and without requiring u to have compact support. This fact can be obtained

by following the arguments in section 4.3 of [6]. Even though those arguments are much longer, they allow one to prove a very general result that, roughly speaking, "whatever estimate can be established for solutions of the heat equation in Banach function spaces with norms that are invariant under time dependent shifting of the x coordinate, the same estimate with the same constant also holds for solutions of the parabolic equations with no lower order terms and with the matrix of the second order coefficients depending only on t and dominating (in the matrix sense) the unit matrix" (see [5]).

The next step is to consider equations with lower order terms. The following lemma and its corollary are stated in a slightly more general form than it is needed in the present article. The point is that we intend to use them in a subsequent article about equations in half spaces.

LEMMA 5.4. *Let $G \subset \mathsf{R}^d$ be a domain (perhaps, $G = \mathsf{R}^d$) and take $\varepsilon \geq \varepsilon_1 > 0$ and $\varepsilon_2 \in (0, \varepsilon/4]$.*

(i) *Let $f^j, g \in \mathbf{L}_p(\tau)$ and let $u \in \mathcal{W}^1_{p,0}(\tau)$ satisfy (1.1) for $t \leq \tau$ and be such that $u_t(x) = 0$ if $x \notin G$.*

(ii) *Suppose that Assumption 2.1 is satisfied.*

(iii) *Suppose that assumption (iii) of Lemma 5.2 is satisfied for any $t_0 \geq 0$ and x_0 such that $\operatorname{dist}(x_0, G) \leq \varepsilon_2$.*

Then there exist constants $N, \lambda_0 \geq 0$, depending only on d, p, K, δ, ε, ε_1, and ε_2, such that estimate (5.1) holds true whenever $\lambda \geq \lambda_0$.

Proof. As usual we will use partitions of unity. Take a nonnegative $\xi \in C_0^\infty(B_{\varepsilon_2})$ with unit L_p-norm and take a nonnegative $\eta \in C_0^\infty((0, \varepsilon_1^2))$ with unit L_p-norm. For $s \in \mathsf{R}$ and $y \in \mathsf{R}^d$ introduce

$$\zeta(t,x) = \xi(x)\eta(t), \quad \zeta^{s,y}(t,x) = \zeta(t-s, x-y), \quad u_t^{s,y}(x) = \zeta^{s,y}(t,x)u_t(x)$$

so that, in particular,

$$|u_t(x)|^p = \int_{\mathsf{R}^{d+1}} |u_t^{s,y}(x)|^p \, dyds. \tag{5.12}$$

Observe that, for each s, y,

$$\begin{aligned} du_t^{s,y} = &\left(\sigma_t^{ik} D_i u_t^{s,y} + \hat{g}_t^{s,y,k}\right) dw_t^k \\ &+ \left(D_j\left(a_t^{ij} D_i u_t^{s,y}\right) - \lambda u_t^{s,y} + D_j \hat{f}_t^{s,y,j} + \hat{f}_t^{s,y,0}\right) dt \end{aligned} \tag{5.13}$$

for $t \leq \tau$, where we dropped the argument x (and ω) and

$$\begin{aligned} \hat{g}_t^{s,y,k} &= \zeta^{s,y}\left(\nu_t^k u_t + g_t^k\right) - u_t \sigma_t^{ik} D_i \zeta^{s,y}, \\ \hat{f}_t^{s,y,j} &= \zeta^{s,y}\left(a_t^j u_t + f_t^j\right) - a_t^{ij} u_t D_i \zeta^{s,y}, \quad j = 1, \dots, d, \\ \hat{f}_t^{s,y,0} &= \zeta^{s,y}\left(f_t^0 + b_t^i D_i u_t + c_t u_t\right) - f_t^j D_j \zeta^{s,y} - \left(a_t^{ij} D_i u_t + a_t^j u_t\right) D_j \zeta^{s,y} + \zeta_t^{s,y} u_t, \end{aligned}$$

and $\zeta_t^{s,y}(t,x) = \xi(x-y)\eta'(t-s)$.

As is easy to see $u^{s,y}(t,x) = 0$ for $(t,x) \notin (s_+, s_+ + \varepsilon_1^2) \times B_{\varepsilon_2}(y)$. Therefore, by Lemma 5.2 if $\operatorname{dist}(y, G) \leq \varepsilon_2$, then

$$\begin{aligned} &\lambda^{p/2} \|u^{s,y}\|^p_{\mathbf{L}_p(\tau)} + \|Du^{s,y}\|^p_{\mathbf{L}_p(\tau)} \\ &\qquad \leq N\left(\sum_{j=1}^d \left\|\hat{f}^{s,y,j}\right\|^p_{\mathbf{L}_p(\tau)} + \|\hat{g}^{s,y}\|^p_{\mathbf{L}_p(\tau)}\right) + N\lambda^{-p/2}\left\|\hat{f}^{s,y,0}\right\|^p_{\mathbf{L}_p(\tau)} \end{aligned} \tag{5.14}$$

provided that $\lambda \geq \lambda_0$, where N and λ_0 depend only on d, δ, p, and ε. This estimate also, obviously, holds if $\operatorname{dist}(y, G) > \varepsilon_2$ since then $u_t^{s,y} \equiv 0$.

Next,

$$\begin{aligned}
&\left|\hat{f}_t^{s,y,j}\right| \leq N\bar{\zeta}^{s,y}|u_t| + \zeta^{s,y}\left|f_t^j\right|, \quad j = 1, \dots, d,\\
&\left|\hat{f}_t^{s,y,0}\right| \leq N\bar{\zeta}^{s,y}(|Du_t| + |u_t|) + N\bar{\zeta}^{s,y}\sum_{j=0}^{d}|f_t^j|,\\
&|\hat{g}_t^{s,y}|_{\ell_2} \leq N\bar{\zeta}^{s,y}|u_t| + \zeta^{s,y}|g_t|_{\ell_2},
\end{aligned}$$

where $\bar{\zeta} = \zeta + |D\zeta| + |\zeta_t|$, $\bar{\zeta}^{s,y}(t,x) = \bar{\zeta}(t-s, x-y)$, and here and below we allow the constants N to depend only on d, p, δ, K, ε, ε_1, and ε_2.

We also notice that $|\zeta^{s,y}Du_t| \leq |D(\zeta^{s,y}u_t)| + \bar{\zeta}^{s,y}|u_t|$. Then we find that

$$\begin{aligned}
&\lambda^{p/2}\|\zeta^{s,y}u\|_{\mathbf{L}_p(\tau)}^p + \|\zeta^{s,y}Du\|_{\mathbf{L}_p(\tau)}^p\\
&\leq N\left(\sum_{i=1}^{d}\left\|\bar{\zeta}^{s,y}f^i\right\|_{\mathbf{L}_p(\tau)}^p + \|\zeta^{s,y}g\|_{\mathbf{L}_p(\tau)}^p + \left\|\bar{\zeta}^{s,y}u\right\|_{\mathbf{L}_p(\tau)}^p\right)\\
&\quad + N\lambda^{-p/2}\left(\left\|\bar{\zeta}^{s,y}f^0\right\|_{\mathbf{L}_p(\tau)}^p + \left\|\bar{\zeta}^{s,y}Du\right\|_{\mathbf{L}_p(\tau)}^p\right).
\end{aligned}$$

We integrate through this estimate and use formulas like (5.12). Then we obtain

$$\begin{aligned}
&\lambda^{p/2}\|u\|_{\mathbf{L}_p(\tau)}^p + \|Du\|_{\mathbf{L}_p(\tau)}^p\\
&\leq N_1\left(\sum_{i=1}^{d}\left\|f^i\right\|_{\mathbf{L}_p(\tau)}^p + \|g\|_{\mathbf{L}_p(\tau)}^p + \|u\|_{\mathbf{L}_p(\tau)}^p\right) + N_1\lambda^{-p/2}\left(\left\|f^0\right\|_{\mathbf{L}_p(\tau)}^p + \|Du\|_{\mathbf{L}_p(\tau)}^p\right).
\end{aligned}$$

Finally, we increase $\lambda_0 \geq 0$, if necessary, in such a way that $N_1\lambda^{-p/2} \leq 1/2$ for $\lambda \geq \lambda_0$. Then we obviously arrive at (5.1) with $N = 2N_1$. The lemma is proved. □

To the best of the author's knowledge, the following multiplicative estimate is new even in the deterministic case.

COROLLARY 5.5. *Let $\lambda = 0$. Then under the assumptions of Lemma 5.4 we have*

$$\|Du\|_{\mathbf{L}_p(\tau)} \leq N\left(\sum_{i=1}^{d}\left\|f^i\right\|_{\mathbf{L}_p(\tau)} + \|g\|_{\mathbf{L}_p(\tau)} + \left\|f^0\right\|_{\mathbf{L}_p(\tau)}^{1/2}\|u\|_{\mathbf{L}_p(\tau)}^{1/2} + \|u\|_{\mathbf{L}_p(\tau)}\right),$$

where N depends only on d, p, K, δ, ε, ε_1, and ε_2.

Indeed, take $\lambda > 0$ and add and subtract the term $(\lambda_0 + \lambda)u_t\,dt$ on the right in (1.1), thus introducing λ into the equation and modifying f_t^0 by including into it one of $(\lambda_0 + \lambda)u_t$. Then after applying (5.1), we see that

$$\begin{aligned}
\|Du\|_{\mathbf{L}_p(\tau)} \leq N\Bigg(&\sum_{i=1}^{d}\left\|f^i\right\|_{\mathbf{L}_p(\tau)} + \|g\|_{\mathbf{L}_p(\tau)}\\
&+ (\lambda_0 + \lambda)^{-1/2}\left\|f^0\right\|_{\mathbf{L}_p(\tau)} + (\lambda_0 + \lambda)^{1/2}\|u\|_{\mathbf{L}_p(\tau)}\Bigg).
\end{aligned}$$

Now it remains only to take the inf with respect to $\lambda > 0$.

Proof of Lemma 2.1. By bearing in mind an obvious shifting of time, we see that in the proof of assertions (i)–(iii) we may assume that $s = 0$.

(i) First of all observe that uniqueness of solution of (2.7) is well known even in a much wider class than $\mathcal{W}^1_p(\infty)$.

Let $u_0 \in \mathrm{tr}_0 \mathcal{W}^1_p$, then $u_0 \in W^{1-2/p}$ for almost each ω and there is a unique solution of the heat equation

$$dv_t = \Delta v_t \, dt$$

of class $L_p((0,1), W^1_p)$ with initial condition u_0. Furthermore,

$$\|v\|_{L_p((0,1),W^1_p)} \sim \|u_0\|_{W^{1-2/p}_p}.$$

Next, take a $\zeta \in C_0^\infty(\mathsf{R})$ such that $\zeta_0 = 1$ and $\zeta_t = 0$ for $t \geq 1/2$ and define $\psi_t(x) = e^{-t} v_t(x) \zeta_t$ for $t \in [0,1]$ and as zero if $t \geq 1/2$. Notice that (a.s.)

$$\psi \in L_p\left(\mathsf{R}_+, W^1_p\right),$$

and

$$\frac{\partial}{\partial t} \psi_t = \Delta \psi_t - \psi_t + e^{-t} \zeta_t' v_t.$$

Then it is a classical result that there exists a unique $\phi \in L_p(\mathsf{R}_+, W^2_p)$ which solves the equation

$$d\phi_t = \left(\Delta \phi_t - \phi_t + e^{-t} \zeta_t' v_t\right) dt$$

with *zero* initial condition. In addition,

$$\|\phi\|_{L_p(\mathsf{R}_+,W^2_p)} \leq N \|\zeta' v\|_{L_p(\mathsf{R}_+,L_p)} \leq N \|u_0\|_{W^{1-2/p}_p},$$

where the constants N depend only on d and p. Owing to these estimates and uniqueness, the operators mapping u_0 into v and ϕ are continuous (and nonrandom). Since u_0 is $\mathcal{F}_0$-measurable, the same is true for ψ, ϕ, and $u = \psi - \phi$, which is of class $L_p(\mathsf{R}_+, W^1_p)$, satisfies (2.7), and equals u_0 for $t = 0$. Also, for each ω

$$\|u\|_{L_p(\mathsf{R}_+,W^1_p)} \leq \|\psi\|_{L_p(\mathsf{R}_+,W^1_p)} + \|\phi\|_{L_p(\mathsf{R}_+,W^1_p)} \leq N \|u_0\|_{W^{1-2/p}_p},$$

where N depends only on d and p. By raising the extreme terms to the pth power and taking expectations we get the first inequality in (2.8) and also finish proving the "only if" part of (i).

To prove the "if" part assume that we have a $v \in \mathcal{W}^1_p(\infty)$ satisfying (2.7) and equal u_0 at $t = 0$. Then $u_t = v_t e^t$ satisfies $\partial u_t / \partial t = \Delta u_t$ and is of class $\mathcal{W}^1_p(1)$. It follows that almost all ω we have $u \in L_p((0,1), W^1_p)$, $u_0 \in W^{1-2/p}_p$, and

$$\|u_0\|_{W^{1-2/p}_p} \leq N \|u\|_{L_p((0,1),W^1_p)} \leq N \|v\|_{L_p(\mathsf{R}_+,W^1_p)}.$$

By raising all expressions to the power p and taking expectations we arrive at the second estimate in (2.8). Assertion (i) is proved.

The "only if" part in (ii) is, actually, proved above. To prove the "if" part write

$$dv_t = \left(D_i f^i_t + f^0_t\right) dt + g^k_t \, dw^k_t = \left(\Delta v_t - \lambda v_t + D_i \hat{f}^i_t + \hat{f}^0_t\right) dt + g^k_t \, dw^k_t,$$

where the constant $\lambda > 0$ will be chosen later, $\hat{f}^i_t = f^i_t - D_i v_t$, $i = 1, \dots, d$, $\hat{f}^0_t = f^0_t + \lambda v_t$, and $\hat{f}^j, g \in \mathbf{L}_p(1)$. Next, take the function ζ as above, set $u = v\zeta$, and observe that

$$du_t = \left(\Delta u_t - \lambda u_t + D_i \check{f}^i_t + \check{f}^0_t\right) dt + \check{g}^k_t \, dw^k_t, \tag{5.15}$$

where $\check{f}^0 = \zeta \hat{f}^0 + v\zeta'$, $\check{f}^i_t = \zeta \hat{f}^i_t$, $i = 1, \dots, d$, $\check{g}^k = \zeta g^k$, and $\check{f}^j, \check{g} \in \mathbf{L}_p(\infty)$ and $u \in \mathcal{W}^1_p(\infty)$.

By Lemma 5.1, for λ fixed and large enough (actually, one can take $\lambda = 1$, which is shown by using dilations), (5.15) with zero initial condition admits a unique solution $\psi \in \mathcal{W}^1_p(\infty)$ and

$$\begin{aligned} \|\psi\|_{\mathbf{W}^1_p(\infty)} &\le N \left(\sum_{j=0}^d \left\| \check{f}^j \right\|_{\mathbf{L}_p(\infty)} + \|\check{g}\|_{\mathbf{L}_p(\infty)} \right) \\ &\le N \left(\sum_{j=0}^d \left\| f^j \right\|_{\mathbf{L}_p(1)} + \|g\|_{\mathbf{L}_p(1)} + \|v\|_{\mathbf{W}^1_p(1)} \right). \end{aligned}$$

Then the difference $\phi = u - \psi$ satisfies (2.7), is of class $\mathcal{W}^1_p(\infty)$, and $\phi_0 = u_0$. By assertion (i) we have $u_0 \in \mathrm{tr}_0 \mathcal{W}^1_p$, which proves the "if" part in (ii). Furthermore,

$$\begin{aligned} \|u_0\|_{\mathrm{tr}_0 \mathcal{W}^1_p} &\le N \|\phi\|_{\mathbf{W}^1_p(\infty)} \le N \|u\|_{\mathbf{W}^1_p(\infty)} + N \|\psi\|_{\mathbf{W}^1_p(\infty)} \\ &\le N \|v\|_{\mathbf{W}^1_p(1)} + N \|\psi\|_{\mathbf{W}^1_p(\infty)} \\ &\le N \left(\sum_{j=0}^d \left\| f^j \right\|_{\mathbf{L}_p(1)} + \|g\|_{\mathbf{L}_p(1)} + \|v\|_{\mathbf{W}^1_p(1)} \right). \end{aligned}$$

This proves assertion (iii).

To prove (iv) observe that obvious dilations of the t axis allow us to assume that $s = 1$. Then write (2.2) for $t \in [0, 1]$ and notice that tu_t admits representation (2.2) with new f^j and g^k having simple relations with u_t and the original f^j and g^k. It follows that in the rest of the proof we may assume that $u_0 = 0$.

In that case, take a sufficiently large $\lambda > 0$ and consider the equation

$$dv_t = \left(\Delta v_t - \lambda v_t + D_i \bar{f}^i_t + \bar{f}^0_t\right) dt + \bar{g}^k_t \, dw^k_t$$

for $t \ge 0$ with zero initial condition, where

$$\begin{aligned} \bar{f}^i_t &= f^i_t I_{(0,1)}(t) - D_i u_t I_{(0,1)}(t), \quad i = 1, \dots, d, \\ \bar{f}^0_t &= \left(f^0_t + \lambda u_t\right) I_{(0,1)}(t), \quad \bar{g}^k_t = g^k_t I_{(0,1)}(t). \end{aligned}$$

By uniqueness, $v_t = u_t$ for $t \in [0, 1]$ and by assertion (iii) we have $v_1 \in \mathrm{tr}_1 \mathcal{W}^1_p$. This fact combined with already known estimates of v proves assertion (iv). The lemma is proved. □

6. Proof of Theorem 2.2. Owing to Lemma 2.1 we may assume that we are given a v as in assertion (i) of the lemma. By introducing a new unknown function $\bar{u} = u - v$ we see that u satisfies (1.1) and $u_0 = v_0$ if and only if $\bar{u}_0 = 0$ and

$$d\bar{u}_t = \left(L_t \bar{u}_t - \lambda \bar{u}_t + D_j \bar{f}^j_t + \bar{f}^0_t \right) dt + \left(\Lambda^k_t \bar{u}_t + \bar{g}^k_t \right) dw^k_t,$$

where

$$\bar{f}^j_t = f^j_t - D_j v_t + a^{ij}_t D_i v_t + a^j_t v_t, \quad j = 1, \dots, d,$$
$$\bar{f}^0_t = f^0_t + b^i_t D_i v_t + (c_t - \lambda + 1) v_t,$$
$$\bar{g}^k_t = g^k_t + \sigma^{ik}_t D_i v_t + \nu^k_t v_t.$$

By Lemma 2.1 we have $\bar{f}^j, \bar{g} \in \mathbf{L}_p(\tau)$ and the problem of finding solutions of (1.1) with initial data u_0 is thus reduced to the same problem but with zero initial data.

Furthermore, if estimate (2.10) holds for solutions with zero initial condition, then (for $\lambda \geq \lambda_0$)

$$\begin{aligned}
\lambda^{1/2}\|u\|_{\mathbf{L}_p(\tau)} + \|Du\|_{\mathbf{L}_p(\tau)} - \lambda^{1/2}\|v\|_{\mathbf{L}_p(\tau)} - \|Dv\|_{\mathbf{L}_p(\tau)} & \\
&\leq \lambda^{1/2}\|\bar{u}\|_{\mathbf{L}_p(\tau)} + \|D\bar{u}\|_{\mathbf{L}_p(\tau)} \\
&\leq N\left(\sum_{i=1}^d \|\bar{f}^i\|_{\mathbf{L}_p(\tau)} + \|\bar{g}\|_{\mathbf{L}_p(\tau)}\right) + N\lambda^{-1/2}\left\|\bar{f}^0\right\|_{\mathbf{L}_p(\tau)} \\
&\leq N\left(\sum_{i=1}^d \left\|f^i\right\|_{\mathbf{L}_p(\tau)} + \|g\|_{\mathbf{L}_p(\tau)} + \|v\|_{\mathbf{W}^1_p(\tau)}\right) \\
&\quad + N\lambda^{-1/2}\left(\left\|f^0\right\|_{\mathbf{L}_p(\tau)} + \|v\|_{\mathbf{W}^1_p(\tau)}\right) + N\lambda^{1/2}\|v\|_{\mathbf{L}_p(\tau)},
\end{aligned}$$

which yields (2.10) in full generality.

It follows that while proving (2.10) we may also assume that $u_0 = 0$. Therefore, in the rest of the proof of assertions (i) and (ii) we assume that $u_0 = 0$. Having in mind the substitution $u_t = v_t e^{-\mu t}$, we see that while proving assertion (i) it suffices to concentrate on large λ and prove only the second part of the assertion.

We recall that we suppose that Assumption 2.2 is satisfied with $\beta = \beta_0/3$ and β_0 from Lemma 5.1 and Assumption 2.3 is satisfied with β_1 defined in Lemma 5.2. It follows that assumption (iii) of Lemma 5.2 is satisfied for any (t_0, x_0).

Now we take λ_0 larger than the one in Lemma 3.3 and the one in Lemma 5.4. In that case uniqueness follows from Lemma 3.3. In the proof of existence we will rely on the method of continuity and the a priori estimate (5.1) which is established in Lemma 5.4. For $\lambda \geq \lambda_0$ and $\theta \in [0,1]$ we consider the equation

$$du_t = [(\theta L_t + (1-\theta)\Delta)u_t - \lambda u_t + D_i f^i_t + f^0_t]\,dt + \left(\theta \Lambda^k_t u_t + g^k_t\right)\,dw^k_t. \tag{6.1}$$

We call a $\theta \in [0,1]$ "good" if the assertions of the theorem hold for (6.1). Observe that 0 is a "good" point by Lemma 5.1. Now to prove the theorem it suffices to show that there exists a $\gamma > 0$ such that if θ_0 is a good point, then all points of the interval $[\theta_0 - \gamma, \theta_0 + \gamma] \cap [0,1]$ are "good". So fix a "good" θ_0 and for any $v \in \mathbf{W}^1_p(\tau)$ consider the equation

$$\begin{aligned}
du_t = [(\theta_0 L_t + (1-\theta_0)\Delta)u_t - \lambda u_t + (\theta - \theta_0)(L_t - \Delta)v_t + D_i f^i_t + f^0_t]\,dt & \\
+ \left(\theta_0 \Lambda^k_t u_t + (\theta - \theta_0)\Lambda^k v_t + g^k_t\right)\,dw^k_t. &
\end{aligned} \tag{6.2}$$

Observe that

$$(L_t - \Delta)v_t = D_j\left(\left(a^{ij} - \delta^{ij}\right)D_i v_t + a^j_t v_t\right) + b^i_t D_i v_t + c v_t$$

and recall that $v \in \mathbf{W}^1_p(\tau)$. It follows by assumption that (6.2) has a unique solution $u \in \mathcal{W}^1_{p,0}(\tau)$ $(\subset \mathbf{W}^1_p(\tau))$.

In this way, for f^j and g being fixed, we define a mapping $v \to u$ in the space $\mathbf{W}^1_p(\tau)$. It is important to keep in mind that the image u of $v \in \mathbf{W}^1_p(\tau)$ is always in $\mathcal{W}^1_{p,0}(\tau)$. Take $v', v'' \in \mathbf{W}^1_p(\tau)$ and let u', u'' be their corresponding images. Then $u := u' - u''$ satisfies

$$du_t = [(\theta_0 L_t + (1-\theta_0)\Delta)u_t - \lambda u_t + (\theta - \theta_0)(L_t - \Delta)v_t)]\,dt + \left(\theta_0 \Lambda^k_t u_t + (\theta - \theta_0)\Lambda^k v_t\right)\,dw^k_t,$$

where $v = v' - v''$. It follows by Lemma 5.4 that

$$\|u\|_{\mathbf{W}^1_p(\tau)} \leq N|\theta - \theta_0|\,\|v\|_{\mathbf{W}^1_p(\tau)}$$

with a constant N independent of f, g, v', v'', θ_0, and θ. For θ sufficiently close to θ_0, our mapping is a contraction and, since $\mathbf{W}^1_p(\tau)$ is a Banach space, it has a fixed point. This fixed point is in $\mathcal{W}^1_{p,0}(\tau)$ and, obviously, satisfies (6.1). This proves assertion (i) of the theorem.

Estimate (2.10) is proved above in Lemma 5.4 and assertion (iii) follows from Theorem 3.1. The theorem is proved. □

REFERENCES

[1] Kyeong-Hun Kim, *On stochastic partial differential equations with variable coefficients in C^1 domains*, Stochastic Process. Appl., 112 (2004), pp. 261–283.

[2] Kyeong-Hun Kim, *On L_p-theory of stochastic partial differential equations of divergence form in C^1 domains*, Probab. Theory Related Fields, 130 (2004), pp. 473–492.

[3] Kyeong-Hun Kim, *L_p estimates for SPDE with discontinuous coefficients in domains*, Electron. J. Probab., 10 (2005), pp. 1–20.

[4] N.V. Krylov, *A generalization of the Littlewood-Paley inequality and some other results related to stochastic partial differential equations*, Ulam Quarterly, 2 (1994), pp. 16–26, http://www.ulam.usm.edu/VIEW2.4/krylov.ps

[5] N.V. Krylov, *A parabolic Littlewood-Paley inequality with applications to parabolic equations*, Topological Methods in Nonlinear Analysis, Journal of the Juliusz Schauder Center, 4 (1994), pp. 355–364.

[6] N.V. Krylov, *An analytic approach to SPDEs*, in Stochastic Partial Differential Equations: Six Perspectives, Mathematical Surveys and Monographs, Vol. 64, AMS, Providence, RI, 1999, pp. 185–242.

[7] N.V. Krylov, *On the foundation of the L_p-theory of SPDEs*, in Stochastic Partial Differential Equations and Applications-VII, G. Da Prato, L. Tubaro eds., A Series of Lecture Notes in Pure and Applied Math., Chapman & Hall/CRC, London, 2006, pp. 179–191.

[8] N.V. Krylov, *Parabolic equations with VMO coefficients in Sobolev spaces with mixed norms*, J. Function. Anal., 250 (2007), pp. 521–558.

[9] N.V. Krylov, *Maximum principle for SPDEs and its applications*, in Stochastic Differential Equations: Theory and Applications, A Volume in Honor of Professor Boris L. Rozovskii, P.H. Baxendale, and S.V. Lototsky, eds., Interdisciplinary Mathematical Sciences, Vol. 2, World Scientific, 2007, pp. 311–338.

[10] N.V. Krylov, *Filtering equations for partially observable diffusion processes with Lipschitz continuous coefficients*, The Oxford Handbook of Nonlinear Filtering, Oxford University Press, to appear.

[11] N.V. Krylov, *Lectures on elliptic and parabolic equations in Sobolev spaces*, Amer. Math. Soc., Graduate Studies in Math. 96, Providence, RI, 2008.

[12] N.V. Krylov, *Itô's formula for the L_p-norm of stochastic W^1_p-valued processes*, http://arxiv.org/abs/0806.1557

[13] B.L. Rozovskii, *Stochastic Evolution Systems*, Kluwer, Dordrecht, 1990.

SIAM J. MATH. ANAL.
Vol. 40, No. 6, pp. 2286–2315

SEMILINEAR STOCHASTIC EQUATIONS IN A HILBERT SPACE WITH A FRACTIONAL BROWNIAN MOTION*

T. E. DUNCAN†, B. MASLOWSKI‡, AND B. PASIK-DUNCAN†

Abstract. The solutions of a family of semilinear stochastic equations in a Hilbert space with a fractional Brownian motion are investigated. The nonlinear term in these equations has primarily only a growth condition assumption. An arbitrary member of the family of fractional Brownian motions can be used in these equations. Existence and uniqueness for both weak and mild solutions are obtained for some of these semilinear equations. The weak solutions are obtained by a measure transformation that verifies absolute continuity with respect to the measure for the solution of the associated linear equation. Some examples of stochastic differential and partial differential equations are given that satisfy the assumptions for the solutions of the semilinear equations.

Key words. semilinear stochastic equations, fractional Brownian motion, stochastic partial differential equations, absolute continuity of measures

AMS subject classifications. 60H15, 60G18, 60G15

DOI. 10.1137/08071764X

1. Introduction.

Fractional Brownian motion denotes a family of Gaussian processes with continuous sample paths that are indexed by the Hurst parameter $H \in (0,1)$ and that have properties that appear empirically in a wide variety of physical phenomena, such as hydrology, economic data, telecommunications, and medicine. Since some physical phenomena are naturally modeled by stochastic partial differential equations and the randomness can be described by a fractional Gaussian noise, it is important to study the problems of the solutions of stochastic differential equations in a Hilbert space with a fractional Brownian motion. A significant family of these stochastic equations is the set of semilinear equations, so it is important to investigate the existence and the uniqueness of the solutions of the equations and the sample path properties of the solutions. If primarily only some growth assumptions are made on the nonlinear terms in the semilinear equations, then it is natural to investigate weak solutions, especially those that arise by an absolutely continuous transformation of the measure of the solution of the associated linear stochastic equation.

The study of the solutions of stochastic equations in an infinite-dimensional space with a (cylindrical) fractional Brownian motion (for example, stochastic partial differential equations) has been relatively limited. For the Hurst parameter $H \in (1/2,1)$, linear and semilinear equations with an additive fractional Gaussian noise, the formal derivative of a fractional Brownian motion, are considered in [8, 13, 15, 28]. Random dynamical systems described by such stochastic equations and their fixed points are studied in [22]. A pathwise (or nonprobabilistic) approach is used in [21] to study a parabolic equation with a fractional Gaussian noise where the stochastic term is a nonlinear function of the solution. Strong solutions of bilinear evolution equations

*Received by the editors March 4, 2008; accepted for publication (in revised form) November 11, 2008; published electronically February 20, 2009. Research supported in part by NSF grants DMS 0204669, DMS 0505706, ANI 0124510, and GACR 201/04/0750.
http://www.siam.org/journals/sima/40-6/71764.html
†Department of Mathematics, University of Kansas, Lawrence, KS 66045 (duncan@math.ku.edu, bozenna@math.ku.edu).
‡Institute of Mathematics, Czech Academy of Sciences, Prague, Czech Republic (maslow@math.cas.cz).

with a fractional Brownian motion are considered in [11, 12], and the same type of equation is studied in [33], where a fractional Feynman–Kac formula is obtained. A stochastic wave equation with a fractional Gaussian noise is considered in [2], and a stochastic heat equation with a multiparameter fractional Gaussian noise is studied in [16, 18].

One facet of the motivation for the study of weak solutions in an infinite-dimensional space follows from some results [7, 27] for weak solutions in finite-dimensional spaces that use an absolutely continuous transformation of measures which generalize the result of Girsanov [14] for Brownian motion.

In this paper, a similar analysis is made for infinite-dimensional state spaces. While the structure of the infinite-dimensional Girsanov theorem is analogous to the finite-dimensional case, significant distinct difficulties arise when the application of this theorem is used for stochastic equations in infinite-dimensional spaces. First the driving process is only cylindrical, so the Girsanov theorem can only be used to transform the semilinear equation to a linear equation that is a fractional Ornstein–Uhlenbeck process. Since there is no classical strong solution to the linear equation, the mild solution must be used, making the analysis of the transformation of the measures by the Radon–Nikodým derivative more difficult because a suitable sample path regularity of the Ornstein–Uhlenbeck process must be verified. Unlike the finite-dimensional case, this regularity is not immediate and some assumptions on the coefficients in the linear equation must be made which are known for the case of Brownian motion. The sample regularity requirement increases as the Hurst parameter H increases. Dually the operators that appear in the Radon–Nikodým derivative are less regular as H increases. Thus the applicability of the Girsanov theorem is not immediate in this case and some conditions must be determined for the whole procedure to succeed. Furthermore, for $H > \frac{1}{2}$ in the finite-dimensional case it is assumed that the nonlinear term in the semilinear equation satisfies a global Hölder condition, but this assumption is not satisfied in many typical examples in infinite-dimensional spaces, such as reaction-diffusion equations. Thus this Hölder condition is relaxed here as well.

In section 2, some results from fractional calculus are given, and these results are used to describe a kernel function for an integral operator that provides an isometry of the second moment of Wiener-type stochastic integrals with respect to a fractional Brownian motion and the Lebesgue space of square integrable functions. Furthermore, some recent results for the solution of a linear stochastic equation in a Hilbert space [28] are described. In section 3, semilinear stochastic equations in a Hilbert space are studied. Initially, an absolute continuity of measures result for transforming the solution of a linear stochastic equation is verified that can be viewed as an analogue of the result of Girsanov [14] for a transformation of a finite-dimensional standard Brownian motion. For a semilinear stochastic equation where the nonlinear term satisfies a linear growth condition and some additional conditions are satisfied, it is shown that there is one and only one weak solution. The weak solution is obtained by verifying an absolute continuity of the measure of the solution with respect to the measure of the solution of the associated linear equation. The cases $H \in (0, 1/2)$ and $H \in (1/2, 1)$ are treated separately. Absolute continuity of the above measures is verified when the nonlinearity satisfies a power growth condition and some additional assumptions are made. In section 4, some examples of stochastic differential and partial differential equations are given that satisfy the assumptions of the theorems.

2. Preliminaries. In this section, a cylindrical fractional Brownian motion in a separable Hilbert space is introduced, a Wiener-type stochastic integral with respect to this process is defined, and some basic properties of this integral are noted. Initially, some facts from the theory of fractional integration (cf. [31]) are described. Let $(V, \|\cdot\|, \langle\cdot,\cdot\rangle)$ be a separable Hilbert space. If $\varphi \in L^1([0,T],V)$, then for $\alpha > 0$ the left-side and the right-side fractional (Riemann–Liouville) integrals of φ are defined (for almost all $t \in [0,T]$) by

$$\left(I^{\alpha}_{0+}\varphi\right)(t) = \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1}\,\varphi(s)\,ds$$

and

$$\left(I^{\alpha}_{T-}\varphi\right)(t) = \frac{1}{\Gamma(\alpha)} \int_t^T (s-t)^{\alpha-1}\varphi(s)\,ds,$$

respectively, where $\Gamma(\cdot)$ is the gamma function. For $\alpha \in (0,1)$ the inverse operators of these fractional integrals are called fractional derivatives and can be given by their respective Weyl representations

$$\left(D^{\alpha}_{0+}\psi\right)(t) = \frac{1}{\Gamma(1-\alpha)} \left(\frac{\psi(t)}{t^{\alpha}} + \alpha \int_0^t \frac{\psi(t)-\psi(s)}{(t-s)^{\alpha+1}}\,ds\right)$$

and

$$\left(D^{\alpha}_{T-}\psi\right)(t) = \frac{1}{\Gamma(1-\alpha)} \left(\frac{\psi(t)}{(T-t)^{\alpha}} + \alpha \int_t^T \frac{\psi(s)-\psi(t)}{(s-t)^{\alpha+1}}\,ds\right),$$

where $\psi \in I^{\alpha}_{0+}\left(L^1([0,T],V)\right)$ and $\psi \in I^{\alpha}_{T-}\left(L^1([0,T],V)\right)$, respectively.

Let $K_H(t,s)$ for $0 \le s \le t \le T$ be the real-valued kernel function

$$K_H(t,s) = \frac{\tilde{c}_H (t-s)^{H-\frac{1}{2}}}{\Gamma\left(H+\frac{1}{2}\right)} + \frac{\tilde{c}_H\left(\frac{1}{2}-H\right)}{\Gamma\left(H+\frac{1}{2}\right)} \int_s^t (u-s)^{H-\frac{3}{2}} \left(1-\left(\frac{s}{u}\right)^{\frac{1}{2}-H}\right) du \tag{2.1}$$

for $H \in (0,1/2)$. If $H \in (1/2,1)$, then K_H has a simpler form as

$$K_H(t,s) = \frac{\hat{c}_H}{\Gamma\left(H-\frac{1}{2}\right)} s^{\frac{1}{2}-H} \int_s^t (u-s)^{H-\frac{3}{2}} u^{H-\frac{1}{2}}\,du. \tag{2.2}$$

The terms $\tilde{c}_H$ and $\hat{c}_H$ are constants that depend only on H.

Define the integral operator $\mathbb{K}_H$ induced from the kernel K_H by

$$\mathbb{K}_H\varphi(t) = \int_0^t K_H(t,s)h(s)\,ds \tag{2.3}$$

for $h \in L^2([0,T],V)$. It is well known [31] that

$$\mathbb{K}_H \colon L^2([0,T],V) \to I^{H+\frac{1}{2}}_{0+}\left(L^2([0,T],V)\right)$$

is a bijection and $\mathbb{K}_H$ can be described as

$$\mathbb{K}_H h(s) = \overline{c}_H I^{2H}_{0+}\left(u_{\frac{1}{2}-H} I^{\frac{1}{2}-H}_{0+}\left(u_{H-\frac{1}{2}}h\right)\right)(s) \tag{2.4}$$

for $H \in (0, 1/2]$ and

$$\mathbb{K}_H h(s) = c_H I_{0+}^1 \left(u_{H-\frac{1}{2}} I_{0+}^{H-\frac{1}{2}} \left(u_{\frac{1}{2}-H} h \right) \right)(s) \tag{2.5}$$

for $H \in [1/2, 1)$, where

$$c_H = \left[\frac{2H\Gamma\left(H+\frac{1}{2}\right)\Gamma\left(\frac{3}{2}-H\right)}{\Gamma(2-2H)} \right]^{\frac{1}{2}}, \tag{2.6}$$

$$\overline{c}_H = c_H \Gamma(2H),$$

and

$$u_a(s) = s^a I$$

for $s \geq 0$ and $a \in \mathbb{R}$. The inverse operator

$$\mathbb{K}_H^{-1} \colon I_{0+}^{H+\frac{1}{2}} \left(L^2([0,T],V) \right) \to L^2([0,T],V)$$

is given by

$$\mathbb{K}_H^{-1} \varphi(s) = \overline{c}_H^{-1} s^{\frac{1}{2}-H} D_{0+}^{\frac{1}{2}-H} \left(u_{H-\frac{1}{2}} D_{0+}^{2H} \varphi \right)(s) \tag{2.7}$$

for $H \in (0, 1/2]$ and

$$\mathbb{K}_H^{-1} \varphi(s) = c_H^{-1} s^{H-\frac{1}{2}} D_{0+}^{H-\frac{1}{2}} \left(u_{\frac{1}{2}-H} D\varphi \right)(s) \tag{2.8}$$

for $H \in [1/2, 1)$ and $\varphi \in I_{0+}^{H+\frac{1}{2}} \left(L^2([0,T],V) \right)$. Note that if $\varphi \in H^1([0,T],V)$, the Sobolev space, then

$$\mathbb{K}_H^{-1} \varphi(s) = \overline{c}_H^{-1} s^{H-\frac{1}{2}} I_{0+}^{\frac{1}{2}-H} \left(u_{\frac{1}{2}-H} \varphi' \right)(s) \tag{2.9}$$

for $H \in (0, 1/2]$.

Since the operator $\mathbb{K}_H^{-1}$ plays an important role in what follows, it is desirable to have some information about its domain $I_{0+}^{H+\frac{1}{2}}\left(L^2([0,T],V)\right)$. It is straightforward that $I_{0+}^{H+\frac{1}{2}} \left(L^2([0,T],V) \right) \subset C^\beta([0,T],V)$ for $\beta > \frac{1}{2} - H$ and $H \in (0, 1/2)$. However, in section 3, a more refined result is needed. If $H \in (1/2, 1)$, then $I_{0+}^{H+\frac{1}{2}} \left(L^2([0,T],V) \right) \subset L^2([0,T],V)$.

A definition of the stochastic integral of a deterministic V-valued function with respect to a scalar fractional Brownian motion $(\beta(t), t \geq 0)$ is given. This definition uses the methods in [1, 6, 11, 30]. An alternative, equivalent method is given in [10].

A family of linear operators $(\mathcal{K}_H^*, H \in (0,1))$ is defined which provides an isometry between Wiener-type integrals of a fractional Brownian motion and $L^2([0,T],V)$. It is written as an adjoint because the linear operator $\mathcal{K}_H$ occurs naturally in the factorization of the covariance for a fractional Brownian motion in L^2.

Let $\mathcal{K}_H^* \colon \mathcal{E} \to L^2([0,T],V)$ be the linear map given by

$$\mathcal{K}_H^* \varphi(t) = \varphi(t) K_H(T,t) + \int_t^T (\varphi(s) - \varphi(t)) \frac{\partial K_H}{\partial s}(s,t)\, ds \tag{2.10}$$

for $\varphi \in \mathcal{E}$ and K_H given by (2.1), where $\mathcal{E}$ is the linear space of V-valued step functions on $[0,T]$, that is, $\varphi \in \mathcal{E}$ if

$$\varphi(t) = \sum_{i=1}^{n-1} x_i \mathbb{1}_{[t_i,t_{i+1})}(t),$$

where $x_i \in V$ for $i \in \{1,\dots,n-1\}$ and $0 = t_1 < t_2 < \cdots < t_n = T$.

Define the stochastic integral as

$$\int_0^T \varphi \, d\beta := \sum_{i=1}^{n} x_i \left(\beta(t_{i+1}) - \beta(t_i)\right).$$

It follows directly that

$$\mathbb{E}\left\| \int_0^T \varphi \, d\beta \right\|^2 = |\mathcal{K}_H^* \varphi|^2_{L^2([0,T],V)}, \tag{2.11}$$

where $|\cdot|_{L^2([0,T],V)}$ is the norm in $L^2([0,T],V)$ induced by the inner product. Let $(\mathcal{H}, |\cdot|_{\mathcal{H}}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ be the Hilbert space obtained by the completion of the pre-Hilbert space $\mathcal{E}$ with the inner product

$$\langle \varphi, \psi \rangle_{\mathcal{H}} := \langle \mathcal{K}_H^* \varphi, \mathcal{K}_H^* \psi \rangle_{L^2([0,T],V)} \tag{2.12}$$

for $\varphi, \psi \in \mathcal{E}$. The stochastic integral is extended to an arbitrary $\varphi \in \mathcal{H}$ by the isometry (2.11). Thus $\mathcal{H}$ is a linear space of integrable functions, and it is useful to obtain some more specific information about $\mathcal{H}$. If $H \in (1/2,1)$, then it is easily verified that $\mathcal{H} \supset \tilde{\mathcal{H}}$, where $\tilde{\mathcal{H}}$ is the Banach space of Borel measurable functions with the norm $|\cdot|_{\tilde{\mathcal{H}}}$ given by

$$|\varphi|^2_{\tilde{\mathcal{H}}} := \int_0^T \int_0^T \|\varphi(u)\| \, \|\varphi(v)\| \, \phi_H(u-v) \, du \, dv, \tag{2.13}$$

where $\phi_H(u) = H(2H-1)|u|^{2H-2}$, and it can be verified that $\tilde{\mathcal{H}} \supset L^{\frac{1}{H}}([0,T],V)$ and, in particular, $\tilde{\mathcal{H}} \supset L^2([0,T],V)$ (cf. [12]). If $\varphi \in \tilde{\mathcal{H}}$ and $H > 1/2$, then

$$\mathbb{E}\left\| \int_0^T \varphi \, d\beta \right\|^2 = \int_0^T \int_0^T \langle \varphi(u), \varphi(v) \rangle \, \phi_H(u-v) \, du \, dv. \tag{2.14}$$

If $H \in (0,1/2)$, then the space of integrable functions is smaller than for $H \in (1/2,1)$. For $H \in (0,1/2)$ it is known that $\mathcal{H} \supset H^1([0,T],V)$ (cf. [17, Theorem 5.20]) and $\mathcal{H} \supset C^\beta([0,T],V)$ for each $\beta > 1/2 - H$ (a more specific result is given in the next section). If $H \in (0,1/2)$, then the linear operator $\mathcal{K}_H^*$ can be described by a fractional derivative

$$\mathcal{K}_H^* \varphi(t) = c_H t^{\frac{1}{2}-H} D_{T-}^{\frac{1}{2}-H} \left(u_{H-\frac{1}{2}} \varphi \right)(t), \tag{2.15}$$

where its domain is $\mathcal{H} = I_{T-}^{\frac{1}{2}-H}\left(L^2([0,T],V)\right)$ (cf. [1, Proposition 6]). If $H \in (1/2,1)$, then

$$\mathcal{K}_H^* \varphi(t) = c_H t^{\frac{1}{2}-H} I_{T-}^{H-\frac{1}{2}} \left(u_{H-\frac{1}{2}} \varphi \right)(t). \tag{2.16}$$

A standard cylindrical fractional Brownian motion is defined now.

DEFINITION 2.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space. A cylindrical process $\langle B, \cdot\rangle \colon \Omega \times \mathbb{R}_+ \times V \to \mathbb{R}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ is called a standard cylindrical fractional Brownian motion with the Hurst parameter $H \in (0,1)$ if*

(1) *for each $x \in V \setminus \{0\}$, $\frac{1}{\|x\|}\langle B(\cdot), x\rangle$ is a standard scalar fractional Brownian motion with the Hurst parameter H;*
(2) *for $\alpha, \beta \in \mathbb{R}$ and $x, y \in V$*

$$\langle B(t), \alpha x + \beta y\rangle = \alpha\langle B(t), x\rangle + \beta\langle B(t), y\rangle \quad \text{a.s. } \mathbb{P}.$$

Note that $\langle B(t), x\rangle$ has the interpretation of the evaluation of the functional $B(t)$ at x though the process $B(\cdot)$ does not take values in V.

For $H = 1/2$, this definition is the usual one for a standard cylindrical Wiener process in V. For a complete orthonormal basis $(e_n, n \in \mathbb{N})$ of V, letting $\beta_n(t) = \langle B(t), e_n\rangle$ for $n \in \mathbb{N}$, the sequence of scalar processes $(\beta_n, n \in \mathbb{N})$ is independent and B can be represented by the formal series

$$B(t) = \sum_{n=1}^{\infty} \beta_n(t) e_n \tag{2.17}$$

that does not converge a.s. in V.

Naturally associated with a standard cylindrical fractional Brownian motion is a standard cylindrical Wiener process $(W(t), t \geq 0)$ in V such that, formally, $\dot{B}(t) = \mathcal{K}_H \dot{W}(t)$. For $x \in V \setminus \{0\}$, let $\beta_x(t) = \langle B(t), x\rangle$. It is elementary to verify from (2.1) that there is a scalar Wiener process $(w_x(t), t \geq 0)$ such that

$$\beta_x(t) = \int_0^t K_H(t,s)\, dw_x(s) \tag{2.18}$$

for $t \in \mathbb{R}_+$. Dually, $w_x(t) = \beta_x\big((\mathcal{K}_H^*)^{-1} \mathbb{1}_{[0,t)}\big)$, where $\mathcal{K}_H^*$ is given by (2.15) or (2.16) and $V = \mathbb{R}$. Thus there is a formal expansion of W,

$$W(t) = \sum_{n=1}^{\infty} w_n(t) e_n, \tag{2.19}$$

where $(e_n, n \in \mathbb{N})$ is a complete orthonormal basis for V and $w_n = \langle W, e_n\rangle$ for $n \in \mathbb{N}$.

Now, the stochastic integral $\int_0^T G\, dB$ is defined for a suitable operator-valued function $G \colon [0,T] \to \mathcal{L}(V)$ so that the integral is a V-valued random variable.

DEFINITION 2.2. *Let $G \colon [0,T] \to \mathcal{L}(V)$ be Borel measurable, let $(e_n, n \in \mathbb{N})$ be a complete orthonormal basis in V, let $G(\cdot)e_n \in \mathcal{H}$ for each $n \in \mathbb{N}$, and let B be a standard cylindrical fractional Brownian motion for some fixed $H \in (0,1)$. The stochastic integral $\int_0^T G\, dB$ is defined as*

$$\int_0^T G\, dB := \sum_{n=1}^{\infty} \int_0^T G e_n\, d\beta_n, \tag{2.20}$$

provided the infinite series converges in $L^2(\Omega, V)$.

It is elementary to verify that this definition does not depend on the complete orthonormal basis that is used.

The following proposition describes some $\mathcal{L}(V)$-valued functions G that can be used as integrands in Definition 2.2.

PROPOSITION 2.3. *Let $G\colon [0,T] \to \mathcal{L}(V)$ be Borel measurable and let $G(\cdot)x \in \mathcal{H}$ for each $x \in V$. Let $\Gamma_T\colon V \to L^2([0,T],V)$ be given as*

$$(\Gamma_T x)(t) = (\mathcal{K}_H^* G x)(t) \tag{2.21}$$

for $t \in [0,T]$ and $x \in V$. If $\Gamma_T \in \mathcal{L}_2\left(V, L^2([0,T],V)\right)$, the linear space of Hilbert–Schmidt operators, then the stochastic integral (2.20) is a centered Gaussian V-valued random variable with the covariance operator $\tilde{Q}_T$ given by

$$\tilde{Q}_T x = \int_0^T \sum_{n=1}^{\infty} \langle (\Gamma_T e_n)(s), x\rangle (\Gamma_T e_n)(s)\, ds. \tag{2.22}$$

This integral does not depend on the choice of the complete orthonormal basis $(e_n, n \in \mathbb{N})$ in V.

Proof. Substituting G in the definition of the stochastic integral (2.20), it is clear that the terms of the summation on the right-hand side are V-valued Gaussian random variables by the construction of the integral for a scalar fractional Brownian motion, and the sequence of random variables $\left(\int_0^T G e_n\, d\beta_n, n \in \mathbb{N}\right)$ is independent. Computing the second moment of the tail of the series in (2.20) yields

$$\begin{aligned}\mathbb{E}\left\|\sum_{k=m}^{\infty} G e_k\, d\beta_k\right\|^2 &= \sum_{k=m}^{\infty} \mathbb{E}\left\|\int_0^T G e_k\, d\beta_k\right\|^2 = \sum_{k=m}^{\infty} \int_0^T \left\|(\mathcal{K}_H^* G e_k)(s)\right\|^2 ds \\ &= \sum_{k=m}^{\infty} \int_0^T \left\|(\Gamma_T e_k)(s)\right\|^2 ds = \sum_{k=m}^{\infty} \left|\Gamma_T e_k\right|^2_{L^2([0,T],V)}.\end{aligned}$$

It is clear that this final series tends to zero as m tends to infinity. Thus there is convergence in $L^2(\Omega, V)$ of the partial sums of the infinite series in (2.20).

To verify that (2.20) is a Gaussian random variable and the form of the covariance $\tilde{Q}_T$, initially note that for any $\varphi \in \mathcal{H}$ and $x \in V$, there is the equality

$$\int_0^T \varphi\, d\beta_x = \int_0^T \mathcal{K}_H^* \varphi\, dw_x, \tag{2.23}$$

where w_x is the Wiener process given by (2.18). The terms in the infinite series on the right-hand side of (2.20) are V-valued, independent centered Gaussian random variables with the sequence of covariance operators $\left(\tilde{Q}_T^{(n)}, n \in \mathbb{N}\right)$

$$\tilde{Q}_T^{(n)} x = \int_0^T \langle (\mathcal{K}_H^* G e_n)(s), x\rangle (\mathcal{K}_H^* G e_n)(s)\, ds \tag{2.24}$$

for each $n \in \mathbb{N}$ and $x \in V$. Thus

$$\begin{aligned}\tilde{Q}_T x &= \sum_{n=1}^{\infty} \int_0^T \langle (\mathcal{K}_H^* G e_n)(s), x\rangle (\mathcal{K}_H^* G e_n)(s)\, ds \\ &= \int_0^T \sum_{n=1}^{\infty} \langle (\Gamma_T e_n)(s), x\rangle (\Gamma_T e_n)(s)\, ds.\end{aligned} \tag{2.25}$$

The summability of the infinite series on the right-hand side follows from the Hilbert–Schmidt property of Γ_T. The independence of the stochastic integral from the choice of the complete orthonormal basis follows from (2.23) and the analogous property for stochastic integrals with respect to a standard cylindrical Wiener process. □

Since for almost all $t \in [0,T]$ the linear operator $\Gamma_T(\cdot)(t) : V \to V$ is Hilbert–Schmidt, so we denote for almost all $t \in [0,T]$ the adjoint of $\Gamma_T(\cdot)(t)$ as $\Gamma_T^*(\cdot)(t) : V \to V$. It follows by (2.25) that for $x, y \in V$,

$$\begin{aligned}\left\langle \tilde{Q}_T x, y \right\rangle &= \int_0^T \sum_{n=1}^{\infty} \langle (\Gamma_T e_n)(s), x\rangle \langle (\Gamma_T e_n)(s), y\rangle \, ds \\ &= \int_0^T \sum_{n=1}^{\infty} \langle e_n, (\Gamma_T^* x)(s)\rangle \langle e_n, (\Gamma_T^* y)(s)\rangle \, ds \\ &= \int_0^T \langle (\Gamma_T^* x)(s), (\Gamma_T^* y)(s)\rangle \, ds \\ &= \int_0^T \langle \Gamma_T \Gamma_T^* x(s), y\rangle \, ds.\end{aligned}$$

If $H \in (1/2, 1)$, then $\tilde{Q}_T$ satisfies

$$\tilde{Q}_T = \int_0^T \int_0^T G(u) G^*(v) \phi_H(u-v) \, du \, dv,$$

where $\phi_H(u) = H(2H-1)|u|^{2H-2}$ and G is assumed to satisfy

$$\int_0^T \int_0^T |G(u)|_{\mathcal{L}_2(V)} |G(v)|_{\mathcal{L}_2(V)} \phi_H(u-v) \, du \, dv < \infty$$

(cf. [13, Proposition 2.2]).

The next proposition shows that some densely defined linear operators commute with the stochastic integration.

PROPOSITION 2.4. *If $\tilde{A}$: $\mathrm{Dom}(\tilde{A}) \to V$ is a closed linear operator,* $\mathrm{Dom}(\tilde{A}) \subset V$, *and G: $[0,T] \to \mathcal{L}(V)$ is Borel measurable such that $G([0,T]) \subset \mathrm{Dom}(\tilde{A})$ and both G and $\tilde{A}G$ satisfy the conditions for G in Proposition* 2.3, *then*

$$\int_0^T G \, dB \subset \mathrm{Dom}(\tilde{A}) \quad \textit{a.s. } \mathbb{P}$$

and

$$\tilde{A} \int_0^T G \, dB = \int_0^T \tilde{A} G \, dB \quad \textit{a.s. } \mathbb{P}. \tag{2.26}$$

Proof. By the assumptions on G and $\tilde{A}G$, it follows that $Ge_n \in \mathcal{H}$ and $\tilde{A}Ge_n \in \mathcal{H}$ for $n \in \mathbb{N}$, so by a standard argument using a sequence of step function integrands, the following equality is satisfied:

$$\tilde{A} \int_0^T Ge_n \, d\beta_n = \int_0^T \tilde{A} G e_n \, d\beta_n.$$

Since the sequence of integrals that are obtained from a complete orthonormal basis $(e_n, n \in \mathbb{N})$ are Gaussian random variables it follows that

$$\lim_{m\to\infty} \sum_{n=1}^{m} \int_0^T Ge_n \, d\beta_n = \int_0^T G \, dB \quad \text{a.s. } \mathbb{P} \tag{2.27}$$

and

$$\lim_{m\to\infty} \tilde{A} \left(\sum_{n=1}^{m} \int_0^T Ge_n \, d\beta_n \right) = \lim_{m\to\infty} \sum_{n=1}^{m} \int_0^T \tilde{A}Ge_n \, d\beta_n = \int_0^T \tilde{A}G \, dB \quad \text{a.s. } \mathbb{P}.$$

Since $\tilde{A}$ is a closed linear operator it follows that $\int_0^T G \, dB \in \mathrm{Dom}(\tilde{A})$ a.s. $\mathbb{P}$ and equality (2.26) is satisfied. □

Some results are reviewed for a linear stochastic differential equation with a cylindrical fractional Brownian motion whose solution is often called a fractional Ornstein–Uhlenbeck process. This process is a mild solution of the linear stochastic equation

$$\begin{aligned} dZ(t) &= AZ(t) \, dt + \Phi \, dB(t), \\ Z(0) &= x, \end{aligned} \tag{2.28}$$

where $Z(t)$, $x \in V$, $(B(t), t \geq 0)$ is a standard cylindrical fractional Brownian with $H \in (0,1)$, $\Phi \in \mathcal{L}(V)$, A: $\mathrm{Dom}(A) \to V$, $\mathrm{Dom}(A) \subset V$, and A is the infinitesimal generator of a strongly continuous semigroup $(S(t), t \geq 0)$ on V. A mild solution of (2.28) is

$$\begin{aligned} Z(t) &= S(t)x + \int_0^t S(t-r)\Phi \, dB(r) \\ &= S(t)x + \hat{Z}(t), \end{aligned} \tag{2.29}$$

where the stochastic integral in (2.29) is given by Definition 2.2.

Typically it is assumed that $(S(t), t \geq 0)$ is an analytic semigroup. In this case, there is a $\hat{\beta} \in \mathbb{R}$ such that the operator $\hat{\beta}I - A$ is uniformly positive on V; that is, the resolvent set contains $\{\lambda \in \mathbb{C}; |\arg \lambda| < \pi/2 + \delta\} \cup U$, where $\delta > 0$ and U is a neighborhood of zero.

For each $\delta \geq 0$, $(V_\delta, \|\cdot\|_\delta)$ is a Hilbert space where $V_\delta = \mathrm{Dom}\left((\hat{\beta}I - A)^\delta\right)$ with the graph norm topology so that

$$\|x\|_\delta = \left\| \left(\hat{\beta}I - A\right)^\delta x \right\|.$$

For the mild solution of (2.28), the cases $H \in (0, 1/2)$ and $H \in (1/2, 1)$ have been treated separately [13, 28] because the conditions for similar results are somewhat different. The case $H = 1/2$ (Brownian motion) has been studied extensively (cf. [4]).

For $H \in (1/2, 1)$, the following sample path property of the solution is described in [13].

PROPOSITION 2.5. *If $H \in (1/2, 1)$, $S(t)\Phi \in \mathcal{L}_2(V)$ for each $t > 0$ and*

$$\int_0^{T_0} \int_0^{T_0} u^{-\alpha} v^{-\alpha} |S(u)\Phi|_{\mathcal{L}_2(V)} |S(v)\Phi|_{\mathcal{L}_2(V)} \phi_H(u-v) \, du \, dv < \infty \tag{2.30}$$

for some $T_0 > 0$ and $\alpha > 0$, then there is a Hölder continuous V-valued version of the process $(\hat{Z}(t), t \geq 0)$ with Hölder exponent $\beta < \alpha$, where $\hat{Z}$ is the stochastic convolution in (2.29) and ϕ_H is given in (2.13). If $(S(t), t \geq 0)$ is an analytic semigroup, then there is a version of the process $(\hat{Z}(t), t \in [0,T])$ with $C^{\beta}([0,T], V_{\delta})$ sample paths for each $T > 0$ and $\beta + \delta < \alpha$.

For each $H \in (0,1)$, there are the following results for the sample path behavior of the mild solution [28].

PROPOSITION 2.6. *Let $(S(t), t \geq 0)$ be an analytic semigroup, let $H \in (0,1)$, and let*

$$|S(t)\Phi|_{\mathcal{L}_2(V)} \leq ct^{-\gamma} \tag{2.31}$$

for $t \in [0,T]$, some $c > 0$, and $\gamma \in [0, H)$. Let $\alpha \geq 0$ and $\delta \geq 0$ satisfy

$$\alpha + \delta + \gamma < H. \tag{2.32}$$

Then there is a version of the process $(\hat{Z}(t), t \in [0,T])$ with $C^{\alpha}([0,T], V_{\delta})$ sample paths. If it is assumed instead of (2.31) and (2.32) that $\Phi \in \mathcal{L}_2(V)$ and $\alpha + \delta < H$, then the process $\left(\hat{Z}(t), t \in [0,T]\right)$ has a $C^{\alpha}([0,T], V_{\delta})$ version. In particular, there is a $C^{\alpha}([0,T], V)$ version for $0 < \alpha < H$.

3. Semilinear stochastic equations. In this section, both weak and mild solutions are obtained for various semilinear stochastic equations with a fractional Brownian motion. The cases $H \in (0, 1/2)$ and $H \in (1/2, 1)$ are treated separately as in the case of the linear stochastic equations (Propositions 2.5 and 2.6). The weak solution of a semilinear equation is obtained by an absolutely continuous transformation of the measure for the solution of the associated linear equation. The absolute continuity methods given here are an analogue of the results for the measure of a finite-dimensional fractional Brownian motion [7, 9, 25, 26] and the results for Wiener measure [3,14]. For a fixed $H \in (0,1)$ and $T > 0$, let $(\mathcal{F}_t, t \in [0,T])$ be the filtration for the standard cylindrical fractional Brownian motion $(B(t), t \in [0,T])$ with the Hurst parameter H. The sub-σ-algebra $\mathcal{F}_t \subset \mathcal{F}$ can be generated by $\sigma(\beta_n(s), s \in [0,t], n \in \mathbb{N})$, where $(\beta_n, n \in \mathbb{N})$ is a sequence of independent scalar fractional Brownian motions with the Hurst parameter H that is given in the definition of a standard cylindrical fractional Brownian motion (Definition 2.1).

The following result describes an absolute continuity for a transformation of a standard cylindrical fractional Brownian motion.

THEOREM 3.1. *Let $H \in (0,1)$ and $T > 0$ be fixed and let $(u(t), t \in [0,T])$ be a V-valued, $(\mathcal{F}_t)$-adapted process such that*

1.
$$\int_0^T \|u(t)\| \, dt < \infty \quad \text{a.s. } \mathbb{P}$$

and

2.
$$U(\cdot) := \int_0^{\cdot} u(s) \, ds \in I_{0+}^{H+\frac{1}{2}}\left(L^2([0,T], V)\right) \quad \text{a.s. } \mathbb{P}.$$

Furthermore, it is assumed that

$$\mathbb{E}\xi(T) = 1,$$

where

$$\xi(T) = \exp\left[\int_0^T \left\langle \mathbb{K}_H^{-1}(U)(t), dW(t)\right\rangle - \frac{1}{2}\int_0^T \left\|\mathbb{K}_H^{-1}(U)(t)\right\|^2 dt\right], \tag{3.1}$$

where $(W(t), t \in [0,T])$ *is a standard cylindrical Wiener process in* V *given by* (2.19) *and* $\mathbb{K}_H^{-1}$ *is the inverse of the integral operator* $\mathbb{K}_H$ *in* (2.3). *Then the process* $\left(\tilde{B}(t), t \in [0,T]\right)$ *given by*

$$\tilde{B}(t) := B(t) - U(t)$$

is a standard cylindrical fractional Brownian motion in V *with the Hurst parameter* H *on the probability space* $\left(\Omega, \mathcal{F}, \tilde{\mathbb{P}}\right)$, *where*

$$\frac{d\tilde{\mathbb{P}}}{d\mathbb{P}} = \xi(T) \quad a.s. \tag{3.2}$$

Proof. Initially, it is noted that for an $(\mathcal{F}_t)$-adapted process, $(\eta(t), t \in [0,T])$ with $\eta \in L^2([0,T], V)$ a.s. $\mathbb{P}$, $\int_0^T \langle \eta, dW\rangle$ is defined by

$$\int_0^T \langle \eta, dW\rangle = \sum_{n=1}^{\infty} \int_0^T \langle \eta, e_n\rangle \, dw_n,$$

where the sequences $(\beta_n, n \in \mathbb{N})$ and $(w_n, n \in \mathbb{N})$ are related by (2.18). It is shown that $\mathbb{K}_H^{-1}U$ satisfies the conditions of η so that the stochastic integral in (2.20) is well-defined. Recall that the linear operator $\mathbb{K}_H$ given in (2.3) is a bijection

$$\mathbb{K}_H \colon L^2([0,T], V) \to I_{0+}^{H+\frac{1}{2}}\left(L^2([0,T], V)\right),$$

so by assumption 1 in Theorem 3.1, $\mathbb{K}_H^{-1}(U) \in L^2([0,T], V)$ a.s. $\mathbb{P}$. From the definition of $\mathbb{K}_H$, it follows that $\left(\mathbb{K}_H^{-1}(U)(t), t \in [0,T]\right)$ is an $(\mathcal{F}_t)$-adapted process because U is $(\mathcal{F}_t)$-adapted. By the construction of the standard cylindrical Wiener process W, it is a Wiener process with respect to $(\mathcal{F}_t)$ so ξ_T is a well-defined random variable. By a Girsanov theorem for Wiener processes in infinite dimensions (cf. [4, 24]), equality (3.2) defines a probability $\tilde{\mathbb{P}}$ on $(\Omega, \mathcal{F})$ such that

$$\tilde{W}(t) := W(t) - \int_0^t \mathbb{K}_H^{-1}(U)(s)\, ds$$

is a standard cylindrical Wiener process in V. Let

$$\tilde{\beta}_n(t) := \langle B(t), e_n\rangle - \langle U(t), e_n\rangle$$

and

$$\tilde{w}_n(t) = \langle W(t), e_n\rangle - \left\langle \int_0^t \mathbb{K}_H^{-1}(U)(s)\, ds, e_n \right\rangle.$$

It follows that

$$\begin{aligned}
\int_0^t K_H(t,s)\, d\tilde{w}_n(s) &= \int_0^t K_H(t,s)\, dw_n(s) - \int_0^t K_H(t,s)\left\langle \mathbb{K}_H^{-1}(U)(s), e_n\right\rangle ds \\
&= \beta_n(t) - \left\langle \int_0^t K_H(t,s)\left(\mathbb{K}_H^{-1}(U)(s)\right) ds, e_n \right\rangle \\
&= \beta_n(t) - \left\langle \mathbb{K}_H\mathbb{K}_H^{-1}(U)(t), e_n\right\rangle \\
&= \beta_n(t) - \langle U(t), e_n\rangle = \tilde{\beta}_n(t).
\end{aligned} \tag{3.3}$$

Thus $(\tilde{B}(t), t \in [0,T])$ is a standard cylindrical fractional Brownian motion in V with the Hurst parameter H on $\big(\Omega, \mathcal{F}, \tilde{\mathbb{P}}\big)$. □

In this section, the following semilinear stochastic equation is considered:

$$dX(t) = (AX(t) + F(X(t)))\, dt + \Phi\, dB(t), \tag{3.4}$$

where $t \in \mathbb{R}_+$, $X(t)$, $X_0 \in V$ is nonrandom, $(B(t), t \geq 0)$ is a standard cylindrical fractional Brownian motion with the Hurst parameter $H \in (0,1)$, $\Phi \in \mathcal{L}(V)$, $A\colon \mathrm{Dom}(A) \to V$, $\mathrm{Dom}(A) \subset V$, and A is the infinitesimal generator of a strongly continuous semigroup $(S(t), t \geq 0)$ on V. The function $F\colon V \to V$ is nonlinear, and for the applications to stochastic partial differential equations it is more useful to assume that F is defined only on a (dense) subspace of V. So, let $(E, \|\cdot\|_E)$ be a separable Banach space that is continuously embedded in V and $F\colon E \to E$ with $X_0 \in E$. Subsequently, it is assumed that $F\colon E \to E$ is Borel measurable, $\mathrm{Im}(F) \subset \mathrm{Im}(\Phi)$, for $G := \Phi^{-1}F$, $G\colon E \to V$, and

$$\|G(x)\| \leq \hat{k}\,(1 + \|x\|_E^{\rho}) \tag{3.5}$$

and

$$\|F(x)\|_E \leq \hat{k}\,(1 + \|x\|_E^{\rho}) \tag{3.6}$$

for each $x \in E$ and some $\rho \geq 1$. Furthermore, it is assumed that there is a constant $\bar{K}$ such that for each pair (x,y) in $\mathrm{Dom}(A)$, there is a $z^* \in \partial\|z\|_E$ such that

$$\langle Ax - Ay + F(x) - F(y), z^* \rangle_{E,E^*} \leq \bar{K}\|x - y\|_E, \tag{3.7}$$

where $\partial\|z\|_E$ is the subdifferential of the norm $\|z\|_E$ at the point $z = x - y$ and $\langle \cdot, \cdot, \rangle_{E,E^*}$ is the pairing between E and E^*. The basic results on subdifferentials can be found in [32]. Inequality (3.7) is a one-sided growth condition that ensures the absence of explosions of solutions of (3.4) in a finite time. Some subsequent examples should clarify its interpretation.

The notions of a weak and a mild solution of (3.4) are given now.

Definition 3.2. *A weak solution of* (3.4) *is a triple* $(X(t), B(t), (\tilde{\Omega}, \tilde{\mathcal{F}}, (\tilde{\mathcal{F}}_t), \tilde{\mathbb{P}}), t \geq 0)$, *where* $(B(t), t \geq 0)$ *is a standard cylindrical fractional Brownian motion in* V *that is defined on the probability space* $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$, $(B(t), t \geq 0)$ *and* $(X(t), t \geq 0)$ *are adapted to the filtration* $(\tilde{\mathcal{F}}_t)$, *and* $(X(t), t \geq 0)$ *is an* E*-valued process satisfying*

$$X(t) = S(t)X_0 + \int_0^t S(t-r)F(X(r))\, dr + \int_0^t S(t-r)\Phi\, dB(r). \tag{3.8}$$

A mild solution, $(X(t), t \geq 0)$ *of* (3.4), *is an* E*-valued process on a fixed probability space* $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$ *with a given standard cylindrical fractional Brownian motion that is the fractional Brownian motion in* (3.8), $(B(t), t \geq 0)$ *and* $(X(t), t \geq 0)$ *are adapted to the filtration* $(\mathcal{F}_t)$, *and the process* $(X(t), t \geq 0)$ *satisfies* (3.8).

Equation (3.8) *has a unique weak solution if, for any two weak solutions* $(X(t), B(t), (\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P}), t \geq 0)$ *and* $(\tilde{X}(t), \tilde{B}(t), (\tilde{\Omega}, \tilde{\mathcal{F}}, (\tilde{\mathcal{F}}_t), \tilde{\mathbb{P}}), t \geq 0)$, *the processes* $(X(t), t \geq 0)$ *and* $(\tilde{X}(t), t \geq 0)$ *have the same probability law.*

The equation has a unique mild solution if, for any two processes $(X_1(t), t \geq 0)$ *and* $(X_2(t), t \geq 0)$ *that satisfy* (3.8) *on the same probability space* $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$ *with the same standard cylindrical fractional Brownian motion,* $\mathbb{P}\,(X_1(t) = X_2(t), t \geq 0) = 1$.

A primary goal in this section is to verify weak existence and weak uniqueness of a solution of (3.4). Note that (H1) alone is not sufficient to ensure that the stochastic convolution has values in E. While the assumption (H2) is given in a rather general form, it is verified for particular examples in section 4. Since the cases $H \in (0, 1/2)$ and $H \in (1/2, 1)$ require different methods, they are treated separately.

The following three assumptions are made to construct a solution of (3.4):

(H1) The semigroup $(S(t), t \geq 0)$ generated by A is analytic on V and for each $t \geq 0$, $S(t)|_E \in \mathcal{L}(E)$ and $\| S(t)|_E \|_{\mathcal{L}(E)}$ is bounded on compact time intervals.

(H2) $\Phi \in \mathcal{L}(V)$ is injective and for $T > 0$ the stochastic convolution process

$$\left(\int_0^t S(t-r)\Phi \, dB(r), t \in [0,T] \right)$$

has a version with $C([0,T], E)$ sample paths.

(H3) The function $F\colon E \to V$ in (3.4) is Borel measurable, $\mathrm{Im}(F) \subset \mathrm{Im}(\Phi)$, and the function $G = \Phi^{-1}F\colon E \to V$ satisfies

$$\|G(x)\| \leq k\,(1 + \|x\|_E) \tag{3.9}$$

for some $k > 0$ and all $x \in E$.

The following result verifies a weak solution for $H \in (0, 1/2)$.

THEOREM 3.3. *If $H \in (0, 1/2)$ and the conditions* (H1)–(H3) *are satisfied, then the semilinear equation* (3.4) *has a weak solution. If additionally $F\colon E \to E$ and*

$$\|F(x)\|_E \leq k_1\,(1 + \|x\|_E) \tag{3.10}$$

for some $k_1 > 0$ and all $x \in E$, then the weak solution is unique.

Proof. Initially, existence of a weak solution is verified. By a standard method that has been used for equations of the form (3.4) with a standard cylindrical Brownian motion (cf. [4, 24]), it suffices to verify that the cylindrical process

$$\tilde{B}(t) = B(t) - \int_0^t G(Z(s))\, ds$$

is a standard cylindrical fractional Brownian motion in a suitable probability space where

$$Z(t) = S(t)X_0 + \tilde{Z}(t)$$

satisfies the associated linear equation. To use Theorem 3.1 it is necessary to verify that $G = \Phi^{-1}F$ satisfies the conditions of U in this theorem, that is,

$$\int_0^{\cdot} G(Z(s))\, ds \in I_{0+}^{H+\frac{1}{2}} \left(L^2([0,T], V) \right) \tag{3.11}$$

and

$$\mathbb{E} \exp\left[\rho(Z)\right] = 1, \tag{3.12}$$

where

$$\rho(Z) = \int_0^T \left\langle \mathbb{K}_H^{-1} \left(\int_0^{\cdot} G(Z) \right))(t), dW(t) \right\rangle - \frac{1}{2} \int_0^T \left\| \mathbb{K}_H^{-1} \left(\int_0^{\cdot} G(Z) \right)(t) \right\|^2 dt, \tag{3.13}$$

$\mathbb{K}_H^{-1}$ is the inverse of $\mathbb{K}_H$ in (2.3), and $(W(t), t \geq 0)$ is a standard cylindrical Wiener process in V by (2.19).

From (2.9), it follows that

$$
\begin{aligned}
(3.14)\qquad & \left| \mathbb{K}_H^{-1} \left(\int_0^{\cdot} G(Z) \right) \right|^2_{L^2([0,T],V)} \\
&= \bar{c}_H^{-2} \left| u_{H-\frac{1}{2}} I_{0+}^{\frac{1}{2}-H} \left(u_{\frac{1}{2}-H} G(Z) \right) \right|^2_{L^2([0,T],V)} \\
&= \hat{c}_H \int_0^T \left(s^{H-\frac{1}{2}} \left\| \int_0^s r^{\frac{1}{2}-H} (s-r)^{-\frac{1}{2}-H} G(Z(r))\, dr \right\| \right)^2 ds \\
&\leq \hat{c}_H k^2 \left(1 + |\tilde{Z}|_{C([0,T],E)} + \sup_{t\in[0,T]} \|S(t)X_0\|_E \right)^2 \int_0^T s^{2H-1} \\
&\quad \cdot \left(\int_0^s r^{\frac{1}{2}-H} (s-r)^{-\frac{1}{2}-H}\, dr \right)^2 ds \\
&\leq c_T \left(1 + |\tilde{Z}|^2_{C([0,T],E)} \right)
\end{aligned}
$$

for some $c_T > 0$ that depends only on T. This inequality verifies (3.11). By (3.14) it follows directly that

$$
(3.15)\qquad \mathbb{E} \exp \left[\hat{k} \int_0^T \left\| K_H^{-1} \left(\int_0^{\cdot} G(Z) \right)(t) \right\|^2 dt \right] \leq c \mathbb{E} \exp \left[\hat{k} c_T |\tilde{Z}|^2_{C([0,T],E)} \right]
$$

for some $c > 0$. Substituting $v = \frac{r}{s}$ in the integral with respect to r on the right-hand side of (3.14), it easily follows that $c_T \downarrow 0$ as $T \downarrow 0$. Since $\tilde{Z}$ is a $C([0,T],E)$-valued Gaussian random variable, it follows that

$$
(3.16)\qquad \mathbb{E} \exp \left[\hat{k} c_T |\tilde{Z}|^2_{C([0,T],E)} \right] < \infty
$$

is satisfied for $T > 0$ sufficiently small by the Fernique inequality. Clearly, (3.16) is the Novikov condition [19] which implies the equality (3.12) for $T > 0$ sufficiently small. For arbitrary $T > 0$, a simple iteration verifies the result, that is,

$$
(3.17)\qquad \mathbb{E} \exp \left[\hat{k} \int_{T_{m-1}}^{T_m} \left\| K_H^{-1} \left(\int_0^{\cdot} G(Z) \right)(t) \right\|^2 dt \right] < \infty
$$

for a sufficiently fine partition $0 = T_0 < T_1 < \cdots < T_n = T$. Using a downward induction procedure from the well-known proofs of the martingale property for the Radon–Nikodým derivative in (3.12) for an arbitrary $T > 0$ (see, e.g., [20, Example 6.2.3]), the verification of the equality in (3.12) is obtained.

Now, uniqueness of the weak solution is verified. Uniqueness in law can be proved in a standard way by removing the term F in (3.4) by absolute continuity of measures, which is a suitable inverse of the above construction of a weak solution.

Let $(\tilde{X}(t), t \in [0,T])$ be a solution to the equation

$$
(3.18)\qquad \tilde{X}(t) = S(t)x_0 + \int_0^t S(t-r)F(X(r))\, dr + \tilde{Z}(t),
$$

where $\tilde{Z}(t) = \int_0^t S(t-r)\Phi\, dB(r)$ and $(B(t), t \in [0,T])$ is some standard cylindrical fractional Brownian motion on a probability space $\left(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}}\right)$.

The process $(\tilde{X}(t), t \in [0,T])$ is defined on the same probability space as $(B(t), t \in [0,T])$. Let $(W(t), t \in [0,T])$ be the Wiener process associated with $(B(t), t \in [0,T])$ by (2.18). It suffices to show that

(3.19)

$$\exp\left[\tilde{\rho}(\tilde{X})\right] := \exp\left[-\int_0^T \left\langle \mathbb{K}_H^{-1}\left(\int_0^{\cdot} G(\tilde{X})\right)(t), dW(t)\right\rangle - \frac{1}{2}\int_0^T \left\|\mathbb{K}_H^{-1}\left(\int_0^{\cdot} G(\tilde{X})\right)(t)\right\|^2 dt\right]$$

is a Radon–Nikodým derivative on $\left(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}}\right)$, so $\tilde{\mathbb{P}}$ is the measure for a fractional Ornstein–Uhlenbeck process and uniqueness in law follows. Thus it is necessary to show that

$$\int_0^{\cdot} G\left(\tilde{X}(s)\right) ds \in I_{0+}^{H+\frac{1}{2}}\left(L^2([0,T], V)\right) \tag{3.20}$$

and

$$\tilde{\mathbb{E}} \exp\left[\tilde{\rho}(\tilde{X})\right] = 1, \tag{3.21}$$

where $\tilde{E}$ is integration with respect to $\tilde{\mathbb{P}}$. The verifications of (3.20) and (3.21) are analogous to the verifications of (3.11) and (3.12), respectively. However, since $\tilde{X}$ is not a Gaussian process, the Fernique inequality cannot be used directly. Initially, it is verified that there is a $c > 0$ such that

$$|\tilde{X}|_{C([0,T],E)} \le c\left(1 + \|X_0\|_E + |\tilde{Z}|_{C([0,T],E)}\right), \tag{3.22}$$

where $\tilde{Z}$ is the stochastic process described in (H2). Let

$$\begin{aligned} u(t) &= \tilde{X}(t) - \tilde{Z}(t) \\ &= S(t)X_0 + \int_0^t S(t-r)F\left(u(r) + \tilde{Z}(r)\right) dr. \end{aligned}$$

Thus

$$\|u(t)\|_E \le c_1\|X_0\| + c_2\int_0^t \left(1 + \|u(r)\|_E + \|\tilde{Z}(r)\|_E\right) dr \tag{3.23}$$

for some positive constants c_1 and c_2. By the Gronwall lemma it follows that

$$\|u(t)\|_E \le c_1\left(1 + \|X_0\|_E + |\tilde{Z}|_{C([0,T],E)}\right) \tag{3.24}$$

for $t \in [0,T]$, so the inequality (3.22) is verified. The exponential that usually occurs in the Gronwall inequality is bounded by $e^{c_2 T}$. Making the analogous computations in (3.14), it follows that

(3.25)

$$\left|\mathbb{K}_H^{-1}\left(\int_0^{\cdot} G(Z)\right)(s)\right|^2_{L^2([0,T],V)} \le c_T\left(1 + |X|^2_{C([0,T],E)}\right) \le \tilde{c}_T\left(1 + |\tilde{Z}|^2_{C([0,T],E)}\right),$$

where $\tilde{c}_T \downarrow 0$ as $T \downarrow 0$, so (3.20) is satisfied. Thus the method in (3.15)–(3.17) can be used to verify (3.21).

The random variable $\exp(\tilde{\rho}(\tilde{X}))$ in (3.19) is a Radon–Nikodým derivative and it defines a probability measure $\mathbb{Q}$ on $\tilde{\Omega}$. By this Girsanov-type theorem the process defined by $\tilde{B}(t) = B(t) + \int_0^t u(s)ds$, where $u(s) = G(\tilde{X}(s))$, is a standard cylindrical fractional Brownian motion with respect to the measure $\mathbb{Q}$. Let $(\tilde{W}(t), t \in [0,T])$ be the Wiener process associated with $(\tilde{B}(t), t \in [0,T])$. Let $U(s) = \mathbb{K}_H^{-1}(\int_0^{\cdot} G(\tilde{X}))(s)$ for $s \in [0,T]$ and let $\mathbb{E}$ and $\mathbb{E}_{\mathbb{Q}}$ denote the expectations with respect to the measures $\tilde{\mathbb{P}}$ and $\mathbb{Q}$, respectively. For a bounded measurable function Ψ on $C([0,T],V)$ it follows that

$$\begin{aligned}
\mathbb{E}[\Psi(\tilde{X})] &= \int_{\tilde{\Omega}} \Psi \frac{d\tilde{\mathbb{P}}}{d\mathbb{Q}} d\mathbb{Q} = \mathbb{E}_{\mathbb{Q}}[\Psi(\tilde{X}) \exp(-\tilde{\rho}(\tilde{X}))] \\
&= \mathbb{E}_{\mathbb{Q}} \left[\Psi(\tilde{X}) \exp\left(\int_0^T \langle U(r), d\tilde{W}(r) \rangle - \frac{1}{2} \int_0^T \|U(r)\|^2 dr \right) \right] \\
&= \mathbb{E}_{\mathbb{Q}} \left[\Psi(S(\cdot)X_0 + \int_0^{\cdot} S(\cdot - r) d\tilde{B}(r)) \exp\left(\int_0^T \left\langle \mathbb{K}_H^{-1} \left(\int_0^{\cdot} G\left(S(\cdot)X_0 \right.\right.\right.\right. \\
&\qquad \left.\left.\left.\left. + \int_0^{\cdot} S(\cdot - r)\Phi d\tilde{B}(r) \right) \right)(s), d\tilde{W}(s) \right\rangle \right. \\
&\quad \left. - \frac{1}{2} \int_0^T \left\| \mathbb{K}_H^{-1} \left(\int_0^{\cdot} G\left(S(\cdot)X_0 + \int_0^{\cdot} S(\cdot - r)\Phi d\tilde{B}(r) \right) \right)(s) \right\|^2 ds \right) \right].
\end{aligned}$$

Since the processes $\tilde{W}$ and $\tilde{B}$ are standard cylindrical Brownian motions and standard cylindrical fractional Brownian motions, respectively, the final expectation on the right-hand side above does not depend on the realization of $\tilde{X}$, so the uniqueness in law is verified. □

Now the existence and the uniqueness of a weak solution of (3.4) is verified for $H \in (1/2, 1)$.

THEOREM 3.4. *If $H \in (1/2,1)$,* (H1)–(H3) *are satisfied, and*

$$\|G(x) - G(y)\| \le k_G \|x - y\|^{\gamma} \tag{3.26}$$

for all $x, y \in E$, some $\gamma \in (0,1]$, $k_G > 0$, and $\tilde{Z} \in C^{\beta}([0,T],V)$ for some β satisfying

$$\beta > \frac{H - \frac{1}{2}}{\gamma}, \tag{3.27}$$

where $\tilde{Z}$ is the stochastic convolution process in (H2)*, then* (3.4) *has a weak solution. If, additionally,* (3.10) *is satisfied, then the weak solution is unique.*

Proof. Initially, the existence of a solution is verified as in the proof of Theorem 3.3. It is shown that

$$\int_0^{\cdot} G(Z(s))\, ds \in I_{0+}^{H+\frac{1}{2}} \left(L^2([0,T],V) \right) \quad \text{a.s.} \tag{3.28}$$

and

$$\mathbb{E} \exp[\rho(Z)] = 1, \tag{3.29}$$

where ρ is given by (3.13). By (2.8) it follows that

$$
\begin{aligned}
(3.30)\qquad & \left|\mathbb{K}_H^{-1}\left(\int_0^{\cdot} G(Z)\right)\right|^2_{L^2([0,T],V)} \\
&= c_H^{-2}\left|u_{H-\frac{1}{2}} D_{0+}^{H-\frac{1}{2}}\left(u_{\frac{1}{2}-H} G(Z)\right)\right|^2_{L^2([0,T],V)} \\
&= c_H^{-2}\int_0^T \left\| \frac{s^{H-\frac{1}{2}}}{\Gamma\left(\frac{3}{2}-H\right)}\left(\frac{s^{\frac{1}{2}-H}G(Z(s))}{s^{H-\frac{1}{2}}}\right)\right. \\
&\quad \left. + \left(H-\frac{1}{2}\right)\int_0^s \frac{s^{\frac{1}{2}-H}G(Z(s)) - r^{\frac{1}{2}-H}G(Z(r))}{(s-r)^{H+\frac{1}{2}}}\,dr\right\|^2 ds \\
&\le c\int_0^T \left(s^{\frac{1}{2}-H}\,\|G(Z(s))\| + s^{H-\frac{1}{2}}\int_0^s \frac{s^{\frac{1}{2}-H} - r^{\frac{1}{2}-H}}{(s-r)^{H+\frac{1}{2}}}\|G(Z(r))\|\,dr\right. \\
&\quad \left. + \int_0^s \frac{\|G(Z(s)) - G(Z(r))\|}{(s-r)^{H+\frac{1}{2}}}\,dr\right)^2 ds.
\end{aligned}
$$

Using (3.9) and (3.26), the analyticity of the semigroup $S(\cdot)$ on V, and the inequality

$$
\int_0^s \frac{s^{\frac{1}{2}-H} - r^{\frac{1}{2}-H}}{(s-r)^{H+\frac{1}{2}}}\,dr \le cs^{1-2H},
$$

where c is a generic constant, it follows that

(3.31)

$$
\begin{aligned}
&\left|\mathbb{K}_H^{-1}\left(\int_0^{\cdot} G(Z)\right)\right|_{L^2([0,T],V)} \\
&\quad \le c_T\left(1 + \|X_0\|_E^2 + |\tilde{Z}|^2_{C([0,T],E)}\right) \\
&\quad + c_T\int_0^T \left[\left(\int_0^s \frac{\|S(s)X_0 - S(r)X_0\|^\gamma}{(s-r)^{H+\frac{1}{2}}}\,dr\right)^2 + \left(\int_0^s \frac{\|\tilde{Z}(s) - \tilde{Z}(r)\|^\gamma}{(s-r)^{H+\frac{1}{2}}}\,dr\right)^2\right] ds \\
&\quad \le c_T\left(1 + \|X_0\|_E^2 + |\tilde{Z}|^2_{C([0,T],E)}\right) + c_T\int_0^T \left(\|X_0\|^\gamma \int_0^s \frac{(s-r)^{\gamma\lambda}}{r^{\gamma\lambda}(s-r)^{H+\frac{1}{2}}}\,dr\right)^2 \\
&\quad + c_T\int_0^T |\tilde{Z}|^2_{C^\beta([0,T],V)}\left(\int_0^s \frac{(s-r)^{\gamma\beta}}{(s-r)^{H+\frac{1}{2}}}\,dr\right)^2 ds,
\end{aligned}
$$

where $\lambda > 0$ satisfies $\gamma\lambda < 1$ and $H + 1/2 - \gamma\lambda < 1$. The first integral term in the initial inequality in (3.31) is obtained by the analyticity of the semigroup $S(.)$ on V, which implies that

$$
||(S(s) - S(r))x|| \le c(s-r)^\lambda r^\lambda\ ||x||, \quad x \in V,\ \lambda \ge 0,\ 0 < r < s \le T,
$$

for some $c = c(\lambda)$. It follows that

(3.32)
$$\left|\mathbb{K}_H^{-1}\left(\int_0^{\cdot} G(Z)\right)\right|^2_{L^2([0,T],V)} \leq c_T\left(1 + \|X_0\|_E^2 + \|\tilde{Z}\|^2_{C([0,T],E)} + |\tilde{Z}|^2_{C^\beta([0,T],V)}\right),$$

where $c_T \downarrow 0$ as $T \downarrow 0$, so (3.28) is verified and by the Fernique inequality (3.29) is also verified.

Now the uniqueness of the weak solution is verified. Let $(\tilde{X}(t), t \in [0,T])$ be the solution to (3.4) on a probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}.\tilde{\mathbb{P}})$. As in the proof of Theorem 3.3, it is shown that

$$\int_0^{\cdot} G(\tilde{X})\, ds \in I_{0+}^{H+\frac{1}{2}}\left(L^2([0,T],V)\right) \quad \text{a.s.} \tag{3.33}$$

and

$$\tilde{\mathbb{E}} \exp\left[\tilde{\rho}(\tilde{X})\right] = 1, \tag{3.34}$$

where $\tilde{\rho}$ is given by (3.13). It is necessary to obtain inequality (3.22), used in the proof of Theorem 3.3. Inequality (3.22) is verified by verifying the inequality

$$\|\hat{X}\|_{C^\beta([0,T],V)} \leq L\left(1 + \|X_0\|_E + |\tilde{Z}|_{C([0,T],E)} + |\tilde{Z}|_{C^\beta([0,T],V)}\right), \tag{3.35}$$

where $\hat{X}(t) = \tilde{X}(t) - S(t)X_0$ and $L > 0$. Let $w(t) = \tilde{X}(t) - S(t)X_0 - \tilde{Z}(t)$ for $t \geq 0$. The process w satisfies

$$w(t) = \int_0^t S(t-r)\psi(r)\, dr \tag{3.36}$$

for $t \in [0,T]$, where

$$\psi(t) = F(w(t) + S(t)X_0 + \tilde{Z}(t)).$$

By inequalities (3.10) and (3.35) it follows that $\psi \in L^\infty([0,T],V)$ a.s. $\mathbb{P}$. Since the semigroup $S(\cdot)$ is analytic on V, w is α-Hölder continuous for each $\alpha \in (0,1)$, and using the method of proof of [29, Theorem 4.3.1] there are constants $c_i > 0$ for $i = 1,2$ such that

(3.37)
$$|w|_{C^\beta([0,T],V)} \leq c_1|\psi|_{L^\infty([0,T],V)} \leq c_2\left(|w|_{L^\infty([0,T],E)} + \|X_0\|_E + |\tilde{Z}|_{C([0,T],E)}\right).$$

Thus

$$\begin{aligned} |\tilde{X}|_{C^\beta([0,T],V)} &\leq |w|_{C^\beta([0,T],V)} + |\tilde{Z}|_{C^\beta([0,T],V)} \\ &\leq c_2\left(|w|_{L^\infty([0,T],E)} + \|X_0\|_E + |\tilde{Z}|_{C([0,T],E)} + |\tilde{Z}|_{C^\beta([0,T],V)}\right). \end{aligned} \tag{3.38}$$

Using (3.22) again to bound $|w|_{L^\infty([0,T],E)}$, inequality (3.35) follows.

Now, using the methods for inequalities (3.30)–(3.32), where $Z(t)$ is replaced by $\tilde{X}(t) = S(t)X_0 + \hat{X}(t)$, it follows that

(3.39)
$$\left| \mathbb{K}_H^{-1} \left(\int_0^{\cdot} G(Z) \right) \right|^2_{L^2([0,T],V)} \leq c_T \left(1 + \|X_0\|_E^2 + |\tilde{X}|^2_{C([0,T],V)} + |\tilde{X}|^2_{C^\beta([0,T],V)} \right),$$

where $c_T \downarrow 0$ as $T \downarrow 0$, which by (3.22) and (3.35) verifies (3.33). Equality (3.34) is obtained from (3.39) by the Fernique inequality as in the proof of Theorem 3.3. ∎

Remark 3.5. The proofs of Theorems 3.3 and 3.4 have verified, in addition to weak existence and uniqueness of a solution to (3.4), the mutual absolute continuity (equivalence) of the probability laws of the solution to (3.4) and the solution of (3.4) with $F \equiv 0$ (the fractional Ornstein–Uhlenbeck process) in the path space.

The next objective is to relax the linear growth conditions (3.9) and (3.10) and the Hölder continuity (3.26). The linear growth condition is replaced by a dissipativity condition of the drift term of (3.4), but some other conditions are also imposed so that there is existence and (strong) uniqueness of a mild solution. The main contribution of the following two theorems is a mutual absolute continuity of the probability laws of the solutions of (3.4) with a nonzero F and (3.4) with $F \equiv 0$.

Initially, the case $H \in (0, 1/2)$ is considered.

THEOREM 3.6. *Let $H \in (0, 1/2)$ and let* (H1) *and* (H2) *be satisfied. Let $\Phi \in \mathcal{L}(V)$ be injective, let $\Phi^{-1} \in \mathcal{L}(E, V)$, and let $(S(t)|_E, t \geq 0)$ be a strongly continuous semigroup on E such that*

$$|S(t)|_E|_{\mathcal{L}(E)} \leq e^{\tilde{w}t} \tag{3.40}$$

for $t \geq 0$ and some $\tilde{w} \in \mathbb{R}$. Let $F\colon E \to E$ be continuous and satisfy

$$\|F(x)\|_E \leq k_1 \left(1 + \|x\|_E^\rho\right) \tag{3.41}$$

for $x \in E$ for some $k_1 \geq 0$ and $\rho \geq 1$, and for each pair $x, y \in E$, there is a $z^* \in \partial\|x - y\|_E$ where $\partial\|z\|_E$ is the subdifferential of the norm $\|\cdot\|_E$ at $z \in E$ such that

$$\langle F(x) - F(y), z^* \rangle_{E,E^*} \leq k_2 \|x - y\|_E \tag{3.42}$$

for some $k_2 \in \mathbb{R}$; that is, $F - k_2 I$ is dissipative on E. Then there is one and only one mild solution of (3.4)*, and its probability law on the Borel σ-algebra of $\breve{\Omega} = C([0,T], E)$ is mutually absolutely continuous with respect to the probability law of the fractional Ornstein–Uhlenbeck process* (3.24) *on Ω.*

Proof. Let $(F_\lambda, \lambda > 0)$ be a family of Lipschitz continuous functions from E to E such that each F_λ satisfies inequalities (3.41) and (3.42) for F with the same constants ρ, k_1, k_2. It is shown that there is a $\bar{k} > 0$ depending only on $\tilde{w}$, k_1, and k_2 such that

$$\|v_\lambda(t)\|_E \leq \bar{k} \left(1 + \|X_0\|_E + \|\phi\|^\rho_{C([0,T],E)} \right) \tag{3.43}$$

for $t \in [0,T]$ is satisfied for each $\lambda > 0$ and $\phi \in C([0,T], E)$, where v_λ is a solution of the equation

$$v_\lambda(t) = S(t)X_0 + \int_0^t S(t-r) F_\lambda \left(v_\lambda(r) + \phi(r) \right) dr \tag{3.44}$$

for $t \in [0,T]$.

To verify inequality (3.43), it can be assumed by translation that $k_2 = 0$ in (3.42) (replace F_λ and A by $F_\lambda - k_2 I$ and $A + k_2 I$, respectively). Thus F_λ is dissipative on E for each $\lambda > 0$ and by the assumptions

$$\langle A_E z, z^* \rangle_{E,E^*} \leq \tilde{w} \|z\|_E^2 \tag{3.45}$$

for each $z \in \mathrm{Dom}(A_E)$ and $z^* \in \partial \|z\|_E$, where A_E is the restriction of A to E that generates the semigroup $S(\cdot)|_E$. For each pair $x, y \in \mathrm{Dom}(A_E)$ and $\lambda > 0$, there is a $z_\lambda^* \in \partial \|x - y\|_E$ such that

$$\langle A_E(x-y) + F_\lambda(x) - F_\lambda(y), z_\lambda^* \rangle_{E,E^*} \leq \tilde{w} \|x - y\|_E.$$

By [5, Proposition 5.5.6], there is a sequence $(v_\lambda^n, n \in \mathbb{N})$ such that $v_\lambda^n \in C^1([0,T], E) \cap C([0,T], \mathrm{Dom}(A_E))$ such that $v_\lambda^n \to v_\lambda$ and $\delta_\lambda^n = \frac{d}{dt} v_\lambda^n - A_E v_\lambda^n - F_\lambda\left(v_\lambda^n + \phi\right) \to 0$ in $C([0,T], E)$ as $n \to \infty$. It follows that

$$\begin{aligned}
\frac{d^-}{dt} \left\| v_\lambda^n(t) \right\|_E &\leq \left\langle A_E v_\lambda^n(t) + F_\lambda\left(v_\lambda^n(t) + \phi(t)\right), (v_\lambda^n(t))^* \right\rangle_{E,E^*} + \left\| \delta_\lambda^n(t) \right\|_E \\
&= \left\langle A_E v_\lambda^n(t) + F_\lambda\left(v_\lambda^n(t) + \phi(t)\right) - F_\lambda(\phi(t)), (v_\lambda^n(t))^* \right\rangle \\
&\qquad + \left\langle F_\lambda(\phi(t)), (v_\lambda^n(t))^* \right\rangle_{E,E^*} + \left\| \delta_x^n(t) \right\|_E \\
&\leq \bar{w} \left\| v_\lambda^n(t) \right\|_E + k_2 \left(1 + |\phi|_{C([0,T],E)}^\rho + \left\| \delta_\lambda^n(t) \right\|_E\right)
\end{aligned} \tag{3.46}$$

for $t \in [0,T]$. Using the Gronwall lemma, and letting $n \to \infty$, verifies inequality (3.43).

The mild solution to (3.4) can be expressed as $X(t) = v(t) + \tilde{Z}(t)$, where v satisfies the equation

$$v(t) = S(t)X_0 + \int_0^t S(t-r) F(v(r) + \tilde{Z}(r))\, dr \tag{3.47}$$

for $t \in [0,T]$. Thus the existence and the uniqueness of a mild solution follows from the corresponding pathwise deterministic result (cf. [5, Proposition 5.5.6]).

The equivalence of the probability laws is shown by application of Theorem 3.1. As in the proof of Theorem 3.3, it suffices to show that

$$\int_0^\cdot G(Z(s))\, ds \in I_{0+}^{H+\frac{1}{2}} \left(L^2([0,T], V)\right) \tag{3.48}$$

and

$$\mathbb{E} \exp\left[\rho(Z)\right] = 1, \tag{3.49}$$

where ρ is given by (3.13). While G is not assumed to have at most linear growth as in Theorem 3.3, there is the growth condition

$$\|G(x)\| \leq \hat{k}\left(1 + \|x\|_E^\rho\right) \tag{3.50}$$

for all $x \in E$ and a constant $\hat{k}$. Proceeding as in (3.14), it follows that

(3.51)
$$\begin{aligned}
&\left|\mathbb{K}_H^{-1}\left(\int_0^{\cdot} G(Z)\right)\right|^2_{L^2([0,t],V)} \\
&\quad \le c_1 \int_0^T \left(s^{H-\frac{1}{2}} \left\| \int_0^s r^{\frac{1}{2}-H}(s-r)^{-\frac{1}{2}-H} G(Z(r))\, dr \right\|\right)^2 ds \\
&\quad \le c_2 \left(1 + |\tilde{Z}|^{\rho}_{C([0,T],E)} + \sup_{t\in[0,T]} \|S(t)X_0\|^{\rho}_E\right) \int_0^T s^{2H-1} \\
&\quad \cdot \left(\int_0^s r^{\frac{1}{2}-H}(s-r)^{-\frac{1}{2}-H}\, dr\right)^2 ds \\
&\quad \le c_3 \left(1 + \|X_0\|^{2\rho}_E + |\tilde{Z}|^{2\rho}_{C([0,T],E)}\right)
\end{aligned}$$

for suitable constants c_1, c_2, c_3. This inequality verifies (3.48). To verify equality (3.49), it suffices to assume that F is dissipative (that is, $k_2 = 0$ in (3.42)). Recall that m-dissipative mapping F is defined as a dissipative mapping satisfying $Range(I - \lambda F) = E$ for each $\lambda > 0$; cf. [22]. Since F is continuous, it is m-dissipative (cf. [23]), so the family $(F_\lambda, \lambda > 0)$ of Yosida approximations of F is defined as

$$F_\lambda(x) = F\left(R_\lambda(x)\right) = \frac{1}{\lambda}\left(R_\lambda(x) - x\right) \tag{3.52}$$

for $x \in E$, where

$$R_\lambda(x) = (I - \lambda F)^{-1}(x). \tag{3.53}$$

It is well known that $F_\lambda \colon E \to E$ for $\lambda > 0$ is Lipschitz continuous, so by Theorem 3.3, there is the equality

$$\mathbb{E} \exp\left[\rho_\lambda(Z)\right] = 1 \tag{3.54}$$

for $\lambda > 0$, where

(3.55)
$$\rho_\lambda(Z) = \int_0^T \left\langle \mathbb{K}_H^{-1}\left(\int_0^{\cdot} G_\lambda(Z)\right)(t), dW(t)\right\rangle - \frac{1}{2}\int_0^T \left\|\mathbb{K}_H^{-1}\left(\int_0^{\cdot} G_\lambda(Z)\right)(t)\right\|^2 dt$$

and $G_\lambda := \Phi^{-1} F_\lambda$. As in (3.51), it follows that

(3.56)
$$\begin{aligned}
&\mathbb{E}\left|\mathbb{K}_H^{-1}\left(\int_0^{\cdot} (G_\lambda(Z) - G(Z))\right)\right|_{L^2([0,T],V)} \\
&\quad \le c_T \mathbb{E} \int_0^T \left(s^{H-\frac{1}{2}} \int_0^s r^{\frac{1}{2}-H}(s-r)^{-\frac{1}{2}-H} \left\|G_\lambda(Z(r)) - G(Z(r))\right\| dr\right)^2 ds.
\end{aligned}$$

By some well-known properties of the Yosida approximations and for $x \in E$,

$$\left\|G_\lambda(x) - G(x)\right\| \le \left|\Phi^{-1}\right|_{\mathcal{L}(E,V)} \left\|F_\lambda(x) - F(x)\right\|, \tag{3.57}$$

it follows that $F_\lambda \to F$ as $\lambda \to 0$ and the right-hand side of (3.57) tends to zero as $\lambda \downarrow 0$, and

$$\begin{aligned} \|G_\lambda(x)\| &\le \left|\Phi^{-1}\right|_{\mathcal{L}(E,V)} \|F_\lambda(x)\|_E \\ &\le \left|\Phi^{-1}\right|_{\mathcal{L}(E,V)} \|F(x)\|_E \\ &\le \left|\Phi^{-1}\right|_{\mathcal{L}(E,V)} k_1 \left(1 + \|x\|_E^\rho\right), \end{aligned} \tag{3.58}$$

so the right-hand side of (3.56) tends to zero as $\lambda \downarrow 0$. For a sequence $(\lambda_n, n \in \mathbb{N})$ that decreases to zero, it follows that

$$\lim_{n\to\infty} \exp\left[\rho_{\lambda_n}(Z)\right] = \exp\left[\rho(Z)\right] \quad \text{a.s. } \mathbb{P}. \tag{3.59}$$

To obtain equality (3.54) from equality (3.59) for $\lambda_n, n \in \mathbb{N}$, it is necessary and sufficient to show that the sequence $(\exp\left[\rho_{\lambda_n}(Z)\right], n \in \mathbb{N})$ is uniformly integrable. A sufficient condition for this uniform integrability is to verify that

$$\sup_n \mathbb{E}\left[\left(\exp\left[\rho_{\lambda_n}(Z)\right]\right) \left|\log\left(\exp\left[\rho_{\lambda_n}(Z)\right]\right)\right|\right] = \sup_n \mathbb{E}\left[\left(\exp\left[\rho_{\lambda_n}(Z)\right]\right) \rho_{\lambda_n}(Z)\right] < \infty. \tag{3.60}$$

By Theorem 3.3,

$$\mathbb{E}\left[\rho_{\lambda_n}(Z) \exp\left[\rho_{\lambda_n}(Z)\right]\right] \le \tilde{\mathbb{E}}_{\lambda_n}\left[2\int_0^T \left\|\mathbb{K}_H^{-1}\left(\int_0^\cdot G_{\lambda_n}(Z)\right)(t)\right\|^2 dt\right], \tag{3.61}$$

where $\tilde{\mathbb{E}}_{\lambda_n}$ is the expectation with respect to $\tilde{\mathbb{P}}_{\lambda_n}$ and

$$\frac{d\tilde{\mathbb{P}}_{\lambda_n}}{d\mathbb{P}} = \exp\left[\rho_{\lambda_n}(Z)\right],$$

and $Z(\cdot)$ satisfies (2.28). On the probability space with the measure $\mathbb{P}_{\lambda_n}$, $Z(\cdot)$ satisfies the following semilinear equation, where $B(\cdot)$ is a fractional Brownian motion with respect to $\mathbb{P}_{\lambda_n}$:

$$\begin{aligned} dX_{\lambda_n}(t) &= \left(AX(t) + F_{\lambda_n}(X(t))\right) dt + \Phi\, dB(t), \\ X_{\lambda_n}(0) &= X_0. \end{aligned} \tag{3.62}$$

Since F_{λ_n} is Lipschitz continuous, there is a unique mild solution on a given probability space, so it suffices to show

$$\mathbb{E}\int_0^T \left\|\mathbb{K}_H^{-1}\left(\int_0^\cdot G_{\lambda_n}\left(X_{\lambda_n}\right)\right)(t)\right\|^2 dt \le c \tag{3.63}$$

for some $c \in \mathbb{R}_+$ that does not depend on λ_n. Repeating inequalities (3.51), where G and Z are replaced by G_{λ_n} and X_{λ_n}, respectively, and using inequality (3.58), it follows that

$$\int_0^T \left\|\mathbb{K}_H^{-1}\left(\int_0^\cdot G_{\lambda_n}\left(X_{\lambda_n}\right)\right)(t)\right\|^2 dt \le c_5\left(1 + \|X_0\|_E^{2\rho} + \left|\tilde{X}_{\lambda_n}\right|^{2\rho}_{C([0,T],E)}\right) \tag{3.64}$$

for a constant c_5 that does not depend on $n \in \mathbb{N}$ where $\tilde{X}_{\lambda_n}(t) = X_{\lambda_n}(t) - S(t)X_0$. By inequality (3.43) there is a constant c_6 that does not depend on n such that

(3.65)
$$\mathbb{E}\int_0^T \left\| \mathbb{K}_H^{-1}\left(\int_0^{\cdot} G_{\lambda_n}\left(X_{\lambda_n}\right)\right)(t)\right\|^2 dt \le c_6\left(1 + \|X_0\|_E^{2\rho} + \mathbb{E}\,|Z|^{4\rho^2}_{C([0,T],E)}\right) = C < \infty.$$

This inequality verifies (3.60). Thus the sequence $(\exp[\rho_{\lambda_n}(Z)], n \in \mathbb{N})$ converges in L^1 and equality (3.54) is satisfied. □

Now the case $H \in (1/2, 1)$ is considered.

THEOREM 3.7. *Let $H \in (1/2, 1)$ and the other assumptions in Theorem 3.6 be satisfied. Let $\Phi^{-1} \in \mathcal{L}(V)$, $\tilde{Z} \in C^\beta([0,T], V)$ for some $\beta \in (0,1)$,*

$$\langle F(x) - F(y), x - y\rangle \le k_2 \|x - y\|^2 \tag{3.66}$$

for each pair $x, y \in E$ and a $k_2 \in \mathbb{R}_+$ (that is, $F - k_2 I$ is dissipative on E with respect to the norm on V) and

$$\|F(x) - F(y)\| \le k_3\left(1 + \|x\|_E^q + \|y\|_E^q\right)\|x - y\|^\gamma \tag{3.67}$$

for each $x, y \in E$, with some $k_3 > 0$, $q \ge 1$, and $\gamma \in (0,1]$ such that

$$\gamma\beta > H - \frac{1}{2}. \tag{3.68}$$

Then there is one and only one mild solution to (3.4), and its probability law is mutually absolutely continuous with respect to the probability law of the fractional Ornstein–Uhlenbeck process (2.28) on $\check{\Omega}$.

Proof. As in the proof of Theorem 3.6, it is shown that

$$\int_0^{\cdot} G(Z(s))\,ds \in I_{0+}^{H+\frac{1}{2}}\left(L^2([0,T], V)\right) \tag{3.69}$$

and

$$\mathbb{E}\exp[\rho(Z)] = 1. \tag{3.70}$$

The methods to verify (3.69) and (3.70) are similar to those used in the proof of Theorem 3.6, but now the operator $\mathbb{K}_H^{-1}$ has a different form. Using inequality (3.41) and the Hölder continuity condition (3.67), it follows that

$$\begin{aligned}
&\left|\mathbb{K}_H^{-1}\left(\int_0^{\cdot} G(Z)\right)\right|^2_{L^2([0,T],V)} \\
&\quad\le c_1\int_0^T\left(s^{\frac{1}{2}-H}\|G(Z(s))\| + s^{H-\frac{1}{2}}\int_0^s \frac{s^{\frac{1}{2}-H} - r^{\frac{1}{2}-H}}{(s-r)^{H+\frac{1}{2}}}\|G(Z(r))\|\,dr\right. \\
&\quad\left. + \int_0^s \frac{\|G(Z(s)) - G(Z(r))\|}{(s-r)^{H+\frac{1}{2}}}\,dr\right)^2 ds \\
&\quad\le c_2\left[1 + |Z|^{2\rho}_{C([0,T],E)} + c_3\left(1 + |Z|^{2q}_{C([0,T],E)}\right)\right. \\
&\quad\left.\cdot\int_0^T\left(\int_0^s \frac{\|S(s)X_0 - S(r)X_0\|^\gamma + \left\|\tilde{Z}(s) - \tilde{Z}(r)\right\|^\gamma}{(s-r)^{H+\frac{1}{2}}}\,dr\right)^2 ds\right]
\end{aligned} \tag{3.71}$$

for some constants c_1, c_2, c_3. By the analyticity of the semigroup $S(\cdot)$ on V, it follows that

$$\begin{aligned}(3.72)\qquad &\left|\mathbb{K}_H^{-1}\left(\int_0^{\cdot} G(Z)\right)\right|^2_{L^2([0,T],V)}\\ &\quad\le c_4\left[1+\|X_0\|_E^{2\rho}\right.\\ &\quad+\left.\left(\|X_0\|_E^{2q}+|Z|^{2q}_{C([0,T],V)}+1\right)\left(\|X_0\|_E^{2\gamma}+|Z|^{2\gamma}_{C^\beta([0,T],V)}\right)\right]\\ &\quad\le c_5\left(1+\|X_0\|_E^m+|Z|^m_{C([0,T],V)}+|Z|^m_{C^\beta([0,T],V)}\right)\end{aligned}$$

for some constants c_4 and c_5 and m sufficiently large. Thus (3.69) is verified. To verify equality (3.70) consider the family of Yosida approximations $(F_\lambda, \lambda>0)$ of F as in the proof of Theorem 3.6. By the dissipativity of F in the norm on V, $F_\lambda\colon V\to V$ is Lipschitz continuous for each $\lambda>0$ and has at most polynomial growth, so F_λ satisfies the assumptions of Theorem 3.4 so that

$$(3.73)\qquad \mathbb{E}\exp\left[\rho_\lambda(Z)\right]=1,$$

where ρ_λ is given by (3.55). By the method used to obtain inequality (3.71), it follows that

$$\begin{aligned}(3.74)\quad &\mathbb{E}\left|\mathbb{K}_H^{-1}\left(\int_0^{\cdot}\left(G_\lambda(Z)-G(Z)\right)\right)\right|^2_{L^2([0,T],V)}\\ &\qquad\le c_6\mathbb{E}\int_0^T\left[s^{\frac12-H}\left\|G_\lambda(Z(s))-G(Z(s))\right\|\right.\\ &\quad+s^{H-\frac12}\int_0^s\frac{s^{\frac12-H}-r^{\frac12-H}}{(s-r)^{H+\frac12}}\left\|G_\lambda(Z(r))-G(Z(r))\right\|\,dr\\ &\qquad\left.+\int_0^s\frac{\left\|G_\lambda(Z(s))-G(Z(s))-G_\lambda(Z(r))+G(Z(r))\right\|}{(s-r)^{H+\frac12}}\,dr\right]^2 ds.\end{aligned}$$

By inequalities (3.57) and (3.58), it follows that $\|G_\lambda(x)-G(x)\|\to 0$ as $\lambda\downarrow 0$ for each $x\in E$ and the family $(G_\lambda,\lambda>0)$ satisfies the growth condition

$$(3.75)\qquad \|G_\lambda(x)\|\le c_7\left(1+\|x\|_E^\rho\right)$$

for $x\in E$ and some $c_7>0$. From the V-dissipativity of F, it follows by [5, Proposition 5.5.3] that

$$\|R_\lambda(x)-R_\lambda(y)\|\le\|x-y\|$$

for $x,y\in E$, so that

$$\begin{aligned}(3.76)\qquad \|F_\lambda(x)-F_\lambda(y)\|&=\|F\left(R_\lambda(x)\right)-F\left(R_\lambda(y)\right)\|\\ &\le k_3\left(1+\|R_\lambda(x)\|_E^q+\|R_\lambda(y)\|_E^q\right)\|x-y\|^\gamma\end{aligned}$$

for $x,y\in E$. Since

$$\|R_\lambda(x)\|_E\le\|x\|_E+\lambda\|F(x)\|_E\le c_8\left(1+\|x\|_E^\rho\right)$$

for $x \in E$, $c_8 \in \mathbb{R}_+$, and $\lambda \in (0,1]$, there is the inequality

$$\|F_\lambda(x) - F_\lambda(y)\| \le c_9 \left(1 + \|x\|_E^m + \|y\|_E^m\right) \|x - y\|^\gamma \tag{3.77}$$

for $x, y \in E$, $c_0 \in \mathbb{R}_+$, $m \ge 1$, and $\lambda \in (0,1]$. So F_λ and G_λ satisfy inequality (3.67) uniformly in $\lambda \in (0,1]$. Thus the right-hand side of inequality (3.74) tends to zero as $\lambda \downarrow 0$ by the dominated convergence theorem where a majorizing function is provided by the estimates (3.75) and (3.77), whose integrability is shown in (3.71) and (3.72), and there is a decreasing sequence $(\lambda_n, n \in \mathbb{N})$ whose limit is zero such that

$$\lim_{n\to\infty} \exp\left[\rho_{\lambda_n}(Z)\right] = \exp\left[\rho(Z)\right] \quad \text{a.s. } \mathbb{P}. \tag{3.78}$$

The uniform integrability of the sequence $(\exp\left[\rho_{\lambda_n}(Z)\right], n \in \mathbb{N})$ is shown by verifying the analogue of (3.60). Equivalently,

$$\sup_n \mathbb{E} \int_0^T \left\| \mathbb{K}_H^{-1} \left(\int_0^\cdot G_{\lambda_n}\left(X_{\lambda_n}\right) \right)(t) \right\|^2 dt \le c < \infty, \tag{3.79}$$

where $X_{\lambda_n}(\cdot)$ is the unique mild solution to (3.62). The analogous inequalities (3.71)–(3.74) are obtained by replacing G by G_λ using the polynomial growth bound and the local Hölder continuity that are uniform in $(\lambda_n, n \in \mathbb{N})$, and $Z(\cdot)$ is replaced by $X_{\lambda_n}(\cdot)$. For some constants c_{10} and $m \ge 1$,

$$\begin{aligned} &\left| \mathbb{K}_H^{-1} \left(\int_0^\cdot G_{\lambda_n}\left(X_{\lambda_n}\right) \right) \right|_{L^2([0,T],V)}^2 \\ &\qquad \le c_{10} \left(1 + \|X_0\|_E^m + \left|\tilde{X}_{\lambda_n}\right|_{C([0,T],E)}^m + \left|\tilde{X}_{\lambda_n}\right|_{C^\beta([0,T],V)}^m \right), \end{aligned} \tag{3.80}$$

where $\tilde{X}_{\lambda_m}(t) = X_{\lambda_m}(t) - S(t)X_0$. By inequality (3.43), it follows that

$$\left|\tilde{X}_{\lambda_n}\right|_{C([0,T],E)} \le c_{11} \left(1 + \|X_0\| + |Z|_{C(0,T],E)}^\rho \right) \tag{3.81}$$

for some $c_{11} > 0$. Let $w_{\lambda_n}(t) = \tilde{X}_{\lambda_n}(t) - \tilde{Z}(t)$ so that

$$w_{\lambda_n}(t) = \int_0^t S(t-r) F_{\lambda_n} \left(w_{\lambda_n}(s) + S(s)X_0 + \tilde{Z}(s) \right) ds \tag{3.82}$$

for $t \in [0,T]$. Inequality (3.81) provides a uniform bound on $|w_{\lambda_n}|_{C([0,T],E)}$, so by repeating the arguments for inequalities (3.37) and (3.38), it follows that

$$|X_{\lambda_n}|_{C^\beta([0,T],V)} \le c_{12} \left(1 + \|X_0\| + |\tilde{Z}|_{C([0,T],E)}^\rho + |\tilde{Z}|_{C^\beta([0,T],V)} \right) \tag{3.83}$$

for some $c_{12} > 0$. Inequalities (3.80) and (3.81) verify inequality (3.79), so the sequence $(\exp\left[\rho_{\lambda_n}(Z)\right], n \in \mathbb{N})$ is uniformly integrable and equality (3.73) is verified. □

4. Some examples. The first example is a finite-dimensional stochastic equation with a nonlinear drift. Consider the equation

$$dX(t) = f(X(t))\, dt + \Phi\, dB(t), \tag{4.1}$$

where $f\colon \mathbb{R}^n \to \mathbb{R}^n$, $\Phi \in \mathcal{L}(\mathbb{R}^n)$, and $(B(t), t \geq 0)$ is an $\mathbb{R}^n$-valued standard fractional Brownian motion with Hurst parameter $H \in (0,1)$. This case can be subsumed in the infinite-dimensional results given here, though some of the assumptions and the results simplify significantly. Let $E = V = \mathbb{R}^n$, $S(t) = I$ for $t \in \mathbb{R}_+$, and assume that $Q = \Phi\Phi^*$ is positive definite. The process

$$\left(\int_0^t \Phi \, dB, t \in [0,T] \right)$$

has sample paths in $C^\beta([0,T], \mathbb{R}^n)$ for $0 < \beta < H$. If $f\colon \mathbb{R}^n \to \mathbb{R}^n$ is Borel measurable and

$$\|f(x)\| \leq k_1(1 + \|x\|) \tag{4.2}$$

for some $k_1 > 0$ and all $x \in \mathbb{R}^n$, then for $H \in (0, 1/2)$ there is one and only one weak solution of (4.1) by Theorem 3.3. If, additionally, it is assumed that

$$\|f(x) - f(y)\| \leq k\|x - y\|^\gamma \tag{4.3}$$

for all $x, y \in \mathbb{R}^n$ and some $\gamma > 1 - \frac{1}{2H}$, then for $H \in (1/2, 1)$ there is one and only one weak solution. In each of these cases, the probability measure of the solution is mutually absolutely continuous with respect to the probability measure of the process $(\Phi B(t), t \in [0,T])$.

Now, replace the inequality in (4.2) by

$$\|f(x)\| \leq k_1(1 + \|x\|^\rho) \tag{4.4}$$

for some $\rho \geq 1$ and $k_1 > 0$. Assume that $f\colon \mathbb{R}^n \to \mathbb{R}^n$ is continuous and satisfies

$$\langle f(x) - f(y), x - y \rangle \leq k_3\|x - y\|^2 \tag{4.5}$$

for some $k_3 > 0$ and all $x, y \in \mathbb{R}^n$. If $H \in (1/2, 1)$, then assume that

$$\|f(x) - f(y)\| \leq k_4 \left(1 + \|x\|^q + \|y\|^q\right) \|x - y\|^\gamma \tag{4.6}$$

for some $q \geq 1$, $k_4 > 0$, $\gamma > 1 - \frac{1}{2H}$. For $H \in (0, \frac{1}{2})$ Theorem 3.6 can be used to verify that the probability law of the solution of (4.1) is mutually absolutely continuous with respect to the probability law of $(\Phi B(t), t \in [0,T])$. Furthermore, there is one and only one mild solution of (4.1); in fact, since the state space is finite-dimensional, the mild solution is a strong solution. For $H \in (\frac{1}{2}, 1)$ Theorem 3.7 can be used to verify mutual absolute continuity and one and only one mild solution as for the case $H \in (0, \frac{1}{2})$. Note that inequalities (4.4)–(4.6) are satisfied for the important case of models where f is a polynomial of odd degree with a negative leading coefficient.

The second example is a stochastic parabolic equation of $2m$th order:

$$\frac{\partial u}{\partial t}(t, \xi) = [L_{2m}u](t, \xi) + f(u(t, \xi)) + \eta(t, \xi) \tag{4.7}$$

for $(t, \xi) \in [0,T] \times \mathcal{O}$ with the initial condition

$$u(0, \xi) = x(\xi) \tag{4.8}$$

for $\xi \in \mathcal{O}$ and the Dirichlet boundary condition

$$\frac{\partial^k u}{\partial v^k}(t, \xi) = 0 \tag{4.9}$$

for $(t,\xi) \in [0,T] \times \partial\mathcal{O}$, $k \in \{0,\dots,m-1\}$, with $\frac{\partial}{\partial v}$ denoting the conormal derivative, $\mathcal{O}$ a bounded domain in $\mathbb{R}^d$ with a smooth boundary, and L_{2m} a $2m$th order uniformly elliptic operator

$$L_{2m} = \sum_{|\alpha| \le 2m} a_\alpha(\xi) D^\alpha \tag{4.10}$$

with $a_\alpha \in C_b^\infty(\mathcal{O})$. For example, if $m = 1$, then this equation is called the stochastic heat equation. The process η denotes a space-dependent noise process that is fractional in time with the Hurst parameter $H \in (0,1)$ and, possibly, in space. The system (4.7)–(4.9) is modeled as

$$\begin{aligned} dX(t) &= AX(t)dt + F(X(t))dt + \Phi dB(t), \\ X(0) &= x \end{aligned} \tag{4.11}$$

in the space $V = L^2(\mathcal{O})$, where $A = L_{2m}$,

$$\mathrm{Dom}(A) = \left\{ \varphi \in H^{2m}(\mathcal{O}) \Big| \frac{\partial^k}{\partial v^k}\varphi = 0 \text{ on } \partial D \text{ for } k \in \{0,\dots,m-1\} \right\},$$

$F : V \to V$ is the operator, $F(x)(\xi) = f(x(\xi)), x \in V, \xi \in \mathcal{O}$, $\Phi \in \mathcal{L}(V)$ defines the space correlation of the noise process, and $(B(t), t \ge 0)$ is a cylindrical standard fractional Brownian motion in V (formally, $\eta(t,\cdot) = \Phi(\partial/\partial t)B(t,\cdot)$). For $\Phi = I$, the noise process is uncorrelated in space. It is well known that A generates an analytic semigroup $(S(t), t \ge 0)$. Furthermore

$$|S(t)\Phi|_{\mathcal{L}_2(V)} \le |S(t)|_{\mathcal{L}_2(V)} |\Phi|_{\mathcal{L}(V)} \le c t^{-\frac{d}{4m}} \tag{4.12}$$

for $t \in [0,T]$, so if

$$H > \frac{d}{4m}, \tag{4.13}$$

then the conditions of Proposition 2.6 are satisfied with $\gamma = \frac{d}{4m}$. Therefore, for any $\Phi \in \mathcal{L}(V)$, the stochastic convolution process

$$\left(\int_0^t S(t-r)\Phi \, dB(r), t \in [0,T] \right)$$

is well-defined and has a version with $C^\beta([0,T],V)$ sample paths for $\beta \ge 0$ satisfying

$$\beta < H - \frac{d}{4m}. \tag{4.14}$$

Note that the condition (4.13) extends the well-known result for a standard Wiener process ($H = \frac{1}{2}$).

Theorems 3.3 and 3.4 are applied to the present example. Assume inequality (4.13) and let Φ be boundedly invertible on V. Furthermore, let $f : \mathbb{R} \to \mathbb{R}$ be measurable and satisfy

$$|f(\xi)| \le k_1(1 + |\xi|), \qquad \xi \in \mathbb{R}. \tag{4.15}$$

By the preceding part of this example, conditions (H1)–(H3) are satisfied for $E = V = L^2(\mathcal{O})$ and the map $F : V \to V$ has at most linear growth. Thus by Theorem 3.3 if $H < \frac{1}{2}$, then there exists a unique weak solution to (4.11).

If $H > \frac{1}{2}$, some additional conditions are required. Assume that

$$\frac{d}{4m} < \frac{1}{2} \tag{4.16}$$

(which is more restrictive than (4.13)) and suppose that

$$|f(\xi) - f(\lambda)| \le k|\xi - \lambda|^\gamma, \qquad \xi, \lambda \in \mathbb{R}, \tag{4.17}$$

for some $k > 0$ and $\gamma > 0$,

$$\frac{H - 1/2}{H - d/4m} < \gamma \le 1. \tag{4.18}$$

Then, letting β be such that $\beta < H - \frac{d}{4m}$ and $\gamma\beta > H - \frac{1}{2}$, it is clear that all of the conditions of Theorem 3.4 are verified so there is a unique weak solution to (4.11).

The third example is a one-dimensional stochastic equation of reaction-diffusion type. Consider the equation

$$\frac{\partial u}{\partial t}(t,\xi) = \frac{\partial^2 u}{\partial \xi^2}(t,\xi) + f(u(t,\xi)) + \eta(t,\xi) \tag{4.19}$$

for $(t,\xi) \in (0,T) \times (0,1)$ and

$$u(0,\xi) = x_0(\xi),$$
$$\frac{\partial u}{\partial \xi}(t,0) = \frac{\partial u}{\partial \xi}(t,1) = 0$$

for $(t,\xi) \in (0,T) \times (0,1)$, where f and η are given in the previous example (with $\mathcal{O} = (0,1)$). The above formal equation can be rewritten in the form (4.11) with $V = L^2([0,1])$, $A = \frac{\partial^2}{\partial \xi^2}$,

$$\mathrm{Dom}(A) = \left\{ \phi \in H^2([0,1]) : \frac{\partial}{\partial \xi}\phi(0) = \frac{\partial}{\partial \xi}\phi(1) = 0 \right\},$$

where $\Phi \in \mathcal{L}(V)$ and F is as given in the preceding example. The semigroup generated by A satisfies the estimate (4.12) (with $m = d = 1$), so if f satisfies the conditions of the previous example, the same conclusions on existence and uniqueness of the weak solution are obtained.

However, it is desirable to relax condition (4.15) of the linear growth of the function f, which is very restrictive in view of reaction-diffusion models, where f is often a polynomial. Let $H > 1/2$ and assume that

$$|f(\xi)| \le k(1 + |\xi|^\rho), \tag{4.20}$$

$$(f(\xi) - f(\lambda))\,\mathrm{sgn}(\xi - \lambda) \le k(\xi - \lambda), \tag{4.21}$$

$$|f(\xi) - f(\lambda)| \le k(1 + |\xi|^q + |\lambda|^q|\xi - \lambda|^\gamma \tag{4.22}$$

for all $\xi, \lambda \in \mathbb{R}$ and some universal constants $\rho > 0, q > 0, k > 0$ and γ satisfying

$$\frac{H - 1/2}{H - 1/4} < \gamma \leq 1. \tag{4.23}$$

Note that these conditions are satisfied if f is Lipschitz or if f is a polynomial of odd degree with a negative leading coefficient.

The conditions of Theorem 3.7 are verified now. Take the state space $E = C([0,1])$. It is well known that the restriction of A to E generates a strongly continuous semigroup of contractions on E. By Proposition 2.6 the stochastic convolution

$$\left(\int_0^t S(t-r)\Phi dB(r), t \in [0,T] \right) \tag{4.24}$$

has $C^\beta([0,T], V_\delta)$ sample paths for $\beta + \delta < H - 1/4$, and, hence, by the Sobolev embedding theorem, in the space $C([0,T], E) \cap C^\beta([0,T], V)$ for $0 < \beta < H - 1/4$ (by (4.23) β can be chosen such that $\beta\gamma > H - 1/2$). It remains to verify the conditions imposed on F. The polynomial growth condition (3.41), the "dissipativity of $F - kI$ on V" (3.66), and the local Hölder continuity of the form (3.67) follow easily from the corresponding conditions on f, that is, (4.20), (4.21), and (4.22). The dissipativity of $F - kI$ on E (3.42) is a well-known consequence of (4.21) by the characterization of the subdifferential of the norm on $E = C([0,1])$ (cf. [32]). The characterization of the subdifferentials for this example is as follows: Given $x \in E$, let $M_x = \{\xi \in [0,1]; |x(\xi)| = ||x||_E\}$. Then $\mu \in \delta||x||_E$ if and only if the following three conditions are satisfied: (i) μ is a Radon measure on $[0,1]$ with $||\mu|| = 1$; (ii) the support of μ is contained in M_x; and (iii) $\int_\Gamma sgn\ x(\xi)\mu(d\xi) \geq 0$ for each Borel set Γ in $[0,1]$. In particular if $x \in E$ has the property that $M_x = \{\xi_0\}$, then $\delta||x||_E = \delta_{\xi_0}$ for $x(\xi_0) = ||x||_E$ and $\delta||x||_E = -\delta_{\xi_0}$ for $x(\xi_0) = -||x||_E$, where δ_{ξ_0} is the Dirac distribution at ξ_0. The family of $x \in E$ with this latter property is dense in E. Therefore, all of the conditions of Theorem 3.7 are satisfied, and it follows that there is a unique weak solution in the present case.

REFERENCES

[1] E. Alòs, O. Mazet, and D. Nualart, *Stochastic calculus with respect to Gaussian processes*, Ann. Probab., 29 (2001), pp. 766–801.

[2] P. Caithamer, *The stochastic wave equation driven by fractional Brownian noise and temporally correlated smooth noise*, Stoch. Dyn., 5 (2005), pp. 45–64.

[3] R. H. Cameron and W. T. Martin, *Transformations of Wiener integrals under translations*, Ann. of Math. (2), 2 (1944), pp. 386–396.

[4] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Encyclopedia Math. Appl. 44, Cambridge University Press, Cambridge, UK, 1992.

[5] G. Da Prato and J. Zabczyk, *Ergodicity for Infinite-Dimensional Systems*, London Math. Soc. Lecture Note Ser. 229, Cambridge University Press, Cambridge, UK, 1996.

[6] L. Decreusefond and A. S. Üstünel, *Stochastic analysis of the fractional Brownian motion*, Potential Anal., 10 (1999), pp. 177–214.

[7] L. Denis, M. Erraoui, and Y. Ouknine, *Existence and uniqueness for solutions of one dimensional SDE's driven by an additive fractional noise*, Stoch. Stoch. Rep., 76 (2004), pp. 409–427.

[8] T. E. Duncan, *Some stochastic semilinear equations in Hilbert space with fractional Brownian motion*, in Optimal Control and Partial Differential Equations, J. Menaldi, E. Rofman, and A Sulem, eds., IOS Press, Amsterdam, 2000, pp. 241–247.

[9] T. E. Duncan, *Some processes associated with a fractional Brownian motion*, in Mathematics of Finance, Contemp. Math. 351, AMS, Providence, RI, 2004, pp. 93–101.

[10] T. E. Duncan, Y. Hu, and B. Pasik-Duncan, *Stochastic calculus for fractional Brownian motion* I. *Theory*, SIAM J. Control Optim., 38 (2000), pp. 582–612.

[11] T. E. Duncan, J. Jakubowski, and B. Pasik-Duncan, *Stochastic integration for fractional Brownian motion in a Hilbert space*, Stoch. Dyn., 6 (2006), pp. 53–75.
[12] T. E. Duncan, B. Maslowski, and B. Pasik-Duncan, *Stochastic equations in Hilbert space with a multiplicative fractional Gaussian noise*, Stochastic Process. Appl., 115 (2005), pp. 1357–1383.
[13] T. E. Duncan, B. Pasik-Duncan, and B. Maslowski, *Fractional Brownian motion and stochastic equations in Hilbert spaces*, Stoch. Dyn., 2 (2002), pp. 225–250.
[14] I. V. Girsanov, *On transforming a class of stochastic processes by absolutely continuous substitution of measures*, Teor. Verojatnost. i Primenen., 5 (1960), pp. 314–330.
[15] W. Grecksch and V. V. Anh, *A parabolic stochastic differential equation with fractional Brownian motion input*, Statist. Probab. Lett., 41 (1999), pp. 337–346.
[16] Y. Hu, *Heat equations with fractional white noise potentials*, Appl. Math. Optim., 43 (2001), pp. 221–243.
[17] Y. Hu, *Integral transformations and anticipative calculus for fractional Brownian motions*, Mem. Amer. Math. Soc., 175 (2005).
[18] Y. Hu, B. Øksendal, and T. Zhang, *General fractional multiparameter white noise theory and stochastic partial differential equations*, Comm. Partial Differential Equations, 29 (2004), pp. 1–23.
[19] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, Grad. Texts in Math. 113, Springer-Verlag, New York, 1988.
[20] R. S. Liptser and A. N. Shiryaev, *Statistics of Random Processes* I: *General Theory*, 2nd ed., Springer-Verlag, Berlin, 2001.
[21] B. Maslowski and D. Nualart, *Evolution equations driven by a fractional Brownian motion*, J. Funct. Anal., 202 (2003), pp. 277–305.
[22] B. Maslowski and B. Schmalfuss, *Random dynamical systems and stationary solutions of differential equations driven by the fractional Brownian motion*, Stochastic Anal. Appl., 22 (2004), pp. 1577–1607.
[23] R. H. Martin, Jr., *A global existence theorem for autonomous differential equations in a Banach space*, Proc. Amer. Math. Soc., 26 (1970), pp. 307–314.
[24] R. Mikulevicius and B. L. Rozovsky, *Martingale paths for stochastic PDE's*, in Stochastic Partial Differential Equations: Six Perspectives, Math. Survey Monogr. 64, AMS, Providence, RI, 1999, pp. 243–325.
[25] S. Moret and D. Nualart, *Onsager-Machlup functional for the fractional Brownian motion*, Probab. Theory Related Fields, 124 (2002), pp. 227–260.
[26] I. Norros, E. Valkeila, and J. Virtamo, *An elementary approach to a Girsanov formula and other analytical results on fractional Brownian motions*, Bernoulli, 5 (1999), pp. 571–587.
[27] D. Nualart and Y. Ouknine, *Regularization of differential equations by fractional noises*, Stoch. Process. Appl., 102, (2002) pp. 103–116.
[28] B. Pasik-Duncan, T. E. Duncan, and B. Maslowski, *Linear stochastic equations in a Hilbert space with a fractional Brownian motion*, in Stochastic Processes, Optimization, and Control Theory: Applications in Financial Engineering, Queueing Networks, and Manufacturing Systems, H. Yan, G. Yin, and Q. Zhang, eds., Springer-Verlag, New York, 2006, pp. 201–221.
[29] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, New York, 1983.
[30] V. Pipiras and M. S. Taqqu, *Integration questions related to fractional Brownian motion*, Probab. Theory Related Fields, 118 (2000), pp. 251–291.
[31] S. G. Samko, A. A. Kilbas, and O. I. Marichev, *Fractional Integrals and Derivatives*, Gordon and Breach, Yverdon, 1993.
[32] E. Sinestrari, *Accretive differential operators*, Boll. Un. Mat. Ital. B (5), 13 (1976), pp. 19–31.
[33] S. Tindel, C. A. Tudor, and F. Viens, *Stochastic evolution equations with fractional Brownian motion*, Probab. Theory Related Fields, 127 (2003), pp. 186–204.

SIAM J. MATH. ANAL.
Vol. 40, No. 6, pp. 2316–2327

ON THE EXISTENCE OF SOLUTIONS OF EQUILIBRIA IN LUBRICATED JOURNAL BEARINGS*

I. CIUPERCA†, M. JAI‡, AND J. I. TELLO§

Abstract. In this paper we study a system of equations concerning equilibrium positions of journal bearings. The problem consists of two surfaces in relative motion separated by a small distance filled with a lubricant. The shape of the inlet surface is circular, while the other surface has a more general shape. Our result shows the existence of at least one equilibrium by using degree theory.

Key words. Reynolds variational inequality, inverse problem, existence of solutions, degree theory

AMS subject classifications. 35J20, 47H11, 49J10

DOI. 10.1137/080724228

1. Introduction.

We consider in this paper a lubricated system called a journal bearing consisting of two cylinders in relative motion. An incompressible fluid, the lubricant, is introduced in the narrow space between the cylinders. An exterior force $F = (F_1, F_2) \in \mathbb{R}^2$ is applied on the inner cylinder (shaft) which turns with a given velocity ω.

The wedge between the two cylinders is assumed to satisfy the thin-film hypothesis, so that the pressure (assumed time-independent) does not depend on the normal coordinate to the bodies and obeys the Reynolds equation.

In order to introduce the Reynolds equation we need to describe the geometry and the dynamic of the system. We suppose that the interior cylinder has a circular form of constant radius (assumed 1) which rotates with known velocity. The transversal axis of the shaft is assumed to have only two degrees of freedom in the transversal plane, i.e., parallel to the exterior cylinder (bush) which is fixed and not necessarily of constant radius.

Let us consider (O, y_1, y_2) a reference system in the transversal plane to the cylinders, and suppose that the distance between O and the surface of the bush is larger than the radius of the shaft (see Figure 1.1). We also assume that the representation in polar coordinates (r, θ) of the bush is given by

$$r = 1 + \delta\rho(\theta), \tag{1.1}$$

where $\rho : [0, 2\pi] \to [1, +\infty[$ is a known function and $\delta > 0$ (the clearance) is a small parameter, representing the distance between the two cylinders when O and the center of the shaft coincide.

*Received by the editors May 14, 2008; accepted for publication (in revised form) November 12, 2008; published electronically February 20, 2009.
http://www.siam.org/journals/sima/40-6/72422.html

†Université de Lyon, Université Lyon 1, CNRS, UMR 5208, Institut Camille Jordan, Bat. Braconnier, 43, blvd du 11 novembre 1918, F-69622 Villeurbanne Cedex, France (ciuperca@math.univ-lyon1.fr).

‡ICJ CNRS-UMR 5208, INSA de Lyon, Bât. Léonard de Vinci, 20 A. A. Einstein, 69621 Villeurbanne Cedex, France (Mohammed.Jai@insa-lyon.fr).

§Matemática Aplicada, E.U.I. Informática, Universidad Politécnica de Madrid, 28031 Madrid, Spain (jtello@eui.upm.es). This author was partially supported by project MTM2005-03463 of DGISGPI of Spain and projects of CAM: CCG07-UPM/000-3199 at UPM and CCG07-UCM/ESP-2787 at UCM.

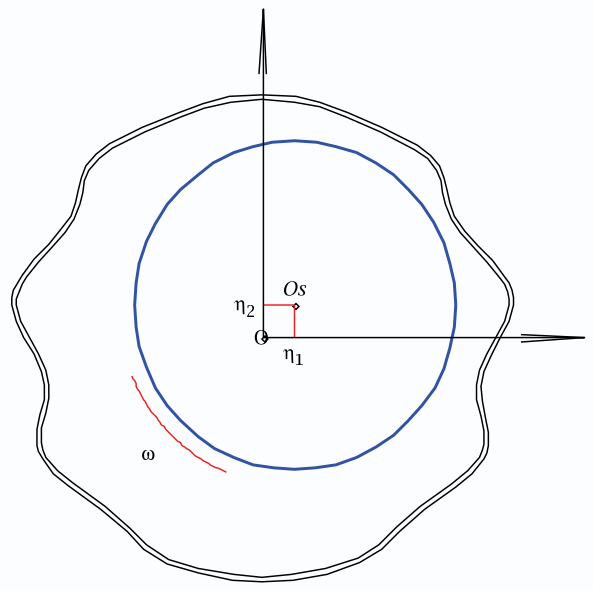


FIG. 1.1. *Scheme of the journal bearing.*

Remark 1.1. The particular case $\rho \equiv 1$ corresponds to a circular bush with radius $1+\delta$ and O the center of the bush.

Let us now denote by O_s the center of the shaft. The position of O_s is given in cartesian coordinates by $(\delta\eta_1, \delta\eta_2)$ and in polar coordinates by $(\delta\eta, \alpha)$, that is,

$$\begin{aligned} \eta_1 &= \eta\cos\alpha, \\ \eta_2 &= \eta\sin\alpha. \end{aligned} \tag{1.2}$$

It is well known (see, for instance, [5]) that the distance between the two cylinders is given by $\delta h(\theta,\eta,\alpha) + O(\delta^2)$ with

$$h(\theta,\eta,\alpha) = \rho(\theta) - \eta\cos(\theta-\alpha) = \rho(\theta) - \eta_1\cos\theta - \eta_2\sin\theta. \tag{1.3}$$

The formulation of the problem is complete when the distance between the surfaces is of order δ (i.e., $h \in O(1)$) and second order terms are neglected. Admissible forces in that case are of order $\frac{1}{\delta^2}$ or smaller. If the distance becomes smaller (i.e., $h \ll 1$) or forces are large (larger than $O(\frac{1}{\delta^2})$), second order terms cannot be neglected and the formulation loses its physical meaning.

Now, the problem will be posed in a fixed domain $\Omega =]0, 2\pi[\times]0,1[$ which parametrizes the space between shaft and bush. In fact the gap is approximated by the following domain given in cylindrical coordinates by (r,θ,x) with respect to O_s:

$$1 \le r \le 1 + \delta h(\theta,\eta,\alpha), \quad \theta \in [0,2\pi[, \quad x \in [0,1],$$

where $O_s = (\delta\eta\cos(\alpha), \delta\eta\sin(\alpha))$.

Since the wedge between the two cylinders satisfies the thin-film hypothesis, the pressure of the lubricant fluid (assumed time-independent) does not depend on the

normal coordinate to the bodies and obeys the Reynolds equation (see [9]). We also consider that there is an alimentation region along the circles $\{x = 0\}$ and $\{x = 1\}$, respectively, where the pressure in the fluid equals the atmospheric pressure supposed to be 0 by translation. Then the pressure $p : (\theta, x) \in \Omega \to \mathbb{R}$ satisfies the following problem written in nondimensional form:

$$\begin{cases} \nabla \cdot (h^3 \nabla p) = \frac{\partial h}{\partial \theta} & \text{on } \Omega, \\ p & \text{is } 2\pi\text{-periodic in } \theta, \\ p = 0 & \text{on }]0, 2\pi[\times \{0\} \cup]0, 2\pi[\times \{1\}. \end{cases} \tag{1.4}$$

In general the solution of (1.4) is not always nonnegative and we must replace (1.4) by the corresponding variational inequality.

In this work we are interested in an equilibrium problem which entails finding the position (η_1, η_2) of the shaft such that the hydrodynamic force (load) created by the pressure film equilibrates the exterior force F. Thus the problem is formulated as follows:

$$\text{Find } p \in K, (\eta_1, \eta_2) \in A \text{ such that}$$

$$\int_\Omega h^3 \nabla p \cdot \nabla(\varphi - p) \geq \int_\Omega h \frac{\partial}{\partial \theta}(\varphi - p) \quad \forall \varphi \in K, \tag{1.5}$$

$$\int_\Omega p \cos\theta \, d\theta dx = F_1, \tag{1.6}$$

$$\int_\Omega p \sin\theta \, d\theta dx = F_2, \tag{1.7}$$

where h is given in (1.3),

$$K = \left\{ \varphi \in H_0^1(\Omega) : \varphi \geq 0 \right\},$$

and $A \subset \mathbb{R}^2$. The set of admissible positions of the shaft (to be defined later) is such that $h(\theta, \eta, \alpha) > 0 \ \ \forall \theta \in [0, 2\pi]$.

As far as we know, very few works can be found in the literature concerning existence of equilibrium in the lubricated devices in spite of a larger number of references to numerical simulations; see, for instance, [3] or [4].

Some results exist in the case of sliders, that is, mechanisms consisting of an almost plane surface sliding above a horizontal plane surface.

Exact solutions for the case in which the upper surface is an inclined plane of angle θ and infinite width have been known for a long time, since the problem becomes one-dimensional and is easily integrated [7].

For more general shapes of the upper surfaces an existence result is obtained in [1] for the equation case and in [2] for the variational inequality in the one-dimensional case.

The content of the paper is as follows: In section 2 we give the main result. In section 3 we recall some elements of the degree theory which will be used. In section 4 some preliminary results are given, and section 5 is devoted to the proof of the main result.

2. Main results and assumptions. For the rest of the paper we assume that

$$\rho \in C^3(\mathbb{R}), \quad \rho,\ \rho',\ \text{and } \rho'' \text{ are } 2\pi\text{-periodic}, \tag{2.1}$$

$$\rho''(\theta) + \rho(\theta) > 0 \qquad \forall\, \theta \in \mathbb{R}, \tag{2.2}$$

$$\min_{0\le\theta\le 2\pi} \rho(\theta) = 1. \tag{2.3}$$

We denote the following:

$$\begin{aligned} \rho_M &= \max_{0\le\theta\le 2\pi} \rho(\theta), \\ m &= \min_{0\le\theta\le 2\pi} (\rho''(\theta) + \rho(\theta)) > 0, \\ M &= \max_{0\le\theta\le 2\pi} (\rho''(\theta) + \rho(\theta)) > 0. \end{aligned} \tag{2.4}$$

Remark 2.1. All these assumptions are clearly satisfied in the particular case of a circular bush, corresponding to $\rho \equiv 1$.

Remark 2.2. Assumption (2.2) is the most restrictive of the assumptions. It is introduced for technical reasons and guarantees that in the limit case the set where $h = 0$ is a single line.

Let us introduce the function $a(\alpha) : \mathbb{R} \to \mathbb{R}_+$ given by

$$a(\alpha) := \min_{\alpha-\frac{\pi}{2}<\theta<\alpha+\frac{\pi}{2}} \left\{ \frac{\rho(\theta)}{\cos(\theta-\alpha)} \right\}. \tag{2.5}$$

The fact that

$$\lim_{\theta\to\alpha\pm\frac{\pi}{2}} \frac{\rho(\theta)}{\cos(\theta-\alpha)} = \infty$$

and the continuity and boundedness of ρ guarantee the existence of at least one minimum.

Now we define the set A by

$$A = \Big\{ (\eta_1, \eta_2) \in \mathbb{R}^2 : 0 \le \eta < a(\alpha) \Big\},$$

where (η, α) are given by (1.2).

It is clear that $h(\theta) > 0\ \forall \theta \in [0, 2\pi]$ if and only if $(\eta_1, \eta_2) \in A$.

For any fixed $(\eta_1, \eta_2) \in A$, problem (1.5) has been studied by several authors. The existence and uniqueness of solutions can be obtained by using the direct methods in the calculus of variations along with strict convexity of the associated functional. Then (1.5) admits a unique classical solution $p \in K$; see, for instance, Kinderlehrer and Stampacchia [6].

Thus problem (1.5)–(1.7) is equivalent to the following problem:

$$\begin{cases} \text{Find } (\eta_1, \eta_2) \in A & \text{such that} \\ G(\eta_1, \eta_2) = (0, 0), \end{cases} \tag{2.6}$$

where $G = (G_1, G_2) : A \to \mathbb{R}^2$ is given by

$$G_1(\eta_1, \eta_2) = \int_\Omega p \cos\theta \, d\theta \, dx - F_1, \tag{2.7}$$

$$G_2(\eta_1, \eta_2) = \int_\Omega p \sin\theta \, d\theta \, dx - F_2, \tag{2.8}$$

where p (depending on η_1, η_2) is the unique solution of (1.5).

The main result of the paper is presented in the following theorem.

THEOREM 2.1. *Under assumptions* (2.1)–(2.3) *and for any* $F \in \mathbb{R}^2$ *there exists at least one solution* $(\eta_1, \eta_2) \in A$ *of* (2.6).

3. Known results on degree theory. In order to prove Theorem 2.1 we use the topological degree theory, which is rapidly recalled in the following.

The topological degree for continuous mappings between n-dimensional Euclidean spaces was first introduced by L. E. J. Brouwer in 1912. We first introduce the definition of degree for C^1 maps.

Let S be a bounded open subset of $\mathbb{R}^n$ and a $C^1(S)$ function $f : S \to \mathbb{R}^n$. Let $y_0 \in \mathbb{R}^n$ such that $y_0 \notin f(\partial S)$, and suppose that $f \in C^1(S)$ and that $Df(x)$ is invertible for all $x \in f^{-1}(y_0)$. Then $f(x) = y_0$ has either no solutions in S or a finite number r of solutions, say $x_1, x_2, \ldots, x_r$ and $\det(Df(x_i)) \neq 0$ for $i = 1, 2, \ldots, r$.

DEFINITION 3.1. *We define the degree of* f *in* S *at* y_0 *as follows:*

If $r = 0$, *then* $d(f, S, y_0) := 0$; *else*

$$d(f, S, y_0) := \sum_{i=1}^{r} sign(\det(Df(x_i))).$$

Remark 3.1. In the particular case where f is a linear function, i.e., $f(x) = Ax$, where A is an invertible matrix, the general formula for calculating the degree of linear functions is the following:

$$deg(Ax, S, 0) = sgn(\det A) \quad \text{if} \quad 0 \text{ is contained in } S.$$

The definition of degree can be extended to continuous maps; see [8, Extension Lemma, p. 60]. For the reader's convenience we state the following result (see, for instance, [8, Corollary 4 to Theorem 1.12, p. 81]) adapted to the finite-dimensional case.

THEOREM 3.1. *Let* S *be a bounded open set in* $\mathbb{R}^n$. *Let* f_0 *and* f_1 *be two continuous functions from* $\overline{S}$ *to* $\mathbb{R}^n$. *We assume moreover that* S *is a star domain with respect to the point* $y_0 \in S$ *and that*

$$[f_0(x) - y_0] \cdot [f_1(x) - y_0] > 0 \text{ for any } \ x \in \partial S.$$

Then

$$d(f_0, S, y_0) = d(f_1, S, y_0).$$

Remark 3.2. It is clear that if $d(f, S, y_0) \neq 0$, then there exists at least a solution $x \in S$ of the equation $f(x) = y_0$.

4. Preliminary results.

LEMMA 4.1. *Let* $a(\alpha)$ *be the function defined in* (2.5). *Then*

$$1 = \min_{0 \leq \theta \leq 2\pi} \rho \leq a(\alpha) \leq \max_{0 \leq \theta \leq 2\pi} \rho = \rho_M.$$

Proof. Since

$$\min_{0 \leq \theta \leq 2\pi} \rho \leq \rho \leq \max_{0 \leq \theta \leq 2\pi} \rho$$

we have that

$$\min_{0\le\theta\le 2\pi}\{\rho\}\ \min_{\alpha-\frac{\pi}{2}<\theta<\alpha+\frac{\pi}{2}}\left\{\frac{1}{\cos(\theta-\alpha)}\right\}\le a(\alpha)$$

and

$$a(\alpha)\le\max_{0\le\theta\le 2\pi}\{\rho\}\ \min_{\alpha-\frac{\pi}{2}<\theta<\alpha+\frac{\pi}{2}}\left\{\frac{1}{\cos(\theta-\alpha)}\right\}.$$

The fact that

$$\min_{\alpha-\frac{\pi}{2}<\theta<\alpha+\frac{\pi}{2}}\left\{\frac{1}{\cos(\theta-\alpha)}\right\}=1$$

implies

$$\min\{\rho\}\le a(\alpha)\le\max\{\rho\},$$

and the proof is complete. □

LEMMA 4.2. *There exists s with $0<s<\frac{\pi}{2}$ such that*

$$|\tilde{\theta}_\alpha-\alpha|<s$$

for any $\alpha\in\mathbb{R}$ and for any $\tilde{\theta}_\alpha\in]\alpha-\frac{\pi}{2},\ \alpha+\frac{\pi}{2}[$ satisfying

$$\frac{\rho(\tilde{\theta}_\alpha)}{\cos(\tilde{\theta}_\alpha-\alpha)}=a(\alpha).$$

Proof. From Lemma 4.1 we deduce

$$\cos(\tilde{\theta}_\alpha-\alpha)\ge\frac{1}{\rho_M},$$

which implies

$$|\tilde{\theta}_\alpha-\alpha|\le\arccos\frac{1}{\rho_M}<\frac{\pi}{2}. \tag{4.1}$$

Now taking $s=\frac{1}{2}(\frac{\pi}{2}+\arccos\frac{1}{\rho_M})$ we obtain the result. □

LEMMA 4.3. *Under assumptions (2.1)–(2.3), for any $\alpha\in\mathbb{R}$, the set of $\theta\in]\alpha-\frac{\pi}{2},\alpha+\frac{\pi}{2}[$, which minimizes the function $\theta\mapsto\frac{\rho(\theta)}{\cos(\theta-\alpha)}$, is a single point which we denote by θ_α satisfying*

$$\frac{\rho(\theta_\alpha)}{\cos(\theta_\alpha-\alpha)}=a(\alpha)$$

and

$$\rho'(\theta_\alpha)\cos(\theta_\alpha-\alpha)+\rho(\theta_\alpha)\sin(\theta_\alpha-\alpha)=0.$$

Moreover the function $\alpha\mapsto\theta_\alpha$ belongs to $C^1(\mathbb{R})$.

Proof. Let α be a fixed value of $\mathbb{R}$ and consider the critical points of the function $\theta\in]\alpha-\frac{\pi}{2},\alpha+\frac{\pi}{2}[\mapsto\frac{\rho(\theta)}{\cos(\theta-\alpha)}$, which satisfies

$$\left(\frac{\rho(\theta)}{\cos(\theta-\alpha)}\right)'=\frac{\rho'(\theta)}{\cos(\theta-\alpha)}+\frac{\rho(\theta)\sin(\theta-\alpha)}{\cos(\theta-\alpha)^2}=0. \tag{4.2}$$

Consider

$$f(\theta,\alpha) := \rho' \cos(\theta-\alpha) + \rho \sin(\theta-\alpha).$$

Notice that if θ is a minimum of

$$\frac{\rho(\theta)}{\cos(\theta-\alpha)},$$

then $f(\theta,\alpha)=0$. We now have that

$$\frac{\partial f}{\partial \theta} = (\rho'' + \rho)\cos(\theta-\alpha),$$

which is positive in $|\theta-\alpha| < \frac{\pi}{2}$ (by hypothesis (2.2)). The assertion follows using also the implicit function theorem on the open set $(\theta,\alpha) \in]-s,s[\times \mathbb{R}$ with s given in Lemma 4.2. □

We now introduce the function $h_0 : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$h_0(\theta,\alpha) = h(\theta, a(\alpha), \alpha) = \rho(\theta) - a(\alpha)\cos(\theta-\alpha).$$

Remark 4.1. We have from Lemma 4.3

$$h_0(\theta_\alpha,\alpha) = \frac{\partial h_0}{\partial \theta}(\theta_\alpha,\alpha) = 0, \tag{4.3}$$

$$\frac{\partial^2 h_0}{\partial \theta^2}(\theta_\alpha,\alpha) = \rho''(\theta_\alpha) + \rho(\theta_\alpha). \tag{4.4}$$

LEMMA 4.4. *Let us denote for any* $\alpha \in [0,2\pi]$ *and* $\epsilon > 0$ *small enough*

$$I_{\alpha,\epsilon} := [\theta_\alpha - 2\sqrt{\epsilon}, \theta_\alpha - \sqrt{\epsilon}].$$

Then for any $\alpha \in [0,2\pi]$ *and* $\theta \in I_{\alpha,\epsilon}$ *we have*

(i) $h(\theta, a(\alpha)-\epsilon, \alpha) \le (2M+2)\epsilon$,

(ii) $-\frac{\partial h}{\partial \theta}(\theta, a(\alpha)-\epsilon,\alpha) \ge \frac{m}{2}\sqrt{\epsilon}$,

with m *and* M *defined in* (2.4).

Proof.

(i) We have

$$h(\theta, a(\alpha)-\epsilon,\alpha) = h_0(\theta,\alpha) + \epsilon\cos(\theta-\alpha).$$

From Taylor development of h_0 at $\theta=\theta_\alpha$, using (4.3) and (4.4) we have (i).

(ii) We have

$$-\frac{\partial h}{\partial \theta}(\theta, a(\alpha)-\epsilon,\alpha) = -\frac{\partial h_0}{\partial \theta}(\theta,\alpha) + \epsilon\sin(\theta-\alpha).$$

Using the Taylor development of $\frac{\partial h_0}{\partial \theta}$ at $\theta=\theta_\alpha$ we obtain

$$-\frac{\partial h}{\partial \theta}(\theta, a(\alpha)-\epsilon,\alpha) = \frac{\partial^2 h_0}{\partial \theta^2}(\hat{\theta},\alpha)(\theta_\alpha-\theta) + \epsilon\sin(\theta-\alpha)$$

with $\hat{\theta} \in I_{\alpha,\epsilon}$.

From (4.4), the uniform continuity of $\frac{\partial^2 h_0}{\partial\theta^2}$, and (2.4) we deduce

$$\frac{\partial^2 h_0}{\partial\theta^2}(\hat{\theta},\alpha) \geq \frac{2}{3}m.$$

Since $\theta_\alpha - \theta \geq \sqrt{\epsilon}$, we obtain for ϵ small enough

$$-\frac{\partial h}{\partial\theta}(\theta, a(\alpha)-\epsilon, \alpha) \geq \frac{2}{3}m\sqrt{\epsilon} - \epsilon,$$

which ends the proof. □

LEMMA 4.5. *Let ϵ be small enough and $\eta := a(\alpha) - \epsilon$. There exists a constant $c > 0$ independent of ϵ and α such that*

$$\inf_{0\leq\alpha\leq 2\pi} \int_\Omega h^3(\theta,\eta,\alpha)|\nabla p|^2 d\theta dx > c\epsilon^{-\frac{1}{2}} \qquad \text{as } \epsilon \to 0.$$

Proof. We take in (1.5) $\phi := p + \varphi$ with $\varphi \in K$ arbitrary. Then

$$\int_\Omega h\frac{\partial\varphi}{\partial\theta} d\theta dx \leq \int_\Omega h^3 \nabla p \nabla\varphi d\theta dx.$$

By the Cauchy–Schwarz inequality we have

$$\int_\Omega h\frac{\partial\varphi}{\partial\theta} d\theta dx \leq \left|\int_\Omega h^3|\nabla p|^2 d\theta dx\right|^{\frac{1}{2}} \left|\int_\Omega h^3|\nabla\varphi|^2 d\theta dx\right|^{\frac{1}{2}}$$

and we deduce

$$\left|\int_\Omega h^3|\nabla p|^2 d\theta dx\right|^{\frac{1}{2}} \geq \sup_{\varphi\in K,\varphi\neq 0} \frac{-\int_\Omega \varphi\frac{\partial h}{\partial\theta} d\theta dx}{\left|\int_\Omega h^3|\nabla\varphi|^2 d\theta dx\right|^{\frac{1}{2}}}. \tag{4.5}$$

Let m be given by (2.4) and $\psi \in C^2(\mathbb{R})$ such that

(i) $\operatorname{supp}(\psi) \subset [-2,-1]$;
(ii) $\psi \geq 0$;
(iii) $\int_{\mathbb{R}} \psi > 0$.

As a consequence of (i)–(iii) we have

$$\frac{\int_{\mathbb{R}} \psi}{\left|\int_{\mathbb{R}} |\psi'|^2\right|^{\frac{1}{2}}} = c_1 > 0. \tag{4.6}$$

Let

$$\varphi_\epsilon(\theta, x) := \psi\left(\frac{\theta-\theta_\alpha}{\sqrt{\epsilon}}\right) x(1-x)$$

with θ_α defined in Lemma 4.3.

It is clear that $\operatorname{supp}(\varphi_\epsilon) = I_{\alpha,\epsilon} \times [0,1]$ with $I_{\alpha,\epsilon}$ as in Lemma 4.4. Using Lemma 4.4 (ii) we deduce

$$\begin{aligned} -\int_\Omega \frac{\partial h}{\partial\theta}\varphi_\epsilon d\theta dx &\geq \min_{\theta\in I_{\alpha,\epsilon}} \left(-\frac{\partial h}{\partial\theta}\right) \int_\Omega \varphi_\epsilon d\theta dx \\ &\geq \frac{m}{2}\epsilon \int_{\mathbb{R}} \psi dy \int_0^1 x(1-x)dx = \frac{m}{12}\epsilon\int_{\mathbb{R}}\psi dy. \end{aligned} \tag{4.7}$$

On the other hand, using Lemma 4.4 (i) we have

$$\left|\int_\Omega h^3|\nabla\varphi_\epsilon|^2 d\theta dx\right|^{\frac{1}{2}} \le \max_{I_{\alpha,\epsilon}}\{h^3\}^{\frac{1}{2}}\left|\int_\Omega |\nabla\varphi_\epsilon|^2 d\theta dx\right|^{\frac{1}{2}}$$
$$\le ((2M+2)\epsilon)^{3/2}\left|\int_\Omega |\nabla\varphi_\epsilon|^2 d\theta dx\right|^{\frac{1}{2}}.$$

A simple calculation gives

$$\int_\Omega |\nabla\varphi_\epsilon|^2 d\theta dx \le c_3\epsilon^{-1/2}.$$

We then have

$$\left|\int_\Omega h^3|\nabla\varphi_\epsilon|^2 d\theta dx\right|^{\frac{1}{2}} \le c_4\epsilon^{5/4}, \tag{4.8}$$

which, also using (4.7), gives us

$$\frac{-\int_\Omega \varphi_\epsilon \frac{\partial h}{\partial\theta} d\theta dx}{\left|\int_\Omega h^3|\nabla\varphi_\epsilon|^2 d\theta dx\right|^{1/2}} \ge c_5\epsilon^{-\frac{1}{4}}. \tag{4.9}$$

Finally from (4.5) and (4.9) we have the result. □

We now introduce for any ϵ small the function $\tilde{h}:\mathbb{R}^2\to\mathbb{R}$ defined by

$$\tilde{h}(\theta,\alpha) := h(\theta_\alpha, a(\alpha)-\epsilon,\alpha) + [\rho''(\theta_\alpha)+\rho(\theta_\alpha)](1-\cos(\theta-\theta_\alpha)). \tag{4.10}$$

Notice that

$$\begin{cases} h(\theta_\alpha, a(\alpha)-\epsilon,\alpha) = \tilde{h}(\theta_\alpha,\alpha) = \epsilon\cos(\theta_\alpha-\alpha), \\ \frac{\partial h}{\partial\theta}(\theta_\alpha, a(\alpha)-\epsilon,\alpha) - \frac{\partial\tilde{h}}{\partial\theta}(\theta_\alpha,\alpha) = \frac{\partial h}{\partial\theta}(\theta_\alpha, a(\alpha)-\epsilon,\alpha) = -\epsilon\sin(\theta_\alpha-\alpha), \\ \frac{\partial^2 h}{\partial\theta^2}(\theta_\alpha, a(\alpha)-\epsilon,\alpha) - \frac{\partial^2\tilde{h}}{\partial\theta^2}(\theta_\alpha,\alpha) = -\epsilon\cos(\theta_\alpha-\alpha). \end{cases} \tag{4.11}$$

LEMMA 4.6. *There exists a constant c independent of ϵ and α such that*

$$\int_0^{2\pi} \frac{\Big(h(\theta,a(\alpha)-\epsilon,\alpha)-\tilde{h}(\theta,\alpha)\Big)^2}{h^3(\theta,a(\alpha)-\epsilon,\alpha)} d\theta \le c.$$

Proof. By the Taylor polynomial of $h-\tilde{h}$ and h at θ_α and (4.11) we have

$$|h(\theta,a(\alpha)-\epsilon,\alpha)-\tilde{h}(\theta,\alpha)| \le \epsilon|\theta-\theta_\alpha| + \epsilon|\theta-\theta_\alpha|^2 + c|\theta-\theta_\alpha|^3 \tag{4.12}$$

and

$$\begin{aligned} h(\theta,a(\alpha)-\epsilon,\alpha) &= \epsilon\cos(\theta_\alpha-\alpha) - \epsilon\sin(\theta_\alpha-\alpha)(\theta-\theta_\alpha) \\ &+\frac{1}{2}\Big(\rho''(\theta_\alpha)+\rho(\theta_\alpha)-\epsilon\cos(\theta_\alpha-\alpha)\Big)(\theta-\theta_\alpha)^2 \\ &+\frac{1}{6}\frac{\partial^3 h}{\partial\theta^3}(\hat{\theta},a(\alpha)-\epsilon,\alpha)(\theta-\theta_\alpha)^3 \end{aligned} \tag{4.13}$$

with $\hat{\theta}\in\mathbb{R}$.

From Lemma 4.2, (4.13), and (2.2) there exists m_1, m_2, and m_3 positive and independent of ϵ and α such that

$$h(\theta, a(\alpha) - \epsilon, \alpha) \geq m_1\epsilon + m_2(\theta - \theta_\alpha)^2 \tag{4.14}$$

$\forall\ \theta$ such that $|\theta - \theta_\alpha| \leq m_3$.

We have from (4.12)

$$\int_0^{2\pi} \frac{(h - \tilde{h})^2}{h^3} d\theta \leq 2(\epsilon^2 I_1 + \epsilon^2 I_2 + c^2 I_3)$$

with $I_k = \int_0^{2\pi} \frac{(\theta - \theta_\alpha)^{2k}}{h^3} d\theta, k = 1, 2, 3$.

Now from (4.14) it is clear that I_3 is bounded uniformly in ϵ and α.

We will prove the uniform estimate for $\epsilon^2 I_1$. For $\epsilon^2 I_2$ the proof is similar.

We have

$$I_1 = I_1^1 + I_1^2 + I_1^3,$$

where I_1^1, I_1^2, I_1^3 are the subintegrals, respectively, in the intervals $|\theta - \theta_\alpha| \geq m_3, \epsilon^{1/3} \leq |\theta - \theta_\alpha| \leq m_3$, and $0 \leq |\theta - \theta_\alpha| \leq \epsilon^{1/3}$.

It is clear that $\epsilon^2 I_1^1$ is bounded uniformly in ϵ and α since h is lower bounded by a positive constant on the interval $|\theta - \theta_\alpha| \geq m_3$.

From (4.14) we have

$$I_1^2 \leq m_2^{-3} \int_{\epsilon^{1/3} \leq |\theta - \theta_\alpha| \leq m_3} \frac{d\theta}{|\theta - \theta_\alpha|^4} \leq c\epsilon^{-4/3}$$

and

$$I_1^3 \leq m_1^{-3}\epsilon^{-3} \int_{0 \leq |\theta - \theta_\alpha| \leq \epsilon^{1/3}} |\theta - \theta_\alpha|^2 d\theta \leq m_1^{-3}\epsilon^{-2},$$

which proves the lemma. □

5. Proof of Theorem 2.1. We apply the degree theory recalled in section 3 to the function G.

Since G is not defined on $\bar{A}$ we introduce for any $\epsilon > 0$ small enough the domain

$$A_\epsilon = \Big\{(\eta_1, \eta_2) : 0 \leq \eta \leq a(\alpha) - \epsilon\Big\}.$$

Let us now introduce the vector field

$$W : (\eta_1, \eta_2) \in A \to (\eta \sin\theta_\alpha, -\eta \cos\theta_\alpha) \in \mathbb{R}^2$$

with (η, α) defined in (1.2) and θ_α as in Lemma 4.3.

We observe that

$$W \cdot (-\eta_2, \eta_1) = -\eta^2 \cos(\theta_\alpha - \alpha).$$

From Lemmas 4.2 and 4.3 we have that

$$W \cdot (-\eta_2, \eta_1) < 0 \quad \forall (\eta_1, \eta_2) \in A.$$

From Theorem 3.1 we deduce that

$$deg(W, A_\epsilon, 0) = deg((-\eta_2, \eta_1), A_\epsilon, 0) = 1.$$

Notice that $deg(f, S, 0) = (-1)^n deg(-f, S, 0)$, where n is the dimension of the domain (for the application here, $n = 2$, so the sign does not change).

It suffices now to prove the inequality

$$G(\eta_1, \eta_2) \cdot W(\eta_1, \eta_2) > 0 \quad \forall (\eta_1, \eta_2) \in \partial A_\epsilon, \tag{5.1}$$

and the proof is finished using again Theorem 3.1.

We have, $\forall (\eta_1, \eta_2) \in \partial A_\epsilon$, $\eta = a(\alpha) - \epsilon$. Then

$$G(\eta_1, \eta_2) \cdot W(\eta_1, \eta_2) = -\eta \int_\Omega p \sin(\theta - \theta_\alpha) d\theta dx + \eta(-F_1 \sin\theta_\alpha + F_2 \cos\theta_\alpha). \tag{5.2}$$

Now taking $\varphi = 0$ and $\varphi = 2p$, respectively, in (1.5) we obtain

$$\int_\Omega h^3 |\nabla p|^2 d\theta dx = -\int_\Omega \frac{\partial h}{\partial \theta} p d\theta dx. \tag{5.3}$$

Notice that

$$\int_\Omega \frac{\partial h}{\partial \theta} p d\theta dx = \int_\Omega \frac{\partial}{\partial \theta}(h - \tilde{h}) p d\theta dx + \int_\Omega \frac{\partial \tilde{h}}{\partial \theta} p d\theta dx \tag{5.4}$$

with $\tilde{h}$ given by (4.10).

We now have

$$\begin{aligned} \int_\Omega \frac{\partial}{\partial \theta}(h - \tilde{h}) p d\theta dx &= -\int_\Omega (h - \tilde{h}) \frac{\partial p}{\partial \theta} d\theta dx \\ &\geq -\int_\Omega \frac{|h - \tilde{h}|}{h^{3/2}} h^{3/2} |\nabla p| d\theta dx \\ &\geq -\frac{1}{2} \int_\Omega \frac{|h - \tilde{h}|^2}{h^3} d\theta dx - \frac{1}{2} \int_\Omega h^3 |\nabla p|^2 d\theta dx. \end{aligned} \tag{5.5}$$

Since

$$\frac{\partial \tilde{h}}{\partial \theta} = (\rho''(\theta_\alpha) + \rho(\theta_\alpha)) \sin(\theta - \theta_\alpha)$$

we obtain from (5.3) and (5.5)

$$\begin{aligned} \int_\Omega h^3 |\nabla p|^2 d\theta dx \leq \frac{1}{2} \int_\Omega \frac{|h - \tilde{h}|^2}{h^3} + \frac{1}{2} \int_\Omega h^3 |\nabla p|^2 d\theta dx \\ - (\rho''(\theta_\alpha) + \rho(\theta_\alpha)) \int_\Omega \sin(\theta - \theta_\alpha) p d\theta dx, \end{aligned}$$

which implies, using also hypotheses (2.2),

$$-\int_\Omega \sin(\theta - \theta_\alpha) p d\theta dx \geq \frac{\int_\Omega h^3 |\nabla p|^2 d\theta dx - \int_\Omega \frac{|h - \tilde{h}|^2}{h^3} d\theta dx}{2(\rho''(\theta_\alpha) + \rho(\theta_\alpha))}.$$

Since ρ is in C^2 and from Lemmas 4.5 and 4.6, we obtain for ϵ small enough

$$-\int_\Omega \sin(\theta-\theta_\alpha)p \geq c(\epsilon^{-1/2}-1)$$

with $c>0$ a constant independent of ϵ and α.

Since $\eta = a(\alpha)-\epsilon$ and thanks to (5.2) we deduce

$$G(\eta_1,\eta_2)\cdot W(\eta_1,\eta_2) \geq \eta\Big(c(\epsilon^{-1/2}-1)-\|F\|\Big).$$

Taking ϵ small enough, we prove the theorem.

Remark 5.1. In the same manner we can prove the existence of at least one equilibrium solution for some other similar problems in this context. For instance, the case of Dirichlet boundary conditions on every boundary of Ω, or the case of a one-dimensional domain with Dirichlet boundary conditions.

Acknowledgments. The third author thanks INSA-Lyon and Institut Camille Jordan at University of Lyon for their hospitality.

REFERENCES

[1] G. Buscaglia, I. Ciuperca, I. Hafidi, and M. Jai, *Existence of equilibria in articulated bearings*, J. Math. Anal. Appl., 328 (2007), pp. 24–45.

[2] G. Buscaglia, I. Ciuperca, I. Hafidi, and M. Jai, *Existence of equilibria in articulated bearings in presence of cavity*, J. Math. Anal. Appl., 335 (2007), pp. 841–859.

[3] J. Durany, G. García, and C. Vázquez, *Numerical simulation of a lubricated Hertzian contact problem under imposed load*, Finite Elem. Anal. Des., 38 (2002), pp. 645–658.

[4] J. Durany, J. Pereira, and F. Varas, *Numerical solution of steady and transient problems in thermohydrodynamic lubrication using a combination of finite element, finite volume and boundary element methods*, Finite Elem. Anal. Des., 44 (2008), pp. 686–695.

[5] J. Frêne, D. Nicolas, B. Deguerce, D. Berthe, and M. Godet, *Lubrification Hydrodynamique*, Eyrolles, Paris, 1990.

[6] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.

[7] L. Leloup, *Etude de la Lubrification et Calcul des Paliers*, Dunod, Paris, 1962.

[8] E. H. Rothe, *Introduction to Various Aspects of Degree Theory in Banach Spaces*, Math. Surveys Monogr. 23, AMS, Providence, RI, 1986.

[9] O. Reynolds, *On the theory of lubrication and its application to Mr Beauchamp Tower's experiments, including an experimental determination of the viscosity of olive oil*, Phil. Trans. Roy. Soc. A, 117 (1886), pp. 157–234.

SIAM J. MATH. ANAL.
Vol. 40, No. 6, pp. 2328–2350

POINTWISE GREEN FUNCTION BOUNDS AND LONG-TIME STABILITY OF LARGE-AMPLITUDE NONCHARACTERISTIC BOUNDARY LAYERS*

SHANTIA YARAHMADIAN† AND KEVIN ZUMBRUN†

Abstract. Using pointwise semigroup techniques of Zumbrun–Howard and Mascia–Zumbrun, we obtain sharp global pointwise Green function bounds for noncharacteristic boundary layers of arbitrary amplitude. These estimates allow us to analyze linearized and nonlinearized stability of noncharacteristic boundary layers of one-dimensional systems of conservation laws, showing that both follow from (and linearized stability is equivalent to) a numerically checkable Evans function condition. Our results extend to the large-amplitude case results obtained for small amplitudes by Matsumura, Nishihara, and others using energy estimates.

Key words. boundary layer stability, Evans function, pointwise Green function bounds

AMS subject classifications. Primary, 35B35; Secondary, 76N20

DOI. 10.1137/080714804

1. Introduction.

Boundary layers appear in many physical settings, such as gas dynamics, MHD, and rotating fluids; see, for example, the physical discussion in [34]. In this paper, we study the stability of boundary layers assuming that the boundary layer solution is noncharacteristic, which means that signals are transmitted into or out of, but not along the boundary. Specifically, we consider a boundary layer, or stationary solution,

$$u = \bar{u}(x), \quad \lim_{x\to+\infty} \bar{u}(x) = u_+, \quad \bar{u}(0) = u_0 \tag{1.1}$$

of a system of conservation laws on the quarter-plane

$$u_t + f(u)_x = (B(u)u_x)_x, \quad x, t > 0, \tag{1.2}$$

$u, f \in \mathbb{R}^n$, $B \in \mathbb{R}^{n\times n}$, with initial data $u(x,0) = g(x)$ and Dirichlet boundary condition

$$u(0,t) = h(t). \tag{1.3}$$

A fundamental question is whether or not such boundary layer solutions are *stable* in the sense of PDE, i.e., whether or not a sufficiently small perturbation of $\bar{u}$ remains close to $\bar{u}$, or converges time-asymptotically to $\bar{u}$, under the evolution of (1.2).

Long-time stability of boundary layers has been considered for scalar equations in [24, 25] and for the equations of isentropic gas dynamics in [26, 23]. The latter results, obtained by energy estimates, apply to arbitrary amplitude layers of "expansive inflow" type analogous to rarefaction waves, but only to small-amplitude layers of "compressive inflow or outflow" type analogous to shock waves or "expansive outflow" type. For general symmetric hyperbolic-parabolic systems, stability of small-amplitude noncharacteristic boundary layers has been shown in multidimensions for

*Received by the editors January 31, 2008; accepted for publication (in revised form) October 9, 2008; published electronically February 25, 2009. This work was supported in part by the National Science Foundation grant number DMS-0300487.
http://www.siam.org/journals/sima/40-6/71480.html
†Department of Mathematics, Indiana University, Bloomington, IN 47405 (syarahma@indiana.edu, kzumbrun@indiana.edu).

strictly parabolic systems in [11], and in one dimension for partially parabolic ("real viscosity") systems in [30].

Here, in the spirit of results obtained for shock waves in [37, 27, 28], we show for general strictly parabolic systems of conservation laws that linearized and nonlinear stability follow from and linearized stability is equivalent to a generalized spectral stability condition phrased in terms of the Evans function associated with the linearized equations about the wave, *independent of the amplitude of the boundary layer in question.* The Evans function, introduced for boundary layers in [32], is described further in section 2; for its origins in the study of stability of traveling waves, see for example, [1, 10] and references therein.

The Evans condition is readily checkable numerically, and in some cases analytically; see [3, 4, 5, 6, 20, 2, 17]. In particular, stability of small-amplitude uniformly noncharacteristic boundary layers has been shown for general hyperbolic–parabolic systems in multidimensions in [13] using elementary Evans function arguments (convergence to the constant layer). An exhaustive numerical study for isentropic gas layers in one dimension has been carried out in [7], with the conclusion of stability for arbitrary amplitudes. Results of instability in some cases have been shown in [32, 33] using a mod two stability index in the spirit of [10].

Our method of analysis [35] is by pointwise Green function methods like those used in [37, 27, 28], and especially [16], to analyze the stability of viscous shock layers. Similar results have been obtained for the related small-viscosity-limit problem in [12, 29, 14]. In particular, we point to the analysis of Grenier and Rousset [12] as using pointwise Green function estimates very similar to those that we use here, though adapted for different purposes.

1.1. Equations and assumptions. Consider a *viscous boundary layer*, a standing-wave solution (1.1) of a general parabolic system of conservation laws (1.2).

Assume, similarly, as in the treatment of the viscous shock case in [16]:

(H0) $f, B \in C^3$.

(H1) $Re\,\sigma(B) > 0$.

(H2) $\sigma(f'(u_+))$ real, distinct, and nonzero.

(H3) $Re\,\sigma(-ikf'(u_+) - k^2 B(u_+)) < -\theta k^2$ for all real k and $\theta > 0$.

LEMMA 1.1 ([27, 36]). *Assuming* (H0)–(H3), *a solution* $\bar u$ *of* (1.1) *if it exists is also unique; moreover,*

$$\big|(d/dx)^k(\bar u - u_+)\big| \le Ce^{-\theta x}, \quad k = 0, \dots, 4, \tag{1.4}$$

as $x \to +\infty$, *for some* $\theta > 0$.

Proof. As in the shock case [28, 36], (1.4) follows by the observation that, under hypotheses (H0)–(H3), u_+ is a hyperbolic rest point of the layer profile ODE. Uniqueness follows by the observation [27] that the standing-wave ODE may be integrated from x to $+\infty$ and rearranged to yield

$$B(u)u' = f(u) - f(u_+) \tag{1.5}$$

together with the boundary conditions at $x = 0$, thus determining a unique solution for all $x \ge 0$. ∎

1.2. Linearized stability and the Evans function. After linearizing (1.2) about the stationary solution $\bar u$, we obtain the linearized equation

$$u_t = Lu := -(Au)_x + (Bu_x)_x, \quad A, B \in C^2, \tag{1.6}$$

where

$$B := B(\bar{u}) \tag{1.7}$$

and

$$Au := dF(\bar{u})u - dB(\bar{u})(u, \bar{u}_x). \tag{1.8}$$

DEFINITION 1.2. *The boundary layer $\bar{u}$ is said to be linearly asymptotically stable if $u(\cdot,t)$ approaches 0 as $t \to \infty$, for any solution u of (1.6) with initial data bounded in some specified norm.*

We define the following *stability criterion*, where $D(\lambda)$, described below, denotes the Evans function associated with the linearized operator L about the layer, an analytic function analogous to the characteristic polynomial of a finite-dimensional operator, whose zeroes away from the essential spectrum agree in location and multiplicity with the eigenvalues of L:

(1.9) There exist no zeroes of $D(\cdot)$ in the nonstable half-plane $Re\lambda \geq 0$.

As discussed, e.g., in [31], under assumptions (H0)–(H3), this is equivalent to *strong spectral stability*, $\sigma(L) \subset \{Re\lambda < 0\}$, *transversality* of $\bar{u}$ as a solution of the connection problem in the associated standing-wave ODE, and *hyperbolic stability* of an associated boundary value problem obtained by formal matched asymptotics. Here, and elsewhere, σ denotes spectrum of a linearized operator or matrix.

Our first main result is as follows.

THEOREM 1.3. *Assuming* (H0)–(H3), *linearized asymptotic $L^1 \cap L^p \to L^p$ stability, $p > 1$, is equivalent to* (1.9).

Theorem 1.3 is obtained as a consequence of the following detailed, pointwise bounds on the Green function $G(x,t;y)$ of the linearized evolution equations (1.6) with homogeneous boundary conditions (more properly speaking, a distribution), defined by:

(i) $(\partial_t - L_x)G = 0$ in the distributional sense, for all $x, y, t > 0$;
(ii) $G(x,t;y) \to \delta(x-y)$ as $t \to 0$;
(iii) $G(0,t;y) \equiv 0$, for all $y, t > 0$.

Denote by

$$a_1^+ < a_2^+ < \cdots < a_n^+ \tag{1.10}$$

the eigenvalues of of the limiting convection matrix $A_+ := df(u_+)$.

Then, our second main result is as follows.

THEOREM 1.4. *Assuming* (H0)–(H3) *and stability condition* (1.9),

$$\begin{aligned} |\partial_x^\gamma \partial_y^\alpha G(x,t;y)| &\leq Ce^{-\eta(|x-y|+t)} \\ &+ C\left(t^{-|\alpha|/2} + |\alpha|e^{-\eta|y|} + |\gamma|e^{-\eta|x|}\right)\left(\sum_{k=1}^{n} t^{-1/2} e^{-(x-y-a_k^- t)^2/Mt}\right. \\ &\left. + \sum_{a_k^+<0,\, a_j^+>0} \chi_{\{|a_k^+ t| \geq |y|\}} t^{-1/2} e^{-\left(x-a_j^+\left(t-|y/a_k^+|\right)\right)^2/Mt}\right), \end{aligned} \tag{1.11}$$

$0 \leq |\alpha|, |\gamma| \leq 1$, *for some η, C, $M > 0$, where x^+ denotes the positive/negative part of x, indicator function $\chi_{\{|a_k^- t| \geq |y|\}}$ is 1 for $|a_k^- t| \geq |y|$ and 0 otherwise.*

1.3. Nonlinear stability.

DEFINITION 1.5. *The boundary layer $\bar u$ is said to be nonlinearly asymptotically stable if $\tilde u(\cdot,t)$ exists for all $t\geq 0$ and approaches $\bar u$ as $t\to\infty$, for any solution $\tilde u$ of* (1.2) *with initial data sufficiently close in some norm to the original layer $\bar u$.*

Denoting by

$$a_1^+ < a_2^+ < \cdots < a_n^+ \tag{1.12}$$

the eigenvalues of of the limiting convection matrix $A_+ := df(u_+)$, define

$$\theta(x,t) := \sum_{a_j^+>0} (1+t)^{-1/2} e^{-|x-a_j^+ t|^2/Lt}, \tag{1.13}$$

$$\psi_1(x,t) := \chi(x,t) \sum_{a_j^+>0} (1+|x|+t)^{-1/2} \left(1+\left|x-a_j^+ t\right|\right)^{-1/2}, \tag{1.14}$$

and

$$\psi_2(x,t) := (1-\chi(x,t)) \left(1+\left|x-a_n^+ t\right|+t^{1/2}\right)^{-3/2}, \tag{1.15}$$

where $\chi(x,t)=1$ for $x\in[0,a_n^+ t]$ and $\chi(x,t)=0$ otherwise and $L>0$ is a sufficiently large constant. For simplicity, take B identically constant. Then, our third and final main result is as follows.

THEOREM 1.6. *Assuming* (H0)–(H3), *$B\equiv$ constant, and the linear stability condition* (1.9), *the profile $\bar u$ is nonlinearly asymptotically stable with respect to perturbations g, h in initial and boundary data satisfying*

$$|g(x)|\leq E_0(1+|x|)^{-3/2}, \quad |h(t)|\leq E_0(1+|t|)^{-3/2}, \quad |h'(t)|\leq E_0(1+|t|)^{-1}$$

for E_0 sufficiently small. More precisely,

$$|\tilde u(x,t)-\bar u(x)| \leq CE_0(\theta+\psi_1+\psi_2)(x,t), \tag{1.16}$$

where $\tilde u$ denotes the solution of (1.2) *with initial data $\tilde g=\bar u+g$ and boundary data $\tilde h = u_0+h$.*

Remark 1.7. Pointwise bound (1.16) yields as a corollary the sharp L^p decay rate

$$|\tilde u(x,t)-\bar u(x)|_{L^p} \leq CE_0(1+t)^{-\frac12(1-\frac1p)}, \quad 1\leq p\leq\infty. \tag{1.17}$$

1.4. Discussion and open problems. The case of boundary layers is quite analogous to the undercompressive shock case; in particular, pointwise estimates as in [16] appear to be necessary to close the one-dimensional analysis by a linearized semigroup approach suitable for large-amplitude layers. (On the other hand, small-amplitude stability has been established using energy estimates in, e.g., [26, 11].) A new feature of the present analysis as compared to those of [16, 18, 19] is the admission of perturbations in boundary as well as initial data. Open problems are extensions to systems with physical (partial) or quasilinear viscosity and to multidimensional boundary layers.

2. The Evans function. Before starting the analysis, we review the basic Evans function methods and gap/conjugation lemma.

2.1. The gap/conjugation lemma. Consider a family of first-order ODE systems on the half-line:

$$\begin{aligned} W' &= \mathbb{A}(x,\lambda)W, \quad \lambda \in \Omega \quad \text{and} \quad x > 0, \\ \mathbb{B}(\lambda)W &= 0, \quad \lambda \in \Omega \quad \text{and} \quad x = 0. \end{aligned} \tag{2.1}$$

These systems of ODEs should be considered as a generalized eigenvalue equation, with λ representing frequency. We assume that the boundary matrix $\mathbb{B}$ is analytic in λ and that the coefficient matrix $\mathbb{A}$ is analytic in λ as a function from Ω into $L^\infty(x)$, C^K in x, and approaches exponentially to a limit $\mathbb{A}_+(\lambda)$ as $x \to \infty$, with uniform exponentially decay estimates

$$\left|(\partial/\partial x)^k(\mathbb{A} - \mathbb{A}_+)\right| \le C_1 e^{-\theta|x|/C_2}, \quad \text{for } x > 0,\ 0 \le k \le K, \tag{2.2}$$

C_j, $\theta > 0$, on compact subsets of Ω. Now we can state a refinement of the "Gap Lemma" of [10, 21], relating solutions of the variable-coefficient ODE to the solutions of its constant-coefficient limiting equations

$$Z' = \mathbb{A}_+(\lambda)Z \tag{2.3}$$

as $x \to +\infty$.

LEMMA 2.1 (conjugation Lemma [29]). *Under assumption* (2.2)*, there exists locally to any given $\lambda_0 \in \Omega$ a linear transformation $P_+(x,\lambda) = I + \Theta_+(x,\lambda)$ on $x \ge 0$, Φ_+ analytic in λ as functions from Ω to $L^\infty[0,+\infty)$, such that:*

(i) *$|P_+|$ and their inverses are uniformly bounded, with*

$$\left|(\partial/\partial\lambda)^j(\partial/\partial x)^k\Theta_+\right| \le C(j)C_1C_2e^{-\theta|x|/C_2} \quad \text{for } x > 0,\ 0 \le k \le K+1, \tag{2.4}$$

$j \ge 0$, where $0 < \theta < 1$ is an arbitrary fixed parameter, and $C > 0$ and the size of the neighborhood of definition depend only on θ, j, the modulus of the entries of $\mathbb{A}$ at λ_0, and the modulus of continuity of $\mathbb{A}$ on some neighborhood of $\lambda_0 \in \Omega$.

(ii) *The change of coordinates $W := P_+Z$ reduces* (2.1) *on $x \ge 0$ to the asymptotic constant-coefficient equations* (2.3)*. Equivalently, solutions of* (2.1) *may be conveniently factorized as*

$$W = (I + \Theta_+)Z_+, \tag{2.5}$$

where Z_+ are solutions of the constant-coefficient equations, and Θ_+ satisfy (2.4)*.*

Proof. As described in [27], for $j = k = 0$ this is a straightforward corollary of the gap lemma as stated in [Z.3], applied to the "lifted" matrix-valued ODE

$$P' = \mathbb{A}_+P - P\mathbb{A} + (\mathbb{A} - \mathbb{A}_+)P$$

for the conjugating matrices P_+. The x-derivative bounds $0 < k \le K+1$ then follow from the ODE and its first K derivatives. Finally, the λ-derivative bounds follow from standard interior estimates for analytic functions. ∎

DEFINITION 2.2. *Following* [1]*, we define the* domain of consistent splitting *for the ODE system $W' = \mathbb{A}(x,\lambda)W$ as the (open) set of λ such that the limiting matrix $\mathbb{A}_+$ is hyperbolic (has no center subspace) and the boundary matrix $\mathbb{B}$ is full rank, with* $\dim S_+ = \operatorname{rank}\mathbb{B}$.

LEMMA 2.3. *On any simply connected subset of the domain of consistent splitting, there exists an analytic basis $\{v_1,\dots,v_k\}$ for the subspace S_+ defined in Definition* 2.2.

Proof. By spectral separation of S_+ from the complementary unstable subspace U_+, the associated (group) eigenprojection is analytic. The existence of an analytic eigenbasis then follows by a standard result of Kato; see [22], pp. 99–102. ∎

COROLLARY 2.4. *By the Conjugation Lemma, on the domain of consistent splitting, the stable manifold of solutions decaying as $x \to +\infty$ of* (2.1) *is*

$$\mathcal{S}^+ := \text{span}\left\{P_+v_1^+, \dots, P_+v_k^+\right\}, \tag{2.6}$$

where $W_+^j := P_+v_j^+$ are analytic in λ and C^{K+1} in x for $\mathbb{A} \in C^K$.

2.2. Definition of the Evans function. On any simply connected subset of the domain of consistent splitting, let $W_1^+, \dots, W_k^+ = P_+v_1^+, \dots, P_+v_k^+$ be the analytic basis described in Corollary 2.4 of the subspace $\mathcal{S}^+$ of solutions W of (2.1) satisfying the boundary condition $W \to 0$ at $+\infty$. Then, the *Evans function* for the ODE systems $W' = \mathbb{A}(x,\lambda)W$ associated with this choice of limiting bases is defined as the $k \times k$ Gramian determinant

$$\begin{aligned} D(\lambda) &:= \det\left(\mathbb{B}W_1^+, \dots, \mathbb{B}W_k^+\right)_{|x=0,\lambda} \\ &= \det\left(\mathbb{B}P_+v_1^+, \dots, \mathbb{B}P_+v_k^+\right)_{|x=0,\lambda}. \end{aligned} \tag{2.7}$$

Remark 2.5. Note that D is independent of the choice of P_+ as, by uniqueness of stable manifolds, the exterior products (minors) $P_+v_1^+ \wedge \cdots \wedge P_+v_k^+$ are uniquely determined by their behavior as $x \to +\infty$.

PROPOSITION 2.6. *Both the Evans function and the subspace $\mathcal{S}^+$ are analytic on the entire simply connected subset of the domain of consistent splitting on which they are defined. Moreover, for λ within this region, equation* (2.1) *admits a nontrivial solution $W \in L^2(x > 0)$ if and only if $D(\lambda) = 0$.*

Proof. Analyticity follows by uniqueness, and local analyticity of P_+, v_k^+. Noting that the first $P_+v_j^+$ are a basis for the stable manifold of (2.1) at $x \to +\infty$, we find that the determinant of $\mathbb{B}P_+v_j^+$ vanishes if and only if $\mathbb{B}(\lambda)$ has nontrivial kernel on $\mathcal{S}_+(\lambda, 0)$, whence, the second assertion follows. ∎

Remark 2.7. In the case that the ODE system describes an eigenvalue equation associated with an ordinary differential operator L, Proposition 2.6 implies that eigenvalues of L agree in location with zeroes of D. (Indeed, they agree also in multiplicity; see [8, 9]; Lemma 6.1, [37]; or Proposition 6.15 of [27].)

When $\ker \mathbb{B}$ has an analytic basis $v_{k+1}^0, \dots, v_{N-k}^0$, for example, in the commonly occurring case, as here, that $\mathbb{B} \equiv$ constant, we have the following useful alternative formulation. This is the version that we will use in our analysis of the Green function and resolvent kernel.

PROPOSITION 2.8. *Let $v_{k+1}^0, \dots, v_{N-k}^0$ be an analytic basis of $\ker \mathbb{B}$, normalized so that $\det(\mathbb{B}^*, v_{k+1}^0, \dots v_N^0) \equiv 1$. Then, the solutions W_j^0 of* (2.1) *determined by initial data $W_j^0(\lambda, 0) = v_j^0$ are analytic in λ and C^{K+1} in x, and*

$$D(\lambda) := \det\left(W_1^+, \dots, W_k^+, W_{k+1}^0, \dots, W_N^0\right)_{|x=0,\lambda}. \tag{2.8}$$

Proof. Analyticity/smoothness follow by analytic/smooth dependence on initial data/parameters. By the chosen normalization, and standard properties of Grammian determinants, $D(\lambda) = \det(W_1^+, \dots, W_k^+, v_{k+1}^0, \dots, v_N^0)_{|x=0,\lambda}$, yielding (2.8). ∎

3. Construction of the resolvent kernel. In this section we construct the explicit form of the resolvent kernel, which is nothing more than the Green function

$G_\lambda(x,y)$ associated with the elliptic operator $(L-\lambda I)$, where

$$(L-\lambda I)G_\lambda(.,y)=\delta_y I,\quad G_\lambda(0,y)\equiv 0. \tag{3.1}$$

Let Λ be the region of consistent splitting for L. It is an established fact (see [15]) that the resolvent $(L-\lambda I)^{-1}$ and the Green function $G_\lambda(x,y)$ are meromorphic in λ on Λ, with isolated poles of finite order. G_λ, in fact, admits a meromorphic extension to a sector

$$\Omega_\theta=\{\lambda : Re(\lambda)\geq -\theta_1-\theta_2|Im(\lambda)|\},\quad \theta_1,\theta_2>0. \tag{3.2}$$

Writing the associated eigenvalue equation in the form of a first-order system (2.1), we obtain

$$W'=\mathbb{A}(\lambda,x)W,\quad \mathbb{B}W(0)=0, \tag{3.3}$$

where

$$W=\begin{pmatrix} w\\ w'\end{pmatrix}\in\mathbb{C}^{2n},\qquad \mathbb{A}=\begin{pmatrix} 0 & I\\ \lambda B^{-1}+A'B^{-1} & AB^{-1}-B'B^{-1}\end{pmatrix},$$

and $\mathbb{B}\equiv$ constant is the rank-n projection onto the first coordinate w of W, with kernel spanned by the constant basis $v^0_{n+j}=e_{n+j}$, $j=1,\dots,n$ and e_j the jth standard basis element.

Denote by

$$\Phi^0=\begin{pmatrix}\phi^0_1(x;\lambda) & \cdots & \phi^0_n(x;\lambda)\end{pmatrix}=\begin{pmatrix}W^0_1 & \cdots & W^0_n\end{pmatrix} \tag{3.4}$$

and

$$\Phi^+=\begin{pmatrix}\phi^+_1(x;\lambda) & \cdots & \phi^+_n(x;\lambda)\end{pmatrix}=\begin{pmatrix}W^+_{n+1} & \cdots & W^+_{2n}\end{pmatrix}=\begin{pmatrix}P_+v^+_1 & \cdots & P_+v^+_k\end{pmatrix} \tag{3.5}$$

the matrices whose columns span the subspaces of solutions of (2.1) decaying at $x=0,+\infty$, respectively, denoting (analytically chosen) complementary subspaces by

$$\Psi^0=\begin{pmatrix}\psi^0_1(x;\lambda) & \cdots & \psi^0_n(x;\lambda)\end{pmatrix}=\begin{pmatrix}W^0_{n+1} & \cdots & W^0_{2n}\end{pmatrix} \tag{3.6}$$

and

$$\Psi^+=\begin{pmatrix}\psi^+_1(x;\lambda) & \cdots & \psi^+_n(x;\lambda)\end{pmatrix}=\begin{pmatrix}W^+_1 & \cdots & W^+_n\end{pmatrix}. \tag{3.7}$$

As described in the previous subsection, eigenfunctions decaying at both $0,+\infty$ occur precisely when the subspaces span Φ^0 and span Φ^+ intersect, i.e., at zeros of the Evans function defined in (2.8):

$$D_L(\lambda):=\det\left(\Phi^0,\Phi^+\right)_{|x=0}=\left(\phi^0_1\wedge\cdots\wedge\phi^0_n\wedge\phi^+_1\wedge\cdots\wedge\phi^+_n\right)_{|x=0}. \tag{3.8}$$

LEMMA 3.1 ([10, 37]). *For $\theta_1,\theta_2>0$ sufficiently small, D_L is locally analytic on sector Ω_θ as defined in* (3.2).

Proof. Direct calculation showing that the domain Λ of consistent splitting is contained in $\Omega_\theta - B(0,r)$ for $r>0$ arbitrary and θ sufficiently small, with v^+_j extending analytically to $B(0,r)$. ∎

LEMMA 3.2. *Let $H_\lambda(x,y)$ denote the Green function for the adjoint operator $(L-\lambda I)^*$ on the half-plane $x \geq 0$. Then $G_\lambda(x,y) = H^*_\lambda(x,y)$. In particular, for $x \neq y$, the matrix $z = G_\lambda(x,.)$ satisfies*

$$(z'B)' = -z'A + z\lambda. \tag{3.9}$$

Proof. Standard duality argument; see [37] for operators on the whole line. ∎

Considering (3.9) as an ODE system for the vector $Z = (z, z')$, it becomes

$$Z' = Z\tilde{\mathbb{A}}(\lambda, x), \tag{3.10}$$

where

$$\tilde{\mathbb{A}} = \begin{pmatrix} 0 & \lambda B^{-1} - A'B^{-1} \\ I & -AB^{-1} - B'B^{-1} \end{pmatrix}. \tag{3.11}$$

LEMMA 3.3 ([37]). *Z is a solution of (3.11) if and only if $Z\mathcal{S}W \equiv$ constant for any solution W of (2.1), where $\mathcal{S} = \begin{pmatrix} -A & B \\ -B & 0 \end{pmatrix}$.*

Proof. Direct computation/comparison with 0 of $(Z\mathcal{S}W)'$; see [37]. ∎

Using Lemma 3.3, we can define dual bases $\tilde{W}^0_j$ and $\tilde{W}^+_j$ by the relations

$$\tilde{W}^{0,+}_j \mathcal{S} W^{0,+}_k = \delta^j_k. \tag{3.12}$$

Likewise, $\tilde{\mathbb{A}}_{0,+}$ can be defined as

$$\tilde{\mathbb{A}}_{0,+} = \begin{pmatrix} 0 & \lambda B^{-1}_{0,+} \\ I & -A_{0,+}B^{-1}_{0,+} \end{pmatrix}. \tag{3.13}$$

We define also the dual subspaces

$$\tilde{\Phi}^0 = \begin{pmatrix} \tilde{\phi}^0_1(x;\lambda) & \cdots & \tilde{\phi}^0_n(x;\lambda) \end{pmatrix} = \begin{pmatrix} \tilde{W}^0_{n+1} & \cdots & \tilde{W}^0_{2n} \end{pmatrix}, \tag{3.14}$$

$$\tilde{\Phi}^+ = \begin{pmatrix} \tilde{\phi}^+_1(x;\lambda) & \cdots & \tilde{\phi}^+_n(x;\lambda) \end{pmatrix} = \begin{pmatrix} \tilde{W}^+_1 & \cdots & \tilde{W}^+_n \end{pmatrix}, \tag{3.15}$$

$$\tilde{\Psi}^0 = \begin{pmatrix} \tilde{\psi}^0_1(x;\lambda) & \cdots & \tilde{\psi}^+_n(x;\lambda) \end{pmatrix} = \begin{pmatrix} \tilde{W}^0_1 & \cdots & \tilde{W}^0_n \end{pmatrix}, \tag{3.16}$$

$$\tilde{\Psi}^+ = \begin{pmatrix} \tilde{\psi}^+_1(x;\lambda) & \cdots & \tilde{\psi}^0_n(x;\lambda) \end{pmatrix} = \begin{pmatrix} \tilde{W}^+_{n+1} & \cdots & \tilde{W}^+_{2n} \end{pmatrix}. \tag{3.17}$$

With these preparations, the construction of the resolvent kernel goes exactly as in the construction performed in [37, 27] on the whole line.

LEMMA 3.4. *We have the the representation*

$$\begin{pmatrix} G_\lambda & G_{\lambda_y} \\ G_{\lambda_x} & G_{\lambda_{xy}} \end{pmatrix} = \begin{cases} \Phi^+(\lambda,x)M^+(\lambda)\tilde{\Psi}^0(\lambda,y) & \text{for} \quad x > y, \\ \Phi^0(\lambda,x)M^0(\lambda)\tilde{\Psi}^+(\lambda,y) & \text{for} \quad x < y, \end{cases} \tag{3.18}$$

where $M^{0,+}$ are to be determined.

Proof. See [37] Lemma 4.6. ∎

Using Lemma 3.4, we find the explicit coordinate-free representation for $x > y$:

$$\begin{pmatrix} G_\lambda & G_{\lambda_y} \\ G_{\lambda_x} & G_{\lambda_{xy}} \end{pmatrix} = \mathcal{F}^{z\to x}\Pi_+(z)\mathcal{S}^{-1}(z)\tilde{\Pi}_0(z)\tilde{\mathcal{F}}^{z\to y}, \tag{3.19}$$

where

$$\Pi_+(y) = \big(\Phi^+(y), 0\big)\big(\Phi^+(y), \Phi^-(y)\big)^{-1}, \tag{3.20}$$

$$\tilde{\Pi}_0(y) = \begin{pmatrix} \tilde{\Psi}^0(y) \\ \tilde{\Psi}^+(y) \end{pmatrix}^{-1} \begin{pmatrix} \tilde{\Psi}^0(y) \\ 0 \end{pmatrix}, \tag{3.21}$$

$$\mathcal{F}^{z\to x} = \big(\Phi^+(x), \Phi^0(x)\big)\big(\Phi^+(z), \Phi^0(z)\big)^{-1}, \tag{3.22}$$

$$\tilde{\mathcal{F}}^{z\to y} = \begin{pmatrix} \tilde{\Psi}^0(z) \\ \Phi^+(z) \end{pmatrix} \begin{pmatrix} \Psi^0(y) \\ \Psi^+(y) \end{pmatrix}^{-1}, \tag{3.23}$$

and similarly for $x < y$.

COROLLARY 3.5. *The resolvent kernel may be expressed as*

$$G_\lambda(x,y) = \begin{cases} (I_n, 0)\Phi^+(x;\lambda)M^+(\lambda)\tilde{\Psi}^{0*}(y;\lambda)(I_n,0)^{tr} & x > y, \\ -(I_n,0)\Phi^0(x;\lambda)M^0(\lambda)\tilde{\Psi}^{+*}(y;\lambda)(I_n,0)^{tr} & x < y, \end{cases} \tag{3.24}$$

where

$$M(\lambda) := \operatorname{diag}\big(M^+(\lambda), M^0(\lambda)\big) = \Phi^{-1}(z;\lambda)\bar{\mathcal{S}}^{-1}(z)\tilde{\Psi}^{-1*}(z;\lambda). \tag{3.25}$$

4. Low-frequency bounds. Our goal in this section is the estimation of the resolvent kernel in the critical regime $|\lambda| \to 0$, i.e., the large time behavior of the Green function G, or global behavior in space and time. We are basically following the same treatment as that carried out for viscous shock waves of strictly parabolic conservation laws in [37, 27]; we refer to those references for details. In the low frequency case the behavior is essentially governed by the equation

$$U_t = L_+U := -A_+U_x + B_+U_{xx}. \tag{4.1}$$

PROPOSITION 4.1. *Assuming* (H0)–(H3), *let K be the order of the pole of G_λ at $\lambda = 0$ and r be sufficiently small that there are no other poles in $B(0,r)$. Then for $\lambda \in \Omega_\theta$ such that $|\lambda| \le r$ and for $x > y > 0$ we have*

$$\begin{pmatrix} G_\lambda & G_{\lambda_y} \\ G_{\lambda_x} & G_{\lambda_{xy}} \end{pmatrix} = \sum_{j,k} d_{jk}(\lambda)\phi_j^+(x)\tilde{\psi}_k^+(y) + \sum_k \phi_k^+(x)\tilde{\phi}_k^+(y), \tag{4.2}$$

where $d_{jk}(\lambda) = \mathcal{O}(\lambda^{-K})$ is a scalar meromorphic function, moreover, $K \le$ order of vanishing of the Evans function $D(\lambda)$ at $\lambda = 0$.

Proof. See [37], Proposition 7.1 for the first statement and Theorem 6.3 for the second statement linking order K of the pole to multiplicity of the zero of the Evans function. ∎

LEMMA 4.2. *Assuming* (H0)–(H3), *for $|\lambda|$ sufficiently small, the eigenvalue equation $(L_+ - \lambda)W = 0$ associated with the limiting, constant-coefficient operator L_+ has a basis of $2n$ solutions $\bar{W}_j^+ = e^{\mu_j^+(\lambda)x}V_j(\lambda)$ where μ_j^+ and V_j^+ are analytic in λ, consisting of n fast modes*

$$\begin{aligned} \mu_j^+ &= \gamma_j^+ + \mathcal{O}(\lambda), \\ V_j^+ &= S_j^+ + \mathcal{O}(\lambda), \end{aligned} \tag{4.3}$$

where γ_j^+, S_j^+ are eigenvalues and associated right eigenvectors of $B_+^{-1}A_+$, and n slow modes

$$\begin{aligned}\mu_{r+j}^+(\lambda) &:= -\lambda/a_j^+ + \lambda^2\beta_j^+/{a_j^+}^3 + \mathcal{O}\left(\lambda^3\right),\\ V_{r+j}^+(\lambda) &:= r_j^+ + \mathcal{O}(\lambda),\end{aligned} \tag{4.4}$$

where a_j^+, l_j^+, r_j^+ are eigenvalues and left and right eigenvectors of $A_+ := dF(u_+)$, and $\beta_j^+ := l_j^+ B_+ r_j^+ > 0$ with $B_+ := B(u_+)$. The same is true for the adjoint eigenvalue equation

$$\left(L^+ - \lambda\right)^* Z = 0;$$

i.e, it has a basis of solutions

$$\bar{\tilde{W}}_j^+ = e^{-\mu_j^+(\lambda)x}\tilde{V}_j(\lambda)$$

with

$$\tilde{V}_j^+(\lambda) = \tilde{T}_j^+ + \mathcal{O}(\lambda), \tag{4.5}$$

$$\tilde{V}_{r+j}^+(\lambda) = l_j^+ + \mathcal{O}(\lambda), \tag{4.6}$$

$\tilde{V}^+$ analytic in λ.

Proof. See [27]. □

PROPOSITION 4.3. *Assume* (H0)–(H3) *and* (1.9), *then, for $r > 0$ sufficiently small, the resolvent kernel G_λ associated with the linearized evolution equation*

$$U_t = L_+U := -A_+U_x + B_+U_{xx} \tag{4.7}$$

satisfies, for $0 \le y \le x$:

(4.8)

$$\begin{aligned}|\partial_x^\gamma\partial_y^\alpha G_\lambda(x,t;y)| \le C\left(|\lambda|^\gamma + e^{-\theta|x|}\right)\left(|\lambda|^\alpha + e^{-\theta|y|}\right)\Bigg(&\sum_{a_k^+>0}\left|e^{\left(-\lambda/a_k^+ + \lambda^2\beta_k^+/{a_k^+}^3\right)(x-y)}\right|\\ &+ \sum_{a_k^+<0,\, a_j^+>0}\left|e^{\left(-\lambda/a_j^+ + \lambda^2\beta_j^+/{a_j^+}^3\right)x + \left(\lambda/a_k^+ - \lambda^2\beta_k^+/{a_k^+}^3\right)y}\right|\Bigg),\end{aligned}$$

$0 \le |\alpha|, |\gamma| \le 1$, $\theta > 0$, with similar bounds for $0 \le x \le y$. Moreover, each term in the summation on the right-hand side of (4.8) *bounds a separately analytic function.*

Proof. By 1.8 D does not vanish on $Re(\lambda) \ge 0$, hence, by continuity, on $|\lambda| \le r$. Thus, according to (4.2), all $|d_{jk}(\lambda)|$ are uniformly bounded on $|\lambda| \le r$, and so it is enough to find estimates for fast and slow modes ϕ_j^+, $\tilde{\phi}_j^+$, ψ_j^+, and $\tilde{\psi}_j^+$. By using (3.5) we find:

$$\begin{pmatrix}\phi_j^+\\ \partial_x\phi_j^+\end{pmatrix} = e^{\mu_j(\lambda)x}P^+\begin{pmatrix}V_j\\ \mu_jV_j\end{pmatrix} = e^{\mu_j(\lambda)x}(I+\Theta)\begin{pmatrix}V_j\\ \mu_jV_j\end{pmatrix} \tag{4.9}$$

and similarly for $\tilde{\phi}_j^+$, ψ_j^+, and $\tilde{\psi}_j^+$. Now using (2.4) and the fact, by (4.4), that $e^{\mu_j(\lambda)x}$ is of order $e^{-|\theta x|}$ for fast modes and order $e^{-\lambda/a_j^+ + \lambda^2\beta_j^+/{a_j^+}^3 + \mathcal{O}(\lambda^3)}$ for slow modes, substituting this and corresponding dual estimates in (4.9) and grouping terms, we obtain the result. □

5. High frequency bounds. To analyze the high frequency behavior of the Green function of the boundary layer, we first establish some bounds for the projection terms in the Green function, using the symmetric formula

$$\begin{pmatrix} G_\lambda(x,y) & \partial_y G_\lambda(x,y) \\ \partial_x G_\lambda(x,y) & \partial_x\partial_y G_\lambda(x,y) \end{pmatrix} = \begin{cases} \mathcal{F}^{y\to x}\Pi_+(x)\mathcal{S}^{-1}(y) & if \quad x>y, \\ \tilde{\mathcal{F}}^{x\to y}\tilde{\Pi}_+(x)\mathcal{S}^{-1}(x) & if \quad x<y. \end{cases} \tag{5.1}$$

By setting $\bar{x} = |\lambda^{\frac{1}{2}}|x$, $\bar{\lambda} = \frac{\lambda}{|\lambda|}$, $\bar{B}(\bar{x}) = B(\frac{\bar{x}}{\lambda^{\frac{1}{2}}})$, $\bar{w}(\bar{x}) = w(\frac{x}{\lambda^{\frac{1}{2}}})$ in the eigenvalue equation $Lw = \lambda w$ associated with (1.6) we obtain

$$\bar{W}' = \bar{\mathbb{B}}\bar{W} + \mathcal{O}\left(\left|\lambda^{-\frac{1}{2}}\right|\right)\bar{W} \tag{5.2}$$

where

$$\mathbb{B} = \begin{pmatrix} 0 & I \\ \bar{\lambda}\bar{B} & 0 \end{pmatrix} \tag{5.3}$$

and $\mathbb{B}' = \mathcal{O}(|\lambda^{-\frac{1}{2}}|)$ and $|\bar{\lambda}| = 1$. Since $\mathbb{B}(\lambda, \bar{x})$ varies within a compact set, then there are C^1 eigenprojections P_0 and P_+ with property $|P'_+| = \mathcal{O}(|\lambda^{-\frac{1}{2}}|)$ and $|P'_0| = \mathcal{O}(|\lambda^{-\frac{1}{2}}|)$ taking $\bar{W}$ onto the stable and unstable subspaces. By using the two new coordinates $Y_+ = P_+\bar{W}$ and $Y_0 = P_0\bar{W}$, we obtain

$$\begin{pmatrix} Y_+ \\ Y_0 \end{pmatrix}' = \begin{pmatrix} A_+ & 0 \\ 0 & A_0 \end{pmatrix}\begin{pmatrix} Y_+ \\ Y_0 \end{pmatrix} + \mathcal{O}\left(\left|\lambda^{-\frac{1}{2}}\right|\right)\begin{pmatrix} y \\ y \end{pmatrix}. \tag{5.4}$$

Equivalently, we can find continuous invertible transformations Q_+, Q_0 such that $E_+ = Q_+A_+Q_+^{-1}$ and $E_0 = Q_0A_0Q_0^{-1}$, where

$$Re(E) := \frac{1}{2}(E_+ + E_+^*) < -\beta^{-\frac{1}{2}}I \tag{5.5}$$

in the sense of quadratic forms.

Again, by coordinate change $Z_+ = Q_+Y_+$, $Z_0 = Q_0Y_0$, we find

$$\begin{pmatrix} z_+ \\ z_0 \end{pmatrix}' = \begin{pmatrix} E_+ & 0 \\ 0 & E_0 \end{pmatrix}\begin{pmatrix} z_+ \\ z_0 \end{pmatrix} + \mathcal{O}\left(\left|\lambda^{-\frac{1}{2}}\right|\right)\begin{pmatrix} z \\ z \end{pmatrix}, \tag{5.6}$$

where

$$\frac{|\bar{w}|}{C} \le |z| \le C|\bar{w}|. \tag{5.7}$$

From this we find by energy estimate that

$$\left(|z_+|^2\right)' < -2\beta^{-\frac{1}{2}}|z_+|^2 \tag{5.8}$$

and, hence,

$$\frac{|z_+(x)|}{|z_+(y)|} \le e^{-\tilde{\beta}^{-\frac{1}{2}}|(x-y)|} \tag{5.9}$$

for any solution z_+ decaying at ∞, where $\tilde{\beta} < \beta$ and, thus,

$$\frac{|z_(x)|}{|z_(y)|} \le e^{-\beta^{-\frac{1}{2}}|(x-y)|} \tag{5.10}$$

for $x > y$, provided that $|\lambda|$ is sufficiently large. From this we obtain

$$\frac{\left|\bar{W}(x)\right|}{\left|\bar{W}(y)\right|} \leq C^2 e^{-\beta^{-\frac{1}{2}}|(x-y)|} \tag{5.11}$$

where C is as in (5.7).

Applying a symmetric argument for the adjoint equation, we obtain the following lemma.

LEMMA 5.1. *On the manifolds Φ_+ and $\tilde{\Psi}_+$ defined in* (3.5) *and* (3.17)*, for λ sufficiently large, within the sector $\Omega_\theta = \{\lambda : Re(\lambda) \geq -\theta_1 - \theta_2 |Im(\lambda)|\}$, $\theta_1, \theta_2 > 0$, we have in rescaled coordinates $\bar{x}$, for some uniform $C > 0$,*

$$|\mathcal{F}^{y\to x}|, \left|\tilde{\mathcal{F}}^{x\to y}\right| \leq Ce^{-\frac{|y-x|}{C}} \tag{5.12}$$

for $x > y$ and $x < y$, respectively.

LEMMA 5.2. *In rescaled coordinates $\bar{x}$, $\bar{\lambda}$, for the projection terms $\Pi_+(y)$ and $\tilde{\Pi}_+(x)$, the projection along Φ_0 onto Φ_+, for λ sufficiently large,*

$$|\Pi_+(y)|, \left|\tilde{\Pi}_+(x)\right| < C \tag{5.13}$$

for some uniform $C > 0$.

Proof. Choosing the coordinates $\begin{pmatrix} W_1 \\ W_2 \end{pmatrix} \in \mathbb{C}^{2n}$ where $W^j = \begin{pmatrix} W_1^j \\ W_2^j \end{pmatrix}$, we show for small enough ϵ and fixed $c > 0$ such that $\frac{|W_2^j|}{|W_1^j|} \leq c\epsilon$ and $\frac{|W_1^j|}{|W_2^j|} < c$ the projection along $E = \operatorname{span}(W^1, \ldots, W^n)$ onto $F = \operatorname{span}(W^{n+1}, \ldots, W^{2n})$

$$\Pi := \left(w^1, \ldots, w^{2n}\right)(O_n, I_n)\left(w^1, \ldots, w^{2n}\right)^{-1} \tag{5.14}$$

satisfies

$$|\Pi| \leq C_2(c, \epsilon). \tag{5.15}$$

To show this without loss of generality we assume that

$$\left(w^{n+1}, \ldots, w^{2n}\right) = \begin{pmatrix} I_n \\ \mathcal{O}(\epsilon) \end{pmatrix}\left(w^1, \ldots, w^n\right) = \begin{pmatrix} \mathcal{O}(1) \\ I_n \end{pmatrix}. \tag{5.16}$$

Now it is sufficient to show that

$$\left|\left(w^1, \ldots, w^{2n}\right)\right| \leq C_2(c, \epsilon). \tag{5.17}$$

However, this amounts to showing that

$$\left|\begin{pmatrix} M & I_n \\ I_n & O \end{pmatrix} + \mathcal{O}(\epsilon))^{-1}\right| \leq C, \tag{5.18}$$

which amounts to showing that

$$\left|\begin{pmatrix} M & I_n \\ I_n & O \end{pmatrix}^{-1}\right| \leq C_2(c), \tag{5.19}$$

where $|M| \leq c$. However, this is easy to show because

$$\begin{pmatrix} M & I_n \\ I_n & O \end{pmatrix}^{-1} = \begin{pmatrix} I_n & -M \\ O & I_n \end{pmatrix},$$

and so

$$\left| \begin{pmatrix} M & I_n \\ I_n & O \end{pmatrix}^{-1} \right| \leq 1 + |M| \leq C. \qquad \square$$

PROPOSITION 5.3. *Assume* (H0)–(H3) *and* (1.9). *Then, for* $R > 0$ *sufficiently large, the resolvent kernel* G_λ *associated with the linearized evolution equation* (4.7) *satisfies, for* $c, C > 0$ *and* $0 \leq |\alpha|, |\gamma| \leq 1$*:*

$$|\partial_x^\gamma \partial_y^\alpha G_\lambda(x,y)| \leq C|\lambda|^{\left(\frac{|\alpha|+|\gamma|-1}{2}\right)} e^{-\sqrt{\lambda}\frac{|y-x|}{c}}. \tag{5.20}$$

Proof. Recalling the coordinate-free representation (5.1) and combining with (5.12) and (5.13), we find that the Green function $\bar{G}_{\bar{\lambda}}$ in rescaled coordinates $\bar{x}$, $\bar{\lambda}$ satisfies

$$\left|\partial_x^\gamma \partial_y^\alpha \bar{G}_{\bar{\lambda}}(\bar{x},\bar{y})\right| \leq C e^{-\frac{|y-x|}{C}}, \tag{5.21}$$

whence, (5.20) follows in the original coordinates. $\square$

Remark 5.4. The argument of Lemma 5.2 is the key new ingredient in the resolvent estimates for the boundary layer case as compared to the analysis on the whole line carried out for viscous shock layers in [37], making essential use of compatibility of the boundary condition with high-frequency behavior. On the whole line, there is no such requirement and high-frequency stability is automatic.

6. Pointwise Green function bounds. With the pointwise bounds established on the resolvent kernel G_λ, we obtain pointwise bounds on the Green function through the inverse Laplace transform formula by a simplified version of the stationary-phase arguments used in [37] for the shock case, repeated here for completeness.

Proof of Theorem 1.4. By sectoriality of L, we have the inverse Laplace transform representation (see [37]):

$$G(t;x,y) = \int_\Gamma e^{\lambda t} G_\lambda(x,y) d\lambda. \tag{6.1}$$

Let $\theta_1 > 0$, $\theta_2 > 0$ be chosen sufficiently small, in particular so small as to satisfy the hypotheses of all previous assertions. By assumption (1.9), the large-$|\lambda|$ bounds on the resolvent kernel, and analyticity of the Evans function $D_L(\lambda)$, it follows that G_λ has finitely many poles in Ω_θ (corresponding to roots of D_L), each with strictly negative real part. Choosing θ_1, θ_2 still smaller, if necessary, we can, thus, arrange that G_λ *is analytic on* Ω_θ. It follows from Cauchy's Theorem that

$$G(x,t;y) = \int_\Gamma e^{\lambda t} G_\lambda(x,y)\, d\lambda \tag{6.2}$$

for any contour Γ that can be expressed as $\Gamma = \partial(\Omega_\theta \setminus \mathcal{S})$ for $\mathcal{S} \subset \mathbb{C}$ open.

Case I. $|x-y|/t$ *large.* We first treat the trivial case that $|x-y|/t \geq S$, S sufficiently large, the regime in which standard short-time parabolic theory applies. Set

$$\bar{\alpha} := \frac{|x-y|}{2\beta t}, \quad R := \beta \bar{\alpha}^2, \tag{6.3}$$

where β is as in (5.5), and consider again the representation of G, that is

$$G(x,t;y) = \int_{\Gamma_1\cup\Gamma_2} e^{\lambda t} G_\lambda(x,y)\, d\lambda, \tag{6.4}$$

where $\Gamma_1 := \partial B(0,R) \cap \bar\Omega_\theta$ and $\Gamma_2 := \partial\Omega_\theta \setminus B(0,R)$. Note that the intersection of Γ with the real axis is $\lambda_{min} = R = \beta\bar\alpha^2$. □

By the large $|\lambda|$ estimates of Proposition 5.3, we have for all $\lambda \in \Gamma_1 \cup \Gamma_2$ that

$$|G_\lambda(x,y)| \le C \frac{e^{-\sqrt{|\lambda|}\frac{|y-x|}{c}}}{\sqrt{|\lambda|}}.$$

Further, we have

$$Re\lambda \le R\left(1-\eta\omega^2\right), \quad \lambda\in\Gamma_1, Re\lambda \quad \le Re\lambda_0 - \eta(|Im\lambda| - |Im\lambda_0|), \quad \lambda\in\Gamma_2 \tag{6.5}$$

for R sufficiently large, where ω is the argument of λ and λ_0 and λ_0^* are the two points of intersection of Γ_1 and Γ_2, for some $\eta > 0$ independent of $\bar\alpha$. Combining these estimates, we obtain

$$\begin{aligned}\left|\int_{\Gamma_1} e^{\lambda t} G_\lambda d\lambda\right| &\le \int_{\Gamma_1} C|\lambda|^{-\frac12}\, e^{Re\lambda t - \beta^{-\frac12}|\lambda|^{-\frac12}|x-y|}\, d\lambda \\ &\le Ce^{-\beta\bar\alpha^2 t}\int_{-L}^{+L} R^{-\frac12} e^{-\beta R\eta\omega^2 t}\, Rd\omega \qquad \le Ct^{-\frac12}e^{-\beta\bar\alpha^2 t}.\end{aligned} \tag{6.6}$$

Likewise,

$$\begin{aligned}\left|\int_{\Gamma_2} e^{\lambda t} G_\lambda d\lambda\right| &\le \int_{\Gamma_2} C|\lambda|^{-\frac12}\, Ce^{Re\lambda t - \beta^{-\frac12}|\lambda|^{-\frac12}|x-y|} d\lambda \\ &\le Ce^{Re(\lambda_0)t - |\beta|^{-\frac12}|\lambda_0|^{-\frac12}|x-y|}\int_{\Gamma_2} |\lambda|^{-\frac12} e^{(Re\lambda - Re\lambda_0)t}\ |d\lambda| \\ &\le Ce^{-\beta\bar\alpha^2 t}\int_{\Gamma_2} |Im\,\lambda|^{-\frac12} e^{-\eta|Im\,\lambda - Im\,\lambda_0|t}\ |d\,Im\lambda| \\ &\le Ct^{-\frac12}e^{-\beta\bar\alpha^2 t}.\end{aligned} \tag{6.7}$$

Combining these last two estimates, we have

$$|G(x,t;y)| \le Ct^{-\frac12} e^{\frac{-\beta\bar\alpha^2 t}{2}} e^{\frac{-(x-y)^2}{8\beta t}} \le Ct^{-\frac12} e^{-\eta t} e^{\frac{-(x-y)^2}{8\beta t}}, \tag{6.8}$$

for $\eta > 0$ independent of $\bar\alpha$. Observing that $\frac{|x-at|}{2t} \le \frac{|x-y|}{t} \le \frac{2|x-at|}{t}$ for any bounded a, for $\frac{|x-y|}{t}$ sufficiently large, we find that this contribution may be absorbed in any summand $t^{\frac{-1}{2}} e^{\frac{-(x-y-a_k^+ t)^2}{Mt}}$.

Case II. $|x-y|/t$ *bounded.* We now turn to the critical case that $|x-y|/t \le S$. A few remarks are in order at the outset. Our goal is to bound $|G|$ by terms of form $Ct^{-1/2}e^{-\bar\alpha^2 t/M}$, where $\bar\alpha := (x - a_j^+(t - |y/a_k^+|))/2t$ or $\bar\alpha := (x-y-a_k^+ t)/2t$ are now *uniformly bounded*, by

$$|x-y|/2t + \max_j \left\{\left|a_j^+\right|\right\}/2 \le S/2 + \max\left|a_j^+\right|/2. \tag{6.9}$$

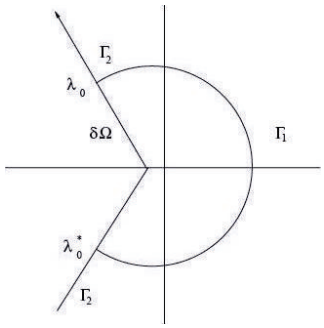


FIG. 1.

Thus, in particular, *contributions of order* $t^{-1/2}e^{-\eta t}$, $\eta > 0$, *can be absorbed in any summand* $t^{-1/2}e^{-(x-y-a_k^+ t)^2/Mt}$) if we take M sufficiently large. Likewise, for G_x and G_y, contributions of order $t^{-1}e^{-\eta t}$ can be absorbed. We will use this observation repeatedly.

In contrast to the previous case of large characteristic speed $|x-y|/t \geq S$, we are not trying to show rapid time-exponential decay. Rather, we are trying to show that the rate of exponential decay of the solution does not degrade too rapidly as $\bar{\alpha} \to 0$: precisely, that it vanishes to order $\bar{\alpha}^2$ and no more. Thus, the crucial part of our analysis will be for small $\bar{\alpha}$. All other situations can be estimated crudely as described just above.

Let r be sufficiently small that the small-$|\lambda|$ bounds hold on $B(0,r)$. Next, choose θ_1 and θ_2 still smaller than before, if necessary, so that $\Omega_\theta \setminus B(0,r) \subset \Lambda$. This implies that $\partial\Omega_\theta \cap B(0,r) \neq \emptyset$, giving the configuration pictured in Figure 1. Similarly as in the previous case, define $\Gamma = \Gamma_1 \cup \Gamma_2$, where Γ_1 is the portion of the circle $\partial B(0,r)$ contained in $\bar{\Omega}_\theta$, and Γ_2 is the portion of $\partial\Omega_\theta$ outside $B(0,r)$.

$$G(x,t;y) = \int_{\Gamma_1} e^{\lambda t} G_\lambda(x,y)\, d\lambda + \int_{\Gamma_2} e^{\lambda t} G_\lambda(x,y)\, d\lambda. \tag{6.10}$$

We separately estimate the terms $\int_{\Gamma_1}$ and $\int_{\Gamma_2}$.

Large- and medium-λ *estimates.* The $\int_{\Gamma_2}$ term is straightforward. The points λ_0, λ_0^* where Γ_1 meets Γ_2 satisfy $Re(\lambda_0) = -\eta < 0$. Moreover, combining the results low-frequency case, we have the bound $|G_\lambda| \leq C|\lambda|^{-\frac{1}{2}}$ for $\lambda \in \Gamma_2$. Thus, we have

$$\begin{aligned} \left| \int_{\Gamma_2} e^{\lambda t} G_\lambda d\lambda \right| &\leq C e^{-Re\,\lambda_0 t} \int_{\Gamma_2} |Im\,\lambda|^{-\frac{1}{2}} e^{-\eta |Im\,\lambda - Im\,\lambda_0| t}\, |d\,Im\lambda| \\ &\leq C t^{-\frac{1}{2}} e^{-\eta t}. \end{aligned} \tag{6.11}$$

This contribution can be absorbed as described above. An analogous computation using $|G_{\lambda_x}|$, $|G_{\lambda_x}| \leq C|\lambda|^{-1}$ shows that the Γ_2 contribution to G_x and G_y is $O(t^{-1}e^{-\eta t})$, and can likewise be absorbed.

Small $|\lambda|$ *estimates.* It remains to estimate the critical term $\int_{\Gamma_1} e^{\lambda t} G_\lambda d\lambda$. This we will estimate in different ways, depending on the size of t.

Bounded time. For t bounded, we can use the medium-λ bounds $|G_\lambda|$, $|G_{\lambda_x}|$, $|G_{\lambda_y}| \leq C$ to obtain $|\int_{\Gamma_1} e^{\lambda t} G_\lambda d\lambda| \leq C_2 |\Gamma_1|$. This contribution is order $Ce^{-\eta t}$ for bounded time, and hence can be absorbed.

Large time. For t large, we must instead estimate $\int_{\Gamma_1} e^{\lambda t} G_\lambda d\lambda$ using the small-$|\lambda|$ expansions. First, observe that all coefficient functions $d_{jk}(\lambda)$ are uniformly bounded (since $|\lambda|$ is bounded in this case).

Expanding $G = \int_\Gamma e^{\lambda t} G_\lambda(x,y) d\lambda$ as

$$\begin{pmatrix} G & G_x \\ G_y & G_{xy} \end{pmatrix} = \int_\Gamma e^{\lambda t} \begin{pmatrix} G_\lambda & G_{\lambda_x} \\ G_{\lambda_y} & G_{\lambda_{xy}} \end{pmatrix} d\lambda,$$

we estimate the $\int_{\Gamma_1}$ contributions to G, G_x, and G_y simultaneously.

Case II(i). $(0 < y < x)$. By our low-frequency estimates, we have

$$\begin{aligned} \int_\Gamma e^{\lambda t} \begin{pmatrix} G_\lambda & G_{\lambda_x} \\ G_{\lambda_y} & G_{\lambda_{xy}} \end{pmatrix} d\lambda &= \int_\Gamma \sum_{j,k} e^{\lambda t} \phi_j^+(x) d_{jk} \tilde{\psi}_k^+(y) d\lambda \\ &\quad + \int_\Gamma \sum_{j,k} e^{\lambda t} \psi_k^+(x) \tilde{\psi}_k^+(y) d\lambda, \end{aligned} \tag{6.12}$$

where each d_{jk} is analytic and, hence, bounded. We estimate separately each of the terms

$$\int_{\Gamma_1} e^{\lambda t} \phi_j^+(x) d_{jk} \tilde{\psi}_k^+(y) d\lambda$$

on the right-hand side of (6.12). Estimates for terms

$$\int_\Gamma \sum_{j,k} e^{\lambda t} \psi_k^+(x) \tilde{\psi}_k^+(y) d\lambda$$

go similarly.

Case II(ia). First, consider the critical case $a_j^+ > 0$, $a_k^+ < 0$. For this case,

$$\left| \phi_{j(x)}^+ d_{jk} \tilde{\psi}_k^+(y) \right| \le C e^{Re\left(\rho_j^+ x - \nu_k^+ y\right)},$$

where

$$\begin{cases} \nu_k^+(\lambda) = -\lambda/a_k^+ + \lambda^2 \beta_k^+ / \left(a_k^+\right)^3 + \mathcal{O}\left(\lambda^3\right) \\ \rho_j^+(\lambda) = -\lambda/a_j^+ + \lambda^2 \beta_j^+ / \left(a_j^+\right)^3 + \mathcal{O}\left(\lambda^3\right). \end{cases}$$

Set

$$\bar{\alpha} = \frac{a_k^+ x/a_j^+ - y - a_k^+ t}{2t}, \quad p := \frac{\beta_j^+ a_k^+ x/\left(a_j^+\right)^3 - \beta_k^+ y/\left(a_k^+\right)^2}{t} > 0.$$

Define Γ'_{1a} to be the portion contained in Ω_θ of the hyperbola

(6.13)

$$\begin{aligned} &Re\left(\rho_j^+ x - \nu_k^+ y\right) + \mathcal{O}\left(\lambda^3\right)(|x| + |y|) \\ &\quad = \left(1/a_k^+\right) Re\left[\lambda\left(-a_k^+ x/a_j^+ + y\right) + \lambda^2\left(x\beta_j^+ a_k^+/\left(a_j^+\right)^3 - y\beta_k^-/\left(a_k^+\right)^2\right)\right] \\ &\quad \equiv \text{constant} \\ &\quad = \left(1/a_k^-\right)\left[\lambda_{min}\left(-a_k^- x/a_j^+ + y\right) + \lambda_{min}^2\left(x\beta_j^+ a_k^+/\left(a_j^+\right)^3 - y\beta_k^+/\left(a_k^+\right)^2\right)\right], \end{aligned}$$

where

$$\lambda_{min} := \begin{cases} \frac{\bar\alpha}{p} & if \quad \left|\frac{\bar\alpha}{p}\right| \le \epsilon, \\ \pm\epsilon & if \quad \frac{\bar\alpha}{p} \gtrless \epsilon. \end{cases} \tag{6.14}$$

Denoting by λ_1, λ_1^* the intersections of this hyperbola with $\partial\Omega_\theta$, define Γ'_{1_b} to be the union of $\lambda_1\lambda_0$ and $\lambda_0^*\lambda_1^*$, and define $\Gamma'_1 = \Gamma'_{1_a} \cup \Gamma'_{1_b}$. Note that $\lambda = \bar\alpha/p$ minimizes the left-hand side of (6.13) for λ real. Note also that p is bounded for $\bar\alpha$ sufficiently small, since $\bar\alpha \le \epsilon$ implies that

$$\left(\left|a_k^+ x/a_j^+\right| + |y|\right)/t \le 2\left|a_k^+\right| + 2\epsilon;$$

i.e., $(|x| + |y|)/t$ is controlled by $\bar\alpha$.

With these definitions, we readily obtain that

$$\begin{aligned} Re\left(\lambda t + \rho_j^+ x - \nu_k^+ y\right) &\le -\left(t/a_k^-\right)\left(\bar\alpha^2/4p\right) - \eta Im(\lambda)^2 t \\ &\le -\bar\alpha^2 t/M - \eta Im(\lambda)^2 t, \end{aligned} \tag{6.15}$$

for $\lambda \in \Gamma'_{1_a}$ (note: here, we have used the crucial fact that $\bar\alpha$ controls $(|x| + |y|)/t$, in bounding the error term $\mathcal{O}(\lambda^3)(|x| + |y|)/t$ arising from expansion). Likewise, we obtain for any q that

$$\int_{\Gamma'_{1_a}} |\lambda|^q e^{Re\left(\lambda t + \rho_j^+ x - \nu_k^- y\right)} d\lambda \le C t^{-\frac{1}{2}-\frac{q}{2}} e^{-\bar\alpha^2 t/M}, \tag{6.16}$$

for suitably large C, $M > 0$ (depending on q). Observing that

$$\bar\alpha = \left(a_k^+/a_j^+\right)\left(x - a_j^+\left(t - \left|y/a_k^+\right|\right)\right)/2t,$$

we find that the contribution of (6.16) can be absorbed in the described bounds for $t \ge |y/a_k^-|$. At the same time, we find that $\bar\alpha \ge x > 0$ for $t \le |y/a_k^+|$, whence,

$$\bar\alpha \ge \left(x - y - a_j^+ t\right)/Mt + |x|/M,$$

for some $\epsilon > 0$ sufficiently small and $M > 0$ sufficiently large.

This gives

$$e^{-\bar\alpha^2/p} \le e^{-\left(x-y-a_k^+ t\right)^2/Mt} e^{-\eta|x|}$$

provided $|x|/t > a_j^+$, a contribution which can again be absorbed. On the other hand, if $t \le |x/a_j^+|$, we can use the dual estimate

$$\begin{aligned} \bar\alpha &= \left(-y - a_k^+\left(t - \left|x/a_j^+\right|\right)\right)/2t \\ &\ge \left(x - y - a_k^+ t\right)/Mt + |y|/M, \end{aligned} \tag{6.17}$$

together with $|y| \ge |a_k^- t|$, to obtain

$$e^{-\bar\alpha^2/p} \le e^{-\left(x-y-a_j^+ t\right)^2/Mt} e^{-\eta|y|},$$

a contribution that can likewise be absorbed.

Case II(ib). In case $a_j^+ < 0$ or $a_k^+ > 0$, terms $|\varphi_j^+| \le Ce^{-\eta|x|}$ and $|\tilde\psi_j^+| \le Ce^{-\eta|y|}$ are strictly smaller than those already treated in Case II(ia), so may be absorbed in previous terms.

Case II(ii) $(0 < x < y)$. The case $0 < x < y$ can be treated very similarly to the previous one; see [37] for details. This completes the proof of Case II, and the theorem.

7. Nonlinear analysis. Introducing the perturbation variable

$$u(x,t) := \tilde{u}(x,t) - \bar{u}(x), \tag{7.1}$$

we obtain

$$u_t - Lu = Q(u)_x, \tag{7.2}$$

where the second-order Taylor remainder satisfies

$$Q(u) := f(\bar{u}+u) - f(\bar{u}) - df(\bar{u})u = \mathcal{O}\left(|u|^2\right) \tag{7.3}$$

so long as $|u|$ remains bounded.

LEMMA 7.1 (integral formulation). *Under the assumptions of Theorem* 1.6, *there exists a classical solution of* (7.2) *for* $0 < t \le T$, $T > 0$, *continuous in* $L^\infty(x)$ *at* $t = 0$, *extending for all* $t > 0$ *such that* $u(\cdot,t)$ *remains sufficiently small in* $L^1 \cap L^\infty$, *given by*

$$\begin{aligned} u(x,t) = &\int_0^\infty G(x,t;y)g(y)\,dy + \int_0^t G_y(x,t-s;0)Bh(s)\,ds \\ &- \int_0^t \int_0^\infty G_y(x,t-s;y)Q(u)(y,s)\,dy\,ds. \end{aligned} \tag{7.4}$$

Proof. From Lemma 3.2 and the inverse Laplace representation (6.1) we find that $G(x,t-s;y)$ considered as a function of y,s satisfies the adjoint equation

$$(\partial_s - L_y)^* G^*(x,t-\cdot;\cdot) = 0, \tag{7.5}$$

or

$$-G_s - (GA)_y + GA_y = (G_y B)_y. \tag{7.6}$$

Likewise, reviewing the construction of the resolvent, we find $G_\lambda(x,0) \equiv 0$, yielding

$$G(x,t-s;0) \equiv 0. \tag{7.7}$$

That is, $G^*(x,t-\cdot;\cdot)$ is the Green function for the adjoint equation, as may alternatively be seen directly by a duality argument analogous to the proof of Lemma 3.2.

Thus, integrating G against (7.2), integrating by parts, and using the fact that $G = 0$ and $u = h$ on the boundary $y = 0$, we obtain for any classical solution of (7.2) that

$$\begin{aligned} &\int_0^t \int_0^\infty G(x,t-s;y)Q(u(y,s))_y\,dy\,ds = \\ &\qquad \int_0^t \int_0^\infty G(x,t-s;y)(\partial_s - L_y)u(y,s)\,dy\,ds \\ &\qquad = \int_0^t \int_0^\infty ((\partial_s - L_y)^* G^*)^*(x,t-s;y)u(y,s)\,dy\,ds \\ &\qquad + u(x,t) - \int_0^\infty G(x,t;y)g(y)\,dy - \int_0^t G_y(x,t-s;0)Bh(s)\,ds, \end{aligned} \tag{7.8}$$

from which we obtain (7.4) by rearranging and integrating by parts the term $\int_0^t \int_0^\infty G(x,t-s;y)Q(u(y,s))_y\,dy\,ds$.

Indeed, (7.4) may be taken as the definition of a weak solution in $L^\infty(x,t)$. (One can see using convolution identities that this agrees with the usual definition in terms of integration against test functions $\phi \in C_0^\infty(\mathbb{R}\times\mathbb{R})$.) Existence of weak solutions can be obtained by a standard contraction mapping/continuation argument using the convolution bounds of Lemmas 7.2–7.4 below; we omit the details, since we shall carry out quite similar but more difficult estimates in the proof of stability. Smoothness of solutions may then be obtained by a bootstrapping argument as sketched in Appendix A. ∎

To establish stability, we use the following lemmas proved in [16].

LEMMA 7.2 (linear estimates [16]). *Under the assumptions of Theorem* 1.6,

$$\int_0^{+\infty} |G(x,t;y)|(1+|y|)^{-3/2}\,dy \le C(\theta+\psi_1+\psi_2)(x,t), \tag{7.9}$$

for $0\le t\le+\infty$, *some* $C>0$.

LEMMA 7.3 (nonlinear estimates [16]). *Under the assumptions of Theorem* 1.6,

$$\int_0^t\int_0^{+\infty} |G_y(x,t-s;y)|\Psi(y,s)\,dyds \le C(\theta+\psi_1+\psi_2)(x,t), \tag{7.10}$$

for $0\le t\le+\infty$, *some* $C>0$, *where*

$$\begin{aligned}\Psi(y,s) := {}& (1+s)^{1/2}s^{-1/2}(\theta+\psi_1+\psi_2)^2(y,s)\\ &+(1+s)^{-1}(\theta+\psi_1+\psi_2)(y,s).\end{aligned} \tag{7.11}$$

We require also the following estimate accounting boundary effects.

LEMMA 7.4 (boundary estimate). *Under the assumptions of Theorem* 1.6,

$$\left|\int_0^t G_y(x,t-s;0)Bh(s)\,ds\right| \le CE_0(\theta+\psi_1+\psi_2)(x,t), \tag{7.12}$$

for $0\le t\le+\infty$, *some* $C>0$.

Proof. The estimate on $\int_0^{t-1}$, where $G_y(x,t-s;0)$ is nonsingular, follows readily by estimates similar to but somewhat simpler than those of Lemma (7.3), which we, therefore, omit.

To bound the singular part $\int_{t-1}^t$, we integrate (7.6) in y from 0 to $+\infty$, recalling that $G(x,t-s;0)\equiv 0$, to obtain

$$G_yB = -\int_0^{+\infty} A_y(y)G(x,t-s;y)\,dy - \int_0^{+\infty} G_s(x,t-s;y)\,dy. \tag{7.13}$$

Substituting in the left-hand side of (7.12), and integrating by parts in s, we obtain

$$\begin{aligned}\int_{t-1}^t G_yBh(s)\,ds = {}& \int_0^1\left(\int_0^{+\infty} A_y(y)G(x,\tau;y)\,dy\right)h(t-\tau)\,d\tau\\ &-\int_0^1\left(\int_0^{+\infty} G(x,\tau;y)\,dy\right)h'(t-\tau)\,d\tau\\ &+\left(\int_0^{+\infty} G(x,1;y)\,dy\right)h(t-1),\end{aligned} \tag{7.14}$$

which, by $\int |G|dy\le C$, has norm bounded by $\max_{0\le\tau\le1}(|h|+|h'|)(t-\tau)$.

Combining this with the more straightforward estimate

$$\begin{aligned}\left|\int_{t-1}^{t} G_y(x,t;0)Bh(s)\,ds\right| &\le \int_0^1 |G_y(x,\tau;0)|Bh(s)\,ds\\ &\le C\max_{0\le\tau\le1}|h(t-\tau)|\int_0^1 \tau^{-1}e^{-|x|^2/C\tau}\,d\tau\\ &= C|x|^{-2}\max_{0\le\tau\le1}|h(t-\tau)|\\ &\quad\times\int_0^1 \left(|x|^2/\tau\right)e^{-|x|^2/C\tau}\,d\tau\\ &\le C\max_{0\le\tau\le1}|h(t-\tau)||x|^{-2},\end{aligned}\tag{7.15}$$

we find that the contribution from $\int_{t-1}^{t}$ has norm bounded by

$$\max_{0\le\tau\le1}(|h|+|h'|)(t-\tau)(1+|x|)^{-2}.$$

Combining this estimate with the one for $\int_0^{t-1}$, we obtain (7.12). □

With these preparations, the proof of stability is straightforward.

Proof of Theorem 1.6. Define

$$\zeta(t) := \sup_{y,0\le s\le t}|u|(\theta+\psi_1+\psi_2)^{-1}(y,t).\tag{7.16}$$

We will establish:

Claim. For all $t\ge 0$ for which a solution exists with ζ uniformly bounded by some fixed, sufficiently small constant, there holds

$$\zeta(t)\le C_2\left(E_0+\zeta(t)^2\right).\tag{7.17}$$

From this result, provided $E_0 < 1/4C_2^2$, we have that $\zeta(t)\le 2C_2E_0$ implies $\zeta(t) < 2C_2E_0$, and so we may conclude by continuous induction that

$$\zeta(t) < 2C_2E_0\tag{7.18}$$

for all $t\ge 0$. (By Lemma 7.1 and standard short-time estimates, $u\in C^0(x)$ exists and ζ remains continuous so long as ζ remains bounded by some uniform constant; hence (7.18) is an open condition.) From (7.18) and the definition of ζ in (7.16) we then obtain the bounds of (1.16). Thus, it remains only to establish the claim above.

Proof of Claim. We must show that $u(\theta+\psi_1+\psi_2)^{-1}$ is bounded by $C(E_0+\zeta(t)^2)$, for some $C>0$, all $0\le s\le t$, so long as ζ remains sufficiently small. By (7.16), we have for all $t\ge 0$ and some $C>0$ that

$$|u(x,t)|\le\zeta(t)(\theta+\psi_1+\psi_2)(x,t),\tag{7.19}$$

and, therefore,

$$|Q(u)(y,s)|\le C\zeta(t)^2\Psi(y,s)\tag{7.20}$$

with Ψ as defined in (7.11), for $0\le s\le t$. Combining (7.20) with representation (7.4)

and applying Lemmas 7.2–7.4, we obtain

$$
\begin{aligned}
|u(x,t)| &\le \int_0^\infty \left|\tilde G(x,t;y)\right| |g(y)|\,dy + \left|\int_0^t G_y(x,t-s;0)Bh(s)\,ds\right| \\
&\quad + \int_0^t \int_0^\infty \left|\tilde G_y(x,t-s;y)\right| |(Q(u))(y,s)|dy\,ds \\
&\le E_0 \int_0^\infty \left|\tilde G(x,t;y)\right| (1+|y|)^{-3/2}\,dy \\
&\quad + \left|\int_0^t G_y(x,t-s;0)Bh(s)\,ds\right| \\
&\quad + C\zeta(t)^2 \int_0^t \int_0^\infty \left|\tilde G_y(x,t-s;y)\right| \Psi(y,s)dy\,ds \\
&\le C\left(E_0+\zeta(t)^2\right)(\theta+\psi_1+\psi_2)(x,t).
\end{aligned}
\tag{7.21}
$$

Dividing by $(\theta+\psi_1+\psi_2)(x,t)$, we obtain (7.17) as claimed. This completes the proof of the claim, and the theorem. ∎

Appendix A. Smoothness of solutions.

In this appendix, we briefly sketch the proof that weak solutions defined by (7.4) are necessarily smooth, classical solutions as well, by indicating how to get the necessary derivative bounds.

Time-derivative. Rewriting the second boundary term on the right-hand side of (7.4) using its convolution structure, as

$$
\int_0^t G_y(x,\tau;0)Bh(t-\tau)\,d\tau,
$$

and differentiating in t, we obtain

$$
G_y(x,t;0)Bh(0) + \int_0^t G_y(x,\tau;0)Bh'(t-\tau)\,d\tau,
$$

for which the first term is bounded and smooth for $x,t>0$, and the second by the same estimate as in (7.15) is bounded by

$$
C|x|^{-2}\int_0^t |h'(t-\tau)|\,d\tau \le C|x|^{-2}\log(1+t).
$$

Differentiating the first and third terms, with respect to t and integrating the third term by parts in y, yields

$$
\begin{aligned}
&\int_0^\infty G(x,t;y)g(y)\,dy - \int_{t/2}^t \int_0^\infty G_y(x,t-s;y)Q(u)_s(y,s)\,dy\,ds \\
&\qquad - \int_0^{t/2}\int_0^\infty G_{yt}(x,t-s;y)Q(u)(y,s)\,dy\,ds,
\end{aligned}
\tag{A.1}
$$

from which, in combination with the boundary estimate already performed, we may readily obtain a short-time bound $|u_t| \le C|x|^{-2}t^{-1}$ by Picard iteration.

Spatial-derivatives. Likewise, differentiating (7.14) with respect to x, we may bound the x-derivative of the boundary term $\int_0^t G_y B\,ds$ by

$$
C\int_0^t \tau^{-1/2}(|h'|+|h|)(t-\tau)|\,d\tau \le C\log(1+t).
$$

Differentiating the first and third terms of the right-hand side of (7.4), with respect to x and integrating the third term by parts in y, yields

$$\int_0^\infty G(x,t;y)g(y)\,dy - \int_{t/2}^t \int_0^\infty G_x(x,t-s;y)Q(u)_y(y,s)\,dy\,ds - \int_0^{t/2}\int_0^\infty G_{yx}(x,t-s;y)Q(u)(y,s)\,dy\,ds, \tag{A.2}$$

from which, in combination with the boundary estimate already performed, we obtain a short-time bound $|u_x| \le Ct^{-1/2}$ by Picard iteration. From the bounds on $|u_t|$ and $|u_x|$, finally, we obtain bounds on $|u_{xx}|$ by the equation satisfied by u.

REFERENCES

[1] J. Alexander, R. Gardner, and C.K.R.T. Jones, *A topological invariant arising in the analysis of traveling waves*, J. Reine Angew. Math., 410 (1990), pp. 167–212.

[2] B. Barker, J. Humpherys, K. Rudd, and K. Zumbrun, *Stability of viscous shocks in isentropic gas dynamics,* Comm. Math. Phys., 281 (2008), pp. 231–249.

[3] L.Q. Brin, *Numerical testing of the stability of viscous shock waves*, Doctoral thesis, Indiana University (1998).

[4] L.Q. Brin, *Numerical testing of the stability of viscous shock waves,* Math. Comput., 70 (2001), pp. 1071–1088.

[5] L. Brin and K. Zumbrun, *Analytically varying eigenvectors and the stability of viscous shock waves*, Seventh Workshop on Partial Differential Equations, Part I (Rio de Janeiro, 2001). Mat. Contemp., 22 (2002), pp. 19–32.

[6] T. Bridges, G. Derks, and G. Gottwald, *Stability and instability of solitary waves of the fifth-order KdV equation: A numerical framework,* Phys. D, 172 (2002), pp. 190–216.

[7] N. Costanzino, J. Humpherys, T. Nguyen, and K. Zumbrun, *Spectral stability of noncharacteristic boundary layers of isentropic Navier–Stokes equations,* Arch. Ration. Mech. Anal., to appear.

[8] R. Gardner and C.K.R.T. Jones, *A stability index for steady state solutions of boundary value problems for parabolic systems*, J. Differential Equations, 91 (1991), pp. 181–203.

[9] R. Gardner and C.K.R.T. Jones, *Traveling waves of a perturbed diffusion equation arising in a phase field model*, Indiana Univ. Math. J., 38 (1989), pp. 1197–1222.

[10] R. Gardner and K. Zumbrun, *The Gap Lemma and geometric criteria for instability of viscous shock profiles*, Commun. Pure Appl. Math., 51 (1998), pp. 797–855.

[11] E. Grenier and O. Gues, *Boundary layers for viscous perturbations of noncharacteristic quasilinear hyperbolic problems*, J. Differential Equations, 143 (1998), pp. 110–146.

[12] E. Grenier and F. Rousset, *Stability of one-dimensional boundary layers by using Green's function*, Commun. Pure Appl. Math., 54 (2001), pp. 1343–1385.

[13] O. Guès, G. Métivier, M. Williams, and K. Zumbrun, *Multidimensional stability of small-amplitude noncharacteristic boundary layers,* preprint (2007).

[14] O. Guès, G. Métivier, M. Williams, and K. Zumbrun, *Viscous boundary value problems for symmetric systems with various multiplicities,* J. Differential Equations, 244 (2008), pp. 309–387.

[15] D. Henry, *Geometric theory of semilinear parabolic equations,* Lecture Notes in Mathematics, Springer-Verlag, Berlin, 1981.

[16] P. Howard and K. Zumbrun, *Stability of undercompressive viscous shock waves*, in press, J. Differential Equations, 225 (2006), pp. 308–360.

[17] J. Humpherys, O. Lafitte, and K. Zumbrun, *Stability of viscous shock profiles in the high Mach number limit,* preprint (2007).

[18] P. Howard and M. Raoofi, *Pointwise asymptotic behavior of perturbed viscous shock profiles*, Adv. Differential Equations, 11 (2006), pp. 1031–1080.

[19] P. Howard, M. Raoofi, and K. Zumbrun, *Sharp pointwise bounds for perturbed shock waves*, J. Hyperbolic Differential Equations, 3 (2006), pp. 297–374.

[20] J. Humpherys and K. Zumbrun, *Spectral stability of small amplitude shock profiles for dissipative symmetric hyperbolic–parabolic systems,* Z. Angew, ed., Math. Phys., 53 (2002), pp. 20–34.

[21] T. Kapitula and B. Sandstede, *Stability of bright solitary-wave solutions to perturbed nonlinear Schrödinger equations,* Phys. D, 124 (1998), pp. 58–103.
[22] T. Kato, *Perturbation Theory for Linear Operators*, Springer–Verlag, Berlin, Heidelberg (1985).
[23] S. Kawashima, S. Nishibata, and P. Zhu, *Asymptotic stability of the stationary solution to the compressible Navier-Stokes equations in the half space,* Commun. Math. Phys., 240 (2003), pp. 483–500.
[24] T.P. Liu and K. Nishihara, *Asymptotic behavior for scalar viscous conservation laws with boundary effect*, J. Differential Equations, 133 (1997), pp. 296–320.
[25] T.P. Liu and S.-H. Yu, *Propagation of stationary viscous Burgers shock under the effect of boundary,* Arch. Ration. Mech. Anal., 139 (1997), pp. 57–82.
[26] A. Matsumura and K. Nishihara, *Large-time behaviors of solutions to an inflow problem in the half space for a one-dimensional system of compressible viscous gas,* Commun. Math. Phys., 222 (2001), pp. 449–474.
[27] C. Mascia and K. Zumbrun, *Pointwise Green's function bounds for shock profiles with degenerate viscosity,* Arch. Ration. Mech. Anal., 169 (2003), pp. 177–263.
[28] C. Mascia and K. Zumbrun, *Stability of large-amplitude shock profiles of hyperbolic–parabolic systems*, Arch. Ration. Mech. Anal., 172 (2004), pp. 93–131.
[29] G. Métivier and K. Zumbrun, *Large viscous boundary layers for noncharacteristic nonlinear hyperbolic problems,* Mem. Amer. Math. Soc., 175(826):vi+107 (2005).
[30] F. Rousset, *Stability of small amplitude boundary layers for mixed hyperbolic-parabolic systems*, Trans. Amer. Math. Soc., 355 (2003), pp. 2291–3008.
[31] F. Rousset, *Inviscid boundary conditions and stability of viscous boundary layers*, Asymptot. Anal., 26 (2001), pp. 285–306.
[32] D. Serre, *Sur la stabilité des couches limites de viscosité,* Annales de l'institut Fourier, 51 (2001), pp. 109–130.
[33] D. Serre and K. Zumbrun, *Boundary layer stability in real vanishing-viscosity limit*, Commun. Math. Phys., 221 (2001), pp. 267–292.
[34] H. Schlichting, K. Gersten, E. Krause, and H. Oertel, Jr., *Boundary-Layer Theory*, Springer; 8th ed. 2000. Corr. 2nd printing edition (March 22, 2004).
[35] S. Yarahmadian, *Pointwise Green function bounds and long-time stability of large-amplitude noncharacteristic boundary layers,* Doctoral thesis, Indiana University (2008).
[36] K. Zumbrun, *Stability of large-amplitude shock waves of compressible Navier-Stokes equations,* In Handbook of mathematical fluid dynamics. Vol. III, pp. 311–533, North-Holland, Amsterdam, 2004. With an appendix by Helge Kristian Jenssen and Gregory Lyng.
[37] K. Zumbrun and P. Howard, *Pointwise semigroup methods and stability of viscous shock waves,* Indiana Math. J., 47 (1998), pp. 741–871.

SIAM J. MATH. ANAL.
Vol. 40, No. 6, pp. 2351–2391

A HIGHER ORDER MODEL FOR IMAGE RESTORATION: THE ONE-DIMENSIONAL CASE*

G. DAL MASO†, I. FONSECA‡, G. LEONI‡, AND M. MORINI†

Abstract. The higher order total variation-based model for image restoration proposed by Chan, Marquina, and Mulet in [*SIAM J. Sci. Comput.*, 22 (2000), pp. 503–516] is analyzed in one dimension. A suitable functional framework in which the minimization problem is well posed is being proposed, and it is proved analytically that the higher order regularizing term prevents the occurrence of the *staircase effect*. The generalized version of the model considered here includes, as particular cases, some curvature dependent functionals.

Key words. image segmentation, total variation models, staircase effect, higher order regularization, relaxation, curvature dependent functionals

AMS subject classifications. 49J45, 26A45, 65K10, 68U10

DOI. 10.1137/070697823

1. Introduction.

Deblurring and denoising of images are fundamental problems in image processing and gave rise in the past few years to a vast variety of techniques and methods touching different fields of mathematics. Among them, variational methods based on the minimization of some energy functional have been successfully employed to treat a fairly general class of image restoration problems. Typically, such functionals present a *fidelity term*, which penalizes the distance between the reconstructed image u and the noisy image g with respect to a suitable metric, and a regularizing term, which makes high frequency noise energetically unfavorable.

When the fidelity term is given by the squared L^2 distance multiplied by a parameter $\lambda > 0$ and the regularizing term is represented by the total variation, we are led to the following minimization problem:

$$\min\left\{|Du|(\Omega) + \lambda \int_\Omega |u-g|^2\,dx : u \in BV(\Omega)\right\}, \tag{1.1}$$

which was proposed by Rudin, Osher, and Fatemi in [13]. Here Ω is an open bounded domain in one or two dimensions, $BV(\Omega)$ denotes the space of functions of bounded variations in Ω, and $|Du|(\Omega)$ stands for the total variation of u in Ω. The main feature of the total variation-based image restoration is perhaps represented by the tendency to yield (almost) piecewise constant solutions or, in other words, "blocky" images. Typically, one observes that *ramps* (i.e., affine regions) in the original image give rise to staircase-like structures in the reconstructed image, a phenomenon which is often referred to as the *staircase effect*. This means that the original edges are well preserved by this method but also that many artificial discontinuities can be generated by the presence of noise, while the finer details of the objects contained in the image may not be properly recovered.

Several variants of (1.1) have been subsequently proposed in order to fix these drawbacks. In this paper we follow the approach of Chan, Marquina, and Mulet [6]:

*Received by the editors July 9, 2007; accepted for publication (in revised form) October 10, 2008; published electronically February 25, 2009.
http://www.siam.org/journals/sima/40-6/69782.html
†SISSA, Via Beirut 2, 34014 Trieste, Italy (dalmaso@sissa.it, morini@sissa.it).
‡Carnegie Mellon University, Pittsburgh, PA 15213-3890 (fonseca@andrew.cmu.edu, giovanni@andrew.cmu.edu).

Since the total variation does not distinguish between jumps and smooth transitions, their idea is to consider an additional penalization of the discontinuities by taking second derivatives into account. More precisely, they propose a regularizing term of the form

$$\int_\Omega |\nabla u|\,dx + \int_\Omega \psi(|\nabla u|)h(\Delta u)\,dx, \tag{1.2}$$

where ψ is a function that must satisfy suitable conditions at infinity in order to allow jumps.

In this paper we consider the following one-dimensional (1D) version of (1.2):

$$\mathcal{F}_p(u) := \int_a^b |u'|\,dx + \int_a^b \psi(|u'|)|u''|^p\,dx, \tag{1.3}$$

where $a < b$ are real numbers, $p \in [1, +\infty)$, and $\psi : \mathbb{R} \to]0, +\infty[$ is a bounded Borel function satisfying suitable integrability assumptions (see (2.1) and (3.1)). We remark that by taking

$$\psi(t) := \frac{1}{(1+t^2)^{\frac{1}{2}(3p-1)}}$$

the functional (1.3) takes the form

$$\int_a^b |u'|\,dx + \int_{\text{Graph } u} |k|^p\,d\mathcal{H}^1,$$

where k denotes the curvature of the graph of u and $\mathcal{H}^1$ stands for the 1D Hausdorff measure. This seems to suggest a broader applicability of our results: Functionals of this kind are often encountered in many computer vision and graphics applications, such as, among the others, corner preserving geometry denoising and segmentation with depth (see, e.g., [3], [15], and the references contained therein).

Our main analytical objective is twofold:

(i) to set up a proper functional framework where the minimization problem corresponding to

$$\mathcal{F}_p(u) + \lambda \int_a^b |u-g|^2\,dx$$

is well posed;

(ii) to give an analytical proof of the fact that the higher order regularizing term eliminates the staircase effect.

We carry out the first part of this program in sections 2 and 3 by using the theory of relaxation (see [7] for a general introduction): We regard $\mathcal{F}_p$ as defined for all functions in the Sobolev space $W^{2,p}(]a,b[)$, we extend it to $L^1(]a,b[)$ by setting $\mathcal{F}_p(u) := +\infty$ if $u \in L^1(]a,b[) \setminus W^{2,p}(]a,b[)$, and then we identify its lower semicontinuous envelope with respect to the strong L^1 convergence. The two cases $p = 1$ and $p > 1$ require a different treatment, and, in fact, the analysis turns out to be considerably more delicate in the case $p = 1$. Moreover, the domains of the relaxed functionals are quite peculiar (see Definitions 2.1 and 3.1) and display properties that are qualitatively different in the two cases. In particular, it turns out that piecewise constant functions corresponding to images with genuine edges are approximable by sequences with bounded energy

only for $p = 1$. The extension of these relaxation results to higher dimensions will be the subject of a subsequent paper.

The second part of the program is carried out in section 4. We start by exhibiting an analytical example of staircasing for the Rudin–Osher–Fatemi model (Theorem 4.3). More precisely, we show that if g has the form $g = g_1 + h$, with g_1 an affine function (which represents the original "clean" signal) and h a highly oscillating noise, then, for certain choices of h, the reconstructed signal u may display a stairlike or piecewise constant structure, even if h is uniformly small. In particular, we note that

(a) the discontinuity set of u is much larger than the discontinuity set of the original signal g_1,

(b) the derivatives of u and u_1 are very far apart,

where u_1 denotes the solution to the Rudin–Osher–Fatemi minimization problem corresponding to the datum g_1. Although we do not attempt to give a formal definition of staircasing, we consider the presence/absence of (a) and (b) as a way to detect the presence/absence of the staircase effect. In Theorems 4.5 and 4.8 we prove that the new model eliminates the effect by showing that (a) and (b) do not occur. More precisely, we show that whenever the datum g is of the form $g = g_1 + h$, with g_1 a regular signal and h a highly oscillating noise close to 0 in the L^∞-weak* topology, the reconstructed image u is regular as well (in particular, no new artificial discontinuities are created) and u and u_1 are close in some strong norm ($W^{1,q}$-norm for every $q > 1$ or C^1-norm, depending on whether $p = 1$ or $p > 1$), where u_1 is the solution corresponding to the clean signal g_1.

For completeness we conclude by mentioning that other approaches have been considered to avoid staircasing: The works by Geman and Reynolds [9] and Chambolle and Lions [5] contain a different use of higher order derivatives as regularizing terms; in [2], Blomgren, Chan, and Mulet propose a BV-H^1 interpolation approach, while Kindermann, Osher, and Jones avoid in [11] the use of second derivatives by considering a sort of nonlocal total variation.

2. The case $p = 1$. We start by studying the compactness properties and the relaxation of (1.3) in the case $p = 1$. Throughout this section $\psi\colon \mathbb{R} \to]0, +\infty[$ will be a bounded Borel function such that

$$M := \int_{-\infty}^{+\infty} \psi(t)\, dt < +\infty \tag{2.1}$$

and

$$\inf_{t\in K} \psi(t) > 0 \quad \text{for every compact set } K \subset \mathbb{R}. \tag{2.2}$$

Let $\Psi_1\colon \overline{\mathbb{R}} \to [0, M]$ be the increasing function defined by

$$\Psi_1(t) := \int_{-\infty}^{t} \psi(s)\, ds,$$

and let $\Psi_1^{-1}\colon [0, M] \to \overline{\mathbb{R}}$ be its inverse function.

Given a bounded open interval $]a, b[$ in $\mathbb{R}$, we let $\mathcal{F}_1\colon L^1(]a, b[) \to [0, +\infty]$ be the functional defined by

$$\mathcal{F}_1(u) := \begin{cases} \displaystyle\int_a^b |u'|\, dx + \int_a^b \psi(u')|u''|\, dx & \text{if } u \in W^{2,1}(]a, b[), \\ +\infty & \text{otherwise.} \end{cases} \tag{2.3}$$

The first step in the study of (2.3) will consist in identifying the subspace of L^1 functions which can be approximated by energy bounded sequences. In order to do so, we need to introduce some notation and recall some basic facts about BV functions of one variable. This will be the content of the next subsection.

2.1. *BV* functions of one variable. We recall that a function $u \in L^1(]a,b[)$ belongs to $BV(]a,b[)$ if and only if

$$\sup\left\{\int_a^b u\varphi'\,dx : \varphi \in C_c^1(]a,b[),\ |\varphi| \le 1\right\} < +\infty. \tag{2.4}$$

Note that this implies that the distributional derivative u' of u is a bounded Radon measure in $]a,b[$. We will often consider the Lebesgue decomposition

$$u' = (u')^a\mathcal{L}^1 + (u')^s,$$

where $(u')^a$ is the density of the absolutely continuous part of u' with respect to the Lebesgue measure $\mathcal{L}^1$ on $]a,b[$ while $(u')^s$ is its singular part. We will denote the total variation measure of u' by $|u'|$. In particular, $|u'|(]a,b[)$ equals the value of the supremum in (2.4). For every function $u \in BV(]a,b[)$ the following left and right approximate limits

$$u_-(y) := \lim_{\varepsilon\to 0+}\frac{1}{\varepsilon}\int_{y-\varepsilon}^{y} u(x)\,dx, \qquad u_+(y) := \lim_{\varepsilon\to 0+}\frac{1}{\varepsilon}\int_{y}^{y+\varepsilon} u(x)\,dx,$$

respectively, are well defined at every point $y \in]a,b[$. In fact, $u_-(y)$ is well defined also at $y = b$ while $u_+(y)$ exists also at $y = a$. The functions u_- and u_+ coincide $\mathcal{L}^1$-a.e. with u and are left and right continuous, respectively. Moreover, it turns out that the set $S_u := \{y \in]a,b[: u_-(y) \ne u_+(y)\}$ is at most countable. The set S_u is often referred to as the set of essential discontinuities or *jump points* of u.

It is well known that, in turn, the singular part $(u')^s$ splits into the sum of an atomic measure concentrated on S_u and a singular diffuse measure $(u')^c$, called the *Cantor part* of u':

$$(u')^s = [u]\mathcal{H}^0 \llcorner S_u + (u')^c,$$

where we set $[u] := u_+ - u_-$ and $\mathcal{H}^0$ stands for the counting measure. Finally, we recall that every $u \in BV(]a,b[)$ is differentiable at $\mathcal{L}^1$-a.e. y in $]a,b[$ with derivative given by $(u')^a(y)$. In this case, we will often write, with a slight abuse of notation, $u'(y)$ instead of $(u')^a(y)$.

We say that a sequence $\{u_k\}$ of functions in $BV(]a,b[)$ *weakly star converges in* $BV(]a,b[)$ to a function $u \in BV(]a,b[)$ if $u_n \to u$ in $L^1(]a,b[)$ and $u_k' \rightharpoonup u'$ weakly* in $M_b(]a,b[)$, where $M_b(]a,b[)$ is the space of bounded Radon measures.

We will also need sometimes the notion of total variation for a function defined everywhere. We recall that $u\colon]a,b[\to \mathbb{R}$ has bounded *pointwise total variation* over the interval $]c,d[\subset]a,b[$ if

$$\mathrm{Var}\,(u;]c,d[) := \sup\sum_{i=1}^{k}|u(y_i) - u(y_{i-1})| < +\infty,$$

where the supremum is taken over all finite families $y_0, y_1, \dots, y_k$ such that $c < y_0 < y_1 < \cdots < y_k < d$, $k \in \mathbb{N}$. It is easy to see that if u has bounded pointwise total

variation in $]a,b[$, then it admits left and right limits at every point, it belongs to $BV(]a,b[)$, and $|u'|(]c,d[) \le \mathrm{Var}\,(u;]c,d[)$ for every interval $]c,d[\subset]a,b[$. Conversely, if $u \in BV(]a,b[)$, the *precise representatives* u_- and u_+ have bounded pointwise total variation and satisfy

$$|u'|(]c,d[) = \mathrm{Var}\,(u_-;]c,d[) = \mathrm{Var}\,(u_+;]c,d[)$$

for every interval $]c,d[\subset]a,b[$.

Finally, we recall the Helly theorem: For every bounded sequence of functions $u_k :]a,b[\to \mathbb{R}$ such that $\sup_k \mathrm{Var}\,(u_k;]a,b[) < +\infty$, there exist u, with pointwise total variation in $]a,b[$, and a subsequence (not relabeled) such that $u_k \to u$ pointwise.

We refer to [14] and [10] for an exhaustive exposition of the properties of BV functions of one variable.

2.2. Compactness. To define the subspace of L^1 functions that can be approximated by energy bounded sequences, for every function $u \in BV\,(]a,b[)$ we consider the sets

$$Z^+[(u')^a] := \left\{ x \in]a,b[: \lim_{\varepsilon\to 0+} \frac{1}{2\varepsilon} \int_{x-\varepsilon}^{x+\varepsilon} (u')^a \, dx = +\infty \right\}, \tag{2.5}$$

$$Z^-[(u')^a] := \left\{ x \in]a,b[: \lim_{\varepsilon\to 0+} \frac{1}{2\varepsilon} \int_{x-\varepsilon}^{x+\varepsilon} (u')^a \, dx = -\infty \right\}. \tag{2.6}$$

It is also convenient to define

$$Z[(u')^a] := Z^+[(u')^a] \cup Z^-[(u')^a].$$

Definition 2.1. *Let $X^1_\psi(]a,b[)$ be the set of all functions $u \in BV(]a,b[)$ such that $v := \Psi_1 \circ (u')^a$ belongs to $BV(]a,b[)$ and the positive part $((u')^c)^+$ and the negative part $((u')^c)^-$ of the measure $(u')^c$ are concentrated on $Z^+[(u')^a]$ and $Z^-[(u')^a]$, respectively.*

Remark 2.2. Note that if $u \in X^1_\psi(]a,b[)$, then the limits

$$(u')^a_-(y) := \lim_{\varepsilon\to 0+} \frac{1}{\varepsilon} \int_{y-\varepsilon}^{y} (u')^a \, dx, \qquad (u')^a_+(y) := \lim_{\varepsilon\to 0+} \frac{1}{\varepsilon} \int_{y}^{y+\varepsilon} (u')^a \, dx \tag{2.7}$$

exist in $\overline{\mathbb{R}}$ for every y. More precisely, $(u')^a_-$ exists also at $y = b$, while $(u')^a_+$ is well defined also at $y = a$. Indeed, since $v = \Psi_1 \circ (u')^a$ is a BV function, it admits a precise representative $\tilde v$ such that the right and left limits exist at every point, and the same property holds for $\Psi_1^{-1}(\tilde v)$. As $\Psi_1^{-1}(\tilde v) = (u')^a$ $\mathcal{L}^1$-a.e. in $]a,b[$, the limits considered in (2.7) are everywhere well defined. Moreover, the set $S_{(u')^a} := S_v$ is at most countable and

$$(u')^a_- = (u')^a_+ \quad \text{on }]a,b[\setminus S_{(u')^a}. \tag{2.8}$$

We also remark that $(u')^a_-$ and $(u')^a_+$ are left and right continuous, respectively, which, in turn, implies that the functions defined by

$$(u')^a_\vee(x) := \max\left\{(u')^a_+(x), (u')^a_-(x)\right\}, \ (u')^a_\wedge(x) := \min\left\{(u')^a_+(x), (u')^a_-(x)\right\}$$

if $x \in]a,b[$ and by

$$(u')^a_\vee(a) = (u')^a_\wedge(a) := (u')^a_+(a), \qquad (u')^a_\vee(b) = (u')^a_\wedge(b) := (u')^a_-(b)$$

are upper and lower semicontinuous in $[a,b]$, respectively. By (2.8) we have

$$Z^+[(u')^a] \setminus S_{(u')^a} = \{x \in]a,b[: (u')^a_\wedge(x) = +\infty\} \setminus S_{(u')^a},$$
$$Z^-[(u')^a] \setminus S_{(u')^a} = \{x \in]a,b[: (u')^a_\vee(x) = -\infty\} \setminus S_{(u')^a}.$$

Therefore, $((u')^c)^+$ is concentrated on the set $\{x \in]a,b[: (u')^a_\wedge(x) = +\infty\}$ and $((u')^c)^-$ is concentrated on the set $\{x \in]a,b[: (u')^a_\vee(x) = -\infty\}$.

Before we proceed we show that the space $X^1_\psi(]a,b[)$ contains functions with nontrivial Cantor part when ψ satisfies suitable decay estimates at infinity.

PROPOSITION 2.3. *Assume that* $\psi\colon \mathbb{R} \to]0,+\infty[$ *is a bounded Borel function satisfying* (2.1), (2.2), *and*

$$\psi(t) \le \frac{c}{t^\alpha} \tag{2.9}$$

for all $t \ge 1$ *and for some* $c > 0$, $\alpha > 1$. *Then there exists* $u \in X^1_\psi(]a,b[)$ *with* $(u')^c \neq 0$.

Proof. For simplicity we take $]a,b[=]0,1[$.

Step 1. We start by recalling the definition of the generalized Cantor set $\mathbb{D}_\delta$, where $\delta \in]0,\frac{1}{2}[$ (see, for instance, [8, Chapter 1, section 2.4]). The construction is entirely similar to the one of the (ternary) Cantor set with the only difference that the middle intervals removed at each step have length $1-2\delta$ times the length of the intervals remaining from the previous step. To be more precise, remove from $[0,1]$ the interval $I_{11} := (\delta, 1-\delta)$. At the second step remove from each of the remaining closed intervals $[0,\delta]$ and $[1-\delta,1]$ the middle intervals, denoted by I_{12} and I_{22}, of length $\delta(1-2\delta)$. Continuing in this fashion at each step n we remove 2^{n-1} middle intervals $I_{1n},\dots,I_{2^{n-1}n}$, each of length $\delta^{n-1}(1-2\delta)$. The *generalized Cantor set* $\mathbb{D}_\delta$ is defined as

$$\mathbb{D}_\delta := [0,1] \setminus \bigcup_{n=1}^{\infty} \bigcup_{k=1}^{2^{n-1}} I_{kn}.$$

The set $\mathbb{D}_\delta$ is closed (since its complement is given by a family of open intervals), and

$$\mathcal{L}^1(\mathbb{D}_\delta) = 1 - \sum_{n=1}^{\infty} \sum_{k=1}^{2^{n-1}} \mathcal{L}^1(I_{kn}) = 1 - \sum_{n=1}^{\infty} \sum_{k=1}^{2^{n-1}} \delta^{n-1}(1-2\delta)$$
$$= 1 - (1-2\delta) \sum_{n=1}^{\infty} (2\delta)^{n-1} = 0.$$

Next, we recall the definition of the corresponding Cantor function f_δ. Set

$$g_n := \frac{1}{(2\delta)^n} \left(1 - \sum_{j=1}^{n} \sum_{k=1}^{2^{j-1}} \chi_{I_{kj}} \right),$$

and define $f_n(x) := \int_0^x g_n(t)\,dt$. It can be shown that $\{f_n\}$ converges uniformly to a continuous nondecreasing function f_δ such that $f_\delta(0) = 0$, $f_\delta(1) = 1$, and $f'_\delta = (f'_\delta)^c$ is supported on $\mathbb{D}_\delta$.

Step 2. We claim that it is enough to find a constant $\delta \in]0,\frac{1}{2}[$ for which it is possible to construct a continuous integrable function $w_\delta :]0,1[\to [0,+\infty]$ such

that $\Psi_1 \circ w_\delta \in BV(]0,1[)$ and $w_\delta(x) = +\infty$ if and only if $x \in \mathbb{D}_\delta$. Indeed, setting $u_\delta(x) := \int_0^x w_\delta(t)\,dt + f_\delta(x)$, we have that $u_\delta \in BV(]0,1[)$, u_δ is continuous, and $(u'_\delta)^a = w_\delta$ so that $Z^+[(u'_\delta)^a] = Z[(u'_\delta)^a] = \mathbb{D}_\delta$ and $\Psi_1 \circ (u'_\delta)^a \in BV(]0,1[)$. Moreover, $(u'_\delta)^c = (f'_\delta)^c$ is supported on $\mathbb{D}_\delta = Z^+[(u'_\delta)^a]$. Hence, u_δ belongs to $X^1_\psi(]a,b[)$.

Step 3. It remains to construct w_δ for a suitable $\delta \in]0, \frac{1}{2}[$. Consider a convex function $\phi :]0,1[\to [0,+\infty)$ such that

$$\lim_{x\to 0^+} \phi(x) = \lim_{x\to 1^-} \phi(x) = +\infty, \quad \phi\left(\frac{1}{2}\right) = 0, \tag{2.10}$$

and

$$\int_0^1 \phi(x)\,dx = 1. \tag{2.11}$$

Choose $s > 0$ so large that

$$\alpha > \frac{s+1}{s}. \tag{2.12}$$

For $x \in I_{kn}$ (see Step 1) define

$$\phi_{kn}(x) := 2^{sn} + \phi\left(\frac{x - a_{kn}}{\delta^{n-1}(1-2\delta)} + \frac{1}{2}\right), \tag{2.13}$$

where a_{kn} is the midpoint of the interval I_{kn}. Finally, set

$$w_\delta := \sum_{n=1}^{\infty} \sum_{k=1}^{2^{n-1}} \phi_{kn}\chi_{I_{kn}} + I_{\mathbb{D}_\delta},$$

where $I_{\mathbb{D}_\delta}$ is the indicator function of the set $\mathbb{D}_\delta$, that is,

$$I_{\mathbb{D}_\delta}(x) := \begin{cases} +\infty & \text{if } x \in \mathbb{D}_\delta, \\ 0 & \text{otherwise.} \end{cases}$$

Using the fact that

$$\int_{I_{kn}} \phi_{kn}\,dx = (2^{sn}+1)\,\delta^{n-1}(1-2\delta),$$

which follows from (2.11) and a change of variables, we have

$$\int_0^1 w_\delta\,dx = \sum_{n=1}^{\infty} \sum_{k=1}^{2^{n-1}} (2^{sn}+1)\,\delta^{n-1}(1-2\delta) < \infty$$

for $\delta < \frac{1}{2^{s+1}}$. To estimate the total variation of $v := \Psi_1 \circ w_\delta$, we consider the approximating sequence

$$v_m(x) := \begin{cases} \Psi_1 \circ \phi_{kn}(x) & \text{if } x \in I_{kn},\ 1 \le k \le 2^{n-1},\ 1 \le n \le m, \\ M & \text{otherwise.} \end{cases}$$

By (2.9), (2.10), (2.13), and the convexity of ϕ, it can be seen that

$$\operatorname{Var}(v_m; I_{kn}) = 2(M - \Psi_1(2^{sn})) = 2\int_{2^{sn}}^{+\infty} \psi(t)\,dt \le \frac{2c}{\alpha - 1}\frac{1}{2^{sn(\alpha-1)}}.$$

It follows that

$$\mathrm{Var}\,(v_m;]0,1[) \le \frac{2c}{\alpha-1}\sum_{n=1}^{m}\sum_{k=1}^{2^{n-1}}\frac{1}{2^{sn(\alpha-1)}} \le \frac{2c}{\alpha-1}\sum_{n=1}^{\infty}\frac{1}{2^{sn(\alpha-1)-n+1}}.$$

The last series is finite, thanks to (2.12). Therefore, the v_m's have equibounded total variations, and, since $v_m \to v$ in $L^1(]0,1[)$, we conclude that $v \in BV(]0,1[)$. □

Energy bounded sequences are compact in $X^1_\psi(]a,b[)$, as made precise by the following theorem.

THEOREM 2.4. *Let $\{u_k\}$ be a sequence of functions bounded in $L^1(]a,b[)$ such that*

$$C := \sup_k \mathcal{F}_1(u_k) < +\infty. \tag{2.14}$$

Then there exist a subsequence (not relabeled) $\{u_k\}$ and a function $u \in X^1_\psi(]a,b[)$ such that

$$u_k \rightharpoonup u \quad \textit{weakly}^* \textit{ in } BV(]a,b[), \tag{2.15}$$

$$\Psi_1 \circ u_k' \rightharpoonup \Psi_1 \circ (u')^a \quad \textit{weakly}^* \textit{ in } BV(]a,b[), \tag{2.16}$$

$$u_k' \to (u')^a \quad \textit{pointwise } \mathcal{L}^1\textit{-a.e. in }]a,b[. \tag{2.17}$$

Proof. By (2.3) and (2.14) we have that each u_k belongs to $W^{2,1}(]a,b[)$ and

$$C_1 := \sup_k \int_a^b [\,|u_k| + |u_k'| + \psi(u_k')|u_k''|\,]\,dx < +\infty. \tag{2.18}$$

Let us define

$$v_k := \Psi_1 \circ u_k'. \tag{2.19}$$

As Ψ_1 is Lipschitz in $\mathbb{R}$, the functions v_k belong to $W^{1,1}(]a,b[)$ and

$$v_k' = \psi(u_k')u_k'' \qquad \mathcal{L}^1\text{-a.e. on }]a,b[. \tag{2.20}$$

It follows from (2.1) and (2.14) that

$$\int_a^b [\,|v_k| + |v_k'|\,]\,dx \le M(b-a) + C. \tag{2.21}$$

By (2.18) and (2.21) and the Helly theorem, passing to a subsequence if necessary, we may assume that

$$u_k \rightharpoonup u \quad \text{weakly}^* \text{ in } BV(]a,b[)$$

and

$$v_k(x) \to v(x) \quad \text{for all } x \in]a,b[\tag{2.22}$$

for some $u \in BV(]a,b[)$ and $v\colon]a,b[\to [0,M]$ with pointwise bounded variation. Note that (2.22) determines the values of v at every $x \in]a,b[$.

Since Ψ_1^{-1} is continuous, we obtain

$$u_k' \to w := \Psi_1^{-1}(v) \quad \text{pointwise in }]a,b[. \tag{2.23}$$

Since by (2.18) we have $\sup_k \|u_k'\|_1 < +\infty$, it follows from Fatou's lemma and (2.23) that w is integrable, and so

$$w \text{ is finite } \mathcal{L}^1\text{-a.e. in }]a,b[. \tag{2.24}$$

Moreover, w has left and right limits in $\overline{\mathbb{R}}$ at each point $x \in]a,b[$, denoted by $w_-(x)$ and $w_+(x)$, respectively, and

$$w(x) = w_-(x) = w_+(x) \quad \text{except for a countable set of points } x. \tag{2.25}$$

We now split the remaining part of the proof into two steps.

Step 1. We prove that

$$w = (u')^a \quad \mathcal{L}^1\text{-a.e. in }]a,b[. \tag{2.26}$$

If not, we have $\mathcal{L}^1(\{w \neq (u')^a\}) > 0$. By (2.24) there exists $t_0 > 0$ such that

$$\mathcal{L}^1(\{w \neq (u')^a\} \cap \{|w| < t_0\}) > 0,$$

and, in particular, we may find an infinite number of disjoint open intervals I such that

$$\mathcal{L}^1(\{w \neq (u')^a\} \cap \{|w| < t_0\} \cap I) > 0. \tag{2.27}$$

By a change of variables we obtain

$$\int_I \psi(u_k')|u_k''|\,dx \geq \int_{m_k}^{M_k} \psi(t)\,dt, \tag{2.28}$$

where

$$m_k := \inf_I u_k' \quad \text{and} \quad M_k := \sup_I u_k'.$$

We claim that at least one of the two sequences $\{m_k\}$ and $\{M_k\}$ is divergent. Indeed, if not, a subsequence of $\{u_k'\}$ would be bounded in $L^\infty(I)$. This implies that $u' \in L^\infty(I)$ and that $u_k' \rightharpoonup u'$ weakly* in $L^\infty(I)$. As $u_k' \to w$ pointwise $\mathcal{L}^1$-a.e. in I, we deduce that $u' = w$ $\mathcal{L}^1$-a.e. in I, which contradicts (2.27). Hence, the claim holds. If

$$\lim_{k\to\infty} M_k = +\infty, \tag{2.29}$$

then by (2.23) and (2.27)

$$\limsup_{k\to\infty} m_k < t_0. \tag{2.30}$$

From (2.28), (2.30), and (2.29) we obtain

$$\liminf_{k\to\infty} \int_I \psi(u_k')|u_k''|\,dx \geq \int_{t_0}^{+\infty} \psi(t)\,dt > 0.$$

Analogously, if $\lim_k m_k = -\infty$, then

$$\liminf_{k\to\infty} \int_I \psi(u_k')|u_k''|\,dx \geq \int_{-\infty}^{-t_0} \psi(t)\,dt > 0.$$

In any case, we can choose an arbitrarily large number m of disjoint intervals I satisfying (2.27). Adding the contributions of each interval we obtain

$$\liminf_{k\to\infty} \int_a^b \psi(u_k')|u_k''|\,dx \geq m \min\left\{\int_{t_0}^{+\infty} \psi(t)\,dx, \int_{-\infty}^{-t_0} \psi(t)\,dt\right\},$$

which contradicts (2.18) for m large enough. This concludes the proof of (2.26).

Step 2. To prove that $u \in X^1_\psi(]a,b[)$, it remains to show that the positive part $((u')^c)^+$ and the negative part $((u')^c)^-$ of the measure $(u')^c$ are concentrated on $Z^+[(u')^a]$ and $Z^-[(u')^a]$, respectively, that is,

$$((u')^c)^\pm (]a,b[\setminus Z^\pm[(u')^a]) = 0. \tag{2.31}$$

To this purpose we introduce the sets

$$E^+[u'] := \left\{x \in]a,b[: \lim_{\varepsilon\to 0+} \frac{(u')^+(]x-\varepsilon,x+\varepsilon[)}{2\varepsilon} = +\infty\right\}, \tag{2.32}$$

$$E^-[u'] := \left\{x \in]a,b[: \lim_{\varepsilon\to 0+} \frac{(u')^-(]x-\varepsilon,x+\varepsilon[)}{2\varepsilon} = +\infty\right\}, \tag{2.33}$$

$$E[u'] := \left\{x \in]a,b[: \lim_{\varepsilon\to 0+} \frac{|u'|(]x-\varepsilon,x+\varepsilon[)}{2\varepsilon} = +\infty\right\}.$$

Since $((u')^s)^+ = ((u')^+)^s$ is concentrated on $E^+[u']$ and $((u')^s)^- = ((u')^-)^s$ is concentrated on $E^-[u']$ (see, e.g., [1, Theorem 2.22]), to prove (2.31) it is enough to show that

$$E^+[u'] \setminus Z^+[(u')^a] \text{ and } E^-[u'] \setminus Z^-[(u')^a] \text{ are at most countable.} \tag{2.34}$$

We show only that $E^+[u'] \setminus Z^+[(u')^a]$ is at most countable, since the other property can be proved in a similar way. Assume, by contradiction, that $E^+[u'] \setminus Z^+[(u')^a]$ is not countable. Since by (2.5) and (2.26)

$$Z^+[(u')^a] \subset \{x \in]a,b[: \max\{w_-(x), w_+(x)\} = +\infty\},$$

by (2.25) there exists $t_0 > 0$ such that

$$(E^+[u'] \setminus Z^+[(u')^a]) \cap \{w < t_0\} \quad \text{is uncountable.}$$

Fix $t_1 > t_0$, and let $x_1, \dots, x_m$ be m distinct points in $(E^+[u'] \setminus Z^+[(u')^a]) \cap \{w < t_0\}$. By (2.32) there exists $\varepsilon > 0$ such that the intervals $I_j :=]x_j - \varepsilon, x_j + \varepsilon[$ are pairwise disjoint and

$$\frac{(u')^+(]x_j-\varepsilon,x_j+\varepsilon[)}{2\varepsilon} > t_1 \quad \text{for } i = 1,\dots,m. \tag{2.35}$$

By a change of variables we obtain

$$\int_{I_j} \psi(u_k')|u_k''|\,dx \geq \int_{m_{kj}}^{M_{kj}} \psi(t)\,dt, \tag{2.36}$$

where

$$m_{kj} := \inf_{I_j} u_k' \quad \text{and} \quad M_{kj} := \sup_{I_j} u_k'.$$

By (2.23) and the fact that $w(x_j) < t_0$, we deduce that

$$\limsup_{k\to\infty} m_{kj} < t_0 \tag{2.37}$$

for $j = 1, \dots, m$. On the other hand, (2.15) and (2.35) yield

$$\liminf_{k\to\infty} \frac{1}{2\varepsilon} \int_{x_j-\varepsilon}^{x_j+\varepsilon} (u_k')^+ \, dx \geq \frac{(u')^+(]x_j-\varepsilon, x_j+\varepsilon[)}{2\varepsilon} > t_1$$

(this can be seen as a particular case of the Reshetnyak lower semicontinuity theorem, with $f = (\cdot)^+$). This implies that $\liminf_{k\to\infty} M_{kj} > t_1$ for $j = 1, \dots, m$. Hence, also by (2.36) and (2.37) we obtain

$$\liminf_{k\to\infty} \sum_{j=1}^{m} \int_{I_j} \psi(u_k')|u_k''| \, dx \geq \sum_{j=1}^{m} \liminf_{k\to\infty} \int_{I_j} \psi(u_k')|u_k''| \, dx \geq m \int_{t_0}^{t_1} \psi(t) \, dt,$$

which contradicts (2.18) for m large enough. This shows (2.34) and concludes the proof of the theorem. ∎

2.3. Relaxation. The following theorem, which is the main result of the section, is devoted to the characterization of the relaxation of $\mathcal{F}_1$ with respect to strong convergence in $L^1(]a,b[)$.

THEOREM 2.5. *Let $\overline{\mathcal{F}}_1 \colon L^1(]a,b[) \to [0,+\infty]$ be defined by*

$$\overline{\mathcal{F}}_1(u) := \inf \left\{ \liminf_{k\to\infty} \mathcal{F}_1(u_k) : u_k \to u \text{ in } L^1(]a,b[) \right\}$$

for every $u \in L^1(]a,b[)$. Then

$$\overline{\mathcal{F}}_1(u) = \begin{cases} |u'|(]a,b[) + |v'|(]a,b[\setminus S_u) + \displaystyle\sum_{x\in S_u} \Phi(\nu_u, (u')^a_-, (u')^a_+) & \text{if } u \in X^1_\psi(]a,b[), \\ +\infty & \text{otherwise,} \end{cases} \tag{2.38}$$

where $v := \Psi_1 \circ (u')^a$, $\nu_u := \operatorname{sign}(u_+ - u_-)$, and

$$\begin{aligned} \Phi(1, t_1, t_2) &:= \int_{t_1}^{+\infty} \psi(t) \, dt + \int_{t_2}^{+\infty} \psi(t) \, dt, \\ \Phi(-1, t_1, t_2) &:= \int_{-\infty}^{t_1} \psi(t) \, dt + \int_{-\infty}^{t_2} \psi(t) \, dt. \end{aligned} \tag{2.39}$$

Remark 2.6. For every $x \in S_u$ we have

$$\Phi(\nu_u(x), (u')^a_-(x), (u')^a_+(x)) = |v'|(\{x\}) + \hat{\Phi}(\nu_u(x), (u')^a_-(x), (u')^a_+(x)),$$

where

$$\hat{\Phi}(1, t_1, t_2) := \int_{\max\{t_1,t_2\}}^{+\infty} \psi(t) \, dt \qquad \text{and} \qquad \hat{\Phi}(-1, t_1, t_2) := \int_{-\infty}^{\min\{t_1,t_2\}} \psi(t) \, dt.$$

In particular, for every Borel set $B \subset]a,b[$

$$\begin{aligned} &|v'|(B \setminus S_u) + \sum_{x\in S_u\cap B} \Phi(\nu_u, (u')^a_-, (u')^a_+) \\ &\quad = |v'|(B) + \sum_{x\in S_u\cap B} \hat{\Phi}(\nu_u, (u')^a_-, (u')^a_+) \geq |v'|(B). \end{aligned}$$

Proof of Theorem 2.5. Let $\mathcal{G}$ be the functional defined by the right-hand side of (2.38). We prove that, for every $u_k \to u$ in $L^1(]a,b[)$, we have

$$\mathcal{G}(u) \le \liminf_{k\to\infty} \mathcal{F}_1(u_k). \tag{2.40}$$

It is enough to consider sequences $\{u_k\}$ for which the liminf is a limit and has a finite value and $u_k \to u$ pointwise $\mathcal{L}^1$-a.e. in $]a,b[$. Then u_k belongs to $W^{2,1}(]a,b[)$ and (2.14) is satisfied. This implies that

$$|u'|(]a,b[) \le \liminf_{k\to\infty} \int_a^b |u_k'|\,dx. \tag{2.41}$$

Moreover, it follows from Theorem 2.4 that $u \in X^1_\psi(]a,b[)$ and that, up to a subsequence, $\{u_k'\}$ converges to $(u')^a$ pointwise $\mathcal{L}^1$-a.e. in $]a,b[$.

Let F be a finite subset of S_u. We want to prove that

$$|v'|(]a,b[\setminus F) + \sum_{x\in F} \Phi(\nu_u, (u')^a_-, (u')^a_+) \le \liminf_{k\to\infty} \int_a^b \psi(u_k')|u_k''|\,dx. \tag{2.42}$$

We write F as $\{x_1,\dots,x_m\}$, with $a < x_1 < \cdots < x_m < b$. For every $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) \in]0,\varepsilon[$ such that $a < x_1 - \delta < x_1 + \delta < x_2 - \delta < x_2 + \delta < \cdots < x_{m-1} - \delta < x_{m-1} + \delta < x_m - \delta < x_m + \delta < b$ and

$$|u(x_j-\delta) - u_-(x_j)| < \varepsilon, \quad |u(x_j+\delta) - u_+(x_j)| < \varepsilon, \tag{2.43}$$

$$|(u')^a(x_j-\delta) - (u')^a_-(x_j)| < \varepsilon, \quad |(u')^a(x_j+\delta) - (u')^a_+(x_j)| < \varepsilon, \tag{2.44}$$

$$u_k(x_j-\delta) \to u(x_j-\delta), \quad u_k(x_j+\delta) \to u(x_j+\delta) \quad \text{as} \quad k\to\infty, \tag{2.45}$$

$$u_k'(x_j-\delta) \to (u')^a(x_j-\delta), \quad u_k'(x_j+\delta) \to (u')^a(x_j+\delta) \quad \text{as} \quad k\to\infty, \tag{2.46}$$

$$|(u')^a(x_j-\delta)| + |(u')^a(x_j+\delta)| + \varepsilon < \frac{|[u](x_j)| - 4\varepsilon}{2\delta}$$

for $j = 1,\dots,m$.

Since $v_k \to v$ pointwise $\mathcal{L}^1$-a.e. in $]a,b[$ and $v_k' = \psi(u_k')u_k''$ $\mathcal{L}^1$-a.e. in $]a,b[$, we obtain

$$|v'|(]x_j+\delta, x_{j+1}-\delta[) \le \liminf_{k\to\infty} \int_{x_j+\delta}^{x_{j+1}-\delta} \psi(u_k')|u_k''|\,dx$$

for $j = 1,\dots,m-1$. A similar result holds for the intervals $]a,x_1-\delta[$ and $]x_m+\delta,b[$. Let F_δ be the union of the intervals $[x_j-\delta, x_j+\delta]$ for $j = 1,\dots,m$. Summing with respect to j and adding the contributions of the intervals $]a,x_1-\delta[$ and $]x_m+\delta,b[$, we obtain

$$|v'|(]a,b[\setminus F_\delta) \le \liminf_{k\to\infty} \int_{]a,b[\setminus F_\delta} \psi(u_k')|u_k''|\,dx. \tag{2.47}$$

We consider now the interval $I_j^\delta := [x_j-\delta, x_j+\delta]$, assuming that $[u](x_j) = u_+(x_j) - u_-(x_j) > 0$. By the mean value theorem there exists $y_{kj}^\delta \in]x_j-\delta, x_j+\delta[$ such that

$$u_k'(y_{kj}^\delta) = \frac{u_k(x_j+\delta) - u_k(x_j-\delta)}{2\delta} \ge \frac{[u](x_j) - 4\varepsilon}{2\delta}, \tag{2.48}$$

where the last inequality follows from (2.43) and (2.45) for k sufficiently large. By a change of variables we obtain

$$\int_{u_k'(x_j-\delta)}^{\frac{[u](x_j)-4\varepsilon}{2\delta}} \psi(t)\,dt \le \int_{x_j-\delta}^{y_{kj}^{\delta}} \psi(u_k')|u_k''|\,dx,$$
$$\int_{u_k'(x_j+\delta)}^{\frac{[u](x_j)-4\varepsilon}{2\delta}} \psi(t)\,dt \le \int_{y_{kj}^{\delta}}^{x_j+\delta} \psi(u_k')|u_k''|\,dx.$$

Adding these inequalities and taking the limit as $k\to\infty$ we obtain, thanks to (2.46),

$$\int_{(u')^a(x_j-\delta)}^{\frac{[u](x_j)-4\varepsilon}{2\delta}} \psi(t)\,dt + \int_{(u')^a(x_j+\delta)}^{\frac{[u](x_j)-4\varepsilon}{2\delta}} \psi(t)\,dt \le \liminf_{k\to\infty} \int_{x_j-\delta}^{x_j-\delta} \psi(u_k')|u_k''|\,dx. \tag{2.49}$$

Similarly, if $[u](x_j)<0$, then we have

$$\int_{\frac{[u](x_j)+4\varepsilon}{2\delta}}^{(u')^a(x_j-\delta)} \psi(t)\,dt + \int_{\frac{[u](x_j)+4\varepsilon}{2\delta}}^{(u')^a(x_j+\delta)} \psi(t)\,dt \le \liminf_{k\to\infty} \int_{x_j-\delta}^{x_j-\delta} \psi(u_k')|u_k''|\,dx. \tag{2.50}$$

From (2.47), (2.49), and (2.50) we deduce that

$$\begin{aligned}|v'|(]a,b[\setminus F_\delta) &+ \sum_{[u](x_j)>0}\left(\int_{(u')^a(x_j-\delta)}^{\frac{[u](x_j)-4\varepsilon}{2\delta}} \psi(t)\,dt + \int_{(u')^a(x_j+\delta)}^{\frac{[u](x_j)-4\varepsilon}{2\delta}} \psi(t)\,dt\right)\\ &+ \sum_{[u](x_j)<0}\left(\int_{\frac{[u](x_j)+4\varepsilon}{2\delta}}^{(u')^a(x_j-\delta)} \psi(t)\,dt + \int_{\frac{[u](x_j)+4\varepsilon}{2\delta}}^{(u')^a(x_j+\delta)} \psi(t)\,dt\right)\\ &\le \liminf_{k\to\infty}\int_a^b \psi(u_k')|u_k''|\,dx.\end{aligned}$$

Taking the limit as $\varepsilon\to 0$ (which implies $\delta(\varepsilon)\to 0$) we obtain (2.42), thanks to (2.44).

Since S_u is at most countable, (2.40) can be obtained from (2.42) by taking the supremum over all finite sets F contained in S_u.

Conversely, let $u\in X^1_\psi(]a,b[)$. We claim that there exists a sequence $\{u_k\}$ in $W^{2,1}(]a,b[)$ such that $u_k\to u$ in $L^1(]a,b[)$ and

$$\mathcal{G}(u) \ge \limsup_{k\to\infty} \mathcal{F}_1(u_k)\ . \tag{2.51}$$

It is clearly enough to consider the case $\mathcal{G}(u)<+\infty$.

We divide the proof into three steps.

Step 1. We prove (2.51) under the additional assumptions that $(u')^a$ is bounded and that $S_u=\{x_1,\dots,x_m\}$, with $x_1<\dots<x_m$. Note that, in this case, $Z[(u')^a]=\emptyset$ and, hence, $(u')^c=0$.

Construct a sequence $\{v_k\}$ in $W^{1,1}(]a,b[)$ such that $v_k\to v=\Psi_1\circ(u')^a$ pointwise $\mathcal{L}^1$-a.e. in $]a,b[$, $\Psi_1(-\|(u')^a\|_\infty)\le v_k\le\Psi_1(\|(u')^a\|_\infty)$, and

$$\int_a^b |v_k'(x)|\ dx \to |v'|(]a,b[)\,.$$

Setting $w_k:=\Psi_1^{-1}(v_k)$, we have $w_k\in W^{1,1}(]a,b[)$, thanks to (2.2),

$$w_k\to (u')^a \quad \text{pointwise } \mathcal{L}^1\text{-a.e. in }]a,b[, \tag{2.52}$$

and $\|w_k\|_\infty \le \left\|(u')^a\right\|_\infty$. Find $\delta_k \to 0^+$ such that

$$w_k\left(x_j-\delta_k\right) \to (u')^a_-\left(x_j\right), \quad w_k\left(x_j+\delta_k\right) \to (u')^a_+\left(x_j\right) \quad \text{for } j=1,\dots,m \tag{2.53}$$

and

$$\begin{gathered} \int_{x_{j-1}+\delta_k}^{x_j-\delta_k} |v_k'|\,dx \to |v'|(]x_{j-1},x_j[) \quad \text{for } j=2,\dots,m, \\ \int_a^{x_1-\delta_k} |v_k'|\,dx \to |v'|(]a,x_1[), \ \int_{x_m+\delta_k}^b |v_k'|\,dx \to |v'|(]x_m+\delta_k,b[). \end{gathered} \tag{2.54}$$

By (2.52) and by the dominated convergence theorem we have

$$u_+(x_{j-1}) + \int_{x_{j-1}+\delta_k}^{x_j-\delta_k} w_k(s)\,ds \longrightarrow u_+(x_{j-1}) + \int_{x_{j-1}}^{x_j} (u')^a\,ds = u_-(x_j) \tag{2.55}$$

for $j=2,\dots,m$, with the obvious changes for $j=1$ and $j=m+1$.

To deal with the jump point x_j, assume first that

$$u_+\left(x_j\right) - u_-\left(x_j\right) > 0. \tag{2.56}$$

In this case, we need to construct functions $f_{kj} \in C^2([x_j-\delta_k, x_j+\delta_k])$ that satisfy the following properties: There exists $y_{kj} \in]x_j-\delta_k, x_j+\delta_k[$ such that

$$f_{kj}\left(x_j-\delta_k\right) = u_+\left(x_{j-1}\right) + \int_{x_{j-1}+\delta_k}^{x_j-\delta_k} w_k\left(s\right)\,ds, \quad f_{kj}\left(x_j+\delta_k\right) = u_+\left(x_j\right), \tag{2.57}$$

$$f_{kj}'\left(x_j-\delta_k\right) = w_k\left(x_j-\delta_k\right), \quad f_{kj}'\left(x_j+\delta_k\right) = w_k\left(x_j+\delta_k\right), \tag{2.58}$$

$$f_{kj}''(x) > 0 \text{ for } x \in]x_j-\delta_k, y_{kj}[, \quad f_{kj}''(x) < 0 \text{ for } x \in]y_{kj}, x_j+\delta_k[, \tag{2.59}$$

$$\left|f_{kj}\left(x_j-\delta_k\right) - \min_{[x_j-\delta_k, y_{kj}]} f_{kj}\right| \le \tfrac{1}{k}, \quad \left|f_{kj}\left(x_j+\delta_k\right) - \max_{[y_{kj}, x_j+\delta_k]} f_{kj}\right| \le \tfrac{1}{k}, \tag{2.60}$$

where we replace x_{j-1} and $x_{j-1}-\delta_k$ by a in the case $j=1$.

We now discuss briefly the existence of such functions. We observe that the latter conditions in (2.57)–(2.59) imply that the graph of f_{kj} in the interval $[y_{kj}, x_j+\delta_k[$ lies below the straight line passing through the point $(x_j+\delta_k, u_+(x_j))$ with slope $w_k(x_j+\delta_k)$, i.e.,

$$f_{kj}(x) \le u_+(x_j) + w_k(x_j+\delta_k)(x-x_j-\delta_k)$$

for $x \in [y_{kj}, x_j+\delta_k[$. It is then easy to see that the inequality

$$u_+\left(x_j\right) - 2w_k\left(x_j+\delta_k\right)\delta_k - u_+\left(x_{j-1}\right) - \int_{x_{j-1}+\delta_k}^{x_j-\delta_k} w_k\left(s\right)\,ds > 0 \tag{2.61}$$

allows us to fulfill also the former conditions in (2.57)–(2.59), as well as (2.60). By (2.53), (2.55), and (2.56), inequality (2.61) is satisfied when δ_k is small enough.

If the left-hand side of (2.56) is negative, then we choose f_{kj} so that (2.57) and (2.58) hold, and there exists $y_{kj} \in]x_j-\delta_k, x_j+\delta_k[$ such that

$$\begin{gathered} f_{kj}''(x) < 0 \text{ for } x \in]x_j-\delta_k, y_{kj}[, \quad f_{kj}''(x) > 0 \text{ for } x \in]y_{kj}, x_j+\delta_k[, \\ \left|f_{kj}\left(x_j-\delta_k\right) - \max_{[x_j-\delta_k, x_j+\delta_k]} f_{kj}\right| \le \tfrac{1}{k}, \quad \left|f_{kj}\left(x_j+\delta_k\right) - \min_{[x_j-\delta_k, x_j+\delta_k]} f_{kj}\right| \le \tfrac{1}{k}. \end{gathered}$$

In the same way the construction is possible if δ_k is small enough.

We are now ready to define the approximating sequence

$$u_k(x) := \begin{cases} u_+(a) + \int_a^x w_k(s)\, ds & \text{if } a \le x < x_1 - \delta_k, \\ f_{kj}(x) & \text{if } x_j - \delta_k \le x < x_j + \delta_k,\ j = 1, \dots, m, \\ u_+(x_{j-1}) + \int_{x_{j-1}+\delta_k}^x w_k(s)\, ds & \text{if } x_{j-1} + \delta_k \le x < x_j - \delta_k,\ j = 2, \dots, m, \\ u_+(x_m) + \int_{x_m+\delta_k}^x w_k(s)\, ds & \text{if } x_m + \delta_k \le x < b. \end{cases}$$

Let us define $x_0 := a$ and $x_{m+1} := b$. Since $w_k \to (u')^a$ in $L^1(]a,b[)$, we have

$$u_k(x) \to u_+(x_{j-1}) + \int_{x_{j-1}}^x (u')^a(s)\, ds = u(x)$$

for every $x \in]x_{j-1}, x_j[$ and $j = 1, \dots, m+1$ and, in turn, $u_k \to u$ in $L^1(]a,b[)$. As

$$\begin{aligned} \int_{x_{j-1}+\delta_k}^{x_j-\delta_k} |u_k'|\, dx + \int_{x_{j-1}+\delta_k}^{x_j-\delta_k} \psi(u_k')\, |u_k''|\, dx &= \int_{x_{j-1}+\delta_k}^{x_j-\delta_k} |w_k|\, dx + \int_{x_{j-1}+\delta_k}^{x_j-\delta_k} \psi(w_k)\, |w_k'|\, dx \\ &\le \int_{x_{j-1}}^{x_j} |w_k|\, dx + \int_{x_{j-1}+\delta_k}^{x_j-\delta_k} |v_k'|\, dx, \end{aligned}$$

by (2.54) and the fact that $w_k \to (u')^a$ in $L^1(]a,b[)$ we have

$$\begin{aligned} &\limsup_{k\to\infty} \left(\int_{x_{j-1}+\delta_k}^{x_j-\delta_k} |u_k'|\, dx + \int_{x_{j-1}+\delta_k}^{x_j-\delta_k} \psi(u_k')\, |u_k''|\, dx \right) \\ &\qquad \le \int_{x_{j-1}}^{x_j} \left|(u')^a\right| dx + |v'|(]x_{j-1}, x_j[). \end{aligned} \tag{2.62}$$

Similarly,

$$\begin{aligned} &\limsup_{k\to\infty} \left(\int_a^{x_1-\delta_k} |u_k'|\, dx + \int_a^{x_1-\delta_k} \psi(u_k')\, |u_k''|\, dx \right) \\ &\qquad \le \int_a^{x_1} \left|(u')^a\right| dx + |v'|(]a, x_1[), \end{aligned} \tag{2.63}$$

$$\begin{aligned} &\limsup_{k\to\infty} \left(\int_{x_m+\delta_k}^{b} |u_k'|\, dx + \int_{x_m+\delta_k}^{b} \psi(u_k')\, |u_k''|\, dx \right) \\ &\qquad \le \int_{x_m}^{b} \left|(u')^a\right| dx + |v'|(]x_m, b[). \end{aligned} \tag{2.64}$$

Assume that $[u](x_j) = u_+(x_j) - u_-(x_j) > 0$. Then (2.56) holds for the functions u_k

if k is sufficiently large. By (2.58), (2.59), (2.60), and a change of variables we obtain

$$\begin{aligned}
&\int_{x_j-\delta_k}^{x_j+\delta_k} |u_k'|\ dx + \int_{x_j-\delta_k}^{x_j+\delta_k} \psi\left(u_k'\right) |u_k''|\ dx \\
&\qquad = \int_{x_j-\delta_k}^{x_j+\delta_k} |f_{kj}'|\, dx + \int_{x_j-\delta_k}^{x_j+\delta_k} \psi\left(f_{kj}'\right) \left|f_{kj}''\right|\ dx \\
&\qquad \le f_{kj}\left(x_j+\delta_k\right) - f_{kj}\left(x_j-\delta_k\right) + \int_{w_k(x_j-\delta_k)}^{f_{kj}'(y_{kj})} \psi\left(t\right)\ dt \\
&\qquad\quad + \int_{w_k(x_j+\delta_k)}^{f_{kj}'(y_{kj})} \psi\left(t\right)\ dt + \frac{2}{k}. \qquad (2.65)
\end{aligned}$$

By (2.59) we have

$$f_{kj}'\left(y_{kj}\right) = \max_{[x_j-\delta_k, x_j+\delta_k]} f_{kj}' \ge \frac{1}{2\delta_k}\left[f_{kj}\left(x_j+\delta_k\right) - f_{kj}\left(x_j-\delta_k\right)\right]. \qquad (2.66)$$

By (2.57) and the fact that $w_k \to (u')^a$ in $L^1\left(]a,b[\right)$ we obtain

$$f_{kj}\left(x_j+\delta_k\right) - f_{kj}\left(x_j-\delta_k\right) \to u_+\left(x_j\right) - \left(u_+\left(x_{j-1}\right) + \int_{x_{j-1}}^{x_j} \left(u'\right)^a\ ds\right) = [u]\left(x_j\right).$$

In turn, using (2.66), we get that $f_{kj}'\left(y_{kj}\right) \to \infty$. Thus, letting $k \to \infty$ in (2.65) and using (2.53), we infer

$$\begin{aligned}
&\limsup_{k\to\infty} \left(\int_{x_j-\delta_k}^{x_j+\delta_k} |u_k'|\ dx + \int_{x_j-\delta_k}^{x_j+\delta_k} \psi\left(u_k'\right) |u_k''|\ dx\right) \\
&\le [u]\left(x_j\right) + \int_{(u')_-^a\left(x_j\right)}^{+\infty} \psi\left(t\right)\ dt + \int_{(u')_+^a\left(x_j\right)}^{+\infty} \psi\left(t\right)\ dt. \qquad (2.67)
\end{aligned}$$

Similarly, if $[u]\left(x_j\right) = u_+\left(x_j\right) - u_-\left(x_j\right) < 0$, we find

$$\begin{aligned}
&\limsup_{k\to\infty} \left(\int_{x_j-\delta_k}^{x_j+\delta_k} |u_k'|\ dx + \int_{x_j-\delta_k}^{x_j+\delta_k} \psi\left(u_k'\right) |u_k''|\ dx\right) \\
&\le \left|[u]\left(x_j\right)\right| + \int_{-\infty}^{\left(u'\right)_-^a\left(x_j\right)} \psi\left(t\right)\ dt + \int_{-\infty}^{\left(u'\right)_+^a\left(x_j\right)} \psi\left(t\right)\ dt. \qquad (2.68)
\end{aligned}$$

Summing over j in (2.62), (2.67), and (2.68) and combining with (2.63) and (2.64), inequality (2.51) follows.

Step 2. Assume only that $u \in X_\psi^1(]a,b[)$ and that S_u is finite. We claim that there exists a sequence $\{u_k\}$ such that $u_k \to u$ in $L^1\left(]a,b[\right)$, each u_k satisfies the hypotheses of Step 1, and

$$\mathcal{G}\left(u\right) \ge \limsup_{k\to\infty} \mathcal{G}\left(u_k\right). \qquad (2.69)$$

Note that, if (2.69) holds, then, by applying Step 1 to each u_k, we may find a sequence $u_{km} \in W^{2,1}\left(]a,b[\right)$ converging to u_k in $L^1\left(]a,b[\right)$ and satisfying

$$\mathcal{G}\left(u_k\right) \ge \limsup_{m\to\infty} \mathcal{F}_1\left(u_{km}\right).$$

By (2.69) we then have

$$\mathcal{G}(u) \geq \limsup_{k\to\infty} \limsup_{m\to\infty} \mathcal{F}_1(u_{km}),$$

and a standard diagonalization argument now yields the existence of a sequence $m_k \to \infty$ such that $u_{km_k} \to u$ in $L^1(]a,b[)$ and

$$\mathcal{G}(u) \geq \limsup_{k\to\infty} \mathcal{F}_1(u_{km_k}).$$

In the construction of the sequence satisfying (2.69), we need to consider the precise representatives $(u')^a_\vee$ and $(u')^a_\wedge$ defined in Remark 2.2. We recall that $(u')^a_\vee$ is upper semicontinuous while $(u')^a_\wedge$ is lower semicontinuous, and so for each $k \in \mathbb{N}$ we may decompose the open sets $\{(u')^a_\wedge > k\}$ and $\{(u')^a_\vee < -k\}$ into the union of two finite sequences of pairwise disjoint open sets U^+_{kj} and U^-_{kj}, that is,

$$\bigcup_j U^+_{kj} = \{(u')^a_\wedge > k\}, \quad \bigcup_j U^-_{kj} = \{(u')^a_\vee < -k\},$$

such that

$$\text{diam}\left(U^+_{kj}\right) \leq \mathcal{L}^1(\{(u')^a_\wedge > k\}),\ \text{diam}\left(U^-_{kj}\right) \leq \mathcal{L}^1(\{(u')^a_\vee < -k\}) \tag{2.70}$$

for every j. Note that, setting $v_\vee := \Psi_1 \circ (u')^a_\vee$ and $v_\wedge := \Psi_1 \circ (u')^a_\wedge$, we have

$$|v'|(]c,d[) = \text{Var}(v_\vee;]c,d[) = \text{Var}(v_\wedge;]c,d[) \tag{2.71}$$

for every interval $]c,d[\subset]a,b[$.

For every set $U^\pm_{kj}$, we fix a nonnegative function $g^\pm_{kj} \in C^1_c(U^\pm_{kj})$ such that

$$\int_{U^\pm_{kj}} g^\pm_{kj}(x)\, dx = ((u')^c)^\pm\left(U^\pm_{kj}\right), \tag{2.72}$$

and $(g^\pm_{kj})'$ has only one zero in the interior of the support of $g^\pm_{kj}$. Then we define

$$g^+_k := \sum\nolimits_j g^+_{kj}, \quad g^-_k := \sum\nolimits_j g^-_{kj}, \quad g_k := g^+_k - g^-_k, \quad w_k := T^k_{-k} \circ (u')^a + g_k, \tag{2.73}$$

where, for any pair of constants $h < k$, the truncation function T^k_h is defined by

$$T^k_h(t) := \begin{cases} h & \text{for } t \leq h, \\ t & \text{for } h \leq t \leq k, \\ k & \text{for } t \geq k. \end{cases}$$

We claim that

$$w_k\, \mathcal{L}^1 \rightharpoonup (u')^a\, \mathcal{L}^1 + (u')^c \ \text{ weakly}^* \text{ in } M_b(]a,b[)\,. \tag{2.74}$$

Define

$$A_k := \left\{(u')^a_\wedge > k\right\} \cup \left\{(u')^a_\vee < -k\right\}.$$

Since by the Chebychev inequality

$$k\mathcal{L}^1(A_k) \to 0, \tag{2.75}$$

it suffices to show that

$$\left(\sum_j g_{kj}^{\pm}\right) \mathcal{L}^1 \rightharpoonup \left((u')^c\right)^{\pm} \text{ weakly}^* \text{ in } M_b\left(]a,b[\right). \tag{2.76}$$

Let $\varphi \in C_0\left(]a,b[\right)$ and $\varepsilon > 0$. By uniform continuity, there exists $\delta = \delta(\varepsilon) > 0$ such that $|\varphi(x) - \varphi(y)| \le \varepsilon$ for all $x,\ y \in]a,b[$ with $|x-y| \le \delta$. In view of (2.70) and (2.75), for all k sufficiently large and for all j we have that $\operatorname{diam}\left(U_{kj}^{\pm}\right) \le \delta$. Let us fix $y_{kj}^{\pm} \in U_{kj}^{\pm}$. Then by (2.72)

$$\begin{aligned}
&\Big| \int_{U_{kj}^{\pm}} \varphi(x)\, g_{kj}^{\pm}(x)\, dx - \int_{U_{kj}^{\pm}} \varphi(x)\, d\left((u')^c\right)^{\pm}(x) \Big| \\
&\qquad = \left| \int_{U_{kj}^{\pm}} \left[\varphi(x) - \varphi\left(y_{kj}^{\pm}\right)\right] g_{kj}^{\pm}(x)\, dx - \int_{U_{kj}^{\pm}} \left[\varphi(x) - \varphi\left(y_{kj}^{\pm}\right)\right]\, d\left((u')^c\right)^{\pm}(x) \right| \\
&\qquad \le \varepsilon \left(\int_{U_{kj}^{\pm}} g_{kj}^{\pm}(x)\, dx + \left((u')^c\right)^{\pm}\left(U_{kj}^{\pm}\right) \right) \le 2\varepsilon \left((u')^c\right)^{\pm}\left(U_{kj}^{\pm}\right).
\end{aligned}$$

Summing over j and using the fact that the measures $\left(\sum_j g_{kj}^{+}\right)\mathcal{L}^1$ and $\left((u')^c\right)^{+}$ are concentrated on $\{(u')_{\wedge}^a > k\}$, while the measures $\left(\sum_j g_{kj}^{-}\right)\mathcal{L}^1$ and $\left((u')^c\right)^{-}$ are concentrated on $\{(u')_{\vee}^a < -k\}$ (see Remark 2.2), we obtain (2.76).

Moreover, we claim that

$$\lim_{k\to\infty} \int_a^b |w_k|\, dx = \int_a^b |\,(u')^a\,|\, dx + |\,(u')^c\,|\,(]a,b[)\ . \tag{2.77}$$

Indeed, using (2.72), (2.73), and Remark 2.2, we deduce that

$$\begin{aligned}
\int_a^b |w_k|\, dx &\le \int_{\{|(u')^a| \le k\}} \left|(u')^a\right|\, dx + k\mathcal{L}^1(A_k) + \sum_j \int_{U_{kj}^{+}} g_{kj}^{+}\, dx + \sum_j \int_{U_{kj}^{-}} g_{kj}^{-}\, dx \\
&\le \int_a^b \left|(u')^a\right|\, dx + k\mathcal{L}^1(A_k) + \sum_j \left((u')^c\right)^{+}\left(U_{kj}^{+}\right) + \sum_j \left((u')^c\right)^{-}\left(U_{kj}^{-}\right) \\
&\le \int_a^b \left|(u')^a\right|\, dx + k\mathcal{L}^1(A_k) + \left|(u')^c\right|\left(]a,b[\right),
\end{aligned}$$

and the limit superior inequality follows from (2.75). The limit inferior inequality follows from (2.74) and the lower semicontinuity of the total variation.

Set

$$u_k(x) := u_+(a) + \int_a^x w_k(s)\, ds + \sum_{x_j < x,\, x_j \in S_u} [u](x_j) \tag{2.78}$$

and $v_k := \Psi_1 \circ (u_k')^a = \Psi_1 \circ w_k$.

We claim that $u_k \to u$ in $L^1(]a,b[)$. For $x \in]a,b[$ by (2.74) and (2.77) it follows that

$$\int_a^x w_k\,dy \to \int_a^x (u')^a\,dy + (u')^c(]a,x[),$$

and so u_k converges to u pointwise $\mathcal{L}^1$-a.e. and, in turn, in $L^1(]a,b[)$.

Next, we show that

$$\limsup_{k\to\infty} \mathcal{G}(u_k) \le \mathcal{G}(u). \tag{2.79}$$

From (2.77) we get

$$|u_k'|(]a,b[) \to |u'|(]a,b[)\ . \tag{2.80}$$

Moreover, as $v_k = T^{\Psi_1(k)}_{\Psi_1(-k)} \circ v$ $\mathcal{L}^1$-a.e. in the open set $V_k :=]a,b[\setminus \operatorname{supp} g_k$, we have $|v_k'| \le |v'|$ as measures in V_k. In particular, this yields $|v_k'|(]a,b[\setminus (A_k \cup S_u)) \le |v'|(]a,b[\setminus (A_k \cup S_u))$ and, hence,

$$|v_k'|(]a,b[\setminus (A_k \cup S_u)) \le |v'|(]a,b[\setminus (A_\infty \cup S_u))\ , \tag{2.81}$$

where

$$A_\infty := \bigcap_k A_k = \left\{(u')^a_\wedge = +\infty\right\} \cup \left\{(u')^a_\vee = -\infty\right\}.$$

Using the properties of g^+_{kj}, we have

$$\begin{aligned} |v_k'|\left(\left\{(u')^a_\wedge > k\right\} \setminus S_u\right) &= \sum_j \int_{U^+_{kj}} \psi\left(k + g^+_{kj}\right) \left|\left(g^+_{kj}\right)'\right| dx \\ &= 2\sum_j \int_k^{k+\sup g^+_{kj}} \psi(t)\,dt \le 2\mathcal{H}^0\left(\left\{j:\ ((u')^c)^+\left(U^+_{kj}\right) > 0\right\}\right) \int_k^\infty \psi(t)\ dt\,. \end{aligned} \tag{2.82}$$

We claim that

$$\begin{aligned} &2\mathcal{H}^0\left(\left\{j:\ ((u')^c)^+\left(U^+_{kj}\right) > 0\right\}\right) \int_k^\infty \psi(t)\ dt \\ &\qquad\qquad \le |v'|\left(\left\{(u')^a_\wedge > k\right\} \setminus S_u\right) + 4\int_k^\infty \psi(t)\ dt. \end{aligned}$$

Indeed, if $((u')^c)^+(U^+_{kj}) > 0$, then there exists a connected component $I^+_{kj} =]a_{kj}, b_{kj}[$ of $U^+_{kj} \setminus S_u$ such that $((u')^c)^+(I^+_{kj}) > 0$. Assume that $I^+_{kj} \subset\subset]a,b[$. Then by Remark 2.2 we may find $c_{kj} \in I^+_{kj}$ such that $(u')^a_\wedge(c_{kj}) = +\infty$, while $(u')^a_\wedge(a_{kj})$, $(u')^a_\wedge(b_{kj}) \le k$. Hence, by (2.71)

$$|v'|\left(U^+_{kj} \setminus S_u\right) \ge |v'|\left(I^+_{kj}\right) \ge 2\int_k^\infty \psi(t)\,dt.$$

Summing over all such intervals and adding the possible contribution of the intervals I^+_{kj} with at least one end point in $\{a,b\}$, we obtain the claim. In turn, by (2.82) we have

$$|v_k'|\left(\left\{(u')^a_\wedge > k\right\} \setminus S_u\right) \le |v'|\left(\left\{(u')^a_\wedge > k\right\} \setminus S_u\right) + 4\int_k^\infty \psi(t)\ dt.$$

A similar estimate holds for the set $\{(u')^a_\vee < -k\} \setminus S_u$, thus, yielding

$$\limsup_{k\to\infty} |v'_k|\,(A_k \setminus S_u) \le |v'|\,(A_\infty \setminus S_u)\ . \tag{2.83}$$

Combining (2.81) with (2.83) we obtain

$$\limsup_{k\to\infty} |v'_k|\,(]a,b[\setminus S_u) \le |v'|\,(]a,b[\setminus S_u)\ .$$

Next, we show that

$$\lim_{k\to\infty} \sum_{x\in S_{u_k}} \Phi\left(\nu_{u_k}, (u'_k)^a_-, (u'_k)^a_+\right) = \sum_{x\in S_u} \Phi\left(\nu_u, (u')^a_-, (u')^a_+\right). \tag{2.84}$$

Note that $S_{u_k} = S_u$ and $\nu_{u_k}(x) = \nu_u(x)$ for all k by (2.78). Moreover, for every $x \in S_u$ if $(u')^a_+(x) \in \mathbb{R}$, then $\left|(u')^a_+(y)\right| \le k_0$ for all y in a right neighborhood of x and for some integer k_0. Thus, by (2.73) and (2.78) we have that $(u'_k)^a(y) = (u')^a(y)$ for $k \ge k_0$ and for $\mathcal{L}^1$-a.e. y in the same right neighborhood. In turn, by (2.7) we infer $(u'_k)^a_+(x) = (u')^a_+(x)$ for all $k \ge k_0$. If $(u')^a_+(x) = \infty$, then for all k we have $(u')^a_+ > k$ in a right neighborhood of x by right continuity (see Remark 2.2). By construction this implies that $(u'_k)^a = w_k \ge k$ $\mathcal{L}^1$-a.e. in the same right neighborhood. Thus, $(u'_k)^a_+(x) \ge k \to (u')^a_+(x)$. Similarly, $(u'_k)^a_-(x) \to (u')^a_-(x)$ so that

$$\Phi\left(\nu_{u_k}(x), (u'_k)^a_-(x), (u'_k)^a_+(x)\right) \to \Phi\left(\nu_u(x), (u')^a_-(x), (u')^a_+(x)\right)\ .$$

Hence, (2.84) follows. This, together with (2.80) and (2.83), yields (2.79).

Step 3. Let now u be an arbitrary function in $X^1_\psi(]a,b[)$ such that $\mathcal{G}(u) < +\infty$. As in the previous step it suffices to construct $u_k \in X^1_\psi(]a,b[)$ satisfying the hypotheses of Step 2, converging to u in $L^1(]a,b[)$ and such that (2.79) holds. Write $S_u = \{x_j\}$, and for each k define $S^k_u := \{x_j : j \le k\}$ and

$$u_k(x) = u_+(a) + \int_a^x (u')^a\ dt + (u')^c(]a,x[) + \sum_{x_j<x,\, x_j\in S^k_u} [u](x_j)\,.$$

It is clear that $\{u_k\}$ converges to u in $L^1(]a,b[)$ and that $|u'_k|(]a,b[) \to |u'|(]a,b[)$. Moreover, $|v'_k|(]a,b[\setminus S_u) = |v'|(]a,b[\setminus S_u)$ and

$$\begin{aligned}\lim_{k\to\infty} \sum_{x\in S_{u_k}} \Phi\left(\nu_{u_k}, (u'_k)^a_-, (u'_k)^a_+\right) &= \lim_{k\to\infty} \sum_{x\in S^k_u} \Phi\left(\nu_u, (u')^a_-, (u')^a_+\right)\\ &= \sum_{x\in S_u} \Phi\left(\nu_u, (u')^a_-, (u')^a_+\right)\ .\end{aligned}$$

This concludes the proof of the theorem. □

We end the section with a compactness result for energy bounded sequences in $X^1_\psi(]a,b[)$.

COROLLARY 2.7. *Let $\{u_k\}$ be a sequence of functions in $X^1_\psi(]a,b[)$ bounded in $L^1(]a,b[)$ and such that*

$$C := \sup_k \overline{\mathcal{F}}_1(u_k) < +\infty. \tag{2.85}$$

Then there exist a subsequence (not relabeled) $\{u_k\}$ and a function $u \in X^1_\psi(]a,b[)$ such that

$$u_k \rightharpoonup u \quad \text{weakly}^* \text{ in } BV(]a,b[), \tag{2.86}$$

$$\Psi_1 \circ (u_k')^a \rightharpoonup \Psi_1 \circ (u')^a \quad \text{weakly}^* \text{ in } BV(]a,b[), \tag{2.87}$$

$$(u_k')^a \to (u')^a \quad \text{pointwise } \mathcal{L}^1\text{-a.e. in }]a,b[.$$

Proof. It is well known that convergence in measure is metrizable with the following metric:

$$d(u_1,u_2) := \int_a^b \frac{|u_1-u_2|}{1+|u_1-u_2|}\,dx,$$

where u_1 and u_2 are (equivalent classes of) measurable functions.

By Theorems 2.4 and 2.5, for every $k \in \mathbb{N}$ we may find $w_k \in W^{2,1}(]a,b[)$ such that

$$\int_a^b |u_k - w_k|\,dx \le \frac{1}{k}, \quad d((u_k')^a, w_k') \le \frac{1}{k}, \tag{2.88}$$

and

$$\mathcal{F}_1(w_k) \le C+1.$$

By Theorem 2.4 we may find a subsequence (not relabeled) of $\{w_k\}$ and a function $u \in X^1_\psi(]a,b[)$ such that (2.15), (2.16), and (2.17) hold (with w_k in place of u_k). It now follows from (2.88) that $u_k \to u$ in $L^1(]a,b[)$ and $(u_k')^a \to (u')^a$ in measure and, hence, pointwise $\mathcal{L}^1$-a.e. in $]a,b[$, up to a further subsequence. From the bound (2.85), the uniqueness of the limit, and the invertibility of Ψ_1, we deduce (2.86) and (2.87). ∎

3. The case $p > 1$. In this section we analyze the functional (1.3) in the case $p > 1$.

Let us state precisely the standing assumptions. Throughout this section p denotes any exponent in $]1,+\infty[$, $\psi\colon \mathbb{R} \to]0,+\infty[$ is a bounded Borel function satisfying

$$M := \int_{-\infty}^{+\infty} (\psi(t))^{1/p}\,dt < +\infty \tag{3.1}$$

in addition to (2.2), and $\Psi_p\colon \overline{\mathbb{R}} \to [0,M]$ denotes the antiderivative of $\psi^{1/p}$ defined by

$$\Psi_p(t) := \int_{-\infty}^{t} (\psi(s))^{1/p}\,ds. \tag{3.2}$$

The function $\Psi_p^{-1}\colon [0,M] \to \overline{\mathbb{R}}$ stands for the inverse function of Ψ_p.

We now consider the functional $\mathcal{F}_p\colon L^1(]a,b[) \to [0,+\infty]$ defined by

$$\mathcal{F}_p(u) := \begin{cases} \displaystyle\int_a^b |u'|\,dx + \int_a^b \psi(u')|u''|^p\,dx & \text{if } u \in W^{2,p}(]a,b[), \\ +\infty & \text{otherwise.} \end{cases} \tag{3.3}$$

It turns out that piecewise smooth functions with bounded derivative and nonempty discontinuity set cannot be approximated by sequences with equibounded energy. This is a consequence of Remark 3.2(i) and Theorem 3.3 below, and to this end we introduce a suitable space of functions. Recall that $Z^{\pm}[(u')^a]$ are the sets defined in (2.5) and (2.6), while $(u')^s$ denotes the singular part of the gradient measure u'.

DEFINITION 3.1. Let $X^p_\psi(]a,b[)$ be the set of all functions $u \in BV(]a,b[)$ such that $v := \Psi_p \circ (u')^a$ belongs to $W^{1,p}(]a,b[)$, and the positive part $((u')^s)^+$ and the negative part $((u')^s)^-$ of the measure $(u')^s$ are concentrated on $Z^+[(u')^a]$ and $Z^-[(u')^a]$, respectively.

Remark 3.2. (i) It follows immediately from the definition that if $u \in X^p_\psi(]a,b[)$, then $(u')^a = \Psi_p^{-1}(v)$ is continuous on $[a,b]$ with values in $\overline{\mathbb{R}}$. In particular, it turns out that

$$Z^{\pm}[(u')^a] = \{x \in]a,b[: (u')^a = \pm\infty\}.$$

By the hypothesis on the support of the singular part $(u')^s$, we have $\lim_{x\to x_0}(u')^a(x) = +\infty$ for every jump point x_0 with $u_+(x_0) - u_-(x_0) > 0$ and $\lim_{x\to x_0}(u')^a(x) = -\infty$ for every jump point x_0 with $u_+(x_0) - u_-(x_0) < 0$. This means that if S_u is nonempty, then u cannot have bounded derivative outside the discontinuity set. In particular, piecewise constant functions are not included in the class $X^p_\psi(]a,b[)$.

(ii) We observe that the function $(u')^a$ is differentiable $\mathcal{L}^1$-a.e. in $]a,b[$ with

$$v' = \psi^{\frac{1}{p}}\left((u')^a\right)\left((u')^a\right)'. \tag{3.4}$$

To see this, we consider the open set

$$A_k := \{x \in]a,b[: -k < (u')^a < k\}.$$

Since by (2.2) the function Ψ_p^{-1} is Lipschitz continuous in the interval $[\Psi_p(-k), \Psi_p(k)]$ and $v \in W^{1,p}(]a,b[)$, by the chain rule, we have that $(u')^a = \Psi_p^{-1} \circ v \in W^{1,p}(A_k)$ and, in particular, it is differentiable $\mathcal{L}^1$-a.e. in A_k, and (3.4) holds. Since $(u')^a$ is integrable, we have that

$$\mathcal{L}^1\left(]a,b[\setminus \bigcup_k A_k\right) = 0,$$

and the conclusion follows.

(iii) It is easy to check that $X^p_\psi(]a,b[)$ may contain discontinuous functions. An example is given by the following construction: Let $\psi\colon \mathbb{R} \to]0,+\infty[$ be defined by

$$\psi(t) := \begin{cases} 1 & \text{if } |t| \le 1, \\ \dfrac{1}{|t|^\alpha} & \text{if } |t| > 1, \end{cases}$$

where α is any number in $]1,+\infty[$, and let $p \in]1, \frac{\alpha+1}{2}[$. Consider now the discontinuous functions $u\colon]-1,1[\to \mathbb{R}$ given by

$$u(x) := \begin{cases} -|x|^\beta & \text{if } x \le 0, \\ 1 + x^\beta & \text{if } x > 0, \end{cases}$$

with

$$0 < \beta < 1 - \frac{p-1}{\alpha - p}.$$

A simple computation shows that the function $\Psi_p \circ (u')^a$ belongs to $W^{1,p}(]-1,1[)$, which in turn implies that $u \in X^p_\psi(]-1,1[)$.

(iv) Finally, the same construction of Proposition 2.3 shows that, for every admissible ψ satisfying (2.9), the space $X^p_\psi(]a,b[)$ contains a function with nontrivial Cantor part if p is sufficiently close to 1. We omit the details of this fact, which can be easily checked following step by step the proof of Proposition 2.3.

The next theorem is the counterpart of Theorem 2.4 for the case $p > 1$. It establishes that energy bounded sequences are relatively compact in $X^p_\psi(]a,b[)$. The proof is similar to the one of Theorem 2.4; nevertheless, since this is the main result of this section we reproduce it in full detail for the reader's convenience.

THEOREM 3.3. *Let $\{u_k\}$ be a sequence of functions bounded in $L^1(]a,b[)$ and such that*

$$C := \sup_k \mathcal{F}_p(u_k) < +\infty. \tag{3.5}$$

Then there exist a subsequence (not relabeled) $\{u_k\}$ and a function $u \in X^p_\psi(]a,b[)$ such that

$$u_k \rightharpoonup u \quad \textit{weakly}^* \textit{ in } BV(]a,b[), \tag{3.6}$$

$$\Psi_p \circ u_k' \rightharpoonup \Psi_p \circ (u')^a \quad \textit{weakly in } W^{1,p}(]a,b[),$$

$$u_k' \to (u')^a \quad \textit{pointwise in }]a,b[. \tag{3.7}$$

Proof. By (3.3) and (3.5) we may assume that each u_k belongs to $W^{2,p}(]a,b[)$ and that

$$C_1 := \sup_k \int_a^b [\,|u_k| + |u_k'| + \psi(u_k')|u_k''|^p\,]\,dx < +\infty. \tag{3.8}$$

Let us define

$$v_k := \Psi_p \circ u_k'. \tag{3.9}$$

As Ψ_p is Lipschitz in $\mathbb{R}$, the functions v_k belong to $W^{1,p}(]a,b[)$ and

$$v_k' = (\psi(u_k'))^{1/p} u_k'' \qquad \mathcal{L}^1\text{-a.e. on }]a,b[. \tag{3.10}$$

It follows from (3.1) and (3.5) that

$$\int_a^b [\,|v_k|^p + |v_k'|^p\,]\,dx \le M^p(b-a) + C_1. \tag{3.11}$$

By (3.8) and (3.11), passing to a subsequence (not relabeled), we may assume that

$$u_k \rightharpoonup u \quad \text{weakly}^* \text{ in } BV(]a,b[)$$

and

$$v_k \rightharpoonup v \quad \text{weakly in } W^{1,p}(]a,b[) \tag{3.12}$$

for some functions $u \in BV(]a,b[)$ and $v \in W^{1,p}(]a,b[;[0,M])$.

Since Ψ_p^{-1} is continuous, we obtain

$$u_k' = \Psi_p^{-1} \circ v_k \to w := \Psi_p^{-1} \circ v \quad \text{pointwise in }]a,b[. \tag{3.13}$$

Note, also, that w is continuous with values in $\overline{\mathbb{R}}$.

We now split the remaining part of the proof into two steps.

Step 1. We prove that

$$w = (u')^a \quad \mathcal{L}^1\text{-a.e. on }]a,b[. \tag{3.14}$$

If not, arguing as for (2.27), we may find $t_0 > 0$ and an infinite number of disjoint open intervals I such that

$$\mathcal{L}^1\left(\{w \neq (u')^a\} \cap \{|w| < t_0\} \cap I\right) > 0. \tag{3.15}$$

By Hölder's inequality and a change of variables, we obtain

$$\begin{aligned} \int_I \psi(u_k')|u_k''|^p\,dx &\geq \frac{1}{\mathcal{L}^1(I)^{p-1}} \left(\int_I (\psi(u_k'))^{1/p}|u_k''|\,dx\right)^p \\ &\geq \frac{1}{(b-a)^{p-1}} \left(\int_{m_k}^{M_k} (\psi(t))^{1/p}\,dt\right)^p, \end{aligned} \tag{3.16}$$

where $m_k := \inf_I u_k'$ and $M_k := \sup_I u_k'$.

Reasoning as in the first step of the proof of Theorem 2.4, we can show that at least one of the two sequences $\{m_k\}$ and $\{M_k\}$ is divergent. If $\lim_k M_k = +\infty$, then by (3.13) $\limsup_k m_k < t_0$ and, in turn, from (3.16) we obtain

$$\liminf_{k\to\infty} \int_I \psi(u_k')|u_k''|^p\,dx \geq \frac{1}{(b-a)^{p-1}} \left(\int_{t_0}^{+\infty} (\psi(t))^{1/p}\,dt\right)^p > 0.$$

Analogously, if $\lim_k m_k = -\infty$, then

$$\liminf_{k\to\infty} \int_I \psi(u_k')|u_k''|^p\,dx \geq \frac{1}{(b-a)^{p-1}} \left(\int_{-\infty}^{t_0} (\psi(t))^{1/p}\,dt\right)^p > 0.$$

In any case, for an arbitrarily large number m of disjoint intervals I satisfying (3.15), adding the contributions of each interval we obtain

$$\begin{aligned} &\liminf_{k\to\infty} \int_a^b \psi(u_k')|u_k''|^p\,dx \\ &\qquad \geq \frac{m}{(b-a)^{p-1}} \min\left\{\left(\int_{t_0}^{+\infty} (\psi(t))^{1/p}\,dt\right)^p, \left(\int_{-\infty}^{t_0} (\psi(t))^{1/p}\,dt\right)^p\right\}, \end{aligned}$$

which contradicts (3.8) for m large enough. This concludes the proof of (3.14) and, in turn, of (3.7).

Step 2. To prove that $u \in X_\psi^p(]a,b[)$, it remains to show that the positive part $((u')^s)^+$ and the negative part $((u')^s)^-$ of the measure $(u')^s$ are concentrated on $Z^+[(u')^a]$ and $Z^-[(u')^a]$, respectively.

Arguing as in Step 2 of the proof of Theorem 2.4, one can see that it is enough to show

$$E^+[u'] \setminus Z^+[(u')^a] \text{ and } E^-[u'] \setminus Z^-[(u')^a] \text{ are empty,} \tag{3.17}$$

where $E^+[u']$ and $E^-[u']$ are the sets introduced in (2.32) and (2.33). We show only that $E^+[u'] \setminus Z^+[(u')^a]$ is empty, since the other property can be proved in the same way.

Assume, by contradiction, that $E^+[u'] \setminus Z^+[(u')^a]$ contains a point x_0. Denote $t_0 := 2|w(x_0)|$, fix any $t_1 > t_0$, and choose $\varepsilon_0 > 0$ such that

$$\frac{1}{(2\varepsilon_0)^{p-1}} \left(\int_{t_0}^{t_1} (\psi(t))^{1/p}\, dt \right)^p > C, \tag{3.18}$$

where C is the constant appearing in (3.5). By (2.32) there exists $0 < \varepsilon < \varepsilon_0$ such that

$$\frac{(u')^+ (]x_0 - \varepsilon, x_0 + \varepsilon[)}{2\varepsilon} > t_1. \tag{3.19}$$

Set $I :=]x_0 - \varepsilon, x_0 + \varepsilon[$. By Hölder's inequality and a change of variables (see (3.16)), we obtain

$$\int_I \psi(u_k')|u_k''|^p\, dx \geq \frac{1}{(2\varepsilon_0)^{p-1}} \left(\int_{m_k}^{M_k} (\psi(t))^{1/p}\, dt \right)^p, \tag{3.20}$$

where $m_k := \inf_I u_k'$ and $M_k := \sup_I u_k'$. By (3.13) and the fact that $w(x_0) < t_0$, we deduce that

$$\limsup_{k\to\infty} m_k < t_0. \tag{3.21}$$

On the other hand, reasoning as at the end of the proof of Theorem 2.4, we deduce from (3.6) and (3.19) that

$$\liminf_{k\to\infty} \frac{1}{2\varepsilon} \int_{x_0-\varepsilon}^{x_0+\varepsilon} (u_k')^+\, dx \geq \frac{(u')^+ (]x_0 - \varepsilon, x_0 + \varepsilon[)}{2\varepsilon} > t_1,$$

which implies that

$$\liminf_{k\to\infty} M_k > t_1. \tag{3.22}$$

From (3.18), (3.20), (3.21), and (3.22) we obtain

$$\liminf_{k\to\infty} \int_I \psi(u_k')|u_k''|^p\, dx \geq \frac{1}{(2\varepsilon_0)^{p-1}} \left(\int_{t_0}^{t_1} (\psi(t))^{1/p}\, dt \right)^p > C,$$

which contradicts (3.8). This shows (3.17) and concludes the proof of the theorem. □

We next identify the relaxation of $\mathcal{F}_p$ with respect to strong convergence in $L^1(]a, b[)$.

THEOREM 3.4. *Let $\overline{\mathcal{F}}_p \colon L^1(]a, b[) \to [0, +\infty]$ be defined by*

$$\overline{\mathcal{F}}_p(u) := \inf \left\{ \liminf_{k\to\infty} \mathcal{F}_p(u_k) : u_k \to u \text{ in } L^1(]a, b[) \right\} \tag{3.23}$$

for every $u \in L^1(]a,b[)$. *Then*

$$\overline{\mathcal{F}}_p(u) = \begin{cases} |u'|(]a,b[) + \displaystyle\int_a^b |v'|^p \, dx & \text{if } u \in X^p_\psi(]a,b[), \\ +\infty & \text{otherwise,} \end{cases} \tag{3.24}$$

where $v := \Psi_p \circ (u')^a$.

Proof. We sketch the proof focusing only on the main changes with respect to the proof of Theorem 2.5. Let $\mathcal{G}_p$ be the functional defined by the right-hand side of (3.24).

We start by showing that

$$\mathcal{G}_p(u) \le \liminf_{k\to\infty} \mathcal{F}_p(u_k) \tag{3.25}$$

whenever $u_k \to u$ in $L^1(]a,b[)$. It is enough to consider sequences $\{u_k\}$ for which the liminf is a limit and has a finite value. Then u_k belongs to $W^{2,1}(]a,b[)$ and (3.5) is satisfied. Setting $v_k := \Psi_p \circ u_k'$, by Theorem 3.3, we have $v_k \rightharpoonup v$ weakly in $W^{1,p}(]a,b[)$. Using the fact that $|v_k'|^p = \psi(u_k')|u_k''|^p$, we deduce that

$$\int_a^b |v'|^p \, dx \le \liminf_{k\to\infty} \int_a^b \psi(u_k')|u_k''|^p \, dx. \tag{3.26}$$

Inequality (3.25) follows now from (3.26) and the lower semicontinuity of the total variation.

We split the proof of the limsup inequality into several steps.

Step 1. Let $u \in X^p_\psi(]a,b[)$ be such that $(u')^s = 0$. We claim that there exists a sequence $\{u_k\}$ in $W^{2,p}(]a,b[)$ such that $u_k \to u$ in $L^1(]a,b[)$ and

$$\limsup_{k\to\infty} \mathcal{F}_p(u_k) \le \mathcal{G}_p(u). \tag{3.27}$$

Define $w_k := ((u')^a \vee -k) \wedge k$. Using the fact that $(u')^a \in W^{1,p}(A_{2k})$, where

$$A_{2k} := \{x \in]a,b[: -2k < (u')^a < 2k\}$$

as observed in Remark 3.2(ii), one sees that $w_k \in W^{1,p}(]a,b[)$. Define

$$u_k(x) := u_+(a) + \int_a^x w_k(y) \, dy.$$

It is easy to see that $u_k \to u$ in $L^1(]a,b[)$ and (3.27) holds.

Step 2. Assume that $u \in X^p_\psi(]a,b[)$, $(u')^c = 0$, and S_u is finite. We claim that there exists a sequence $\{u_k\}$ of functions in $X^p_\psi(]a,b[)$, with $(u_k')^s = 0$, such that $u_k \to u$ in $L^1(]a,b[)$ and

$$\limsup_{k\to\infty} \mathcal{G}_p(u_k) \le \mathcal{G}_p(u). \tag{3.28}$$

Since the construction is local, it is enough to consider the case $S_u = \{x_0\}$ for some $x_0 \in]a,b[$ with $[u](x_0) > 0$. By the properties of $X^p_\psi(]a,b[)$, we can find two sequences $x_k \nearrow x_0$ and $y_k \searrow x_0$ such that

$$u(x_k) \to u_-(x_0), \quad u(y_k) \to u_+(x_0), \text{ and } (u')^a(x_k) = (u')^a(y_k) \to (u')^a(x_0) = +\infty.$$

Consider the affine functions $h_k(x) := u(x_k) + (u')^a(x_k)(x - x_k)$. For every k sufficiently large there exists $z_k \in]x_k, b[$ such that $h_k(z_k) = u_+(x_0)$. Since $(u')^a(x_k) \to +\infty$ and $x_k \to x_0$, we have that $z_k \to x_0$ as $k \to \infty$. Define

$$u_k(x) := \begin{cases} u(x) & \text{if } a < x \le x_k, \\ h_k(x) & \text{if } x_k < x \le z_k, \\ u(x + y_k - z_k) + u_+(x_0) - u(y_k) & \text{if } z_k < x < b. \end{cases}$$

Using the fact that $(u')^a(x_k) = (u')^a(y_k)$, it is easy to check that $u_k \in X^p_\psi(]a, b[)$, with $(u_k')^s = 0$, $u_k \to u$ in $L^1(]a, b[)$, and (3.28) holds.

Step 3. Assume that $u \in X^p_\psi(]a, b[)$ and $(u')^c = 0$. We claim that there exists a sequence of functions u_k in $X^p_\psi(]a, b[)$, with $(u_k')^c = 0$ and S_{u_k} finite, such that $u_k \to u$ in $L^1(]a, b[)$ and (3.28) holds.

To see this, it is enough to consider the same approximation constructed in Step 3 of the proof of Theorem 2.5.

Step 4. Assume that $u \in X^p_\psi(]a, b[)$. We claim that there exists a sequence of functions u_k in $X^p_\psi(]a, b[)$, with $(u_k')^c = 0$, such that $u_k \to u$ in $L^1(]a, b[)$ and (3.28) holds.

Since $(u')^a$ is continuous from $]a, b[$ into $\overline{\mathbb{R}}$ and integrable (see Remark 3.2), we have that $K := \{x \in]a, b[: |(u')^a| = +\infty\}$ is relatively closed in $]a, b[$ with zero $\mathcal{L}^1$ measure. Hence, we may find a sequence of open sets $A_k \subset]a, b[$ such that $A_k \searrow K$. Let $\{I_j^k\}_j$ be the collection of all connected components of A_k intersecting K. Let $c_j^k := (u')^s(I_j^k) > 0$. By the properties of $X^p_\psi(]a, b[)$, for every j we may choose $x_j^k \in I_j^k \cap K$ such that $(u')^a(x_j^k) = +\infty$ if $c_j^k > 0$ and $(u')^a(x_j^k) = -\infty$ if $c_j^k < 0$. Define

$$u_k(x) := u_+(a) + \int_a^x (u')^a(y)\, dy + \sum_{j:\, x_j^k \le x} c_j^k.$$

Using the definition of $X^p_\psi(]a, b[)$, one can check that

$$\sum_j c_j^k \delta_{x_j^k} \rightharpoonup (u')^s \quad \text{weakly}^* \text{ in } M_b(]a, b[)$$

and $|\sum_j c_j^k \delta_{x_j^k}|(]a, b[) \to |(u')^s|(]a, b[)$ as $k \to \infty$. Using this fact it is easy to see that the sequence $\{u_k\}$ meets all of the requirements.

By combining Steps 1–4 with a diagonal argument, one can finally prove that (3.27) holds for every u in $X^p_\psi(]a, b[)$. ∎

Corollary 3.5. *Let $\{u_k\}$ be a sequence of functions in $X^p_\psi(]a, b[)$ bounded in $L^1(]a, b[)$ and such that*

$$C := \sup_k \overline{\mathcal{F}}_p(u_k) < +\infty. \tag{3.29}$$

Then there exists a subsequence (not relabeled) $\{u_k\}$ and a function $u \in X^p_\psi(]a, b[)$ such that

$$u_k \rightharpoonup u \quad \textit{weakly}^* \textit{ in } BV(]a, b[), \tag{3.30}$$

$$\Psi_p \circ (u_k')^a \rightharpoonup \Psi_p \circ (u')^a \quad \textit{weakly in } W^{1,p}(]a, b[), \tag{3.31}$$

$$(u_k')^a \to (u')^a \quad \textit{pointwise in }]a, b[. \tag{3.32}$$

Proof. With an argument entirely similar to the one used in the proof of Corollary 2.7 we can extract a subsequence $\{u_k\}$ which satisfies (3.30) and (3.31). In turn, (3.31) and the continuity of Ψ_p^{-1} in $\overline{\mathbb{R}}$ imply (3.32). □

4. The staircase effect. The purpose of this section is to show analytically that the presence of the higher order term in the functional $\overline{\mathcal{F}}$ prevents the occurrence of the so-called *staircase effect*, as opposed to what happens in image reconstructions based on the total variation functional.

4.1. The Rudin–Osher–Fatemi model. We start by showing that staircase-like structures do appear in solutions to the Rudin–Osher–Fatemi problem, i.e., in minimizers for the functional $\mathrm{ROF}_{\lambda,g} : BV(]a,b[) \to \mathbb{R}$ defined by

$$\mathrm{ROF}_{\lambda,g}(w) := |w'|(]a,b[) + \lambda \int_a^b (w-g)^2\,dx,$$

where $\lambda > 0$ is the *fidelity parameter* and $g \in L^2(]a,b[)$ is the given "signal" to be processed. This fact is well known and numerically observed in many situations. We provide here a simple analytical example. A different example is provided in [12], and a more detailed analysis on the nature of the staircase effect in the Rudin–Osher–Fatemi model can be found in a recent paper of Caselles, Chambolle, and Novaga [4]. Our example will be constructed by means of the following proposition which deals with minimizers of $\mathrm{ROF}_{\lambda,g}$ when g is a monotone function.

PROPOSITION 4.1. *Let $g : [a,b] \to [0,1]$ be a nondecreasing function such that $g_+(a) = 0$ and $g_-(b) = 1$. Let g^{-1} denote the left-continuous generalized inverse of g, defined by*

$$g^{-1}(c) := \inf\{x \in [a,b] :\ g(x) \geq c\} \tag{4.1}$$

for every $c \in [0,1]$, and assume that there exist $0 < c_1 < c_2 < 1$ such that

$$2\lambda \int_a^{g^{-1}(c_1)} (c_1 - g(x))\,dx = 1 \qquad \text{and} \qquad 2\lambda \int_{g^{-1}(c_2)}^b (g(x) - c_2)\,dx = 1. \tag{4.2}$$

Then the function u, defined by

$$u(x) := \begin{cases} c_1 & \text{if } a \leq x \leq g^{-1}(c_1), \\ g(x) & \text{if } g^{-1}(c_1) < x \leq g^{-1}(c_2), \\ c_2 & \text{if } g^{-1}(c_2) < x \leq b, \end{cases}$$

is the unique minimizer of $\mathrm{ROF}_{\lambda,g}$ in $BV(]a,b[)$.

Remark 4.2. Since

$$\int_a^{g^{-1}(c)} (c - g(x))\,dx = \int_0^c g^{-1}(y)\,dy, \qquad \int_{g^{-1}(c)}^b (g(x) - c)\,dx = \int_c^1 g^{-1}(y)\,dy$$

for all $c \in [0,1]$, the continuity of the integral implies that condition (4.2) is satisfied for every λ sufficiently large.

Proof of Proposition 4.1. We split the proof into two steps.

Step 1. We assume first that u is absolutely continuous. In order to prove the minimality of u, by density it suffices to show that $\mathrm{ROF}_{\lambda,g}(u+\varphi) \geq \mathrm{ROF}_{\lambda,g}(u)$ for

every $\varphi \in C^1([a,b])$, which, in turn, due to the convexity of $\mathrm{ROF}_{\lambda,g}$, is equivalent to proving that

$$\frac{d^+}{d\varepsilon}\mathrm{ROF}_{\lambda,g}(u+\varepsilon\varphi)\Big|_{\varepsilon=0} \geq 0 \quad \text{for every } \varphi \in C^1([a,b]), \tag{4.3}$$

where $\frac{d^+}{d\varepsilon}$ denotes the right derivative. By a straightforward computation, we have

$$\frac{d^+}{d\varepsilon}\mathrm{ROF}_{\lambda,g}(u+\varepsilon\varphi)\Big|_{\varepsilon=0} = \int_{\{u'=0\}} |\varphi'|\,dx + \int_{\{u'>0\}} \varphi'\,dx + 2\lambda\int_a^b (u-g)\varphi\,dx. \tag{4.4}$$

Consider now the function $\theta : [a,b] \to [0,1]$ defined by $\theta(x) := 2\lambda\int_a^x (u-g)\,dt$. Using (4.2) and the definition of u, one can check that $\theta(a) = \theta(b) = 0$, $0 \leq \theta \leq 1$, and $\theta \equiv 1$ in $[g^{-1}(c_1), g^{-1}(c_2)]$. In particular, $\{u' > 0\} \subset [g^{-1}(c_1), g^{-1}(c_2)] \subset \{\theta = 1\}$ so that by (4.4)

$$\frac{d^+}{d\varepsilon}\mathrm{ROF}_{\lambda,g}(u+\varepsilon\varphi)\Big|_{\varepsilon=0} \geq \int_a^b \varphi'\theta\,dx + 2\lambda\int_a^b (u-g)\varphi\,dx = 0,$$

where the last equality is obtained by integrating by parts and by using the fact that $\theta' = 2\lambda(u-g)$ and $\theta(a) = \theta(b) = 0$. This shows (4.3) and concludes the proof of Step 1.

Step 2. In the general case, we construct a sequence $\{g_k\} \subset AC([g^{-1}(c_1), g^{-1}(c_2)])$ of nondecreasing functions such that $g_k(g^{-1}(c_1)) = c_1$, $g_k(g^{-1}(c_2)) = c_2$, and $g_k \to g$ in $L^2([g^{-1}(c_1), g^{-1}(c_2)])$. Let $\tilde{g}_k$ be the function coinciding with g_k in $[g^{-1}(c_1), g^{-1}(c_2)]$ and with g elsewhere in $[a,b]$, and, analogously, let u_k coincide with g_k in the interval $[g^{-1}(c_1), g^{-1}(c_2)]$ and with u elsewhere. For any $v \in BV(]a,b[)$, by applying the previous step we obtain

$$\mathrm{ROF}_{\lambda,\tilde{g}_k}(v) \geq \mathrm{ROF}_{\lambda,\tilde{g}_k}(u_k) = \mathrm{ROF}_{\lambda,g}(u).$$

The minimality of u follows by letting $k \to \infty$. Finally, uniqueness is a consequence of the strict convexity of $\mathrm{ROF}_{\lambda,g}$. ∎

As a corollary of the previous result, we can prove analytically the occurrence of the staircase effect in a very simple case. Let $g(x) := x$, $x \in [0,1]$, be the original 1D image to which we add the "noise"

$$h_n(x) := \frac{i}{n} - x \qquad \text{if } \frac{i-1}{n} \leq x < \frac{i}{n}, \ \ i = 1, \ldots, n,$$

where $n \in \mathbb{N}$, so that the resulting degraded 1D image is given by the staircase function

$$g_n(x) := \frac{i}{n} \qquad \text{if } \frac{i-1}{n} \leq x < \frac{i}{n}, \quad i = 1, \ldots, n. \tag{4.5}$$

Note that, even though $h_n \to 0$ uniformly, the reconstructed image u_n preserves the staircase structure of g_n. Indeed, we show that there exists a nondegenerate interval $I \subset [0,1]$ such that each u_n coincides with the degraded 1D image g_n in I for all $n \in \mathbb{N}$. More precisely, we have the following theorem.

THEOREM 4.3 (staircase effect). *Let $\lambda > 4$, let g_n be as in (4.5), and let u_n be the unique minimizer of $\mathrm{ROF}_{\lambda,g_n}$ in $BV(]0,1[)$. Then for all n sufficiently large there exist $0 < a_n < b_n < 1$, with*

$$a_n \to \frac{1}{\sqrt{\lambda}}, \quad b_n \to 1 - \frac{1}{\sqrt{\lambda}}$$

as $n \to \infty$, such that $u_n = g_n$ on $[a_n, b_n]$ and u_n is constant on each interval $[0, a_n)$ and $(b_n, 1]$.

Proof. Let g_n^{-1} denote the generalized inverse function of g_n defined by (4.1) with g replaced by g_n. As both $\{g_n\}$ and $\{g_n^{-1}\}$ converge uniformly to $g(x) = x$ and since $\lambda > 4$, one can check that for n large enough there exist $0 < c_1^{(n)} < c_2^{(n)} < 1$ satisfying

$$2\lambda \int_0^{g_n^{-1}\left(c_1^{(n)}\right)} \left(c_1^{(n)} - g_n\right) dx = 1 \qquad \text{and} \qquad 2\lambda \int_{g_n^{-1}\left(c_2^{(n)}\right)}^1 \left(g_n - c_2^{(n)}\right) dx = 1$$

with $c_1^{(n)} \to c_1$ and $c_2^{(n)} \to c_2$ as $n \to \infty$, where c_1 and c_2 are defined by

$$2\lambda \int_0^{c_1} (c_1 - x)\, dx = 1 \qquad \text{and} \qquad 2\lambda \int_{c_2}^1 (x - c_2)\, dx = 1. \tag{4.6}$$

By Proposition 4.1, the unique minimizer u_n of $\mathrm{ROF}_{\lambda, g_n}$ in $BV(]0,1[)$ takes the form

$$u_n(x) = \begin{cases} c_1^{(n)} & \text{if } 0 \le x \le g_n^{-1}\left(c_1^{(n)}\right), \\ g_n(x) & \text{if } g_n^{-1}\left(c_1^{(n)}\right) < x \le g_n^{-1}\left(c_2^{(n)}\right), \\ c_2^{(n)} & \text{if } g_n^{-1}\left(c_2^{(n)}\right) < x \le 1. \end{cases}$$

The conclusion follows by observing that $a_n := g_n^{-1}(c_1^{(n)}) \to c_1$ and $b_n := g_n^{-1}(c_2^{(n)}) \to c_2$ and that $c_1 = \frac{1}{\sqrt{\lambda}}$ and $c_2 = 1 - \frac{1}{\sqrt{\lambda}}$, thanks to (4.6). ∎

4.2. Absence of the staircase effect: The case $p = 1$. Next, we show that the presence of the higher order term in the functional $\overline{\mathcal{F}}_1$ prevents the occurrence of the staircase effect. We begin with the case $p = 1$. We consider the minimization problem

$$\min\left\{\overline{\mathcal{F}}_1(u) + \lambda \int_a^b (u - g)^2\, dx :\ u \in X^1_\psi(]a,b[)\right\}, \tag{4.7}$$

where $\overline{\mathcal{F}}_1$ is the relaxed functional given in (2.38). To prove the absence of the staircase effect we need the following auxiliary result that is of independent interest.

PROPOSITION 4.4. *Assume that $\psi\colon \mathbb{R} \to]0, +\infty[$ is a bounded Borel function satisfying* (2.1) *and* (2.2). *Let $g : [a,b] \to \mathbb{R}$ be Lipschitz continuous, and let $u \in X^1_\psi(]a,b[)$ be a solution of the minimization problem* (4.7). *Then u is Lipschitz continuous and $u' \in BV(]a,b[)$.*

Proof. The plan of the proof is the following. We will show that the discontinuity set S_u is empty and that the left and right limits $(u')^a_-$ and $(u')^a_+$, respectively, defined in (2.7), are finite everywhere on $]a,b]$ and on $[a,b[$, respectively. Note that this will imply that the sets $Z^{\pm}[(u')^a]$ (see (2.5) and (2.6)) are empty and, in turn, that $u \in W^{1,1}(]a,b[)$ by the properties of the space $X^1_\psi(]a,b[)$. Moreover, recalling that the functions $(u')^a_\vee$ and $(u')^a_\wedge$ defined in Remark 2.2 are upper and lower semicontinuous on $[a,b]$, it will also follow that both $(u')^a_-$ and $(u')^a_+$ are bounded, yielding the Lipschitz continuity of u. In turn, the fact that $u' \in BV(]a,b[)$ is a consequence of the local Lipschitz continuity of Ψ_1^{-1}.

Step 1. We start by showing that S_u is empty. We argue, by contradiction, assuming that S_u contains a point x_0. Without loss of generality, we may suppose that

$\nu_u(x_0) = 1$; i.e., $u_+(x_0) > u_-(x_0)$. We also assume that $\frac{1}{2}(u_+(x_0) + u_-(x_0)) \geq g(x_0)$. In the following, it is convenient to think of u as coinciding everywhere with its lower semicontinuous representative $u_\wedge := \min\{u_-, u_+\}$.

Find $\varepsilon > 0$ so small that

$$\sum_{\substack{x \in S_u \\ x \in]x_0, x_0+\varepsilon[}} |[u](x)| < \frac{[u](x_0)}{4}, \tag{4.8}$$

and let $C > 0$ satisfy

$$C > 2\|g'\|_\infty \quad \text{and} \quad \frac{1}{2}(u_+(x_0) + u_-(x_0)) + C\varepsilon > u_-(x_0 + \varepsilon). \tag{4.9}$$

For $t \in [0, 1]$ consider the affine function

$$h^t(x) := \frac{(1-t)}{2}(u_+(x_0) + u_-(x_0)) + t\left(\frac{1}{4}u_-(x_0) + \frac{3}{4}u_+(x_0)\right) + C(x - x_0)$$

and note that, by (4.9), there exists $x^t \in]x_0, x_0 + \varepsilon[$ such that

$$\left(x^t, h^t\left(x^t\right)\right) \in \Gamma_u \quad \text{and} \quad g < h^t < u \text{ in } \left]x_0, x^t\right[, \tag{4.10}$$

where Γ_u stands for the extended graph of u defined by

$$\Gamma_u := \{(x, t) \in]a, b[\times \mathbb{R} : \min\{u_-(x), u_+(x)\} \leq t \leq \max\{u_-(x), u_+(x)\}\}.$$

Let u^t be the function defined by

$$u^t(x) := \begin{cases} h^t(x) & \text{if } x \in]x_0, x^t[, \\ u(x) & \text{otherwise,} \end{cases} \tag{4.11}$$

and note that

$$\lambda\left(\int_a^b |u - g|^2\, dx - \int_a^b |u^t - g|^2\, dx\right) \geq \lambda\left(\int_a^b |u - g|^2\, dx - \int_a^b |u^1 - g|^2\, dx\right) =: \eta > 0 \tag{4.12}$$

for every $t \in [0, 1]$. Now it is convenient to approximate u with functions having only finitely many jump points. Hence, the following approximation procedure is needed only when S_u is infinite. In this case, write $S_u = \{x_0, x_1, \ldots, x_j, \ldots\}$, for each k define $S_u^k := \{x_j : 0 \leq j \leq k\}$, and for $x \in]a, b[$ set

$$u_k(x) = u_+(a) + \int_a^x (u')^a\, dt + (u')^c(]a, x[) + \sum_{x_j < x,\, x_j \in S_u^k} [u](x_j).$$

Note that, since $u_k \to u$ in $L^\infty(]a, b[)$, for k large enough it follows from (4.9) and (4.10) that for every $t \in [0, 1]$ there exists $x_k^t \in]x_0, x_0 + \varepsilon[$ such that

$$\left(x_k^t, h^t\left(x_k^t\right)\right) \in \Gamma_{u_k} \quad \text{and} \quad g < h^t < u_k \text{ in } \left]x_0, x_k^t\right[,$$

where Γ_{u_k} denotes the extended graph of u_k. For all such k, we consider the comparison function u_k^t defined as in (4.11), with u and x^t replaced by u_k and x_k^t, respectively. Using the uniform convergence of $\{u_k\}$ to u and (4.10), we have that $x^t \le \liminf_k x_k^t$, which yields $u^t \ge \limsup_k u_k^t$ $\mathcal{L}^1$-a.e. on $]a,b[$. Moreover, $u_k \to u$ in $\overline{\mathcal{F}}_1$ energy. Hence, also by (4.12) we may find k so large that for $t \in [0,1]$

$$\lambda\left(\int_a^b |u_k-g|^2\,dx-\int_a^b |u_k^t-g|^2\,dx\right) \ge \lambda\left(\int_a^b |u_k-g|^2\,dx-\int_a^b |u_k^1-g|^2\,dx\right) \ge \frac{\eta}{2}, \tag{4.13}$$

$$\overline{\mathcal{F}}_1(u_k) + \lambda\int_a^b |u_k-g|^2\,dx \le \overline{\mathcal{F}}_1(u) + \lambda\int_a^b |u-g|^2\,dx + \frac{\eta}{4}. \tag{4.14}$$

Let us fix k satisfying (4.13) and (4.14). We claim that there exists $\bar t \in [0,1]$ such that $x_k^{\bar t}$ is a continuity point for u_k. Indeed, if not, then for every $t \in [0,1]$ there exists a jump point x_j, with $1 \le j \le k$, such that $x_k^t = x_j$ and the point $(x_k^t, h^t(x_k^t))$ belongs to the corresponding vertical segment of the extended graph of u_k. Setting $I_j := \{t \in [0,1] : x_k^t = x_j\}$ and $\sigma_j := \{(x_j, h^t(x_j)) : t \in I_j\}$, it is clear that $[0,1] = \cup_{j=1}^k I_j$ and $\mathcal{H}^1(\sigma_j) = \mathcal{H}^1(\{(x_0, h^t(x_0)) : t \in I_j\})$. Thus,

$$\sum_{\substack{x\in S_u\\ x\in]x_0,x_0+\varepsilon[}} |[u](x)| \ge \sum_{j=1}^k \mathcal{H}^1(\sigma_j) = \mathcal{H}^1\big(\{(x_0,h^t(x_0)) : t\in[0,1]\}\big) = \frac{[u](x_0)}{4},$$

in contradiction with (4.8).

Since from now on $\bar t$ and k are fixed, to simplify the notation we set $\hat x := x_k^{\bar t}$, $\hat u := u_k^{\bar t}$, $\hat h := h^{\bar t}$, and $\hat v := \Psi_1 \circ (\hat u')^a$. By construction (see (4.11)) we have

$$|\hat u'|(]a,b[) \le |u_k'|(]a,b[). \tag{4.15}$$

Next, we claim that

$$(u')_-^a(\hat x) \le \hat h'(\hat x) = C. \tag{4.16}$$

If $(u')_-^a(\hat x) \le 0$, there is nothing to prove. If $(u')_-^a(\hat x) > 0$, then by left continuity $(u')_-^a(y) > 0$ for y sufficiently close to $\hat x$, which, in turn, implies $(u')^c(]y,\hat x[) \ge 0$ by the properties of $X_\psi^1(]a,b[)$. Since S_{u_k} is finite and $\hat x$ is a continuity point, for y in a left neighborhood of $\hat x$ we can write

$$\hat h(\hat x) = u_k(\hat x) = u_k(y) + \int_y^{\hat x} (u')^a(s)\,ds + (u')^c(]y,\hat x[) > \hat h(y) + \int_y^{\hat x} (u')^a(s)\,ds,$$

where we have used the fact that $u_k(\hat x) = \hat h(\hat x)$ and $\hat h < u_k$ in a left neighborhood of $\hat x$. Claim (4.16) follows.

Now, recalling that $\Phi(1,t_1,t_2) = 2\Psi_1(+\infty) - \Psi_1(t_1) - \Psi_1(t_2)$ for every t_1, $t_2 \in \overline{\mathbb{R}}$

by (2.39) and using Remark 2.6, we estimate

$$
\begin{aligned}
&|v'|([x_0,\hat{x}]\setminus S_u)+\sum_{x\in S_u\cap[x_0,\hat{x}]}\Phi\left(\nu_u,(u')^a_-,(u')^a_+\right)\\
&\quad\geq |v'|(]x_0,\hat{x}])+\Phi\left(1,(u')^a_-(x_0),(u')^a_+(x_0)\right)\\
&\quad\geq |\Psi_1\left((u')^a_+(x_0)\right)-\Psi_1\left((u')^a_-(\hat{x})\right)|+|\Psi_1\left((u')^a_+(\hat{x})\right)-\Psi_1\left((u')^a_-(\hat{x})\right)|\\
&\qquad+\Phi\left(1,(u')^a_-(x_0),(u')^a_+(x_0)\right)\\
&\quad= |\Psi_1\left((u')^a_+(x_0)\right)-\Psi_1\left((u')^a_-(\hat{x})\right)|+|\Psi_1\left((u')^a_+(\hat{x})\right)-\Psi_1\left((u')^a_-(\hat{x})\right)|\\
&\qquad+2\Psi_1(+\infty)-\Psi_1\left((u')^a_-(x_0)\right)-\Psi_1\left((u')^a_+(x_0)\right)\\
&\quad\geq -\Psi_1\left((u')^a_-(\hat{x})\right)+2\Psi_1(+\infty)-\Psi_1\left((u')^a_-(x_0)\right)\\
&\qquad+|\Psi_1\left((u')^a_+(\hat{x})\right)-\Psi_1\left((u')^a_-(\hat{x})\right)|\\
&\quad= \Psi_1(C)-\Psi_1\left((u')^a_-(\hat{x})\right)+2\Psi_1(+\infty)-\Psi_1\left((u')^a_-(x_0)\right)-\Psi_1(C)\\
&\qquad+|\Psi_1\left((u')^a_+(\hat{x})\right)-\Psi_1\left((u')^a_-(\hat{x})\right)|\\
&\quad\geq |\Psi_1(C)-\Psi_1\left((u')^a_+(\hat{x})\right)|+\Phi\left(1,(\hat{u}')^a_-(x_0),(\hat{u}')^a_+(x_0)\right)\\
&\quad= |\hat{v}'|\left([x_0,\hat{x}]\setminus S_{\hat{u}}\right)+\sum_{x\in S_{\hat{u}}\cap[x_0,\hat{x}]}\Phi\left(\nu_{\hat{u}},(\hat{u}')^a_-,(\hat{u}')^a_+\right),
\end{aligned}
\tag{4.17}
$$

where in the last inequality we have used (4.11) and (4.16). Collecting (4.13), (4.15), and (4.17) we deduce that

$$
\overline{\mathcal{F}}_1(\hat{u})+\lambda\int_a^b|\hat{u}-g|^2+\frac{\eta}{2}\leq\overline{\mathcal{F}}_1(u_k)+\lambda\int_a^b|u_k-g|^2
$$

and, in turn, by (4.14)

$$
\overline{\mathcal{F}}_1(\hat{u})+\lambda\int_a^b|\hat{u}-g|^2\,dx<\overline{\mathcal{F}}_1(u)+\lambda\int_a^b|u-g|^2\,dx,
\tag{4.18}
$$

which contradicts the minimality of u.

If $\frac{1}{2}(u_+(x_0)+u_-(x_0))<g(x_0)$, then we proceed in a similar manner: The comparison function $\hat{u}$ is now constructed by replacing u_k with an affine function (defined as before and with C and t properly chosen) in a left neighborhood of x_0. The argument is completely analogous to the previous one, and we omit the details.

Step 2. We finally show that $(u')^a_-$ and $(u')^a_+$ are finite everywhere in $]a,b]$ and in $[a,b[$, respectively. We give the details only for $(u')^a_-$, since one can argue for $(u')^a_+$ in an entirely similar way.

Recall that, by the previous step, u is continuous. Once again, we reason, by contradiction, by assuming that there exists $\bar{x}\in]a,b]$ such that $|(u')^a_-(\bar{x})|=+\infty$. Without loss of generality, we may suppose that $(u')^a_-(\bar{x})=+\infty$. Using Remark 2.2 and the differentiability properties of BV functions we may choose a point $x_1\in]a,\bar{x}[$ such that u is differentiable at x_1 and

$$
u(x_1)\neq g(x_1),\ u'(x_1)=(u')^a_-(x_1)=(u')^a_+(x_1),\ u'(x_1)>2\|g'\|_\infty,\ |v'|(]a,x_1])>0.
$$

The first condition is a consequence of the fact that g is Lipschitz and u cannot be Lipschitz in any left neighborhood of $\bar{x}$, since $|(u')^a_-(\bar{x})|=+\infty$. The last condition follows easily from the fact that $(u')^a$ cannot be constant $\mathcal{L}^1$-a.e. on $]a,\bar{x}[$. Assume

that $u(x_1) > g(x_1)$. Then, by our choice of x_1 and by the previous step, we can find $\varepsilon \in [0, \frac{1}{2}[$, with

$$\Psi_1(u'(x_1)) - \Psi_1((1-\varepsilon)u'(x_1)) < |v'|([x_1, b[), \tag{4.19}$$

such that the affine function $h(x) := u(x_1) + (1-\varepsilon)u'(x_1)(x - x_1)$ satisfies one of the following conditions: Either there exists a point $x_2 \in]x_1, b[$ for u such that

$$h(x_2) = u(x_2) \text{ and } g < h < u \text{ in }]x_1, x_2[\tag{4.20}$$

or

$$g < h < u \text{ in }]x_1, b[. \tag{4.21}$$

In the latter case, we set $x_2 := b$. We now consider the comparison function

$$\hat{u}(x) := \begin{cases} h(x) & \text{if } x \in]x_1, x_2[, \\ u(x) & \text{otherwise,} \end{cases}$$

and we denote $\hat{v} := \Psi_1 \circ (\hat{u}')^a$. We claim that (4.18) holds, contradicting the minimality of u. By (4.20) and (4.21) in any case we have

$$\lambda \int_a^b |\hat{u} - g|^2 \, dx < \lambda \int_a^b |u - g|^2 \, dx.$$

Moreover, if $x_2 < b$, we have $|\hat{u}'|([x_1, x_2]) = u(x_2) - u(x_1) \le |u'|([x_1, x_2])$, while if $x_2 = b$, we have $|\hat{u}'|([x_1, b[) = u_-(b) - u(x_1) \le |u'|([x_1, b[)$ so that in both cases $|\hat{u}'|([a, b[) \le |u'|(]a, b[)$. Hence, (4.18) will follow if we show that $|\hat{v}'|([x_1, x_2]) \le |v'|([x_1, x_2])$, where $[x_1, x_2]$ is replaced by $[x_1, b[$ if $x_2 = b$. To see this, we first assume that (4.20) holds. Arguing as for (4.16), we deduce $(u')^a_-(x_2) \le h'(x_2) = (1-\varepsilon)u'(x_1)$. Therefore, using the properties of x_1, we have

$$\begin{aligned} |v'|([x_1, x_2]) &= |v'|(]x_1, x_2[) + |v'|(\{x_2\}) \\ &\ge \Psi_1(u'(x_1)) - \Psi_1((u')^a_-(x_2)) + |\Psi_1((u')^a_-(x_2)) - \Psi_1((u')^a_+(x_2))| \\ &= \Psi_1(u'(x_1)) - \Psi_1((1-\varepsilon)u'(x_1)) + \Psi_1((1-\varepsilon)u'(x_1)) - \Psi_1((u')^a_-(x_2)) \\ &\quad + |\Psi_1((u')^a_-(x_2)) - \Psi_1((u')^a_+(x_2))| \\ &\ge \Psi_1(u'(x_1)) - \Psi_1((1-\varepsilon)u'(x_1)) + |\Psi_1((1-\varepsilon)u'(x_1)) - \Psi_1((u')^a_+(x_2))| \\ &= |\hat{v}'|([x_1, x_2]). \end{aligned}$$

If (4.21) holds, then by (4.19) we obtain

$$|v'|([x_1, b[) > \Psi_1(u'(x_1)) - \Psi_1((1-\varepsilon)u'(x_1)) = |\hat{v}'|([x_1, b[).$$

If $u(x_1) < g(x_1)$, we modify the previous argument in the following way. We now choose $\varepsilon \in [0, \frac{1}{2}[$ satisfying (4.19) with $|v'|([x_1, b[)$ replaced by $|v'|(]a, x_1])$ and such that the affine function $h(x)$ defined before satisfies one of the following conditions: Either there exists a point $x_2 \in]a, x_1[$ such that $h(x_2) = u(x_2)$ and $u < h < g$ in $]x_2, x_1[$ or $u < h < g$ in $]a, x_1[$. In the latter case, we set $x_2 := a$. We now consider the comparison function

$$\hat{u}(x) := \begin{cases} h(x) & \text{if } x \in]x_2, x_1[, \\ u(x) & \text{otherwise,} \end{cases}$$

and we proceed exactly as before to show (4.18). ∎

We now turn to the main theorem of this subsection.

THEOREM 4.5. *Assume that $\psi\colon \mathbb{R} \to]0,+\infty[$ is a bounded Borel function satisfying* (2.1) *and* (2.2)*, let $g : [a,b] \to \mathbb{R}$ be Lipschitz continuous, and let $\{h_n\}$ satisfy*

$$h_n \rightharpoonup 0 \quad \text{weakly}^* \text{ in } L^\infty(]a,b[)\,. \tag{4.22}$$

Define $\mathcal{A}_n$ as the class of all solutions to (4.7)*, with g replaced by $g_n := g + h_n$. Then for n large enough every solution $u_n \in \mathcal{A}_n$ is Lipschitz continuous. Moreover,*

$$\limsup_{n\to\infty} \sup_{w\in\mathcal{A}_n} \|w\|_{1,\infty} < +\infty \tag{4.23}$$

and for every sequence $\{u_n\} \subset \mathcal{A}_n$ there exists a subsequence (not relabeled) and a solution u to (4.7) *such that $u_n \to u$ in $W^{1,p}(]a,b[)$ for all $p \in [1,+\infty[$.*

Proof. It will be enough to prove that for any (sub)sequence $\{u_n\} \subset \mathcal{A}_n$ we may extract a further subsequence (not relabeled) and find a solution u to (4.7) such that u_n is Lipschitz continuous for n large enough,

$$\limsup_{n\to\infty} \|u_n\|_{1,\infty} < +\infty, \tag{4.24}$$

and $u_n \to u$ in $W^{1,p}(]a,b[)$ for all $p \in [1,+\infty[$. Since the sequence h_n is bounded in $L^\infty(]a,b[)$, for any $w \in X^1_\psi(]a,b[)$ we have

$$\sup_n \left(\overline{\mathcal{F}}_1(u_n) + \lambda \int_a^b (u_n - g_n)^2\,dx \right) \le \overline{\mathcal{F}}_1(w) + \lambda \int_a^b (w - g_n)^2\,dx \le C < \infty$$

for a suitable constant $C > 0$ independent of n. By Corollary 2.7 there exist a subsequence not relabeled and a function $u \in X^1_\psi(]a,b[)$ such that

$$u_n \rightharpoonup u \quad \text{weakly}^* \text{ in } BV(]a,b[) \tag{4.25}$$

and

$$u_n' \to (u')^a \quad \text{pointwise } \mathcal{L}^1\text{-a.e. in }]a,b[. \tag{4.26}$$

Moreover, since also the functions h_n^2 are equibounded, upon extracting a further subsequence we may find $f \in L^\infty(]a,b[)$ such that

$$h_n^2 \rightharpoonup f \quad \text{weakly}^* \text{ in } L^\infty(]a,b[). \tag{4.27}$$

It is convenient to "localize" the functional $\overline{\mathcal{F}}_1$: For every Borel set $B \subset]a,b[$ and for $w \in X^1_\psi(]a,b[)$ we set

$$\overline{\mathcal{F}}_1(w;B) := |w'|(B) + |v'|(B \setminus S_w) + \sum_{x\in S_w\cap B} \Phi\left(\nu_w, (w')^a_-, (w')^a_+\right), \tag{4.28}$$

where $v := \Psi_1 \circ (w')^a$. We divide the remaining part of the proof into two steps.

Step 1. We claim that u is a solution of the minimization problem (4.7) and that, for every open interval $I =]c,d[$, with $a \le c < d \le b$ and $c,\, d \in [a,b] \setminus S_{(u')^a}$,

$$\lim_{n\to\infty} \overline{\mathcal{F}}_1(u_n; I) = \overline{\mathcal{F}}_1(u; I)\,. \tag{4.29}$$

To see this, note that for each $n \in \mathbb{N}$

$$\lambda \int_I (u_n - g_n)^2 \, dx = \lambda \int_I (u_n - g)^2 \, dx - 2\lambda \int_I (u_n - g)\, h_n \, dx + \lambda \int_I h_n^2 \, dx.$$

By (4.22), (4.25), and (4.27) it follows that

$$\lim_{n\to\infty} \int_I (u_n - g_n)^2 \, dx = \int_I (u - g)^2 \, dx + \int_I f \, dx. \tag{4.30}$$

Recall, also, that by lower semicontinuity

$$\liminf_{n\to\infty} \overline{\mathcal{F}}_1 (u_n; A) \geq \overline{\mathcal{F}}_1 (u; A) \tag{4.31}$$

for every open set $A \subset]a, b[$.

By the minimality of u_n, for every $w \in X^1_\psi(]a, b[)$ we have

$$\begin{aligned} \overline{\mathcal{F}}_1 (w) &+ \lambda \int_a^b (w - g)^2 \, dx - 2\lambda \int_a^b (w - g)\, h_n \, dx + \lambda \int_a^b h_n^2 \, dx \\ &= \overline{\mathcal{F}}_1 (w) + \lambda \int_a^b (w - g_n)^2 \, dx \geq \overline{\mathcal{F}}_1 (u_n) + \lambda \int_a^b (u_n - g_n)^2 \, dx. \end{aligned}$$

Using (4.31) (with $A =]a, b[$) and once again (4.22) and (4.27), we get

$$\begin{aligned} \overline{\mathcal{F}}_1 (w) + \lambda \int_a^b (w - g)^2 \, dx + \lambda \int_a^b f \, dx &\geq \limsup_{n\to\infty} \left(\overline{\mathcal{F}}_1 (u_n) + \lambda \int_a^b (u_n - g_n)^2 \, dx \right) \\ &\geq \liminf_{n\to\infty} \left(\overline{\mathcal{F}}_1 (u_n) + \lambda \int_a^b (u_n - g_n)^2 \, dx \right) \\ &\geq \overline{\mathcal{F}}_1 (u) + \lambda \int_a^b (u - g)^2 \, dx + \lambda \int_a^b f \, dx. \end{aligned}$$

Given the arbitrariness of $w \in X^1_\psi(]a, b[)$ this implies that u is a solution of the minimization problem (4.7). Moreover, taking $w = u$ in the previous inequalities and using (4.30) we deduce (4.29) for $I =]a, b[$; i.e.,

$$\lim_{n\to\infty} \overline{\mathcal{F}}_1(u_n) = \overline{\mathcal{F}}_1(u). \tag{4.32}$$

It remains to prove (4.29) for every open interval of the form $I =]c, d[$, with c, $d \in [a, b] \setminus S_{(u')^a}$. To this end, fix one such interval and assume, by contradiction, that

$$\limsup_{n\to\infty} \overline{\mathcal{F}}_1 (u_n; I) > \overline{\mathcal{F}}_1 (u; I)\,. \tag{4.33}$$

As u is continuous by Proposition 4.4, our assumption on I implies that the end points c and d do not charge $\overline{\mathcal{F}}_1 (u; \cdot)$ so that $\overline{\mathcal{F}}_1 (u; I) = \overline{\mathcal{F}}_1 \left(u; \overline{I} \cap]a, b[\right)$. Therefore, combining (4.31), (4.32), and (4.33) we obtain

$$\begin{aligned} \overline{\mathcal{F}}_1(u) &= \overline{\mathcal{F}}_1 \left(u; \overline{I} \cap]a, b[\right) + \overline{\mathcal{F}}_1 \left(u;]a, b[\setminus \overline{I}\right) = \overline{\mathcal{F}}_1 (u; I) + \overline{\mathcal{F}}_1 \left(u;]a, b[\setminus \overline{I}\right) \\ &< \limsup_{n\to\infty} \overline{\mathcal{F}}_1 (u_n; I) + \liminf_{n\to\infty} \overline{\mathcal{F}}_1 \left(u_n;]a, b[\setminus \overline{I}\right) \leq \lim_{n\to\infty} \overline{\mathcal{F}}_1(u_n) = \overline{\mathcal{F}}_1(u), \end{aligned}$$

which is a contradiction. This concludes the proof of (4.29).

Step 2. We now show that u_n is Lipschitz continuous for n large enough and that (4.24) holds. Note that the convergence of u_n to u in $W^{1,p}(]a,b[)$ for all $p \in [1,+\infty[$ will then easily follow from (4.24) and (4.26). Assume, by contradiction, that the conclusion is false. Then, arguing as at the beginning of the proof of Proposition 4.4, we may find a subsequence (not relabeled) and points $x_n \in]a,b[$ such that one of the following two cases holds:

(i) $x_n \notin S_{(u_n')^a}$ and $|(u_n')^a(x_n)| \to +\infty$;

(ii) $x_n \in S_{u_n}$ for every $n \in \mathbb{N}$.

Assume that (i) holds and, without loss of generality, that $(u_n')^a(x_n) \to +\infty$. Upon extracting a further subsequence, we may also assume that $x_n \to x_0 \in [a,b]$. Recall that, by Proposition 4.4 and by the previous step, the function u is Lipschitz continuous. Hence, there are two cases: Either

$$\overline{\mathcal{F}}_1(u;\{x_0\}\cap]a,b[) = 0 \tag{4.34}$$

or

$$x_0 \in S_{(u')^a},\ (u')^a_{\pm}(x_0) \in \mathbb{R},\ \overline{\mathcal{F}}_1(u;\{x_0\}) = |\Psi_1((u')^a_+(x_0)) - \Psi_1((u')^a_-(x_0))|. \tag{4.35}$$

Assume first that (4.34) holds. Set $L := \|u'\|_\infty$ and fix ε so small that

$$\overline{\mathcal{F}}_1(u;I_\varepsilon) < \int_{L+1}^{+\infty} \psi(t)\,dt,$$

where $I_\varepsilon :=]x_0-\varepsilon, x_0+\varepsilon[\cap]a,b[$. By (4.29) we also have

$$\overline{\mathcal{F}}_1(u_n;I_\varepsilon) < \int_{L+1}^{+\infty} \psi(t)\,dt \tag{4.36}$$

for n large enough. On the other hand, by (4.26) there exists $y \in I_\varepsilon$ such that $(u_n')^a(y) < L+1$ for n large. Moreover, taking into account (i), we also have $(u_n')^a(x_n) > L+1$ for n large enough. Thus,

$$\overline{\mathcal{F}}_1(u_n;I_\varepsilon) \geq |v_n'|(I_\varepsilon) \geq |\Psi_1((u_n')^a(x_n)) - \Psi_1((u_n')^a(y))| \geq \Psi_1((u_n')^a(x_n)) - \Psi_1(L+1).$$

Passing to the limit as $n \to \infty$, we then obtain

$$\liminf_{n\to\infty} \overline{\mathcal{F}}_1(u_n;I_\varepsilon) \geq \Psi_1(+\infty) - \Psi_1(L+1) = \int_{L+1}^{+\infty} \psi(t)\,dt,$$

which contradicts (4.36).

In case (4.35) holds, then $x_0 \in]a,b[$. Set

$$\eta := 2\Psi_1(+\infty) - \Psi_1((u')^a_+(x_0)) - \Psi_1((u')^a_-(x_0)) - |\Psi_1((u')^a_+(x_0)) - \Psi_1((u')^a_-(x_0))| > 0, \tag{4.37}$$

and choose ε such that both $x_0-\varepsilon$ and $x_0+\varepsilon$ belong to $]a,b[\setminus S_{(u')^a}$ and

$$\overline{\mathcal{F}}_1(u;I_\varepsilon) < |\Psi_1((u')^a_+(x_0)) - \Psi_1((u')^a_-(x_0))| + \frac{\eta}{3}, \tag{4.38}$$

$$|\Psi_1((u')^a_\pm(y)) - \Psi_1((u')^a_\pm(x_0))| < \frac{\eta}{4} \quad \text{for } y \in I^\pm_\varepsilon, \tag{4.39}$$

where $I_\varepsilon :=]x_0-\varepsilon, x_0+\varepsilon[$, $I_\varepsilon^+ :=]x_0, x_0+\varepsilon[$, and $I_\varepsilon^- :=]x_0-\varepsilon, x_0[$. Note that by (4.29) and (4.38) we have

$$\overline{\mathcal{F}}_1(u_n; I_\varepsilon) < |\Psi_1((u')_+^a(x_0)) - \Psi_1((u')_-^a(x_0))| + \frac{\eta}{3} \tag{4.40}$$

for n large enough. Moreover, by (4.26) and (4.39) we may find y^-, $y^+ \in I_\varepsilon$, with $y^- < x_0 < y^+$, such that

$$y^\pm \notin S_{(u_n')^a} \quad \text{and} \quad \left|\Psi_1\left((u_n')^a\left(y^\pm\right)\right) - \Psi_1\left((u')_\pm^a(x_0)\right)\right| < \frac{\eta}{4} \tag{4.41}$$

for n large enough. As $y^- < x_n < y^+$ for n sufficiently large, we have

$$\begin{aligned}\overline{\mathcal{F}}_1(u_n; I_\varepsilon) &\geq |v_n'|(I_\varepsilon) \geq |\Psi_1((u_n')^a(x_n)) - \Psi_1((u_n')^a(y^-))| \\ &\quad + |\Psi_1((u_n')^a(x_n)) - \Psi_1((u_n')^a(y^+))| \\ &\geq |\Psi_1((u_n')^a(x_n)) - \Psi_1((u')_-^a(x_0))| + |\Psi_1((u_n')^a(x_n)) - \Psi_1((u')_+^a(x_0))| - \frac{\eta}{2},\end{aligned} \tag{4.42}$$

where the last inequality follows from (4.41). Letting $n \to \infty$ in (4.42) and recalling (4.37) we deduce

$$\begin{aligned}\liminf_{n\to\infty} \overline{\mathcal{F}}_1(u_n; I_\varepsilon) &\geq 2\Psi_1(+\infty) - \Psi_1((u')_+^a(x_0)) - \Psi_1((u')_-^a(x_0)) - \frac{\eta}{2} \\ &= |\Psi_1((u')_+^a(x_0)) - \Psi_1((u')_-^a(x_0))| + \frac{\eta}{2},\end{aligned}$$

which contradicts (4.40). This concludes the proof of (4.24) if (i) holds.

Assume now that (ii) holds, and let $x_0 \in [a,b]$ be the limit of a subsequence (not relabeled) of $\{x_n\}$. Passing to a further subsequence, we may also assume that $\nu_{u_n}(x_n)$ is constant, say, 1 (the other case is analogous). Again, either (4.34) or (4.35) holds. If (4.34) holds, we may choose, as before, ε so small that (4.36) holds. On the other hand, by (4.26) there exists $y \in I_\varepsilon$ such that $(u_n')^a(y) < L+1$ for n large, and, thus, taking into account (ii), we have

$$\begin{aligned}\overline{\mathcal{F}}_1(u_n; I_\varepsilon) &\geq |v_n'|(I_\varepsilon \setminus \{x_n\}) + \Phi(1, (u_n')_-^a(x_n), (u_n')_+^a(x_n)) \\ &\geq |\Psi_1((u_n')^a(y)) - \Psi_1(+\infty)| = \int_{(u_n')^a(y)}^{+\infty} \psi(t)\,dt > \int_{L+1}^{+\infty} \psi(t)\,dt,\end{aligned}$$

which contradicts (4.36). If (4.35) holds, then we may argue exactly as for case (i), with the only difference that (4.42) must be replaced by

$$\begin{aligned}\overline{\mathcal{F}}_1(u_n; I_\varepsilon) &\geq |v_n'|(I_\varepsilon \setminus \{x_n\}) + \Phi(1, (u_n')_-^a(x_n), (u_n')_+^a(x_n)) \\ &\geq |\Psi_1((u_n')_-^a(x_n)) - \Psi_1((u_n')^a(y^-))| + \left|\Psi_1((u_n')_+^a(x_n)) - \Psi_1\left((u_n')^a\left(y^+\right)\right)\right| \\ &\quad + \Phi(1, (u_n')_-^a(x_n), (u_n')_+^a(x_n)) \\ &\geq 2\Psi_1(+\infty) - \Psi_1((u')_-^a(x_0)) - \Psi_1((u')_+^a(x_0)) - \frac{\eta}{2} \\ &= |\Psi_1((u')_+^a(x_0)) - \Psi_1((u')_-^a(x_0))| + \frac{\eta}{2},\end{aligned}$$

where we have used (4.37) and (4.41). The last chain of inequalities contradicts (4.40) and concludes the proof of the theorem. ∎

4.3. Absence of the staircase effect: The case $p > 1$. We now turn to the case $p > 1$. We consider the minimization problem

$$\min\left\{\overline{\mathcal{F}}_p(u) + \lambda\int_a^b |u-g|^2\,dx :\ u \in X^p_\psi(]a,b[)\right\}, \tag{4.43}$$

where $\overline{\mathcal{F}}_p$ is the relaxed functional given in (3.24). We start with two auxiliary results.

PROPOSITION 4.6. *Let $p > 1$, and assume that $\psi\colon \mathbb{R} \to]0,+\infty[$ is a bounded Borel function satisfying* (2.2) *and* (3.1). *Let g be Lipschitz continuous, and let u_n be a sequence in $X^p_\psi(]a,b[)$ such that $\sup_n \overline{\mathcal{F}}_p(u_n) < +\infty$ and $u_n \to g$ in $L^2(]a,b[)$. Then $g \in C^1([a,b]) \cap X^p_\psi(]a,b[)$. Moreover, $u_n \in C^1([a,b])$ for n large enough and $u_n \to g$ in $C^1([a,b])$.*

Proof. By the assumptions and by Corollary 3.5 we deduce that $g \in X^p_\psi(]a,b[)$. The fact that $g \in C^1([a,b])$ now follows from Remark 3.2(i). To prove the last part of the statement we start by showing that $(u_n')^a \to g'$ uniformly in $]a,b[$. Again, by Corollary 3.5, the whole sequence u_n satisfies

$$\Psi_p \circ (u_n')^a \rightharpoonup \Psi_p \circ g' \quad \text{weakly in } W^{1,p}(]a,b[), \tag{4.44}$$

which implies, in particular, that

$$(\Psi_p \circ (u_n')^a)([a,b]) \subset [\Psi_p(-2\|g'\|_\infty), \Psi_p(2\|g'\|_\infty)] \quad \text{for } n \text{ large enough.} \tag{4.45}$$

Since by (2.2) Ψ_p^{-1} is Lipschitz continuous on $[\Psi_p(-2\|g'\|_\infty), \Psi_p(2\|g'\|_\infty)]$, it follows from (4.44) and (4.45) that $(u_n')^a \to g'$ uniformly in $]a,b[$. In turn, by Definition 3.1 we have that $u_n' = (u_n')^a$ in $]a,b[$. In particular, $u_n \in C^1([a,b])$ by Remark 3.2(i) and $u_n \to g$ in $C^1([a,b])$. ∎

PROPOSITION 4.7. *Let p and ψ be as in the previous proposition. Then, for every $C > 0$, there exists $\overline{\lambda} = \overline{\lambda}(C)$ with the following property: For all $g \in C^1([a,b]) \cap X^p_\psi(]a,b[)$, with $\|g\|_{C^1([a,b])} \le C$ and $\overline{\mathcal{F}}_p(g) \le C$, and for all $\lambda \ge \overline{\lambda}$, every solution u to* (4.43) *belongs to $C^1([a,b])$.*

Proof. Assume, by contradiction, that for every $n \in \mathbb{N}$ there exist $g_n \in C^1([a,b]) \cap X^p_\psi(]a,b[)$, with $\|g_n'\|_\infty \le C$ and $\overline{\mathcal{F}}_p(g_n) \le C$, and a solution u_n to

$$\min\left\{\overline{\mathcal{F}}_p(u) + n\int_a^b |u-g_n|^2\,dx :\ u \in X^p_\psi(]a,b[)\right\}$$

which does not belong to $C^1([a,b])$. Owing to Proposition 4.6 we may assume, without loss of generality, that $g_n \to g$ in $C^1([a,b])$ for a suitable function $g \in C^1([a,b]) \cap X^p_\psi(]a,b[)$. Moreover, by minimality, we have

$$\overline{\mathcal{F}}_p(u_n) + n\int_a^b |u_n - g_n|^2\,dx \le \overline{\mathcal{F}}_p(g_n) \le C.$$

It follows, in particular, that $\sup_n \overline{\mathcal{F}}_p(u_n) < +\infty$ and $u_n \to g$ in $L^2(]a,b[)$. By Proposition 4.6 we conclude that $u_n \in C^1([a,b])$ for n large enough, which gives a contradiction. ∎

The next theorem shows that also in the case $p > 1$ the staircase effect does not occur.

THEOREM 4.8. *Let ψ and p be as in Proposition* 4.6, *let $g \in C^1([a,b]) \cap X^p_\psi(]a,b[)$, and let h_n satisfy* (4.22). *For $\lambda > 0$ and $n \in \mathbb{N}$ let $\mathcal{A}_{\lambda,n} \subset X^p_\psi(]a,b[)$ be the class of the solutions to the minimization problem* (4.43), *with g replaced by $g_n := g + h_n$. Let $\overline{\lambda}$ be as in Proposition* 4.7, *with $C := \max\{\|g\|_{C^1([a,b])}, \overline{\mathcal{F}}_p(g)\}$. Then for all $\lambda \geq \overline{\lambda}$ we have $\mathcal{A}_{\lambda,n} \subset C^1([a,b])$ for n sufficiently large. Moreover,*

$$\lim_{\lambda\to\infty} \limsup_{n\to\infty} \sup_{u\in\mathcal{A}_{\lambda,n}} \|u - g\|_{C^1([a,b])} = 0. \tag{4.46}$$

Proof. We start by showing the second part of the statement. Assume, by contradiction, that (4.46) does not hold. Then there exist $\delta > 0$, a sequence of real numbers $\lambda_k \to +\infty$, and, for every k, a sequence of integers $n^k_j \to \infty$ as $j \to \infty$ such that for every k, j

$$\|u_{\lambda_k,n^k_j} - g\|_{C^1([a,b])} \geq \delta \tag{4.47}$$

for a suitable function $u_{\lambda_k,n^k_j} \in \mathcal{A}_{\lambda_k,n^k_j}$, with the understanding that

$$\|u_{\lambda_k,n^k_j} - g\|_{C^1([a,b])} = +\infty \text{ if } u_{\lambda_k,n^k_j} \notin C^1([a,b]).$$

Arguing exactly as in Step 1 of the proof of Theorem 4.5 we can show that for every k there exist a subsequence (still denoted by n^k_j) and a solution u_k to (4.43), with λ replaced by λ_k, such that

$$u_{\lambda_k,n^k_j} \rightharpoonup u_k \quad \text{weakly}^* \text{ in } BV(]a,b[) \qquad \text{and} \qquad \overline{\mathcal{F}}_p(u_{\lambda_k,n^k_j}) \to \overline{\mathcal{F}}_p(u_k) \tag{4.48}$$

as $j \to \infty$. Moreover, since $g \in C^1([a,b]) \cap X^p_\psi(]a,b[)$, we have, by minimality, that

$$\overline{\mathcal{F}}_p(u_k) + \lambda_k \int_a^b |u_k - g|^2\, dx \leq \overline{\mathcal{F}}_p(g), \tag{4.49}$$

which shows, in particular, that $u_k \to g$ in $L^2(]a,b[)$. Combining (4.48) and (4.49) and using a diagonal argument, we may find a subsequence $n^k_{j_k}$ such that

$$\sup_k \overline{\mathcal{F}}_p(u_{\lambda_k,n^k_{j_k}}) < +\infty \qquad \text{and} \qquad u_{\lambda_k,n^k_{j_k}} \to g \text{ in } L^2(]a,b[).$$

Proposition 4.6 then implies that $u_{\lambda_k,n^k_{j_k}} \to g$ in $C^1([a,b])$, which contradicts (4.47).

Finally, the first part of the statement follows from a similar argument, by contradiction, as a consequence of Propositions 4.6 and 4.7 and from the fact that if $u_n \in \mathcal{A}_{\lambda,n}$ then, up to subsequences, u_n converges to a solution of (4.43). ∎

Acknowledgments. The authors thank the Center for Nonlinear Analysis (NSF grants DMS-0405343 and DMS-0635983) for its support during the preparation of this paper. The research of G. Dal Maso and M. Morini was partially supported by the projects "Calculus of Variations" 2004 and "Problemi di Calcolo delle Variazioni in Meccanica e in Scienza dei Materiali" 2006–2008, supported by the Italian Ministry of Education, University, and Research, and by the project "Variational Problems with Multiple Scales" 2006, supported by the Italian Ministry of University and Research. The research of I. Fonseca was partially supported by the NSF under grant DMS-040171 and that of G. Leoni under grants DMS-0405423 and DMS-0708039.

REFERENCES

[1] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford Math. Monogr., Clarendon Press, Oxford University Press, New York, 2000.

[2] P. Blomgren, T. F. Chan, and P. Mulet, *Extensions to total variation denoising*, in Proceedings of the SPIE 1997, San Diego, International Society for Optical Engineering, 1997.

[3] A. Braides and R. March, *Approximation by* Γ*-convergence of a curvature-depending functional in visual reconstruction*, Commun. Pure Appl. Math., 59 (2006), pp. 71–121.

[4] V. Caselles, A. Chambolle, and M. Novaga, *The discontinuity set of solutions of the TV denoising problem and some extensions*, Multiscale Model. Simul., 6 (2007), pp. 879–894.

[5] A. Chambolle and P. L. Lions, *Image recovery via total variation minimization and related problems*, Numer. Math., 76 (1997), pp. 167–188.

[6] T. Chan, A. Marquina, and P. Mulet, *High-order total variation-based image restoration*, SIAM J. Sci. Comput., 22 (2000), pp. 503–516.

[7] G. Dal Maso, *An Introduction to* Γ*-Convergence*, Birkhäuser, Boston, 1993.

[8] E. DiBenedetto, *Real Analysis*, Birkhäuser, Boston, 2002.

[9] D. Geman and G. Reynolds, *Constrained restoration and the recovery of discontinuities*, IEEE Trans. on Pattern Anal. and Mach. Intell., 14 (1992), pp. 367–383.

[10] E. Hewitt and K. Stromberg, *Real and Abstract Analysis. A Modern Treatment of the Theory of Functions of a Real Variable*, Springer-Verlag, New York, Heidelberg, 1975.

[11] S. Kindermann, S. Osher, and P. W. Jones, *Deblurring and denoising of images by nonlocal functionals*, Multiscale Model. Simul., 4 (2005), pp. 1091–1115.

[12] W. Ring, *Stuctural properties of solutions to total variation regularization problems*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 799–810.

[13] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.

[14] W. Rudin, *Real and Complex Analysis*, 2nd ed., McGraw–Hill, New York, Düsseldorf, Johannesburg, 1974.

[15] W. Zhu, T. Chan, and S. Esedoglu, *Segmentation with depth: A level set approach*, SIAM J. Sci. Comput., 28 (2006), pp. 1957–1973.

SIAM J. MATH. ANAL.
Vol. 40, No. 6, pp. 2392–2420

ENTIRE SOLUTIONS IN DELAYED LATTICE DIFFERENTIAL EQUATIONS WITH MONOSTABLE NONLINEARITY*

ZHI-CHENG WANG†, WAN-TONG LI‡, AND JIANHONG WU§

Abstract. We construct new types of entire solutions for a class of monostable delayed lattice differential equations with global interaction by mixing a heteroclinic orbit of the spatially averaged ordinary differential equations with traveling wave fronts with different speeds. We also establish the uniqueness of entire solutions and the continuous dependence of such an entire solution on parameters, such as wave speeds, for the spatially discrete Fisher-KPP equation.

Key words. entire solution, traveling wave front, heteroclinic orbit, delayed lattice differential equation, monostable nonlinearity

AMS subject classifications. 35B40, 35R10, 37L60, 58D25

DOI. 10.1137/080727312

1. Introduction.

We consider the following delayed lattice differential equations:

$$u_n'(t) = D \sum_{i\in\mathbb{Z}\setminus\{0\}} I(i)\left[u_{n-i}(t) - u_n(t)\right] - du_n(t) + \sum_{i\in\mathbb{Z}} J(i)\, b\left(u_{n-i}(t-\tau)\right), \tag{1.1}$$

where $D > 0$ is a given constant, $\tau \geq 0$, $I(i) = I(-i) \geq 0$, $J(i) = J(-i) \geq 0$, $\sum_{i\in\mathbb{Z}\setminus\{0\}} I(i) = 1$, $\sum_{i\in\mathbb{Z}} J(i) = 1$, $\sum_{i\in\mathbb{Z}\setminus\{0\}} e^{\lambda|i|} I(i) < \infty$, and $\sum_{i\in\mathbb{Z}} e^{\lambda|i|} J(i) < \infty$ for every $\lambda \geq 0$. The birth function $b \in C^2(\mathbb{R})$, and we assume that there exists a constant $K > 0$ such that

$$b(0) = dK - b(K) = 0$$

and that

(H1) for $u \in (0, K)$, there hold $b(u) > du$, $b'(u) \geq 0$, and $b(u) \leq b'(0)u$;

(H2) $b'(K) < d < b'(0)$.

A specific function $b(u) = pue^{-\alpha u}$ with $p > 0$ and $\alpha > 0$, which has been widely used in the mathematical biology literature, satisfies the above conditions for a wide range of parameters p and α.

A special case when $I(i) = 0$ for $|i| \neq 1$ and $I(1) = \frac{1}{2}$ is

$$u_n' = \frac{D}{2}\left[u_{n+1} + u_{n-1} - 2u_n\right] - du_n + \sum_{i\in\mathbb{Z}} J(i)\, b\left(u_{n-i}(t-\tau)\right). \tag{1.2}$$

*Received by the editors June 14, 2008; accepted for publication (in revised form) October 24, 2008; published electronically February 25, 2009.

http://www.siam.org/journals/sima/40-6/72731.html

†School of Mathematics and Statistics, Lanzhou University, Lanzhou, Gansu 730000, People's Republic of China (wangzhch@lzu.edu.cn); Department of Mathematics and Statistics, York University, Toronto, Ontario, M3J 1P3, Canada. This work was partially supported by the NSF of Gansu Province of China (0710RJZA020) and The Fundamental Research Fund for Physics and Mathematics of Lanzhou University (LZULL200807).

‡School of Mathematics and Statistics, Lanzhou University, Lanzhou, Gansu, 730000, People's Republic of China (wtli@lzu.edu.cn). This work was partially supported by the NSFC (10871085).

§Laboratory for Industrial and Applied Mathematics, Department of Mathematics and Statistics, York University, Toronto, Ontario, M3J 1P3, Canada (wujh@mathstat.yorku.ca). This work was partially supported by Canada Research Chairs Program, by Natural Sciences and Engineering Research Council of Canada, and by Mathematics for Information Technology and Complex Systems.

This system was derived by Weng, Huang, and Wu [40] for the dynamics of growth of a single species population with two age classes distributed over a patchy environment consisting of all integer nodes of a one-dimensional (1-D) lattice. Another special case when $\tau = 0$, $J(0) = 1$, and $J(i) = 0$ for $|i| > 0$ is

$$u_n'(t) = D \sum_{i \in \mathbb{Z} \setminus \{0\}} I(i) [u_{n-i}(t) - u_n(t)] + f(u_n(t)), \tag{1.3}$$

which was derived by Bates and Chmaj [1] as an l_2-gradient flow for a Helmholtz-free energy functional with general long range interactions. Both lattice systems include, as a special example, the following spatially discrete Fisher-KPP equation:

$$u_n'(t) = \frac{D}{2} [u_{n+1} + u_{n-1} - 2u_n] + f(u_n(t)). \tag{1.4}$$

It is shown in [25] that (1.1) admits a nondecreasing traveling wave front $\phi_c(n+ct)$ satisfying $\phi_c(-\infty) = 0$ and $\phi_c(+\infty) = K$ for every $c \geq c^* > 0$. Furthermore, $\lim_{\xi \to -\infty} \phi_c(\xi) e^{-\lambda_1(c)\xi} = 1$ and $\lim_{\xi \to -\infty} \phi_c'(\xi) e^{-\lambda_1(c)\xi} = \lambda_1(c)$ for $c > c^*$, where c and $\lambda_1(c)$ satisfy

$$\Delta(\lambda, c) = D \sum_{i \in \mathbb{Z} \setminus \{0\}} I(i) \left(e^{-\lambda i} - 1\right) - c\lambda - d + b'(0) e^{-\lambda c\tau} \sum_{i \in \mathbb{Z}} J(i) e^{-\lambda i} = 0 \tag{1.5}$$

and c^* is determined by $\Delta(\lambda, c) = 0$ and $\frac{\partial}{\partial \lambda} \Delta(\lambda, c) = 0$. More precisely, there exist $c^* > 0$ and $\lambda^* > 0$ such that

(D1) if $0 < c < c^*$ and $\lambda > 0$, then $\Delta(\lambda, c) > 0$;

(D2) if $c = c^*$, then the equation $\Delta(\lambda, c^*) = 0$ has a double real root $\lambda_1(c^*) = \lambda_2(c^*)$ with $0 < \lambda_1(c^*) = \lambda_2(c^*) = \lambda^*$ such that $\Delta(\lambda, c^*) > 0$ for $\lambda \neq \lambda^*$;

(D3) if $c > c^*$, then the equation $\Delta(\lambda, c) = 0$ has two positive real roots $\lambda_1(c)$ and $\lambda_2(c)$ with $0 < \lambda_1(c) < \lambda^* < \lambda_2(c)$ such that $\lambda_1'(c) < 0$, $\lambda_2'(c) > 0$, $\frac{d}{dc}\{c\lambda_1(c)\} < 0$, and

$$\Delta(\lambda, c) = \begin{cases} > 0 & \text{for } \lambda < \lambda_1(c), \\ < 0 & \text{for } \lambda \in (\lambda_1(c), \lambda_2(c)), \\ > 0 & \text{for } \lambda > \lambda_2(c). \end{cases}$$

We note that in [25] (see also [1], where bistable waves were considered) there is a further assumption on the kernel I; that is, *the support of I contains either $i = 1$ or two relatively prime integers*, to ensure $\phi_c'(\xi) > 0$ for every $c \geq c^*$. It is interesting to note that, for $c > c^*$, we can confirm $\phi_c'(\xi) > 0$ for any $\xi \in \mathbb{R}$ without this assumption. In fact, for a fixed $c > c^*$, their proof of the existence of nondecreasing traveling wave fronts ϕ_c with $\phi_c(-\infty) = 0$, $\phi_c(+\infty) = K$, and $\lim_{\xi \to -\infty} \phi_c(\xi) e^{-\lambda_1(c)\xi} = 1$ is independent of this assumption; so is the proof of $\lim_{\xi \to -\infty} \phi_c'(\xi) e^{-\lambda_1(c)\xi} = \lambda_1(c)$. We now note that $\phi_c'(\xi) \geq 0$ for $\xi \in \mathbb{R}$. Assume that there exists $\xi_0 \in \mathbb{R}$ such that $\phi_c'(\xi_0) = 0$. Then there must be $\phi_c''(\xi_0) = 0$. It is obvious that ϕ_c satisfies

$$\begin{aligned} c\phi_c''(\xi) = D &\sum_{i \in \mathbb{Z} \setminus \{0\}} I(i) [\phi_c'(\xi - i) - \phi_c'(\xi)] - d\phi_c'(\xi) \\ &+ \sum_{i \in \mathbb{Z}} J(i) b' (\phi_c(\xi - i - c\tau)) \phi_c'(\xi - i - c\tau), \end{aligned}$$

which implies that $\sum_{i \in \mathbb{Z} \setminus \{0\}} I(i) \phi_c'(\xi_0 - i) = 0$. By $\sum_{i \in \mathbb{Z} \setminus \{0\}} I(i) = 1$, there exists a $i_0 \in \mathbb{N}$ such that $I(i_0) > 0$ and $\phi_c'(\xi_0 - i_0) = 0$. Let $\xi_1 = \xi_0 - i_0$. A similar argument yields $\phi_c'(\xi_0 - 2i_0) = \phi_c'(\xi_1 - i_0) = 0$. Continuing this procedure, we have

$\phi_c'(\xi_0 - mi_0) = 0$ for all $m \in \mathbb{N}$, a contradiction to the fact $\lim_{\xi\to-\infty} \phi_c'(\xi)e^{-\lambda_1(c)\xi} = \lambda_1(c)$. Therefore, we have $\phi_c'(\xi) > 0$ for $\xi \in \mathbb{R}$. Of course, to prove $\phi_c'(\xi) > 0$ for the case $c = c^*$ the above assumption on the support of I seems necessary. Since in what follows in this paper we use only the traveling wave fronts ϕ_c with $c > c^*$, we shall not require this assumption.

In the remainder of this paper, we always normalize the traveling wave front $\phi_c(n + ct)$ so that $\phi_c(0) = \frac{K}{2}$. Then, for each $c \in (c^*, +\infty)$, we set

$$\alpha_c = \lim_{z\to-\infty} \phi_c(z)e^{-\lambda_1(c)z}. \tag{1.6}$$

Furthermore, we define $A_c > 0$ for each $c \in (c^*, +\infty)$ by

$$A_c = \inf\left\{A > 0 : A \geq \phi_c(z)e^{-\lambda_1(c)z} \quad \text{for any} \quad z \in \mathbb{R}\right\}. \tag{1.7}$$

It is easy to see that $A_c \geq \alpha_c$.

Our focus is on the so-called entire solutions; here an entire solution of (1.1) is a solution defined for all $(n, t) \in \mathbb{Z} \times \mathbb{R}$. In what follows, we say that a sequence of functions $\Phi_p(t) = \{\Phi_{n,p}(t)\}_{n\in\mathbb{Z}}$ converges to a function $\Phi_{p_0}(t) = \{\Phi_{n,p_0}(t)\}_{n\in\mathbb{Z}}$ in $\mathcal{T}$ if, for every compact set $S \subset \mathbb{Z} \times \mathbb{R}$, the functions $\Phi_{n,p}(t)$ and $\frac{d}{dt}\Phi_{n,p}(t)$ converge uniformly in $(n, t) \in S$ to $\Phi_{n,p_0}(t)$ and $\frac{d}{dt}\Phi_{n,p_0}(t)$ as $p \to p_0$.

One of our main results can be stated as follows.

THEOREM 1.1. *Let $\Gamma(t)$ be a heteroclinic orbit of the following functional differential equation:*

$$\frac{d}{dt}u(t) = -du(t) + b(u(t - \tau)),$$

which is increasing and satisfies $\Gamma(-\infty) = 0$, $\Gamma(+\infty) = K$, $\lim_{t\to-\infty} e^{-\lambda_* t}\Gamma(t) = K$, and $\Gamma(t) \leq Ke^{\lambda_* t}$ for all $t \in \mathbb{R}$, where $\lambda_* > 0$ is the unique real root of the equation $\lambda + d - b'(0)e^{-\lambda\tau} = 0$. Then for every $c_1, \ldots, c_m, c_1', \ldots, c_l' > c^*$, $\theta_0, \theta_1, \ldots, \theta_m, \theta_1', \ldots, \theta_l' \in \mathbb{R}$, and $\chi \in \{0, 1\}$, there exists an entire solution $\Phi(t) = \{\Phi_n(t)\}_{n\in\mathbb{Z}}$ of (1.1) such that

$$\begin{aligned} &\max\left\{\max_{1\leq i\leq m} \phi_{c_i}\left(n + c_i t + \theta_i\right), \max_{1\leq j\leq l} \phi_{c_j'}\left(-n + c_j' t + \theta_j'\right), \chi\Gamma\left(t + \theta_0\right)\right\} \\ &\quad \leq \Phi_n(t) \leq \min\left\{\vartheta_m^+(n, t), \vartheta_l^-(n, t), \vartheta^0(n, t)\right\} \end{aligned} \tag{1.8}$$

on $(n, t) \in \mathbb{Z} \times \mathbb{R}$, where

$$\begin{aligned} \vartheta_m^+(n, t) = \min_{1\leq i\leq m} \Bigg\{ &\phi_{c_i}\left(n + c_i t + \theta_i\right) + \chi K e^{\lambda_*(t+\theta_0)} \\ &+ \sum_{1\leq j\leq m, j\neq i} A_{c_j} e^{\lambda_1(c_j)(n+c_j t+\theta_j)} + \sum_{1\leq j\leq l} A_{c_j'} e^{\lambda_1\left(c_j'\right)\left(-n+c_j' t+\theta_j'\right)} \Bigg\}, \end{aligned}$$

$$\begin{aligned} \vartheta_l^-(n, t) = \min_{1\leq i\leq l} \Bigg\{ &\phi_{c_i'}\left(-n + c_i' t + \theta_i'\right) + \chi K e^{\lambda_*(t+\theta_0)} \\ &+ \sum_{1\leq j\leq m} A_{c_j} e^{\lambda_1(c_j)(n+c_j t+\theta_j)} + \sum_{1\leq j\leq l, j\neq i} A_{c_j'} e^{\lambda_1\left(c_j'\right)\left(-n+c_j' t+\theta_j'\right)} \Bigg\}, \end{aligned}$$

$$\vartheta^0(n, t) = \chi\Gamma\left(t + \theta_0\right) + \sum_{1\leq j\leq m} A_{c_j} e^{\lambda_1(c_j)(n+c_j t+\theta_j)} + \sum_{1\leq j\leq l} A_{c_j'} e^{\lambda_1\left(c_j'\right)\left(-n+c_j' t+\theta_j'\right)},$$

and $m, l \in \mathbb{N} \cup \{0\}$ with $\chi + m + l \geq 2$. Moreover, the following statements hold:

(i) *For any* $n \in \mathbb{Z}$, $\Phi_n'(t) > 0$ *for* $t \in \mathbb{R}$.

(ii) $\lim_{t\to\infty} \sup_{n\in\mathbb{Z}} |\Phi_n(t) - K| = 0$ *and* $\lim_{t\to-\infty} \sup_{|n|\le N_0} |\Phi_n(t)| = 0$ *for every given* $N_0 \in \mathbb{N}$.

(iii) *If* $m \ge 1$, *then* $\lim_{n\to\infty} \|\Phi_n(\cdot) - K\|_{L^\infty[a,+\infty)} = 0$ *for every* $a \in \mathbb{R}$; *if* $l \ge 1$, *then* $\lim_{n\to-\infty} \|\Phi_n(\cdot) - K\|_{L^\infty[a,+\infty)} = 0$ *for every* $a \in \mathbb{R}$.

(iv) *If* $\chi = 1$ *and* $m = 0$ ($l = 0$, *respectively*), *then* $\Phi_n(t)$ *converges uniformly on* $t \in [a,b]$ *to* $\Gamma(t+\theta_0)$ *as* $n \to +\infty$ ($n \to -\infty$, *respectively*) *for any* $a, b \in \mathbb{R}$ *with* $a < b$.

(v) *If* $\chi = 1$, *then* $\Phi_n(t) \sim Ke^{\lambda_*(t+\theta_0)}$ *as* $t \to -\infty$ *for every* $n \in \mathbb{Z}$.

(vi) *If* $\chi = 0$, *then, for every* $n \in \mathbb{Z}$, *there exist* $B_2(n) > B_1(n) > 0$ *such that*

$$B_1(n)e^{c_{\max}\lambda_1(c_{\max})t} < \Phi_n(t) < B_2(n)e^{c_{\max}\lambda_1(c_{\max})t} \quad \text{for every } t \ll -1,$$

where $c_{\max} = \max\{\max_{1\le i\le m} c_i, \max_{1\le j\le l} c_j'\}$.

(vii) *If we denote* $\Phi(t)$ *by* $\Phi_{c_1,\dots,c_m;c_1',\dots,c_l';\theta_1,\dots,\theta_m;\theta_1',\dots,\theta_l';\theta_0}(t)$ *when* $\chi = 1$ *and denote* $\Phi(t)$ *by* $\Phi_{c_1,\dots,c_m;c_1',\dots,c_l';\theta_1,\dots,\theta_m;\theta_1',\dots,\theta_l'}(t)$ *when* $\chi = 0$, *then*

$$\Phi_{c_1,\dots,c_m;c_1',\dots,c_l';\theta_1,\dots,\theta_m;\theta_1',\dots,\theta_l';\theta_0}(t)$$

converges to $\Phi_{c_1,\dots,c_m;c_1',\dots,c_l';\theta_1,\dots,\theta_m;\theta_1',\dots,\theta_l'}(t)$ *as* $\theta_0 \to -\infty$ *in* $\mathcal{T}$ *and uniformly on* $(n,t) \in \mathbb{Z}\times(-\infty,a]$ *for every* $a \in \mathbb{R}$; $\Phi_{c_1,\dots,c_m;c_1',\dots,c_l';\theta_1,\dots,\theta_m;\theta_1',\dots,\theta_l';\theta_0}(t)$ *converges to* K *as* $\theta_0 \to +\infty$ *in* $\mathcal{T}$ *and uniformly on* $(n,t) \in \mathbb{Z}\times[a,+\infty)$ *for every* $a \in \mathbb{R}$.

(viii) $\Phi_{c_1,\dots,c_m;c_1',\dots,c_l';\theta_1,\dots,\theta_m;\theta_1',\dots,\theta_l';\theta_0}(t)$ *converges to*

$$\Phi_{c_1,\dots,c_{i-1},c_{i+1},\dots,c_m;c_1',\dots,c_l';\theta_1,\dots,\theta_{i-1},\theta_{i+1},\dots,\theta_m;\theta_1',\dots,\theta_l';\theta_0}(t)$$

as $\theta_i \to -\infty$ *in* $\mathcal{T}$ *and uniformly on* $(n,t) \in \{n : n \le N_0, n \in \mathbb{Z}\}\times(-\infty,a]$ *for every* $N_0 \in \mathbb{Z}$ *and* $a \in \mathbb{R}$. $\Phi_{c_1,\dots,c_m;c_1',\dots,c_l';\theta_1,\dots,\theta_m;\theta_1',\dots,\theta_l';\theta_0}(t)$ *converges to*

$$\Phi_{c_1,\dots,c_m;c_1',\dots,c_{j-1}',c_{j+1}',\dots,c_l';\theta_1,\dots,\theta_m;\theta_1',\dots,\theta_{j-1}',\theta_{j+1}',\dots,\theta_l';\theta_0}(t)$$

as $\theta_j' \to -\infty$ *in* $\mathcal{T}$ *and uniformly on* $(n,t) \in \{n : n \ge N_0, n \in \mathbb{Z}\}\times(-\infty,a]$ *for every* $N_0 \in \mathbb{Z}$ *and* $a \in \mathbb{R}$. *Similar results hold for* $\Phi_{c_1,\dots,c_m;c_1',\dots,c_l';\theta_1,\dots,\theta_m;\theta_1',\dots,\theta_l'}(t)$.

(ix) $\Phi(t)$ *converges to* K *as* $\theta_i \to +\infty$ *in* $\mathcal{T}$ *and uniformly on* $(n,t) \in \{n : n \ge N_0, n \in \mathbb{Z}\}\times[a,+\infty)$ *for every* $N_0 \in \mathbb{Z}$ *and* $a \in \mathbb{R}$; $\Phi(t)$ *converges to* K *as* $\theta_j' \to +\infty$ *in* $\mathcal{T}$ *and uniformly on* $(n,t) \in \{n : n \le N_0, n \in \mathbb{Z}\}\times[a,+\infty)$ *for every* $N_0 \in \mathbb{Z}$ *and* $a \in \mathbb{R}$.

From (iv) and (v) of Theorem 1.1 and the fact $\lambda_* < c_{\max}\lambda_1(c_{\max})$, it follows that $\Phi_{c_1,\dots,c_m;c_1',\dots,c_l';\theta_1,\dots,\theta_m;\theta_1',\dots,\theta_l';\theta_0}(t)$ are completely different from

$$\Phi_{c_1,\dots,c_m;c_1',\dots,c_l';\theta_1,\dots,\theta_m;\theta_1',\dots,\theta_l'}(t).$$

Theorem 1.1 applies to the spatially discrete Fisher-KPP equation (1.4), where $f \in C^2$ satisfies $f(0) = f(1) = 0$, $f'(0) > 0$, $f'(1) < 0$, $f(u) > 0$, and $f(u) \le f'(0)u$ for $u \in (0,1)$. In this case, $K = 1$, $d = \max_{u\in[0,1]} |f'(u)|$, and $b(u) = du + f(u)$. The existence, uniqueness, and stability of traveling wave fronts of (1.4) were studied in Chen, Fu, and Guo [3], Chen and Guo [4, 5], and Zinner [44]; the entire solutions of (1.4) were studied by Guo and Morita [18] and Guo [19]. However, there seem to be no results on the uniqueness of entire solutions of (1.4) and the continuous dependence on parameters $c_1, \dots, c_m, c_1', \dots, c_l', \theta_0, \theta_1, \dots, \theta_m, \theta_1', \dots, c_l'$ given by Theorem 1.1. The following theorem is devoted to this topic and is a spatially discrete version of results of Hamel and Nadirashvili [20], where the reaction-diffusion Fisher-KPP equation was considered.

THEOREM 1.2. *For any $c, c' > c^*$, $\theta_0, \theta, \theta' \in \mathbb{R}$, and $\varrho, \varrho', \chi \in \{0,1\}$ with $\varrho + \varrho' + \chi \geq 2$, there exists a unique entire solution $\Phi(t) = \{\Phi_n(t)\}_{n\in\mathbb{Z}}$ of (1.4) such that* (i)–(ix) *of Theorem* 1.1 *hold and*

$$
\begin{aligned}
(1.9)\qquad & \max\left\{\varrho\phi_c\left(n+ct+\theta\right), \varrho'\phi_{c'}\left(-n+c't+\theta'\right), \chi\Gamma\left(t+\theta_0\right)\right\} \\
& \leq \Phi_n(t) \\
& \leq \min\Big\{\varrho\phi_c\left(n+ct+\theta\right) + \chi e^{f'(0)(t+\theta_0)} + \varrho' A_{c'} e^{\lambda_1(c')\left(-n+c't+\theta'\right)}, \\
& \qquad\qquad \varrho'\phi_{c'}\left(-n+c't+\theta'\right) + \chi e^{f'(0)(t+\theta_0)} + \varrho A_c e^{\lambda_1(c)(n+ct+\theta)}, \\
& \qquad\qquad \chi\Gamma\left(t+\theta_0\right) + \varrho A_c e^{\lambda_1(c)(n+ct+\theta)} + \varrho' A_{c'} e^{\lambda_1(c')\left(-n+c't+\theta'\right)}\Big\}
\end{aligned}
$$

on $(n,t) \in \mathbb{Z}\times\mathbb{R}$. In particular, when $\varrho = \varrho' = \chi = 1$, the entire solutions $\Phi = \Phi_{c,c',\theta,\theta',\theta_0}$ depend continuously on $(c,c',\theta,\theta',\theta_0) \in (c^*,+\infty)^2\times\mathbb{R}^3$ in $\mathcal{T}$; when $\varrho = \varrho' = 1$ and $\chi = 0$, the entire solutions $\Phi = \Phi_{c,c',\theta,\theta'}$ depend continuously on $(c,c',\theta,\theta') \in (c^*,+\infty)^2\times\mathbb{R}^2$ in $\mathcal{T}$; when $\varrho = \chi = 1$ and $\varrho' = 0$, the entire solutions $\Phi = \Phi_{c,\theta,\theta_0}$ depend continuously on $(c,\theta,\theta_0) \in (c^*,+\infty)\times\mathbb{R}^2$ in $\mathcal{T}$; when $\varrho' = \chi = 1$ and $\varrho = 0$, the entire solutions $\Phi = \Phi_{c',\theta',\theta_0}$ depend continuously on $(c',\theta',\theta_0) \in (c^*,+\infty)\times\mathbb{R}^2$ in $\mathcal{T}$.

We note that when $\varrho' = \chi = 0$ and $\varrho = 1$, $\Phi_n(t) = \Phi_{n;c,\theta}(t) = \phi(n+ct+\theta)$ for $(n,t) \in \mathbb{Z}\times\mathbb{R}$; when $\varrho = \chi = 0$ and $\varrho' = 1$, $\Phi_n(t) = \Phi_{n;c',\theta'}(t) = \phi(-n+c't+\theta')$ for $(n,t) \in \mathbb{Z}\times\mathbb{R}$; and when $\varrho = \varrho' = 0$ and $\chi = 1$, $\Phi_n(t) = \Phi_{n;\theta_0}(t) = \Gamma(t+\theta_0)$ for $(n,t) \in \mathbb{Z}\times\mathbb{R}$. Therefore, similarly to the discussions in Hamel and Nadirashvili [20], it follows from Theorems 1.1 and 1.2 that the functions $\Phi_{c,c',\theta,\theta',\theta_0}(t)$ ($\Phi_{c,c',\theta,\theta'}(t)$, $\Phi_{c,\theta,\theta_0}(t)$, $\Phi_{c',\theta',\theta_0}(t)$, respectively) established by Theorem 1.2 are the 5-D (4-D, 3-D, and 3-D, respectively) manifold of entire solutions of (1.4). In addition, (1.4) possesses two 2-D manifolds of entire solutions of traveling wave type, namely, $\Phi^+_{c,\theta}(t) = \{\phi_c(n+ct+\theta)\}_{n\in\mathbb{Z}}$ and $\Phi^-_{c',\theta'}(t) = \{\phi_{c'}(-n+c't+\theta')\}_{n\in\mathbb{Z}}$, and a 1-D manifold of spatially homogeneous entire solutions, namely, $\Gamma(t+\theta_0)$. Let $\mathcal{M}_5$ ($\mathcal{M}_4$, $\mathcal{M}_3^+$, $\mathcal{M}_3^-$, $\mathcal{M}_2^+$, $\mathcal{M}_2^-$, and $\mathcal{M}_1$, respectively) be the above 5-D (4-D, 3-D, 3-D, 2-D, 2-D, and 1-D, respectively) manifold of entire solutions. Then, from Theorems 1.1 and 1.2, it follows that $\mathcal{M}_4$ is on the boundary of $\mathcal{M}_5$ (via taking the limit $\theta_0 \to -\infty$) and $\mathcal{M}_3^+$ (or $\mathcal{M}_3^-$) is on the boundary of $\mathcal{M}_5$ (via taking the limit $\theta \to -\infty$) (or $\theta' \to -\infty$). $\mathcal{M}_2^+$ (or $\mathcal{M}_2^-$) is on the boundary of $\mathcal{M}_4$ (via taking the limit $\theta' \to -\infty$) (or $\theta \to -\infty$) and is also on the boundary of $\mathcal{M}_3^+$ (or $\mathcal{M}_3^-$) (via taking the limit $\theta_0 \to -\infty$). $\mathcal{M}_1$ is on the boundary of $\mathcal{M}_3^+$ (or $\mathcal{M}_3^-$) (via taking the limit $\theta \to -\infty$) (or $\theta' \to -\infty$). In particular, $\mathcal{M}_2^+$ (or $\mathcal{M}_2^-$) is on the boundary of $\mathcal{M}_5$ (via taking the limits $\theta' \to -\infty$ and $\theta_0 \to -\infty$) (or $\theta \to -\infty$ and $\theta_0 \to -\infty$), and $\mathcal{M}_1$ is on the boundary of $\mathcal{M}_5$ (via taking the limits $\theta \to -\infty$ and $\theta' \to -\infty$). We can also easily show that the functions $\Phi_{c,c',\theta,\theta',\theta_0}$ converge to $\Phi^+_{c,\theta}$ as $\theta' \to -\infty$ and $\theta_0 \to -\infty$ in $\mathcal{T}$ and to $\Phi^-_{c',\theta'}$ as $\theta \to -\infty$ and $\theta_0 \to -\infty$ in $\mathcal{T}$, and that $\Phi_{c,c',\theta,\theta',\theta_0}$ converge to Φ_{θ_0} as $\theta \to -\infty$ and $\theta' \to -\infty$ in $\mathcal{T}$.

Contrasting to [18, 20], we require only $f(u) \leq f'(0)u$ for any $u \in (0,1)$ other than $f'(u) \leq f'(0)$. We also note some differences on the uniqueness of entire solutions up to a spatial-temporal translation between a reaction-diffusion equation and its spatially discrete analogue (see a similar remark for the bistable nonlinearity reported by Wang, Li, and Ruan [39]). Namely, consider the reaction-diffusion KPP equation

$$
(1.10)\qquad \frac{d}{dt}u(x,t) = D\Delta u(x,t) + f(u),
$$

for which the existence of entire solutions was established by Hamel and Nadirashvili ([20, Theorems 1.1, 1.3, and 1.4 and Corollary 1.5]). For comparison, we consider only the entire solutions established by [20, Theorem 1.3], corresponding to the case $\chi = 0$ and $\varrho = \varrho' = 1$ in our Theorem 1.2. The entire solution $v_{c,c',h,h'}(x,t)$ of (1.10) established by [20, Theorem 1.3] and satisfying (1.4) of [20] is unique for each given $(c,c',h,h') \in (c^*,+\infty)^2 \times \mathbb{R}^2$. Consequently, it is easy to see that for any $(\bar{h},\bar{h}') \neq (h,h')$,

$$v_{c,c',\bar{h},\bar{h}'}(x,t) = v_{c,c',\bar{h},\bar{h}'}(x+x_0,t+t_0) \quad \text{for} \quad (x,t) \in \mathbb{R}^2,$$

where

$$x_0 = \frac{c\left(\bar{h}'-h'\right) - c'\left(\bar{h}-h\right)}{c+c'}, \quad t_0 = \frac{\left(\bar{h}'-h'\right)+\left(\bar{h}-h\right)}{c+c'}.$$

But for (1.4) if $(\bar{\theta},\bar{\theta}') \neq (\theta,\theta')$, then $\Phi_{n;c,c',\bar{\theta},\bar{\theta}'}(t) = \Phi_{n+n_0;c,c',\theta,\theta'}(t+t_0)$ for any $(n,t) \in \mathbb{Z} \times \mathbb{R}$ if and only if

$$\frac{c\left(\bar{h}'-h'\right) - c'\left(\bar{h}-h\right)}{c+c'} \in \mathbb{Z},$$

and, hence, $n_0 = \frac{c(\bar{h}'-h') - c'(\bar{h}-h)}{c+c'}$, $t_0 = \frac{(\bar{h}'-h')+(\bar{h}-h)}{c+c'}$. When $(\bar{c},\bar{c}') \neq (c,c')$, as proved by [24, Theorem 1.1], there exists no $(x_0,t_0) \in \mathbb{R}^2$ such that $v_{\bar{c},\bar{c}',h,h'}(\cdot,\cdot) = v_{c,c',h,h'}(\cdot + x_0, \cdot + t_0)$ on $\mathbb{R}^2$ for (1.10). Similarly, for (1.4), there exists no $(n_0,t_0) \in \mathbb{Z} \times \mathbb{R}$ such that $\Phi_{n;\bar{c},\bar{c}',h,h'}(t) = v_{n+n_0;c,c',h,h'}(t)$ for any $(n,t) \in \mathbb{Z} \times \mathbb{R}$.

There have been extensive studies about the dynamics of lattice delay systems (1.1), as reported in a recent survey by Gourley and Wu [16]. In particular, the asymptotic speed of propagation and the existence of monotone traveling waves were studied in [25, 40]. The existence, uniqueness, and stability of traveling wave solutions of (1.1) and (1.2) with monostable and bistable nonlinearities have henceforth been studied; see Ma and Zou [26] for the bistable case and Ma and coworkers [25, 27] for the monostable case. Also, Gourley and Wu [17] proved for (1.2) that if the birth rate is so small that a patch alone cannot sustain a positive equilibrium, then the whole population in the patchy environment will become extinct; and if the birth rate is large enough that each patch can sustain a positive equilibrium and if the maturation time is moderate, then the model exhibits nonlinear oscillations characterized by the occurrence of multiple periodic traveling waves. A stage-structured model for a single species on a finite 1-D spatial lattice was also studied in [22]. Related results on traveling waves of lattice differential equations (without delay) can be found in Cahn, Chow, and Van Vleck [2], Chen, Fu, and Guo [3], Chen and Guo [4, 5], Chow [10], Mallet-Paret [28], Wu and Zou [42], and references therein. We note that some progress has been made as well for 2-D lattice delay differential equations; see, for example, Cheng, Li, and Wang [8, 9], Shi, Li, and Cheng, [32], and Weng et al. [41]. In addition, Wang, Li, and Ruan [35, 36, 37] studied traveling wave solutions of reaction-diffusion equations with spatial-temporal delay.

The aforementioned studies also suggest that these wave solutions $\phi_c(n+ct)$ are defined for all $t \in \mathbb{R}$. They often determine the long time behavior of the solutions of Cauchy-type problems and constitute an important part of global attractors, which consist of *entire solutions*. However, the global attractors can be quite complicated, and recent studies for reaction-diffusion equations with continuous spatial variables have showed the existence of many new types of entire solutions arising from the

simple traveling wave fronts, and these entire solutions combined provide essential information about the global attractors; see Chen and Guo [6], Chen, Guo, and Nimomiya [7], Fukao, Morita, and Nimomiya [14], Guo and Morita [18], Hamel and Nadirashvili [20, 21], and Yagisita [43]. For the Fisher-KPP nonlinearity and bistable nonlinearity, these entire solutions behave as two (opposite) wave fronts of positive speed(s) approaching each other from both sides of the x-axis and then annihilate in a finite time. Similar results hold true for nonlocal reaction-diffusion equations with delayed monostable and bistable nonlinearities ([24, 38]). Morita and Ninomiya [30] and Guo [19] have constructed other types of entire solutions for reaction-diffusion equations and discrete diffusive equations with bistable nonlinearity, respectively, which are different from those obtained in [6, 7, 14, 18, 20, 21, 24, 38, 43]. In particular, Li, Liu, and Wang [23] established the existence of entire solutions for reaction-advection-diffusion equations in cylinders, where the ignition temperature nonlinearity has been studied. As reported in [30], entire solutions play also very important roles in some other areas, for example, transient dynamics, distinct history of two solutions, etc.

The remainder of this paper is organized as follows: In section 2, we show how systems (1.1) arise from some areas, such as population biology. In section 3, we establish some existence and comparison results, which are needed in what follows. In section 4, we show the existence of heteroclinic orbit $\Gamma(t)$ connecting two equilibria 0 and K. Section 5 is devoted to Theorem 1.1, and then Theorem 1.2 is proved in section 6.

2. Important particular cases. In this section, we derive from a structured population model a particular case of systems along with an explicit formula to calculate $J(i)$.

Consider a single species population with age structures distributed over a patchy environment consisting of all integer nodes of a 1-D lattice. Let $w_n(t)$ be the density of juvenile individuals in the nth patch and at time t, $v_n(t,a)$ be the density of individuals with age a in the nth patch and at time t, and $\tau > 0$ the length of a juvenile period. Then

$$w_n(t) = \int_0^\tau v_n(t,a)\, da.$$

Let $u_n(t)$ be the density of mature individuals in the nth patch and at time t. Assume that the spatial dispersal of juvenile individuals and mature individuals is isotropic and can be long range (see Murray [31]). Assume that the diffusion rate of juvenile individuals with age a is $\overline{D}(a) \geq 0$ and the diffusion rate of mature individuals is a constant $D > 0$. Let $K(n-i)$ and $I(n-i)$ be the probability distributions of juvenile individuals and mature individuals traveling from the ith patch to the nth patch, respectively. Then we have

$$K(i) \geq 0,\ I(i) \geq 0,\ K(i) = K(-i),\ I(i) = I(-i),\quad \sum_{i\in\mathbb{Z}\setminus\{0\}} K(i) = 1,\quad \sum_{i\in\mathbb{Z}\setminus\{0\}} I(i) = 1.$$

Since only the mature population can reproduce, we have

$$\begin{cases} \frac{\partial}{\partial t} v_n(t,a) + \frac{\partial}{\partial a} v_n(t,a) = \overline{D}(a) \sum_{i\in\mathbb{Z}\setminus\{0\}} K(n-i)\left[v_i(t,a) - v_n(t,a)\right] \\ \qquad\qquad -\mu(a)\, v_n(t,a), \quad 0 < a < \tau, \\ v_n(t,0) = \widehat{b}\left(u_n(t)\right), \\ \frac{d}{dt} u_n(t) = D \sum_{i\in\mathbb{Z}\setminus\{0\}} I(n-i)\left[u_i(t) - u_n(t)\right] - \widehat{d}\left(u_n(t)\right) + v_n(t,\tau), \end{cases} \tag{2.1}$$

where $\mu(a)$ denotes the death rate of the juvenile individuals with age $a \in (0,\tau)$, $\widehat{b}:\mathbb{R}_+ \to \mathbb{R}_+$ is the birth function, and $\widehat{d}:\mathbb{R}_+ \to \mathbb{R}_+$ is the death function of mature individuals.

For fixed $s \geq -\tau$, let $V_n^s(t) = v_n(t, t-s)$ for $s \leq t \leq s+\tau$. Then $V_n^s(s) = v_n(s,0) = \widehat{b}(u_n(s))$. From (2.1),

$$\begin{aligned}(2.2)\quad \frac{d}{dt}V_n^s(t) &= \frac{\partial}{\partial t}v_n(t,a)\Big|_{a=t-s} + \frac{\partial}{\partial a}v_n(t,a)\Big|_{a=t-s} \\ &= \overline{D}(t-s)\sum_{i\in\mathbb{Z}\setminus\{0\}} K(i)\left[V_{n-i}^s(t) - V_n^s(t)\right] - \mu(t-s)V_n^s(t).\end{aligned}$$

Note that the grid function $V_n^s(t)$ can be viewed as the discrete spectral of a periodic function $v^s(t,\omega)$ by discrete Fourier transform [15, 34]:

$$(2.3)\qquad v^s(t,\omega) = \frac{1}{\sqrt{2\pi}}\sum_{n=-\infty}^{\infty} e^{-i(n\omega)}V_n^s(t),$$

$$(2.4)\qquad V_n^s(t) = \frac{1}{\sqrt{2\pi}}\int_{-\pi}^{\pi} e^{i(n\omega)}v^s(t,\omega)\,d\omega,$$

where i is the imaginary unit. Applying (2.2) and (2.3) yields

$$\begin{aligned}\frac{\partial}{\partial t}v^s(t,\omega) &= \left[\overline{D}(t-s)\sum_{k\in\mathbb{Z}\setminus\{0\}} K(k)\left(e^{-ik\omega}-1\right) - \mu(t-s)\right]v^s(t,\omega) \\ &= \left[-2\overline{D}(t-s)\sum_{k\in\mathbb{Z}\setminus\{0\}} K(k)\sin^2\left(\frac{k\omega}{2}\right) - \mu(t-s)\right]v^s(t,\omega).\end{aligned}$$

Solving the equation, we get

$$\begin{aligned}v^s(t,\omega) = \exp&\left\{-2\sum_{k\in\mathbb{Z}\setminus\{0\}} K(k)\sin^2\left(\frac{k\omega}{2}\right)\int_s^t \overline{D}(z-s)\,dz - \int_s^t \mu(z-s)\,dz\right\} \\ &\times v^s(s,\omega).\end{aligned}$$

By the inverse discrete Fourier transform (2.4), we obtain

$$\begin{aligned}V_n^s(t) = &\frac{1}{\sqrt{2\pi}}e^{-\int_s^t \mu(z-s)dz} \\ &\times\int_{-\pi}^{\pi} e^{i(n\omega)}\exp\left\{-2\alpha_s\sum_{k\in\mathbb{Z}\setminus\{0\}} K(k)\sin^2\left(\frac{k\omega}{2}\right)\right\}v^s(s,\omega)\,d\omega,\end{aligned}$$

where $\alpha_{t-s} = \int_s^t \overline{D}(z-s)dz = \int_0^{t-s}\overline{D}(z)dz$. Noting that $V_n^s(s) = v_n(s,0) = \widehat{b}(u_n(s))$, by (2.3) we have

$$v_s(s,\omega) = \frac{1}{\sqrt{2\pi}}\sum_{j\in\mathbb{Z}} e^{-i(j\omega)}V_j^s(t) = \frac{1}{\sqrt{2\pi}}\sum_{j\in\mathbb{Z}} e^{-i(j\omega)}\widehat{b}(u_j(s)).$$

Hence,

$$V_n^s(t) = \frac{1}{2\pi} e^{-\int_s^t \mu(z-s)dz} \sum_{j\in\mathbb{Z}} \widehat{b}(u_j(s)) \times \int_{-\pi}^{\pi} e^{i((n-j)\omega)} \exp\left\{-2\alpha_s \sum_{k\in\mathbb{Z}\setminus\{0\}} K(k) \sin^2\left(\frac{k\omega}{2}\right)\right\} d\omega. \tag{2.5}$$

Let $t = s+\tau$, $\widehat{\mu} = e^{-\int_s^t \mu(z-s)dz} = e^{-\int_0^\tau \mu(z)dz}$, and $\alpha = \int_0^\tau \overline{D}(z)dz$. Then (2.5) yields

$$v_n(t,\tau) = \frac{\widehat{\mu}}{2\pi} \sum_{j\in\mathbb{Z}} \beta_\alpha(n-j)\widehat{b}(u_j(s)) = \frac{\widehat{\mu}}{2\pi} \sum_{j\in\mathbb{Z}} \beta_\alpha(n-j)\widehat{b}(u_j(t-\tau)),$$

where

$$\beta_\alpha(j) = \int_{-\pi}^{\pi} e^{i(j\omega)} \exp\left\{-2\alpha \sum_{k\in\mathbb{Z}\setminus\{0\}} K(k) \sin^2\left(\frac{k\omega}{2}\right)\right\} d\omega.$$

Thus, the last equality of (2.1) becomes

$$\frac{d}{dt}u_n(t) = D \sum_{i\in\mathbb{Z}\setminus\{0\}} I(i)[u_{n-i}(t) - u_n(t)] - \widehat{d}(u_n(t)) + \frac{\widehat{\mu}}{2\pi} \sum_{j\in\mathbb{Z}} \beta_\alpha(n-j)\widehat{b}(u_j(t-\tau)), \quad t>0. \tag{2.6}$$

Let $\widehat{d}(u) = du$, $b(u) = \widehat{\mu}\widehat{b}(u)$, and $J(i) = \frac{1}{2\pi}\beta_\alpha(i)$; then (2.6) reduces to (1.1).

In particular, the case when $I(i) = K(i) = 0$ for $|i| \neq 1$ and $I(1) = K(1) = \frac{1}{2}$ was studied by Weng, Huang, and Wu [40]. In this case, (2.6) reduces to (1.2). When $\overline{D}(a) \equiv 0$, we have $\alpha = 0$, and it follows that (2.6) reduces to

$$\frac{d}{dt}u_n(t) = D \sum_{i\in\mathbb{Z}\setminus\{0\}} I(i)[u_{n-i}(t) - u_n(t)] - \widehat{d}(u_n(t)) + \widehat{\mu}\widehat{b}(u_n(t-\tau)), \quad t>0.$$

When $\tau = 0$, $\alpha = 0$, and $\widehat{\mu} = 1$, we have

$$\frac{d}{dt}u_n(t) = D \sum_{i\in\mathbb{Z}\setminus\{0\}} I(i)[u_{n-i}(t) - u_n(t)] - \widehat{d}(u_n(t)) + \widehat{b}(u_j(t)), \quad t>0,$$

which coincides with (1.3).

When the diffusion rate $\overline{D}(a)$ and death rate $\mu(a)$ of the juvenile individuals are independent of age a, namely, $D_0 \equiv \overline{D}(a)$ and $\gamma \equiv \mu(a)$ for $a \in [0,\tau]$, we have

$$\frac{d}{dt}w_n(t) = D_0 \sum_{i\in\mathbb{Z}\setminus\{0\}} K(i)[w_{n-i}(t) - w_n(t)] - \gamma u_n(t) + \widehat{b}(u_n(t)) - \frac{e^{-\gamma\tau}}{2\pi} \sum_{j\in\mathbb{Z}} \beta_\alpha(n-j)\widehat{b}(u_j(t-\tau)), \quad t>0. \tag{2.7}$$

We note that it is easy to prove that $\sum_{j\in\mathbb{Z}} \frac{1}{2\pi}\beta_\alpha(j) = 1$. It seems difficult to prove $\beta_\alpha(j) \geq 0$ for general kernel $\sum_{i\in\mathbb{Z}\setminus\{0\}} K(i) = 1$ though it was proved by Weng, Huang, and Wu [40] for the case when $K(i) = 0$ for $|i| \neq 1$ and $K(\pm 1) = \frac{1}{2}$. Nevertheless, in the remainder of this paper, we consider (1.1) for general kernel functions $I(i)$ and $J(i)$ satisfying the assumptions in section 1.

3. Preliminaries. Consider the initial value problem

$$\begin{cases} u_n'(t) = D\sum_{i\in\mathbb{Z}\setminus\{0\}} I(i)\left[u_{n-i}(t) - u_n(t)\right] - du_n(t) + \sum_{i\in\mathbb{Z}} J(i)\, b\left(u_{n-i}(t-\tau)\right), \\ u_n(s) = \varphi_n(s), \end{cases} \tag{3.1}$$

where $n \in \mathbb{Z}$, $t > 0$, and $s \in [-\tau, 0]$.

DEFINITION 3.1. *A sequence of continuous differentiable functions $\{v_n(t)\}_{n\in\mathbb{Z}}$, $t \in [-\tau, l)$, $l > 0$, is called a supersolution (subsolution) of* (3.1) *on* $[0, l)$ *if*

$$\begin{aligned} v_n'(t) \geq (\leq)\, D \sum_{i\in\mathbb{Z}\setminus\{0\}} I(i)\left[v_{n-i}(t) - v_n(t)\right] - dv_n(t) \\ + \sum_{i\in\mathbb{Z}} J(i)\, b\left(v_{n-i}(t-\tau)\right) \end{aligned} \tag{3.2}$$

for all $t \in [0, l)$.

LEMMA 3.2. *For any $\varphi = \{\varphi_n\}_{n\in\mathbb{Z}}$ with $\varphi_n \in C([-\tau, 0], [0, K])$,* (3.1) *admits a unique solution $u(t;\varphi) = \{u_n(t;\varphi)\}_{n\in\mathbb{Z}}$ on $[0, +\infty)$ satisfying $u_n(s) = \varphi_n(s)$ and $0 \leq u_n(t) \leq K$ for $s \in [-\tau, 0]$, $t \in [-\tau, +\infty)$, and $n \in \mathbb{Z}$. For any pair of supersolution $w_n^+(t)$ and subsolution $w_n^-(t)$ of* (3.1) *on $[0, +\infty)$ with $0 \leq w_n^-(t) \leq K$, $0 \leq w_n^+(t) \leq K$ for $t \in [-\tau, +\infty)$, $n \in \mathbb{Z}$, and $w_n^+(s) \geq w_n^-(s)$ for $s \in [-\tau, 0]$, $n \in \mathbb{Z}$, there holds $w_n^+(t) \geq w_n^-(t)$ for $t \geq 0$, $n \in \mathbb{Z}$.*

Note that (3.1) is equivalent to

$$\begin{cases} u_n(t) = \varphi_n(0)\, e^{-(D+d)t} + \int_0^t e^{(D+d)(s-t)} H_n[u](s)\, ds, & t > 0, \\ u_n(t) = \varphi_n(t), & t \in [-\tau, 0], \end{cases}$$

where $H_n[u](t) = D\sum_{i\in\mathbb{Z}\setminus\{0\}} I(i)u_{n-i}(t) + \sum_{i\in\mathbb{Z}} J(i)b(u_{n-i}(t-\tau))$. Lemma 3.2 can be proved using an argument used in [26, Lemma 4.1].

Consider also the following linear initial value problem:

$$\begin{cases} u_n'(t) = D\sum_{i\in\mathbb{Z}\setminus\{0\}} I(i)\left[u_{n-i}(t) - u_n(t)\right] - du_n(t) + b'(0)\sum_{i\in\mathbb{Z}} J(i)\, u_{n-i}(t-\tau), \\ u_n(s) = \varphi_n(s) \in C\left([-\tau, 0], \mathbb{R}\right), \end{cases} \tag{3.3}$$

where $n \in \mathbb{Z}$, $t > 0$, and $s \in [-\tau, 0]$.

Before stating the following theorem, we first define a Banach space l^∞ by

$$l^\infty = \left\{ \xi = \{\xi_i\}_{i\in\mathbb{Z}}, \xi_i \in \mathbb{R} : \sup_{i\in\mathbb{Z}} |\xi_i| < \infty \right\}$$

with the norm $\|\xi\|_{l^\infty} = \sup_{i\in\mathbb{Z}} |\xi_i|$.

THEOREM 3.3. *For any $\varphi = \{\varphi_n\}_{n\in\mathbb{Z}}$ with $\varphi \in C([-\tau, 0], l^\infty)$,* (3.3) *admits a unique solution $u(t) := u(t;\varphi) = \{u_n(t;\varphi)\}_{n\in\mathbb{Z}}$ on $[0, +\infty)$. Furthermore, if $\varphi^1, \varphi^2 \in C([-\tau, 0], l^\infty)$ satisfy $\varphi_n^1(s) \leq \varphi_n^2(s)$ for any $n \in \mathbb{Z}$ and $s \in [-\tau, 0]$, then $u_n(t;\varphi^1) \leq u_n(t;\varphi^2)$ holds for any $n \in \mathbb{Z}$ and $t > 0$.*

Proof. Let $X = l^\infty$. Set $X^+ = \{\xi \in l^\infty : \xi_i \geq 0 \text{ for each } i \in \mathbb{Z}\}$. Then it is easy to see that X^+ is a closed cone of X. Let $T(t) = e^{-(D+d)t}$; it is obvious that $\{T(t)\}$ is a strongly continuous semigroup on X. In particular, it is strongly positive. Now let $\mathcal{C} = C([-\tau, 0], X)$ be the Banach space of continuous functions from $[-\tau, 0]$ into X with the supremum norm. Set $\mathcal{C}^+ = \{\Phi \in \mathcal{C} : \Phi(s) \in X^+, s \in [-\tau, 0]\}$. Then $\mathcal{C}^+$ is a positive cone of $\mathcal{C}$. For any continuous function $w : [-\tau, +\infty) \to X$, define $w_t \in \mathcal{C}$,

$t \in [0, +\infty)$, by $w_t(s) = w(t+s)$, $s \in [-\tau, 0]$. Then the map $t \mapsto w_t$ is a continuous function from $[0, +\infty)$ to $\mathcal{C}$.

Define $f : \mathcal{C} \to X$ by

$$f(w) = \{f_i(w)\}_{i\in\mathbb{Z}}$$

for $w = \{w_i\}_{i\in\mathbb{Z}} \in \mathcal{C}$, where

$$f_i(w) = D \sum_{j\in\mathbb{Z}\setminus\{0\}} I(j)\, w_{i-j}(0) + b'(0) \sum_{k\in\mathbb{Z}} J(k)\, w_{i-k}(-\tau).$$

It is not difficult to verify that $f : \mathcal{C} \to X$ is globally Lipschitz continuous. Furthermore, since for any $v, w \in \mathcal{C}$ with $v \geq w$ in $\mathcal{C}$,

$$\begin{aligned} f_i(v) - f_i(w) &= D \sum_{j\in\mathbb{Z}\setminus\{0\}} I(j)\, v_{i-j}(0) - D \sum_{j\in\mathbb{Z}\setminus\{0\}} I(j)\, w_{i-j}(0) \\ &\quad + b'(0) \sum_{k\in\mathbb{Z}} J(k)\, v_{i-k}(-\tau) - b'(0) \sum_{k\in\mathbb{Z}} J(k)\, w_{i-k}(-\tau) \\ &\geq 0, \end{aligned}$$

it follows that $f(v) \geq f(w)$ in X for any $v, w \in \mathcal{C}$ with $v \geq w$, which implies that $f : \mathcal{C} \to X$ is quasi-monotone in the sense that

$$\lim_{h\to 0} \frac{1}{h} \operatorname{dist}\left((v(0) - w(0)) + h\left[f(v) - f(w)\right], X^+\right) = 0$$

for any $v, w \in \mathcal{C}$ with $v \geq w$.

Note that (3.3) is equivalent to

$$\begin{cases} u(t) = T(t)\, u(0) + \int_0^t T(t-s)\, f(u_s)\, ds, & t > 0, \\ u(t) = \varphi(t), & t \in [-\tau, 0]. \end{cases} \tag{3.4}$$

Take $M_0 = \max_{t\in[-\tau,0]} \|\varphi(t)\|_{l^\infty}$. Furthermore, define a vector-valued function $v^+(\cdot) = \{v_n^+(\cdot)\}_{n\in\mathbb{Z}} : [-\tau, +\infty) \to X$ by

$$\begin{cases} v_n^+(t) = M_0, & t \in [-\tau, 0], \\ v_n^+(t) = M_0 e^{(b'(0)-d)t}, & t > 0 \text{ for any } n \in \mathbb{Z}. \end{cases} \tag{3.5}$$

It is easy to verify that v^+ satisfies

$$v^+(t) \geq T(t)\, v^+(s) + \int_s^t T(t-r)\, f\left(v_r^+\right) dr \text{ for any } t > s \geq 0. \tag{3.6}$$

Define $v^-(\cdot) = \{v_n^-(\cdot)\}_{i\in\mathbb{Z}} : [-\tau, +\infty) \to X$ by $v^-(\cdot) = -v^+(\cdot)$. Then v^- satisfies

$$v^-(t) \leq T(t)\, v^-(s) + \int_s^t T(t-r)\, f\left(v_r^-\right) dr \text{ for any } t > s \geq 0. \tag{3.7}$$

Now we use the conclusions of [29]. By setting $S(t,s) = T(t,s) = T(t-s)$ for any $t \geq s \geq 0$ and $B(t, \Phi) = f(\Phi)$, the existence and uniqueness of the solution $u(t;\varphi)$ follows from [29, Corollary 5].

For any $\varphi^1, \varphi^2 \in \mathcal{C}$ with $\varphi^1 \le \varphi^2$ in $\mathcal{C}$, again applying [29, Corollary 5], we have

$$v^-(t) \le u\left(t;\varphi^1\right) \le u\left(t;\varphi^2\right) \le v^+(t) \text{ in } X \text{ for any } t \ge 0,$$

via letting

$$M_0 = \max\left\{ \max_{s\in[-\tau,0]} \left\|\varphi^1(s)\right\|_{l^\infty}, \max_{s\in[-\tau,0]} \left\|\varphi^2(s)\right\|_{l^\infty} \right\}$$

in (3.5). This implies the solution semiflow is order preserving. The proof is complete. □

Remark 3.1. Assume that the continuous functions $w^\pm = \{w_n^\pm\}_{n\in\mathbb{Z}} : [-\tau,+\infty) \to l^\infty$ satisfy (3.6) and (3.7), respectively, and $w_n^+(s) \ge w_n^-(s)$ for any $(n,s) \in \mathbb{Z}\times[-\tau,0]$; then we have $w_n^+(t) \ge w_n^-(t)$ for any $(n,t) \in \mathbb{Z}\times[0,+\infty)$.

THEOREM 3.4. *Assume that*

$$w_n^-(t) \in C\left([-\tau,\infty),(-\infty,K]\right) \quad \text{and} \quad w_n^+(t) \in C\left([-\tau,\infty),[0,\infty)\right)$$

satisfy $w_n^-(t) \le w_n^+(t)$ *for any* $t \in [-\tau,0]$ *and* $n \in \mathbb{Z}$ *and*

$$\begin{aligned} \frac{d}{dt}w_n^+(t) \ge D &\sum_{i\in\mathbb{Z}\setminus\{0\}} I(i)\left[w_{n-i}^+(t) - w_n^+(t)\right] - dw_n^+(t) \\ &+ b'(0)\sum_{i\in\mathbb{Z}} J(i)\, w_{n-i}^+(t-\tau), \end{aligned} \tag{3.8}$$

$$\begin{aligned} \frac{d}{dt}w_n^-(t) \le D &\sum_{i\in\mathbb{Z}\setminus\{0\}} I(i)\left[w_{n-i}^-(t) - w_n^-(t)\right] - dw_n^-(t) \\ &+ b'(0)\sum_{i\in\mathbb{Z}} J(i)\, w_{n-i}^-(t-\tau) \end{aligned} \tag{3.9}$$

for any $t > 0$ *and* $n \in \mathbb{Z}$. *Then there holds* $w_n^+(t) \ge w_n^-(t)$ *for any* $t > 0$ *and* $n \in \mathbb{Z}$.

Proof. Put $w_n(t) := w_n^-(t) - w_n^+(t)$, $n \in \mathbb{Z}$, $t \in [-\tau,+\infty)$. Then $w_n(t)$ is continuous and bounded from above by K, and $\overline{w}(t) := \sup_{n\in\mathbb{Z}} w_n(t)$ is continuous on $[-\tau,\infty)$. We use a contradiction argument to prove the assertion. Suppose that the assertion is not true. Let $M_0 > 0$ be such that $M_0 + d - b'(0)e^{-M_0\tau} > 0$. Then there exists $t_0 > 0$ such that $\overline{w}(t_0) > 0$ and

$$\overline{w}(t_0)\, e^{-M_0 t_0} = \sup_{t\ge-\tau}\left\{\overline{w}(t)\, e^{-M_0 t}\right\} > \overline{w}(s)\, e^{-M_0 s} \text{ for all } s \in [-\tau,t_0). \tag{3.10}$$

Let $\{n_j\}_{j\in\mathbb{N}}$ be a sequence so that $w_{n_j}(t_0) > 0$ for all $j \ge 1$ and $\lim_{j\to\infty} w_{n_j}(t_0) = \overline{w}(t_0)$. Let $\{t_j\}_{j\in\mathbb{N}} \subset (0,t_0]$ so that

$$w_{n_j}(t_j)\, e^{-M_0 t_j} = \max_{t\in[0,t_0]}\left\{w_{n_j}(t)\, e^{-M_0 t}\right\}. \tag{3.11}$$

Since

$$w_{n_j}(t_0)\, e^{-M_0 t_0} \le w_{n_j}(t_j)\, e^{-M_0 t_j} \le \overline{w}(t_j)\, e^{-M_0 t_j} \le \overline{w}(t_0)\, e^{-M_0 t_0},$$

we have $\lim_{j\to+\infty} \overline{w}(t_j)e^{-M_0 t_j} = \overline{w}(t_0)e^{-M_0 t_0}$. Then there must be $\lim_{j\to+\infty} t_j = t_0$ due to (3.10). In view of $w_{n_j}(t_0)e^{-M_0(t_0-t_j)} \le w_{n_j}(t_j) \le \overline{w}(t_0)e^{-M_0(t_0-t_j)}$, we obtain $\lim_{j\to+\infty} w_{n_j}(t_j) = \overline{w}(t_0)$.

Following (3.11), for each $j \geq 1$, we have

$$0 \leq \frac{d}{dt}\left\{w_{n_j}(t)\, e^{-M_0 t}\right\}_{t=t_j} = \left[w'_{n_j}(t_j) - M_0 w_{n_j}(t_j)\right] e^{-M_0 t_j},$$

and, hence, $w'_{n_j}(t_j) \geq M_0 w_{n_j}(t_j)$. Then it follows from (3.8) and (3.9) that

$$\begin{aligned}
0 &\geq w'_{n_j}(t_j) - D\sum_{i\in\mathbb{Z}\setminus\{0\}} I(i)\left[w_{n_j-i}(t_j) - w_{n_j}(t_j)\right] + d w_{n_j}(t_j) \\
&\quad - b'(0)\sum_{i\in\mathbb{Z}} J(i)\left[w^-_{n_j-i}(t_j-\tau) - w^+_{n_j-i}(t_j-\tau)\right] \\
&\geq (M_0 + D + d)\, w_{n_j}(t_j) - D\sum_{i\in\mathbb{Z}\setminus\{0\}} I(i)\, w_{n_j-i}(t_j) - b'(0)\,\overline{w}(t_j-\tau) \\
&\geq (M_0 + D + d)\, w_{n_j}(t_j) - D\overline{w}(t_j) - b'(0)\,\overline{w}(t_j-\tau).
\end{aligned}$$

Taking $j \to +\infty$, we have

$$\begin{aligned}
0 &\geq (M_0 + D + d)\,\overline{w}(t_0) - D\overline{w}(t_0) - b'(0)\, e^{M_0(t_0-\tau)}\left[\overline{w}(t_0-\tau)\, e^{-M_0(t_0-\tau)}\right] \\
&\geq (M_0 + d)\,\overline{w}(t_0) - b'(0)\, e^{M_0(t_0-\tau)}\overline{w}(t_0)\, e^{-M_0 t_0} \\
&= \left[M_0 + d - b'(0)\, e^{-M_0\tau}\right]\overline{w}(t_0).
\end{aligned}$$

In view of $M_0 + d - b'(0)e^{-M_0\tau} > 0$, we obtain that $\overline{w}(t_0) \leq 0$, which contradicts to $\overline{w}(t_0) > 0$. Consequently, we conclude that $w_n^+(t) \geq w_n^-(t)$ for all $n \in \mathbb{Z}$ and $t \in (0, +\infty)$. This completes the proof. □

4. Existence of heteroclinic orbits. In this section, we show the existence of a heteroclinic orbit connecting the equilibria $u \equiv 0$ and $u \equiv K$ for the following functional differential equation:

$$\frac{d}{dt}u(t) = -du(t) + b(u(t-\tau)). \tag{4.1}$$

There are now various methods developed to establish the existence of such a heteroclinic orbit, for example, Faria, Huang, and Wu [11], Faria and Trofimchuk [12, 13], Li, Wang, and Wu [24], and Smith [33]. However, except for Faria and Trofimchuk [13], these results do not give the exponential decay rate of the heteroclinic orbit connecting the equilibria $u \equiv 0$ and $u \equiv K$ at minus infinity. At the same time, the results in [13] are not directly applicable (see the condition (A1) of [13]) and do not ensure the monotonicity of the heteroclinic orbit.

Define

$$\Lambda(\lambda) = \lambda + d - b'(0)\, e^{-\lambda\tau};$$

then it is easy to prove that the equation $\Lambda(\lambda) = 0$ has one and only one real root $\lambda_* > 0$ such that $\Lambda(\lambda) < 0$ for any $\lambda < \lambda_*$ and $\Lambda(\lambda) > 0$ for any $\lambda > \lambda_*$.

Define an operator $S : C(\mathbb{R}, [0, K]) \to C(\mathbb{R}, [0, K])$ by

$$S(u)(t) = e^{-dt}\int_{-\infty}^{t} e^{ds} b(u(s-\tau))\, ds \text{ for any } u \in C(\mathbb{R}, [0, K]).$$

PROPOSITION 4.1.
(i) *If* $u \in C(\mathbb{R}, [0, K])$, *then* $S(u) \in C^1(\mathbb{R}, [0, K])$.
(ii) *For any* $u, v \in C(\mathbb{R}, [0, K])$ *with* $u \le v$, $S(u) \le S(v)$.
(iii) *For any* $u \in C(\mathbb{R}, [0, K])$, *if* $u(\cdot)$ *is increasing in* $\mathbb{R}$, *then so is* $S(u)(\cdot)$.
Let $b''_{\max} = \max_{u\in[0,K]} |b''(u)|$. Define

$$\overline{u}(t) = K \min\left\{e^{\lambda_* t}, 1\right\} \text{ and } \underline{u}(t) = \max\left\{Ke^{\lambda_* t}\left(1 - Me^{\epsilon t}\right), 0\right\},$$

where $\epsilon \in (0, \lambda_*)$ and $M > 1$ with

$$1 - \frac{(d+\lambda_*)\, e^{-\epsilon\tau}}{d+\lambda_*+\epsilon} - \frac{(d+\lambda_*)\, Ke^{-\lambda_*\tau} b''_{\max}}{Mb'(0)\,(d+2\lambda_*)} > 0.$$

LEMMA 4.2. *For any* $t \in \mathbb{R}$, $S(\overline{u})(t) \le \overline{u}(t)$ *and* $\underline{u}(t) \le S(\underline{u})(t)$.
Proof. First, we prove $S(\overline{u})(t) \le \overline{u}(t)$. When $t \ge 0$, $\overline{u}(t) = K$. Therefore,

$$\begin{aligned} S(\overline{u})(t) &= e^{-dt}\int_{-\infty}^{t} e^{ds} b(\overline{u}(s-\tau))\, ds \\ &\le e^{-dt}\int_{-\infty}^{t} e^{ds} b(K)\, ds = dKe^{-dt}\int_{-\infty}^{t} e^{ds} ds = K = \overline{u}(t). \end{aligned}$$

When $t < 0$, $\overline{u}(t) = Ke^{\lambda_* t}$. Noting that $d + \lambda_* = b'(0)e^{-\lambda_*\tau}$, we have

$$\begin{aligned} S(\overline{u})(t) &= e^{-dt}\int_{-\infty}^{t} e^{ds} b(\overline{u}(s-\tau))\, ds \le e^{-dt}\int_{-\infty}^{t} e^{ds} b'(0)\overline{u}(s-\tau)\, ds \\ &\le b'(0)\, Ke^{-dt}\int_{-\infty}^{t} e^{ds} e^{\lambda_*(s-\tau)} ds = \frac{b'(0)\, e^{-\lambda_*\tau} K}{d+\lambda_*} e^{\lambda_* t} = \overline{u}(t). \end{aligned}$$

Now we prove $\underline{u}(t) \le S(\underline{u})(t)$. Let $t_0 = \frac{1}{\epsilon}\ln\frac{1}{M} < 0$ such that $1 - Me^{\epsilon t_0} = 0$. When $t \ge t_0$, $\underline{u}(t) = 0$, and, hence, $\underline{u}(t) \le S(\underline{u})(t)$. When $t < t_0$, $\underline{u}(t) = Ke^{\lambda_* t}(1 - Me^{\epsilon t}) \le Ke^{\lambda_* t}$. In this case, we have

$$\begin{aligned} S(\underline{u})(t) &= e^{-dt}\int_{-\infty}^{t} e^{ds} b(\underline{u}(s-\tau))\, ds \\ &\ge e^{-dt}\int_{-\infty}^{t} e^{ds}\left[b'(0)\,\underline{u}(s-\tau) - b''_{\max}\underline{u}^2(s-\tau)\right] ds \\ &\ge e^{-dt}\int_{-\infty}^{t} e^{ds}\left[b'(0)\, Ke^{\lambda_*(s-\tau)}\left(1 - Me^{\epsilon(s-\tau)}\right) - b''_{\max}K^2 e^{2\lambda_*(s-\tau)}\right] ds \\ &= \frac{b'(0)\, Ke^{-\lambda_*\tau}}{d+\lambda_*} e^{\lambda_* t} - \frac{Mb'(0)\, Ke^{-(\lambda_*+\epsilon)\tau}}{d+\lambda_*+\epsilon} e^{(\lambda_*+\epsilon)t} - \frac{b''_{\max}K^2 e^{-2\lambda_*\tau}}{d+2\lambda_*} e^{2\lambda_* t} \\ &\ge \underline{u}(t) + \frac{Mb'(0)\, Ke^{-\lambda_*\tau}}{d+\lambda_*}\left[1 - \frac{(d+\lambda_*)\, e^{-\epsilon\tau}}{d+\lambda_*+\epsilon} - \frac{(d+\lambda_*)\, Ke^{-\lambda_*\tau} b''_{\max}}{Mb'(0)\,(d+2\lambda_*)}\right] e^{(\lambda_*+\epsilon)t} \\ &\ge \underline{u}(t). \end{aligned}$$

The proof is complete. □

THEOREM 4.3. *There exists a heteroclinic solution* $\Gamma(t)$ *of* (4.1)*, which is increasing on* $\mathbb{R}$ *and satisfies* $\lim_{t\to-\infty} e^{-\lambda_* t}\Gamma(t) = K$*,* $\Gamma(+\infty) = K$*,* $\Gamma(t) \le Ke^{\lambda_* t}$*, and* $\Gamma'(t) > 0$ *for every* $t \in \mathbb{R}$*.*

Proof. By an argument similar to that of [42, Theorem 3.1], we can get a nondecreasing solution $\Gamma(t)$ which meets the theorem except $\Gamma'(t) > 0$ for any $t \in \mathbb{R}$. Since $\Gamma'(t)$ satisfies

$$\Gamma''(t) = -d\Gamma'(t) + b'\left(\Gamma(t-\tau)\right)\Gamma'(t-\tau) \quad \forall t \in \mathbb{R},$$

we have

$$\Gamma'(t) = e^{-d(t-s)}\Gamma'(s) + \int_s^t e^{-d(t-r)} b'\left(\Gamma(t-\tau)\right)\Gamma'(r-\tau)dr \quad \text{for any} \quad s < t.$$

Note that $\Gamma'(t) \ge 0$ for any $t \in \mathbb{R}$. Then it is easy to see that if $\Gamma'(t_0) > 0$ for some $t_0 \in \mathbb{R}$, then $\Gamma'(t) > 0$ for all $t > t_0$. In view of $\lim_{t\to-\infty} e^{-\lambda_* t}\Gamma(t) = K$, we know that there exists a sequence $\{t_i\}$ with $t_i \to -\infty$ as $i \to +\infty$ such that $\Gamma(t_i) > 0$ for any $i \in \mathbb{N}$. Hence, we conclude $\Gamma'(t) > 0$ for any $t \in \mathbb{R}$. This completes the proof. □

5. Proof of Theorem 1.1. In this section, we prove Theorem 1.1.

LEMMA 5.1. *Suppose that* $u(t;\varphi) = \{u_n(t;\varphi)\}_{n\in\mathbb{Z}}$ *is a solution of* (1.1) *with initial value* $\varphi = \{\varphi_n\}_{n\in\mathbb{Z}}$ *with* $\varphi_n \in C([-\tau, 0], [0, K])$*; then there exists a positive constant* $M_* > 0$ *such that for any* $\varphi = \{\varphi_n\}_{n\in\mathbb{Z}}$ *with* $\varphi_n \in C([-\tau, 0], [0, K])$ *and* $t > \tau$*,* $|u_n'(t;\varphi)| \le M_*$ *and* $|u_n''(t;\varphi)| \le M_*$*.*

Proof. Denote $u_n(t;\varphi)$ by $u_n(t)$. Let $M' = 2DK + 2dK$. It is easy to see that $|u_n'(t;\varphi)| \le M'$ for any $t > 0$. For $t > \tau$, there is

$$\begin{aligned} u_n''(t) = {} & D\sum_{i\in\mathbb{Z}\setminus\{0\}} I(i)\left[u_{n-i}'(t) - u_n'(t)\right] \\ & - du_n'(t) + \sum_{i\in\mathbb{Z}} J(i)\, b'\left(u_{n-i}(t-\tau)\right) u_{n-i}'(t-\tau). \end{aligned}$$

Set $M'' = 2DM' + dM' + M'b'(0)$. Then $|u_n''(t;\varphi)| \le M''$. Note that M' and M'' are independent of φ and $t > \tau$. Take $M_* = \max\{M', M''\}$. This completes the proof. □

LEMMA 5.2. *Let* $u^k(t;\varphi^k) = \{u_n^k(t;\varphi^k)\}_{n\in\mathbb{Z}}$ *be a solution of the following initial value problem:*

$$\begin{cases} \frac{d}{dt}u_n^k(t) = D\sum_{i\in\mathbb{Z}\setminus\{0\}} I(i)\left[u_{n-i}(t) - u_n(t)\right] \\ \qquad\qquad -du_n(t) + \sum_{i\in\mathbb{Z}} J(i)\, b\left(u_{n-i}(t-\tau)\right), \ t > 0, \\ u_n^k(t) = \varphi_n^k(t), \ t \in [-\tau, 0], \end{cases}$$

where

$$\begin{aligned} \varphi_n^k(t) = \max\Big\{ & \max_{1\le i\le m} \phi_{c_i}\left(n + c_i(t-k) + \theta_i\right), \\ & \max_{1\le j\le l} \phi_{c_j'}\left(-n + c_j'(t-k) + \theta_j'\right), \chi\Gamma\left((t-k) + \theta_0\right)\Big\}. \end{aligned}$$

Then $v^k\left(t;\varphi^k\right)=\left\{u_n^k\left(t+k;\varphi^k\right)\right\}_{n\in\mathbb{Z}}$ *satisfies*

$$\begin{aligned}\limsup_{t>-k,k\to+\infty} v_n^k(t)\le{}&\phi_{c_i}(n+c_it+\theta_i)+\chi Ke^{\lambda_*(t+\theta_0)}+\sum_{1\le j\le m,j\ne i}A_{c_j}e^{\lambda_1(c_j)(n+c_jt+\theta_j)}\\&+\sum_{1\le j\le l}A_{c'_j}e^{\lambda_1\left(c'_j\right)\left(-n+c'_jt+\theta'_j\right)}\qquad \text{for } 1\le i\le m,\end{aligned}\tag{5.1}$$

$$\begin{aligned}\limsup_{t>-k,k\to+\infty} v_n^k(t)\le{}&\phi_{c'_i}(-n+c'_it+\theta'_i)+\chi Ke^{\lambda_*(t+\theta_0)}+\sum_{1\le j\le m}A_{c_j}e^{\lambda_1(c_j)(n+c_jt+\theta_j)}\\&+\sum_{1\le j\le l,j\ne i}A_{c'_j}e^{\lambda_1\left(c'_j\right)\left(-n+c'_jt+\theta'_j\right)}\qquad \text{for } 1\le i\le l,\end{aligned}\tag{5.2}$$

$$\begin{aligned}\limsup_{t>-k,k\to+\infty} v_n^k(t)\le{}&\chi\Gamma(t+\theta_0)+\sum_{1\le j\le m}A_{c_j}e^{\lambda_1(c_j)(n+c_jt+\theta_j)}\\&+\sum_{1\le j\le l}A_{c'_j}e^{\lambda_1\left(c'_j\right)\left(-n+c'_jt+\theta'_j\right)}.\end{aligned}\tag{5.3}$$

Proof. We prove only (5.1), because the proofs of (5.2) and (5.3) are similar to that of (5.1). Assume $m\ge 1$. Consider $i=1$. Let

$$w_n^k(t)=u_n^k(t)-\phi_{c_1}(n+c_1(t-k)+\theta_1).$$

Then $w_n^k(t)$ satisfies

$$\begin{aligned}\frac{d}{dt}w_n^k(t)&=D\sum_{i\in\mathbb{Z}\setminus\{0\}}I(i)\left[w_{n-i}(t)-w_n(t)\right]-dv_n(t)+\sum_{i\in\mathbb{Z}}J(i)\,b\left(u_{n-i}(t-\tau)\right)\\&\quad-\sum_{i\in\mathbb{Z}}J(i)\,b\left(\phi_{c_1}(n-i+c_1(t-\tau)+\theta_1)\right)\\&\le D\sum_{i\in\mathbb{Z}\setminus\{0\}}I(i)\left[w_{n-i}(t)-w_n(t)\right]-dw_n(t)+b'(0)\sum_{i\in\mathbb{Z}}J(i)\,w_{n-i}(t-\tau).\end{aligned}$$

Since

$$\chi Ke^{\lambda_*((t-k)+\theta_0)}+\sum_{2\le j\le m}A_{c_j}e^{\lambda_1(c_j)(n+c_j(t-k)+\theta_j)}+\sum_{1\le j\le l}A_{c'_j}e^{\lambda_1\left(c'_j\right)\left(-n+c'_j(t-k)+\theta'_j\right)}$$

is a solution of (3.3) with

$$\begin{aligned}\varphi_n(s)={}&\chi Ke^{\lambda_*((s-k)+\theta_0)}+\sum_{2\le j\le m}A_{c_j}e^{\lambda_1(c_j)(n+c_j(s-k)+\theta_j)}\\&+\sum_{1\le j\le l}A_{c'_j}e^{\lambda_1\left(c'_j\right)\left(-n+c'_j(s-k)+\theta'_j\right)}\end{aligned}$$

for any $s \in [-\tau, 0]$ and $n \in \mathbb{Z}$, then by Theorem 3.4 and

$$\begin{aligned}
&\chi K e^{\lambda_*((s-k)+\theta_0)} + \sum_{2\le j\le m} A_{c_j} e^{\lambda_1(c_j)(n+c_j(s-k)+\theta_j)} \\
&\quad + \sum_{1\le j\le l} A_{c'_j} e^{\lambda_1(c'_j)\left(-n+c'_j(s-k)+\theta'_j\right)} \\
&\ge \max\left\{ \max_{1\le i\le m} \phi_{c_i}\left(n + c_i\left(t-k\right) + \theta_i\right), \right. \\
&\qquad \left. \max_{1\le j\le l} \phi_{c'_j}\left(-n + c'_j\left(t-k\right) + \theta'_j\right), \chi\Gamma\left(\left(t-k\right)+\theta_0\right)\right\} \\
&\quad - \phi_{c_1}\left(n + c_1(s-k) + \theta_1\right)
\end{aligned}$$

for any $s \in [-\tau, 0]$ and $n \in \mathbb{Z}$, we have

$$\begin{aligned}
w_n^k\left(t\right) &\le \chi K e^{\lambda_*((t-k)+\theta_0)} + \sum_{2\le j\le m} A_{c_j} e^{\lambda_1(c_j)(n+c_j(t-k)+\theta_j)} \\
&\quad + \sum_{1\le j\le l} A_{c'_j} e^{\lambda_1(c'_j)\left(-n+c'_j(t-k)+\theta'_j\right)} \quad \text{for any } t > 0.
\end{aligned}$$

That is,

$$\begin{aligned}
u_n^k\left(t\right) &\le \phi_{c_1}\left(n + c_1(t-k) + \theta_1\right) + \chi K e^{\lambda_*((t-k)+\theta_0)} + \sum_{2\le j\le m} A_{c_j} e^{\lambda_1(c_j)(n+c_j(t-k)+\theta_j)} \\
&\quad + \sum_{1\le j\le l} A_{c'_j} e^{\lambda_1(c'_j)\left(-n+c'_j(t-k)+\theta'_j\right)} \quad \text{for any } t > 0.
\end{aligned}$$

By the arbitrariness of $k \in \mathbb{N}$, we have that (5.1) holds.

When $m = 0$, the inequality (5.1) reduces to the following:

$$\limsup_{t>-k, k\to+\infty} v_n^k\left(t\right) \le \chi K e^{\lambda_*(t+\theta_0)} + \sum_{1\le j\le l} A_{c'_j} e^{\lambda_1(c'_j)\left(-n+c'_j t+\theta'_j\right)},$$

which holds obviously. This completes the proof. □

Proof of Theorem 1.1. Define $v^k(t) = \{v_n^k(t)\}_{n\in\mathbb{Z}}$ with $v_n^k(t) := u_n(t+k;\psi^k)$ for any $(n,t) \in \mathbb{Z}\times[-\tau-k, +\infty)$, where

$$\begin{aligned}
\psi^k &= \left\{\psi_n^k\left(s\right)\right\}_{k\in\mathbb{Z}}, \\
\psi_n^k\left(s\right) &= \max\left\{ \max_{1\le i\le m} \phi_{c_i}\left(n + c_i\left(s-k\right) + \theta_i\right), \right. \\
&\qquad \left. \max_{1\le j\le l} \phi_{c'_j}\left(-n + c'_j\left(s-k\right) + \theta'_j\right), \chi\Gamma\left(\left(s-k\right)+\theta_0\right)\right\} < K
\end{aligned}$$

for any $(n,s) \in \mathbb{Z}\times[-\tau, 0]$. Note that

$$\begin{aligned}
&\max\left\{ \max_{1\le i\le m} \phi_{c_i}\left(n + c_i t + \theta_i\right), \max_{1\le j\le l} \phi_{c'_j}\left(-n + c'_j t + \theta'_j\right), \chi\Gamma\left(t+\theta_0\right)\right\} \\
&\quad \le v_n^k\left(t\right) \le v_n^{k+1}\left(t\right) \le \min\left\{K, \vartheta_m^+(n,t), \vartheta_l^-(n,t), \vartheta^0(n,t)\right\} \qquad (5.4)
\end{aligned}$$

for any $(n,t) \in \mathbb{Z}\times[-\tau-k, +\infty)$. From Lemma 5.1 and by a diagonal extraction process, there exists a subsequence $\{v^{k_i}(t) = \{v_n^{k_i}(t)\}_{n\in\mathbb{Z}} : i \in \mathbb{N}\}$ such that $v^{k_i}(t)$

converges to a function $\Phi(t)=\{\Phi_n(t)\}_{n\in\mathbb{Z}}$ in $\mathcal{T}$; that is, for any compact set $S\subset\mathbb{Z}\times\mathbb{R}$, $v_n^{k_i}(t)$ and $\frac{d}{dt}v_n^{k_i}(t)$ converge uniformly in $(n,t)\in S$ to $\Phi_n(t)$ and $\frac{d}{dt}\Phi_n(t)$, respectively. In view of $v_n^k(t)\le v_n^{k+1}(t)$ for any $t>-k$, we have $\lim_{k\to+\infty}v_n^k(t)=\Phi_n(t)$ for any $(n,t)\in\mathbb{Z}\times\mathbb{R}$. The limit function is unique, whence all of the functions $v^k(t)$ converge to the function $\Phi(t)$ in $\mathcal{T}$ as $k\to+\infty$. Since $u^{k_i}(t)=\{u_n^{k_i}(t)\}_{n\in\mathbb{Z}}$ satisfies (1.1), the limit function $\Phi(t)=\{\Phi_n(t)\}_{n\in\mathbb{Z}}$ is an entire solution of (1.1). In particular, it follows from (5.4) that (1.8) holds on $(n,t)\in\mathbb{Z}\times\mathbb{R}$.

Now we show (i); that is, $\frac{d}{dt}\Phi_n(t)>0$ on $\mathbb{R}$ for every $n\in\mathbb{Z}$. Since

$$\begin{aligned}\psi_n^k(s)&=\max\left\{\max_{1\le i\le m}\phi_{c_i}\left(n+c_i\left(s-k\right)+\theta_i\right),\right.\\&\qquad\left.\max_{1\le j\le l}\phi_{c_j'}\left(-n+c_j'\left(s-k\right)+\theta_j'\right),\chi\Gamma\left(\left(s-k\right)+\theta_0\right)\right\}\\&\le\max\left\{\max_{1\le i\le m}\phi_{c_i}\left(n+c_i\left(s+\varepsilon-k\right)+\theta_i\right),\right.\\&\qquad\left.\max_{1\le j\le l}\phi_{c_j'}\left(-n+c_j'\left(s+\varepsilon-k\right)+\theta_j'\right),\chi\Gamma\left(\left(s+\varepsilon-k\right)+\theta_0\right)\right\}\\&=\psi_n^k\left(s+\varepsilon\right)\end{aligned}$$

for any $\varepsilon>0$, $s\in[-\tau,0]$, and $n\in\mathbb{Z}$, we have $u_n^k(t;\psi^k(\cdot))\le u_n^k(t;\psi^k(\cdot+\varepsilon))$ for any $(n,t)\in\mathbb{Z}\times[-\tau,+\infty)$. On the other hand, $\psi_n^k(s+\varepsilon)\le u_n^k(s+\varepsilon;\psi^k(\cdot))$ for any $\varepsilon>0$, $s\in[-\tau,0]$, and $n\in\mathbb{Z}$, and, hence,

$$u_n^k\left(t;\psi^k\left(\cdot\right)\right)\le u_n^k\left(t;u_n^k\left(\cdot+\varepsilon;\psi^k\left(\cdot\right)\right)\right)=u_n^k\left(t+\varepsilon;\psi^k\left(\cdot\right)\right)$$

for any $(n,t)\in\mathbb{Z}\times[-\tau,+\infty)$. Thus, it follows from the arbitrariness of $\epsilon>0$; that $u_n^k(t)$ is increasing on t; that is, $v^k(t)$ is increasing on t. Therefore, $\Phi_n'(t)\ge0$ on $\mathbb{R}$ for every $n\in\mathbb{Z}$. Since $\Phi_n'(t)$ satisfies

$$\begin{aligned}\Phi_n''(t)&=D\sum_{i\in\mathbb{Z}\setminus\{0\}}I(i)\left[\Phi_{n-i}'(t)-\Phi_n'(t)\right]\\&\quad-d\Phi_n'(t)+\sum_{i\in\mathbb{Z}}J(i)\,b'\left(\Phi_{n-i}(t-\tau)\right)\Phi_{n-i}'(t-\tau),\end{aligned}\tag{5.5}$$

we have that $\Phi_n'(t)$ satisfies

$$\Phi_n'(t)=\Phi_n'(s)\,e^{-(D+d)(t-s)}+\int_s^t e^{-(D+d)(t-r)}R_n(\Phi)(r)\,dr\quad\text{for any } s<t,$$

where $R_n(\Phi)(t)=D\sum_{i\in\mathbb{Z}\setminus\{0\}}I(i)\Phi_{n-i}'(t)+\sum_{i\in\mathbb{Z}}J(i)b'(\Phi_{n-i}(t-\tau))\Phi_{n-i}'(t-\tau)\ge0$. Obviously, for each $n\in\mathbb{Z}$, if there exists $t_0\in\mathbb{R}$ such that $\Phi_n'(t_0)>0$, then $\Phi_n'(t)>0$ for any $t>t_0$. Therefore, there must be $\Phi_n'(t)>0$ for any $(n,t)\in\mathbb{Z}\times\mathbb{R}$. We argue by a contradiction. In fact, assume that, for some $n_1\in\mathbb{Z}$, there is t_1 such that $\Phi_{n_1}'(t_1)=0$ and, hence, then $\Phi_{n_1}'(t)=0$ for any $t\le t_1$, which implies that $\lim_{t\to-\infty}\Phi_{n_1}(t)=\Phi_{n_1}(t_1)>0$. But following from (1.8), we have $\lim_{t\to-\infty}\Phi_{n_1}(t)=0$, which yields a contradiction.

Now we prove (vii). For the sake of convenience, we denote

$$\Phi_{c_1,\dots,c_m;c_1',\dots,c_l';\theta_1,\dots,\theta_m;\theta_1',\dots,\theta_l';\theta_0}(t)$$

by $\Phi(t;\theta_0)$ and $\Phi_{c_1,\dots,c_m;c_1',\dots,c_l';\theta_1,\dots,\theta_m;\theta_1',\dots,\theta_l'}(t)$ by $\Phi(t;-\infty)$. For $\chi\in\{0,1\}$, let

$$\begin{aligned}
\psi^k(t)_\chi &= \left\{\psi_n^k(s)_\chi\right\}_{k\in\mathbb{Z}},\\
\psi_n^k(s)_\chi &= \max\left\{\max_{1\le i\le m}\phi_{c_i}\left(n+c_i(s-k)+\theta_i\right),\right.\\
&\qquad \left.\max_{1\le j\le l}\phi_{c_j'}\left(-n+c_j'(s-k)+\theta_j'\right),\chi\Gamma\left((s-k)+\theta_0\right)\right\},
\end{aligned}$$

and $v^k(t)_\chi=\{v_n^k(t)_\chi\}_{n\in\mathbb{Z}}$ with $v_n^k(t)_\chi := u_n(t+k;\psi^k(\cdot)_\chi)$ for any $(n,t)\in\mathbb{Z}\times[-\tau-k,+\infty)$. Set $\bar v^k(t)=v^k(t)_1-v^k(t)_0=\{v_n^k(t)_1-v_n^k(t)_0\}_{n\in\mathbb{Z}}$. Then $\bar v^k(t)$ satisfies $0\le\bar v^k(t)\le K$ for any $t\in[-\tau-k,+\infty)$ and

$$\frac{d}{dt}\bar v_n^k(t)\le D\sum_{i\in\mathbb{Z}\setminus\{0\}}I(i)\left[\bar v_{n-i}^k(t)-\bar v_n^k(t)\right]-d\bar v_n^k(t)+b'(0)\sum_{i\in\mathbb{Z}}J(i)\,\bar v_{n-i}^k(t-\tau).$$

Noting that $\bar v_n^k(s)\le Ke^{\lambda_*(s+\theta_0)}$ for any $s\in[-\tau-k,-k]$ and $w_n^k(t)=Ke^{\lambda_*(t+\theta_0)}$ satisfies

$$\frac{d}{dt}w_n^k(t)=D\sum_{i\in\mathbb{Z}\setminus\{0\}}I(i)\left[w_{n-i}^k(t)-w_n^k(t)\right]-dw_n^k(t)+b'(0)\sum_{i\in\mathbb{Z}}J(i)\,w_{n-i}^k(t-\tau)$$

for any $t\in[-\tau-k,+\infty)$, it follows from Theorem 3.4 that $0\le\bar v_n^k(t)\le Ke^{\lambda_*(t+\theta_0)}$ for any $(n,t)\in\mathbb{Z}\times[-\tau-k,+\infty)$ and $k\in\mathbb{N}$. Note that $\lim_{k\to+\infty}v_n^k(t)_0=\Phi_n(t;-\infty)$ and $\lim_{k\to+\infty}v_n^k(t)_1=\Phi_n(t;\theta_0)$ for any $(n,t)\in\mathbb{Z}\times\mathbb{R}$. Therefore, there must be $0<\Phi_n(t;\theta_0)-\Phi_n(t;-\infty)\le Ke^{\lambda_*(t+\theta_0)}$ for all $(n,t)\in\mathbb{Z}\times\mathbb{R}$, which implies that $\Phi(t;\theta_0)$ converges uniformly on $(n,t)\in\mathbb{Z}\times(-\infty,a]$ to $\Phi(t;-\infty)$ as $\theta_0\to-\infty$ for any $a\in\mathbb{R}$. For any sequence $\theta_0^k\to-\infty$ $(k\to+\infty)$, the functions $\Phi(t;\theta_0^k)$ converge to a solution of (1.1) in $\mathcal{T}$, which turns out to be $\Phi(t;-\infty)$. The limit does not depend on the sequence θ_0^k, whence all of the functions $\Phi(t;\theta_0)$ converge to $\Phi(t;-\infty)$ in $\mathcal{T}$ as $\theta_0\to-\infty$. The assertion as $\theta_0\to+\infty$ is obvious.

We next prove (viii). Assume $\chi=1$. Similarly to that in (vii), we denote

$$\Phi_{c_1,\dots,c_m;c_1',\dots,c_l';\theta_1,\dots,\theta_m;\theta_1',\dots,\theta_l';\theta_0}(t)$$

by $\Phi(t)_{\theta_i}$ and $\Phi_{c_1,\dots,c_{i-1},c_{i+1},\dots,c_m;c_1',\dots,c_l';\theta_1,\dots,\theta_{i-1},\theta_{i+1},\dots,\theta_m;\theta_1',\dots,\theta_l';\theta_0}(t)$ by $\Phi(t)_\infty$. Set

$$\begin{aligned}
\psi^k(t)_{\theta_i} &= \left\{\psi_n^k(s)_{\theta_i}\right\}_{k\in\mathbb{Z}},\\
\psi_n^k(s)_{\theta_i} &= \max\left\{\max_{1\le j\le m}\phi_{c_j}\left(n+c_j(s-k)+\theta_j\right),\right.\\
&\qquad \left.\max_{1\le j\le l}\phi_{c_j'}\left(-n+c_j'(s-k)+\theta_j'\right),\Gamma\left((s-k)+\theta_0\right)\right\},
\end{aligned}$$

and $v^k(t)_{\theta_i}=\{v_n^k(t)_{\theta_i}\}_{n\in\mathbb{Z}}$ with $v_n^k(t)_{\theta_i} := u_n(t+k;\psi^k(\cdot)_{\theta_i})$ for any $(n,t)\in\mathbb{Z}\times[-\tau-k,+\infty)$. Take

$$\begin{aligned}
\psi^k(t)_\infty &= \left\{\psi_n^k(s)_\infty\right\}_{k\in\mathbb{Z}},\\
\psi_n^k(s)_\infty &= \max\left\{\max_{j\in\{1,\dots,i-1,i+1,\dots,m\}}\phi_{c_j}\left(n+c_j(s-k)+\theta_j\right),\right.\\
&\qquad \left.\max_{1\le j\le l}\phi_{c_j'}\left(-n+c_j'(s-k)+\theta_j'\right),\Gamma\left((s-k)+\theta_0\right)\right\},
\end{aligned}$$

and $v^k(t)_\infty = \{v_n^k(t)_\infty\}_{n\in\mathbb{Z}}$ with $v_n^k(t)_\infty := u_n(t+k;\psi^k(\cdot)_\infty)$ for any $(n,t)\in\mathbb{Z}\times[-\tau-k,+\infty)$. Set $\hat{v}^k(t) = v^k(t)_{\theta_i} - v^k(t)_\infty = \{v_n^k(t)_{\theta_i} - v_n^k(t)_\infty\}_{n\in\mathbb{Z}}$. Then $\hat{v}^k(t)$ satisfies $0\le \hat{v}_n^k(t)\le K$ for any $(n,t)\in\mathbb{Z}\times[-\tau-k,+\infty)$ and

$$\frac{d}{dt}\hat{v}_n^k(t) \le D\sum_{i\in\mathbb{Z}\setminus\{0\}} I(i)\left[\hat{v}_{n-i}^k(t) - \hat{v}_n^k(t)\right] - d\hat{v}_n^k(t) + b'(0)\sum_{i\in\mathbb{Z}} J(i)\,\hat{v}_{n-i}^k(t-\tau).$$

Noting that $\hat{v}_n^k(s)\le \phi_{c_i}(n+c_is+\theta_i)\le A_{c_i}e^{\lambda_1(c_i)(n+c_is+\theta_i)}$ for any $s\in[-\tau-k,-k]$ and that $\bar{w}_n^k(t) = A_{c_i}e^{\lambda_1(c_i)(n+c_it+\theta_i)}$ satisfies

$$\frac{d}{dt}\bar{w}_n^k(t) = D\sum_{i\in\mathbb{Z}\setminus\{0\}} I(i)\left[\bar{w}_{n-i}^k(t) - \bar{w}_n^k(t)\right] - d\bar{w}_n^k(t) + b'(0)\sum_{i\in\mathbb{Z}} J(i)\,\bar{w}_{n-i}^k(t-\tau)$$

for any $t\in[-\tau-k,+\infty)$, it follows from Theorem 3.4 that $0\le\hat{v}_n^k(t)\le A_{c_i}e^{\lambda_1(c_i)(n+c_it+\theta_i)}$ for any $(n,t)\in\mathbb{Z}\times[-\tau-k,+\infty)$ and $k\in\mathbb{N}$. Since $\lim_{k\to+\infty}v^k(t)_{\theta_i} = \Phi(t)_{\theta_i}$ and $\lim_{k\to+\infty}v^k(t)_\infty = \Phi(t)_\infty$, we have $0 < \Phi_n(t)_{\theta_i} - \Phi_n(t)_\infty \le A_{c_i}e^{\lambda_1(c_i)(n+c_it+\theta_i)}$ for all $(n,t)\in\mathbb{Z}\times\mathbb{R}$, which implies that $\Phi(t)_{\theta_i}$ converges uniformly on $(n,t)\in\{n: n\le N_0, n\in\mathbb{Z}\}\times(-\infty,a]$ to $\Phi(t)_\infty$ as $\theta_i\to-\infty$ for any $N_0\in\mathbb{Z}$ and $a\in\mathbb{R}$. For any sequence $\theta_i^k\to-\infty$ $(k\to+\infty)$, the functions $\Phi(t)_{\theta_i^k}$ converge to a solution of (1.1), which must be $\Phi(t)_\infty$. Since the limit is independent of the sequence θ_i^k, all of the functions $\Phi(t)_{\theta_i}$ converge to $\Phi(t)_\infty$ in $\mathcal{T}$ as $\theta_i\to-\infty$. The assertion as $\theta_j'\to-\infty$ and the case $\chi=0$ can be proved similarly.

Using the inequality (1.8), we can prove (ii)–(vi) and (ix) of Theorem 1.1. □

6. Proof of Theorem 1.2. In this section, we prove Theorem 1.2. We prove only the continuous dependence of the entire solution on the parameters c, c', θ, θ', and θ_0 and the uniqueness of the entire solutions satisfies (1.9). Other conclusions follow immediately from Theorem 1.1.

Consider (1.4) or

$$u_n'(t) = \frac{D}{2}\left[u_{n+1}+u_{n-1}-2u_n\right] + f\left(u_n(t)\right), \tag{6.1}$$

where f satisfies the conditions given after (1.4). Let $\phi_c(n+ct)$ be a traveling wave front of (6.1) with wave speed $c>c^*$. As done in section 1, we normalize $\phi_c(n+ct)$ so that $\phi(0)=\frac{1}{2}$. Then the functions $\phi_c(z)$ are continuous with respect to $c\in(c^*,+\infty)$ in the norms $C^1_{loc}(\mathbb{R})$ (see [20, p. 1267] for the definition of these norms). Indeed, if $c_l\to c\in(c^*,+\infty)$, then by the unique boundedness of $|\phi_{c_l}'(z)|$ and $|\phi_{c_l}''(z)|$ in $z\in\mathbb{R}$ on $l\in\mathbb{N}$ and by a diagonal extraction process, there exists a subsequence c_{l_i} such that $\phi_{c_{l_i}}\to\phi$ in $C^1_{loc}(\mathbb{R})$, where ϕ is a solution of

$$c\phi'(z) = \frac{D}{2}\left[\phi(z+1)+\phi(z-1)-2\phi(z)\right] + f\left(\phi(z)\right) \text{ in } z\in\mathbb{R}.$$

Obviously, ϕ is nondecreasing in $\mathbb{R}$ and is not a constant and $\phi(0)=\frac{1}{2}$. By the assumptions of f, we have $\phi(-\infty)=0$ and $\phi(+\infty)=1$. Thus, ϕ is a traveling wave front of (6.1) with wave speed c. Following Chen and Guo [5], we have $\phi\equiv\phi_c$. Finally, the whole sequence $\phi_{c_l}\to\phi_{c_0}$ in $C^1_{loc}(\mathbb{R})$ as $l\to+\infty$. In view of [4, 5], we know that α_c and A_c defined by (1.6) and (1.7), respectively, are still valid for (6.1). In addition, there exactly is $\lambda_* = f'(0)$ for (6.1).

LEMMA 6.1. *For* (6.1), $\alpha_c = \lim_{z\to-\infty}\phi_c(z)\,e^{-\lambda_1(c)z}$ *is continuous in* $c\in(c^*,+\infty)$.

Proof. Fix $c_0 \in (c^*, +\infty)$ and let $c_l \to c_0$ as $l \to +\infty$ with $c_l > c^*$ for each $l \in \mathbb{N}$. Then by Chen and Guo [4] (see also Ma, Weng, and Zou [25]), we know that, for each $c \in (c^*, +\infty)$, there exists a unique traveling wave front $\widetilde{\phi}_c$ such that $\widetilde{\phi}_c'(\cdot) > 0$, $\widetilde{\phi}_c(-\infty) = 0$, $\widetilde{\phi}_c(+\infty) = 1$, and

$$\lim_{z \to -\infty} \widetilde{\phi}_c(z)\, e^{-\lambda_1(c)z} = 1.$$

Then we have that $\widetilde{\phi}_{c_l} \to \widetilde{\phi}_{c_0}$ in $C^1_{loc}(\mathbb{R})$ as $l \to +\infty$. In fact, since $c_l \to c_0$ as $l \to +\infty$, then there exist a subsequence c_{l_i} and a function $\widetilde{\phi}$ such that $\widetilde{\phi}_{c_{l_i}} \to \widetilde{\phi}$ in $C^1_{loc}(\mathbb{R})$, where $\widetilde{\phi}$ is nondecreasing and satisfies

$$c\widetilde{\phi}'(z) = \frac{D}{2}\left[\widetilde{\phi}(z+1) + \widetilde{\phi}(z-1) - 2\widetilde{\phi}(z)\right] + f\left(\widetilde{\phi}(z)\right) \text{ in } z \in \mathbb{R}.$$

On the other hand, by Chen and Guo [4], there exist two constants $q > 1$ and $\beta > 1$, independent of c_l, such that

$$e^{\lambda_1(c_l)z} - qe^{\beta\lambda_1(c_l)z} \le \phi_{c_l}(z) \le e^{\lambda_1(c_l)z} + qe^{\beta\lambda_1(c_l)z} \text{ for any } z \in \mathbb{R}$$

and therefore, as $l \to \infty$,

$$e^{\lambda_1(c_0)z} - qe^{\beta\lambda_1(c_0)z} \le \widetilde{\phi}(z) \le e^{\lambda_1(c_0)z} + qe^{\beta\lambda_1(c_0)z} \text{ for any } z \in \mathbb{R},$$

which implies that $\widetilde{\phi}(z)$ is not a constant and satisfies $\widetilde{\phi}(z)e^{-\lambda_1(c_0)z} = 1$. Then it follows from Chen and Guo [5] (see also Ma, Weng, and Zou [25]) that $\widetilde{\phi} \equiv \widetilde{\phi}_{c_0}$. Consequently, the whole sequence $\widetilde{\phi}_{c_l} \to \widetilde{\phi}_{c_0}$ in $C^1_{loc}(\mathbb{R})$ as $l \to +\infty$.

Now let $\widetilde{\phi}_{c_0}(\varsigma_0) = \frac{1}{2}$ and $\widetilde{\phi}_{c_l}(\varsigma_l) = \frac{1}{2}$. Then we have that $\varsigma_l \to \varsigma_0$ as $l \to +\infty$. Assume that this assertion is not true. Take $\varsigma_l \to \overline{\varsigma} \ne \varsigma_0$ as $l \to +\infty$ (up to extraction of some subsequence). If $|\overline{\varsigma}| < \infty$, then, by $\widetilde{\phi}_{c_l} \to \widetilde{\phi}_{c_0}$ in $C^1_{loc}(\mathbb{R})$ as $l \to +\infty$, we have $\widetilde{\phi}_{c_0}(\overline{\varsigma}) = \frac{1}{2} = \widetilde{\phi}_{c_0}(\varsigma_0)$, which is impossible since $\frac{d}{dz}\widetilde{\phi}_{c_0}(z) > 0$ for any $z \in \mathbb{R}$ and $\overline{\varsigma} \ne \varsigma_0$. If $\overline{\varsigma} = +\infty$, then $\widetilde{\phi}_{c_l}(\varsigma_0 + 1) < \widetilde{\phi}_{c_l}(\varsigma_l) = \frac{1}{2}$ for sufficiently large l implies that $\widetilde{\phi}_{c_0}(\varsigma_0 + 1) \le \frac{1}{2} = \widetilde{\phi}_{c_0}(\varsigma_0)$, which is also impossible. Similarly, $\overline{\varsigma} = -\infty$ is impossible. Thus, we conclude that $\varsigma_l \to \varsigma_0$ as $l \to +\infty$.

Again applying Chen and Guo [5], we have that $\phi_{c_0}(\cdot) = \widetilde{\phi}_{c_0}(\varsigma_0 + \cdot)$ and $\phi_{c_l}(\cdot) = \widetilde{\phi}_{c_l}(\varsigma_l + \cdot)$. Since

$$\lim_{z \to -\infty} \phi_{c_0}(z)\, e^{-\lambda_1(c_0)(z+\varsigma_0)} = \lim_{z \to -\infty} \widetilde{\phi}_{c_0}(z + \varsigma_0)\, e^{-\lambda_1(c_0)(z+\varsigma_0)} = 1,$$

we have $\lim_{z \to -\infty} \phi_{c_0}(z)e^{-\lambda_1(c_0)z} = e^{\lambda_1(c_0)\varsigma_0} = \alpha_{c_0}$. Similarly, we have

$$\lim_{z \to -\infty} \phi_{c_l}(z)\, e^{-\lambda_1(c_l)z} = e^{\lambda_1(c_l)\varsigma_l} = \alpha_{c_l}.$$

Finally, there holds $\alpha_{c_l} \to \alpha_{c_0}$ as $l \to +\infty$. This completes the proof. ∎

Recall that

$$A_c = \inf\left\{A > 0 : Ae^{\lambda_1(c)z} \ge \phi_c(z) \text{ in } z \in \mathbb{R}\right\}.$$

LEMMA 6.2. *For* (6.1), A_c *is continuous on* $c \in (c^*, +\infty)$.

Proof. Fix $c_0 \in (c^*, +\infty)$ and let $c_l \to c_0$ as $l \to +\infty$ with $c_l > c^*$ for each $l \in \mathbb{N}$. We prove the theorem by way of contradiction. Assume $A_{c_l} \to A_0 \in \mathbb{R} \cup \{\infty\}$

as $l \to \infty$ (up to extraction of some subsequence) and $A_0 \neq A_{c_0}$. Since $A_{c_l} \geq e^{-\lambda_1(c_l)z}\phi_{c_l}(z)$ for any $z \in \mathbb{R}$, $A_0 \geq e^{-\lambda_1(c_0)z}\phi_{c_0}(z)$ and, hence, $A_0 > A_{c_0}$. Fix $b = \min\{\frac{A_0+A_{c_0}}{2}, A_{c_0}+1\}$. Then there exists $L \in \mathbb{N}$ such that for any $l > L$, $A_{c_l} > b$. On the other hand, since $\alpha_{c_l} \to \alpha_{c_0} \leq A_{c_0}$ and $\lambda_1(c_l) \to \lambda_1(c_0)$, there exists a constant $Z_0 > 0$, independent of c_l, such that $\phi_{c_l}(z)e^{-\lambda_1(c_l)z} \leq b$ for any $|z| > Z_0$. For $z \in [-Z_0, Z_0]$, by $\phi_{c_0}(z)e^{-\lambda_1(c_0)z} \leq A_{c_0}$, $\phi_{c_l}(z) \to \phi_{c_0}(z)$ in $C^1_{loc}(\mathbb{R})$, and the equicontinuity of $e^{-\lambda_1(c_l)z}$ on l, there exists $L' > L$ such that $\phi_{c_l}(z)e^{-\lambda_1(c_l)z} \leq b$ for any $l > L'$ and $z \in [-Z_0, Z_0]$. Therefore, $\phi_{c_l}(z)e^{-\lambda_1(c_l)z} \leq b$ for any $l > L'$ and $z \in \mathbb{R}$, which contradicts $A_{c_l} > b$ for any $l > L$. □

Before proving Theorem 1.2, we first consider the following linear Cauchy problem:

$$\begin{cases} \frac{d}{dt}u_n(t) = \frac{D}{2}\left[u_{n+1}(t) + u_{n-1}(t) - 2u_n(t)\right] + f'(0)\,u_n(t), & t > 0, \\ u_n(0) = u_n^0, \end{cases} \tag{6.2}$$

where $u^0 = \{u_n^0\}_{n\in\mathbb{Z}} \in l^\infty$. By Theorem 3.3, we know that (6.2) admits a unique solution $u(t) = u(t;u^0) = \{u_n(t)\}_{n\in\mathbb{Z}}$ on $t \in [0, +\infty)$. By using the discrete Fourier transformation, we can exactly solve the solution $u(t) = \{u_n(t)\}_{n\in\mathbb{Z}}$ of (6.2) as follows:

$$u_n(t) = \frac{1}{\pi}e^{f'(0)t}\sum_{i=-\infty}^{+\infty} u_i^0 \int_0^\pi \cos\left((i-n)\omega\right) e^{Dt(\cos\omega - 1)}d\omega. \tag{6.3}$$

This formulation is very crucial for the proof of Theorem 1.2.

Proof of Theorem 1.2. We prove only the case $\varrho = \varrho' = \chi = 1$. Consider a sequence

$$(c_k, c_k', \theta_k, \theta_k', \theta_{0,k}) \to (c, c', \theta, \theta', \theta_0) \in (c^*, +\infty)^2 \times \mathbb{R}^3 \quad \text{as} \quad k \to +\infty.$$

For given $(c_k, c_k', \theta_k, \theta_k', \theta_{0,k})$ and $(c, c', \theta, \theta', \theta_0)$, it follows from Theorem 1.1 that there exist entire solutions $\Phi_{c_k;c_k';\theta_k;\theta_k';\theta_{0,k}}(t)$ and $\Phi_{c;c';\theta;\theta';\theta_0}(t)$ of (1.4) satisfying (1.9). For the sake of convenience, set $\Phi^k(t) = \{\Phi_n^k(t)\}_{n\in\mathbb{Z}} = \Phi_{c_k;c_k';\theta_k;\theta_k';\theta_{0,k}}(t)$ and $\Phi(t) = \{\Phi_n(t)\}_{n\in\mathbb{Z}} = \Phi_{c;c';\theta;\theta';\theta_0}(t)$.

Using Lemma 5.1, there exists a function $\widetilde{\Phi}(t) = \{\widetilde{\Phi}_n(t)\}_{n\in\mathbb{Z}}$ such that $\Phi_n^k(t) \to \widetilde{\Phi}_n(t)$ as $k \to \infty$ (up to extraction of some subsequence) in $\mathcal{T}$. In particular, the function $\widetilde{\Phi}(t) = \{\widetilde{\Phi}_n(t)\}_{n\in\mathbb{Z}}$ is also an entire solution of (6.1) (or (1.4)). By passage to the limit $k \to +\infty$ in (1.9), the function $\widetilde{\Phi}(t) = \{\widetilde{\Phi}_n(t)\}_{n\in\mathbb{Z}}$ fulfills the estimates

$$\begin{aligned} &\max\left\{\phi_c(n+ct+\theta), \Gamma(t+\theta_0), \phi_{c'}(-n+c't+\theta')\right\} \\ &\leq \widetilde{\Phi}_n(t) \leq \min\Big\{1, \phi_c(n+ct+\theta) + e^{\lambda_*(t+\theta_0)} + A_{c'}e^{\lambda_1(c')(-n+c't+\theta')}, \\ &\qquad \Gamma(t+\theta_0) + A_c e^{\lambda_1(c)(n+ct+\theta)} + A_{c'}e^{\lambda_1(c')(-n+c't+\theta')}, \\ &\qquad \phi_{c'}(-n+c't+\theta') + e^{\lambda_*(t+\theta_0)} + A_c e^{\lambda_1(c)(n+ct+\theta)}\Big\} \end{aligned} \tag{6.4}$$

for any $(n,t) \in \mathbb{Z}\times\mathbb{R}$.

Let us now prove that $\widetilde{\Phi}_n(t) = \Phi_n(t)$ for any $(n,t) \in \mathbb{Z}\times\mathbb{R}$. Recall that the functions $v^k(t) = \{v_n^k(t)\}_{n\in\mathbb{N}}$, which are solutions of the Cauchy problems

$$\frac{d}{dt}v_n^k(t) = \frac{D}{2}\left[v_{n+1}^k(t) + v_{n-1}^k(t) - 2v_n^k(t)\right] + f\left(v_n^k(t)\right), \; t > -k, \; n \in \mathbb{N},$$

with the initial conditions

$$v_n^k(-k) = v_{n,0}^k = \max\left\{\phi_c(n-ck+\theta), \Gamma(-k+\theta_0), \phi_{c'}(-n-c'k+\theta')\right\},$$

converge to the function $\Phi(t) = \{\Phi_n(t)\}_{n\in\mathbb{N}}$ in $\mathcal{T}$; see the proof of Theorem 1.1. Let us now compare the functions $\widetilde{\Phi}(t)$ to the functions $v^k(t)$ for $t > -k$. Following (6.4), we get that

$$\begin{aligned}(6.5)\qquad &\left|\widetilde{\Phi}_n(-k) - v_n^k(-k)\right| \\ &\le \begin{cases} e^{f'(0)(-k+\theta_0)} + A_{c'}e^{\lambda_1(c')(-n-c'k+\theta')} & \text{if } n \ge n_k^+ + 2, \\ A_c e^{\lambda_1(c)(n-ck+\theta)} + A_{c'}e^{\lambda_1(c')(-n-c'k+\theta')} & \text{if } n_k^- + 1 \le n \le n_k^+ - 1, \\ e^{f'(0)(-k+\theta_0)} + A_c e^{\lambda_1(c)(n-ck+\theta)} & \text{if } n \le n_k^- - 2 \end{cases}\end{aligned}$$

for sufficiently large k, where n_k^- and n_k^+ are two integers defined as follows:

$$\begin{aligned} n_k^- &= \operatorname{int}\left[-c'k + \frac{f'(0)}{\lambda_1(c')}k + \theta' - \frac{1}{\lambda_1(c')}\ln\frac{1}{\alpha_{c'}} - \frac{f'(0)}{\lambda_1(c')}\theta_0\right], \\ n_k^+ &= \operatorname{int}\left[ck - \frac{f'(0)}{\lambda_1(c)}k - \theta + \frac{1}{\lambda_1(c)}\ln\frac{1}{\alpha_c} + \frac{f'(0)}{\lambda_1(c)}\theta_0\right]. \end{aligned}$$

n_k^- is obtained by comparing $\phi_{c'}(-n-c'k+\theta')$ and $\Gamma(-k+\theta_0)$ and using the asymptotic behaviors of $\phi_{c'}$ and Γ. Similarly, we obtain n_k^+.

For any $x \in \mathbb{R}$, define

$$\operatorname{int}[x] = \max\{m : m \in \mathbb{Z}, m \le x\}.$$

Fix $t_0 > -k$. For any $n \in \mathbb{Z}$ and $k \in \mathbb{N}$ with $k > -t_0$, define

$$a_n^k = \frac{1}{2\pi}e^{-D(t_0+k)}\int_0^\pi \cos(nw)\, e^{D(t_0+k)\cos w}dw.$$

By Weng, Huang, and Wu [40], we know that $a_n^k = a_{-n}^k > 0$ for any $n \in \mathbb{Z}$ and $\sum_{n=-\infty}^{+\infty} a_n^k = 1$. Furthermore, for $n > 1$ there is

$$\begin{aligned} a_n^k &= \frac{1}{2\pi}\int_0^\pi \cos(nw)\, e^{D(t_0+k)[\cos w-1]}dw = \frac{1}{2n\pi}\int_0^\pi e^{D(t_0+k)[\cos w-1]}d\sin(nw) \\ &= \frac{D(t_0+k)}{2n\pi}\int_0^\pi \sin(nw)\sin w e^{D(t_0+k)[\cos w-1]}dw \\ &= \frac{D(t_0+k)}{4n\pi}\int_0^\pi \left[\cos((n-1)w) - \cos((n+1)w)\right]e^{D(t_0+k)[\cos w-1]}dw \\ &= \frac{D(t_0+k)}{2n}\left(a_{n-1}^k - a_{n+1}^k\right), \end{aligned}$$

which implies that $a_{n-1}^k > a_{n+1}^k$ for any $n \in \mathbb{N}$. By the symmetry, $a_{-n-1}^k < a_{-n+1}^k$ for any $n \in \mathbb{N}$.

We claim that, for any given $\eta \in (0, +\infty)$, there hold

$$(6.6)\qquad a_{\operatorname{int}[\eta k]}^k \to 0 \quad \text{and} \quad a_{\operatorname{int}[\eta k]-1}^k \to 0 \quad \text{as} \quad k \to +\infty.$$

Consider $a^k_{\mathrm{int}[\eta k]} \to 0$ as $k \to +\infty$. If the assertion is false, then we can assume that $a^k_{\mathrm{int}[\eta k]} \to \delta > 0$ as $k \to +\infty$ (up to extraction of some subsequence). Taking k sufficiently large such that $\frac{\mathrm{int}[\eta k]}{2} > \frac{4}{\delta} + 1$ and $a^k_{\mathrm{int}[\eta k]} > \frac{\delta}{2}$, by $a^k_{n-1} > a^k_{n+1}$ for $n \in \mathbb{N}$, we have $\sum_{n=0}^{\mathrm{int}[\eta k]} a^k_n > 1$, which contradicts the fact $\sum_{n=-\infty}^{+\infty} a^k_n < 1$. Therefore, $\lim_{k\to+\infty} a^k_{\mathrm{int}[\eta k]} = 0$. Similarly, we have $\lim_{k\to+\infty} a^k_{\mathrm{int}[\eta k]-1} = 0$. For any $m > N$,

$$\begin{aligned}
\sum_{n=N\geq 1}^{m} a^k_n &= \frac{D(t_0+k)}{2N}\left(a^k_{N-1} - a^k_{N+1}\right) + \frac{D(t_0+k)}{2(N+1)}\left(a^k_N - a^k_{N+2}\right) \\
&\quad + \frac{D(t_0+k)}{2(N+2)}\left(a^k_{N+1} - a^k_{N+3}\right) \\
&\quad + \cdots + \frac{D(t_0+k)}{2(m-1)}\left(a^k_{m-2} - a^k_m\right) + \frac{D(t_0+k)}{2m}\left(a^k_{m-1} - a^k_{m+1}\right) \\
&= \frac{D(t_0+k)}{2N}a^k_{N-1} + \frac{D(t_0+k)}{2(N+1)}a^k_N - \frac{D(t_0+k)}{N(N+2)}a^k_{N+1} \\
&\quad - \frac{D(t_0+k)}{(N+1)(N+3)}a^k_{N+2} - \cdots - \frac{D(t_0+k)}{(m-3)(m-1)}a^k_{m-2} \\
&\quad - \frac{D(t_0+k)}{(m-2)m}a^k_{m-1} - \frac{D(t_0+k)}{2(m-1)}a^k_m - \frac{D(t_0+k)}{2m}a^k_{m+1} \\
&\leq \frac{D(t_0+k)}{2N}a^k_{N-1} + \frac{D(t_0+k)}{2(N+1)}a^k_N.
\end{aligned}$$

Therefore,

$$\sum_{n=N\geq 1}^{\infty} a^k_n \leq \frac{D(t_0+k)}{2N}a^k_{N-1} + \frac{D(t_0+k)}{2(N+1)}a^k_N.$$

Consequently, we obtain that for any given $\eta > 0$,

$$\sum_{n=\mathrm{int}[\eta k]}^{\infty} a^k_n \to 0 \quad \text{and} \quad \sum_{n=\mathrm{int}[\eta k]+1}^{\infty} a^k_n \to 0 \quad \text{as} \quad k \to +\infty \tag{6.7}$$

due to $\frac{D(t_0+k)}{[\eta k]} \to \frac{D}{\eta}$ as $k \to +\infty$.

Now fix $(n_0, t_0) \in \mathbb{Z} \times \mathbb{R}$; we estimate $\widetilde{\Phi}_{n_0}(t_0) - v^k_{n_0}(t_0)$ as $k \to +\infty$. We compare $\widetilde{\Phi}_n(t) - v^k_n(t)$ with a solution of the linear equation

$$\frac{d}{dt}u_n(t) = \frac{D}{2}\left[u_{n+1}(t) + u_{n-1}(t) - 2u_n(t)\right] + f'(0)\,u_n(t), \quad t > -k,$$

with the initial condition $u_n(-k) = |\widetilde{\Phi}_n(-k) - v^k_n(-k)|$. Using (6.3), we deduce that

$$
\begin{aligned}
&\left|\widetilde{\Phi}_{n_0}(t_0)-v^k_{n_0}(t_0)\right| \\
&\le \frac{1}{\pi}e^{f'(0)(t_0+k)}\sum_{n=-\infty}^{+\infty}\int_0^\pi \cos\left((n-n_0)w\right)e^{D(t_0+k)[\cos w-1]}\left|\widetilde{\Phi}_n(-k)-v^k_n(-k)\right|dw \\
&\le \frac{1}{\pi}e^{f'(0)(t_0+k)}\sum_{n=n_k^++2}^{+\infty}\int_0^\pi \cos\left((n-n_0)w\right)e^{D(t_0+k)[\cos w-1]}\left|\widetilde{\Phi}_n(-k)-v^k_n(-k)\right|dw \\
&\quad+\frac{1}{\pi}e^{f'(0)(t_0+k)}\int_0^\pi \cos\left(\left(n_k^++1-n_0\right)w\right)e^{D(t_0+k)[\cos w-1]}\left|\widetilde{\Phi}_n(-k)-v^k_n(-k)\right|dw \\
&\quad+\frac{1}{\pi}e^{f'(0)(t_0+k)}\int_0^\pi \cos\left(\left(n_k^+-n_0\right)w\right)e^{D(t_0+k)[\cos w-1]}\left|\widetilde{\Phi}_n(-k)-v^k_n(-k)\right|dw \\
&\quad+\frac{1}{\pi}e^{f'(0)(t_0+k)}\sum_{n=n_k^-+1}^{n_k^+-1}\int_0^\pi \cos\left((n-n_0)w\right)e^{D(t_0+k)[\cos w-1]}\left|\widetilde{\Phi}_n(-k)-v^k_n(-k)\right|dw \\
&\quad+\frac{1}{\pi}e^{f'(0)(t_0+k)}\int_0^\pi \cos\left(\left(n_k^--n_0\right)w\right)e^{D(t_0+k)[\cos w-1]}\left|\widetilde{\Phi}_n(-k)-v^k_n(-k)\right|dw \\
&\quad+\frac{1}{\pi}e^{f'(0)(t_0+k)}\int_0^\pi \cos\left(\left(n_k^--1-n_0\right)w\right)e^{D(t_0+k)[\cos w-1]}\left|\widetilde{\Phi}_n(-k)-v^k_n(-k)\right|dw \\
&\quad+\frac{1}{\pi}e^{f'(0)(t_0+k)}\sum_{n=-\infty}^{n_k^--2}\int_0^\pi \cos\left((n-n_0)w\right)e^{D(t_0+k)[\cos w-1]}\left|\widetilde{\Phi}_n(-k)-v^k_n(-k)\right|dw.
\end{aligned}
$$

Call I, II, III, IV, V, VI, and VII the seven terms on the right-hand side of this last inequality. We have

$$
\begin{aligned}
\mathrm{I} &= \frac{1}{\pi}e^{f'(0)(t_0+k)}\sum_{n=n_k^++2}^{+\infty}\int_0^\pi \cos\left((n-n_0)w\right)e^{D(t_0+k)[\cos w-1]}\left|\widetilde{\Phi}_n(-k)-v^k_n(-k)\right|dw \\
&\le \frac{1}{\pi}e^{f'(0)(t_0+k)}\sum_{n=n_k^++2}^{+\infty}\int_0^\pi \cos\left((n-n_0)w\right)e^{D(t_0+k)[\cos w-1]}e^{f'(0)(-k+\theta_0)}dw \\
&\quad+\frac{A_{c'}}{\pi}e^{f'(0)(t_0+k)}\sum_{n=n_k^++2}^{+\infty}\int_0^\pi \cos\left((n-n_0)w\right)e^{D(t_0+k)[\cos w-1]}e^{\lambda_1(c')\left(-n-c'k+\theta'\right)}dw \\
&= \mathrm{I}_1+\mathrm{I}_2.
\end{aligned}
$$

Obviously,

$$
\begin{aligned}
\mathrm{I}_1 &= \frac{1}{\pi}e^{f'(0)(t_0+k)}\sum_{n=n_k^++2}^{+\infty}\int_0^\pi \cos\left((n-n_0)w\right)e^{D(t_0+k)[\cos w-1]}e^{f'(0)(-k+\theta_0)}dw \\
&= 2e^{f'(0)(t_0+\theta_0)}\sum_{n=n_k^++2}^{+\infty}a^k_{n-n_0}\to 0 \text{ as } k\to+\infty
\end{aligned}
$$

due to the fact $\frac{n_k^++2-n_0}{k} > \frac{c\lambda_1(c)-f'(0)}{2\lambda_1(c)} > 0$ for sufficiently large k and (6.7). In

addition, we have

$$\begin{aligned}
\mathrm{I}_2 &= \frac{A_{c'}}{\pi} e^{f'(0)(t_0+k)} \sum_{n=n_k^++2}^{+\infty} \int_0^{\pi} \cos\left((n-n_0)\,w\right) e^{D(t_0+k)[\cos w-1]} e^{\lambda_1(c')\left(-n-c'k+\theta'\right)} dw \\
&= 2A_{c'} e^{\left(f'(0)-\lambda_1(c')c'\right)k} e^{f'(0)\left(t_0+\theta'\right)} \sum_{n=n_k^++2}^{+\infty} a_{n-n_0}^k e^{-\lambda_1(c')n} \\
&\le 2A_{c'} e^{f'(0)\left(t_0+\theta'\right)} \sum_{n=n_k^++2}^{+\infty} a_{n-n_0}^k \to 0 \text{ as } k \to \infty.
\end{aligned}$$

Consider II. In this case, let

$$0 \le \delta = \left(ck - \frac{f'(0)}{\lambda_1(c)}k - \theta + \frac{1}{\lambda_1(c)} \ln \frac{1}{\alpha_c} + \frac{f'(0)}{\lambda_1(c)}\theta_0\right) - n_k^+ < 1.$$

Then we have

$$\begin{aligned}
\mathrm{II} &= 2e^{f'(0)(t_0+k)} a_{n_k^++1-n_0}^k \left|\widetilde{\Phi}_{n_k^++1}(-k) - v_{n_k^++1}^k(-k)\right| \\
&\le 2e^{f'(0)(t_0+k)} a_{n_k^++1-n_0}^k \left[e^{f'(0)(-k+\theta_0)} + A_{c'} e^{\lambda_1(c')\left(-n_k^+-1-c'k+\theta'\right)}\right] \\
&\quad + 2e^{f'(0)(t_0+k)} a_{n_k^++1-n_0}^k \left[A_{c'} e^{\lambda_1(c')\left(-n_k^+-1-c'k+\theta'\right)} + A_c e^{\lambda_1(c)\left(n_k^++1-ck+\theta\right)}\right] \\
&= 2e^{f'(0)(t_0+k)} a_{n_k^++1-n_0}^k \left[e^{f'(0)(-k+\theta_0)} + 2A_{c'} e^{\lambda_1(c')\left(-n_k^+-1-c'k+\theta'\right)}\right] \\
&\quad + 2e^{f'(0)(t_0+k)} a_{n_k^++1-n_0}^k A_c e^{\lambda_1(c)\left\{\left(ck-\frac{f'(0)}{\lambda_1(c)}k-\theta+\frac{1}{\lambda_1(c)}\ln\frac{1}{\alpha_c}+\frac{f'(0)}{\lambda_1(c)}\theta_0\right)-\delta+1-ck+\theta\right\}} \\
&= 2e^{f'(0)(t_0+k)} a_{n_k^++1-n_0}^k \left[e^{f'(0)(-k+\theta_0)} + 2A_{c'} e^{\lambda_1(c')\left(-n_k^+-1-c'k+\theta'\right)}\right] \\
&\quad + \frac{2A_c}{\alpha_c} e^{\lambda_1(c)(1-\delta)+f'(0)(t_0+\theta_0)} a_{n_k^++1-n_0}^k \\
&\to 0 \qquad \text{as} \qquad k \to +\infty.
\end{aligned}$$

Similarly, we have that III$\to 0$ as $k \to +\infty$. For IV, we have

$$\begin{aligned}
\mathrm{IV} &= \frac{1}{\pi} e^{f'(0)(t_0+k)} \\
&\quad \times \sum_{n=n_k^-+1}^{n_k^+-1} \int_0^{\pi} \cos\left((n-n_0)\,w\right) e^{D(t_0+k)[\cos w-1]} \left|\widetilde{\Phi}_n(-k) - v_n^k(-k)\right| dw \\
&\le 2e^{f'(0)(t_0+k)} \sum_{n=n_k^-+1}^{n_k^+-1} a_{n-n_0}^k \left[A_c e^{\lambda_1(c)(n-ck+\theta)} + A_{c'} e^{\lambda_1(c')\left(-n-c'k+\theta'\right)}\right] \\
&= 2e^{f'(0)(t_0+k)} \sum_{n=0}^{n_k^+-1} a_{n-n_0}^k \left[A_c e^{\lambda_1(c)(n-ck+\theta)} + A_{c'} e^{\lambda_1(c')\left(-n-c'k+\theta'\right)}\right] \\
&\quad + 2e^{f'(0)(t_0+k)} \sum_{n=n_k^-+1}^{-1} a_{n-n_0}^k \left[A_c e^{\lambda_1(c)(n-ck+\theta)} + A_{c'} e^{\lambda_1(c')\left(-n-c'k+\theta'\right)}\right] \\
&= \mathrm{IV}_1 + \mathrm{IV}_2.
\end{aligned}$$

Consider IV_1. Then

$$\begin{aligned}
\mathrm{IV}_1 &= 2e^{f'(0)(t_0+k)} \sum_{n=0}^{n_k^+-1} a_{n-n_0}^k \left[A_c e^{\lambda_1(c)(n-ck+\theta)} + A_{c'} e^{\lambda_1(c')(-n-c'k+\theta')} \right] \\
&\le 2A_{c'} e^{-\left(\lambda_1(c')c'-f'(0)\right)k} e^{f'(0)t_0+\lambda_1(c')\theta'} \sum_{n=0}^{n_k^+} a_{n-n_0}^k \\
&\quad + 2A_c e^{f'(0)(t_0+k)} \sum_{n=\mathrm{int}\left[n_k^+/2\right]}^{n_k^+-1} a_{n-n_0}^k e^{\lambda_1(c)(n-ck+\theta)} \\
&\quad + 2A_c e^{f'(0)(t_0+k)} \sum_{n=0}^{\mathrm{int}\left[n_k^+/2\right]} a_{n-n_0}^k e^{\lambda_1(c)(n-ck+\theta)} \\
&\le 2A_{c'} e^{-\left(\lambda_1(c')c'-f'(0)\right)k} e^{f'(0)t_0+\lambda_1(c')\theta'} \\
&\quad + 2A_c e^{f'(0)(t_0+k)} e^{\lambda_1(c)\left(\left(n_k^+-1\right)-ck+\theta\right)} \sum_{n=\mathrm{int}\left[n_k^+/2\right]}^{\infty} a_{n-n_0}^k \\
&\quad + 2A_c e^{\lambda_1(c)\left(\mathrm{int}\left[n_k^+/2\right]-ck+\theta\right)} e^{f'(0)(t_0+k)} \\
&= \mathrm{IV}_1^1 + \mathrm{IV}_1^2 + \mathrm{IV}_1^3.
\end{aligned}$$

It is easy to see that $\mathrm{IV}_1^1 \le 2A_{c'} e^{-(\lambda_1(c')c'-f'(0))k} e^{f'(0)t_0+\lambda_1(c')\theta'} \to 0$ as $k \to +\infty$, because $\lambda_1(c')c' - f'(0) > 0$. By virtue of

$$\begin{aligned}
& e^{f'(0)(t_0+k)} e^{\lambda_1(c)\left(\left(n_k^+-1\right)-ck+\theta\right)} \\
&\quad \le A_c e^{f'(0)(t_0+k)} e^{\lambda_1(c)\left\{\left(ck-\frac{f'(0)}{\lambda_1(c)}k-\theta+\frac{1}{\lambda_1(c)}\ln\frac{1}{\alpha_c}+\frac{f'(0)}{\lambda_1(c)}\theta_0\right)-1-ck+\theta\right\}} \\
&\quad = \frac{1}{\alpha_c} e^{f'(0)(t_0+\theta_0)-\lambda_1(c)}
\end{aligned}$$

and $\frac{n_k^+}{2} - n_0 > \frac{c\lambda_1(c)-f'(0)}{4\lambda_1(c)}k$ for sufficiently large k, we have

$$\mathrm{IV}_1^2 = 2A_c e^{f'(0)(t_0+k)} e^{\lambda_1(c)\left(\left(n_k^+-1\right)-ck+\theta\right)} \sum_{n=\mathrm{int}\left[n_k^+/2\right]}^{\infty} a_{n-n_0}^k \to 0 \text{ as } k \to +\infty.$$

Moreover, we have

$$\begin{aligned}
\mathrm{IV}_1^3 &= 2A_c e^{\lambda_1(c)\left(\mathrm{int}\left[n_k^+/2\right]-ck+\theta\right)} e^{f'(0)(t_0+k)} \\
&\le 2A_c e^{\lambda_1(c)\left(\frac{1}{2}\left[ck-\frac{f'(0)}{\lambda_1(c)}k-\theta+\frac{1}{\lambda_1(c)}\ln\frac{1}{\alpha_c}+\frac{f'(0)}{\lambda_1(c)}\theta_0\right]-ck+\theta\right)} e^{f'(0)(t_0+k)} \\
&= 2A_c e^{f'(0)t_0+\frac{1}{2}\ln\frac{1}{\alpha_c}+\frac{1}{2}f'(0)\theta_0+\frac{1}{2}\lambda_1(c)\theta} e^{-\frac{1}{2}\left[\lambda_1(c)c-f'(0)\right]k} \\
&\to 0 \text{ as } k \to +\infty.
\end{aligned}$$

We can prove that IV_2, V, VI, and VII converge to zero as $k \to +\infty$ by arguments similar to those for IV_1, III, II, and I, respectively.

Eventually, $|\widetilde{\Phi}_{n_0}(t_0) - v_{n_0}^k(t_0)| \to 0$ as $k \to +\infty$. Since $v_{n_0}^k(t_0) \to \Phi_{n_0}(t_0)$ and $(n_0, t_0) \in \mathbb{Z} \times \mathbb{R}$ is arbitrary, we obtain that $\Phi_n(t) = \widetilde{\Phi}_n(t)$ for any $(n,t) \in \mathbb{Z} \times \mathbb{R}$. The limit function being unique, the whole sequence Φ^k converges to Φ as $k \to +\infty$.

Using the same estimates as above, we can prove that the entire solution of (1.4) satisfying (1.9) is unique. ∎

Acknowledgments. The authors thank the referees for their valuable comments and suggestions on the original manuscript.

REFERENCES

[1] P.W. Bates and A. Chmaj, *A discrete convolution model for phase transitions*, Arch. Ration. Mech. Anal., 150 (1999), pp. 281–305.

[2] J.W. Cahn, S.-N. Chow, and E.S. Van Vleck, *Spatially discrete nonlinear diffusion equations*, Rocky Mountain J. Math., 25 (1995), pp. 87–118.

[3] X. Chen, S.-C. Fu, and J.-S. Guo, *Uniqueness and asymptotics of traveling waves of monostable dynamics on lattices*, SIAM J. Math. Anal., 38 (2006), pp. 233–258.

[4] X. Chen and J.-S. Guo, *Existence and asymptotic stability of travelling waves of discrete quasilinear monostable equations*, J. Differential Equations, 184 (2002), pp. 549–569.

[5] X. Chen and J.-S. Guo, *Uniqueness and existence of travelling waves of discrete quasilinear monostable dynamics*, Math. Ann., 326 (2003), pp. 123–146.

[6] X. Chen and J.-S. Guo, *Existence and uniqueness of entire solutions for a reaction-diffusion equation*, J. Differential Equations, 212 (2005), pp. 62–84.

[7] X. Chen, J.-S. Guo and H. Ninomiya, *Entire solutions of reaction-diffusion equations with balanced bistable nonlinearities*, Proc. Roy. Soc. Edinburgh Sect. A, 136 (2006), pp. 1207–1237.

[8] C.P. Cheng, W.T. Li, and Z.C. Wang, *Spreading speeds and traveling waves in a delayed population model with stage structure on a two-dimensional spatial lattice*, IMA J. Appl. Math., 73 (2008), pp. 592–618.

[9] C.P. Cheng, W.T. Li, and Z.C. Wang, *Asymptotic Stability of Traveling Wavefronts in a Delayed Population Model with Stage Structure on a Two-dimensional Spatial Lattice*, submitted.

[10] S.-N. Chow, *Lattice dynamical systems*, in Dynamical Systems, Lecture Notes in Math. 1822, J.W. Macki and P. Zecca, eds., Springer, Berlin, 2003, pp. 1–102.

[11] T. Faria, W. Huang, and J. Wu, *Travelling waves for delayed reaction-diffusion equations with global response*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 462A (2006), pp. 229–261.

[12] T. Faria and S. Trofimchuk, *Nonmonotone travelling waves in a single species reaction diffusion equation with delay*, J. Differential Equations, 228 (2006), pp. 357–376.

[13] T. Faria and S. Trofimchuk, *Positive heteroclinics and traveling waves for scalar population models with a single delay*, Appl. Math. Comput., 185 (2007), pp. 594–603.

[14] Y. Fukao, Y. Morita, and H. Ninomiya, *Some entire solutions of the Allen-Cahn equation*, Taiwanese J. Math., 8 (2004), pp. 15–32.

[15] R.R. Goldberg, *Fourier Transform*, Cambridge University Press, New York, 1965.

[16] S.A. Gourley and J. Wu, *Delayed nonlocal diffusive systems in biological invasion and disease spread*, in Nonlinear Dynamics and Evolution Equations, Fields Inst. Commun. 48, American Mathematical Society, Providence, RI, 2006, pp. 137–200.

[17] S.A. Gourley and J. Wu, *Extinction and periodic oscillations in an age-structured population model in a patchy environment*, J. Math. Anal. Appl., 289 (2004), pp. 431–445.

[18] J.-S. Guo and Y. Morita, *Entire solutions of reaction-diffusion equations and an application to discrete diffusive equations*, Discrete Contin. Dyn. Syst., 12 (2005), pp. 193–212.

[19] Y.-J.L. Guo, *Entire solutions for a discrete diffusive equation*, J. Math. Anal. Appl., 347 (2008), pp. 450–458.

[20] F. Hamel and N. Nadirashvili, *Entire solutions of the KPP Equation*, Comm. Pure Appl. Math., 52 (1999), pp. 1255–1276.

[21] F. Hamel and N. Nadirashvili, *Travelling fronts and entire solutions of the Fisher-KPP equation in* $\mathbb{R}^N$, Arch. Ration. Mech. Anal., 157 (2001), pp. 91–163.

[22] Y. Kyrychko, S.A. Gourley, and M.V. Bartuccelli, *Dynamics of a stage-structured population model on an isolated finite lattice*, SIAM J. Math. Anal., 37 (2006), pp. 1688–1708.

[23] W.T. Li, N.W. Liu, and Z.C. Wang, *Entire solutions in reaction-advection-diffusion equations in cylinders*, J. Math. Pures Appl., 90 (2008), pp. 492–504.

[24] W.T. Li, Z.C. Wang, and J. Wu, *Entire solutions of monostable reaction-diffusion equations with delayed nonlinearity*, J. Differential Equations, 245 (2008), pp. 102–129.

[25] S. Ma, P. Weng, and X. Zou, *Asymptotic speeds of propagation and traveling wavefronts in a non-local delayed lattice differential equation*, Nonlinear Anal., 65 (2006), pp. 1858–1890.

[26] S. Ma and X. Zou, *Propagation and its failure in a lattice delayed differential equation with global interaction*, J. Differential Equations, 212 (2005), pp. 129–190.

[27] S. Ma and X. Zou, *Existence, uniqueness and stability of traveling waves in a discrete reaction-diffusion monostable equation with delay*, J. Differential Equations, 217 (2005), pp. 54–87.
[28] J. Mallet-Paret, *The global structure of traveling waves in spatially discrete dynamical systems*, J. Dynam. Differential Equations, 11 (1999), pp. 49–127.
[29] R.H. Martin and H.L. Smith, *Abstract functional equations and reaction-diffusion systems*, Trans. Amer. Math. Soc., 321 (1990), pp. 1–44.
[30] Y. Morita and H. Ninomiya, *Entire solutions with merging fronts to reaction-diffusion equations*, J. Dynam. Differential Equations, 18 (2006), pp. 841–861.
[31] J.D. Murray, *Mathematical Biology*, Springer, Berlin, 1989.
[32] Z.X. Shi, W.T. Li, and C.P. Cheng, *Stability and Uniqueness of Traveling Wavefronts in a Two-dimensional Lattice Differential Equation with Delay*, Appl. Math. Comput., to appear.
[33] H.L. Smith, *Monotone Dynamical Systems: An Introduction to the Theory of Competitive and Cooperative Systems*, Math. Surveys and Monogr. 41, American Mathematical Society, Providence, RI, 1995.
[34] E.C. Titchmarsh, *Introduction to the Theory of Fourier Integrals*, Clarendon Press, Oxford, 1962.
[35] Z.C. Wang, W.T. Li, and S. Ruan, *Travelling wave fronts of reaction-diffusion systems with spatio-temporal delays*, J. Differential Equations, 222 (2006), pp. 185–232.
[36] Z.C. Wang, W.T. Li, and S. Ruan, *Existence and stability of traveling wave fronts in reaction advection diffusion equations*, J. Differential Equations, 238 (2007), pp. 153–200.
[37] Z.C. Wang, W.T. Li, and S. Ruan, *Traveling fronts in monostable equations with nonlocal delayed effects*, J. Dynam. Differential Equations, 20 (2008), pp. 573–607.
[38] Z.C. Wang, W.T. Li, and S. Ruan, *Entire solutions in bistable reaction-diffusion equations with nonlocal delayed nonlinearity*, Trans. Amer. Math. Soc., 361 (2009), pp. 2047–2084.
[39] Z.C. Wang, W.T. Li, and S. Ruan, *Entire Solutions in Lattice Delayed Differential Equations with Global Interaction: Bistable Case*, submitted.
[40] P. Weng, H. Huang, and J. Wu, *Asymptotic speed of propagation of wave fronts in a lattice delay differential equation with global interaction*, IMA J. Appl. Math., 68 (2003), pp. 409–439.
[41] P. Weng, J. Wu, H. Huang, and J. Ling, *Asymptotic speed of propagation of wave fronts in a* $2D$ *lattice delay differential equation with global interaction*, Can. Appl. Math. Q., 11 (2003), pp. 377–414.
[42] J. Wu and X. Zou, *Asymptotic and periodic boundary value problems of mixed FDEs and wave solutions of lattice differential equations*, J. Differential Equations, 135 (1997), pp. 315–357.
[43] H. Yagisita, *Backward global solutions characterizing annihilation dynamics of travelling fronts*, Publ. Res. Inst. Math. Sci., 39 (2003), pp. 117–164.
[44] B. Zinner, *Existence of travelling wavefront solutions for the discrete Nagumo equation*, J. Differential Equations, 96 (1992), pp. 1–27.

SIAM J. MATH. ANAL.
Vol. 40, No. 6, pp. 2421–2436

AN ELEMENTARY APPROACH TO A MODEL PROBLEM OF LAGERSTROM*

S. P. HASTINGS† AND J. B. MCLEOD‡

Abstract. The equation studied is $u'' + \frac{n-1}{r}u' + \varepsilon u\, u' + ku'^2 = 0$, with boundary conditions $u(1) = 0$, $u(\infty) = 1$. This model equation has been studied by many authors since it was introduced in the 1950s by P. A. Lagerstrom. We use an elementary approach to show that there is an infinite series solution which is uniformly convergent on $1 \le r < \infty$. The first few terms are easily derived, from which one quickly deduces the inner and outer asymptotic expansions, with no matching procedure or a priori assumptions about the nature of the expansion. We also give a short and elementary existence and uniqueness proof which covers all $\varepsilon > 0$, $k \ge 0$, and $n \ge 1$.

Key words. matched asymptotics, singular perturbation, boundary value problem

AMS subject classifications. Primary, 34E05; Secondary, 34E10, 34E15

DOI. 10.1137/080718759

1. Introduction.

The main problem is to investigate the asymptotics as $\varepsilon \to 0$ of the boundary value problem

$$u'' + \frac{n-1}{r}u' + \varepsilon u u' + ku'^2 = 0, \tag{1}$$

with

$$u(1) = 0, u(\infty) = 1. \tag{2}$$

We consider the cases $k = 0$ and $k = 1$. Our interest in these problems, originally due to Lagerstrom in the 1950s [6], [7], was stimulated by two recent papers by Popovic and Szmolyan [10], [11], who adopt a geometric approach to the problem when $k = 0$, and there are many papers which use methods of matched asymptotics or multiple scales, with varying degrees of rigor. We will review some of this work below. The point of this paper is to give a completely rigorous and relatively short answer to the problem without making any appeal either to geometric methods or to matched asymptotics. We can express the solution as an infinite series, uniformly convergent for all values of the independent variable. From this series we obtain the inner and outer asymptotic expansions with no a priori assumption about the nature of these expansions. An important and, as far as we know, original feature is that there is no "matching."

Lagerstrom came up with these problems as models of viscous incompressible ($k = 0$) and compressible ($k = 1$) flow, so much of his work centered on $n = 2$ or 3, but he also discussed general $n \ge 1$ [8]. The infinite series we develop can be obtained for any real number n. What n controls is the rate of convergence of the series.

*Received by the editors March 18, 2008; accepted for publication (in revised form) October 31, 2008; published electronically February 25, 2009.
http://www.siam.org/journals/sima/40-6/71875.html
†Department of Mathematics, University of Pittsburgh, Thackeray Hall, Pittsburgh, PA 15360 (sph@pitt.edu).
‡Mathematics Institute, University of Oxford, 24-29 St. Giles, Oxford Ox13LB, United Kingdom (mcleod@maths.ox.ac.uk).

For $\varepsilon = k = 0$, there is an obvious distinction between $n > 2$ and $n \leq 2$. If $n > 2$, then the problem (1)–(2) has the unique solution

$$u = 1 - \frac{1}{r^{n-2}}, \tag{3}$$

so that the solution with ε small is presumably some sort of perturbation of this. If $n \leq 2$, then there is no such solution. A consequence is that the convergence as $\varepsilon \to 0$ is more subtle when $n \leq 2$ than when $n > 2$. Our analysis will show that there is little prospect of discussing the behavior for small ε if $n < 2$, but fortunately we can handle all $n \geq 2$. Although it has been thought that finding the asymptotics when $k = 1$ is considerably more difficult than when $k = 0$ [3], we will show that our technique covers each case with comparable effort.

Our methods are not restricted to Lagerstrom's problems (1)–(2). In subsequent work (in preparation), we will show that there is a general method which can yield similar results for a class of singularly perturbed boundary value problems.

We start in section 2 by showing that each of these problems has one and only one solution for any $n \geq 1$ and any $\varepsilon > 0$. This is based on a simple shooting argument plus a comparison principle. These results have been obtained before, but our proof is quite short. In the subsequent sections we develop the integral equation referred to above and show how it leads with relative ease to the inner and outer expansions. These expansions go back to Lagerstrom and Kaplun, with rigorous justification of some of the features to be found in [1] or [11], for example. We find the exposition in Hinch's book [3] particularly clear (though nonrigorous), and make that work our point of comparison in checking that we get the same expansions as were found previously.

2. Existence and uniqueness. As far as we know, the first existence proof was by Hsiao [4], who considered only $n = 1$ and sufficiently small $\varepsilon > 0$. Subsequently Tam gave what seems to be the first proof valid for all $\varepsilon > 0$ and $k \geq 0$ [14]. Subsequent proofs by MacGillivray [9], Cohen, Lagerstrom, and Fokas [1], Hunter, Tajdari, and Boyer [5], each of which covers all $\varepsilon > 0$, and by several other authors, e.g., [12], [10], for restricted ranges of ε, add to the variety of techniques which have been shown to work. Uniqueness is proved in [5] (for $k = 0$) by use of a contraction mapping theorem, and in [1] by essentially a comparison method. The goal of [10] is not to give a short proof, but to illustrate the application of geometric perturbation theory to a much studied problem in matched asymptotic expansions. The proofs we give of existence and uniqueness are considerably shorter than the others we have seen.

THEOREM 1. *There exists a unique solution to the problem* (1)–(2) *for any* $k \geq 0$, $\varepsilon > 0$, *and* $n \geq 1$.

Proof. Like some others, starting with [14], we prove existence using a shooting method by considering the initial value problem

$$u'' + \frac{n-1}{r}u' + \varepsilon u u' + k u'^2 = 0, \tag{4}$$

$$u(1) = 0,\ u'(1) = c, \tag{5}$$

for each $c > 0$. Since $u' = 0$ implies that $u'' = 0$ and u is constant, any solution to this problem is positive and increasing. As was observed in [14],

$$u'' + \varepsilon u u' \leq 0,$$

and so from (5),

$$u' + \frac{1}{2}\varepsilon u^2 \leq c.$$

In particular, since $u' \geq 0$,

$$u \leq \sqrt{\frac{2c}{\varepsilon}}, \tag{6}$$

so the solution exists, and satisfies this bound, on $[1, \infty)$. Therefore, $\lim_{r\to\infty} u(r)$ exists. Writing the equation in the form

$$\left(r^{n-1}u'\right)' + (\varepsilon u + ku')\left(r^{n-1}u'\right) = 0 \tag{7}$$

and integrating twice gives

$$r^{n-1}u'(r) = ce^{-ku(r)-\varepsilon\int_1^r u(s)ds}, \tag{8}$$

$$u(r) = \int_1^r \frac{c}{s^{n-1}} e^{-ku-\int_1^s \varepsilon u dt} ds. \tag{9}$$

If $u(2) < 1$, then since u is increasing, (9) implies that

$$u(2) > p(c) = \int_1^2 \frac{c}{s^{n-1}} e^{-\varepsilon-k} ds.$$

From this and (6), we see that there are c_1 and c_2, with $0 < c_1 < c_2$, such that if $c = c_1$, then $u(\infty) < 1$, while if $c = c_2$, then $u(\infty) > u(2) \geq 1$. Further, from (8) for any $r > R > 2$,

$$u(r) = u(R) + c\int_R^r \frac{1}{s^{n-1}} e^{-ku-\int_1^s \varepsilon u dt} ds.$$

If $n > 2$, the second term is bounded above by $\frac{c}{(n-2)R^{n-2}}$, while if $1 \leq n \leq 2$ and $R \geq 2$, it is bounded by $c\int_R^\infty e^{-\varepsilon(s-2)p(c)} ds$. Hence, this term tends to zero as $R \to \infty$, uniformly for $r \geq R$, $c_1 \leq c \leq c_2$. Since $u(R)$ is a continuous function of c, for any R, it follows that $u(\infty)$ is also continuous in c, and so there is a c with $u(\infty) = 1$, giving a solution to (1)–(2).

For uniqueness, suppose that there are two solutions of (1)–(2), say u_1 and u_2, with $u_1'(1) > u_2'(1) > 0$. Then $u_1 > u_2$ on some maximal interval, say $(1, X)$ where $X \leq \infty$. For the same initial conditions, if $\varepsilon = k = 0$, then direct integration shows that $u_1 > u_2$ on $(1, \infty)$, and, moreover, $u_1(\infty) > u_2(\infty)$. We then raise ε and k, looking for a pair (ε_1, k_1) such that $u_1(X) = u_2(X)$ for some $X \leq \infty$, and if $0 \leq \varepsilon < \varepsilon_1$ or $0 \leq k < k_1$, no such X exists. Hence, at (ε_1, k_1), $u_1 \geq u_2$ on $[0, \infty)$. If, at (ε_1, k_1), $X < \infty$, then u_1 and u_2 must be tangent at X, since $u_1 - u_2$ has a minimum there, contradicting the uniqueness of initial value problems for (1). Hence, $X = \infty$, and $u_1 > u_2$ on $(1, \infty)$.

Observe from (7) that if $u_1'(r) = u_2'(r)$ for some r, then $\left(r^{n-1}(u_1' - u_2')\right)' < 0$, since $u_1 > u_2$, so that there cannot be oscillations in $u_1' - u_2'$. Hence, $u_1(\infty) = u_2(\infty)$ implies that there is an R with $u_1'(R) = u_2'(R)$ and $u_1' < u_2'$ on (R, ∞). Integrating

(7), and recalling that $u_1(\infty) = u_2(\infty)$, gives

$$r^{n-1}(u_1' - u_2')|_R^\infty = \frac{1}{2}\varepsilon R^{n-1}\left(u_1^2 - u_2^2\right)|_R + \frac{1}{2}\varepsilon(n-1)\int_R^\infty s^{n-2}\left(u_1^2 - u_2^2\right)ds - k\int_R^\infty s^{n-1}\left(u_1'^2 - u_2'^2\right)ds.$$

The left-hand side is zero, and all the terms on the right are positive, giving the necessary contradiction. □

Remark 1. The existence theorem in [10] has one added part. It is shown there that as $\varepsilon \to 0$, the solution tends to a so-called singular solution obtained by taking a formal limit as $\varepsilon \to 0$. See [10] for details. This limit result follows from our rigorous asymptotic expansions given below.

Remark 2. There would seem to be no difficulty in extending the existence proof even to $n < 1$, but the uniqueness proof does use essentially the fact that $n \geq 1$.

3. The infinite series (with $k = 0$, $n \geq 2$). Starting again with (1), and $u(1) = 0$, we first consider the case $k = 0$ and obtain

$$r^{n-1}u' = Be^{-\varepsilon\int_1^r u(t)dt} \tag{10}$$

for some constant B. Since $u(\infty) = 1$, (10) implies that $u'(r)$ is exponentially small as $r \to \infty$. Hence we can rewrite (10) as

$$r^{n-1}u' = Ce^{-\varepsilon r - \varepsilon\int_\infty^r (u-1)dt},$$

so that

$$u - 1 = C\int_\infty^r \frac{1}{t^{n-1}}e^{-\varepsilon t - \varepsilon\int_\infty^t (u-1)ds}dt.$$

Setting $\varepsilon r = \rho$, $\varepsilon t = \tau$, and $\varepsilon s = \sigma$, we obtain

$$u(\rho) - 1 = C\varepsilon^{n-2}\int_\infty^\rho \frac{1}{\tau^{n-1}}e^{-\tau}e^{-\int_\infty^\tau (u(\sigma)-1)d\sigma}d\tau, \tag{11}$$

where we use the arguments ρ and σ to indicate that we mean the rescaled version of u. Here C is a constant satisfying

$$-1 = C\varepsilon^{n-2}\int_\infty^\varepsilon \frac{1}{\tau^{n-1}}e^{-\tau}e^{-\int_\infty^\tau (u-1)d\sigma}d\tau. \tag{12}$$

Since for each ε there is a unique solution, this determines a unique C, dependent on ε.

We now consider the integral

$$\int_\tau^\infty (1 - u(\sigma))\,d\sigma, \tag{13}$$

which appears in the exponent in (12). This integral has been seen to converge for each ε, but we need a bit more, namely, that it is bounded uniformly in $\varepsilon \leq \tau < \infty$ as $\varepsilon \to 0$. To see this, we note that as a function of σ, u satisfies

$$\frac{d^2u}{d\sigma^2} + \frac{n-1}{\sigma}\frac{du}{d\sigma} + u\frac{du}{d\sigma} = 0$$

$$u = 0 \text{ when } \sigma = \varepsilon,\ u(\infty) = 1.$$

Denoting the unique solution by $u_\varepsilon(\sigma)$, we claim that if $0 < \varepsilon_1 < \varepsilon_2$, then $u_{\varepsilon_1} > u_{\varepsilon_2}$ for $\varepsilon_2 \le \sigma < \infty$. If this is false, then ε_1 and ε_2 can be chosen so that $u_{\varepsilon_1}(\sigma_0) = u_{\varepsilon_2}(\sigma_0)$ for some $\sigma_0 \ge \varepsilon_2$. But then the problem

$$\frac{d^2u}{d\sigma^2} + \frac{n-1}{\sigma}\frac{du}{d\sigma} + u\frac{du}{d\sigma} = 0,$$

$$u(\sigma_0) = u_{\varepsilon_1}(\sigma_0),\ u(\infty) = 1,$$

has two solutions, contradicting our earlier uniqueness proof.

A consequence of this is that $\int_{\varepsilon_2}^{\infty}(1 - u_{\varepsilon_1}(\sigma))\,d\sigma < \int_{\varepsilon_2}^{\infty}(1 - u_{\varepsilon_2}(\sigma))\,d\sigma$, which implies that the integral in the exponent in (12), including the minus sign in front, is bounded below independently of $\tau \ge \varepsilon$ and of ε. We then see that the τ-integral in (12) approaches $-\infty$ as $\varepsilon \to 0$, and hence that

$$\lim_{\varepsilon \to 0^+} C\varepsilon^{n-2} = 0.$$

Since $\int_{\infty}^{\tau}(u-1)\,d\sigma > 0$, it follows from (11) that if

$$E_{n-1}(\rho) = \int_{\rho}^{\infty} \frac{1}{\tau^{n-1}} e^{-\tau} d\tau,$$

then

$$|u(\rho) - 1| < C\varepsilon^{n-2} E_{n-1}(\rho). \tag{14}$$

For purposes of future estimates, we make the obvious remark that

$$E_{n-1}(\rho) = \begin{cases} O\left(\rho^{2-n}\right) \text{ as } \rho \to 0 \text{ if } n > 2, \\ O(\log \rho) \text{ as } \rho \to 0 \text{ if } n = 2, \\ O\left(\rho^{1-n} e^{-\rho}\right) \text{ as } \rho \to \infty. \end{cases} \tag{15}$$

Hence if $n > 2$, there is a constant K such that

$$E_{n-1}(\rho) \le K \min\left(\rho^{2-n}, \rho^{1-n} e^{-\rho}\right). \tag{16}$$

The method now is to work from (11). As observed before, since $u'(r)$ is exponentially small as $r \to \infty$, the integral term $\int_{\rho}^{\infty}(u-1)\,d\sigma$ converges. Hence, for given $\varepsilon > 0$ and $\rho_0 > 0$, and any $\rho \ge \rho_0$,

(17)
$$u(\rho) - 1 = C\varepsilon^{n-2} \int_{\infty}^{\rho} \frac{1}{\tau^{n-1}} e^{-\tau} \left\{ 1 - \int_{\infty}^{\tau}(u-1)\,d\sigma + \frac{1}{2}\left(\int_{\infty}^{\tau}(u-1)\,d\sigma\right)^2 - \cdots \right\} d\tau,$$

where the series in the integrand converges uniformly for $\rho_0 \le \tau < \infty$.

In fact, we will need to use this series for all $\rho \ge \varepsilon$. Thus we need to check its convergence in this interval. This follows from (14) and (15), which imply that for any $\rho \ge \varepsilon$, if $n \ge 2$, then

$$\left| \int_{\rho}^{\infty} (u(s) - 1)\,ds \right| < C\varepsilon^{n-2} \int_{\varepsilon}^{\infty} E_{n-1}(s)\,ds \tag{18}$$

and

$$\varepsilon^{n-2}\int_{\varepsilon}^{\infty} E_{n-1}(s)\,ds = \begin{cases} o(1) \text{ as } \varepsilon \to 0 \text{ if } n > 2, \\ O(1) \text{ as } \varepsilon \to 0 \text{ if } n = 2. \end{cases}$$

Hence for $n > 2$ and any C, the series in the integrand of (17) converges uniformly on $[\varepsilon, \infty)$.

Now set

$$\Phi = C\varepsilon^{n-2}\int_{\varepsilon}^{\infty} E_{n-1}(s)\,ds.$$

We note that if $n > 2$, then $\Phi \to 0$ as $\varepsilon \to 0$, while if $n = 2$, then $\Phi \to 0$ as $C \to 0$.

We proceed to solve (17) by iteration. Thus, the first approximation is, from (16),

$$u(\rho) - 1 = C\varepsilon^{n-2}\int_{\infty}^{\rho} \frac{1}{\tau^{n-1}} e^{-\tau} d\tau + O\left(\Phi^2\right),$$

and we obtain the second approximation by substituting this back into (17). Repeating this, we reach

(19)
$$\begin{aligned} u - 1 = {} & -C\varepsilon^{n-2}E_{n-1} + \left(C\varepsilon^{n-2}\right)^2 \int_{\infty}^{\rho} \frac{1}{\tau^{n-1}} e^{-\tau} \left(\int_{\infty}^{\tau} E_{n-1}\,d\sigma\right) d\tau \\ & + \frac{1}{2}\left(C\varepsilon^{n-2}\right)^3 \int_{\infty}^{\rho} \frac{1}{\tau^{n-1}} e^{-\tau} \left(\int_{\infty}^{\tau} E_{n-1}\,d\sigma\right)^2 d\tau \\ & - \left(C\varepsilon^{n-2}\right)^3 \int_{\infty}^{\rho} \frac{1}{\tau^{n-1}} e^{-\tau} \int_{\infty}^{\tau} \left\{\int_{\infty}^{\sigma} \frac{1}{s^{n-1}} e^{-s} \left(\int_{\infty}^{s} E_{n-1}\,dt\right) ds\right\} d\sigma\,d\tau + O\left(\Phi^4\right), \end{aligned}$$

as $\Phi \to 0$.

To obtain C, we need to be able to evaluate each of these terms for small ρ (in particular, for $\rho = \varepsilon$), and this is a matter of integration by parts. Thus, for nonintegral n,

$$\begin{aligned} E_{n-1}(\rho) &= \int_{\rho}^{\infty} \frac{e^{-\tau}}{\tau^{n-1}} d\tau = -\frac{\rho^{2-n}}{2-n} e^{-\rho} + \frac{1}{2-n}\int_{\rho}^{\infty} \frac{e^{-\tau}}{\tau^{n-2}} d\tau \\ &= -\frac{\rho^{2-n}}{2-n} e^{-\rho} + \frac{1}{2-n} E_{n-2}, \end{aligned} \tag{20}$$

and this can be repeated to give E_{n-1} as a sum of terms of the form $c_k \rho^k e^{-\rho}$ and E_{n-p}, until $0 < n - p < 1$. Then

$$\begin{aligned} E_{n-p} &= \int_{0}^{\infty} \frac{e^{-\tau}}{\tau^{n-p}} d\tau - \int_{0}^{\rho} \frac{e^{-\tau}}{\tau^{n-p}} d\tau \\ &= \Gamma(p + 1 - n) - \int_{0}^{\rho} \frac{e^{-\tau}}{\tau^{n-p}} d\tau, \end{aligned}$$

and we can then continue to integrate by parts as far as we like. (If n is an integer, we will reach $\int_{\rho}^{\infty} \frac{e^{-\tau}}{\tau} d\tau$, which introduces a logarithm.)

Thus $E_{n-1}(\rho)$ can be expressed as a sum of terms of the form $c_k\rho^k e^{-\rho}$, and so obviously the same is true of E_{n-1}^2 , with $e^{-2\rho}$ in place of $e^{-\rho}$. Also,

$$\begin{aligned}\int_\infty^\rho E_{n-1}(\tau)\,d\tau &= \int_\infty^\rho \left(\int_\tau^\infty \frac{e^{-\sigma}}{\sigma^{n-1}}d\sigma\right)d\tau \\ &= \left[\tau\left(\int_\tau^\infty \frac{e^{-\sigma}}{\sigma^{n-1}}d\sigma\right)\right]\Big|_\infty^\rho + \int_\infty^\rho \frac{e^{-\tau}}{\tau^{n-2}}d\tau \\ &= \rho E_{n-1} - E_{n-2}, \qquad (21)\end{aligned}$$

so that $\int_\infty^\rho E_{n-1}d\tau$ can be expressed as the same type of sum. Hence the second term in (19) gives a sum of terms of the form $E_k(2\rho)$ and the third and fourth terms a sum involving $E_k(3\rho)$.

We now carry the process through in the most interesting cases, $n = 2, 3$.

4. The case $k = 0, n = 2$. When $n = 2$ we are interested in

$$\begin{aligned}E_1(\rho) &= \int_\rho^\infty \frac{1}{\tau}e^{-\tau}d\tau \\ &= -e^{-\rho}\log\rho + \int_\rho^\infty e^{-\tau}\log\tau d\tau \\ &= -e^{-\rho}\log\rho + \int_0^\infty e^{-\tau}\log\tau d\tau - \int_0^\rho e^{-\tau}\log\tau d\tau \\ &= -e^{-\rho}\log\rho - \gamma - \rho(\log\rho - 1)e^{-\rho} + O\left(\rho^2\log\rho\right), \text{ for small } \rho, \\ &= -\log\rho - \gamma + \rho + O\left(\rho^2\log\rho\right). \qquad (22)\end{aligned}$$

(See, for example, [2, Chapter 1].) Also, for future purposes, using (20) we obtain

$$E_2(\rho) = \frac{e^{-\rho}}{\rho} - E_1(\rho) \qquad (23)$$

$$= \frac{1}{\rho} + \log\rho + (\gamma - 1) - \frac{1}{2}\rho + O\left(\rho^2\log\rho\right) \text{ as } \rho \to 0. \qquad (24)$$

Looking now at (19), with $\rho = \varepsilon$, we see that as $\varepsilon \to 0$,

$$C\log\varepsilon \to -1$$

and

$$C = \frac{1}{\log\frac{1}{\varepsilon}} + O\left(\frac{1}{\left(\log\frac{1}{\varepsilon}\right)^2}\right).$$

Hence the series in (19) is in powers of $\frac{1}{\log\frac{1}{\varepsilon}}$.

Also, we will work our approximations (in order to compare the results with those of Hinch in [3]) to order $\frac{1}{\log^2\left(\frac{1}{\varepsilon}\right)}$, so that (for example)

$$u = \frac{a(r)}{\log\left(\frac{1}{\varepsilon}\right)} + \frac{b(r)}{\log^2\left(\frac{1}{\varepsilon}\right)} + O\left(\log^{-3}\frac{1}{\varepsilon}\right)$$

for any fixed value of r (ρ of order ε). This, as we shall see, necessitates finding

$$C = \frac{1}{\log\left(\frac{1}{\varepsilon}\right)}\left\{1 + \frac{A}{\log\left(\frac{1}{\varepsilon}\right)} + \frac{B}{\log^2\left(\frac{1}{\varepsilon}\right)} + O\left(\log^{-3}\left(\frac{1}{\varepsilon}\right)\right)\right\}$$

and requires use of all the terms in (19).

With this in mind, we look at the second term of (19). Thus from (21),

$$\int_\rho^\infty E_1 d\tau = -\rho E_1 + e^{-\rho}, \tag{25}$$

so that the second term is

$$\begin{aligned} C^2 \int_\infty^\rho \frac{1}{\tau} e^{-\tau}\left(\tau E_1 - e^{-\tau}\right) d\tau &= C^2\left\{\int_\infty^\rho e^{-\tau} E_1 d\tau - \int_\infty^\rho \frac{e^{-2\tau}}{\tau} d\tau\right\} \\ &= C^2\left\{\left[-e^{-\tau} E_1\right]|_\infty^\rho - 2\int_\infty^\rho \frac{e^{-2\tau}}{\tau} d\tau\right\} \\ &= C^2\left(-e^{-\rho} E_1(\rho) + 2E_1(2\rho)\right). \end{aligned} \tag{26}$$

From (22), the second term is therefore

$$\begin{aligned} &C^2\left(\log\rho + \gamma - 2\log 2\rho - 2\gamma + O(\rho)\right) \\ &= C^2\left(-\log\rho - \gamma - 2\log 2 + O(\rho)\right) \end{aligned} \tag{27}$$

as $\rho \to 0$.

In the third and fourth terms of (19) we need only the leading terms; i.e., we can ignore the equivalent of $\gamma + 2\log 2$ in (27). Using (25) the third term becomes

$$\frac{1}{2} C^3 \int_\infty^\rho \frac{e^{-\tau}}{\tau}\left(e^{-\tau} - \tau E_1\right)^2 d\tau = \frac{1}{2} C^3\left(\log\rho + O(1)\right) \text{ as } \rho \to 0. \tag{28}$$

Finally, in the fourth term, the integrand in the τ-integral is just the second term (as a function of σ), so that from (25), the fourth term is

$$M = -C^3 \int_\infty^\rho \frac{e^{-\tau}}{\tau}\left[\int_\infty^\tau \left\{-e^{-\sigma} E_1(\sigma) + 2E_1(2\sigma)\right\} d\sigma\right] d\tau. \tag{29}$$

It is seen from (25) that for any $\tau \leq \infty$, $\int_0^\tau E_1(\sigma)\, d\sigma$ converges. Hence we can write the inner integral above in the form $\int_\infty^0 + \int_0^\tau$, and it follows that

$$M = -C^3 \int_\infty^\rho \frac{e^{-\tau}}{\tau}\left\{K + r(\tau)\right\} d\tau,$$

where K is a constant, r is bounded, and $r(\tau) = O(\tau\log\tau)$ as $\tau \to 0$. It further follows that

$$M = C^3\left(K E_1(\rho) + O(1)\right) \text{ as } \rho \to 0.$$

We can evaluate K using (25) and (22):

$$\int_0^\infty E_1(2\sigma)\,d\sigma = \frac{1}{2}\int_0^\infty E_1(u)\,du = \frac{1}{2},$$

$$\int_0^\infty e^{-\sigma}E_1(\sigma)\,d\sigma = \left[-\left(e^{-\sigma}-1\right)E_1\right]_0^\infty - \int_0^\infty \left(e^{-\sigma}-1\right)\frac{e^{-\sigma}}{\sigma}d\sigma$$

$$= \lim_{\sigma\to 0}\left\{-E_1(2\sigma)+E_1(\sigma)\right\} = \lim_{\sigma\to 0}\left(\log 2\sigma - \log\sigma\right) = \log 2. \tag{30}$$

Hence, from (29), the fourth term of (19) is

$$C^3\left\{E_1(\rho)(\log 2-1)+O(1)\right\} = -C^3\left\{(\log 2-1)\log\rho + O(1)\right\} \text{ as } \rho\to 0. \tag{31}$$

Now setting $\rho=\varepsilon$ and using (27), (28), and (31), we obtain that

$$-1 = -C\left(-\log\varepsilon-\gamma+O(\varepsilon)\right) + C^2\left(-\log\varepsilon-\gamma-2\log 2+O(\varepsilon)\right)$$

$$+\frac{1}{2}C^3\left(\log\varepsilon+O(1)\right) - C^3\left\{(\log 2-1)\log\varepsilon+O(1)\right\}$$

as $\varepsilon\to 0$. Hence,

$$\frac{1}{\log\left(\frac{1}{\varepsilon}\right)} = C\left(1-\frac{\gamma}{\log\left(\frac{1}{\varepsilon}\right)}\right) - C^2\left(1-\frac{\gamma+2\log 2}{\log\left(\frac{1}{\varepsilon}\right)}\right)$$

$$+C^3\left(\frac{3}{2}-\log 2\right) + O\left(\log^{-4}\left(\frac{1}{\varepsilon}\right)\right), \tag{32}$$

and

$$C = \frac{1}{\log\left(\frac{1}{\varepsilon}\right)} + \frac{A}{\log^2\left(\frac{1}{\varepsilon}\right)} + \frac{B}{\log^3\left(\frac{1}{\varepsilon}\right)} + O\left(\frac{1}{\log^4\left(\frac{1}{\varepsilon}\right)}\right),$$

where

$$-\gamma + A - 1 = 0,$$

$$B - \gamma A - 2A + (\gamma + 2\log 2) + \frac{3}{2} - \log 2 = 0.$$

Hence,

$$A = \gamma + 1,$$

$$B = \gamma^2 + 2\gamma + \frac{1}{2} - \log 2.$$

Thus, for fixed r, ρ of order ε, we have, with $\lambda = \log\left(\frac{1}{\varepsilon}\right)$,

$$u-1 = \left(\frac{1}{\lambda} + \frac{\gamma+1}{\lambda^2} + \frac{(\gamma+1)^2 - \frac{1}{2} - \log 2}{\lambda^3}\right)(\log r + \log\varepsilon + \gamma)$$

$$+\frac{1}{\lambda^2}\left(1+\frac{2(\gamma+1)}{\lambda}\right)(-\log r - \log\varepsilon - \gamma - 2\log 2)$$

$$+\frac{1}{\lambda^3}\left(\frac{3}{2}-\log 2\right)(\log r + \log\varepsilon) + O\left(\lambda^{-4}\right),$$

so that, after cancellation,

$$u = \frac{\log r}{\lambda} + \frac{\gamma \log r}{\lambda^2} + O\left(\lambda^{-3}\right).$$

This is the "inner expansion." For the "outer expansion," i.e., fixed ρ, r of order $\frac{1}{\varepsilon}$, we use (19), truncated to second order, to get

$$u - 1 = -E_1(\rho)\left(\frac{1}{\lambda} + \frac{\gamma + 1}{\lambda^2}\right) + \frac{1}{\lambda^2}\left(2E_1(2\rho) - e^{-\rho}E_1(\rho)\right) + O\left(\lambda^{-3}\right).$$

These results are in accordance with those of Hinch and of others on this problem.

5. The case $k = 0, n = 3$. Here we are interested in (from (22) and (23))

$$E_2(\rho) = \frac{e^{-\rho}}{\rho} - E_1(\rho) = \frac{1}{\rho} + \log\rho + (\gamma - 1) - \frac{1}{2}\rho + O\left(\rho^2 \log\rho\right) \text{ as } \rho \to 0.$$

Thus, the first term on the right of (19) evaluated at $\rho = \varepsilon$ is

$$-C\left(1 + \varepsilon\log\varepsilon + (\gamma - 1) + O\left(\varepsilon^2\right)\right) \text{ as } \varepsilon \to 0.$$

The second term is

$$\begin{aligned}
&(C\varepsilon)^2 \int_\infty^\rho \frac{1}{\tau^2} e^{-\tau}\left(\int_\infty^\tau E_2 d\sigma\right) d\tau \\
&= (C\varepsilon)^2 \left\{\left[-E_2(\tau)\int_\infty^\tau E_2(\sigma)\, d\sigma\right]_\infty^\rho + \int_\infty^\rho E_2^2 d\tau\right\} \\
&= (C\varepsilon)^2 \left\{-E_2(\rho)\int_\infty^\rho E_2(\tau)\, d\tau + \int_\infty^\rho E_2^2 d\tau\right\}.
\end{aligned}$$

From (23) we see that

$$\int_\infty^\rho E_2^2 d\tau = -\frac{1}{\rho} + \log^2\rho + O(\log\rho) \text{ as } \rho \to 0,$$

while from (21),

$$\int_\infty^\rho E_2 d\tau = \rho E_2 - E_1 = 1 + \log\rho + \gamma + O(\rho\log\rho),$$

$$E_2 \int_\infty^\rho E_2 d\tau = \frac{1}{\rho}\log\rho + \frac{\gamma + 1}{\rho} + O\left(\log^2\rho\right).$$

In all, the second term is

$$(C\varepsilon)^2\left\{-\frac{1}{\rho}\log\rho - \frac{\gamma + 2}{\rho} + O\left(\log^2\rho\right)\right\}.$$

It is readily verified that the third and fourth terms in (19) give $O\{C^3\varepsilon^3(\frac{1}{\rho}\log^2\rho)\}$, which is negligible. Thus, evaluating (19) at $\rho = \varepsilon$, we have

$$-1 = -C\varepsilon\left(\frac{1}{\varepsilon} + \log\varepsilon + \gamma - 1\right) + (C\varepsilon)^2\left(-\frac{1}{\varepsilon}\log\varepsilon - \frac{\gamma + 2}{\varepsilon}\right) + O\left(C^3\varepsilon^2\log^2\varepsilon\right),$$

so that

$$C = 1 - 2\varepsilon \log \varepsilon - \varepsilon (2\gamma + 1) + O\left(\varepsilon^2 \log^2 \varepsilon\right).$$

Then, for fixed r, ρ of order ε, we have

$$u - 1 = -\varepsilon (1 - 2\varepsilon \log \varepsilon - \varepsilon (2\gamma + 1)) \left(\frac{1}{\varepsilon r} + \log \varepsilon + \log r + \gamma - 1\right)$$

$$+ \varepsilon^2 \left(-\frac{1}{\varepsilon r} (\log \varepsilon + \log r) - \frac{\gamma + 2}{\varepsilon r}\right) + O\left(\varepsilon^2 \log^2 \varepsilon\right),$$

$$u = 1 - \frac{1}{r} - \varepsilon \log \varepsilon \left(1 - \frac{1}{r}\right) - \varepsilon \left(\log r + \frac{\log r}{r}\right) + \varepsilon (1 - \gamma) \left(1 - \frac{1}{r}\right)$$

$$+ O\left(\varepsilon^2 \log^2 \varepsilon\right).$$

For fixed ρ, r of order ε^{-1}, we again use (19), to give

$$u - 1 = -\varepsilon (1 - 2\varepsilon \log \varepsilon - \varepsilon (2\gamma + 1)) E_2(\rho)$$

$$+ \varepsilon^2 \left\{E_1(\rho) E_2(\rho) - \rho E_2^2(\rho) - \int_\rho^\infty E_2^2 d\tau\right\} + O\left(\varepsilon^3\right). \tag{33}$$

Again, these results are in agreement with those of Hinch, and others, although (33) gives one term further.

Remark 3. It is of interest to consider what happens when $n < 2$, since, at least for $n \geq 1$, there still exists a unique solution. Equation (19) is still valid at $\rho = \varepsilon$, but since $E_{n-1}(\rho)$ is no longer singular at $\rho = 0$ for $n < 2$, (19) with $\rho = \varepsilon$ becomes merely an implicit equation for $C\varepsilon^{n-2}$. This tells us that $C \to 0$, since $\varepsilon^{n-2} \to \infty$, but we no longer get an asymptotic expansion. In particular, it is no longer obvious that C is unique. Of course, we know this from Theorem 1 if $n \geq 1$.

6. The case $k = 1$. We can in fact treat a generalization, which causes no further difficulties,

$$u'' + \frac{n-1}{r} u' + f(u) u'^2 + \varepsilon u\, u' = 0, \tag{34}$$

with the same boundary conditions. As before, we will compare our results with those of Hinch in [3].

As remarked in the proof of Theorem 1, the solution will necessarily have $u' > 0$ so that conditions on $f(u)$ are necessary only for $0 \leq u \leq 1$. We require only that f be continuous and positive in this interval.

Then (34) can be written as

$$\frac{\left(r^{n-1} u'\right)'}{r^{n-1} u'} + f(u) u' + \varepsilon u = 0,$$

so that

$$\log\left(r^{n-1} u'\right) = -F(u) - \varepsilon \int_1^r u dt + A$$

for some constant A, where

$$F(u) = \int_0^u f(s)\, ds.$$

This becomes

$$e^{F(u)} u' = \frac{C}{r^{n-1}} e^{-\varepsilon r - \varepsilon \int_\infty^r (u-1) ds},$$

or, on integration,

$$G(u) - G(1) = C \int_\infty^r \frac{1}{t^{n-1}} e^{-\varepsilon t - \varepsilon \int_\infty^t (u-1) ds} dt,$$

where

$$G(u) = \int_0^u e^{F(v)} dv.$$

In order to keep the manipulations simple and effect comparisons, we will consider from here the Lagerstrom model, where $f(u) = 1$, $F(u) = u$, $G(u) = e^u - 1$. Then, with $\varepsilon r = \rho$, $\varepsilon t = \tau$, we have

$$e^u - e = C\varepsilon^{n-2} \int_\infty^\rho \frac{1}{\tau^{n-1}} e^{-\tau} e^{-\int_\infty^\tau (u-1) d\sigma} d\tau, \tag{35}$$

and writing

$$u - 1 = \frac{u-1}{e^u - e} (e^u - e),$$

we get

$$e^u - e = C\varepsilon^{n-2} \int_\infty^\rho \frac{1}{\tau^{n-1}} e^{-\tau} e^{-\int_\infty^\tau \frac{u-1}{e^u-e}(e^u-e) d\sigma} d\tau. \tag{36}$$

As in section 3, we can integrate by parts, and since $0 \le \frac{u-1}{e^u-e} \le 1$ in $0 \le u < 1$, we will develop a convergent series as before. To get the first three terms (necessary to give Hinch's accuracy when $n = 2$), we have from (36) that

$$e^u - e = C\varepsilon^{n-2} \int_\infty^\rho \frac{1}{\tau^{n-1}} e^{-\tau} \left\{ 1 - \int_\infty^\tau \frac{u-1}{e^u-e} (e^u - e)\, d\sigma \right. \tag{37}$$

$$\left. + \frac{1}{2} \left(\int_\infty^\tau \frac{u-1}{e^u-e} (e^u - e)\, d\sigma \right)^2 + \cdots \right\} d\tau.$$

As before, since $e^u - e \to 0$ exponentially fast as $\rho \to \infty$, the series in the integrand converges uniformly for large τ, so that (37) is valid for large ρ. But again we need to extend it down to $\rho = \varepsilon$. From (35) we have

$$e^u - e \le C\varepsilon^{n-2} E_{n-1}(\rho),$$

and so the convergence proof is the same as that preceding (19).

Before proceeding further with $n = 2$, we make a couple of remarks about the simpler case $n > 2$. Then, as we saw in section 5, only two terms are necessary to give the required accuracy, and then (37) gives

$$e^u - u = C\varepsilon^{n-2} \int_\infty^\rho \frac{e^{-\tau}}{\tau^{n-1}} \left\{ 1 - \int_\infty^\tau \frac{u-1}{e^u - e} \left(e^u - e\right) d\sigma + \cdots \right\},$$

and since $\frac{u-1}{e^u-e}$ appears in what is already the highest order term, we can replace it by its limit as $u \to 1$, i.e., $\frac{1}{e}$. Thus we get, to the required order,

$$e^u - e = -C\varepsilon^{n-2} E_{n-1} - \left(\frac{C\varepsilon^{n-2}}{e} \right) \int_\infty^\rho \frac{e^{-\tau}}{\tau^{n-1}} \int_\infty^\tau \left(e^u - e\right) d\sigma d\tau.$$

(We will proceed more carefully for $n = 2$.) This, apart from the factor $\frac{1}{e}$, is the same equation dealt with in section 5 (with $e^u - e$ in place of $u - 1$), and the solution can be written down from there. If we had a general function f in place of 1, we would get

$$e^{F(u)} - e^{F(1)} = -C\varepsilon^{n-2} E_{n-1} - \frac{C\varepsilon^{n-2}}{e^{F(1)} f\left(1\right)} \int_\infty^\rho \frac{e^{-\tau}}{\tau^{n-1}} \left(\int_\infty^\tau \left(e^{F(u)} - e^{F(1)} \right) d\sigma \right) d\tau.$$

Turning now to the case $n = 2$ and $F\left(u\right) = u$, we need three terms on the right of (37). Thus,

$$\frac{u-1}{e^u - e} = \frac{1}{e} - \frac{1}{2e^2} \left(e^u - e\right) + O\left(e^u - e\right)^2 \text{ as } u \to 1. \tag{38}$$

We follow the method used just before (19) and obtain from (37) that

$$\begin{aligned}
&e^u - e \\
&= -C\varepsilon^{n-2} E_{n-1} + \frac{1}{e} \left(C\varepsilon^{n-2}\right)^2 \int_\infty^\rho \frac{1}{\tau^{n-1}} e^{-\tau} \left(\int_\infty^\tau E_{n-1} d\sigma \right) d\tau \\
&+ \frac{1}{2e^2} \left(C\varepsilon^{n-2}\right)^3 \int_\infty^\rho \frac{1}{\tau^{n-1}} e^{-\tau} \left(\int_\infty^\tau E_{n-1}^2 d\sigma \right) d\tau \\
&+ \frac{1}{2e^2} \left(C\varepsilon^{n-2}\right)^3 \int_\infty^\rho \frac{1}{\tau^{n-1}} e^{-\tau} \left(\int_\infty^\tau E_{n-1} d\sigma \right)^2 d\tau \\
&- \frac{1}{e^2} \left(C\varepsilon^{n-2}\right)^3 \int_\infty^\rho \frac{1}{\tau^{n-1}} e^{-\tau} \int_\infty^\tau \left\{ \int_\infty^\sigma \frac{1}{s^{n-1}} e^{-s} \left(\int_\infty^s E_{n-1} dt \right) ds \right\} d\sigma d\tau + O\left(\Phi^4\right) \\
&= -C\varepsilon^{n-2} E_{n-1} + F_1 + F_2 + F_3 + F_4 + O\left(\Phi^4\right),
\end{aligned}$$

say. As before, if $n > 2$, then this is valid for any C as $\varepsilon \to 0$, uniformly in $\rho \geq \varepsilon$, while if $n = 2$, it is valid as $C \to 0$.

For $n = 2$ we can continue to follow the argument in section 4. Thus, as $\rho \to 0$,

$$F_1 = \frac{1}{e} C^2 \left(-\log \rho - \gamma - 2\log 2 + O\left(\rho\right) \right),$$

$$F_3 = \frac{1}{2e^2} C^3 \left(\log \rho + O\left(1\right) \right),$$

$$F_4 = -\frac{1}{e^2} C^3 \left[\left(\log 2 - 1\right) \log \rho + O\left(1\right) \right].$$

The term F_2 did not appear before. Only the highest order term is needed for our expansion and this is

$$-\frac{1}{2e^2}C^3\left(\int_\infty^\rho \frac{1}{\tau}e^{-\tau}d\tau\right)\int_0^\infty E_1^2 d\sigma.$$

Now

$$\begin{aligned}\int_0^\infty E_1^2 d\sigma &= \left[\tau E_1^2\right]_0^\infty + 2\int_0^\infty \tau\frac{e^{-\tau}}{\tau}E_1 d\tau \\ &= 2\int_0^\infty e^{-\tau}E_1 d\tau = 2\log 2 \text{ from } (30).\end{aligned}$$

Thus,

$$F_2 = \frac{1}{e^2}C^3E_1\left(\log 2 + O\left(\rho\log^2\rho\right)\right) = -\frac{1}{e^2}C^3\left((\log 2)\log\rho + O(1)\right),$$

and, evaluating at $\rho = \varepsilon$, we have

$$\begin{aligned}1 - e &= C\left(\log\varepsilon + \gamma + O(\varepsilon)\right) \\ &\quad - \frac{1}{e}C^2\left(\log\varepsilon + \gamma + 2\log 2 + O(\varepsilon)\right) \\ &\quad + \frac{1}{2e^2}C^3\log\varepsilon\left(-2\log 2 + 1 - 2\log 2 + 2\right) + O\left(C^3\right), \\ \frac{e-1}{\log\frac{1}{\varepsilon}} &= C\left(1 - \frac{\gamma}{\log\left(\frac{1}{\varepsilon}\right)}\right) - \frac{1}{e}C^2\left(1 - \frac{\gamma + 2\log 2}{\log\left(\frac{1}{\varepsilon}\right)}\right) \\ &\quad + \frac{1}{2e^2}C^3\left(3 - 4\log 2 + O\left(\frac{C\varepsilon}{\log\varepsilon}\right) + O\left(\frac{C^2\varepsilon}{\log\varepsilon}\right) + O\left(\frac{C^3}{\log\varepsilon}\right)\right).\end{aligned}$$

Hence if

$$C = \frac{e-1}{\log\left(\frac{1}{\varepsilon}\right)} + \frac{A}{\log^2\left(\frac{1}{\varepsilon}\right)} + \frac{B}{\log^3\left(\frac{1}{\varepsilon}\right)} + O\left(\log^{-4}\left(\frac{1}{\varepsilon}\right)\right),$$

then

$$\begin{aligned}-\gamma(e-1) + A - \frac{(e-1)^2}{e} &= 0, \\ A &= \frac{e-1}{e}(\gamma e + e - 1),\end{aligned}$$

$$B - A\gamma + \frac{(e-1)^2}{e}(\gamma + 2\log 2) - \frac{2A(e-1)}{e} + \frac{1}{2e^2}(e-1)^3(3 - 4\log 2) = 0.$$

We can of course calculate B, but in fact its value will be irrelevant to the level of approximation that we take.

Then, for fixed r (ρ of order ε), we have, with $l = \log\left(\frac{1}{\varepsilon}\right)$,

$$\begin{aligned} e^u - e &= (e-1)\left\{\frac{1}{l} + \frac{\gamma + 1 - \frac{1}{e}}{l^2} + \frac{B/(e-1)}{l^3}\right\}(\log\varepsilon + \log r + \gamma) \\ &\quad + \frac{1}{e}(e-1)^2\left\{\frac{1}{l^2} + \frac{2\left(\gamma + 1 - \frac{1}{e}\right)}{l^3}\right\}(-\log\varepsilon - \log r - \gamma - 2\log 2) \\ &\quad + \frac{1}{2e^2}\frac{(e-1)^3}{l^3}(3 - 4\log 2)(\log\varepsilon + \log r) + O\left(l^{-3}\right) \\ &= 1 - e + \frac{(e-1)\log r}{l} + \frac{\gamma(e-1)\log r}{l^2} + O\left(l^{-3}\right). \end{aligned} \tag{39}$$

(Note that the definitions of A and B were such that $u = 0$ at $r = 1$ up to and including order l^{-2}, so that to that order there can be only terms in $\log r$, not constant terms. We do not need the explicit value of B.) To obtain u, we have to invert, so that

$$u = \log\left\{1 + \frac{e-1}{l}\log r + \frac{\gamma(e-1)}{l^2}\log r + O\left(l^{-3}\right)\right\}.$$

For fixed ρ, r of order ε^{-1}, we have

$$e^u - e = -\frac{e-1}{l}\left(1 + \frac{\gamma + 1 - \frac{1}{e}}{l}\right)E_1(\rho) + \frac{(e-1)^2}{el^2}\left(2E_1(2\rho) - e^{-\rho}E_1(\rho)\right) + O\left(l^{-3}\right).$$

Thus

$$\begin{aligned} u - 1 &= \frac{1}{e}(e^u - e) - \frac{1}{2e^2}(e^u - e)^2 + \cdots \\ &= -\frac{e-1}{e}\left(1 + \frac{\gamma + 1 - \frac{1}{e}}{l}\right)\frac{E_1(\rho)}{l} + \frac{(e-1)^2}{e^2}\frac{(2E_1(2\rho) - e^{-\rho}E_1(\rho))}{l^2} \\ &\quad - \frac{(e-1)^2}{2e^2}\frac{E_1^2(\rho)}{l^2} + O\left(l^{-3}\right). \end{aligned} \tag{40}$$

Again, these results are consistent with those of Hinch and others, except that Hinch has an algebraic mistake which in (40) replaces $\gamma + 1 - \frac{1}{e}$ by $\gamma - 1 + \frac{1}{e}$.

7. Final remarks. Starting with Lagerstrom, the terms involving $\log\varepsilon$ in the inner expansions have been considered difficult to explain. They are often called "switchback" terms, because there is nothing obvious in the equation which indicates the need for such terms, and because, starting with an expansion in powers of ε, one finds inconsistent results which are resolved only by adding terms of lower order, that is, powers of $\varepsilon\log\varepsilon$. The recent approach to the problem by geometric perturbation theory explains this by reference to a "resonance phenomenon," which is too complicated for us to describe here [10], [11].

In our work, the necessity for such terms is seen already from (11) and the resulting expansion (17):

$$u(\rho) - 1 = C\varepsilon^{n-2}\int_\infty^\rho \frac{1}{\tau^{n-1}}e^{-\tau}\left\{1 - \int_\infty^\tau (u-1)\,d\sigma + \frac{1}{2}\left(\int_\infty^\tau (u-1)\,d\sigma\right)^2 - \cdots\right\}d\tau.$$

In the existence proof it was seen in (9) that $C = O(1)$ as $\varepsilon \to 0$. On the right of (17) the first term is simply $-C\varepsilon^{n-2}E_{n-1}(\rho)$, and the simple expansions given for E_1 and E_2 show immediately the need for the logarithmic terms. There is no "switchback," because the procedure does not start with any assumption about the nature of the expansion, and there is no need for a "matching."

A number of authors have noted that the outer expansion is a uniformly valid asymptotic expansion on $[1, \infty)$, and therefore it "contains" the inner expansion [5], though this is more subtle when $k = 1$ [13]. Our twist on this is that both expansions are contained in the uniformly convergent series defined implicitly by (17). The simple derivation of this series via the integral equation (11) is new, as far as we know.

Acknowledgment. We thank the referees for some very helpful comments. In particular, they called our attention to earlier proofs of the existence and uniqueness results, in some cases by techniques similar to ours.

REFERENCES

[1] D. S. Cohen, A. Fokas, and P. A. Lagerstrom, *Proof of some asymptotic results for a model equation for low Reynolds number flow*, SIAM J. Appl. Math., 35 (1978), pp. 187–207.
[2] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Higher Transcendental Functions*, Vol. 1, McGraw-Hill, New York, 1953.
[3] E. J. Hinch, *Perturbation Methods*, Cambridge University Press, Cambridge, UK, 1991.
[4] G. C. Hsiao, *Singular perturbations for a nonlinear differential equation with a small parameter*, SIAM J. Math Anal., 4 (1973), pp. 283–301.
[5] C. Hunter, M. Tajdari, and S. D. Boyer, *On Lagerstrom's model of slow incompressible viscous flow*, SIAM J. Appl. Math., 50 (1990), pp. 48–63.
[6] S. Kaplun and P. A. Lagerstrom, *Asymptotic expansions of Navier-Stokes solutions for small Reynolds number*, J. Math. Mech., 6 (1957), pp. 585–593.
[7] P. A. Lagerstrom and R. G. Casten, *Basic concepts underlying singular perturbation techniques*, SIAM Rev., 14 (1972), pp. 63–120.
[8] P. A. Lagerstrom and C. A. Reinelt, *Note on logarithmic switchback terms in regular and singular perturbation expansions*, SIAM J. Appl. Math. 44 (1984), pp. 451–462.
[9] A. D. MacGillivray, *On a model equation of Lagerstrom*, SIAM J. Appl. Math., 34 (1978), pp. 804–812.
[10] N. Popovic and P. Szmolyan, *A geometric analysis of the Lagerstrom model problem*, J. Differential Equations, 199 (2004), pp. 290–325.
[11] N. Popovic and P. Szmolyan, *Rigorous asymptotic expansions for Lagerstrom's model equations, a geometric approach*, Nonlinear Anal., 59 (2004), pp. 531–565.
[12] S. Rosenblat and J. Shepherd, *On the asymptotic solution of the Lagerstrom model equation*, SIAM J. Appl. Math., 29 (1975), pp. 110–120.
[13] L. A. Skinner, *Note on the Lagerstrom singular perturbation models*, SIAM J. Appl. Math., 41 (1981), pp. 362–364.
[14] K. Tam, *On the Lagerstrom model for flow at low Reynolds numbers*, J. Math. Anal. Appl., 49 (1975), pp. 286–294.

SIAM J. MATH. ANAL.
Vol. 40, No. 6, pp. 2437–2450

VECTOR FIELDS FOR MEAN VALUE COORDINATES*

S. L. LEE†

Abstract. We find vector fields $\mathbf{F}$ that provide the representation $\mathbf{v} = \int_{\partial D} \mathbf{r}\,\mathbf{F} \cdot d\mathbf{r} / \int_{\partial D} \mathbf{F} \cdot d\mathbf{r}$ for any compact 2-D manifold $D \subset \mathbb{R}^2$ with piecewise smooth boundary ∂D and $\mathbf{v} \in \mathbb{R}^2$, and the representation $\mathbf{v} = \iint_{\partial M} \mathbf{r}\,\mathbf{F} \cdot d\mathbf{S} / \iint_{\partial M} \mathbf{F} \cdot d\mathbf{S}$ for any compact 3-D manifold $M \subset \mathbb{R}^3$ and $\mathbf{v} \in \mathbb{R}^3$. Our method exploits properties of conservative fields in $\mathbb{R}^2$ and divergence free vector fields in $\mathbb{R}^3$. Discrete versions, which are more general than Floater's mean value coordinates, are derived from the above representations with a special choice of $\mathbf{F}$, either by taking points on the boundaries of $D \subset \mathbb{R}^2$ and $M \subset \mathbb{R}^3$ or by considering representations on boundaries of polygons in $\mathbb{R}^2$ or polyhedra in $\mathbb{R}^3$.

Key words. star-shaped region, mean value coordinates, conservative and divergence free vector fields, Stoke's theorem, divergence theorem, spherical harmonics

AMS subject classifications. 41A05, 52A30, 52B10, 52B70

DOI. 10.1137/070694144

1. Introduction.

Representation of points and functionals on a set by its extreme points or boundary is an important problem in mathematics and its applications. Barycentric coordinates, the Krein–Millman theorem, and Choquet's theorem are examples of such a representation. Recently, in conjunction with the construction of one-one transformations and parametrizations of meshes in $\mathbb{R}^3$, Floater [2] has found an explicit formula for the representation of points that lie in the kernel of star-shaped polygons in $\mathbb{R}^2$ in terms of extreme points of the polygon. Because of its usefulness in computer graphics and its potential applications in functional approximation and interpolation, the idea has been quickly extended to star-shaped polyhedrons in $\mathbb{R}^3$ (see [3], [4], [5], [6], [7]). Floater's original idea was motivated by the mean value property of harmonic functions; i.e., if ϕ is harmonic in a region $\Omega \subset \mathbb{R}^2$, then for any disc $D(\mathbf{v}, r) \subset \Omega$ with center at $\mathbf{v}$ and radius r,

$$\phi(\mathbf{v}) = \frac{1}{2\pi r} \int_{\partial D} \phi(\mathbf{r}) ds, \tag{1.1}$$

where $ds = \left|\frac{d\mathbf{r}}{dt}\right| dt$ for a parametrization $\mathbf{r} = \mathbf{r}(t)$ of ∂D. He observed that a class of real-valued piecewise linear functions defined on a triangular mesh in $\mathbb{R}^2$, which he calls *convex combination functions*, shares discretely some properties of harmonic functions. Forcing the mean value property (1.1) on the convex combination functions produces the new coordinates, which he calls the *mean value coordinates.* The problem of computing these coordinates then reduces to integrating elementary trigonometric functions over the unit circle or the unit sphere (see [2], [3], [7]).

This paper explores the connection between mean value coordinates with conservative vector fields on compact 2-D manifolds in $\mathbb{R}^2$ and divergence free vector fields on compact 3-D manifolds in $\mathbb{R}^3$ that provide a mathematical foundation, and puts

*Received by the editors June 9, 2007; accepted for publication (in revised form) November 8, 2008; published electronically February 25, 2009.
http://www.siam.org/journals/sima/40-6/69414.html
†Department of Mathematics, National University of Singapore, 2 Science Drive 2, Singapore 117543 (matleesl@nus.edu.sg).

mean value coordinates in a more general mathematical setting. In particular we find vector fields $\mathbf{F}$ that provide the representation

$$\mathbf{v} = \frac{\int_{\partial D} \mathbf{r}\ \mathbf{F}(\mathbf{r}-\mathbf{v}) \cdot d\mathbf{r}}{\int_{\partial D} \mathbf{F}(\mathbf{r}-\mathbf{v}) \cdot d\mathbf{r}}$$

for any compact 2-D manifold $D \subset \mathbb{R}^2$ with piecewise smooth boundary ∂D and for any vector $\mathbf{v} \in \mathbb{R}^2$, and

$$\mathbf{v} = \frac{\iint_{\partial M} \mathbf{r}\ \mathbf{F}(\mathbf{r}-\mathbf{v}) \cdot d\mathbf{S}}{\iint_{\partial M} \mathbf{F}(\mathbf{r}-\mathbf{v}) \cdot d\mathbf{S}}$$

for any compact 3-D manifold $M \subset \mathbb{R}^3$ and $\mathbf{v} \in \mathbb{R}^3$. We shall refer to these as *mean value representations.* Our method exploits properties of conservative and divergence free vector fields. With a special choice of $\mathbf{F}$ and restricting to points on the boundaries of the manifolds or by considering the representations on the boundaries of polygons in $\mathbb{R}^2$ and polyhedra in $\mathbb{R}^3$, discrete versions of mean value representation are obtained, which are more general than Floater's mean value coordinates.

In section 2 a search for vector fields that provide the mean value representation in $\mathbb{R}^2$ leads to a class of vector fields

$$\mathbf{F} = \left[\frac{1}{xy}\Phi\left(\frac{x}{r},\frac{y}{r}\right), \frac{-1}{y^2}\Phi\left(\frac{x}{r},\frac{y}{r}\right)\right]^T,$$

where Φ is an arbitrary real-valued differentiable function on the unit circle $\mathbb{S}^1$. For these $\mathbf{F}$, the vector fields $x\mathbf{F}$, $y\mathbf{F}$ are conservative on $\mathbb{R}^2 \setminus \{\mathbf{0}\}$. Further, if

$$\Phi(\cos\theta, \sin\theta) = -\cos\theta \sin^2\theta \left(1 + \sum_{k=2}^{\infty} a_k \cos k\theta + b_k \sin k\theta\right),$$

with $|\sum_{k=2}^{\infty} a_k \cos k\theta + b_k \sin k\theta| \le 1$, then the corresponding $\mathbf{F}$ provides a mean value representation. A more general form of Floater's mean value coordinates is derived from the particular choice of $\Phi(\cos\theta, \sin\theta) = -\cos\theta \sin^2\theta$, i.e., $\Phi(\frac{x}{r}, \frac{y}{r}) = -\frac{xy^2}{r^3}$, and hence $\mathbf{F}(x,y) := [-y/r^3, x/r^3]^T$. The derivation exploits properties of path independent integrals of conservative fields. A similar search in section 3 for vector fields that provide the mean value representations in $\mathbb{R}^3$ leads to a class

$$\mathbf{F}(\mathbf{r}) = \left[\frac{1}{xyz}\Phi\left(\frac{x}{\rho},\frac{y}{\rho},\frac{z}{\rho}\right), \frac{1}{x^2 z}\Phi\left(\frac{x}{\rho},\frac{y}{\rho},\frac{z}{\rho}\right), \frac{1}{x^2 y}\Phi\left(\frac{x}{\rho},\frac{y}{\rho},\frac{z}{\rho}\right)\right]^T,$$

where $\Phi : \mathbb{S}^2 \to \mathbb{R}$ is an arbitrary differentiable function and $\rho := (x^2+y^2+z^2)^{1/2}$. In this case $x\mathbf{F}$, $y\mathbf{F}$, and $z\mathbf{F}$ are divergence free on $\mathbb{R}^3 \setminus \{\mathbf{0}\}$. Further, if

$$\begin{aligned}\Phi\left(\frac{x}{\rho},\frac{x}{\rho},\frac{x}{\rho}\right) &= \frac{x^2 yz}{\rho^4}\left(1 + \Psi\left(\frac{x}{\rho},\frac{x}{\rho},\frac{x}{\rho}\right)\right) \\ &= \sin^3\phi \cos^2\theta \sin\theta \cos\phi\ (1 + \Psi(\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi)),\end{aligned}$$

where Ψ is orthogonal to the first order spherical harmonics and $|\Psi| \le 1$, then $\mathbf{F}$ provides a mean value representation in $\mathbb{R}^3$. The derivation of Floater's type mean value coordinates in $\mathbb{R}^3$ is carried out in section 4. Here we take $\Psi = 0$, and hence

$\Phi(\frac{x}{\rho}, \frac{y}{\rho}, \frac{z}{\rho}) = \frac{x^2yz}{\rho^4}$ and $\mathbf{F} = \frac{\mathbf{r}}{\rho^4}$, and consider compact 3-D manifolds with polyhedral faces of any topology. The integral of $\iint_{\partial M} \mathbf{r}\mathbf{F} \cdot d\mathbf{S}$ over the boundary of such a manifold M is the sum of the integrals over its faces. The integral $\iint_S \mathbf{r}\mathbf{F} \cdot d\mathbf{S}$ over each face S is evaluated using Stoke's theorem. This is done by finding vector fields $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ such that $x\mathbf{F} = curl(\mathbf{A})$ $y\mathbf{F} = curl(\mathbf{B})$, $z\mathbf{F} = curl(\mathbf{C})$, which is possible since $\operatorname{div}(x\mathbf{F}) = \operatorname{div}(y\mathbf{F}) = \operatorname{div}(z\mathbf{F}) = 0$. This leads to a simple formula for the integrals over the faces of the boundary of M, from which discrete versions of the mean value representation or Floater's type mean value coordinates are obtained. If all the faces are triangular, the procedure leads to a unique solution. In all other cases the solutions are not unique.

2. Representation on boundaries of compact 2-D manifolds in $\mathbb{R}^2$. Let $D \subset \mathbb{R}^2$ be a compact 2-D manifold with piecewise smooth boundary ∂D, which is assumed to be oriented in the positive direction; i.e., if one travels along this direction, the region D lies on his left. We want to find a vector field $\mathbf{F} = [F_1(x, y), F_2(x, y)]^T$ such that for any $D \subset \mathbb{R}^2$ and $\mathbf{v} \in \mathbb{R}^2$,

$$\int_{\partial D} (\mathbf{r} - \mathbf{v})\mathbf{F}(\mathbf{r} - \mathbf{v}) \cdot d\mathbf{r} = 0 \tag{2.1}$$

and

$$\int_{\partial D} \mathbf{F}(\mathbf{r} - \mathbf{v}) \cdot d\mathbf{r} \neq 0, \tag{2.2}$$

so that

$$\mathbf{v} = \frac{\int_{\partial D} \mathbf{r}\ \mathbf{F}(\mathbf{r} - \mathbf{v}) \cdot d\mathbf{r}}{\int_{\partial D} \mathbf{F}(\mathbf{r} - \mathbf{v}) \cdot d\mathbf{r}}. \tag{2.3}$$

Without loss of generality, we may assume that $\mathbf{v} = \mathbf{0}$ so that the problem reduces to finding $\mathbf{F}$ such that for any D

(2.4)

$$\begin{aligned}\int_{\partial D} \mathbf{r}\ \mathbf{F} \cdot d\mathbf{r} &= \left[\int_{\partial D} xF_1(x, y)dx + xF_2(x, y)dy, \int_{\partial D} yF_1(x, y)dx + yF_2(x, y)dy\right]^T \\ &= \mathbf{0}\end{aligned}$$

and (2.2) holds, i.e., $\int_{\partial D} \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} \neq 0$.

First, suppose that $\mathbf{F}$ is continuously differentiable on a domain that contains D. By Green's theorem, (2.4) holds if

$$\frac{\partial(xF_1)}{\partial y} = \frac{\partial(xF_2)}{\partial x} \quad \text{and} \quad \frac{\partial(yF_1)}{\partial y} = \frac{\partial(yF_2)}{\partial x},$$

or, equivalently,

$$x\frac{\partial F_2}{\partial x} - x\frac{\partial F_1}{\partial y} = -F_2 \quad \text{and} \quad y\frac{\partial F_2}{\partial x} - y\frac{\partial F_1}{\partial y} = F_1. \tag{2.5}$$

The equations in (2.5) give $F_2 = \frac{-x}{y}F_1$, which together with the first equation of (2.5) lead to

$$x\frac{\partial F_1}{\partial x} + y\frac{\partial F_1}{\partial y} = -2F_1. \tag{2.6}$$

The general solution of (2.6) that vanishes at infinity is $F_1(x,y) = \frac{1}{xy}\Phi\left(\frac{x}{r},\frac{y}{r}\right)$, where Φ is an arbitrary real-valued differentiable function defined on the unit circle $\mathbb{S}^1$ and $r := ||\mathbf{r}|| = (x^2+y^2)^{1/2}$. Further, $\Phi\left(\frac{x}{r},\frac{y}{r}\right)$ is the general solution of the corresponding homogeneous equation.

PROPOSITION 2.1. *Let*

$$\mathbf{F} = [F_1, F_2]^T := \left[\frac{1}{xy}\Phi\left(\frac{x}{r},\frac{y}{r}\right), \frac{-1}{y^2}\Phi\left(\frac{x}{r},\frac{y}{r}\right)\right]^T, \tag{2.7}$$

where Φ is an arbitrary real-valued differentiable function on the unit circle $\mathbb{S}^1$. Then

$$x\mathbf{F} = \left[\frac{1}{y}\Phi\left(\frac{x}{r},\frac{y}{r}\right), \frac{-x}{y^2}\Phi\left(\frac{x}{r},\frac{y}{r}\right)\right]^T \quad \textit{and} \quad y\mathbf{F} = \left[\frac{1}{x}\Phi\left(\frac{x}{r},\frac{y}{r}\right), \frac{-1}{y}\Phi\left(\frac{x}{r},\frac{y}{r}\right)\right]^T \tag{2.8}$$

are conservative vector fields on any region not containing the origin with potential functions ϕ_1 and ϕ_2, respectively, where

$$\phi_1(x,y) = \int \frac{1}{y}\Phi\left(\frac{x}{r},\frac{y}{r}\right)\, dx = -\int \frac{x}{y^2}\Phi\left(\frac{x}{r},\frac{y}{r}\right)\, dy, \tag{2.9}$$

$$\phi_2(x,y) = \int \frac{1}{x}\Phi\left(\frac{x}{r},\frac{y}{r}\right)\, dx = -\int \frac{1}{y}\Phi\left(\frac{x}{r},\frac{y}{r}\right)\, dy. \tag{2.10}$$

Proof. Using the fact that $x\frac{\partial\Phi}{\partial x} + y\frac{\partial\Phi}{\partial y} = 0$, it is straightforward to verify that $x\mathbf{F} = \bigtriangledown\phi_1$ and $y\mathbf{F} = \bigtriangledown\phi_2$. □

It follows from Proposition 2.1 that the integrals of $x\mathbf{F}$ and $y\mathbf{F}$ over the boundary of any 2-D manifold in $\mathbb{R}^2$ that does not contain the origin are zero. We also require their integrals over any closed curve that encloses the origin to be zero, and it suffices to integrate over the unit circle $\mathbb{S}^1$. We have

$$\int_{\mathbb{S}^1} x\mathbf{F}\cdot d\mathbf{r} = -\int_0^{2\pi} \frac{1}{\sin^2\theta}\Phi(\cos\theta,\sin\theta)d\theta, \tag{2.11}$$

$$\int_{\mathbb{S}^1} y\mathbf{F}\cdot d\mathbf{r} = -\int_0^{2\pi} \frac{1}{\cos\theta\sin\theta}\Phi(\cos\theta,\sin\theta)d\theta. \tag{2.12}$$

If

$$\Phi(\cos\theta,\sin\theta) = -\cos\theta\sin^2\theta(1+\psi(\cos\theta,\sin\theta)), \tag{2.13}$$

where

$$\psi(\cos\theta,\sin\theta) := \sum_{k=2}^{\infty} a_k\cos k\theta + b_k\sin k\theta$$

is a differentiable function, then (2.11) and (2.12) give

$$\int_{\mathbb{S}^1} x\mathbf{F}\cdot d\mathbf{r} = \int_0^{2\pi} \cos\theta(1+\psi(\cos\theta,\sin\theta))d\theta = 0, \tag{2.14}$$

$$\int_{\mathbb{S}^1} y\mathbf{F}\cdot d\mathbf{r} = \int_0^{2\pi} \sin\theta(1+\psi(\cos\theta,\sin\theta))d\theta = 0, \tag{2.15}$$

since $1+\psi$ is orthogonal to $\cos\theta$ and $\sin\theta$.

We now state the results and complete the proof.

THEOREM 2.2. *Let* $\mathbf{F}$ *be as in Proposition* 2.1 *with* Φ *in* (2.13). *Then for any compact* 2*-D manifold* $D \subset \mathbb{R}^2$ *with piecewise smooth boundary* ∂D *and for* $\mathbf{v} \in \mathbb{R}^2 \setminus \partial D$,

$$\int_{\partial D} (\mathbf{r}-\mathbf{v})\ \mathbf{F}(\mathbf{r}-\mathbf{v}) \cdot d\mathbf{r} = \mathbf{0}. \tag{2.16}$$

Further, if $|\psi| \le 1$,

$$\mathbf{v} = \frac{\int_{\partial D} \mathbf{r}\ \mathbf{F}(\mathbf{r}-\mathbf{v}) \cdot d\mathbf{r}}{\int_{\partial D} \mathbf{F}(\mathbf{r}-\mathbf{v}) \cdot d\mathbf{r}} \tag{2.17}$$

for any $\mathbf{v} \in \mathbb{R}^2$.

We first prove a lemma, from which one can deduce that (2.2) holds if $|\psi| \le 1$.

LEMMA 2.3. *For any compact* 2*-D manifold* $D \subset \mathbb{R}^2$ *with piecewise smooth boundary,*

$$\int_{\partial D} \mathbf{F} \cdot d\mathbf{r} = \begin{cases} -\iint_D \frac{1}{r^3}\left(1+\psi\left(\frac{x}{r},\frac{y}{r}\right)\right)\ dxdy & \text{if } \mathbf{0} \notin D, \\ \iint_{\mathbb{R}^2\setminus D} \frac{1}{r^3}\left(1+\psi\left(\frac{x}{r},\frac{y}{r}\right)\right)\ dxdy & \text{if } \mathbf{0} \in D. \end{cases} \tag{2.18}$$

In particular, $\int_{\partial D} \mathbf{F} \cdot d\mathbf{r} \neq 0$ *if* $|\psi| \le 1$.

Proof. If $\mathbf{0} \notin D$, by Green's theorem

$$\begin{aligned} \int_{\partial D} \mathbf{F} \cdot d\mathbf{r} &= \iint_D \frac{\partial}{\partial x}\left(\frac{-1}{y^2}\Phi\right) - \frac{\partial}{\partial y}\left(\frac{1}{xy}\Phi\right)\ dxdy \\ &= \iint_D \frac{1}{xy^2}\Phi\left(\frac{x}{r},\frac{y}{r}\right) - \frac{1}{xy^2}\left\{x\frac{\partial\Phi}{\partial x} + y\frac{\partial\Phi}{\partial y}\right\}\ dxdy \\ &= \iint_D \frac{1}{xy^2}\Phi\left(\frac{x}{r},\frac{y}{r}\right)\ dxdy \\ &= -\iint_D \frac{1}{r^3}\left(1+\psi\left(\frac{x}{r},\frac{y}{r}\right)\right)\ dxdy. \end{aligned} \tag{2.19}$$

If $\mathbf{0} \in D$, take a disc D_R, with center at the origin and radius R, that contains D. By (2.19),

$$\int_{\partial D_R} \mathbf{F} \cdot d\mathbf{r} - \int_{\partial D} \mathbf{F} \cdot d\mathbf{r} = -\iint_{D_R\setminus D} \frac{1}{r^3}\left(1+\psi\left(\frac{x}{r},\frac{y}{r}\right)\right)\ dxdy$$

so that

$$\begin{aligned} \int_{\partial D} \mathbf{F} \cdot d\mathbf{r} &= \int_{\partial D_R} \mathbf{F} \cdot d\mathbf{r} + \iint_{D_R\setminus D} \frac{1}{r^3}\left(1+\psi\left(\frac{x}{r},\frac{y}{r}\right)\right)\ dxdy \\ &= -\frac{1}{R}\int_0^{2\pi} 1+\psi(\cos\theta,\sin\theta)d\theta + \iint_{D_R\setminus D} \frac{1}{r^3}\left(1+\psi\left(\frac{x}{r},\frac{y}{r}\right)\right)\ dxdy. \end{aligned}$$

Taking the limit as $R \to \infty$ gives the second formula in (2.18). ∎

Proof of Theorem 2.2. If $\mathbf{v} \notin \partial D$, (2.16) follows from Proposition 2.1 and (2.14) and (2.15). Equation (2.17) follows from (2.16) and Lemma 2.3.

If $\mathbf{v} \in \partial D$, for any $\epsilon > 0$, we replace an appropriate segment of ∂D that contains $\mathbf{v}$ by a circular arc with center at $\mathbf{v}$ and radius ϵ, so that $\mathbf{v}$ lies outside the resulting piecewise smooth curve, which we call ∂D_ϵ. Then (2.17) holds for ∂D_ϵ for all $\epsilon > 0$, and therefore holds in the limit as $\epsilon \to 0$. ∎

If we choose $\psi = 0$ in (2.13) so that

$$\Phi(\cos\theta, \sin\theta) = -\cos\theta \sin^2\theta, \tag{2.20}$$

a more general form of Floater's mean value coordinates [2] can be deduced as a corollary. The proof here, which is different from the existing methods (see [2], [3], [5] [6]), provides a better understanding why they work.

COROLLARY 2.4. *Let $D \subset \mathbb{R}^2$ be a compact 2-D manifold with piecewise smooth boundary ∂D, and let $\mathbf{v}_j \in \partial D$, $j = 1, 2, \ldots, k > 2$, be arranged in the positive orientation. Then for any $\mathbf{v} \in \mathbb{R}^2$,*

$$\mathbf{v} = \frac{\sum_{j=1}^k w_j \mathbf{v}_j}{\sum_{j=1}^k w_j}, \tag{2.21}$$

where

$$w_j := \frac{\tan(\alpha_{j-1}/2) + \tan(\alpha_j/2)}{||\mathbf{v}_j - \mathbf{v}||}, \quad j = 1, 2, \ldots, k-1, \tag{2.22}$$

and α_j is the angle at $\mathbf{v}$ of the oriented triangle $[\mathbf{v}, \mathbf{v}_j, \mathbf{v}_{j+1}]$ with $|\alpha_j| < \pi$ and takes a positive value if the vector $\frac{\mathbf{v}_j}{||\mathbf{v}_j||} - \frac{\mathbf{v}_{j+1}}{||\mathbf{v}_{j+1}||}$ is in the counterclockwise orientation and a negative value otherwise.

Proof. Take $\psi = 0$ in (2.13) so that $\Phi\left(\frac{x}{r}, \frac{y}{r}\right) = -\frac{xy^2}{r^3}$ and $\mathbf{F}(\mathbf{r}) \equiv \mathbf{F}(x, y) := [-y/r^3, x/r^3]^T$. By (2.9) and (2.10), $x\mathbf{F} = \bigtriangledown\phi_1$ and $y\mathbf{F} = \bigtriangledown\phi_2$, where

$$\phi_1(x, y) = -\frac{1}{y} \int \frac{xy^2}{r^3}\, dx = \frac{y}{r}, \tag{2.23}$$

$$\phi_2(x, y) = \int \frac{xy^2}{yr^3}\, dy = \frac{-x}{r}. \tag{2.24}$$

Hence the vector $[\phi_1(x, y), \phi_2(x, y)]^T = [y, -x]^T/r$ is the rotation of the $[x, y]^T/r$ along the unit circle in the clockwise direction through an angle of $\pi/2$. Let $\mathbf{v}_j = [x_j, y_j]$, $j = 1, 2, \ldots, k$, $\mathbf{v}_{k+1} := \mathbf{v}_1$. Then for $j = 1, 2, \ldots, k$,

$$\begin{aligned}\int_{\mathbf{v}_j}^{\mathbf{v}_{j+1}} \mathbf{r}\ \mathbf{F} \cdot d\mathbf{r} &= [\phi_1(\mathbf{v}_{j+1}), \phi_2(\mathbf{v}_{j+1})]^T - [\phi_1(\mathbf{v}_j), \phi_2(\mathbf{v}_j)]^T \\ &= [y_{j+1}, -x_{j+1}]^T/||\mathbf{v}_{j+1}|| - [y_j, -x_j]^T/||\mathbf{v}_j||.\end{aligned}$$

The vector $[y_{j+1}, -x_{j+1}]^T/||\mathbf{v}_{j+1}|| - [y_j, -x_j]^T/||\mathbf{v}_j||$ is parallel to $\mathbf{v}_{j+1}/||\mathbf{v}_{j+1}|| + \mathbf{v}_j/||\mathbf{v}_j||$ and

$$\begin{aligned}\left\| \frac{[y_{j+1}, -x_{j+1}]^T}{||\mathbf{v}_{j+1}||} - \frac{[y_j, -x_j]^T}{||\mathbf{v}_j||} \right\| &= \left\| \frac{\mathbf{v}_{j+1}}{||\mathbf{v}_{j+1}||} - \frac{\mathbf{v}_j}{||\mathbf{v}_j||} \right\| \\ &= \tan(\alpha_j/2) \left\| \frac{\mathbf{v}_{j+1}}{||\mathbf{v}_{j+1}||} + \frac{\mathbf{v}_j}{||\mathbf{v}_j||} \right\|,\end{aligned}$$

where α_j is the angle at the origin, $\mathbf{0}$, of the triangle $\left[\mathbf{0}, \frac{\mathbf{v}_j}{||\mathbf{v}_j||}, \frac{\mathbf{v}_{j+1}}{||\mathbf{v}_{j+1}||}\right]$. Hence

$$\int_{\mathbf{v}_j}^{\mathbf{v}_{j+1}} \mathbf{r}\ \mathbf{F}\cdot d\mathbf{r} = \tan(\alpha_j/2)\ \left\{\frac{\mathbf{v}_{j+1}}{||\mathbf{v}_{j+1}||} + \frac{\mathbf{v}_j}{||\mathbf{v}_j||}\right\}$$

with the convention that α_j takes the positive sign if $\frac{\mathbf{v}_{j+1}}{||\mathbf{v}_{j+1}||} - \frac{\mathbf{v}_j}{||\mathbf{v}_j||}$ is counterclockwise and the negative sign if it is clockwise. Hence for $\mathbf{v}\in\mathbb{R}^2$,

$$\int_{\mathbf{v}_j}^{\mathbf{v}_{j+1}} (\mathbf{r}-\mathbf{v})\ \mathbf{F}(\mathbf{r}-\mathbf{v})\cdot d\mathbf{r} = \tan(\alpha_j/2)\ \left\{\frac{\mathbf{v}_j-\mathbf{v}}{||\mathbf{v}_j-\mathbf{v}||} + \frac{\mathbf{v}_{j+1}-\mathbf{v}}{||\mathbf{v}_{j+1}-\mathbf{v}||}\right\},$$

where α_j is the angle at $\mathbf{v}$ of triangle $[\mathbf{v}, \mathbf{v}_j, \mathbf{v}_{j+1}]$. Substituting this into

$$\sum_{j=1}^{k}\int_{\mathbf{v}_j}^{\mathbf{v}_{j+1}} (\mathbf{r}-\mathbf{v})\ \mathbf{F}(\mathbf{r}-\mathbf{v})\cdot d\mathbf{r} = \int_{\partial D} (\mathbf{r}-\mathbf{v})\ \mathbf{F}(\mathbf{r}-\mathbf{v})\cdot d\mathbf{r} = \mathbf{0}$$

gives (2.21) with w_j given by (2.22). □

3. Representation on boundaries of compact 3-D manifolds in $\mathbb{R}^3$. As in the previous section, we want to find a differentiable vector field $\mathbf{F}(\mathbf{r}) = [F_1(x,y,z), F_2(x,y,z), F_3(x,y,z)]^T$, $\mathbf{r} = [x,y,z]^T \in \mathbb{R}^3\setminus\{\mathbf{0}\}$, that satisfies

$$\iint_{\partial M} \mathbf{r}\ \mathbf{F}(\mathbf{r})\cdot d\mathbf{S} = \mathbf{0} \tag{3.1}$$

and

$$\iint_{\partial M} \mathbf{F}(\mathbf{r})\cdot d\mathbf{S} \neq 0 \tag{3.2}$$

for any compact 3-D manifold $M\in\mathbb{R}^3$ with piecewise smooth boundary ∂M. We assume that the positive orientation of ∂M is one in which its normal points away from M. Then if (3.1) and (3.2) hold, then, for any $\mathbf{v}\in\mathbb{R}^3$,

$$\mathbf{v} = \frac{\iint_{\partial M} \mathbf{r}\ \mathbf{F}(\mathbf{r}-\mathbf{v})\cdot d\mathbf{S}}{\iint_{\partial M} \mathbf{F}(\mathbf{r}-\mathbf{v})\cdot d\mathbf{S}}. \tag{3.3}$$

A similar search as in the previous section requiring that $x\mathbf{F}$, $y\mathbf{F}$, and $z\mathbf{F}$ be divergence free, i.e., $\operatorname{div}(x\mathbf{F}) = \operatorname{div}(y\mathbf{F}) = \operatorname{div}(z\mathbf{F}) = 0$, shows that F_i, $i=1,2,3$, are solutions of the partial differential equation

$$x\frac{\partial U}{\partial x} + y\frac{\partial U}{\partial y} + z\frac{\partial U}{\partial z} = -3U. \tag{3.4}$$

The general solution of (3.4) that vanishes at infinity is

$$U(x,y,z) = \frac{1}{xyz}\Phi\left(\frac{x}{\rho}, \frac{y}{\rho}, \frac{z}{\rho}\right), \tag{3.5}$$

where $\Phi : \mathbb{S}^2 \to \mathbb{R}$ is an arbitrary differentiable function on the unit sphere $\mathbb{S}^2\subset\mathbb{R}^3$, $\rho := (x^2+y^2+z^2)^{1/2}$, and $\Phi\left(\frac{x}{\rho}, \frac{y}{\rho}, \frac{z}{\rho}\right)$ is the general solution of the corresponding homogeneous equation. To state the results, we define spherical harmonics

$$C_\ell^m(\phi,\theta) := N_\ell^m P_\ell^m(\cos\phi)\cos m\theta, \quad m = 0,1,\ldots,\ell,$$

$$S_\ell^m(\phi,\theta) := N_\ell^m P_\ell^m(\cos\phi)\sin m\theta, \quad m = 1,2,\ldots\ell,$$

where P_ℓ^m are the associated Legendre polynomials and N_ℓ^m are normalization constants (see [9]).

THEOREM 3.1. *Let*

(3.6)
$$\mathbf{F}(\mathbf{r}) \equiv \mathbf{F}(x,y,z) := \left[\frac{1}{xyz}\Phi\left(\frac{x}{\rho},\frac{y}{\rho},\frac{z}{\rho}\right), \frac{1}{x^2z}\Phi\left(\frac{x}{\rho},\frac{y}{\rho},\frac{z}{\rho}\right), \frac{1}{x^2y}\Phi\left(\frac{x}{\rho},\frac{y}{\rho},\frac{z}{\rho}\right)\right]^T,$$

where $\Phi : \mathbb{S}^2 \to \mathbb{R}$ is an arbitrary differentiable function. Then $x\mathbf{F}$, $y\mathbf{F}$, and $z\mathbf{F}$ are divergence free on $\mathbb{R}^3 \setminus \{\mathbf{0}\}$.

Further, if

$$\Phi\left(\frac{x}{\rho},\frac{x}{\rho},\frac{x}{\rho}\right) := \frac{x^2yz}{\rho^4}\left(1+\Psi\left(\frac{x}{\rho},\frac{x}{\rho},\frac{x}{\rho}\right)\right)$$

$$= \sin^3\phi\cos^2\theta\sin\theta\cos\phi\ (1+\Psi(\sin\phi\cos\theta,\sin\phi\sin\theta,\cos\phi)) \tag{3.7}$$

and

$$\Psi(\sin\phi\cos\theta,\sin\phi\sin\theta,\cos\phi) := \sum_{\ell=2}^{\infty}\left\{\sum_{m=0}^{\ell} a_\ell^m\ C_\ell^m(\phi,\theta) + \sum_{m=1}^{\ell} b_\ell^m\ S_\ell^m(\phi,\theta)\right\}, \tag{3.8}$$

then for any compact 3-D manifold M in $\mathbb{R}^3$ with piecewise smooth boundary and for $\mathbf{v} \in \mathbb{R}^3 \setminus \partial M$,

$$\iint_{\partial M} (\mathbf{r}-\mathbf{v})\ \mathbf{F}(\mathbf{r}-\mathbf{v}) \cdot d\mathbf{S} = \mathbf{0}. \tag{3.9}$$

In addition, if $|\Psi| \le 1$, then for any $\mathbf{v} \in \mathbb{R}^3$,

$$\mathbf{v} = \frac{\iint_{\partial M} \mathbf{r}\ \mathbf{F}(\mathbf{r}-\mathbf{v}) \cdot d\mathbf{S}}{\iint_{\partial M} \mathbf{F}(\mathbf{r}-\mathbf{v}) \cdot d\mathbf{S}}. \tag{3.10}$$

Proof. Suppose $\mathbf{F}$ is as in (3.6), where $\Phi : \mathbb{S}^2 \to \mathbb{R}$ is an arbitrary differentiable function on the unit sphere $\mathbb{S}^2$. Then

$$\operatorname{div}(x\mathbf{F}) = \frac{1}{xyz}\left(x\frac{\partial\Phi}{\partial x} + y\frac{\partial\Phi}{\partial y} + z\frac{\partial\Phi}{\partial z}\right) = 0, \quad (x,y,z) \in \mathbb{R}^3 \setminus \{\mathbf{0}\},$$

since $\Phi(\frac{x}{\rho},\frac{y}{\rho},\frac{z}{\rho})$ satisfies

$$x\frac{\partial\Phi}{\partial x} + y\frac{\partial\Phi}{\partial y} + z\frac{\partial\Phi}{\partial z} = 0.$$

Similarly, $\operatorname{div}(y\mathbf{F}) = \operatorname{div}(z\mathbf{F}) = 0$.

Suppose Φ is given by (3.7) and (3.8). To prove (3.9) we may assume $\mathbf{v} = \mathbf{0}$, so it reduces to proving (3.1). Since $x\mathbf{F}$, $y\mathbf{F}$, and $z\mathbf{F}$ are divergence free, (3.1) holds for any compact 3-D manifold M that does not contain the origin. To prove that (3.1) holds for any M that contains the origin, it suffices to prove that it holds for the unit

ball B with boundary $\partial B = \mathbb{S}^2$. In spherical polar coordinates,

$$\iint_{\mathbb{S}^2} x\mathbf{F}\cdot d\mathbf{S} = \int_0^{2\pi}\int_0^{\pi} \frac{\Phi(\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi)}{\sin\phi\cos\phi\sin\theta\cos\theta} d\phi d\theta, \tag{3.11}$$

$$\iint_{\mathbb{S}^2} y\mathbf{F}\cdot d\mathbf{S} = \int_0^{2\pi}\int_0^{\pi} \frac{\Phi(\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi)}{\sin\phi\cos\phi\cos^2\theta} d\phi d\theta, \tag{3.12}$$

$$\iint_{\mathbb{S}^2} z\mathbf{F}\cdot d\mathbf{S} = \int_0^{2\pi}\int_0^{\pi} \frac{\Phi(\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi)}{\sin^2\phi\cos^2\theta\sin\theta} d\phi d\theta. \tag{3.13}$$

By (3.7) and (3.11)–(3.13),

$$\begin{aligned}\iint_{\mathbb{S}^2} \mathbf{r}\ \mathbf{F}\cdot d\mathbf{S} &\equiv \iint_{\mathbb{S}^2} [x,y,z]^T\mathbf{F}\cdot d\mathbf{S}\\ &= \int_0^{2\pi}\int_0^{\pi} (1+\Psi)\ [C_1^1(\phi,\theta), S_1^1(\phi,\theta), C_1^0(\phi,\theta)]^T\ \sin\phi d\phi d\theta,\end{aligned}$$

where

$$C_1^0(\phi,\theta) = \cos\phi, \quad C_1^1(\phi,\theta) = \sin\phi\cos\theta, \quad S_1^1(\phi,\theta) = \sin\phi\sin\theta$$

are the first order spherical harmonics. Since $1+\Psi$ is orthogonal to C_1^0, C_1^1, and S_1^1, it follows that $\iint_{\mathbb{S}^2} \mathbf{r}\ \mathbf{F}\cdot d\mathbf{S} = 0$.

To prove (3.10) we first show that (3.2) holds. The same argument as in the proof of Lemma 2.3 using the divergence theorem gives

$$\iint_{\partial M} \mathbf{F}\cdot d\mathbf{S} = \begin{cases} \iiint_M \frac{1}{\rho^4}\left(1+\Psi\left(\frac{x}{\rho},\frac{y}{\rho},\frac{z}{\rho}\right)\right)\ dxdydz & \text{if } \mathbf{0}\notin M,\\ -\iiint_{\mathbb{R}^3\setminus M} \frac{1}{\rho^4}\left(1+\Psi\left(\frac{x}{\rho},\frac{y}{\rho},\frac{z}{\rho}\right)\right)\ dxdydz & \text{if } \mathbf{0}\in M.\end{cases} \tag{3.14}$$

In particular, $\iint_{\partial M}\mathbf{F}\cdot d\mathbf{S} \neq 0$ if $|\Psi|\le 1$. Hence (3.10) follows from (3.9) for $\mathbf{v}\notin\partial M$. A limiting argument as in the proof of Theorem 2.2 shows that it also holds for $\mathbf{v}\in\partial M$. ∎

4. Mean value coordinates in $\mathbb{R}^3$. We now apply the results in Theorem 3.1 to compute mean value coordinates in $\mathbb{R}^3$. An explicit expression for mean value coordinates for points that lie in the kernel of a star-shaped polyhedron in $\mathbb{R}^3$ is given in [3]. An algorithm for computing these coordinates for all points in $\mathbb{R}^3$ can be found in [6], but no explicit expression is given there. In this section we derive a new formula for mean value coordinates for points in $\mathbb{R}^3$ with respect to the boundary of any compact 3-D manifold with polyhedral faces of any topology. In the case where all the faces of ∂M are triangular, the formula is equivalent to that in [3].

Consider $\mathbf{F} = [F_1, F_2, F_3]^T$ in (3.6), where Ψ is given by (3.7) as in Theorem 3.1. Throughout this section we shall assume that $\Psi \equiv 0$ so that Φ takes the simplest form:

$$\Phi\left(\frac{x}{\rho},\frac{y}{\rho},\frac{z}{\rho}\right) = \frac{x^2yz}{\rho^4}$$

and

$$\mathbf{F}\equiv\mathbf{F}(\mathbf{r}) = \frac{\mathbf{r}}{\rho^4}, \quad \rho\neq\mathbf{0}. \tag{4.1}$$

Now

$$curl(\mathbf{F}) = \bigtriangledown\left(\frac{1}{\rho^4}\right) \times \mathbf{r} + \frac{1}{\rho^4}\ curl(\mathbf{r})$$

$$= \frac{-4}{\rho^6}\mathbf{r} \times \mathbf{r} = \mathbf{0}$$

and

$$\text{div}(\mathbf{F}) = \bigtriangledown\left(\frac{1}{\rho^4}\right) \cdot \mathbf{r} + \frac{1}{\rho^4}\ \text{div}(\mathbf{r}) = -\frac{1}{\rho^4}, \quad \rho \neq 0.$$

Since

$$\text{div}(x\mathbf{F}) = \text{div}(y\mathbf{F}) = \text{div}(z\mathbf{F}) = 0,$$

we want to find vector fields $\mathbf{A}, \mathbf{B}, \mathbf{C}$ such that

$$x\mathbf{F} = curl(\mathbf{A}), \quad y\mathbf{F} = curl(\mathbf{B}), \quad z\mathbf{F} = curl(\mathbf{C}), \quad \rho \neq 0, \tag{4.2}$$

so that Stoke's theorem can be applied to evaluate the integrals of $x\mathbf{F}$, $y\mathbf{F}$, $z\mathbf{F}$ over the polyhedral faces of ∂M.

By (4.2)

$$curl(x\mathbf{F}) = \bigtriangledown(\text{div}(\mathbf{A})) - \bigtriangledown^2\mathbf{A},$$

so that if $\text{div}(\mathbf{A}) = 0$, the vector components of $\mathbf{A}$ satisfy Poisson's equations

$$\bigtriangledown^2\mathbf{A} = -curl(x\mathbf{F}), \quad \rho \neq 0. \tag{4.3}$$

Similarly, if $\text{div}(\mathbf{B}) = \text{div}(\mathbf{C}) = 0$,

$$\bigtriangledown^2\mathbf{B} = -curl(y\mathbf{F}), \tag{4.4}$$

$$\bigtriangledown^2\mathbf{C} = -curl(z\mathbf{F}), \quad \rho \neq 0. \tag{4.5}$$

Now

$$curl(x\mathbf{F}) = \bigtriangledown x \times \mathbf{F} + x\ curl(\mathbf{F}) = [0, -F_3, F_2]^T = \frac{1}{\rho^4}[0, -z, y]^T,$$

$$curl(y\mathbf{F}) = \bigtriangledown y \times \mathbf{F} + y\ curl(\mathbf{F}) = [F_3, 0, -F_1]^T = \frac{1}{\rho^4}[z, 0, -x]^T,$$

$$curl(z\mathbf{F}) = \bigtriangledown z \times \mathbf{F} + z\ curl(\mathbf{F}) = [-F_2, F_1, 0]^T = \frac{1}{\rho^4}[-y, x, 0]^T,$$

so that by (4.3)–(4.5),

$$\bigtriangledown^2\mathbf{A} = \frac{1}{\rho^4}[0, z, -y]^T, \tag{4.6}$$

$$\bigtriangledown^2\mathbf{B} = \frac{1}{\rho^4}[-z, 0, x]^T, \tag{4.7}$$

$$\bigtriangledown^2\mathbf{C} = \frac{1}{\rho^4}[y, -x, 0]^T, \quad \rho \neq 0. \tag{4.8}$$

We look for divergence free solutions, $\mathbf{A}, \mathbf{B}, \mathbf{C}$, of Poisson's equations (4.6), (4.7), (4.8), which satisfy (4.2) and vanish at infinity.

PROPOSITION 4.1. *The vector fields*

$$\mathbf{A} = \frac{1}{2\rho^2}[0, z, -y]^T, \quad \mathbf{B} = \frac{1}{2\rho^2}[-z, 0, x]^T, \quad \mathbf{C} = \frac{1}{2\rho^2}[y, -x, 0]^T \tag{4.9}$$

are divergence free and satisfy (4.2). *Further, they are unique divergence free solutions of* (4.6), (4.7), (4.8) *that satisfy* (4.2) *and vanish at infinity.*

Proof. It is straightforward to verify that $\mathbf{A}, \mathbf{B}, \mathbf{C}$ satisfy (4.2) and $\operatorname{div}(\mathbf{A}) = \operatorname{div}(\mathbf{B}) = \operatorname{div}(\mathbf{C}) = 0$. Hence they are particular solutions of (4.6), (4.7), (4.8), respectively. To prove uniqueness, we consider (4.6), in which the general solution is

$$\mathbf{A}_g = \left[U, U + \frac{z}{2\rho^2}, U - \frac{y}{2\rho^2}\right]^T,$$

where U is the general solution of the corresponding homogeneous equation, which is the Laplace equation. The conditions that $\operatorname{div}(\mathbf{A}_g) = 0$, $curl(\mathbf{A}_g) = x\mathbf{F}$, and $\mathbf{A}_g$ vanishes at infinity imply $U = 0$. Hence $\mathbf{A}_g = \mathbf{A}$. The proof of uniqueness of $\mathbf{B}$ and $\mathbf{C}$ is the same. □

Consider a 3-D manifold, $M \subset \mathbb{R}^3$, with polyhedral faces (not necessarily planar) of arbitrary topology, and each face has piecewise linear boundaries. Let $\mathcal{F}$ be the set of all its faces, which are oriented with their normals pointing away from M. The orientation of each face induces an orientation on the edges that form its boundary. Let $\mathcal{V}$ be the set of all vertices of ∂M.

PROPOSITION 4.2. *Let S be a face of ∂M whose boundary is the union of oriented line segments $[\mathbf{v}_j, \mathbf{v}_{j+1}]$, $j = 1, 2, \ldots, k = k(S)$, $\mathbf{v}_{k+1} = \mathbf{v}_1$. Then for any $\mathbf{v} \notin \partial S$,*

$$\iint_S (\mathbf{r} - \mathbf{v})\mathbf{F}(\mathbf{r} - \mathbf{v}) \cdot d\mathbf{S} = \frac{1}{2} \sum_{j=1}^{k} \alpha_j \, \mathbf{n}_j, \tag{4.10}$$

where $\alpha_j \in (0, \pi)$ is the angle at $\mathbf{v}$ of the triangle $[\mathbf{v}, \mathbf{v}_j, \mathbf{v}_{j+1}]$ and

$$\mathbf{n}_j = \frac{(\mathbf{v}_{j+1} - \mathbf{v}) \times (\mathbf{v}_j - \mathbf{v})}{||(\mathbf{v}_{j+1} - \mathbf{v}) \times (\mathbf{v}_j - \mathbf{v})||}. \tag{4.11}$$

Proof. We assume without loss of generality that $\mathbf{v} = \mathbf{0}$. By Stoke's theorem

$$\iint_S \mathbf{r}\ \mathbf{F} \cdot d\mathbf{S} = \left[\iint_S x\ \mathbf{F} \cdot d\mathbf{S}, \iint_S y\ \mathbf{F} \cdot d\mathbf{S}, \iint_S z\ \mathbf{F} \cdot d\mathbf{S}\right]^T$$

$$= \left[\int_{\partial S} \mathbf{A} \cdot d\mathbf{r}, \int_{\partial S} y\mathbf{B} \cdot d\mathbf{r}, \int_{\partial S} \mathbf{C} \cdot d\mathbf{r}\right]^T \tag{4.12}$$

$$= \sum_{j=1}^{k} \left[\int_{[\mathbf{v}_j, \mathbf{v}_{j+1}]} \mathbf{A} \cdot d\mathbf{r}, \int_{[\mathbf{v}_j, \mathbf{v}_{j+1}]} \mathbf{B} \cdot d\mathbf{r}, \int_{[\mathbf{v}_j, \mathbf{v}_{j+1}]} \mathbf{C} \cdot d\mathbf{r}\right]^T. \tag{4.13}$$

Let $\mathbf{v}_j = [x_j, y_j, z_j]^T$. A straightforward computation gives

$$\int_{[\mathbf{v}_j,\mathbf{v}_{j+1}]} \mathbf{A}\cdot d\mathbf{r} = -\begin{vmatrix} y_j & z_j \\ y_{j+1} & z_{j+1}\end{vmatrix}\left(\frac{\alpha_j}{2||\mathbf{v}_{j+1}\times\mathbf{v}_j||}\right),$$

$$\int_{[\mathbf{v}_j,\mathbf{v}_{j+1}]} \mathbf{B}\cdot d\mathbf{r} = \begin{vmatrix} x_j & z_j \\ x_{j+1} & z_{j+1}\end{vmatrix}\left(\frac{\alpha_j}{2||\mathbf{v}_{j+1}\times\mathbf{v}_j||}\right),$$

$$\int_{[\mathbf{v}_j,\mathbf{v}_{j+1}]} \mathbf{C}\cdot d\mathbf{r} = -\begin{vmatrix} x_j & y_j \\ x_{j+1} & y_{j+1}\end{vmatrix}\left(\frac{\alpha_j}{2||\mathbf{v}_{j+1}\times\mathbf{v}_j||}\right),$$

which, by (4.13), leads to (4.10) and (4.11). □

The next step in the computation of the mean value coordinates is to express the sum on the right of (4.10) in terms of $\mathbf{v}_j - \mathbf{v}$, $j = 1, 2, \ldots, k$:

$$\frac{1}{2}\sum_{j=1}^{k}\alpha_j\,\mathbf{n}_j = \sum_{j=1}^{k} a_j(S)(\mathbf{v}_j - \mathbf{v}), \tag{4.14}$$

where $a_j(S) \equiv a_{\mathbf{v}_j}(S)$ so that

$$\mathbf{0} = \sum_{S\in\mathcal{F}}\iint_S (\mathbf{r}-\mathbf{v})\mathbf{F}(\mathbf{r}-\mathbf{v})\cdot d\mathbf{S} = \sum_{S\in\mathcal{F}}\sum_{j=1}^{k(S)} a_j(S)(\mathbf{v}_j - \mathbf{v}),$$

which would lead to

$$\mathbf{v} = \frac{\sum_{S\in\mathcal{F}}\sum_{j=1}^{k(S)} a_j(S)\,\mathbf{v}_j}{\sum_{S\in\mathcal{F}}\sum_{j=1}^{k(S)} a_j(S)} = \frac{\sum_{\mathbf{p}\in\mathcal{V}}\lambda_{\mathbf{p}}\,\mathbf{p}}{\sum_{\mathbf{p}\in\mathcal{V}}\lambda_{\mathbf{p}}}, \tag{4.15}$$

where

$$\lambda_{\mathbf{p}} := \sum_{\substack{S\subset\mathcal{F}\\ \mathbf{p}\in S}} a_{\mathbf{p}}(S). \tag{4.16}$$

If $k = 3$, $a_j(S)$ can be obtained uniquely by taking the inner products of the expressions in (4.14) with $\mathbf{n}_\ell$ or $\mathbf{v}_\ell - \mathbf{v}$, $\ell = 1, 2, 3$. Taking the inner product with $\mathbf{n}_\ell$ gives

$$a_{\ell+2}(S) = \frac{1}{2\,\mathbf{n}_\ell\cdot(\mathbf{v}_{\ell+2}-\mathbf{v})}\sum_{j=1}^{3}\mathbf{n}_\ell\cdot\mathbf{n}_j\,\alpha_j, \quad \ell = 1, 2, 3,$$

where $a_j = a_{j+3}$ and $\mathbf{n}_j = \mathbf{n}_{j+3}$ for all j. These coordinates are the same as those obtained earlier in [3] for the representation of points in the kernel of 3-D star-shaped manifolds. We have shown that they are applicable to all points in $\mathbb{R}^3$ with respect to the boundaries of more general manifolds. If $k > 3$, the representation (4.14) exists, but is not unique. We shall derive a formula for $a_j(S)$ in (4.14) for the general case, which reduces to the above formula when $k = 3$. The idea is to express

$$\frac{1}{2}\sum_{j=1}^{k}\alpha_j\,\mathbf{n}_j = b_{\ell,0}(\mathbf{v}-\mathbf{v}_\ell) + b_{\ell,1}(\mathbf{v}-\mathbf{v}_{\ell+1}) + b_{\ell,2}(\mathbf{v}-\mathbf{v}_{\ell+2}) \tag{4.17}$$

for $\ell = 1, 2, \ldots, k$, where $b_{k+1,i} = b_{1,i}$, $i = 0, 1, 2$, and define

$$a_j(S) := \frac{1}{k}(b_{j,0} + b_{j-1,1} + b_{j-2,2}), \quad j = 1, 2, \ldots, k.$$

Then

$$\begin{aligned}\frac{1}{2}\sum_{j=1}^{k} \alpha_j \, \mathbf{n}_j &= \frac{1}{k}\sum_{\ell=1}^{k}\sum_{i=0}^{2} b_{\ell,i}(\mathbf{v} - \mathbf{v}_{\ell+i}) \\ &= \frac{1}{k}\sum_{j=1}^{k}\sum_{i=0}^{2} b_{j-i,i}(\mathbf{v} - \mathbf{v}_j) \\ &= \sum_{j=1}^{k} a_j(S)(\mathbf{v} - \mathbf{v}_j),\end{aligned}$$

as required by (4.14) so that (4.15) holds.

To compute the coefficients $b_{\ell,i}$, $i = 0, 1, 2$, taking the inner product of (4.17) with $\mathbf{n}_\ell$, $\mathbf{n}_{\ell+1}$, and $\mathbf{n}_{\ell+2}$ leads to

$$b_{\ell,2} = \frac{1}{2\ \mathbf{v}_{\ell+2} \cdot \mathbf{n}_\ell} \sum_{j=1}^{k} \mathbf{n}_\ell \cdot \mathbf{n}_j \ \alpha_j,$$

$$b_{\ell,0} = \frac{1}{2\ \mathbf{v}_\ell \cdot \mathbf{n}_{\ell+1}} \sum_{j=1}^{k} \mathbf{n}_{\ell+1} \cdot \mathbf{n}_j \ \alpha_j,$$

and

$$b_{\ell,1} = \frac{1}{2\ \mathbf{v}_{\ell+1} \cdot \mathbf{n}_{\ell+2}} \sum_{j=1}^{k} \mathbf{n}_{\ell+2} \cdot \mathbf{n}_j \ \alpha_j - \frac{\mathbf{v}_\ell \cdot \mathbf{n}_{\ell+2}}{2\ \mathbf{v}_\ell \cdot \mathbf{n}_{\ell+1}\ \mathbf{v}_{\ell+1} \cdot \mathbf{n}_{\ell+2}} \sum_{j=1}^{k} \mathbf{n}_{\ell+1} \cdot \mathbf{n}_j \ \alpha_j$$

for $\ell = 1, 2, \ldots, k$.

Remark. In general, the functions Φ and Ψ in (3.7) and (3.8), respectively, define the vector field

$$\mathbf{F}(\mathbf{r}) = \frac{g(\mathbf{r})}{\rho^4}\mathbf{r}, \quad \mathbf{r} = (x, y, z) \in \mathbb{R}^3 \setminus \{\mathbf{0}\},$$

where

$$g(\mathbf{r}) := 1 + \Psi\left(\frac{\mathbf{r}}{\rho}\right),$$

and $\mathbf{F}$ provides mean value representations. Since

$$\mathbf{F}(\mathbf{r}) = -g(\mathbf{r}) \bigtriangledown \left(\frac{1}{2\rho^2}\right) \tag{4.18}$$

and

$$\operatorname{div}(x\mathbf{F}(\mathbf{r})) = 0, \quad \operatorname{div}(y\mathbf{F}(\mathbf{r})) = 0, \quad \operatorname{div}(z\mathbf{F}(\mathbf{r})) = 0, \tag{4.19}$$

the system (4.18), (4.19) can be discretized by the mimetic finite difference schemes (see [1], [8], and the references therein) to give a large class of mean value type coordinates.

Acknowledgments. Thanks to an anonymous referee for pointing out the connection with mimetic finite difference schemes. Thanks also to Rick Beatson for useful discussions that generated my interest in the subject and to the Abdus Salam International Centre for Theoretical Physics, Trieste, for providing an excellent environment and financial support during my visit in March/April 2007.

REFERENCES

[1] F. Brezzi, K. Lipnikov, M. Shashkov, and V. Simoncini, *A new discretization methodology for diffusion problems on generalized polyhedral meshes*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 3682–3692.

[2] M. S. Floater, *Mean value coordinates*, Comput. Aided Geom. Design, 20 (2003), pp. 19–27.

[3] M. S. Floater, G. Kos, and M. Reimers, *Mean value coordinates in* $3D$, Comput. Aided Geom. Design, 22 (2005), pp. 623–631.

[4] K. Hormann and M. S. Floater, *Mean value coordinates for arbitrary planar polygons*, ACM Trans. Graph., 25 (2006), pp. 1424–1441.

[5] T. Ju, S. Schaefer, and J. Warren, *Mean value coordinates for closed triangular meshes*, ACM Trans. Graph., 24 (2005), pp. 561–566.

[6] T. Langer, A. Belyaev, and H.-P. Seidel, *Mean value coordinates for arbitrary spherical polygons and polyhedra in* $\mathbb{R}^3$, in Curve and Surface Design: Avignon, 2006, P. Chenin, T. Lyche, and L. L. Schumaker, eds., Nashboro Press, Brentwood, TN, 2007, pp. 193–202.

[7] T. Langer, A. Belyaev, and H.-P. Seidel, *Spherical barycentric coordinates*, in Fourth Eurographics Symposium on Geometry Processing, A. Sheffer and K. Polthier, eds., Eurographics Association, 2006, pp. 81–88.

[8] K. Lipnikov, M. Shashkov, and D. Svyatskiy, *The mimetic finite difference discretization of diffusion problem on unstructured polyhedral meshes*, J. Comput. Phys., 211 (2006), pp. 473–491.

[9] E. T. Whittaker and G. N. Watson, *Solution of Laplace's equation involving Legendre functions*, in A Course in Modern Analysis, 4th ed., Cambridge University Press, 1990, pp. 391–395.

SIAM J. MATH. ANAL.
Vol. 40, No. 6, pp. 2451–2477

COLLISIONS IN THREE-DIMENSIONAL FLUID STRUCTURE INTERACTION PROBLEMS*

MATTHIEU HILLAIRET† AND TAKÉO TAKAHASHI‡

Abstract. This paper deals with a system composed of a rigid ball moving into a viscous incompressible fluid over a fixed horizontal plane. The equations of motion for the fluid are the Navier–Stokes equations, and the equations for the motion of the rigid ball are obtained by applying Newton's laws. We show that for any weak solution of the corresponding system satisfying the energy inequality, the rigid ball never touches the plane. This result is the extension of that obtained in [M. Hillairet, *Comm. Partial Differential Equations*, 32 (2007), pp. 1345–1371] in the two-dimensional setting.

Key words. fluid structure interactions, Cauchy theory, qualitative properties, collisions

AMS subject classifications. 35R35, 76D03, 76D05

DOI. 10.1137/080716074

1. Introduction.

In the last decade, several studies showed that collisions between rigid bodies in a fluid would lead to great difficulties in the mathematical treatment of fluid-structure interaction models. For example, in [4, p. 287], Feireisl constructs a solution in which a sphere remains stuck to the ceiling of the cavity regardless of the intensity of the gravity. This example emphasizes that collisions would lead to unphysical solutions to standard mathematical systems. Indeed, Starovoitov proves also that several solutions exist when contact occurs [11, 12]. Therefore, at least one of these solutions does not represent a physical configuration.

Before handling the description of these collisions, several studies proved they do not occur in fluid structure systems. In [15], Vázquez and Zuazua prove no collision can occur between particles for a one-dimensional toy model. Then Starovoitov obtains a criterion for the velocity field of solutions [12] in the multidimensional setting. Namely, he proves no collision can occur if the gradient of the velocity field is sufficiently integrable. Finally, two parallel studies [8, 9] prove a no-collision result when there is only one body in a bounded (or partially bounded) two-dimensional cavity. In the first case, the author considers a rigid disk inside a bigger disk. In the second case, the author considers a rigid disk above a ramp. The aim of the present study is to extend these two-dimensional results to three-dimensional comparable configurations, i.e., for a rigid sphere above a ramp in $\mathbb{R}^3$.

1.1. Mathematical model.

We consider a homogeneous rigid sphere $\mathcal{B}$ with radius 1 and density $\rho_{\mathcal{B}}$. We denote by $\mathbf{G}$ its center (of mass), by $\mathbf{V}$ (resp., $\boldsymbol{\omega}$) its translation (resp., angular) velocity, and by m (resp., $\mathbb{J}$) its mass (resp., inertia). Notice that $\boldsymbol{\omega}$ is a vector in $\mathbb{R}^3$ and $\mathbb{J} = J\mathrm{I}_3$, where $J \in (0, \infty)$, and I_3 is the 3×3

*Received by the editors February 19, 2008; accepted for publication (in revised form) November 8, 2008; published electronically February 25, 2009.

http://www.siam.org/journals/sima/40-6/71607.html

†Université de Toulouse, UPS, IMT, Laboratoire MIP, Université Paul Sabatier Toulouse 3, 31062 Toulouse Cedex 9, France (matthieu.hillairet@math.univ-toulouse.fr). This author was supported by ANR grant Jeunes Chercheurs "RUGO-Analyse et Calcul des Effets de Rugosité sur les Écoulements."

‡Institut Elie Cartan UMR 7502, INRIA–Nancy-Université–CNRS, POB 239, Vandœuvre-lès-Nancy 54506, France, and Team-Project CORIDA, INRIA Nancy Grand-Est, 615 rue du Jardin Botanique, POB 54600, Villers-lès-Nancy, France (takeo8@gmail.com). This author was supported in part by ANR grants JCJC06_137283 and BLAN07-2_202879.

identity matrix. The velocity field of $\mathcal{B}$ reads $\mathbf{V}+\boldsymbol{\omega}\times(\mathbf{x}-\mathbf{G})$ for all $\mathbf{x}\in\mathcal{B}$. The sphere evolves over a ramp $\mathcal{P}$. The remainder of the cavity $\mathbb{R}^3_+$ is denoted by $\mathcal{F}$. It contains an incompressible viscous and Newtonian fluid which does not slip on boundaries and has constant density $\rho_{\mathcal{F}}=1$ and viscosity μ. The whole system evolves only through the interactions between solid and fluid without any external force field.

The evolution of the fluid is described by $(\mathbf{u},p)$, a velocity/pressure field satisfying the incompressible Navier–Stokes equations:

$$\begin{cases}\partial_t\mathbf{u}+\mathbf{u}\cdot\nabla\mathbf{u}=\operatorname{div}\mathbb{T}(\mathbf{u},p)\\ \operatorname{div}\mathbf{u}=0\end{cases}\qquad\text{in }\mathcal{F},\tag{1.1}$$

where $\mathbb{T}(\mathbf{u},p)$ is the Newtonian stress tensor:

$$\mathbb{T}(\mathbf{u},p)=2\mu D(\mathbf{u})-p\mathrm{I}_3.$$

Here $D(\mathbf{u})$ stands for the symmetric part of the gradient of $\mathbf{u}$.

To describe the evolution of $\mathcal{B}$, we apply Newton's laws assuming continuity of the stress-tensor of the fluid on $\partial\mathcal{B}$. It yields

$$\begin{cases}-\displaystyle\int_{\partial\mathcal{B}}\mathbb{T}(\mathbf{u},p)\mathbf{n}\,\mathrm{d}\sigma=m\dot{\mathbf{V}},\\ -\displaystyle\int_{\partial\mathcal{B}}\mathbb{T}(\mathbf{u},p)\mathbf{n}\times(\mathbf{x}-\mathbf{G})\,\mathrm{d}\sigma=J\dot{\boldsymbol{\omega}}.\end{cases}\tag{1.2}$$

Here $\mathbf{n}$ stands for the normal to $\partial\mathcal{B}$ directed towards $\mathcal{B}$. As $\mathbb{J}$ is a scalar matrix, the inertial term $\mathbb{J}\boldsymbol{\omega}\times\boldsymbol{\omega}$ vanishes in the conservation of momentum.

This system is complemented with the boundary conditions

$$\mathbf{u}_{|\partial\mathcal{B}}=\mathbf{V}+\boldsymbol{\omega}\times(\mathbf{x}-\mathbf{G}),\quad \mathbf{u}_{|\mathcal{P}}=0,\quad \mathbf{u}_{|\infty}=0\tag{1.3}$$

and initial conditions

$$\mathbf{u}(0,\cdot)=\mathbf{u}^0,\quad \mathbf{V}(0)=\mathbf{V}^0,\quad \boldsymbol{\omega}(0)=\boldsymbol{\omega}^0,\quad \mathbf{G}(0)=\mathbf{G}^0.\tag{1.4}$$

For short, we shall refer to the whole system (1.1)–(1.3) as (FSIS) for fluid-solid interaction system. We emphasize that this system is strongly coupled. On the one hand, the position of the sphere $\mathcal{B}$ fixes the domain $\mathcal{F}$ where the incompressible Navier–Stokes equations (1.1) have to be solved and the movement of $\mathcal{B}$ fixes the boundary conditions for (1.1) on $\partial\mathcal{B}$. In particular, we lay stress upon these time-dependences, denoting by $\mathcal{F}(t)$ (resp., $\mathcal{B}(t)$) the domain occupied by the fluid (resp., the solid body) at time t in the following. We shall reserve the notation $\mathcal{B}$ (resp., $\mathcal{F}$) for the sphere (resp., the fluid) as "actors" in the scenarios provided by our solutions to (FSIS). On the other hand, the solution $(\mathbf{u},p)$ prescribes the displacement of $\mathcal{B}$ via the computation of the forces and torques applied to $\mathcal{B}$:

$$-\int_{\partial\mathcal{B}}\mathbb{T}(\mathbf{u},p)\mathbf{n}\,\mathrm{d}\sigma,\quad -\int_{\partial\mathcal{B}}\mathbb{T}(\mathbf{u},p)\mathbf{n}\times(\mathbf{x}-\mathbf{G})\,\mathrm{d}\sigma.$$

Our main result is that no collision can occur between $\mathcal{B}$ and $\mathcal{P}$ in finite time in solutions to (FSIS). This reads as follows.

THEOREM 1.1. *Given $T>0$, let $(\mathbf{u},\mathbf{G})$ be a weak solution to (FSIS) over $(0,T)$ with initial data $(\mathbf{u}^0,\mathbf{G}^0)$. Then there exists a decreasing function $h_{min}\in$*

$\mathcal{C}([0,T];(0,\infty))$ *depending only on initial data* $(\mathbf{u}^0,\mathbf{G}^0)$ *such that* $h(t):=\operatorname{dist}(\mathcal{B}(t),\mathcal{P})$ *satisfies*

$$h(t)\geqslant h_{min}(t)\quad\forall\, t\in(0,T).$$

This result has been expected ever since the computations of Cooley and O'Neill [1] in the slow motion regime. However, until now no rigorous mathematical result has been available in the full nonlinear case. We emphasize that this result still holds true when one adds a reasonable external force field $\mathbf{f}$, for example, the gravity or $\mathbf{f}\in L^2((0,T)\times\mathbb{R}^3_+)$. In section 3, we provide an interpretation for the weak formulation of (FSIS) explaining how the distance can be estimated from below with a suitable test function. Then we construct a test function explicitly. In section 4, the interpretation of the weak formulation is applied to the constructed test function. Technical details are postponed to the appendix.

As mentioned in Theorem 1.1, our result applies to any weak solution to (FSIS). However, the Cauchy theory of weak solutions has been developed only in bounded and in exterior domains (to our knowledge) so that we do not know whether one solution exists to (FSIS). For the sake of completeness, we extend classical results for the Cauchy theory of (FSIS) to a half-space in the next section. Eventually, we obtain that, for a sufficiently large class of initial data $(\mathbf{u}^0,\mathbf{G}^0)$, the system (FSIS) predicts that no collision can occur between $\mathcal{B}$ and $\mathcal{P}$.

1.2. Notation. Throughout, bold symbols stand for vectors. Given $\mathbf{a}\in\mathbb{R}^3$ we denote by $\mathbf{a}\otimes\mathbf{a}$ the symmetric matrix with entries a_ia_j. Coordinates $\mathbf{x}=(x_1,x_2,x_3)$ are centered on $\mathcal{P}$. For example, we have $\mathcal{P}:=\{(x_1,x_2,0),\ (x_1,x_2)\in\mathbb{R}^2\}$. The half-space above $\mathcal{P}$ is $\mathbb{R}^3_+$ and $\mathbb{R}^3_{++}$ stands for $\{(x_1,x_2,x_3)$ with $x_3>1\}$. This is the domain where the center of mass $\mathbf{G}$ can evolve as long as no collision between $\mathcal{B}$ and $\mathcal{P}$ occurs.

Given $\mathbf{G}\in\mathbb{R}^3$ and $\delta>0$, we denote by $\mathcal{B}(\mathbf{G},\delta)$ the sphere with center $\mathbf{G}$ and radius δ. For short, we also set $\mathcal{B}_{\mathbf{G}}=\mathcal{B}(\mathbf{G},1)$. This is the domain occupied by $\mathcal{B}$ when its center of mass meets $\mathbf{G}$. In this case, the fluid domain $\mathcal{F}_{\mathbf{G}}$ is the complementary of $\mathcal{B}_{\mathbf{G}}$ in $\mathbb{R}^3_+$. If the orthogonal projection of $\mathbf{G}$ on $\mathcal{P}$ is the center of coordinates, we have $\mathbf{G}=\mathbf{G}_h=(0,0,1+h)$ with $h=\operatorname{dist}(\mathcal{B}_{\mathbf{G}},\mathcal{P})$. In this case the suitable parameter is h and not $\mathbf{G}$. Thus, when using notation with h as subscript instead of $\mathbf{G}$, we implicitly mean that the subscript should be $\mathbf{G}_h$. For example, $\mathcal{B}_h:=\mathcal{B}_{\mathbf{G}_h}$.

In the whole paper, we denote by $\eta:[0,\infty)\to[0,1]$ a smooth function such that

$$\eta(s)=\begin{cases}1 & \text{if } s<\frac{1}{2},\\ 0 & \text{if } s>1,\end{cases}$$

and we set $\eta_\alpha=\eta(\cdot/\alpha)$ for all parameters $\alpha>0$.

We use the classical Lebesgue and Sobolev spaces $L^\alpha(A)$, $W^{\beta,\alpha}(A)$, $H^\beta(A)$ with A an open set, $\alpha\geqslant 1$, and $\beta\geqslant 0$. We define

$$\mathcal{H}=\{\boldsymbol{\phi}\in L^2(\mathbb{R}^3_+)\ ;\ \operatorname{div}\boldsymbol{\phi}=0,\ \boldsymbol{\phi}\cdot\mathbf{n}=0 \text{ on } \mathcal{P}\},$$

$$\mathcal{V}=\{\boldsymbol{\phi}\in H^1(\mathbb{R}^3_+)\ ;\ \operatorname{div}\boldsymbol{\phi}=0,\ \boldsymbol{\phi}=0 \text{ on } \mathcal{P}\}.$$

We recall that $\mathcal{H}$ and $\mathcal{V}$ are closed subspaces of $L^2(\mathbb{R}^3_+)$ and $H^1_0(\mathbb{R}^3_+)$, respectively. Thus, they form Hilbert spaces with respect to the induced inner products. For an open subset $A\subset\mathbb{R}^3_+$, we also consider the following Hilbert spaces:

$$\mathbb{H}(A)=\{\boldsymbol{\phi}\in\mathcal{H}\ ;\ D(\boldsymbol{\phi})=0 \text{ in } A\},$$

$$\mathbb{V}(A)=\{\boldsymbol{\phi}\in\mathcal{V}\ ;\ D(\boldsymbol{\phi})=0 \text{ in } A\}.$$

To simplify, if $\mathbf{G} \in \mathbb{R}^3_{++}$, we set

$$\mathbb{H}(\mathbf{G}) = \mathbb{H}(\mathcal{B}_{\mathbf{G}}), \quad \mathbb{V}(\mathbf{G}) = \mathbb{V}(\mathcal{B}_{\mathbf{G}}).$$

For all $\mathbf{G} \in \mathbb{R}^3_{++}$, we also denote by $\rho_{\mathbf{G}}$ the function

$$\rho_{\mathbf{G}}(\mathbf{x}) = \begin{cases} \rho_{\mathcal{B}} & \text{if } \mathbf{x} \in \mathcal{B}_{\mathbf{G}}, \\ 1 & \text{if } \mathbf{x} \in \mathcal{F}_{\mathbf{G}}. \end{cases}$$

If $\mathbf{v} \in \mathbb{H}(\mathbf{G})$, from [14, p. 18], there exists a unique pair $(\mathbf{V}[\mathbf{v}], \boldsymbol{\omega}[\mathbf{v}]) \in \mathbb{R}^3 \times \mathbb{R}^3$ such that

$$\mathbf{v}|_{\mathcal{B}_{\mathbf{G}}} = \mathbf{V}[\mathbf{v}] + \boldsymbol{\omega}[\mathbf{v}] \times (\mathbf{x} - \mathbf{G}).$$

In particular, if $(\mathbf{u}, \mathbf{v}) \in \mathbb{H}(\mathbf{G})^2$,

$$\int_{\mathbb{R}^3_+} \rho_{\mathbf{G}} \mathbf{u} \cdot \mathbf{v} \, \mathrm{d}\mathbf{x} = \int_{\mathbb{R}^3_+ \setminus \mathcal{B}_{\mathbf{G}}} \mathbf{u} \cdot \mathbf{v} \, \mathrm{d}\mathbf{x} + m\mathbf{V}[\mathbf{u}] \cdot \mathbf{V}[\mathbf{v}] + J\boldsymbol{\omega}[\mathbf{u}] \cdot \boldsymbol{\omega}[\mathbf{v}].$$

2. Cauchy theory. First, we give the definition of weak solution to (FSIS).

DEFINITION 2.1. *Given $\mathbf{G}^0 \in \mathbb{R}^3_{++}$ and $\mathbf{u}^0 \in \mathbb{H}(\mathbf{G}^0)$, a pair $(\mathbf{u}, \mathbf{G})$ is called a weak solution to (FSIS) on $(0,T)$ with initial data $(\mathbf{u}^0, \mathbf{G}^0)$ if*

$$\mathbf{G} \in W^{1,\infty}(0,T;\mathbb{R}^3_{++}), \quad \text{with} \quad \mathbf{G}(0) = \mathbf{G}^0, \tag{2.1}$$

$$\mathbf{u} \in L^\infty(0,T;\mathcal{H}) \cap L^2(0,T;\mathcal{V}), \tag{2.2}$$

$$\mathbf{u} = \mathbf{V} + \boldsymbol{\omega} \times (\mathbf{x} - \mathbf{G}) \quad \text{in } \mathcal{B}_{\mathbf{G}}, \quad \text{with} \quad \mathbf{V} = \dot{\mathbf{G}}; \tag{2.3}$$

if for all $\mathbf{v} \in \mathcal{C}([0,T];H^1_0(\mathbb{R}^3_+)) \cap H^1(0,T;L^2(\mathbb{R}^3_+))$ with compact support in $(0,T)\times\mathbb{R}^3_+$ and such that $\mathbf{v} \in \mathbb{V}(\mathbf{G}(t))$ for all $t \in [0,T]$,

$$-\int_0^T \int_{\mathbb{R}^3_+} \rho_{\mathbf{G}} \mathbf{u} \cdot \partial_t \mathbf{v} \, \mathrm{d}\mathbf{y} \, \mathrm{d}t + 2\mu \int_0^T \int_{\mathbb{R}^3_+} D(\mathbf{u}) : D(\mathbf{v}) \, \mathrm{d}\mathbf{y} \, \mathrm{d}t - \int_0^T \int_{\mathbb{R}^3_+} \mathbf{u} \otimes \mathbf{u} : D(\mathbf{v}) \, \mathrm{d}\mathbf{y} \, \mathrm{d}t = 0; \tag{2.4}$$

if for all $\mathbf{v} \in \mathcal{C}([0,T];L^2(\mathbb{R}^3_+))$ with compact support in $[0,T) \times \mathbb{R}^3_+$ and such that $\mathbf{v} \in \mathbb{H}(\mathbf{G}(t))$ for all $t \in [0,T]$ we have

$$W : t \mapsto \int_{\mathbb{R}^3_+} \rho_{\mathbf{G}} \mathbf{u} \cdot \mathbf{v} \, \mathrm{d}\mathbf{x} \in \mathcal{C}([0,T]) \quad \text{with} \quad W(0) = \int_{\mathbb{R}^3_+} \rho_{\mathbf{G}^0} \mathbf{u}^0 \cdot \mathbf{v} \, \mathrm{d}\mathbf{x}; \tag{2.5}$$

if the energy estimate holds true:

$$\frac{1}{2} \int_{\mathbb{R}^3_+} \rho_{\mathbf{G}} |\mathbf{u}|^2 \, \mathrm{d}\mathbf{x} + 2\mu \int_0^t \int_{\mathbb{R}^3_+} |D(\mathbf{u})|^2 \, \mathrm{d}\mathbf{x} \, \mathrm{d}s \leqslant \frac{1}{2} \int_{\mathbb{R}^3_+} \rho_{\mathbf{G}^0} |\mathbf{u}^0|^2 \, \mathrm{d}\mathbf{x} \quad \text{for a.a. } t \in (0,T).$$

Before going to our existence result for such weak solutions, let us recall some of their straightforward properties. First, combining (2.2) and (2.3) yields that $\mathbf{u}(t,\cdot) \in$

$\mathbb{V}(\mathbf{G}(t))$ for almost all $t \in (0,T)$. Moreover, it follows from standard arguments that the pair $(\mathbf{V},\boldsymbol{\omega})$ such that (2.3) holds satisfies

$$|\mathbf{V}|^2 + |\boldsymbol{\omega}|^2 \leqslant C \int_{\mathbb{R}^3_+} \rho_{\mathbf{G}} |\mathbf{u}|^2 \, \mathrm{d}\mathbf{x}, \tag{2.6}$$

where C depends only on ρ_B. In particular, the pair $(\mathbf{V},\boldsymbol{\omega})$ associated to a weak solution $(\mathbf{u},\mathbf{G})$ belongs to $L^\infty(0,T)$.

The result of well-posedness we obtain for (FSIS) can be stated as follows.

THEOREM 2.2. *Assuming $\mathbf{G}^0 \in \mathbb{R}^3_{++}$ and $\mathbf{u}^0 \in \mathbb{H}(\mathbf{G}^0)$, there exists at least one maximal weak solution $(T_0,(\mathbf{U},\mathbf{G}))$ to (FSIS) with initial data $(\mathbf{U}^0,\mathbf{G}^0)$. Moreover, we have the alternative:*

- $T_0 = \infty$,
- $T_0 < \infty$ *and* $G_3(t) \to 1$ *as* $t \to T_0$.

The proof of Theorem 2.2 given in what follows is inspired by methods developed in other papers, and since we use many similar arguments, we choose to present here only the main ideas and refer to the appropriate references to avoid repeating technical calculations which are not the main interest of this paper. From now on $(\mathbf{G}^0,\mathbf{u}^0) \in \mathbb{R}^3_{++} \times \mathbb{H}(\mathbf{G}^0)$ is a fixed initial condition. We denote $(\mathbf{V}^0,\boldsymbol{\omega}^0) = (\mathbf{V}[\mathbf{u}^0],\boldsymbol{\omega}[\mathbf{u}^0])$ and $d^0 := \operatorname{dist}(\mathcal{B}_{\mathbf{G}^0},\mathcal{P})$. Due to our assumption, there holds $d^0 > 0$. The remainder of the section is devoted to the proof of Theorem 2.2.

2.1. Strong solutions for an approximate system. As in [7], we prove the existence of weak solutions by first obtaining strong solutions for an approximate problem of (FSIS). More precisely, we consider an even nonnegative function $\kappa \in C_0^\infty(\mathbb{R})$ such that $\kappa(s) = 0$ if $|s| \geqslant 1$. We define for all $\varepsilon > 0$

$$K_\varepsilon(\mathbf{x}) = \frac{c}{\varepsilon^3}\kappa\left(\left|\frac{\mathbf{x}}{\varepsilon}\right|^2\right) \quad (\mathbf{x} \in \mathbb{R}^3). \tag{2.7}$$

The constant c is chosen so that

$$\int_{\mathbb{R}^3} K_\varepsilon(\mathbf{x}) \, \mathrm{d}\mathbf{x} = 1 \quad \forall\, \varepsilon > 0.$$

Then, for all $\mathbf{u} \in L^2((0,T)\times\mathbb{R}^3_+)$ and for all $\varepsilon > 0$, we set

$$\mathbf{u}_\varepsilon(t,\mathbf{x}) = \int_{\mathbb{R}^3_+} K_\varepsilon(\mathbf{x}-\mathbf{y})\mathbf{u}(t,\mathbf{y}) \, \mathrm{d}\mathbf{y}.$$

Let us consider the following problem, which approximates (FSIS):

$$\partial_t \mathbf{u} - \mu\Delta\mathbf{u} + (\mathbf{u}_\varepsilon \cdot \nabla)\mathbf{u} + \nabla p = 0 \quad \text{in } \mathcal{F}_{\mathbf{G}(t)},\ t \in (0,T), \tag{2.8}$$

$$\operatorname{div}\mathbf{u} = 0 \quad \text{in } \mathcal{F}_{\mathbf{G}(t)},\ t \in (0,T), \tag{2.9}$$

$$\mathbf{u} = 0 \quad \text{on } \mathcal{P},\ t \in (0,T), \tag{2.10}$$

$$\mathbf{u}_{|_\infty} = 0, \quad t \in (0,T), \tag{2.11}$$

$$\mathbf{u} = \dot{\mathbf{G}} + \boldsymbol{\omega} \times (\mathbf{x} - \mathbf{G}) \quad \text{on } \partial\mathcal{B}_{\mathbf{G}},\ t \in (0,T), \tag{2.12}$$

$$m\ddot{\mathbf{G}} = -\int_{\partial\mathcal{B}_{\mathbf{G}}} \mathbb{T}(\mathbf{u},p)\mathbf{n}\,\mathrm{d}\sigma + \frac{1}{2}\int_{\partial\mathcal{B}_{\mathbf{G}}} ((\mathbf{u}_\varepsilon - \mathbf{u})\cdot\mathbf{n})\mathbf{u}\,\mathrm{d}\sigma \quad \text{in } (0,T), \tag{2.13}$$

$$\begin{aligned} J\dot{\boldsymbol{\omega}} = &-\int_{\partial\mathcal{B}_{\mathbf{G}}} (\mathbf{x}-\mathbf{G})\times\mathbb{T}(\mathbf{u},p)\mathbf{n}\,\mathrm{d}\sigma \\ &+\frac{1}{2}\int_{\partial\mathcal{B}_{\mathbf{G}}} ((\mathbf{u}_\varepsilon-\mathbf{u})\cdot\mathbf{n})(\mathbf{x}-\mathbf{G})\times\mathbf{u}\,\mathrm{d}\sigma \quad \text{in } (0,T). \end{aligned} \tag{2.14}$$

We complete the system with the initial conditions

$$\mathbf{u}(0,\cdot) = \mathbf{u}^0, \quad \dot{\mathbf{G}}(0) = \mathbf{V}^0, \quad \boldsymbol{\omega}(0) = \boldsymbol{\omega}^0, \quad \mathbf{G}(0) = \mathbf{G}^0. \tag{2.15}$$

We define the space $\widehat{H^1}(A)$ by

$$\widehat{H^1}(A) = \{q \in L^2_{loc}(\overline{A})\ ;\ \nabla q \in L^2(A)\}.$$

We denote

$$\mathcal{F}_T = \{(t,\mathbf{x}) \in [0,T]\times\mathbb{R}^3\ ;\ \mathbf{x} \in \mathcal{F}_{\mathbf{G}(t)}\}.$$

Consider a smooth mapping $\mathbf{X} : \mathbb{R}^3_{++}\times\mathcal{F}_{\mathbf{G}^0} \to \mathbb{R}^3$ such that for all $\mathbf{G}\in\mathbb{R}^3_{++}$, $\mathbf{X}(\mathbf{G},\cdot)$ is a C^∞-diffeomorphism from $\mathcal{F}_{\mathbf{G}^0}$ onto $\mathcal{F}_{\mathbf{G}}$. Moreover, suppose that the mappings

$$(\mathbf{G},\mathbf{y}) \mapsto D_{\mathbf{G}} D_y^\alpha \mathbf{X}(\mathbf{G},\mathbf{y}), \quad \alpha\in\mathbb{N}^3,$$

exist and are continuous and compactly supported in $\mathcal{F}_{\mathbf{G}^0}$. For any $\mathbf{g} : \mathcal{F}_T \to \mathbb{R}^3$, we denote by $\mathbf{g}_\mathbf{X} : [0,T]\times\mathcal{F}_{\mathbf{G}^0} \to \mathbb{R}^3$ the mapping $\mathbf{g}_\mathbf{X}(t,\mathbf{y}) = \mathbf{g}(t,\mathbf{X}(\mathbf{G}(t),\mathbf{y}))$ for all $t \geqslant 0$ for all $\mathbf{y}\in\mathcal{F}_{\mathbf{G}^0}$. We use similar notation for $g : \mathcal{F}_T \to \mathbb{R}$.

We introduce the following function spaces in a variable domain:

$$\begin{aligned} L^2(0,T;H^2(\mathcal{F}(t))) &= \{\mathbf{u}\ ;\ \mathbf{u}_\mathbf{X} \in L^2(0,T;H^2(\mathcal{F}_{\mathbf{G}^0}))\},\\ H^1(0,T;L^2(\mathcal{F}(t))) &= \{\mathbf{u}\ ;\ \mathbf{u}_\mathbf{X} \in H^1(0,T;L^2(\mathcal{F}_{\mathbf{G}^0}))\},\\ C([0,T],H^1(\mathcal{F}(t))) &= \{\mathbf{u}\ ;\ \mathbf{u}_\mathbf{X} \in C([0,T],H^1(\mathcal{F}_{\mathbf{G}^0}))\},\\ L^2(0,T;\widehat{H}^1(\mathcal{F}(t))) &= \{p\ ;\ p_\mathbf{X} \in L^2(0,T;\widehat{H}^1(\mathcal{F}_{\mathbf{G}^0}))\}. \end{aligned}$$

THEOREM 2.3. *Assume the initial conditions satisfy*

$$\operatorname{dist}(\mathcal{B}_{\mathbf{G}^0},\mathcal{P}) = d^0 > 0, \quad \mathbf{u}^0 \in \mathbb{V}(\mathbf{G}^0).$$

Then, given $\varepsilon > 0$, there exist a time $T > 0$ depending only on $\|\mathbf{u}^0\|_{L^2(\mathbb{R}^3_+)}$ and a 4-uplet $(\mathbf{u},\mathbf{G},\omega,p)$ satisfying

$$\mathbf{G} \in H^2(0,T), \quad \operatorname{dist}(\mathcal{B}_{\mathbf{G}(t)},\mathcal{P}) \geqslant \frac{d^0}{2} > 0 \quad \forall\, t\in[0,T], \tag{2.16}$$

$$\mathbf{u} \in L^2(0,T;H^2(\mathcal{F}(t))) \cap C([0,T];H^1(\mathcal{F}(t))) \cap H^1(0,T;L^2(\mathcal{F}(t))), \tag{2.17}$$

$$p \in L^2(0,T;\widehat{H^1}(\mathcal{F}(t))), \quad \boldsymbol{\omega} \in H^1(0,T), \tag{2.18}$$

and satisfying (2.8)–(2.14) *almost everywhere on* $[0,T]$ *or in the trace sense.*

Proof. One can obtain local existence of strong solutions to (2.8)–(2.14) via arguments similar to those in the proof of Theorem 1.1 in [2] (see also [3, 13, 10] for

results obtained applying similar techniques). For the sake of brevity, we compute only energy estimates here in order to prove that these local strong solutions can be continued up to collision between $\mathcal{B}$ and $\mathcal{P}$. We refer the reader to the mentioned articles for technical details.

First, we multiply (2.8) by $\mathbf{u}$, (2.13) by $\dot{\mathbf{G}}$, and (2.14) by $\boldsymbol{\omega}$. We deduce the energy estimate

$$\frac{1}{2}\int_{\mathbb{R}^3_+}\rho_{\mathbf{G}(t)}|\mathbf{u}|^2(t)\,\mathrm{d}\mathbf{x}+2\mu\int_0^t\int_{\mathbb{R}^3_+}|D(\mathbf{u})|^2\,\mathrm{d}\mathbf{x}\,\mathrm{d}s=\frac{1}{2}\int_{\mathbb{R}^3_+}\rho_{\mathbf{G}^0}|\mathbf{u}^0|^2\,\mathrm{d}\mathbf{x}. \tag{2.19}$$

From the above estimate and (2.6) we obtain a time $T>0$ depending on $\int_{\mathbb{R}^3_+}\rho_{\mathbf{G}^0}|\mathbf{u}^0|^2\,\mathrm{d}\mathbf{x}$ such that (2.16) holds for all ε.

Then we introduce a smooth function Υ with compact support in $\mathcal{B}(\mathbf{G}^0,1+d_0/4)$ and such that $\Upsilon=1$ on $\mathcal{B}_{\mathbf{G}^0}$. We also set

$$\mathbf{w}_R=\frac{1}{2}\left(\dot{\mathbf{G}}\times(\mathbf{x}-\mathbf{G})+|\mathbf{x}-\mathbf{G}|^2\boldsymbol{\omega}\right),$$
$$\mathbf{u}_R(t,\mathbf{x})=\operatorname{curl}\left[\Upsilon\left(\mathbf{x}-\mathbf{G}(t)+\mathbf{G}^0\right)\mathbf{w}_R(t,\mathbf{x})\right].$$

The function $\mathbf{u}_R$ has a compact support in $\mathcal{B}(\mathbf{G}^0,1+d_0/4)$ and satisfies

$$\operatorname{div}\mathbf{u}_R=0,\quad \mathbf{u}_R=\dot{\mathbf{G}}+\boldsymbol{\omega}\times(\mathbf{x}-\mathbf{G})\text{ on }\mathcal{B}(t).$$

We multiply (2.8) by

$$\boldsymbol{\varphi}=\partial_t\mathbf{u}+(\mathbf{u}_R\cdot\nabla)\mathbf{u}-(\mathbf{u}\cdot\nabla)\mathbf{u}_R,$$

which yields

$$\int_{\mathcal{F}(t)}(\partial_t\mathbf{u}+\mathbf{u}_\varepsilon\cdot\nabla\mathbf{u})\cdot\boldsymbol{\varphi}\,\mathrm{d}\mathbf{x}=\int_{\mathcal{F}(t)}\operatorname{div}\mathbb{T}(\mathbf{u},p)\cdot\boldsymbol{\varphi}\,\mathrm{d}\mathbf{x}. \tag{2.20}$$

Using the regularization of the nonlinear term, we obtain the existence of a constant C_0 depending only on $\|\mathbf{u}_0\|_{L^2(\mathbb{R}^3_+)}$ such that the left-hand side of (2.20) can be written as

$$LHS=\int_{\mathcal{F}(t)}|\partial_t\mathbf{u}|^2\,\mathrm{d}\mathbf{x}+R_l\quad\text{with}\quad|R_l|\leqslant C_0\int_{\mathcal{F}(t)}|\nabla\mathbf{u}|^2\,\mathrm{d}\mathbf{x}+\frac{1}{2}\int_{\mathcal{F}(t)}|\partial_t\mathbf{u}|^2\,\mathrm{d}\mathbf{x}.$$

Concerning the right-hand side of (2.20), we apply the same arguments as in the proof of Lemma 4.3 in [3] and obtain (see [3, (4.24)])

$$RHS=-\mu\frac{\mathrm{d}}{\mathrm{dt}}\int_{\mathcal{F}(t)}|D(\mathbf{u})|^2\,\mathrm{d}\mathbf{x}+I_r+R_r,$$

where

$$I_r=\int_{\partial\mathcal{B}(t)}\mathbb{T}(\mathbf{u},p)\mathbf{n}\cdot\left[\ddot{\mathbf{G}}+\dot{\boldsymbol{\omega}}\times(\mathbf{x}-\mathbf{G})-\boldsymbol{\omega}\times\dot{\mathbf{G}}\right]\,\mathrm{d}\sigma,$$

and, with the same convention as previously for C_0, we obtain

$$|R_r|\leqslant C_0\left[1+\int_{\mathcal{F}(t)}|\nabla\mathbf{u}|^2\,\mathrm{d}\mathbf{x}\right].$$

Finally, using (2.13)–(2.14), we deduce

$$I_r = -\left[m|\ddot{\mathbf{G}}|^2 + J|\dot{\boldsymbol{\omega}}|^2 - m\ddot{\mathbf{G}}\cdot(\boldsymbol{\omega}\times\dot{\mathbf{G}})\right] + \tilde{R}_r,$$

with

$$\tilde{R}_r = \frac{1}{2}\int_{\partial\mathcal{B}(t)} (\mathbf{u}-\mathbf{u}_\varepsilon)\cdot\mathbf{n}\left(\dot{\mathbf{G}}+\boldsymbol{\omega}\times(\mathbf{x}-\mathbf{G})\right)\cdot\left(\ddot{\mathbf{G}}-\boldsymbol{\omega}\times\dot{\mathbf{G}}+\dot{\boldsymbol{\omega}}\times(\mathbf{x}-\mathbf{G})\right)\,\mathrm{d}\sigma.$$

In this last integral, extending the rigid velocity fields to $\mathcal{B}(\mathbf{G}, 1+d_0/4)$ in a similar fashion to that in $\mathbf{u}_R$, we obtain, after integration by parts,

$$|\tilde{R}_r| \leqslant C_0 + \frac{m}{2}|\ddot{\mathbf{G}}|^2 + \frac{J}{2}|\dot{\boldsymbol{\omega}}|^2.$$

Combining all these estimates, we finally deduce

$$\begin{aligned}\mu\frac{\mathrm{d}}{\mathrm{dt}}\left[\int_{\mathcal{F}(t)}|\nabla\mathbf{u}|^2\,\mathrm{d}\mathbf{x}\right] + \frac{1}{2}\left[m|\ddot{\mathbf{G}}|^2 + J|\dot{\boldsymbol{\omega}}|^2 + \int_{\mathcal{F}(t)}|\partial_t\mathbf{u}|^2\,\mathrm{d}\mathbf{x}\right]\\ \leqslant C_0\left[1+\int_{\mathbb{R}^3_+}|\nabla\mathbf{u}|^2\,\mathrm{d}\mathbf{x}\right].\end{aligned}$$

As a consequence, we have obtained that the mapping $t\mapsto\|\mathbf{u}\|_{H^1(\mathcal{F}(t))}$ is bounded on $[0,T]$. □

2.2. Convergences. As in the assumptions of our theorem, we assume that the initial condition $\mathbf{u}^0$ belongs to $\mathbb{H}(\mathbf{G}^0)$; we introduce a sequence $\mathbf{u}^{0k}\in\mathbb{V}(\mathbf{G}^0)$ such that

$$\mathbf{u}^{0k}\to\mathbf{u}^0 \quad \text{in } \mathbb{H}(\mathbf{G}^0).$$

We also take a sequence $\varepsilon^k\to 0$. Then, applying Theorem 2.3, we can consider a (uniform) time T such that, for all k, the corresponding solutions $(\mathbf{u}^k,\mathbf{G}^k,\omega^k,p^k)$ exist on $[0,T]$ and satisfy

$$\operatorname{dist}(\mathcal{B}_{\mathbf{G}^k(t)},\mathcal{P}) > \frac{d^0}{2} \quad \forall\, t\in[0,T],\ \forall\, k.$$

In the following, we extend $\mathbf{u}^k$ with the value of the rigid velocity field $\mathbf{V}^k+\boldsymbol{\omega}^k\times(\mathbf{x}-\mathbf{G}^k)$ on the solid domain. From (2.19), the following holds:

$$\mathbf{u}^k \text{ is bounded in } L^\infty(0,T;L^2(\mathbb{R}^3_+))\cap L^2(0,T;H^1_0(\mathbb{R}^3_+)),$$

so that, up to a subsequence (which we do not relabel), we can assume

$$\mathbf{u}^k \rightharpoonup \mathbf{u} \quad \text{in } L^2(0,T;H^1(\mathbb{R}^3_+))\text{-weak and } L^\infty(0,T;L^2(\mathbb{R}^3_+))\text{-weak}^*, \tag{2.21}$$

$$\mathbf{G}^k\to\mathbf{G} \quad \text{in } C([0,T];\mathbb{R}^3_{++}). \tag{2.22}$$

Taking any $\mathbf{U}\in\mathcal{D}((0,T)\times\mathbb{R}^3_+)$ such that $D(\mathbf{U})=0$ in a neighborhood of $\mathcal{B}_{\mathbf{G}(t)}$ for all $t\in(0,T)$, we can multiply (2.8) by $\mathbf{U}$ for k sufficiently large. Integrating by

parts and using Reynolds' transport theorem yields

$$
\begin{aligned}
(2.23)\quad & -\int_0^T \int_{\mathbb{R}^3_+} \rho_{\mathbf{G}^k} \mathbf{u}^k \cdot \partial_t \mathbf{U} \, \mathrm{d}\mathbf{x} \, \mathrm{d}s \\
& \quad + \int_0^T \int_{\mathbb{R}^3_+} (\mathbf{u}^k \otimes \mathbf{u}^k) : D(\mathbf{U}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}s + 2\mu \int_0^T \int_{\mathbb{R}^3_+} D(\mathbf{u}^k) : D(\mathbf{U}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}s \\
& = \int_0^T \int_{\mathcal{F}_{\mathbf{G}^k(t)}} \left[((\mathbf{u}^k_{\varepsilon^k} - \mathbf{u}^k) \cdot \nabla) \mathbf{U} \right] \cdot \mathbf{u}^k \, \mathrm{d}\mathbf{x} - \frac{1}{2} \int_0^T \int_{\partial \mathcal{B}_{\mathbf{G}^k(t)}} (\mathbf{U} \cdot \mathbf{u}^k)((\mathbf{u}^k_{\varepsilon^k} - \mathbf{u}^k) \cdot \mathbf{n}) \, \mathrm{d}\sigma.
\end{aligned}
$$

In order to pass to the limit in this weak formulation, we need to prove L^2-compactness of the $\mathbf{u}^k$. As usual, this is the main difficulty of the proof. The procedure to prove this compactness property follows closely the method developed in [7]. The main difference here is that our cavity is unbounded. Therefore, we do not look for a compactness property of the sequence $\mathbf{u}^k$ on the whole domain but only locally in space. Indeed, with the help of Friedrichs' lemma (see [6, Lemma II.4.2]) we have the following: for all relatively compact $\mathcal{O} \subset \mathbb{R}^3_+$, for any $\gamma > 0$ there exist $I = I(\gamma, \mathcal{O}) \in \mathbb{N}$ and functions $\boldsymbol{\psi}_j \in L^\infty(\mathcal{O})$, $j = 1, \ldots, I$, such that

$$
\begin{aligned}
(2.24)\quad & \left\| \mathbf{u}^k - \mathbf{u} \right\|^2_{L^2(0,T;L^2(\mathcal{O}))} \\
& \leqslant \sum_{j=1}^{I} \int_0^T \left(\int_{\mathcal{O}} \rho_{\mathbf{G}} \left(\mathbf{u}^k(t) - \mathbf{u}(t) \right) \cdot \boldsymbol{\psi}_j \, \mathrm{d}\mathbf{y} \right)^2 \mathrm{d}t + \gamma \left\| \nabla \mathbf{u}^k - \nabla \mathbf{u} \right\|^2_{L^2(0,T;L^2(\mathcal{O}))}.
\end{aligned}
$$

Due to the uniform bound on $\mathbf{u}^k$ in $L^2(0,T;H^1_0(\mathbb{R}^3_+))$, our remaining task is to prove that, for any $\boldsymbol{\psi} \in L^\infty(\mathcal{O})$, there exists a subsequence (which we do not relabel) for which there holds

$$
(2.25)\qquad \lim_{k \to \infty} \int_0^T \left(\int_{\mathcal{O}} \rho_{\mathbf{G}} \left(\mathbf{u}^k(t) - \mathbf{u}(t) \right) \cdot \boldsymbol{\psi} \, \mathrm{d}\mathbf{y} \right)^2 \mathrm{d}t = 0.
$$

As a first step, we obtain results similar to those of (2.25) for another family of test functions (not included in $L^\infty(\mathcal{O})$. To this end, we divide the segment $[0,T]$ into N segments $[t_{i-1}, t_i]$, with $\Delta t = t_i - t_{i-1} = T/N$, $i = 1, \ldots, N$. For all i and for $\delta < \frac{d^0}{2}$, we consider an orthonormal basis $(\mathbf{e}^{i,\delta}_j)$ of $\mathbb{V}(\mathcal{B}(\mathbf{G}(t_i), 1+\delta))$. Without further restrictions, we assume all the $\mathbf{e}^{i,\delta}_j$ with compact support. We also consider the set of piecewise linear functions in t:

$$
(2.26)\qquad \mathbf{U}^\delta(t, \mathbf{x}) = \mathbf{e}^{i-1,\delta}_j(\mathbf{x}) + \frac{t - t_i}{\Delta t} \left(\mathbf{e}^{i,\delta}_l(\mathbf{x}) - \mathbf{e}^{i-1,\delta}_j(\mathbf{x}) \right)
$$

for $t \in [t_{i-1}, t_i]$, $j, l \in \mathbb{N}$, $i \in \{1, \ldots, N\}$. There exists a countable set of functions satisfying (2.26).

It is worth noting that, since $\mathbf{G}$ is uniformly continuous, for N big enough, the functions $\mathbf{U}^\delta$ of the above form satisfy $D(\mathbf{U}^\delta(t)) = 0$ in $\mathcal{B}(\mathbf{G}(t), 1 + \delta/2)$ for all t. In particular, these functions are in $C([0,T]; \mathbb{V}(\mathbf{G}(t)))$. These functions are important to approximate continuous functions in time with value in $\mathbb{H}(\mathbf{G}(t))$ and thus functions in $L^2(0,T;\mathbb{H}(\mathbf{G}(t)))$ (see Lemmas 4.1 and 4.2 in [7]). From (2.22), there exists $k_0 = k_0(\delta)$ such that for $k \geqslant k_0$,

$$
D(\mathbf{U}^\delta) = 0 \quad \text{in } \mathcal{B}_{\mathbf{G}^k(t)} \quad \forall\, t \in [0,T]
$$

for all functions satisfying (2.26).

We multiply (2.8) by $\mathbf{U}^\delta$ satisfying (2.26). After similar computations to those leading to (2.23), we obtain

$$\frac{d}{dt}\int_{\mathbb{R}^3_+}\rho_{\mathbf{G}^k}\mathbf{u}^k\cdot\mathbf{U}^\delta\,\mathrm{d}\mathbf{x} = -2\mu\int_{\mathcal{F}^k(t)}D(\mathbf{u}^k):D(\mathbf{U}^\delta)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{F}^k(t)}\left(\partial_t\mathbf{U}^\delta+(\mathbf{u}^k_\varepsilon\cdot\nabla)\mathbf{U}^\delta\right)\cdot\mathbf{u}^k\,\mathrm{d}\mathbf{x} - \frac{1}{2}\int_{\partial\mathcal{B}_{\mathbf{G}^k(t)}}(\mathbf{U}^\delta\cdot\mathbf{u}^k)((\mathbf{u}^k_\varepsilon-\mathbf{u}^k)\cdot\mathbf{n})\,\mathrm{d}\sigma. \tag{2.27}$$

Following the estimates of [7], using Arzelà and Ascoli and the diagonal Cantor procedure, we obtain that for all $\mathbf{U}^\delta$ satisfying (2.26), we have

$$\int_{\mathbb{R}^3_+}\rho_{\mathbf{G}^k}\mathbf{u}^k\cdot\mathbf{U}^\delta\,\mathrm{d}\mathbf{x}\to\int_{\mathbb{R}^3_+}\rho_{\mathbf{G}}\mathbf{u}\cdot\mathbf{U}^\delta\,\mathrm{d}\mathbf{x}\quad\text{in } C([0,T]). \tag{2.28}$$

However, as $\mathbf{G}^k\to\mathbf{G}$ in $\mathcal{C}([0,T])$, and $\mathbf{u}^k$ is bounded in $L^\infty(0,T;L^2(\mathbb{R}^3_+))\cap L^2(0,T;H^1_0(\mathbb{R}^3_+))$, this also leads to

$$\int_{\mathbb{R}^3_+}\rho_{\mathbf{G}}\mathbf{u}^k\cdot\mathbf{U}^\delta\,\mathrm{d}\mathbf{x}\to\int_{\mathbb{R}^3_+}\rho_{\mathbf{G}}\mathbf{u}\cdot\mathbf{U}^\delta\,\mathrm{d}\mathbf{x}\quad\text{in } C([0,T]). \tag{2.29}$$

Via a density argument (see Lemmas 4.1 and 4.2 in [7] for details), we might extend this convergence result to all $\mathbf{U}\in C([0,T];L^2(\mathbb{R}^3_+))$, $\operatorname{div}\mathbf{U}=0$, $D(\mathbf{U})=0$ in $\mathcal{B}(\mathbf{G}(t),1)$,

$$\int_{\mathbb{R}^3_+}\rho_{\mathbf{G}}\mathbf{u}^k\cdot\mathbf{U}\,\mathrm{d}\mathbf{x}\to\int_{\mathbb{R}^3_+}\rho_{\mathbf{G}}\mathbf{u}\cdot\mathbf{U}\,\mathrm{d}\mathbf{x}\quad\text{in } C([0,T]), \tag{2.30}$$

and to all $\mathbf{U}\in L^2(0,T;L^2(\mathbb{R}^3_+))$, $\operatorname{div}\mathbf{U}=0$, $D(\mathbf{U})=0$ in $\mathcal{B}(\mathbf{G}(t),1)$,

$$\int_{\mathbb{R}^3_+}\rho_{\mathbf{G}}\mathbf{u}^k\cdot\mathbf{U}\,\mathrm{d}\mathbf{x}\to\int_{\mathbb{R}^3_+}\rho_{\mathbf{G}}\mathbf{u}\cdot\mathbf{U}\,\mathrm{d}\mathbf{x}\quad\text{in } L^2(0,T). \tag{2.31}$$

Relation (2.30) implies in particular that $\mathbf{u}$ satisfies the initial conditions.

However, in (2.25), no assumption is made on the velocity field ψ over $\mathcal{B}_{\mathbf{G}}$. In order to reduce (2.25) to the above convergence result, we need to project such $\psi\in L^\infty(\mathcal{O})$ on velocity fields which are rigid over $\mathcal{B}_{\mathbf{G}}$. Similarly, $\mathbf{u}^k$ is rigid on $\mathcal{B}_{\mathbf{G}^k}$, so we also modify $\mathbf{u}^k$ slightly to obtain a function with the same properties but which is rigid in $\mathcal{B}_{\mathbf{G}}$. To this end, we extend $\mathbf{u}^k$ by 0 in $\mathbb{R}^3\setminus\mathbb{R}^3_+$, and since $\mathbf{u}^k=0$ on $\mathcal{P}$, we have $\mathbf{u}^k\in L^2(0,T;H^1(\mathbb{R}^3))$. We set

$$\widehat{\mathbf{u}}^k(t,\mathbf{y})=\mathbf{u}^k(t,\mathbf{y}+\mathbf{G}^k(t)-\mathbf{G}(t)).$$

We have

$$\operatorname{div}\widehat{\mathbf{u}}^k=0,\quad \widehat{\mathbf{u}}^k=0\quad\text{on }\{\mathbf{y}\in\mathbb{R}^3\ ;\ y_3=G_3-G_3^k\},\quad D(\mathbf{u}^k)=0\quad\text{in }\mathcal{B}_{\mathbf{G}}$$

and

$$\|\widehat{\mathbf{u}}^k-\mathbf{u}^k\|_{L^2((0,T)\times\mathbb{R}^3)}\leqslant\|\mathbf{G}^k-\mathbf{G}\|_{L^\infty(0,T)}\|\nabla\mathbf{u}^k\|_{L^2((0,T)\times\mathbb{R}^3)}\to 0. \tag{2.32}$$

We define

$$\mathcal{L}^2(\mathbb{R}^3)=\left\{\mathbf{v}\in L^2(\mathbb{R}^3)\ ;\ \operatorname{div}\mathbf{v}=0\quad\text{in }\mathbb{R}^3\right\}.$$

Following [7] (see (4.37) in [7]), there exists a function $\Lambda : \mathcal{L}^2(\mathbb{R}^3) \to \mathcal{L}^2(\mathbb{R}^3)$ such that for all $\mathbf{v} \in \mathcal{L}^2(\mathbb{R}^3)$, $\mathbf{u} = \Lambda(\mathbf{v})$ satisfies

$$\mathbf{u} = \mathbf{v} \quad \text{in } \mathcal{B}_{\mathbf{G}^0}, \quad \mathbf{u} = 0 \quad \text{in } \mathbb{R}^3 \setminus \overline{\mathcal{B}\left(\mathbf{G}^0, 1 + \frac{d^0}{4}\right)},$$

$$\|\mathbf{u}\|_{L^2(\mathbb{R}^3)} \leqslant c\|\mathbf{v}\|_{L^2(\mathbb{R}^3)}.$$

Moreover,

$$\text{if} \quad \mathbf{v} \in C([0,T]; L^2(\mathbb{R}^3)), \quad \text{then} \quad \Lambda\mathbf{v} \in C([0,T]; L^2(\mathbb{R}^3)),$$

and

$$\text{if} \quad \mathbf{v} \in L^2(0,T; L^2(\mathbb{R}^3)), \quad \text{then} \quad \Lambda\mathbf{v} \in L^2(0,T; L^2(\mathbb{R}^3)).$$

We also define $\check{\Lambda}^t : \mathcal{L}^2(\mathbb{R}^3) \to \mathcal{L}^2(\mathbb{R}^3)$ as follows:

$$\check{\Lambda}^t \check{\mathbf{v}}(\mathbf{y}) = [\Lambda\check{\mathbf{v}}]\,(\mathbf{y} + \mathbf{G}^0 - \mathbf{G}(t)),$$

where

$$\check{\mathbf{v}}(\mathbf{x}) = \mathbf{v}(\mathbf{x} + \mathbf{G}(t) - \mathbf{G}^0).$$

Then, as in [7], we consider

$$\overline{\mathbf{u}}^k(\mathbf{x}, t) = \mathbf{u}^k + \check{\Lambda}^t(\widehat{\mathbf{u}}^k - \mathbf{u}^k).$$

This function is rigid in $\mathcal{B}_{\mathbf{G}(t)}$ and

$$\|\overline{\mathbf{u}}^k - \mathbf{u}^k\|_{L^2((0,T)\times\mathbb{R}^3)} \to 0. \tag{2.33}$$

We are now in a position to prove (2.25) by using (2.31). Assume that $\boldsymbol{\psi} \in L^\infty(\mathcal{O})$. We set

$$I_k := \int_0^T \left(\int_{\mathcal{O}} \rho_{\mathbf{G}} \left(\mathbf{u}^k(t) - \mathbf{u}(t)\right) \cdot \boldsymbol{\psi} \, \mathrm{d}\mathbf{y} \right)^2 \mathrm{d}t$$

and

$$\overline{I}_k = \int_0^T \left(\int_{\mathcal{O}} \rho_{\mathbf{G}} \left(\overline{\mathbf{u}}^k(t) - \mathbf{u}(t)\right) \cdot \boldsymbol{\psi} \, \mathrm{d}\mathbf{y} \right)^2 \mathrm{d}t.$$

From (2.33), to prove that I_k goes to 0, it is sufficient to prove that $\overline{I}_k$ goes to 0. Then, given $\mathbf{G} \in \mathbb{R}^3_+$ such that $\operatorname{dist}(\mathcal{B}_{\mathbf{G}}, \mathcal{P}) > 0$, we denote by $P(\mathbf{G})$ the orthogonal projection from $L^2(\mathbb{R}^3_+, \rho_{\mathbf{G}}\mathrm{d}\mathbf{x})$ onto $\mathbb{H}(\mathbf{G})$ and introduce

$$\widetilde{\boldsymbol{\psi}}(t, \mathbf{x}) = P(\mathbf{G}(t))\boldsymbol{\psi}(\mathbf{x}).$$

We notice that $\widetilde{\boldsymbol{\psi}} \in L^\infty(0,T; L^2(\mathbb{R}^3_+))$, $\operatorname{div}\widetilde{\boldsymbol{\psi}} = 0$, $D(\widetilde{\boldsymbol{\psi}}) = 0$ in $\mathcal{B}(\mathbf{G}(t), 1)$ and

$$\overline{I}_k := \int_0^T \left(\int_{\mathcal{O}} \rho_{\mathbf{G}} \left(\overline{\mathbf{u}}^k(t) - \mathbf{u}(t)\right) \cdot \widetilde{\boldsymbol{\psi}} \, \mathrm{d}\mathbf{y} \right)^2 \mathrm{d}t.$$

We set

$$\widetilde{I}_k := \int_0^T \left(\int_{\mathcal{O}} \rho_{\mathbf{G}} \left(\mathbf{u}^k(t) - \mathbf{u}(t) \right) \cdot \widetilde{\boldsymbol{\psi}} \, \mathrm{d}\mathbf{y} \right)^2 \, \mathrm{d}t$$

and notice that, as before, if $\widetilde{I}_k$ goes to 0, then $\overline{I}_k$ goes to 0. However, since $\widetilde{\boldsymbol{\psi}} \in L^\infty(0,T;L^2(\mathbb{R}^3_+))$ we can take $\mathbf{U} = \widetilde{\boldsymbol{\psi}}$ in (2.31). Consequently, we obtain that there exists a subsequence we do not relabel such that $\overline{I}_k \to 0$. This ends the proof of L^2-compactness of the sequence $\mathbf{u}^k$. In particular we conclude from (2.24) that

$$\left\| \mathbf{u}^k - \mathbf{u} \right\|_{L^2(0,T;L^2(\mathcal{O}))} \to 0.$$

Combining (2.33) and the above relation, we obtain that

$$\left\| \overline{\mathbf{u}}^k - \mathbf{u} \right\|_{L^2(0,T;L^2(\mathcal{B}_{\mathbf{G}}))} \to 0.$$

This implies that

$$\dot{\mathbf{G}}^k \to \dot{\mathbf{G}} \quad \text{and} \quad \boldsymbol{\omega}^k \to \boldsymbol{\omega} \quad \text{in } L^2(0,T).$$

Using the above compactness, we can pass to the limit in (2.23) in the first three terms. For the two terms of the last line of (2.23), we proceed as follows. First, we can notice that

$$\int_{\mathcal{F}_{\mathbf{G}^k(t)}} \left| \mathbf{u}^k_{\varepsilon^k} - \mathbf{u}^k \right|^2 \, \mathrm{d}\mathbf{x} \leqslant C\varepsilon^k \|\mathbf{u}^k\|^2_{H^1(\mathcal{F}_{\mathbf{G}^k(t)})}$$

and thus

$$\int_0^T \int_{\mathcal{F}_{\mathbf{G}^k(t)}} \left[((\mathbf{u}^k_{\varepsilon^k} - \mathbf{u}^k) \cdot \nabla) \mathbf{U} \right] \cdot \mathbf{u}^k \, \mathrm{d}\mathbf{x} \to 0,$$

as $k \to \infty$.

Second, from the choice of κ and K_ε (see (2.7)), we can easily check that

$$\mathbf{u}^k_{\varepsilon^k} = \dot{\mathbf{G}}^k + \boldsymbol{\omega}^k \times (\mathbf{x} - \mathbf{G}^k) \quad \text{in } \mathcal{B}(\mathbf{G}^k(t), 1 - \varepsilon^k).$$

As a consequence, using Lemma 4.10 in [5], we conclude that

$$\int_{\partial\mathcal{B}_{\mathbf{G}^k(t)}} \left| \mathbf{u}^k_{\varepsilon^k} - \mathbf{u}^k \right|^2 \, \mathrm{d}\sigma \leqslant C\varepsilon^k \int_{\mathcal{B}_{\mathbf{G}^k(t)}} \left| \nabla \mathbf{u}^k_{\varepsilon^k} - \nabla \mathbf{u}^k \right|^2 \, \mathrm{d}\mathbf{x}.$$

The above inequality implies that

$$-\frac{1}{2} \int_0^T \int_{\partial\mathcal{B}_{\mathbf{G}^k(t)}} (\mathbf{U} \cdot \mathbf{u}^k)((\mathbf{u}^k_{\varepsilon^k} - \mathbf{u}^k) \cdot \mathbf{n}) \, \mathrm{d}\sigma \to 0,$$

as $k \to \infty$.

Finally, we can pass to the limit in (2.23) and obtain the weak formulation (2.4) for smooth test functions. We can pass from smooth test functions to the required regularity for $\mathbf{v}$ by applying the same approximation technique as when we obtained (2.29).

3. Constructing test functions. Let us begin with some notation. We introduce (r, θ, z), the cylindrical coordinates associated to (x_1, x_2, x_3):

$$x_1 = r\cos(\theta), \quad x_2 = r\sin(\theta), \quad x_3 = z.$$

Given $h > 0$ and $l > 0$, we denote by $\Omega_{h,l}$ the cylindric domain under $\mathcal{B}_h$ with radius l:

$$\Omega_{h,l} := \{(r, \theta, z) \in \mathcal{F}_h \text{ such that } r \in [0, l), z \in (0, 1 + h)\}. \tag{3.1}$$

We notice that whenever $l < 1$, the upper boundary of $\Omega_{h,\delta}$ is parametrized by (r, θ):

$$(r, \theta, z) \in \partial\Omega_{h,l} \cap \partial\mathcal{B}_h \Leftrightarrow \{r \in [0, \delta) \text{ and } z = \delta_h(r)\},$$

where, for arbitrary nonnegative h,

$$\delta_h(s) := 1 + h - \sqrt{1 - s^2} \quad \forall\, s \in [0, 1). \tag{3.2}$$

As in [9], we estimate the distance between $\mathcal{B}$ and $\mathcal{P}$ from below with a suitable choice of test function in the weak formulation. To this end, we introduce an approximation of the Stokes solution for a given position of $\mathcal{B}$ in $\mathbb{R}^3_+$ (namely, $\mathcal{B}_h$). We call these approximations "static functions" and denote them by $(\mathbf{w}_h)_{h>0}$. Given a weak solution $(\mathbf{u}, \mathbf{G})$ to (FSIS) in $(0, T)$, we construct admissible test functions by setting

$$\begin{array}{rccl} \widetilde{\mathbf{w}}: & (0,T)\times\mathbb{R}^3_+ & \longrightarrow & \mathbb{R}^3, \\ & (t, \mathbf{x}) & \longmapsto & \zeta(t)\mathbf{w}_{h(t)}(x_1 - G_1(t), x_2 - G_2(t), x_3) \end{array} \tag{3.3}$$

for arbitrary $\zeta \in \mathcal{D}(0, T)$. In this definition $h(t)$ stands for the distance between the sphere and the ramp at time t.

Applying the weak formulation, we obtain

$$\int_0^T \int_{\mathbb{R}^3_+} [\rho_{\mathbf{G}}\mathbf{u} \cdot \widetilde{\mathbf{w}}_t + (\mathbf{u} \otimes \mathbf{u} - 2\mu \mathrm{D}(\mathbf{u})) : \mathrm{D}(\widetilde{\mathbf{w}})] \ \mathrm{d}\mathbf{x}\ \mathrm{d}t = 0.$$

In this equation, the key ingredient is

$$\int_{\mathbb{R}^3_+} \mathrm{D}(\mathbf{u}) : \mathrm{D}(\mathbf{w}_h)\ \mathrm{d}\mathbf{x}.$$

It shall behave like $\dot{h}/h^\alpha$ with an exponent α to be made precise. The other terms appear as remainders. We shall bound them by an integrable (in time) function. This relies on the following lemma.

LEMMA 3.1. *Given $h > 0$, $r_0 > 0$, and $(\mathbf{u}, \mathbf{w}) \in H^1_0(\mathbb{R}^3_+) \times (\mathbb{H}(\mathbf{G}_h) \cap \mathcal{C}^\infty(\mathcal{F}_h))$, we assume $\mathbf{w}$ is with compact support. Then there exists C depending only on the size of the support of $\mathbf{w}$ such that*

$$\left| \int_{\mathbb{R}^3_+} \mathbf{u} \cdot \mathbf{w} \ \ \mathrm{d}\mathbf{x} \right| \leqslant C \|\nabla \mathbf{u}\|_{L^2(\mathbb{R}^3_+)} \left[\|\mathbf{w}\|_{2,2} + \|\mathbf{w}\|_{L^2\left(\mathbb{R}^3_+ \setminus \Omega_{h,r_0}\right)} \right], \tag{3.4}$$

where

$$\|\mathbf{w}\|^2_{2,2} = \int_0^{r_0} \left(\delta_h(r)^2 \left[\int_0^{\delta_h(r)} \sup_{\theta \in (0, 2\pi)} \{|\mathbf{w}(r, \theta, z)|^2)\}\ \mathrm{d}z \right] \right) r\ \mathrm{d}r.$$

If, moreover, $\mathbf{w} \in \mathbb{V}(\mathbf{G}_h)$*, we have*

$$\left| \int_{\mathbb{R}^3_+} \mathbf{u} \otimes \mathbf{u} : D(\mathbf{w}) \ \mathrm{d}\mathbf{x} \right| \leqslant C \|\nabla \mathbf{u}\|^2_{L^2(\mathbb{R}^3_+)} \left[\|\mathbf{w}\|_{\infty,2} + \|D(\mathbf{w})\|_{L^\infty\left(\mathcal{F}_h \setminus \Omega_{h,r_0}\right)} \right], \tag{3.5}$$

where

$$\|\mathbf{w}\|_{\infty,2} = \sup_{r \in (0,r_0)} \left\{ \delta_h(r)^{\frac{3}{2}} \left[\int_0^{\delta_h(r)} \sup_{\theta \in (0,2\pi)} \{ |\nabla \mathbf{w}(r,\theta,z)|^2 \} \ \mathrm{d}z \right]^{\frac{1}{2}} \right\}.$$

Proof. We denote by I_1 and I_2 the two integrals we want to estimate in (3.5) and (3.4).

We first deal with I_1. As $D(\mathbf{w}) = 0$ in $\mathcal{B}_h$, we might restrict the integration domain to $\mathcal{F}_h$. We split the integral into an integral in $\mathcal{F}_h \setminus \Omega_{h,r_0}$ and an integral in Ω_{h,r_0}: $I_1 = I_1^{in} + I_1^{out}$ with

$$|I_1^{out}| = \left| \int_{\mathcal{F}_h \setminus \Omega_{h,r_0}} \mathbf{u} \otimes \mathbf{u} : D(\mathbf{w}) \ \mathrm{d}\mathbf{x} \right| \leqslant \|\mathbf{u}\|^2_{L^2(\mathrm{Supp}(\mathbf{w}))} \|D(\mathbf{w})\|_{L^\infty\left(\mathcal{F}_h \setminus \Omega_{h,r_0}\right)}.$$

Because $\mathrm{Supp}(\mathbf{w})$ is bounded and $\mathbf{u} \in H^1_0(\mathbb{R}^3_+)$, we can use the Poincaré inequality. Concerning the integral in Ω_{h,r_0}, we have

$$I_1^{in} = \int_0^{2\pi} \int_0^{r_0} \int_0^{\delta_h(r)} [\mathbf{u}(r,\theta,z) \otimes \mathbf{u}(r,\theta,z) : D(\mathbf{w})(r,\theta,z)] \, r \ \mathrm{d}z \ \mathrm{d}r \ \mathrm{d}\theta.$$

Using a Hölder inequality with respect to the z-variable, we deduce

$$|I_1^{in}| \leqslant C \int_0^{2\pi} \int_0^{r_0} \left[\int_0^{\delta_h(r)} |\mathbf{u}(r,\theta,z)|^4 \mathrm{d}z \right]^{\frac{1}{2}} \left[\int_0^{\delta_h(r)} |D(\mathbf{w})|^2 \ \mathrm{d}z \right]^{\frac{1}{2}} r \ \mathrm{d}r \ \mathrm{d}\theta.$$

Then a direct generalization of the Poincaré inequality (see Lemma 12 in [9]) implies

$$\left[\int_0^{\delta_h(r)} |\mathbf{u}(r,\theta,z)|^4 \ \mathrm{d}z \right]^{\frac{1}{2}} \leqslant C \delta_h(r)^{\frac{3}{2}} \left[\int_0^{\delta_h(r)} |\nabla \mathbf{u}(r,\theta,z)|^2 \ \mathrm{d}z \right].$$

Substituting this in I_1^{in} and using again a Hölder inequality, we then obtain (3.5).

To estimate I_2, we decompose it in the same manner as I_1, and with the same proof, we deduce that there exists $C = C(\mathrm{Supp}(\mathbf{w}))$ such that

$$|I_2^{out}| \leqslant C \|\nabla \mathbf{u}\|_{L^2(\mathbb{R}^3_+)} \|\mathbf{w}\|_{L^2\left(\mathbb{R}^3_+ \setminus \Omega_{h,r_0}\right)}.$$

It remains to estimate the integral in Ω_{h,r_0}:

$$I_2^{in} = \int_0^{2\pi} \int_0^{r_0} \int_0^{\delta_h(r)} [\mathbf{u}(r,\theta,z) \cdot \mathbf{w}(r,\theta,z)] \, r \ \mathrm{d}z \ \mathrm{d}r \ \mathrm{d}\theta.$$

As above, a Hölder inequality in the z-variable associated to the Poincaré inequality implies

$$|I_2^{in}| \leqslant C \int_0^{2\pi} \int_0^{r_0} \left[\int_0^{\delta_h(r)} |\nabla \mathbf{u}(r,\theta,z)|^2 \ \mathrm{d}z \right]^{\frac{1}{2}} \delta_h(r) \left[\int_0^{\delta_h(r)} |\mathbf{w}|^2 \ \mathrm{d}z \right]^{\frac{1}{2}} r \ \mathrm{d}r \ \mathrm{d}\theta.$$

We conclude by using a Cauchy–Schwarz inequality. ∎

3.1. Explicit formula. From now on h is a fixed positive parameter. As in [9], we introduce a velocity field which is a good approximation (in a sense to be made precise) to the solution to the Stokes problem

$$\left\{\begin{array}{rl} \operatorname{div}\mathbb{T}(\mathbf{w},q)=0 & \\ \operatorname{div}\mathbf{w}=0 & \text{in } \mathcal{F}_h,\\ \mathbf{w}_{|\mathcal{P}}=0, & \\ \mathbf{w}_{|\partial\mathcal{B}_h}=\mathbf{e}_3. & \end{array}\right. \tag{3.6}$$

At first, we focus on the divergence-free and boundary conditions. So we introduce a potential vector field $\mathbf{a}_h$ and set $\mathbf{w}_h=\operatorname{curl}\mathbf{a}_h$. One choice for $\mathbf{a}_h$ could be

$$\mathbf{a}_h^s(\mathbf{x}):=\frac{\eta_{h_0}(|\mathbf{x}-\mathbf{G}_h|-1)}{2}(\mathbf{e}_3\times(\mathbf{x}-\mathbf{G}_h))\quad\forall\,\mathbf{x}\in\mathcal{F}_h,\quad\text{with } h_0>0.$$

The field $\mathbf{w}_h^s:=\operatorname{curl}\mathbf{a}_h^s$ satisfies the divergence-free and boundary conditions regardless of the value of $h_0<h$. However, when h goes to 0, this particular velocity field does not take advantage of the particular shape of the aperture between $\mathcal{B}$ and $\mathcal{P}$. Thus, we need to find another velocity field, especially in this aperture, in $\Omega_{h,1/2}$.

As we want to obtain an approximation of the Stokes problem, we construct a velocity field in which the fluid escapes from under the sphere with the most efficiency. Consequently, we want the velocity field to be planar and radial in each plane. Thus, our potential vector field reads, in cylindrical coordinates,

$$\mathbf{a}_h^d(r,\theta,z)=\big(-\phi_h^d(r,z)\sin(\theta),\phi_h^d(r,z)\cos(\theta),0\big)^\top\quad\forall\,(r,\theta,z)\in\Omega_{h,1/2},$$

so that, for all $(r,\theta,z)\in\Omega_{h,1/2}$,

$$\mathbf{w}_h^d(r,\theta,z)=\left(-\partial_z\phi_h^d(r,z)\cos(\theta),-\partial_z\phi_h^d(r,z)\sin(\theta),\partial_r\phi_h^d(r,z)+\frac{\phi_h^d(r,z)}{r}\right)^\top.$$

We set, in order to fit boundary conditions (this shall be critical in Lemma 3.2),

$$\phi_h^d(r,z)=r\chi_o\left(\frac{z}{\delta_h(r)}\right),\ \text{with } \chi_o(s)=\frac{s^2(3-2s)}{2}\quad\forall\,s\in(0,1).$$

From now on, we set $h_0=(\sqrt{17/16}-1)/2$. It remains to interpolate $\mathbf{w}_h^s$ and $\mathbf{w}_h^d$ so that we obtain

$$\mathbf{a}_h(\mathbf{x})=\left\{\begin{array}{rl} \eta_{1/2}(r)\mathbf{a}_h^d(r,\theta,z)+(1-\eta_{1/2}(r))\mathbf{a}_h^s(\mathbf{x}) & \text{in } \Omega_{h,1/2},\\ \mathbf{a}_h^s(\mathbf{x}) & \text{in } \mathbb{R}^3_+\setminus\Omega_{h,1/2}\end{array}\right.$$

and $\mathbf{w}_h=\operatorname{curl}\mathbf{a}_h$. Explicitly, in $\Omega_{h,1/2}$, we have

$$\mathbf{w}_h(r,\theta,z)=\eta_{1/2}(r)\mathbf{w}_h^d(r,\theta,z)+(1-\eta_{1/2}(r))\mathbf{w}_h^s(\mathbf{x})+\mathbf{rem}_0(\mathbf{x}),\tag{3.7}$$

where, denoting by $\boldsymbol{n\pi}_3(\mathbf{x})=(x_1,x_2,0)^\top/\sqrt{x_1^2+x_2^2}$, we have

$$\mathbf{rem}_0(\mathbf{x})=\eta'_{1/2}(r)\boldsymbol{n\pi}_3(\mathbf{x})\times\big(\mathbf{a}_h^d(r,\theta,z)-\mathbf{a}_h^s(\mathbf{x})\big)\quad\text{in } \Omega_{h,1/2}.$$

3.2. From static to moving test function. The main point in this subsection is to prove that, given a weak solution to (FSIS) $(\mathbf{u}, \mathbf{G})$ and $\zeta \in \mathcal{D}(0,T)$, the function $\widetilde{\mathbf{w}}$ constructed in (3.3) is a suitable test function. To this end, we need to extend $\mathbf{w}_h$ on $\mathbb{R}^3_+$ first. This is possible thanks to the following technical result.

LEMMA 3.2. *Given $h > 0$, we have*

$$\begin{aligned} &\mathbf{w}_h(\mathbf{x}) = \mathbf{e}_3, \quad &&\mathbf{a}_h(\mathbf{x}) = (\mathbf{e}_3 \times \mathbf{x})/2 \quad &&\forall\, \mathbf{x} \in \partial\mathcal{B}_h, \\ &\mathbf{w}_h(\mathbf{x}) = 0, \quad &&\mathbf{a}_h(\mathbf{x}) = 0 \quad &&\forall\, \mathbf{x} \in \mathcal{P}. \end{aligned}$$

Proof. We set $\lambda = z/\delta_h(r)$ and differentiations of λ by subscripts. We have

$$\partial_z \phi_h^d(r,z) = r\lambda_z \chi_o'(\lambda), \quad \partial_r \phi_h^d(r,z) = \chi_o(\lambda) + r\lambda_r \chi_o'(\lambda). \tag{3.8}$$

Computing with the value of λ yields

$$\lambda_z = \frac{1}{\delta_h(r)}, \quad \lambda_r = -\frac{z\delta_h'(r)}{(\delta_h(r))^2}.$$

Our choice for χ_o implies that

$$\chi_o(0) = \chi_o'(0) = 0, \quad \chi_o(1) = \frac{1}{2}, \quad \chi_o'(1) = 0.$$

Replacing λ by 0 in (3.8) yields

$$\phi_h^d(r,z) = \partial_z \phi_h^d(r,z) = \partial_r \phi_h^d(r,z) = 0 \quad \text{ on } \mathcal{P}.$$

Consequently, $\mathbf{a}_h^d = \mathbf{w}_h^d = 0$ on $\mathcal{P}$. Replacing λ by 1,

$$\partial_z \phi_h^d(r,z) = 0, \quad \phi_h^d(r,z) = \frac{r}{2}, \quad \partial_r \phi_h^d(r,z) = \frac{1}{2} \quad \text{ on } \partial\mathcal{B}_h.$$

Consequently, $\mathbf{a}_h^d(\mathbf{x}) = (\mathbf{e}_3 \times \mathbf{x})/2$ and $\mathbf{w}_h^d = \mathbf{e}_3$ on $\mathcal{B}_h$.

Concerning the smooth part, a straightforward computation leads to

$$\mathbf{w}_h^s(\mathbf{x}) = \eta_{h_0}(|\mathbf{x} - \mathbf{G}_h| - 1)\mathbf{e}_3 + \eta_{h_0}'(|\mathbf{x} - \mathbf{G}_h| - 1)\frac{\mathbf{x} - \mathbf{G}_h}{|\mathbf{x} - \mathbf{G}_h|} \times \frac{(\mathbf{e}_3 \times (\mathbf{x} - \mathbf{G}_h))}{2}.$$

Due to our choice, we have

$$\eta_{h_0}(|\mathbf{x} - \mathbf{G}_h| - 1) = 1, \quad \eta_{h_0}'(|\mathbf{x} - \mathbf{G}_h| - 1) = 0 \quad \text{ on } \partial\mathcal{B}_h.$$

Consequently $\mathbf{a}_h^s(\mathbf{x}) = (\mathbf{e}_3 \times \mathbf{x})/2$ and $\mathbf{w}_h^s = \mathbf{e}_3$ on $\partial\mathcal{B}_h$. Then

$$\eta_{h_0}(|\mathbf{x} - \mathbf{G}_h| - 1) = 0, \quad \eta_{h_0}'(|\mathbf{x} - \mathbf{G}_h| - 1) = 0 \quad \text{ if } |\mathbf{x} - \mathbf{G}_h| \geqslant 1 + 2h_0 = \sqrt{17/16}.$$

Moreover, if $\mathbf{x} \in \mathcal{P} \setminus \overline{\Omega_{h,1/4}}$, we have $r > 1/4$ and, as $\mathbf{G}_h = (0,0,1+h)$,

$$|\mathbf{x} - \mathbf{G}_h|^2 > (1+h)^2 + (1/4)^2 > 17/16.$$

Consequently, $\mathbf{a}_h^s(\mathbf{x}) = \mathbf{w}_h^s(\mathbf{x}) = 0$ for $\mathbf{x} \in \mathcal{P} \setminus \overline{\Omega_{h,1/4}}$.

It remains to check that boundary conditions are satisfied in the transition region, i.e., when $\mathbf{x} \in \overline{\Omega_{h,1/2}} \setminus \Omega_{h,1/4}$. On $\mathcal{P}$, we remark that $\mathbf{w}_h^d(\mathbf{x}) = \mathbf{w}_h^s(\mathbf{x}) = 0 = \mathbf{a}_h^s(\mathbf{x}) =$

$\mathbf{a}_h^d(\mathbf{x}) = 0$. Interpolating the potential vector fields, we obtain $\mathbf{w}_h = 0$ on $\mathcal{P}$. Finally, on $\overline{\mathcal{B}_h \cap \overline{\Omega_{h,1/4}}}$ we have already computed

$$\mathbf{w}_h^d(\mathbf{x}) = \mathbf{w}_h^s(\mathbf{x}) = \mathbf{e}_3 \quad \text{and} \quad \mathbf{a}_h^s(\mathbf{x}) = \mathbf{a}_h^d(\mathbf{x}) = (\mathbf{e}_3 \times \mathbf{x})/2.$$

Interpolating the potential vector fields, we deduce $\mathbf{w}_h(\mathbf{x}) = \mathbf{e}_3$. This concludes the proof. ∎

Remark 3.1. According to this lemma, we extend $\mathbf{a}_h$ (resp., $\mathbf{w}_h$) to $\mathbb{R}^3_+$ with the value $(\mathbf{e}_3 \times \mathbf{x})/2$ (resp., $\mathbf{e}_3$) in $\mathcal{B}_h$. In what follows, we consider the functions $\mathbf{a} : (h, \mathbf{x}) \to \mathbf{a}_h(\mathbf{x})$ and $\mathbf{w} : (h, \mathbf{x}) \to \mathbf{w}_h(\mathbf{x})$. Denoting by $\mathcal{Q}_c = \{(h, \mathbf{x}) \in (0, 1) \times \mathbb{R}^3 \ ; \ \mathbf{x} \in \mathcal{B}_h\}$, standard analytic arguments imply $\mathbf{a} \in \mathcal{C}^\infty(\mathcal{Q}_c) \cap \mathcal{C}^\infty(((0, 1) \times \mathbb{R}^3_+) \setminus \overline{\mathcal{Q}_c})$. We note that $\mathbf{w}_h$ vanishes as soon as $|\mathbf{x} - \mathbf{G}_h| > (\sqrt{17/16} - 1)/2$ and $|\mathbf{x}| > 1/2$. Consequently, the above lemma implies $\mathbf{w} \in H^1((\bar{h}, 1) \times \mathbb{R}^3_+)$ for any $\bar{h} > 0$ and, after standard composition arguments, this yields

$$\widetilde{\mathbf{w}} \in \mathcal{C}([0, T]; \mathrm{H}^1(\mathbb{R}^3_+)) \cap H^1(0, T; \mathrm{L}^2(\mathbb{R}^3_+))$$

as long as $h(t) \in (\bar{h}, 1]$ for all $t \in (0, T)$. So, $\widetilde{\mathbf{w}}$ is a suitable test function for the weak formulation as long as $h(0, T) \subset (\bar{h}, 1)$.

3.3. Estimate of remainder terms. In order to exploit the weak formulation with our test function, we need to dominate remainder terms according to Lemma 3.1. We begin with estimates on Sobolev norms of $\mathbf{w}_h$.

By construction, our test functions behave differently under the sphere (in $\Omega_{h,\delta}$) and above the sphere (in $\mathcal{F}_h \setminus \Omega_{h,\delta}$ for arbitrary fixed $\delta > 0$). Above the sphere we have the following.

LEMMA 3.3. *Given $\alpha \geqslant 0$ and $\delta > 0$ there exists $C(\alpha, \delta) < \infty$ such that*

$$\|\mathbf{a}_h\|_{H^\alpha(\mathcal{F}_h \setminus \Omega_{h,\delta})} \leqslant C(\alpha, \delta) \quad \forall\, h \in (0, 1).$$

Proof. By construction the restriction $\mathbf{a} : \mathcal{Q}_{c,\delta} \to \mathbb{R}^3$, with

$$\mathcal{Q}_{c,\delta} := \{(h, \mathbf{x}) \in [0, 1] \times \overline{\mathbb{R}^3_+} \quad \text{with } \mathbf{x} \notin \Omega_{h,\delta}\},$$

is smooth and with compact support. ∎

Inside $\Omega_{h,1/4}$, estimates rely essentially on dominations of integrals:

$$\int_0^{\frac{1}{4}} \frac{r^\alpha \, \mathrm{d}r}{[\delta_h(r)]^\beta}.$$

We refer the reader to the appendix for such computations.

LEMMA 3.4. *The family $(\mathbf{w}_h)_{0<h<1}$ is uniformly bounded in $L^2(\mathbb{R}^3_+)$.*

Proof. Because of the previous lemma, we focus on $\mathbf{w}_h^d$ inside $\Omega_{h,1/4}$.

In this region, we have

$$(\mathbf{w}_h^d(r, \theta, z))_1 = -\partial_z \phi_h^d(r, z) \cos(\theta), \quad (\mathbf{w}_h^d(r, \theta, z))_2 = -\partial_z \phi_h^d(r, z) \sin(\theta),$$

and

$$(\mathbf{w}_h^d(r, \theta, z))_3 = \partial_r \phi_h^d(r, z) + \frac{\phi_h^d(r, z)}{r}.$$

Thus

$$|\mathbf{w}_h^d(r, \theta, z)| \leqslant |\partial_z \phi_h^d(r, z)| + |\partial_r \phi_h^d(r, z)| + \frac{|\phi_h^d(r, z)|}{r}.$$

Applying Lemma A.3, this leads to

$$|\mathbf{w}_h^d(r,\theta,z)| \leqslant C\left(1+\frac{r}{\delta_h(r)}\right) \quad \forall\,(r,\theta,z)\in\Omega_{h,1/4}, \quad \forall\, h\in(0,1).$$

The result then follows from Lemma A.1 with $(\alpha,\beta)=(3,1)$. ∎

As a technical device for applying Lemma 3.1, we have the following.

LEMMA 3.5. *Let us define*

$$w_h(r,\theta,z)=|\partial_r \mathbf{w}_h^d(r,\theta,z)|+\frac{|\partial_\theta \mathbf{w}_h^d(r,\theta,z)|}{r}+|\partial_h \mathbf{w}_h^d(r,\theta,z)| \quad \forall\,(r,\theta,z)\in\Omega_{h,1/4}.$$

Then

$$\int_0^{\frac{1}{4}}\left(\delta_h(r)^2\left[\int_0^{\delta_h(r)}\sup_{\theta\in(0,2\pi)}|w_h(r,\theta,z)|^2\ \mathrm{d}z\right]\right) r\ \mathrm{d}r \tag{3.9}$$

is uniformly bounded for $h\in(0,1)$.

Proof. A straightforward computation yields, for all $(r,\theta,z)\in\Omega_{h,1/4}$,

$$|\partial_r \mathbf{w}_h^d(r,\theta,z)| \leqslant C\left(|\partial_{rz}\phi_h^d(r,z)|+|\partial_{rr}\phi_h^d(r,z)|+\left|\frac{\partial_r\phi_h^d(r,z)}{r}-\frac{\phi_h^d(r,z)}{r^2}\right|\right)$$

and

$$\begin{aligned}\frac{|\partial_\theta \mathbf{w}_h^d(r,\theta,z)|}{r} &\leqslant \frac{|\partial_z\phi_h^d(r,z)|}{r},\\ |\partial_h \mathbf{w}_h^d(r,\theta,z)| &\leqslant \frac{|\partial_h\phi_h^d(r,z)|}{r}+|\partial_{hr}\phi_h^d(r,z)|+|\partial_{hz}\phi_h^d(r,z)|.\end{aligned}$$

Combining the above inequalities with Lemma A.3, we deduce there exists a constant C independent of h such that

$$|w_h(r,\theta,z)| \leqslant C\left(\frac{1}{\delta_h(r)}+\frac{r}{\delta_h(r)^2}\right)$$

for all $(r,\theta,z)\in\Omega_{h,1/4}$. Consequently,

$$\int_0^{1/4}\left(\delta_h(r)^2\left[\int_0^{\delta_h(r)}\sup_{\theta\in(0,2\pi)}|w_h(r,\theta,z)|^2\ \mathrm{d}z\right]\right) r\ \mathrm{d}r \leqslant C\int_0^{1/4}(\delta_h(r)+1)r\ \mathrm{d}r.$$

As the last integral remains bounded when h goes to 0, the same holds for the integral of w_h. ∎

Then, to dominate the trilinear form, we need the following result.

LEMMA 3.6. *We set*

$$dw_h(r,\theta,z)=|\partial_r \mathbf{w}_h^d(r,\theta,z)|+\frac{|\partial_\theta \mathbf{w}_h^d(r,\theta,z)|}{r}+|\partial_z \mathbf{w}_h^d(r,\theta,z)| \quad \forall\,(r,\theta,z)\in\Omega_{h,1/4}.$$

Then the quantity

$$\sup_{r\in(0,1/4)}\left\{\delta_h(r)^{\frac{3}{2}}\left[\int_0^{\delta_h(r)}\sup_{\theta\in(0,2\pi)}\{|dw_h(r,\theta,z)|^2\}\ \mathrm{d}z\right]^{\frac{1}{2}}\right\} \tag{3.10}$$

is uniformly bounded for $h\in(0,1)$.

Proof. As in the previous proof, there exists a constant C independent of h such that

$$|dw_h(r,\theta,z)| \leqslant C\left(\frac{1}{\delta_h(r)} + \frac{r}{\delta_h(r)^2}\right)$$

for all $(r,\theta,z) \in \Omega_{h,1/4}$. Therefore

$$\int_0^{\delta_h(r)} |dw_h(r,z)|^2 \,\mathrm{d}z \leqslant C\left(\frac{1}{\delta_h(r)} + \frac{r^2}{\delta_h(r)^3}\right)$$

and

$$\sup_{r\in(0,1/4)}\left\{\delta_h(r)^{\frac{3}{2}}\left[\int_0^{\delta_h(r)} \sup_{\theta\in(0,2\pi)}\{|dw_h(r,\theta,z)|^2\}\,\mathrm{d}z\right]^{\frac{1}{2}}\right\} \leqslant C \sup_{r\in(0,1/4)}(\delta_h(r)+r),$$

which is uniformly bounded when $h \in (0,1)$. ∎

Finally, there holds the following lemma, which is reminiscent of works by Starovoitov.

LEMMA 3.7. *There exists a constant $C > 0$ such that*

$$|\nabla \mathbf{w}_h|_2^2 \geqslant \frac{C}{h} \quad \forall\, h \in (0,1).$$

Proof. Given $h > 0$ we already noticed that

$$\mathbf{w}_h(\mathbf{x}) = \mathbf{w}_h^d(\mathbf{x}) \quad \forall\, \mathbf{x} \in \Omega_{h,1/4}.$$

Consequently,

$$|\nabla \mathbf{w}_h(\mathbf{x})| \geqslant |\partial_z \mathbf{w}_h^d(\mathbf{x})| \quad \forall\, \mathbf{x} \in \Omega_{h,1/4}.$$

With the explicit formula for $\mathbf{w}_h^d$, we have $|\partial_z \mathbf{w}_h^d| \geqslant |\partial_{zz}\phi_h^d|$, where $|\partial_{zz}\phi_h^d(r,z)| = \dfrac{r}{\delta_h^2(r)}\chi_o''(\lambda)$. Consequently,

$$|\nabla \mathbf{w}_h|_2^2 \geqslant 2\pi \int_0^{\frac{1}{4}} \frac{r^3\,\mathrm{d}r}{\delta_h(r)^3} \int_0^1 |\chi_o''(s)|^2 \,\mathrm{d}s.$$

As χ is a polynomial with degree 3, its second derivative does not vanish, and neither does the s-integral. Then we obtain the result by applying Lemma A.1 with $\alpha = \beta = 3$. ∎

3.4. Our test function and the Stokes problem. First, we prove that our choice is a good one because it is a good approximation of the solution to the Stokes problem.

LEMMA 3.8. *There exist $q_h \in \mathcal{C}^\infty(\overline{\mathcal{F}_h})$ and $\mathbf{f}_h \in \mathcal{C}_c^\infty(\overline{\mathcal{F}_h})$ such that*

$$\left\{\begin{array}{ll} \mu\Delta \mathbf{w}_h - \nabla q_h = \mathbf{f}_h & \\ \operatorname{div} \mathbf{w}_h = 0 & \end{array}\right. \quad \text{in } \mathcal{F}_h, \tag{3.11}$$

where there exists an absolute constant C for which

$$\int_0^{2\pi}\int_0^{\frac{1}{4}}\left(\delta_h(r)^2\left[\int_0^{\delta_h(r)}|\mathbf{f}_h|^2\,\mathrm{d}z\right]\right) r\,\mathrm{d}r\,\mathrm{d}\theta + \|\mathbf{f}_h\|^2_{L^2(\mathcal{F}_h\setminus\Omega_{h,1/4})} \leqslant C.$$

Proof. By construction, we have $\mathbf{w}_h = \operatorname{curl} \widetilde{\mathbf{a}}_h^d + \operatorname{curl} \widetilde{\mathbf{a}}_h^s$, where

$$\widetilde{\mathbf{a}}_h^d(\mathbf{x}) = \begin{cases} \eta_{1/2}(r)\mathbf{a}_h^d(\mathbf{x}), & \mathbf{x} \in \Omega_{h,1/2}, \\ 0 \quad \text{else}, \end{cases} \quad \widetilde{\mathbf{a}}_h^s(\mathbf{x}) = \begin{cases} (1-\eta_{1/2}(r))\mathbf{a}_h^s(\mathbf{x}), & \mathbf{x} \in \Omega_{h,1/2}, \\ \mathbf{a}_h^s(\mathbf{x}) \quad \text{else}. \end{cases}$$

Then, according to Lemma 3.3, the smooth part $\widetilde{\mathbf{a}}_h^s$ is bounded in any Sobolev space uniformly in h. Consequently, $\widetilde{\mathbf{f}}_h = \mu\Delta \operatorname{curl} \widetilde{\mathbf{a}}_h^s$ is bounded in all Sobolev spaces. We have

$$\mu\Delta\mathbf{w}_h = \mu\Delta \operatorname{curl} \widetilde{\mathbf{a}}_h^d + \widetilde{\mathbf{f}}_h.$$

In the following we write $\widetilde{\phi}_h^d(r,z) = \eta_{1/2}(r)\phi_h^d(r,z)$ for all $(r,\theta,z) \in \Omega_{h,1/2}$. Let us recall that in cylindrical coordinates we have

$$\Delta = \frac{\partial_r[r\partial_r]}{r} + \frac{\partial_{\theta\theta}}{r^2} + \partial_{zz}.$$

Consequently,

$$\left[\Delta \operatorname{curl} \widetilde{\mathbf{a}}_h^d\right]_1 = -\left[\partial_{rrz}\widetilde{\phi}_h^d + \frac{\partial_{rz}\widetilde{\phi}_h^d}{r} - \frac{\partial_z\widetilde{\phi}_h^d}{r^2} + \partial_{zzz}\widetilde{\phi}_h^d\right]\cos(\theta)$$

and

$$\left[\Delta \operatorname{curl} \widetilde{\mathbf{a}}_h^d\right]_2 = -\left[\partial_{rrz}\widetilde{\phi}_h^d + \frac{\partial_{rz}\widetilde{\phi}_h^d}{r} - \frac{\partial_z\widetilde{\phi}_h^d}{r^2} + \partial_{zzz}\widetilde{\phi}_h^d\right]\sin(\theta),$$

$$\left[\Delta \operatorname{curl} \widetilde{\mathbf{a}}_h^d\right]_3 = \partial_{rrr}\widetilde{\phi}_h^d + 2\frac{\partial_{rr}\widetilde{\phi}_h^d}{r} - \frac{\partial_r\widetilde{\phi}_h^d}{r^2} + \frac{\widetilde{\phi}_h^d}{r^3} + \partial_{zz}\left[\partial_r\widetilde{\phi}_h^d + \frac{\widetilde{\phi}_h^d}{r}\right].$$

We remark here that

$$\partial_{zzz}\widetilde{\phi}_h^d(r,z) = -6\frac{r\,\eta_{1/2}(r)}{\delta_h^3(r)}.$$

Consequently, denoting by Φ a primitive of $s \mapsto -6s\eta_{1/2}(s)/(\delta_h(s))^3$, we have

$$\nabla\Phi(r) = (\partial_{zzz}\phi_h^d\cos(\theta), \partial_{zzz}\phi_h^d\sin(\theta), 0)^\top.$$

We set

$$q_h(\mathbf{x}) = \mu\Phi(r) + \mu\partial_z\left[\partial_r\widetilde{\phi}_h^d + \frac{\widetilde{\phi}_h^d}{r}\right], \quad \check{\mathbf{f}}_h = \mu\Delta \operatorname{curl} \widetilde{\mathbf{a}}_h^d - \nabla q_h.$$

In particular $\mu\Delta\mathbf{w}_h - \nabla q_h = \widetilde{\mathbf{f}}_h + \check{\mathbf{f}}_h$, so that our result follows from the same result for $\check{\mathbf{f}}_h$. Denoting by $\check{f}_1, \check{f}_2, \check{f}_3$ the Cartesian components of $\check{\mathbf{f}}_h$, straightforward computations show that

$$|\check{f}_1|^2 + |\check{f}_2|^2 \leqslant 4\left[\partial_{rrz}\widetilde{\phi}_h^d + \frac{\partial_{rz}\widetilde{\phi}_h^d}{r} - \frac{\partial_z\widetilde{\phi}_h^d}{r^2}\right]^2.$$

As $\eta'_{1/2}$ vanishes uniformly in $\Omega_{h,1/4}$, Lemma 3.3 implies there exists a universal constant C such that

$$|\check{f}_1|^2 + |\check{f}_2|^2 \leqslant C\left[1 + \left|\partial_{rrz}\phi_h^d + \frac{\partial_{rz}\phi_h^d}{r} - \frac{\partial_z\phi_h^d}{r^2}\right|\right]^2.$$

Then, for the same reasons, we have

$$|\check{f}_3|^2 \leqslant C\left[1 + \left|\partial_{rrr}\phi_h^d\right| + \left|\frac{2\partial_{rr}\phi_h^d}{r}\right| + \left|\frac{\phi_h^d}{r^3} - \frac{\partial_r\phi_h^d}{r^2}\right|\right]^2.$$

Replacing with the size computed in Lemma A.3, we obtain

$$|\check{\mathbf{f}}_h(\mathbf{x})|^2 \leqslant C\left(\frac{r}{\delta_h^2(r)} + \frac{1}{\delta_h(r)}\right)^2.$$

Consequently, for arbitrary $r \in (0, 1/2)$

$$\int_0^{\delta_h(r)} |\check{\mathbf{f}}_h|^2 \,\mathrm{d}z \leqslant C\left(\frac{r^2}{\delta_h(r)^3} + \frac{1}{\delta_h(r)}\right)$$

and

$$\int_0^{2\pi}\int_0^{1/2} \delta_h(r)^2 \int_0^{\delta_h(r)} |\check{\mathbf{f}}_h|^2 \,\mathrm{d}z r \,\mathrm{d}r \,\mathrm{d}\theta \leqslant C\int_0^{1/2}\left(\frac{r^3}{\delta_h(r)} + r\delta_h(r)\right)\,\mathrm{d}r,$$

which is uniformly bounded for $h \in (0,1)$. This concludes the proof. ∎

As a direct corollary, we get the following lemma.

LEMMA 3.9. *There exist $K_m < \infty$ and a function $\widetilde{n}_3 : [0,1] \to \mathbb{R}_+$ such that, for any $h < 1$ and $\mathbf{w} \in \mathbb{V}(\mathbf{G}_h)$ such that $\mathbf{w} = \mathbf{V}_\mathbf{w} + \mathbf{R}_\mathbf{w} \times (\mathbf{x} - \mathbf{G}_h)$ in $\mathcal{B}_h$, we have*

$$\left|2\mu\int_{\mathbb{R}^3_+} D(\mathbf{w}_h) : D(\mathbf{w}) \;\,\mathrm{d}\mathbf{x} - \widetilde{n}_3(h)\mathbf{V}_\mathbf{w}\cdot\mathbf{e}_3\right| \leqslant K_m\|\nabla\mathbf{w}\|_{L^2(\mathbb{R}^3_+)}. \tag{3.12}$$

Moreover, there exist $h_m > 0$ and a constant $c > 0$ such that $\widetilde{n}_3(h) \geqslant c/h$ for all $h < h_m$.

Proof. Given $h > 0$ and $\mathbf{w} \in \mathbb{V}(\mathbf{G}_h)$, we apply the Stokes identity with (3.11) and obtain

$$2\mu\int_{\mathbb{R}^3_+} \mathrm{D}(\mathbf{w}_h) : \mathrm{D}(\mathbf{w}) \,\mathrm{d}\mathbf{x} = \int_{\partial\mathcal{B}_h} \mathbb{T}(\mathbf{w}_h, p_h)\mathbf{n}\cdot\mathbf{w} \,\mathrm{d}\sigma - \int_{\mathcal{F}_h} \mathbf{f}_h\cdot\mathbf{w} \,\mathrm{d}\mathbf{x}. \tag{3.13}$$

For symmetry reasons, there exists $\widetilde{n}_3 : (0,1) \to \mathbb{R}$ such that

$$\int_{\partial\mathcal{B}_h} \mathbb{T}(\mathbf{w}_h, p_h)\mathbf{n} \,\mathrm{d}\sigma = \widetilde{n}_3(h)\mathbf{e}_3$$

and

$$\int_{\partial\mathcal{B}_h} (\mathbf{x} - \mathbf{G}_h) \times \mathbb{T}(\mathbf{w}_h, p_h)\mathbf{n} \,\mathrm{d}\sigma = 0.$$

On the other hand, applying (3.4) and Lemma 3.8, we also deduce

$$\left|\int_{\mathcal{F}_h} \mathbf{f}_h\cdot\mathbf{w} \,\mathrm{d}\mathbf{x}\right| \leqslant C\|\nabla\mathbf{w}\|_{L^2(\mathbb{R}^3_+)} \quad \forall\, h \in (0,1),$$

where C is a positive constant. Finally, we have obtained the existence of a constant K such that, for arbitrary $h \in (0,1)$ and $\mathbf{w} \in \mathbb{V}(\mathbf{G})$,

$$\left| 2\mu \int_{\mathbb{R}^3_+} \mathrm{D}(\mathbf{w}_h) : \mathrm{D}(\mathbf{w}) \, \mathrm{d}\mathbf{x} - \widetilde{n}_3(h) \mathbf{V}_{\mathbf{w}} \cdot \mathbf{e}_3 \right| \leqslant K \|\nabla \mathbf{w}\|_{L^2(\mathbb{R}^3_+)}.$$

In order to estimate $\widetilde{n}_3$, we take $\mathbf{w} = \mathbf{w}_h$ in (3.13) and obtain

$$\widetilde{n}_3 = \int_{\partial \mathcal{B}_h} \mathbb{T}(\mathbf{w}_h, p_h)\mathbf{n} \cdot \mathbf{e}_3 \, \mathrm{d}\sigma = 2\mu \int_{\mathbb{R}^3_+} |D(\mathbf{w}_h)|^2 \, \mathrm{d}\mathbf{x} + \int_{\mathcal{F}_h} \mathbf{f}_h \cdot \mathbf{w}_h \, \mathrm{d}\mathbf{x}. \tag{3.14}$$

Dealing as previously with the last integral, we deduce that

$$\left| \int_{\mathcal{F}_h} \mathbf{f}_h \cdot \mathbf{w}_h \, \mathrm{d}\mathbf{x} \right| \leqslant K \|\nabla \mathbf{w}_h\|_{L^2(\mathbb{R}^3_+)} \quad \forall\, h \in (0,1).$$

But, applying Lemma 3.7, we have that

$$2\mu \int_{\mathbb{R}^3_+} |D(\mathbf{w}_h)|^2 \, \mathrm{d}\mathbf{x} = \mu \int_{\mathbb{R}^3_+} |\nabla \mathbf{w}_h|^2 \, \mathrm{d}\mathbf{x} \geqslant \frac{C}{h} \quad \forall\, h \in (0,1).$$

Consequently, the asymptotic behavior of the right-hand side in (3.14) when h goes to 0 is prescribed by the first integral. Hence, there exist $h_m > 0$ and constants $\widetilde{c}, c > 0$ such that

$$\widetilde{n}_3(h) \geqslant \widetilde{c} \int_{\mathbb{R}^3_+} |D(\mathbf{w}_h)|^2 \, \mathrm{d}\mathbf{x} \geqslant \frac{c}{h} \quad \forall\, h \in (0, h_m). \qquad \square$$

4. Proof of Theorem 1.1. We let the reader convince himself that Theorem 1.1 is a direct consequence of the following theorem.

THEOREM 4.1. *Given $(\mathbf{u}, \mathbf{G})$ a weak solution to (FSIS) on $(0,T)$ with initial data $(\mathbf{u}^0, \mathbf{G}^0)$, we assume there exists $0 \leqslant \tau_0 < \tau_1 \leqslant T$ for which*

$$h(t) := \mathrm{dist}(\mathcal{B}(t), \mathcal{P}) \leqslant 1 \quad \forall\, t \in [\tau_0, \tau_1].$$

Then there exists $C(\|\mathbf{u}^0\|_{L^2(\mathbb{R}^3_+)}) < \infty$ depending only on the L^2-norm of initial data such that

$$h(t) \geqslant h(\tau_0) \exp\left[-C(\|\mathbf{u}^0\|_{L^2(\mathbb{R}^3_+)})(1 + \sqrt{T}) \right] \quad \forall\, t \in (\tau_0, \tau_1).$$

The remainder of this paper is devoted to the proof of this result. From now on $(\mathbf{u}, \mathbf{G})$ is a given weak solution to (FSIS) with initial data $(\mathbf{u}^0, \mathbf{G}^0)$. For simplicity, we assume that $h(t) \leqslant 1$ for all $t \in (0,T)$. This means that $\tau_0 = 0$ and $\tau_1 = T$ in the assumptions of our theorem.

As mentioned before, we estimate the distance h from below with our approximation of the Stokes problem. So, from now on, $(\mathbf{w}_h)_{h \in (0,1)}$ are the approximations constructed in section 3.1. Given $0 < t_0 < t_1 < 1$, we set

$$\zeta_\varepsilon(t) = \eta_\varepsilon(\mathrm{dist}(t, [t_0, t_1])).$$

Then $\zeta_\varepsilon \in \mathcal{D}(0,T)$ whenever ε is sufficiently small. Consequently, according to Remark 3.1, for ε sufficiently small

$$\begin{array}{rccl} \widetilde{\mathbf{w}}_\varepsilon : & (0,T) \times \mathbb{R}^3_+ & \longrightarrow & \mathbb{R}^3, \\ & (t, \mathbf{x}) & \longmapsto & \zeta_\varepsilon(t) \mathbf{w}_{h(t)}(x_1 - G_1(t), x_2 - G_2(t), x_3) \end{array}$$

can be taken as a test function in (2.4):

$$\int_{(0,T)\times\mathbb{R}^3_+} [\rho_{\mathbf{G}}\mathbf{u}\cdot\partial_t\widetilde{\mathbf{w}}_\varepsilon + (\mathbf{u}\otimes\mathbf{u} - 2\mu\mathrm{D}(\mathbf{u})) : \mathrm{D}(\widetilde{\mathbf{w}}_\varepsilon)] \ \ \mathrm{d}\mathbf{x} \ \ \mathrm{d}t = 0. \tag{4.1}$$

In the following, we set

$$\begin{aligned} I_1 &:= \int_{(0,T)\times\mathbb{R}^3_+} \rho_{\mathbf{G}}\mathbf{u}\cdot\partial_t\widetilde{\mathbf{w}}_\varepsilon \ \mathrm{d}t \ \mathrm{d}\mathbf{x}, \\ I_2 &:= \int_{(0,T)\times\mathbb{R}^3_+} \mathbf{u}\otimes\mathbf{u} : D(\widetilde{\mathbf{w}}_\varepsilon) \ \mathrm{d}t \ \mathrm{d}\mathbf{x}, \\ I_3 &:= \int_{(0,T)\times\mathbb{R}^3_+} D(\mathbf{u}) : D(\widetilde{\mathbf{w}}_\varepsilon) \ \mathrm{d}t \ \mathrm{d}\mathbf{x}. \end{aligned}$$

After a change of variables, we have for almost all $t \in (0,T)$

$$\int_{\mathbb{R}^3_+} D(\mathbf{u})(t,\cdot) : D(\widetilde{\mathbf{w}}_\varepsilon)(t,\cdot) \ \mathrm{d}\mathbf{x} = \zeta_\varepsilon(t) \int_{\mathbb{R}^3_+} D(\mathbf{u})(t, x_1+G_1, x_2+G_2, x_3) : D(\mathbf{w}_{h(t)}) \ \mathrm{d}\mathbf{x}.$$

Thus, applying Lemma 3.9,

$$\int_{\mathbb{R}^3_+} D(\mathbf{u})(t, x_1+G_1, x_2+G_2, x_3) : D(\mathbf{w}_{h(t)}) \ \mathrm{d}\mathbf{x} = \dot{h}\widetilde{n}_3(h) + E(t),$$

where $|E(t)| = K_M|\nabla\mathbf{u}(t,\cdot)|_2$. Consequently,

$$I_3 = \int_0^T \zeta_\varepsilon(t)\dot{h}(t)\widetilde{n}_3(h(t)) \ \mathrm{d}t + \widetilde{E}, \tag{4.2}$$

where

$$|\widetilde{E}| \leqslant K_M\sqrt{T}\|\mathbf{u}_0\|_2. \tag{4.3}$$

Similarly, for almost all $t \in (0,T)$,

$$\begin{aligned} &\int_{\mathbb{R}^3_+} [\mathbf{u}\otimes\mathbf{u}](t,\cdot) : D(\widetilde{\mathbf{w}}_\varepsilon)(t,\cdot) \ \mathrm{d}\mathbf{x} \\ &\qquad = \zeta_\varepsilon(t) \int_{\mathbb{R}^3_+} [\rho_{\mathbf{G}_h}\mathbf{u}\otimes\mathbf{u}](t, x_1+G_1, x_2+G_2, x_3) : D(\mathbf{w}_{h(t)})(\mathbf{x}) \ \mathrm{d}\mathbf{x}. \end{aligned}$$

Consequently, applying Lemma 3.1 together with Lemmas 3.6 and 3.3, we obtain

$$\left| \int_{\mathbb{R}^3_+} [\mathbf{u}\otimes\mathbf{u}](t,\cdot) : D(\widetilde{\mathbf{w}}_\varepsilon)(t,\cdot) \ \mathrm{d}\mathbf{x} \right| \leqslant K_m\|\nabla\mathbf{u}\|^2_{L^2(\mathbb{R}^3_+)}.$$

Thus,

$$|I_2| \leqslant K_m\|\mathbf{u}_0\|^2_{L^2(\mathbb{R}^3_+)}. \tag{4.4}$$

Finally, computing $\partial_t\widetilde{\mathbf{w}}_\varepsilon$ we have, after our change of variable,

$$I_1 = I_1^\chi + I_1^w,$$

where

$$I_1^\chi := \int_0^T \int_{\mathbb{R}^3_+} [\rho_{\mathbf{G}_h}\mathbf{u}](t, x_1 + G_1, x_2 + G_2, x_3)] \cdot \zeta_\varepsilon'(t)\mathbf{w}_{h(t)}(\mathbf{x}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}t$$

and

$$I_1^w := \int_0^T \zeta_\varepsilon(t) \int_{\mathbb{R}^3_+} [\rho_{\mathbf{G}_h}\mathbf{u}](x_1 + G_1, x_2 + G_2, x_3) \cdot \left[\dot{h}\partial_h \mathbf{w}_h - V_1 \partial_{x_1}\mathbf{w}_h - V_2\partial_{x_2}\mathbf{w}_h\right](\mathbf{x}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}t.$$

Applying the Cauchy–Schwarz inequality and Lemma 3.4 on $\mathbf{w}_h$, we deduce that

$$|I_1^\chi| \leqslant C \left[\int_{t_0-\varepsilon}^{t_0} |\zeta_\varepsilon'(t)| \ \|\mathbf{u}(t,\cdot)\|_{L^2(\mathbb{R}^3_+)} \, \mathrm{d}t + \int_{t_1}^{t_1+\varepsilon} |\zeta_\varepsilon'(t)| \ \|\mathbf{u}(t,\cdot)\|_{L^2(\mathbb{R}^3_+)} \, \mathrm{d}t\right],$$

and therefore, using the uniform L^2-bound on $\mathbf{u}$, we obtain

$$|I_1^\chi| \leqslant K\|\mathbf{u}^0\|_{L^2(\mathbb{R}^3_+)}. \tag{4.5}$$

Finally, applying (3.4) in Lemma 3.1 together with (3.9) in Lemma 3.5, we conclude that

$$|I_1^w| \leqslant K_m \int_0^T [|\dot{h}|^2 + |\mathbf{V}_1|^2 + |\mathbf{V}_2|^2]^{\frac{1}{2}} \|\nabla \mathbf{u}\|_{L^2(\mathbb{R}^3_+)} \ \mathrm{d}t,$$

so that, with standard energy estimate,

$$|I_1^w| \leqslant K_m\sqrt{T}\|\mathbf{u}_0\|^2_{L^2(\mathbb{R}^3_+)}. \tag{4.6}$$

Gathering (4.2), (4.4), (4.5), (4.6) with (4.1) yields

$$\left|\int_0^T \zeta_\varepsilon(t)\dot{h}(t)\widetilde{n}_3(h(t)) \ \mathrm{d}t\right| \leqslant K_m(1+\sqrt{T})\left\{\|\mathbf{u}_0\|_{L^2(\mathbb{R}^3_+)} + \|\mathbf{u}_0\|^2_{L^2(\mathbb{R}^3_+)}\right\},$$

where we emphasize that K_m depend only on our choice for the approximation of the solution to the Stokes problem, but not on ε. Thus, letting ε go to 0, as h and $\widetilde{n}_3$ are continuous functions, we obtain

$$|N_3(h(t_1)) - N_3(h(t_0))| \leqslant K_m(1+\sqrt{T})\left\{\|\mathbf{u}_0\|_{L^2(\mathbb{R}^3_+)} + \|\mathbf{u}_0\|^2_{L^2(\mathbb{R}^3_+)}\right\},$$

where N_3 is a primitive of $\widetilde{n}_3$ which vanishes in $h = 1$, for example. Applying Lemma 3.9, we have $\widetilde{n}_3(h) \geqslant c/h$ when $0 < h < h_m$ for some $c > 0$ and $h_m > 0$, and we finally deduce

$$|\ln(h(t)/h(t_0))| \leqslant K_m(1+\sqrt{T})\left\{\|\mathbf{u}_0\|_{L^2(\mathbb{R}^3_+)} + \|\mathbf{u}_0\|^2_{L^2(\mathbb{R}^3_+)}\right\}.$$

Because h is continuous, letting t_0 tend to 0, we finally obtain

$$h(t) \geqslant h(0)\exp\left[-K_m(1+\sqrt{T})\left\{\|\mathbf{u}_0\|_{L^2(\mathbb{R}^3_+)} + \|\mathbf{u}_0\|^2_{L^2(\mathbb{R}^3_+)}\right\}\right].$$

This is the expected result.

Appendix. Detailed description of ϕ_h^d. In this section we estimate the size of ϕ_h^d and its derivatives. We recall that

$$\phi_h^d(r,\theta,z) = r\chi_o(z/\delta_h(r)), \quad \text{with} \quad \chi_o(s) = \frac{s^2(3-2s)}{2}.$$

In order to compare functions in the following, we introduce the following conventions. Given families $(f_h : \Omega_{h,1/4} \to \mathbb{R})_{h\in(0,1)}$ and $(g_h : \Omega_{h,1/4} \to \mathbb{R})_{h\in(0,1)}$ we denote $f_h \prec g_h$ if there exists an absolute constant such that

$$|f_h(\mathbf{x})| \leqslant Cg_h(\mathbf{x}) \quad \forall\, \mathbf{x} \in \Omega_{h,1/4} \text{ and } h < 1.$$

Given nonnegative functions $f : (0,1) \to \mathbb{R}^+$ and $g : (0,1) \to \mathbb{R}^+$, we also denote

$$f(s) \sim g(s) \quad \forall s \in (0,1)$$

if there exist two positive constants c and C such that

$$cf(s) \leqslant g(s) \leqslant Cf(s) \quad \forall s \in (0,1).$$

First, we compute typical $L^1(0,1/4)$-sizes of functions $r \mapsto r^\alpha/(\delta_h(r))^\beta$.

LEMMA A.1. *Given $(\alpha,\beta) \in (\mathbb{R}_+)^2$, we have the following estimations for all $h \in (0,1)$:*

$$\int_0^{\frac14} \frac{r^\alpha}{\delta_h(r)^\beta}\,\mathrm{d}r \sim \begin{cases} 1 & \text{if } \alpha > 2\beta - 1, \\ h^{\frac{(\alpha+1)-2\beta}{2}} & \text{if } \alpha < 2\beta - 1. \end{cases}$$

Proof. As in [9], we remark that, for all $h \in (0,1)$, we have

$$h + \frac{s^2}{2} \leqslant \delta_h(s) \leqslant h + s^2 \quad \forall s \in (0,1/4).$$

Consequently, we can replace $\delta_h(r)$ by $h + \gamma r^2$ with some generic parameter $\gamma > 0$, and we are bound to calculate

$$I_{\alpha,\beta} := \int_0^{1/4} \frac{r^\alpha}{(h+\gamma r^2)^\beta}\,\mathrm{d}r,$$

in which we set $r = \sqrt{h}s$. This yields

$$I_{\alpha,\beta} := h^{\frac{(\alpha+1)-2\beta}{2}} \int_0^{\frac{1}{4\sqrt{h}}} \frac{s^\alpha}{(1+\gamma s^2)^\beta}\,\mathrm{d}s.$$

Consequently, if $\alpha > 2\beta - 1$, the integral behaves like $Ch^{-\frac{(\alpha+1)-2\beta}{2}}$, and we obtain the first case, while if $\alpha < 2\beta - 1$, the integral goes to a finite positive value as $h \to \infty$, and we obtain the second case. ∎

We now compare $\lambda(r,z,h) = z/\delta_h(r)$ to member functions $(r,\theta,z) \mapsto r^\alpha/(\delta_h(r))^\beta$ in $\Omega_{h,1/4}$.

LEMMA A.2. *We have the following sizes:*

$$\begin{array}{llll} \lambda \prec 1, & \lambda_r \prec r/\delta_h, & \lambda_z \prec 1/\delta_h, & \lambda_h \prec 1/\delta_h, \\ \lambda_{rh} \prec r/\delta_h^2, & \lambda_{zh} \prec 1/\delta_h^2, & \lambda_{rr} \prec 1/\delta_h, & \lambda_{rz} \prec r/\delta_h^2, \\ \lambda_{rrz} \prec 1/\delta_h^2, & & \lambda_{rrr} \prec r/\delta_h^2. & \end{array}$$

Proof. The reader may rapidly check that all the derivatives of δ_h are independent of h and that all the odd ones are bounded by r over $(0, 1/4)$. Then in $\Omega_{h,1/4}$ we have $z \in (0, \delta_h(r))$, and consequently $\lambda \prec 1$. Then

$$\lambda_r = -\frac{\lambda \delta_h'}{\delta_h}, \quad \lambda_z = \frac{1}{\delta_h}, \quad \lambda_h = -\frac{\lambda}{\delta_h}.$$

As δ' is bounded by r necessarily independent of h, we get $\lambda_r \prec r/\delta_h$ and $\lambda_z \prec 1/\delta_h$, $\lambda_h \prec 1/\delta_h$. To the next order, we get

$$\lambda_{rz} = -\frac{\delta_h'}{\delta_h^2}, \quad \lambda_{rr} = \lambda \left(2\frac{(\delta_h')^2}{\delta_h^2} - \frac{\delta_h''}{\delta_h} \right), \quad \lambda_{rh} = -2\frac{\lambda \delta_h'}{\delta_h^2}, \quad \lambda_{zh} = -\frac{1}{\delta_h^2}.$$

As δ'' is bounded independently of h and $r^2 \leqslant h + r^2$, we obtain

$$\lambda_{rz} \prec \frac{r}{\delta_h^2}, \quad \lambda_{rr} \prec \frac{1}{\delta_h}, \quad \lambda_{rh} \prec \frac{r}{\delta_h}, \quad \lambda_{zh} \prec \frac{1}{\delta_h^2}.$$

Finally,

$$\lambda_{rrz} = \frac{1}{\delta_h} \left(2\frac{(\delta_h')^2}{\delta_h^2} - \frac{\delta_h''}{\delta_h} \right), \quad \lambda_{rrr} = \lambda \left(6\frac{\delta_h'' \delta_h'}{\delta_h^2} - 6\frac{(\delta_h')^3}{\delta_h^3} - \frac{\delta_h^{(3)}}{\delta_h} \right).$$

And, as $\delta_h^{(3)}$ is bounded by r and $r^2 \leqslant \delta_h$,

$$\lambda_{rrz} \prec \frac{1}{\delta_h^2}, \quad \lambda_{rrr} \prec \frac{r}{\delta_h^2}. \qquad \square$$

Then we obtain the following lemma.

LEMMA A.3. *We have the following sizes:*

$$\begin{array}{llll}
\phi_h^d \prec r, & \partial_r \phi_h^d \prec 1, & \partial_z \phi_h^d \prec r/\delta_h, & \partial_r \phi_h^d / r - \phi_h^d / r^2 \prec r/\delta_h, \\
\partial_h \phi_h^d \prec r/\delta_h, & \partial_{rh} \phi_h^d \prec 1/\delta_h, & \partial_{zh} \phi_h^d \prec r/\delta_h^2, & \partial_{rz} \phi_h^d / r - \partial_z \phi_h^d / r^2 \prec r/\delta_h^2, \\
\partial_{rr} \phi_h^d \prec r/\delta_h, & \partial_{rz} \phi_h^d \prec 1/\delta_h, & \partial_{zz} \phi_h^d \prec r/\delta_h^2, & \\
\partial_{rrr} \phi_h^d \prec 1/\delta_h, & \partial_{rzz} \phi_h^d \prec 1/\delta_h^2, & \partial_{rrz} \phi_h^d \prec r/\delta_h^2, & \partial_{zzz} \phi_h^d \prec r/\delta_h^3.
\end{array}$$

Proof. By definition, we have $\phi_h^d(r, z) = r\chi_o(\lambda)$, where χ_o is a fixed polynomial and, according to the previous lemma, λ is bounded. Consequently, we obtain $\phi_h^d \prec r$.

In the following, we shall drop all dependencies of χ_o in λ. Due to the same argument as for χ_o, all those quantities depending only on χ_o are bounded independently of (h, r, z) in $\Omega_{h,1/4}$.

So, we compute

$$\partial_r \phi_h^d = \chi_o + r\lambda_r \chi_o', \quad \partial_z \phi_h^d = r\lambda_z \chi_o', \quad \partial_h \phi_h^d = r\lambda_h \chi_o'.$$

Applying the previous lemma and $r^2 \leqslant \delta_h(r)$, we get

$$\partial_r \phi_h^d \prec 1, \quad \partial_z \phi_h^d \prec r/\delta_h, \quad \partial_h \phi_h^d \prec r/\delta_h, \quad \partial_r \phi_h^d / r - \phi_h^d / r^2 = \lambda_r \chi_o' \prec r/\delta_h.$$

To the next order, we obtain, as λ_z is independent of z,

$$\partial_{rr} \phi_h^d = (2\lambda_r + r\lambda_{rr})\chi_o' + r(\lambda_r)^2 \chi_o'',$$

$$\partial_{rz} \phi_h^d = (\lambda_z + r\lambda_{rz})\chi_o' + r\lambda_r \lambda_z \chi_o'' \quad \partial_{zz} \phi_h^d = r(\lambda_z)^2 \chi_o'',$$

$$\partial_{zh} \phi_h^d = r\lambda_z \lambda_h \chi_o'' + r\lambda_{hz} \chi_o', \quad \text{and} \quad \partial_{rh} \phi_h^d = \lambda_h \chi_o' + r\lambda_{rh} \chi_o' + r\lambda_r \lambda_h \chi_o''.$$

As above,

$$\partial_{rr}\phi_h^d \prec r/\delta_h, \quad \partial_{rz}\phi_h^d \prec 1/\delta_h, \quad \partial_{zz}\phi_h^d \prec r/\delta_h^2, \quad \partial_{rh}\phi_h^d \prec 1/\delta_h, \quad \partial_{zh}\phi_h^d \prec r/\delta_h^2$$

and

$$\frac{\partial_{rz}\phi_h^d}{r} - \frac{\partial_z\phi_h^d}{r^2} = \lambda_{rz}\chi_o' + \lambda_r\lambda_z\chi_o'' \prec r/\delta_h^2.$$

To the next order, we obtain

$$\partial_{rrr}\phi_h^d = (3\lambda_{rr} + r\lambda_{rrr})\chi_o' + (3\lambda_r^2 + 3r\lambda_{rr}\lambda_r))\chi_o'' + r(\lambda_r)^3\chi_o^{(3)},$$

and thus $\partial_{rrr}\phi_h^d \prec 1/\delta_h$; and

$$\partial_{rzz}\phi_h^d = (\lambda_z^2 + 2r\lambda_{rz}\lambda_z)\chi_o'' + r(\lambda_z)^2\lambda_r\chi_o^{(3)},$$

so $\partial_{rzz}\phi_h^d \prec 1/\delta_h^2$; and

$$\partial_{rrz}\phi_h^d = (2\lambda_{rz} + r\lambda_{rrz})\chi_o' + (2\lambda_r\lambda_z + r(\lambda_{rr}\lambda_z + 2\lambda_{rz}\lambda_r)\chi_o'' + r(\lambda_r)^2\lambda_z\chi_o^{(3)},$$

so $\partial_{rrz}\phi_h^d \prec r/\delta_h^2$. Finally, $\partial_{zzz}\phi_h^d = r(\lambda_z)^3\chi_o^{(3)}$, so that $\partial_{zzz}\phi_h^d \prec r/{\delta_h}^3$. □

REFERENCES

[1] M. D. A. Cooley and M. E. O'Neill, *On the slow motion generated in a viscous fluid by the approach of a sphere to a plane wall or stationary sphere*, Mathematika, 16 (1969), pp. 37–49.

[2] P. Cumsille and T. Takahashi, *Wellposedness for the system modelling the motion of a rigid body of arbitrary form in an incompressible viscous fluid*, Czechoslovak Math. J., 58 (2008), pp. 961–992.

[3] P. Cumsille and M. Tucsnak, *Wellposedness for the Navier-Stokes flow in the exterior of a rotating obstacle*, Math. Methods Appl. Sci., 29 (2006), pp. 595–623.

[4] E. Feireisl, *On the motion of rigid bodies in a viscous incompressible fluid*, J. Evol. Equ., 3 (2003), pp. 419–441.

[5] H. Fujita and N. Sauer, *On existence of weak solutions of the Navier-Stokes equations in regions with moving boundaries*, J. Fac. Sci. Univ. Tokyo Sect. I, 17 (1970), pp. 403–420.

[6] G. P. Galdi, *An introduction to the mathematical theory of the Navier-Stokes equations, Vol.* I, *Linearized Steady Problems*, Springer Tracts Nat. Philos. 38, Springer-Verlag, New York, 1994.

[7] M. D. Gunzburger, H.-C. Lee, and G. A. Seregin, *Global existence of weak solutions for viscous incompressible flows around a moving rigid body in three dimensions*, J. Math. Fluid Mech., 2 (2000), pp. 219–266.

[8] T. I. Hesla, *Collisions of Smooth Bodies in Viscous Fluids: A Mathematical Investigation*, Ph.D. thesis, University of Minnesota, revised version, 2005.

[9] M. Hillairet, *Lack of collision between solid bodies in a 2D incompressible viscous flow*, Comm. Partial Differential Equations, 32 (2007), pp. 1345–1371.

[10] A. Inoue and M. Wakimoto, *On existence of solutions of the Navier-Stokes equation in a time dependent domain*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 24 (1977), pp. 303–319.

[11] V. N. Starovoĭtov, *Nonuniqueness of a solution to the problem on motion of a rigid body in a viscous incompressible fluid*, J. Math. Sci., 130 (2005), pp. 4893–4898.

[12] V. N. Starovoitov, *Behavior of a rigid body in an incompressible viscous fluid near a boundary*, in Free Boundary Problems (Trento, 2002), Internat. Ser. Numer. Math. 147, Birkhäuser, Basel, 2004, pp. 313–327.

[13] T. Takahashi, *Analysis of strong solutions for the equations modeling the motion of a rigid-fluid system in a bounded domain*, Adv. Differential Equations, 8 (2003), pp. 1499–1532.

[14] R. Temam, *Problèmes mathématiques en plasticité*, Gauthier-Villars, Montrouge, 1983.

[15] J. L. Vázquez and E. Zuazua, *Lack of collision in a simplified 1D model for fluid-solid interaction*, Math. Models Methods Appl. Sci., 16 (2006), pp. 637–678.

SIAM J. MATH. ANAL.
Vol. 40, No. 6, pp. 2478–2495

ASYMPTOTIC STABILITY OF PERIODIC SOLUTIONS FOR NONSMOOTH DIFFERENTIAL EQUATIONS WITH APPLICATION TO THE NONSMOOTH VAN DER POL OSCILLATOR*

ADRIANA BUICĂ†, JAUME LLIBRE‡, AND OLEG MAKARENKOV§

Abstract. In this paper we study the existence, uniqueness, and asymptotic stability of the periodic solutions of the Lipschitz system $\dot{x} = \varepsilon g(t, x, \varepsilon)$, where $\varepsilon > 0$ is small. Our results extend the classical second Bogoliubov theorem for the existence of stable periodic solutions to nonsmooth differential systems. As an application we prove the existence of asymptotically stable 2π-periodic solutions of the nonsmooth van der Pol oscillator $\ddot{u} + \varepsilon\,(|u| - 1)\,\dot{u} + (1 + a\varepsilon)u = \varepsilon\lambda \sin t$. Moreover, we construct the so-called resonance curves that describe the dependence of the amplitude of these solutions as a function of the parameters a and λ. Finally we compare such curves with the resonance curves of the classical van der Pol oscillator $\ddot{u} + \varepsilon\left(u^2 - 1\right)\dot{u} + (1 + a\varepsilon)u = \varepsilon\lambda \sin t$.

Key words. periodic solution, asymptotic stability, averaging theory, nonsmooth differential system, nonsmooth van der Pol oscillator

AMS subject classifications. 34C29, 34C25, 47H11

DOI. 10.1137/070701091

1. Introduction.

In this paper we study the existence, uniqueness, and asymptotic stability of the T-periodic solutions of the system

$$\dot{x} = \varepsilon g(t, x, \varepsilon), \tag{1.1}$$

where $\varepsilon > 0$ is a small parameter, and the function $g \in C^0(\mathbb{R} \times \mathbb{R}^k \times [0, 1], \mathbb{R}^k)$ is T-periodic in the first variable and locally Lipschitz with respect to the second. For this class of differential systems, the study of the T-periodic solutions can be made using the averaging function

$$g_0(v) = \int_0^T g(\tau, v, 0)d\tau \tag{1.2}$$

and looking for the periodic solutions that starts near some $v_0 \in g_0^{-1}(0)$.

In the case that g is of class C^1, we recall the stable periodic case of the second Bogoliubov's theorem [6, Chap. 1, section 5, Theorem II] which states *If* $\det (g_0)'(v_0) \neq 0$ *and* $\varepsilon > 0$ *is sufficiently small, then system* (1.1) *has a unique* T*-periodic solution in a neighborhood of* v_0*. Moreover, if all the eigenvalues of the Jacobian matrix*

*Received by the editors August 24, 2007; accepted for publication (in revised form) November 8, 2008; published electronically February 25, 2009.
http://www.siam.org/journals/sima/40-6/70109.html

†Department of Applied Mathematics, Babeş-Bolyai University, Cluj-Napoca, Romania (abuica@math.ubbcluj.ro).

‡Departament de Matemàtiques, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain (jllibre@mat.uab.cat). This author was partially supported by MEC/FEDER grant MTM2008-03437 and by CICYT grant 2005SGR 00550.

§Research Institute of Mathematics, Voronezh State University, Voronezh, Russia (omakarenkov@math.vsu.ru). This author was partially supported by grant BF6M10 of the Russian Federation Ministry of Education and U.S. CRDF (BRHE), by RFBR grant 06-01-72552, by the President of the Russian Federation Young Researcher grant MK-1620.2008.1, and by Marie Curie IIF GA-2008-221331.

$(g_0)'(v_0)$ *have negative real part, then this periodic solution is asymptotically stable.* This theorem has a long history and includes results by Fatou [15], Mandelstam and Papaleksi [31], and Krylov and Bogoliubov [25, section 2].

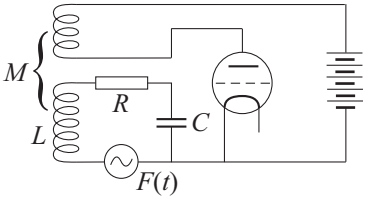


FIG. 1.1. *Circuit scheme for the classical triode oscillator (see* [3, *Chap.* VIII, section 2, *Figure* 348], [30, *Chap.* I, section 5, *Figure* 1], *and* [36, section 3.1.7, *Figures* 3–5]*).*

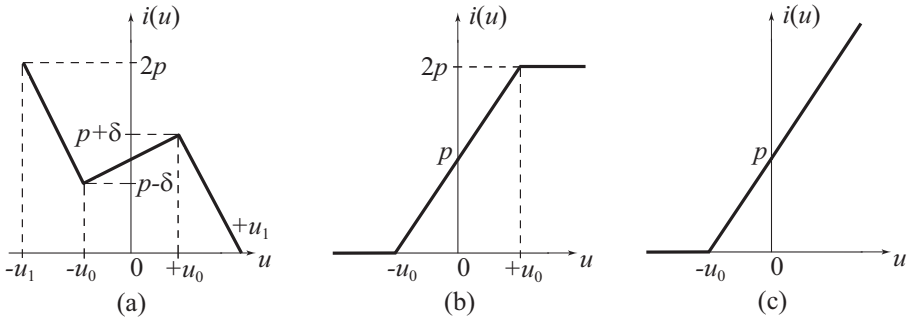


FIG. 1.2. *Characteristics of the triode of the circuit of Figure* 1.1. (a) *Triode in a harsh regime (see* [3, *Chap.* IV, section 7, *Figure* 212b], [30, *Chap.* I, section 5, *comments for eqs.* 5.3–5.4]*);* (b) *triode with saturation (see* [3, *Chap.* VIII, section 3, *Figure* 364]*);* (c) *triode without saturation (see* [3, *Chap.* IX, section 7, *Figure* 482]*).*

The Bogoliubov result provided a theoretical justification for some resonance phenomena which appear in many real physical systems. One of the most significant examples is the classical triode oscillator whose scheme is drawn in Figure 1.1 and whose current u is described by the second order differential equation

$$\ddot{u} + \frac{1}{LC}\left(RC - Mi'(u)\right)\dot{u} + \omega^2 u = \frac{1}{LC}F(t), \tag{1.3}$$

where $R = \varepsilon R_0$, $M = \varepsilon M_0$, $\omega^2 = 1 + \varepsilon b$, $F(t) = \varepsilon\lambda \sin t$, $\varepsilon > 0$, is assumed to be small and the triode characteristic $i(u)$ is drawn in Figure 1.2(a). The analysis of the diagram of bifurcation of the periodic solutions in this system is performed in almost every book on nonlinear oscillations (see Andronov, Witt, and Khaikin [3, Chap. VIII, section 2], Malkin [30, Chap. I, section 5], and Nayfeh and Mook [36, section 3.1.7]) but with the smooth approximation $i(u) = i_{(a)}(u) = S_0 + S_1 u - \frac{1}{3}S_3 u^3$ (leading to the classical van der Pol equation). Therefore it is natural to look for a technique that permits one to avoid this smooth approximation and allows one to work with the original shape of the triode characteristic drawn in Figure 1.2(a).

Though the unforced equation (1.3) (i.e., for $F = 0$) with i described by Figure 1.2(b) and Figure 1.2(c) is well studied (see [3, Chap. VIII, section 3 and Chap. IX, section 7]), the question about resonances in these equations when $F \neq 0$ (e.g., $F(t) = \varepsilon\lambda \sin t$) is still partially open. In this direction Levinson [29] uncovered a

family of solutions of (1.3) of remarkable singular structure and Levi [28] completed the study of the limit behavior of all solutions. The present paper complements these results by describing the location of asymptotically stable periodic solutions of (1.3).

Studying (1.3) with the triode characteristic given by Figure 1.2(a), 1.2(b), or 1.2(c) we finally note that there exists a change of variables (see, for example, how Levinson changed equation 2.0 in [29]) that allows one to rewrite (1.3) into the form (1.1) with some function g that is not C^1 but is Lipschitz with respect to the second variable. Therefore the goal of this work is to generalize the results on the existence of a stable periodic solution of the second Bogoliubov theorem to the case that the function g of (1.1) is only Lipschitz.

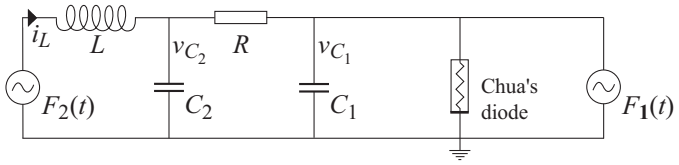


FIG. 1.3. *Forced Chua's circuit (see* [5, 11, 20, 35, 40]*).*

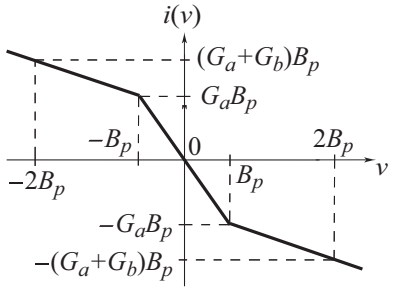


FIG. 1.4. *Nonlinear characteristic of the Chua's diode of the circuit drawn at Figure* 1.3 *given by* $i(v) = G_b v + (1/2)(G_a - G_b)\,(|v + B_p| - |v - B_p|)$, *where* $G_a, G_b, B_p \in \mathbb{R}$ *are some constants depending on the properties of the Chua's diode (see* [10]*).*

Another motivation of this paper comes from the forced Chua's circuit (see Figure 1.3) studied in a large number of papers in the modern electrical engineering. This circuit is described by the three-dimensional system

$$
\begin{aligned}
C_1 \frac{dv_{C_1}}{dt} &= \frac{v_{C_2} - v_{C_1}}{R} - i(v_{C_1}) + F_1(t),\\
C_2 \frac{dv_{C_2}}{dt} &= \frac{v_{C_1} - v_{C_2}}{R} + i_L,\\
L \frac{di_L}{dt} &= -v_{C_2} + F_2(t, v_{C_2}),
\end{aligned}
\tag{1.4}
$$

where $i(v)$ (the characteristic of the Chua's diode) is a piecewise linear function, as it is represented in Figure 1.4. The recent literature provides insight into the numerical simulations of (1.4) (see [40, 20], where $F_1 \neq 0$ and $F_2 \neq 0$, [5, 35], where $F_1 = 0$ and F_2 is periodic, or [11] where both F_1 and F_2 are periodic). Generalization of the Bogoliubov result for (1.1) with Lipschitz right-hand part will allow for the first

time the theoretical detection of asymptotically stable periodic solutions of (1.4) in the case that C_1 is large enough. Eventually this theoretical analysis may provide new interesting parameters of the forced Chua's circuit for doing additional numerical experiments.

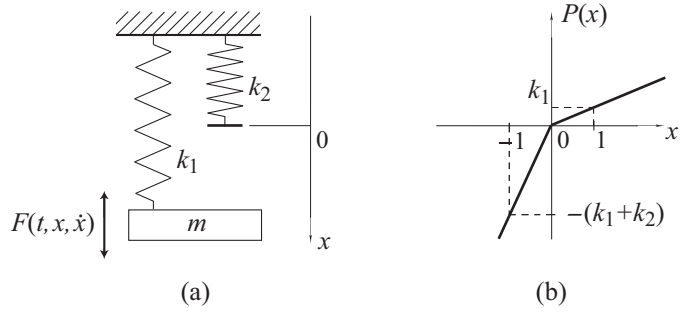


FIG. 1.5. *A prototypic device presented in* (a) *where a driven mass is attached to a immovable beam via a spring with piecewise linear stiffness like in* (b)*; see, e.g.,* [7], [24, *Chap.* I, *p.* 16 *and Chap.* IV, *p.* 100]*, and* [38].

On the other hand, part of the interest in generalizing the Bogoliubov result comes from mechanics, where differential systems with piecewise linear stiffness describe various oscillating processes. One of these systems is exhibited by the device drawn in Figure 1.5(a), where a forced mass is attached to a spring whose stiffness changes from k_1 to $k_1 + k_2$ when the mass coordinate crosses 0 in the negative direction. This device is governed by the second order differential equation

$$m\ddot{x} + P(x) = F(t, x, \dot{x}), \tag{1.5}$$

where the piecewise linear stiffness P is drawn in Figure 1.5(b). Depending on the particular configuration of the device of Figure 1.5(a), different expressions for F in (1.5) must be considered. Thus we have that $F(t, x, \dot{x}) = -f(x)\dot{x} + M\cos\omega t$ with piecewise constant f for a shock-absorber and jigging conveyor (see [24, Chap. I, p. 16 and Chap. IV, p. 100]), where the original Bogoliubov result is employed without justification). The function F takes the simpler form $F(t, x, \dot{x}) = -c\dot{x} + MQ(t)$ for an impact resonator, and $F(t, x, \dot{x}) = -c\dot{x} + M\sin\omega t$ for a cracked-body model (see [38, 7], where only numerical experiments are performed). In each of these situations (1.5) can be rewritten in the form (1.1) with g Lipschitz, provided that the constant k_2 and the amplitude of the force F are sufficiently small. Therefore the extension of the Bogoliubov result to the nonsmooth case that we shall do will allow one to justify the resonances that appeared in all these results. We note that the recent report by Los Alamos National Laboratory [13] describes the increasing interest in a specific form of the model of Figure 1.5(a) called the cracked-body model and, particularly, in the suspension bridge models. Consequently the results of this paper can be applied to such models.

A first model of a one-dimensional suspended bridge is drawn in Figure 1.6(a). It is represented (see [16, 27]) by the beam bending under its own weight and being supported by cables whose restoring force due to elasticity is proportional to u^+ (see Figure 1.6(b)), where $u = u(t, x)$ is the displacement of a point at a distance x from one end of the bridge at time t and u is measured in the downward direction. Looking for u of the form $u(t, x) = z(t)\sin(\pi x/L)$ and considering $F(x, t) = h(t)\sin(\pi x/L)$,

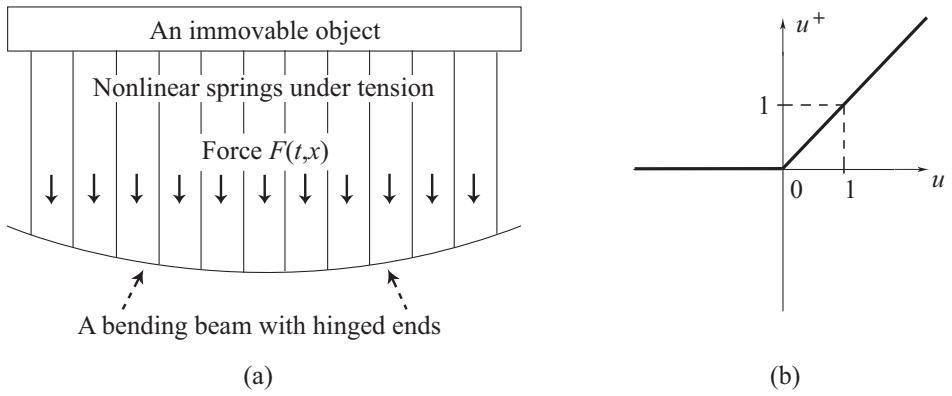


FIG. 1.6. (a) *The first idealization of the suspension bridge: the beam bending under its own weight is supported by the nonlinear cables (see* [27, *Figure* 2]*);* (b) *characteristic of stiffness of nonlinear springs.*

we arrive (see [16]) at the following particular case of differential equation (1.5):

$$m\ddot{z} + \delta\dot{z} + c(\pi/L)^4 z + dz^+ = mg + h(t), \tag{1.6}$$

where the constant $m > 0$ is the mass per unit of length, $\delta > 0$ is a small viscous damping coefficient, $c > 0$ measures the flexibility or stiffness of the bridge, $L > 0$ is the length of the bridge, $d > 0$ represents the stiffness of nonlinear springs, and h is a continuous T-periodic force modelling wind, marching troops, or cattle (see [19] for details). Considering $c > 0$ and $d > 0$ fixed and assuming that either $c > 0$ and $h(t)$ are sufficiently small, or that $c > 0$ is fixed and $h(t)$ is sufficiently large, or that $c > 0$ is sufficiently small and $h(t)$ fixed, Glover, Lazer, and McKenna [16], Lazer and McKenna [27], and Fabry [14] proved various theorems on the location of asymptotically stable T-periodic solutions in (1.6). The question *What happens with these solutions when* $d > 0$*,* $\delta > 0$*, and* $h(t)$ *are all sufficiently small?* was open and can be solved using the generalization of the Bogoliubov result that we provide. Lazer and McKenna proved in [26] that the Poincaré map for (1.6) is differentiable, but we note that this is not sufficient for applying the original Bogoliubov result.

We end the list of possible applications noting that system (1.4) describing the Chua's circuit (Figure 1.3) appeared recently for studying the so-called negative slope mechanical systems (see Awrejcewicz [4, section 8.2.2]). So our results can also be applied to these mechanical systems.

These applications require generalizations of the second Bogoliubov theorem for Lipschitz right-hand parts. To the best of our knowledge Mitropol'skii was the first to consider such a kind of generalization. Assuming that g is Lipschitz, $g_0 \in C^3(\mathbb{R}^k, \mathbb{R}^k)$, and all the eigenvalues of the matrix $(g_0)'(v_0)$ have negative real part, Mitropol'skii [34] developed the Bogoliubov result proving the existence and uniqueness of a T-periodic solution of system (1.1) in a neighborhood of v_0. There was great progress weakening the assumptions of the existence result (see Samoilenko [39] and Mawhin [32]), but this progress did not take place in the case of the uniqueness. Moreover, the asymptotic stability of the T-periodic solution remained unstudied in the case of Lipschitz systems for a long time. It has been done recently by Buică and Daniilidis in [8] for Lipschitz systems (1.1) assuming that the function $v \mapsto g(t, v, 0)$ is differentiable at v_0 for almost any $t \in [0, T]$ and that the eigenvectors of the matrix $(g_0)'(v_0)$ are orthogonal.

In section 2, assuming that g is piecewise differentiable in the second variable, we prove in Theorem 2.5 that the stable periodic solution of the Bogoliubov theorem persists when g is not necessary C^1. Theorem 2.5 follows from this more general Theorem 2.1 whose hypotheses do not use any differentiability—neither of g, nor of g_0. Assuming only continuity for g, we show in Theorem 2.9 the existence of a nonasymptotically stable T-periodic solution of system (1.1) if the Brouwer topological degree of $-g_0$ is negative. In section 3 we illustrate our results constructing the resonance curves of the nonsmooth van der Pol oscillator, also studied in [18], and compare these curves with the resonance curves of the classical van der Pol oscillator, which were constructed by Andronov and Witt [1, 2].[1]

2. Main results. Throughout the paper $\Omega \subset \mathbb{R}^k$ will be an open set. For any $\delta > 0$ we denote $B_\delta(v_0) = \left\{v \in \mathbb{R}^k \ : \ \|v - v_0\| \leq \delta\right\}$. We have the following main result on the existence, uniqueness, and asymptotic stability of T-periodic solutions for system (1.1).

THEOREM 2.1. *Let $g \in C^0(\mathbb{R} \times \Omega \times [0,1], \mathbb{R}^k)$ and $v_0 \in \Omega$. Assume the following four conditions.*

(i) *For some $L > 0$ we have that $\|g(t, v_1, \varepsilon) - g(t, v_2, \varepsilon)\| \leq L \|v_1 - v_2\|$ for any $t \in [0, T]$, $v_1, v_2 \in \Omega$, $\varepsilon \in [0, 1]$.*

(ii) *For any $\gamma > 0$ there exists $\delta > 0$ such that*

$$\left\| \int_0^T g(\tau, v_1 + u(\tau), \varepsilon) d\tau - \int_0^T g(\tau, v_2 + u(\tau), \varepsilon) d\tau \right.$$
$$\left. - \int_0^T g(\tau, v_1, 0) d\tau + \int_0^T g(\tau, v_2, 0) d\tau \right\| \leq \gamma \|v_1 - v_2\|$$

for any $u \in C^0([0,T], \mathbb{R}^k)$, $\|u\| \leq \delta$, $v_1, v_2 \in B_\delta(v_0)$, and $\varepsilon \in [0, \delta]$.

(iii) *Let g_0 be the averaged function given by (1.2) and consider that $g_0(v_0) = 0$.*

(iv) *There exist $q \in [0,1)$, $\alpha, \delta_0 > 0$, and a norm $\|\cdot\|_0$ on $\mathbb{R}^k$ such that $\|v_1 + \alpha g_0(v_1) - v_2 - \alpha g_0(v_2)\|_0 \leq q \|v_1 - v_2\|_0$ for any $v_1, v_2 \in B_{\delta_0}(v_0)$.*

Then there exists $\delta_1 > 0$ such that for every $\varepsilon \in (0, \delta_1]$, system (1.1) has exactly one T-periodic solution x_ε with $x_\varepsilon(0) \in B_{\delta_1}(v_0)$. Moreover, the solution x_ε is asymptotically stable and $x_\varepsilon(0) \to v_0$ as $\varepsilon \to 0$.

When the solution $x(\cdot, v, \varepsilon)$ of system (1.1) with the initial condition $x(0, v, \varepsilon) = v$ is well defined on $[0, T]$ for any $v \in B_{\delta_0}(v_0)$, the map $v \mapsto x(T, v, \varepsilon)$ is also well defined and is called the *Poincaré map* at time T of system (1.1). In order to prove the existence, uniqueness, and stability of the T-periodic solutions of system (1.1) stated in Theorem 2.1, it is sufficient to study the same properties for the fixed points of this Poincaré map.

Before proving Theorem 2.1 we state and prove two lemmas. In order to state the first lemma, we need to introduce the function

$$g_\varepsilon(v) = \int_0^T g(\tau, x(\tau, v, \varepsilon), \varepsilon) d\tau$$

[1] At the final stage of publishing of this paper, we have been informed by Prof. Michael Guevara that the resonance curves mentioned appeared for the first time in [Balth. van der Pol, *Tijdschr. Ned Rad Gen.*, (1924) (in Dutch)], an English translation appeared in [*Phil. Mag.*, vol. 3, 1927, p. 65]. We thank Prof. Guevara for calling our attention to those papers and some other historical background on resonance curves.

and to note that by writing the equivalent integral equation of system (1.1) we have

$$x(T, v, \varepsilon) = v + \varepsilon g_\varepsilon(v).$$

LEMMA 2.2. *Let $g \in C^0(\mathbb{R} \times \Omega \times [0,1], \mathbb{R}^k)$ and $\delta_0 > 0$ be such that $B_{\delta_0}(v_0) \subset \Omega$. If* (i) *is satisfied, then there exist $\delta \in [0, \delta_0]$ and $L_1 > 0$ such that the map $(v, \varepsilon) \mapsto g_\varepsilon(v)$ is well defined and continuous on $B_{\delta_0}(v_0) \times [0, \delta]$ and*

$$\|g_\varepsilon(v_1) - g_\varepsilon(v_2)\| \le L_1 \|v_1 - v_2\| \quad \textit{for any } \varepsilon \in [0, \delta],\ v_1, v_2 \in B_{\delta_0}(v_0).$$

If both (i) *and* (ii) *are satisfied, then for any $\gamma > 0$ there exists $\delta \in [0, \delta_0]$ such that*

$$\|g_\varepsilon(v_1) - g_0(v_1) - g_\varepsilon(v_2) + g_0(v_2)\| \le \gamma \|v_1 - v_2\|$$

for any $v_1, v_2 \in B_\delta(v_0)$ and $\varepsilon \in [0, \delta]$.

Proof. Using the continuity of the solution of a differential system with respect to the initial data and the parameter (see [37, Chap. 4, section 23, statements G and D]), we obtain the existence of $\varepsilon_0 > 0$ such that $x(t, v, \varepsilon) \in \Omega$ for any $t \in [0, T]$, $v \in B_{\delta_0}(v_0)$, and $\varepsilon \in [0, \varepsilon_0]$. Using the Grönwall–Bellman lemma (see [17, Lemma 6.2] or [12, Chap. II, section 11]) from the representation $x(t, v, \varepsilon) = v + \varepsilon \int_0^t g(\tau, x(\tau, v, \varepsilon), \varepsilon) d\tau$ and the property (i), we obtain $\|x(t, v_1, \varepsilon) - x(t, v_2, \varepsilon)\| \le e^{\varepsilon L T} \|v_1 - v_2\|$ for all $t \in [0, T]$, v_1, $v_2 \in B_{\delta_0}(v_0)$, and $\varepsilon \in [0, \varepsilon_0]$. Therefore $y(t, v, \varepsilon) = \int_0^t g(\tau, x(\tau, v, \varepsilon), \varepsilon) d\tau$ satisfies the property

$$\|y(t, v_1, \varepsilon) - y(t, v_2, \varepsilon)\| \le L_1 \|v_1 - v_2\| \tag{2.1}$$

for all $t \in [0, T]$, v_1, $v_2 \in B_{\delta_0}(v_0)$, $\varepsilon \in [0, \varepsilon_0]$, and $L_1 = L T e^{\varepsilon_0 L T}$. Since $g_\varepsilon(v) = y(T, v, \varepsilon)$ the first part of the lemma has been proven.

Taking into account that $x(t, v, \varepsilon) = v + \varepsilon y(t, v, \varepsilon)$, we have

$$y(T, v_1, \varepsilon) - y(T, v_1, 0) - y(T, v_2, \varepsilon) + y(T, v_2, 0) = I_1(v_1, v_2, \varepsilon) + I_2(v_1, v_2, \varepsilon), \tag{2.2}$$

where

$$I_1(v_1, v_2, \varepsilon) = \int_0^T [g(\tau, v_2 + \varepsilon y(\tau, v_1, \varepsilon), \varepsilon) - g(\tau, v_2 + \varepsilon y(\tau, v_2, \varepsilon), \varepsilon)] d\tau,$$

$$I_2(v_1, v_2, \varepsilon) = \int_0^T [(g(\tau, v_1 + \varepsilon y(\tau, v_1, \varepsilon), \varepsilon) - g(\tau, v_2 + \varepsilon y(\tau, v_1, \varepsilon), \varepsilon))] d\tau$$

$$- \int_0^T (g(\tau, v_1, 0) - g(\tau, v_2, 0)) d\tau.$$

Since $(t, v, \varepsilon) \mapsto y(t, v, \varepsilon)$ is bounded on $[0, T] \times B_{\delta_0}(v_0) \times [0, \varepsilon_0]$, we have that $\varepsilon y(t, v, \varepsilon) \to 0$ as $\varepsilon \to 0$ uniformly with respect to $t \in [0, T]$ and $v \in B_{\delta_0}(v_0)$. Decreasing $\varepsilon_0 > 0$, if necessary, we get that $v_2 + \varepsilon y(t, v_1, \varepsilon) \in \Omega$ for any $t \in [0, T]$, $v_1, v_2 \in B_{\delta_0}(v_0)$, $\varepsilon \in [0, \varepsilon_0]$. By assumption (i) and relation (2.1) we obtain that $\|I_1(v_1, v_2, \varepsilon)\| \le T \cdot \varepsilon L L_1 \|v_1 - v_2\|$ for all $\varepsilon \in [0, \varepsilon_0]$, $v_1, v_2 \in B_{\delta_0}(v_0)$.

We fix $\gamma > 0$ and take $\delta > 0$ given by (ii). Without loss of generality we can consider that $\delta \le \min\{\delta_0, \varepsilon_0, \gamma/(2TLL_1)\}$. Therefore assumption (ii) implies that $\|I_2(v_1, v_2, \varepsilon)\| \le (\gamma/2)\|v_1 - v_2\|$ for any $\varepsilon \in [0, \delta]$, $v_1, v_2 \in B_\delta(v_0)$. Substituting the obtained estimations for I_1 and I_2 into (2.2) we have $\|y(T, v_1, \varepsilon) - y(T, v_1, 0) - y(T, v_2, \varepsilon) +$

$y(T, v_2, 0)\| \le (\varepsilon TLL_1 + \gamma/2)\|v_1 - v_2\| \le \gamma\|v_1 - v_2\|$ for any $\varepsilon \in [0, \delta]$, $v_1, v_2 \in B_\delta(v_0)$. Hence the proof is complete. □

LEMMA 2.3. *Let $g_0 : \Omega \to \mathbb{R}^k$, satisfying assumption* (iv) *for some $q \in (0,1)$, $\alpha, \delta_0 > 0$, and a norm $\|\cdot\|_0$ on $\mathbb{R}^k$. Then $\|v_1 + \varepsilon g_0(v_1) - v_2 - \varepsilon g_0(v_2)\|_0 \le (1 - \varepsilon(1-q)/\alpha)\,\|v_1 - v_2\|_0$ for any $v_1, v_2 \in B_{\delta_0}(v_0)$ and any $\varepsilon \in [0, \alpha]$.*

Proof. Indeed, the equality $v + \varepsilon g_0(v) = (1 - \varepsilon/\alpha)v + \varepsilon/\alpha\,(v + \alpha g_0(v))$ implies that the Lipschitz constant of the function $I + \varepsilon g_0$ with respect to the norm $\|\cdot\|_0$ is $(1 - \varepsilon/\alpha) + \varepsilon/\alpha\, q = 1 - \varepsilon(1-q)/\alpha$. □

Proof of Theorem 2.1. By Lemma 2.2 we have that there exists $\delta_1 \in [0, \delta_0]$ such that

$$\|g_\varepsilon(v_1) - g_0(v_1) - g_\varepsilon(v_2) + g_0(v_2)\|_0 \le ((1-q)/(2\alpha))\|v_1 - v_2\|_0 \tag{2.3}$$

for any $\varepsilon \in [0, \delta_1]$, $v_1, v_2 \in B_{\delta_1}(v_0)$. First we prove that there exists $\varepsilon_1 \in [0, \delta_1]$ such that for every $\varepsilon \in [0, \varepsilon_1]$ there exists $v_\varepsilon \in B_{\delta_1}(v_0)$ such that $x(\cdot, v_\varepsilon, \varepsilon)$ is a T-periodic solution of (1.1) by showing that there exists v_ε such that $x(T, v_\varepsilon, \varepsilon) = v_\varepsilon$. Using (iii) and (iv) we have

$$\|v + \alpha g_0(v) - v_0\|_0 \le q\|v - v_0\|_0 \quad \text{for any } v \in B_{\delta_1}(v_0).$$

Therefore we have that the map $I + \alpha g_0$ maps $B_{\delta_1}(v_0)$ into itself. From Lemma 2.2 we have that there exists $\varepsilon_0 > 0$ such that the map $(v, \varepsilon) \mapsto g_\varepsilon(v)$ is well defined and continuous on $B_{\delta_1}(v_0) \times [0, \varepsilon_0]$. We deduce that there exists $\varepsilon_1 > 0$ sufficiently small such that, for every $\varepsilon \in [0, \varepsilon_1]$, the map $I + \alpha g_\varepsilon$ maps $B_{\delta_1}(v_0)$ into itself as well. Therefore, by the Brouwer theorem (see, for example, [23, Theorem 3.1]) we have that $B_{\delta_1}(v_0)$ contains at least one fixed point of the map $I + \alpha g_\varepsilon$ for any $\varepsilon \in [0, \varepsilon_1]$. Denote this fixed point by v_ε. Then we have $g_\varepsilon(v_\varepsilon) = 0$ and $x(T, v_\varepsilon, \varepsilon) = v_\varepsilon$ for any $\varepsilon \in [0, \varepsilon_1]$.

Now we prove that $x(\cdot, v_\varepsilon, \varepsilon)$ is the only T-periodic solution of (1.1) starting near v_0 and that, moreover, it is asymptotically stable. Knowing that $x(T, v, \varepsilon) = v + \varepsilon g_\varepsilon(v)$ we write the following identity:

$$x(T, v, \varepsilon) = v + \varepsilon g_0(v) + \varepsilon\,(g_\varepsilon(v) - g_0(v))\,. \tag{2.4}$$

Using Lemma 2.3 we have from (2.3) and (2.4) that

$$\begin{aligned} \|x(T, v_1, \varepsilon) - x(T, v_2, \varepsilon)\|_0 \le &\quad (1 - \varepsilon(1-q)/\alpha + \varepsilon(1-q)/(2\alpha))\|v_1 - v_2\|_0 \\ = &\quad (1 - \varepsilon(1-q)/(2\alpha))\|v_1 - v_2\|_0 \end{aligned}$$

for all $v_1, v_2 \in B_{\delta_1}(v_0)$ and $\varepsilon \in [0, \delta_1]$. We proved before that there exists $\varepsilon_1 > 0$ such that for every $\varepsilon \in [0, \varepsilon_1]$ there exists $v_\varepsilon \in B_{\delta_1}(v_0)$ with $x(\cdot, v_\varepsilon, \varepsilon)$ a T-periodic solution of (1.1). Since $\varepsilon(1-q)/(2\alpha) > 0$ and $\varepsilon_1 \le \delta_1$, the last inequality implies that for each $\varepsilon \in [0, \delta_1]$, the T-periodic solution $x(\cdot, v_\varepsilon, \varepsilon)$ is the only T-periodic solution of (1.1) in $B_{\delta_1}(v_0)$ and, moreover (see [23, Lemma 9.2]), it is asymptotically stable. □

Remark 2.4. We note that a similar result close to Theorem 2.1 is contained in [8, Theorem 3.5]. But instead of the assumption (iv) with a fixed $\alpha > 0$, it is assumed in [8] with any $\alpha > 0$ sufficiently small. Anyway, notice that Lemma 2.3 implies that it is the same to assume (iv) for only one $\alpha > 0$ or for all $\alpha > 0$ sufficiently small. The advantage of our Theorem 2.1 is that it does not require differentiability of $g(t, \cdot, \varepsilon)$ at any point, while [8] needs it at v_0. See also Remark 2.8.

In general it is not easy to check assumptions (ii) and (iv) in the applications of Theorem 2.1. Thus we also give the following theorem based on Theorem 2.1 which assumes certain type of piecewise differentiability instead of (ii) and deals with properties of the matrix $(g_0)'(v_0)$ instead of the Lipschitz constant of g_0.

For any set $M \subset [0,T]$ measurable in the sense of Lebesgue we denote by $\mathrm{mes}(M)$ the Lebesgue measure of M (see [21, Chap. V, section 3]).

THEOREM 2.5. *Let $g \in C^0(\mathbb{R}\times\Omega\times[0,1],\mathbb{R}^k)$ satisfying* (i). *Let g_0 be the averaged function given by* (1.2) *and consider $v_0 \in \Omega$ such that $g_0(v_0)=0$. Assume that*

(v) *given any $\widetilde{\gamma}>0$ there exist $\widetilde{\delta}>0$ and $M\subset[0,T]$ with $\mathrm{mes}(M)<\widetilde{\gamma}$ such that for every $v\in B_{\widetilde{\delta}}(v_0)$, $t\in[0,T]\setminus M$, and $\varepsilon\in[0,\widetilde{\delta}]$ we have that $g(t,\cdot,\varepsilon)$ is differentiable at v and $\|g'_v(t,v,\varepsilon)-g'_v(t,v_0,0)\|\le\widetilde{\gamma}$.*

Finally assume that

(vi) *g_0 is continuously differentiable in a neighborhood of v_0 and the real parts of all the eigenvalues of $(g_0)'(v_0)$ are negative.*

Then there exists $\delta_1>0$ such that for every $\varepsilon\in(0,\delta_1]$, system (1.1) *has exactly one T-periodic solution x_ε with $x_\varepsilon(0)\in B_{\delta_1}(v_0)$. Moreover, the solution x_ε is asymptotically stable and $x_\varepsilon(0)\to v_0$ as $\varepsilon\to 0$.*

For proving Theorem 2.5 we need two preliminary lemmas.

LEMMA 2.6. *Let $g\in C^0(\mathbb{R}\times\Omega\times[0,1],\mathbb{R}^k)$ satisfying* (i). *If* (v) *holds, then* (ii) *is satisfied.*

Proof. Let $\gamma>0$ be an arbitrary number. We show that (ii) holds with $\delta=\widetilde{\delta}/2$, where $\widetilde{\delta}$ is given by (v) applied with $\widetilde{\gamma}=\min\{\gamma/(4L),\gamma/(4T)\}$. We consider also $M\subset[0,T]$ given by (v) applied with the same value of $\widetilde{\gamma}$.

Let $u\in C^0([0,T],\mathbb{R}^k)$, $\|u\|\le\delta$, and $F(v)=\int_0^T g(\tau,v+u(\tau),\varepsilon)d\tau-\int_0^T g(\tau,v,0)d\tau$. Let $v_1,v_2\in B_\delta(v_0)$ and $\varepsilon\in[0,\delta]$. We have $F(v)=F_1(v)+F_2(v)$, where $F_1(v)=\int_M(g(\tau,v+u(\tau),\varepsilon)-g(\tau,v,0))d\tau$ and $F_2(v)=\int_{[0,T]\setminus M}(g(\tau,v+u(\tau),\varepsilon)-g(\tau,v,0))d\tau$. By (i) we have that $\|F_1(v_1)-F_1(v_2)\|\le 2L\cdot\mathrm{mes}(M)\|v_1-v_2\|<2L\widetilde{\gamma}\|v_1-v_2\|\le(\gamma/2)\|v_1-v_2\|$. On the other hand, using (v), we will prove that a similar relation holds for F_2. In order to do this, we denote $h(\tau,v)=g(\tau,v+u(\tau),\varepsilon)-g(\tau,v,0)$. Notice that for each $\tau\in[0,T]\setminus M$ we can write $h'_v(\tau,v)=(g'_v(\tau,v+u(\tau),\varepsilon)-g'_v(\tau,v_0,0))-(g'_v(\tau,v,0)-g'_v(\tau,v_0,0))$. As a direct consequence of (v) we deduce that $\|h'_v(\tau,v)\|\le 2\widetilde{\gamma}$ for all $v\in B_\delta(v_0)$ and $\tau\in[0,T]\setminus M$. Now applying the mean value theorem for the function $h(\tau,\cdot)$, we have $\|h(\tau,v_1)-h(\tau,v_2)\|\le 2\widetilde{\gamma}\|v_1-v_2\|$ for all $\tau\in[0,T]\setminus M$ and all $v_1,v_2\in B_\delta(v_0)$. Then $\|F_2(v_1)-F_2(v_2)\|\le\int_{[0,T]\setminus M}\|h(\tau,v_1)-h(\tau,v_2)\|d\tau\le 2T\widetilde{\gamma}\|v_1-v_2\|\le(\gamma/2)\|v_1-v_2\|$. Therefore we have proved that $\|F(v_1)-F(v_2)\|\le\gamma\|v_1-v_2\|$, which coincides with (ii). ∎

LEMMA 2.7. *Let $g_0:\Omega\to\mathbb{R}^k$ satisfying assumption* (vi) *for some $v_0\in\Omega$. Then there exist $q\in[0,1)$, α, $\delta_0>0$ and a norm $\|\cdot\|_0$ on $\mathbb{R}^k$ such that* (iv) *is satisfied.*

Proof. If λ is an eigenvalue of $\alpha(g_0)'(v_0)$, then $\lambda+1$ is an eigenvalue of $I+(\alpha g_0)'(v_0)$. Since the eigenvalues of $\alpha(g_0)'(v_0)$ tend to 0 as $\alpha\to 0$ and have negative real parts, then there exists $\alpha\in[0,1)$ such that the absolute values of all the eigenvalues of $I+\alpha(g_0)'(v_0)$ are less than one. Therefore (see [22, p. 90, Lemma 2.2]) there exist $\widetilde{q}\in[0,1)$ and a norm $\|\cdot\|_0$ on $\mathbb{R}^k$ such that $\sup_{\|\xi\|_0\le 1}\|\xi+\alpha(g_0)'(v_0)\xi\|_0\le\widetilde{q}$.

By continuous differentiability of g_0 in a neighborhood of v_0 we have that $\|g_0(v_1)-g_0(v_2)-(g_0)'(v_0)(v_1-v_2)\|/\|v_1-v_2\|\le\|g_0(v_1)-g_0(v_2)-(g_0)'(v_2)(v_1-v_2)\|+\|(g_0)'(v_2)(v_1-v_2)-(g_0)'(v_0)(v_1-v_2)\|/\|v_1-v_2\|\to 0$ as $\max\{\|v_1-v_0\|,\|v_2-v_0\|\}\to 0$. Therefore taking into account that all norms on $\mathbb{R}^k$ are equivalent, there exists $\delta_0>0$ such that $\|g_0(v_1)-g_0(v_2)-(g_0)'(v_0)(v_1-v_2)\|_0\le(1-\widetilde{q})/(2\alpha)\|v_1-v_2\|_0$ for all

$v_1, v_2 \in B_{\delta_0}(v_0)$. Then

$$\begin{aligned}
&\|v_1+\alpha g_0(v_1) - v_2 - \alpha g_0(v_2)\|_0 \\
&\quad\le \alpha\|g_0(v_1) - g_0(v_2) - (g_0)'(v_0)(v_1 - v_2)\|_0 + \|v_1 - v_2 + \alpha(g_0)'(v_0)(v_1 - v_2)\|_0 \\
&\quad\le (1 + \widetilde{q})/2\,\|v_1 - v_2\|_0
\end{aligned}$$

for all $v_1, v_2 \in B_{\delta_0}(v_0)$. □

Proof of Theorem 2.5. Lemmas 2.6 and 2.7 imply that assumptions (ii) and (iv) of Theorem 2.1 are satisfied. Therefore the conclusion of the theorem follows by applying Theorem 2.1. □

It was observed by Mitropol'skii in [34] that in spite of the fact that $g(t, \cdot, \varepsilon)$ in (1.1) is only Lipschitz, sometimes the function g_0 turns out to be differentiable in applications. In particular we will see in section 3 that this is the case for the nonsmooth van der Pol oscillator.

Clearly if $g \in C^1(\mathbb{R} \times \mathbb{R}^k \times [0,1], \mathbb{R}^k)$, then (i) and (v) hold in any open bounded set $\Omega \subset \mathbb{R}^k$. Therefore Theorem 2.5 is a generalization of the stable periodic case of the second Bogoliubov theorem formulated in the introduction.

Remark 2.8. Theorem 2.5 does not require the eigenvectors of $(g_0)'(v_0)$ to be orthogonal as in [8, Theorem 3.6]. Moreover, assumption (H_2) of [8] is more restrictive than (v).

For completeness we also give the following theorem on the existence of nonasymptotically stable T-periodic solutions for (1.1). In the theorem below, $d(F, V)$ denotes the Brouwer topological degree of the vector field $F \in C^0(\mathbb{R}^k, \mathbb{R}^k)$ on the open and bounded set $V \subset \mathbb{R}^k$ (see [23, Chap. 2, section 5.2]).

THEOREM 2.9. *Let $g \in C^0(\mathbb{R} \times \mathbb{R}^k \times [0,1], \mathbb{R}^k)$. Assume that there exists an open bounded set $V \subset \mathbb{R}^k$ such that $g_0(v) \neq 0$ for any $v \in \partial V$ and*

(vii) $d(-g_0, V) < 0$.

Then there exists $\varepsilon_0 > 0$ such that for any $\varepsilon \in (0, \varepsilon_0]$ system (1.1) *has at least one nonasymptotically stable T-periodic solutions x_ε with $x_\varepsilon(0) \in V$.*

Proof. Since $g_0(v) \neq 0$ for any $v \in \partial V$, then from Mawhin's theorem [32] (or [33, section 5]) we have that there exists $\varepsilon_0 > 0$ such that

$$d(-g_0, V) = d(I - x(T, \cdot, \varepsilon), V) \quad \text{for any } \varepsilon \in (0, \varepsilon_0]. \tag{2.5}$$

By [23, Theorem 9.6] for any asymptotically stable T-periodic solution x_ε of (1.1) we have that $d(I - x(T, \cdot, \varepsilon), B_\delta(x_\varepsilon(0))) = 1$ for $\delta > 0$ sufficiently small. Therefore if all the possible T-periodic solutions of (1.1) with $\varepsilon \in (0, \varepsilon_0]$ had been asymptotically stable, then the degree $d(I - x(T, \cdot, \varepsilon), V)$ would have been nonnegative, contradicting (vii) and (2.5). □

Remark 2.10. Assumptions (iii) and (iv) imply that $d(-g_0, V) = 1$ (see [23, Theorem 5.16]).

Finally thinking in terms of the application to the nonsmooth van der Pol oscillator, we formulate the following theorem which combines Mawhin's theorem (see [32] or [33, Theorem 3]) and Theorems 2.5 and 2.9. In this theorem $([g_0]_i)'_{(j)}$ stays for the derivative of the ith component of the function g_0 with respect to the jth variable.

THEOREM 2.11. *Let $g \in C^0(\mathbb{R} \times \Omega \times [0,1], \mathbb{R}^2)$. Let $v_0 \in \Omega$ be such a point that $g_0(v_0) = 0$ and g_0 is continuously differentiable in a neighborhood of v_0.*

(a) *If $\det (g_0)'(v_0) \neq 0$, then there exists $\varepsilon_0 > 0$ such that for any $\varepsilon \in (0, \varepsilon_0]$ system* (1.1) *has at least one T-periodic solution x_ε satisfying $x_\varepsilon(0) \to v_0$ as $\varepsilon \to 0$.*

(b) *If* (i) *and* (v) *hold and*

$$\det (g_0)'(v_0) > 0 \quad \textit{and} \quad ([g_0]_1)'_{(1)}(v_0) + ([g_0]_2)'_{(2)}(v_0) < 0, \tag{2.6}$$

then there exists $\varepsilon_0 > 0$ *such that for any* $\varepsilon \in (0, \varepsilon_0]$ *system* (1.1) *has exactly one* T*-periodic solution* x_ε *such that* $x_\varepsilon(0) \to v_0$ *as* $\varepsilon \to 0$. *Moreover, the solution* x_ε *is asymptotically stable.*

(c) *If* $\det (g_0)'(v_0) < 0$, *then there exists* $\varepsilon_0 > 0$ *such that for any* $\varepsilon \in (0, \varepsilon_0]$ *system* (1.1) *has at least one nonasymptotically stable* T*-periodic solution* x_ε *such that* $x_\varepsilon(0) \to v_0$ *as* $\varepsilon \to 0$.

Proof. Statement (a) is added for the completeness of the formulation of Theorem 2.11 and it follows from Mawhin's theorem (see [32] or [33, Theorem 3]).

On the other hand, it is a simple calculation to show that (2.6) implies that all the eigenvalues of $(g_0)'(v_0)$ have negative real part. Therefore assumption (vi) of Theorem 2.5 is also satisfied and statement (b) follows from this theorem.

Statement (c) follows from Theorem 2.9. Indeed, since $\det (g_0)'(v_0) < 0$, it implies (see [23, Theorem 5.9]) that $d(g_0, B_\rho(v_0))$ is defined for any $\rho > 0$ sufficiently small and that $d(g_0, B_\rho(v_0)) = \det(g_0)'(v_0) < 0$. ∎

3. Application to the nonsmooth van der Pol oscillator. In [18] Hogan first demonstrated the existence of a limit cycle for the nonsmooth van der Pol equation $\ddot{u}+\varepsilon(|u|-1)\dot{u}+u = 0$. This equation governs the circuit drawn at Figure 1.1 with the triode characteristic $i(u) = S_0 + S_1 u - S_2 u|u|$ whose derivative $i'(u) = S_1 - 2S_2|u|$ is nondifferentiable (see Nayfeh and Mook [36, section 3.3.4], where the same stiffness characteristic appears in mechanics). In this paper we extend this study by considering the van der Pol problem on the location of stable and unstable periodic solutions of the perturbed equation

$$\ddot{u} + \varepsilon \left(|u| - 1\right) \dot{u} + (1 + a\varepsilon)u = \varepsilon\lambda \sin t, \tag{3.1}$$

where a is a detuning parameter and $\varepsilon\lambda \sin t$ is an external force. We assume that $\varepsilon > 0$ is sufficiently small, and we consider that the parameters a and λ vary in $\mathbb{R}$.

We finally note that standard change of variables (see example in section 3) brings (1.3) into the form (1.1), with g sufficiently smooth to satisfy the hypotheses of the second Bogoliubov theorem. But we remind the reader that our aim is to apply directly Theorem 2.5, in the same way that Andronov and Witt applied Bogoliubov theorem to the classical van der Pol oscillator

$$\ddot{u} + \varepsilon \left(u^2 - 1\right) \dot{u} + (1 + a\varepsilon)u = \varepsilon\lambda \sin t, \tag{3.2}$$

which can be found in [1, Figure 4] or in [30, Chap. I, section 16, Figure 15].

A function u is a solution of (3.1) if and only if $(z_1, z_2) = (u, \dot{u})$ is a solution of the system

$$\begin{aligned} \dot{z}_1 &= \dot{z}_2, \\ \dot{z}_2 &= -z_1 + \varepsilon[-az_1 - (|z_1| - 1)z_2 + \lambda \sin t]. \end{aligned} \tag{3.3}$$

After the change of variables

$$\begin{pmatrix} z_1(t) \\ z_2(t) \end{pmatrix} = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix},$$

system (3.3) takes the form

$$
\begin{aligned}
\dot{x}_1 &= \varepsilon \sin(-t)\left[-a(x_1\cos t + x_2 \sin t)\right. \\
&\quad \left. -\left(|x_1\cos t + x_2 \sin t| - 1\right)(-x_1 \sin t + x_2 \cos t) + \lambda \sin t\right], \\
\dot{x}_2 &= \varepsilon \cos(-t)\left[-a(x_1\cos t + x_2 \sin t)\right. \\
&\quad \left. -\left(|x_1\cos t + x_2 \sin t| - 1\right)(-x_1 \sin t + x_2 \cos t) + \lambda \sin t\right].
\end{aligned} \tag{3.4}
$$

The corresponding averaged function g_0, calculated using (1.2), is given by

$$
\begin{aligned}
[g_0]_1(M,N) &= \pi a N - \pi\lambda + \pi M - \tfrac{4}{3}M\sqrt{M^2+N^2}, \\
[g_0]_2(M,N) &= -\pi a M + \pi N - \tfrac{4}{3}N\sqrt{M^2+N^2},
\end{aligned} \tag{3.5}
$$

and it is continuously differentiable in $\mathbb{R}^2\backslash\{0\}$.

In short, by statement (a) of Theorem 2.11, the zeros $(M,N) \in \mathbb{R}^2$ of this function with the property that $\det (g_0)'(M,N) \neq 0$ determine the 2π-periodic solutions of (3.3) emanating from the solution of the unperturbed system

$$
\begin{aligned}
u_1(t) &= M\cos t + N \sin t, \\
u_2(t) &= -M \sin t + N \cos t.
\end{aligned} \tag{3.6}
$$

We have the following expression for the determinant:

$$
\det (g_0)'(M,N) = \pi^2(1+a^2) + \frac{32}{9}(M^2+N^2) - 4\pi\sqrt{M^2+N^2}. \tag{3.7}
$$

Following Andronov and Witt [1] we are concerned with the dependence of the amplitude of the solution (3.6) with respect to a and λ. Thus we decompose this solution as follows:

$$
u_1(t) = A\sin(t+\phi), \quad u_2(t) = A\cos(t+\phi), \tag{3.8}
$$

where (M,N) is related to (A,ϕ) by

$$
M = A\sin\phi, \quad N = A\cos\phi. \tag{3.9}
$$

Substituting (3.9) into (3.5) and (3.7) we obtain

$$
\begin{aligned}
[g_0((A\sin\phi, A\cos\phi))]_1 &= -(4/3)\cdot A|A|\sin\phi + \pi a A\cos\phi + \pi A\sin\phi - \pi\lambda, \\
[g_0((A\sin\phi, A\cos\phi))]_2 &= -(4/3)\cdot A|A|\cos\phi - \pi a A\sin\phi + \pi A\cos\phi,
\end{aligned} \tag{3.10}
$$

and, respectively,

$$
\det (g_0)'((A\sin\phi, A\cos\phi)) = \pi^2(1+a^2) + \frac{32}{9}A^2 - 2\pi|A|. \tag{3.11}
$$

Looking for the zeros (A,ϕ) of (3.10), we find the implicit formula

$$
A^2\left(a^2 + \left(1 - \frac{4}{3\pi}|A|\right)^2\right) = \lambda^2 \tag{3.12}
$$

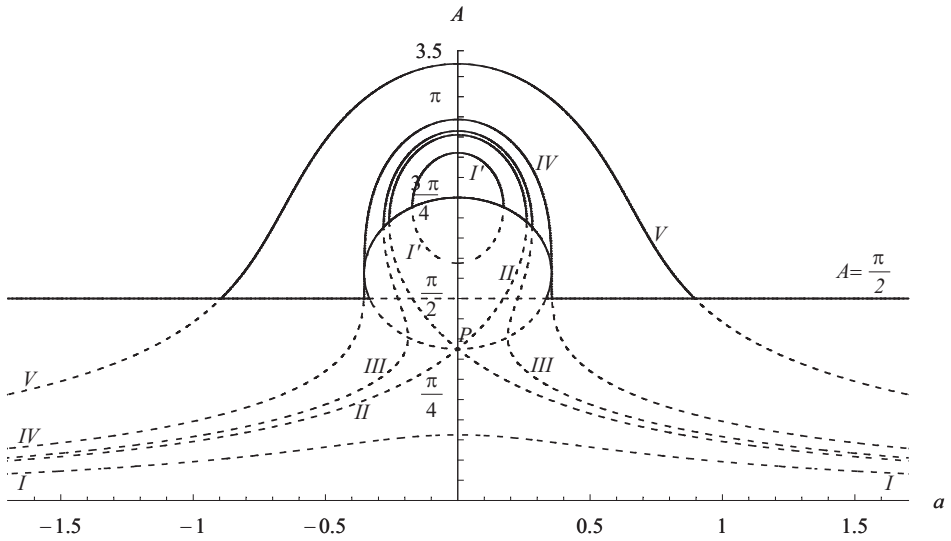


FIG. 3.1. *Dependence of the amplitude of stable (solid curves) and unstable (dash curves)* 2π*-periodic solutions of the nonsmooth periodically perturbed van der Pol equation* (3.1) *on the detuning parameter* a *obtained over formulas* (3.12), (3.16), *and* (3.17) *for different values of* λ. *The curve* I *is plotted with* $\lambda = 0.4$, II *with* $\lambda = 3\pi/16$, III *with some* $\lambda = \sqrt{0.4} \in (3\pi/16, 9\sqrt{3}\pi/64)$, IV *with* $\lambda = 9\sqrt{3}\pi/64$, *and* V *with* $\lambda = 1.5$. *Point* P *is* $2/\sqrt{3}$.

for determining A. Observe that the number of positive zeros of (3.12) coincide with the number of zeros of the equation $A^2\left(a^2 + \left(1 - \frac{4}{3\pi}A\right)^2\right) = \lambda^2$. To estimate this number we define

$$f(A) = A^2\left(a^2 + \left(1 - \frac{4}{3\pi}A\right)^2\right) - \lambda^2,$$

and we have

$$f'(A) = 2A\left(a^2 + \left(1 - \frac{4}{3\pi}A\right)^2\right) - \frac{8}{3\pi}A^2\left(1 - \frac{4}{3\pi}A\right).$$

Since f' has one or two zeros, then (3.12) has one, two, or three positive solutions A for any fixed a and λ. In order to understand the different situations that can appear, we follow Andronov and Witt, who suggested in [1] (see also [2]) to construct the so-called *resonance curves*, namely, the curves A in function of a, for λ fixed. The equation of this curve is given by formula (3.12). Some curves (3.12) corresponding to different values of λ are drawn in Figure 3.1. The way for describing these resonance curves (3.12) is borrowed from [30, Ch. 1, section 5], where the classical van der Pol equation is considered.

When $\lambda = 0$ the curve (3.12) is formed by the axis $A = 0$ and the isolated point $(0, 3\pi/4)$. When $\lambda > 0$ but sufficiently small, the resonance curve consists of two branches: instead of $A = 0$ we have the curve of the type $I - I$, and instead of the point $(0, 3\pi/4)$ we obtain an oval $I' - I'$ surrounding this point. When $\lambda > 0$ increases, the oval $I' - I'$ and the branch $I - I$ tend to each other, and, for a certain λ, there exists only one branch $II - II$ with a double point P. The value of this λ can be obtained assuming that (3.12) has for $a = 0$ a double root and, therefore, (3.11)

should be zero. Solving jointly (3.12) and (3.11) with $a = 0$ we obtain $\lambda = 3\pi/16$ and $P = 2\pi/8$. If $\lambda > 3\pi/16$, then we have curves of the type III which take form V when $\lambda > 0$ crosses the value $\lambda = 9\sqrt{3}\pi/64$. From here, if $\lambda < 3\pi/16$, then (3.12) has three real roots when $|a|$ is sufficiently small, and only one root when $|a|$ is greater than a certain number which depends on λ. When $3\pi/16 < \lambda < 9\sqrt{3}\pi/64$ (3.12) has one, three, and again one solution according to whether $a < a_1$, $a_1 < a < a_2$, and $a > a_2$, respectively, where a_1, a_2 depend on λ. The amplitude curves of type V provide exactly one solution of (3.12) for any value of a. The value $\lambda = 9\sqrt{3}\pi/64$, which separates the curves where (3.12) has three solutions from the curves where (3.12) has one solution, is obtained from the property that (3.12) with this λ has a double root for some a and thus this value of a vanishes (3.11). Therefore $\lambda = 9\sqrt{3}\pi/64$ is the point separating the interval $(0, \lambda)$ where the system formed by (3.12) and

$$\pi^2(1 + a^2) + \frac{32}{9}A^2 - 2\pi|A| = 0 \tag{3.13}$$

has at least one solution from the interval (λ, ∞) where (3.12)–(3.13) has no solutions.

In short, we have studied the amplitudes of the 2π-periodic solutions of system (3.3) depending on a and λ, when a physical system described by (3.3) possesses 2π-periodic oscillations and when some of them are asymptotically stable. To find the answer we have used statement (b) of Theorem 2.11. Assumption (i) is obviously satisfied with $\Omega = \mathbb{R}^2$. The next statement shows that the right-hand side of system (3.4) satisfies (v).

PROPOSITION 3.1. *Let $v_0 \in \mathbb{R}^2$, $v_0 \neq 0$. Then the right-hand side of* (3.4) *satisfies* (v) *for any $a, \lambda \in \mathbb{R}$.*

The proof of Proposition 3.1 is given in section 4.

Thus we have to study the signs of (3.11) and $([g_0]_1)'_M(A\sin\phi, A\cos\phi) + ([g_0]_2)'_N(A\sin\phi, A\cos\phi)$. We have

$$([g_0]_1)'_M(M, N) + ([g_0]_2)'_N(M, N) = 2\left(\pi - 2\sqrt{M^2 + N^2}\right),$$

and therefore the conditions for the asymptotic stability of the 2π-periodic solutions of (3.3) near (3.6) are

$$\pi^2(1 + a^2) + \frac{32}{9}(M^2 + N^2) - 4\pi\sqrt{M^2 + N^2} > 0 \tag{3.14}$$

and

$$2\left(\pi - 2\sqrt{M^2 + N^2}\right) < 0. \tag{3.15}$$

Substituting (3.9) into the inequalities (3.14) and (3.15), we obtain the following equivalent inequalities in terms of the amplitude A:

$$\pi^2(1 + a^2) + \frac{32}{9}A^2 - 2\pi|A| > 0 \tag{3.16}$$

and

$$2\pi - 4|A| < 0. \tag{3.17}$$

Conditions (3.16) and (3.17) mean that the asymptotically stable 2π-periodic solutions of (3.3) correspond to those parts of resonance curves under consideration which are

outside the ellipse (3.13) and above the line $A = \pi/2$. All the results are collected in Figure 3.1, where it is easy to see that for any detuning parameter a and any amplitude $\lambda > 0$, (3.1) possesses at least one asymptotically stable 2π-periodic solution with amplitude close to A obtained from (3.12). Among all the asymptotically stable 2π-periodic solutions of (3.1), there exists exactly one whose fixed neighborhood does not contain any nonasymptotically stable 2π-periodic solution of (3.1) for sufficiently small $\varepsilon > 0$. The amplitude of this asymptotically stable 2π-periodic solution is obtained from (3.16)–(3.17).

To compare the changes due to nonsmoothness in the behavior of the resonance curves, we give in Figure 3.2 the resonance curves of the classical van der Pol oscillator (3.2).

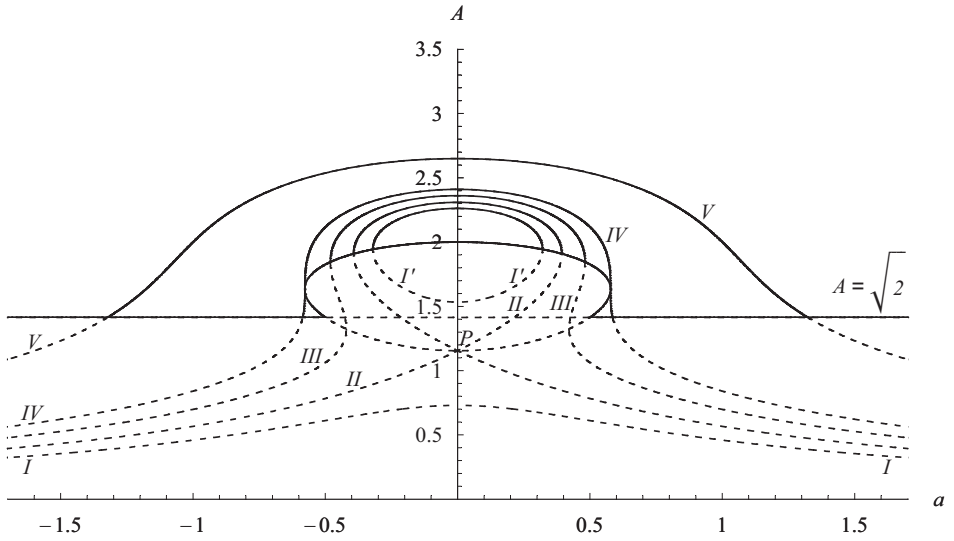


FIG. 3.2. *Dependence of the amplitude of stable (solid curves) and unstable (dash curves) 2π-periodic solutions of the classical periodically perturbed van der Pol equation* (3.2) *on the detuning parameter a for different values of λ. Following Andronov and Witt (see* [1, *Figure* 4]*), curve I is plotted with $\lambda = \sqrt{0.4}$, II with $\lambda = 4\sqrt{3}/9$, III with some $4\sqrt{3}/9 < \lambda < \sqrt{32/27}$, IV with $\lambda = \sqrt{32/27}$, and V with $\lambda = 2$. Point P is $2/\sqrt{3}$.*

The formulas of Figure 3.1 can be compared with the formulas for Figure 3.2. In fact, the corresponding expressions (3.12)–(3.13) and (3.14)–(3.15) are (see the formulas (5.21)–(5.22) and (16.6)–(16.7) of [30])

$$A^2\left(a^2 + \left(1 - \frac{A^2}{4}\right)^2\right) = \lambda^2,$$

$$1 - a^2 - A^2 + \frac{3}{16}A^4 = 0,$$

and

$$1 + a^2 - (M^2 + N^2) + \frac{3}{16}(M^2 + N^2)^2 > 0,$$

$$2 - (M^2 + N^2) < 0,$$

respectively, when we consider the classical van der Pol equation (3.2).

It can be checked that the eigenvectors of the matrix $(g_0)'((A\sin\phi, A\cos\phi))$ are orthogonal only for $A = 0$, so Theorem 3.6 from [8] cannot be applied. At the same time assumption (H2) from [8] is not satisfied for our problem (see Remark 2.8).

4. Appendix.

Proof of Proposition 3.1. As before, $[v]_i$ is the ith component of the vector $v \in \mathbb{R}^2$. Let $g(t,v) = |[v]_1\cos t + [v]_2\sin t|$, and notice that it is enough to prove that $g : [0,2\pi]\times\mathbb{R}^2 \to \mathbb{R}$ satisfies (v). In the case that $[v_0]_2 \neq 0$, denote $\theta(v) = \arctan(-[v]_1/[v]_2)$ if $[v_0]_2 = 0$, denote $\theta(v) = \arctan(-[v]_1/[v]_2)$ if $[v_0]_1[v]_2 < 0$, $\theta(v_0) = \pi/2$, and, respectively, $\theta(v) = \arctan(-[v]_1/[v]_2) + \pi$ if $[v_0]_1[v]_2 > 0$. In any case notice that the function $v \mapsto \theta(v)$ is continuous in every sufficiently small neighborhood of v_0. Fix $\widetilde{\gamma} > 0$. Let M be the union of two intervals centered in $\theta(v_0)$ (respectively, $\theta(v_0) + 2\pi$ if $\theta(v_0) < 0$) and in $\theta(v_0) + \pi$, each of length $\widetilde{\gamma}/2$. Denote them M_1 and M_2. Take $\widetilde{\delta} > 0$ such that $\theta(v) \in M_1$ for all $v \in B_{\widetilde{\delta}}(v_0)$. Of course, also $\theta(v) + \pi \in M_2$ for all $\|v - v_0\| \leq \widetilde{\delta}$. This implies that for fixed $t \in [0,2\pi]\setminus M$, $[v]_1\cos t + [v]_2\sin t$ has constant sign for all $v \in B_{\widetilde{\delta}}(v_0)$, which further gives that $g(t,\cdot)$ is differentiable and $g'_v(t,v) = g'_v(t,v_0)$ for all $v \in B_{\widetilde{\delta}}(v_0)$. Hence (v) is fulfilled. ∎

Acknowledgments. The authors are grateful to the anonymous referee who motivated us to include a list of several applications of Theorem 2.5 in the introduction—that definitely improved the paper. The authors also thank Aris Daniilidis for helpful discussions and Rafael Ortega who called our attention to the change of variables used in the Levinson paper [29]. Part of this work was done during a visit of the first and the third author to the Centre de Recerca Matemàtica, Barcelona (CRM). They express their gratitude to the CRM for providing very nice working conditions.

Finally we thank Martin Golubitsky and Andre Vanderbauwhede, who invited us to present the paper at their minisimposia "Recent developments in bifurcation theory" of Equadiff 2007 (see [9]), giving a significant impact to its recognition.

REFERENCES

[1] A. Andronov and A. Witt, *On mathematical theory of entrainment*, Z. Prikl. Phis., 6 (1930), pp. 3–17 (in Russian).

[2] A. Andronow and A. Witt, *Zur Theorie des Mitnehmens von van der Pol*, Arch. Elektrotechnik, 24 (1930), pp. 99–110 (in German).

[3] A. A. Andronov, A. A. Witt, and S. E. Khaikin, *Theory of Oscillators*, (translated from the Russian by F. Immirzi; translation edited and abridged by W. Fishwick) Pergamon Press, Oxford, 1966.

[4] J. Awrejcewicz and C. H. Lamarque, *Bifurcation and Chaos in Nonsmooth Mechanical Systems*, World Scientific, River Edge, NJ, 2003.

[5] M. S. Baptista, T. P. Silva, J. C. Sartorelli, and I. L. Caldas, *Phase synchronization in the perturbed Chua circuit*, Phys. Rev. E (3), 74 (2006), no. 056707.

[6] N. N. Bogoliubov, *On Some Statistical Methods in Mathematical Physics*, Akademiya Nauk Ukrainskoi SSR, Kiev, 1945 (in Russian).

[7] A. P. Bovsunovskii, *Comparative analysis of nonlinear resonances of a mechanical system with unsymmetrical piecewise characteristic of restoring force*, Strength Matherials, 39 (2007), pp. 159–169.

[8] A. Buică and A. Daniilidis, *Stability of periodic solutions for Lipschitz systems obtained via the averaging method*, Proc. Amer. Math. Soc., 135 (2007), pp. 3317–3327.

[9] A. Buică, J. Llibre, and O. Makarenkov, *Lipschitz constant of integral with respect to a parameter versus derivative of integral with respect to a parameter in the theory of nonsmooth bifurcations*, Book of abstracts, Equadiff 2007, Vienna University of Technology, 2007, pp. 88–89.

[10] L. O. Chua, *Global unfolding of Chua's circuit*, IEICE Trans. Fundamentals, E76-A, 5 (1993), pp. 704–734.
[11] D. Cofagna and G. Grassi, *Chaotic beats in a modified Chua's circuit: Dynamic behaviour and circuit design*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 17 (2007), pp. 209–226.
[12] B. P. Demidovich, *Lectures on the Mathematical Theory of Stability*, Izdat. Nauka, Moscow, 1967 (in Russian).
[13] S. W. S. W. Doebling, C. R. Farrar, M. B. Prime, and D. W. Shevitz, *Damage Identification and Health Monitoring of Structural and Mechanical Systems from Changes in Their Vibration Characteristics: A Literature Review*, LA-13070-MS, UC-900, Los Alamos National Laboratory, 1996.
[14] C. Fabry, *Large-amplitude oscillations of a nonlinear asymmetric oscillator with damping*, Nonlinear Anal. Ser. A: Theory Methods, 44 (2001), pp. 613–626.
[15] P. Fatou, *Sur le mouvement d'un système soumis à des forces à courte période*, Bull. Soc. Math. France, 56 (1928), pp. 98–139 (in French).
[16] J. Glover, A. C. Lazer, and P. J. McKenna, *Existence and stability of large scale nonlinear oscillations in suspension bridges*, Z. Angew. Math. Phys., 40 (1989), pp. 172–200.
[17] J. K. Hale, *Ordinary Differential Equations*, Robert E. Krieger Publishing Co., Inc., Huntington, NY, 1980.
[18] S. J. Hogan, *Relaxation oscillations in a system with a piecewise smooth drag coefficient*, J. Sound Vibration, 263 (2003), pp. 467–471.
[19] A. A. Jakkula, *A history of suspension bridges in bibliographical form*, Bull. Agricultural and Mechanical College of Texas 4th ser., vol. 12 Federal Works Agency, Washington, D.C., 1941.
[20] G. A. Johnston and E. R. Hunt, *Derivative control of the steady state in Chua's circuit driven in the chaotic region*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 40 (2000), pp. 833–835.
[21] A. N. Kolmogorov and S. V. Fomin, *Introductory Real Analysis* (translated and edited by R. A. Silverman, Selected Russian Publications in the Mathematical Sciences, Prentice-Hall, Englewood Cliffs, NJ, 1970.
[22] M. A. Krasnosel'skii, *Positive Solutions of Operator Equations* (translated from the Russian by R. E. Flaherty; edited by Leo F. Boron), P. Noordhoff Ltd., Groningen, The Netherlands, 1964.
[23] M. A. Krasnosel'skii, *The Operator of Translation Along the Trajectories of Differential Equations*, Translations of Mathematical Monographs 19 (translated from the Russian by Scripta Technica), American Mathematical Society, Providence, RI, 1968.
[24] B. I. Kryukov, *Dynamics of Resonance-Type Vibration Machines*, Naukova Dnka, Kiev, 1967 (in Russian).
[25] N. M. Krylov and N. N. Bogoliubov, *Introduction to Non-linear Mechanics*, Akademiya Nauk Ukrainskoi SSR, 1937 (in Russian); English translation: Annals of Mathematics Studies 11, Princeton University Press, Princeton, NJ, 1943.
[26] A. C. Lazer and P. J. McKenna, *Existence, uniqueness, and stability of oscillations in differential equations with asymmetric nonlinearities*, Trans. Amer. Math. Soc., 315 (1989), pp. 721–739.
[27] A. C. Lazer and P. J. McKenna, *Large-amplitude periodic oscillations in suspension bridges: Some new connections with nonlinear analysis*, SIAM Rev., 32 (1990), pp. 537–578.
[28] M. Levi, *Qualitative analysis of the periodically forced relaxation oscillations*, Mem. Amer. Math. Soc., 32 (1981), no. 244.
[29] N. Levinson, *A second order differential equation with singular solutions*, Ann. Math. (2), 50 (1949), pp. 127–153.
[30] I. G. Malkin, *Some problems of the theory of nonlinear oscillations*, Gosudarstv. Izdat. Tehn.-Teor. Lit., Moscow, 1956 (in Russian); English translation: AEC tr 3766 (book 1), US Atomic Energy Commission, 1959.
[31] L. I. Mandelstam and N. D. Papaleksi, *On justification of one approximation method for solving differential equations*, Zh. Eksper. Teoret. Fiz., IV (1934), no. 117.
[32] J. Mawhin, *Le Problème des Solutions Périodiques en Mécanique non Linéaire*, Thèse de doctorat en sciences, Université de Liège, 1969; published in *Degré topologique et solutions périodiques des systèmes différentiels non linéaires*, Bull. Soc. Roy. Sci. Liège, 38 (1969), pp. 308–398.
[33] J. Mawhin, *Periodic solutions in the golden sixties: The birth of a continuation theorem*, in Ten Mathematical Essays on Approximation in Analysis and Topology, Elsevier B. V., Amsterdam, 2005, pp. 199–214.

[34] Yu. A. Mitropol'skii, *On periodic solutions of systems of nonlinear differential equations with non-differentiable right-hand sides*, Ukrain. Mat. Ž., 11 (1959), pp. 366–379 (in Russian).
[35] K. Murali and M. Lakshmanan, *Chaotic dynamics of the driven Chua's circuit*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 50 (2003), pp. 1503–1508.
[36] A. H. Nayfeh and D. T. Mook, *Nonlinear Oscillations*, Wiley-Interscience, New York, 1979.
[37] L. S. Pontrjagin, *Ordinary differential equations* (translated from the Russian by L. Kacinskas and W. B. Counts), Adiwes International Series in Mathematics, Addison–Wesley Publishing, Reading, MA, 1962.
[38] R. Puers, J. Bienstman, and J. Vandewalle, *The autonomous impact resonator: A new operation principle for a silicon resonant strain gauge*, in Int. Conf. on Solid-State Sensors and Actuators, Digest of Technical Papers, Chicago, 1997, pp. 1105–1108.
[39] A. M. Samoilenko, *On periodic solutions of differential equations with nondifferentiable right–hand sides*, Ukrain. Mat. Ž., 15 (1963), pp. 328–332 (in Russian).
[40] E. Sánchez, M. A. Matías, and V. Pérez-Muñuzuri, *Chaotic synchronization in small assemblies of driven Chua's circuits*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 47 (2000), pp. 644–654.

SIAM J. MATH. ANAL.
Vol. 40, No. 6, pp. 2496–2516

SPATIAL DYNAMICS OF A NONLOCAL PERIODIC REACTION-DIFFUSION MODEL WITH STAGE STRUCTURE*

YU JIN† AND XIAO-QIANG ZHAO†

Abstract. In this paper, we investigate a nonlocal periodic reaction-diffusion population model with stage structure. In the case of unbounded spatial domain, we establish the existence of the asymptotic speed of spread and show that it coincides with the minimal wave speed for monotone periodic traveling waves. In the case of bounded spatial domain, we obtain a threshold result on the global attractivity of either zero or a positive periodic solution.

Key words. nonlocal periodic model, spreading speed, traveling waves, positive periodic solution, global attractivity

AMS subject classifications. 34K05, 35K57, 37N25, 92B05

DOI. 10.1137/070709761

1. Introduction.

Age structure has been an interesting topic in population dynamics (see, e.g., [1, 2, 3, 5, 6, 7, 9, 14, 17, 18, 20] and the references therein), since we can investigate the separate quantities of immature and mature populations in an age-structured population model. To derive a model for a single species of population with age structure and diffusion, we usually assume that individuals move around not only after maturity, but also while immature. For a standard argument, Metz and Diekmann [14] give

$$\frac{\partial u}{\partial t}+\frac{\partial u}{\partial a}=D(a)\frac{\partial^2 u}{\partial x^2}-\mu(a)u,$$

where $u(t,a,x)$ is the density of the population of the species at time $t\geq 0$, age $a\geq 0$, and location x in a spatial domain Ω; $D(a)\geq 0$ and $\mu(a)\geq 0$ are the diffusion rate and the death rate of the population at age a, respectively.

To study the behaviors of immature individuals and mature individuals, we can also divide the population of a species into two groups: immature population and mature population. For simplicity, we assume that the maturation time (or the length of the juvenile period) is the same for all juvenile individuals, denoted by $\tau\geq 0$. For distributed maturation delay, see, e.g., [2, 3] and the references therein. Assume that the diffusion rate and death rate are age-dependent for immature individuals, but age-independent for mature individuals. As a result, we have the following system for a single species of population with age structure and diffusion (see also [5, 17, 18, 20]):

(1.1)
$$\begin{cases}\partial_t u(t,a,x)+\partial_a u(t,a,x)=d_j(a)\Delta u-\mu_j(a)u(t,a,x), & t>0, 0<a<\tau, x\in\Omega,\\ u(t,0,x)=f(u_m(t,x)), & t\geq -\tau, x\in\Omega,\\ \partial_t u_m(t,x)=d_m\Delta u_m-g(u_m(t,x))+u(t,\tau,x), & t>0, x\in\Omega,\end{cases}$$

*Received by the editors November 30, 2007; accepted for publication (in revised form) November 17, 2008; published electronically February 25, 2009. This research was supported in part by the NSERC of Canada and MITACS of Canada.
http://www.siam.org/journals/sima/40-6/70976.html
†Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, NL, A1C 5S7, Canada (yuj@math.mun.ca, xzhao@math.mun.ca).

where $u(t,a,x)$ is the density of the population at time $t \geq -\tau$, age $a \geq 0$, location $x \in \Omega$, $u_m(t,x)$ is the density of the mature population, $f(u_m)$ and $g(u_m)$ are the birth rate and the mortality rate of mature individuals, respectively, $d_j(a) \geq 0$ is the diffusion rate of the immature individuals at age $a \in (0,\tau)$, $d_m \geq 0$ is the diffusion rate of the mature individuals, $\mu_j(a) > 0$ denotes the per capita mortality rate of juveniles at age a, $u(t,\tau,x)$ is the adults recruitment term for those of maturation age τ, Δ is the Laplacian operator.

In fact, the dynamics of many populations is influenced greatly by the time varying environments (e.g., due to seasonal variation). For example, in a one year period, the birth rate may be high in spring and summer and low in winter, while in winter more individuals might be at risk of death because of low temperature, lack of food, or some other reasons. Moreover, populations usually like to move in warm weather during the spring and summer time. Therefore, it is more realistic to consider a nonautonomous version of (1.1) for population dynamics. In particular, a periodic model, in which the birth rate, mortality rates, and diffusion rates are assumed to be periodic in time, is probably the simplest but nonetheless interesting and realistic case. In this paper, we consider the following model:

(1.2)
$$\begin{cases} \partial_t u(t,a,x) + \partial_a u(t,a,x) = d_j(t,a)\Delta u - \mu_j(t,a)u(t,a,x), & t > 0, a \in (0,\tau),\ x \in \Omega, \\ u(t,0,x) = f(t,u_m(t,x)), & t \geq -\tau,\ x \in \Omega, \\ \partial_t u_m(t,x) = d_m(t)\Delta u_m - g(t,u_m(t,x)) + u(t,\tau,x), & t > 0,\ x \in \Omega, \end{cases}$$

where $d_j(t,a) \geq 0$ and $\mu_j(t,a) \geq 0$ denote the diffusion rate and the per capita mortality rate of juveniles at age a at time t, respectively; $d_m(t) \geq 0$ denotes the diffusion rate of mature individuals at time t; $f(t,u_m)$ and $g(t,u_m)$ are the birth and mortality rates of mature individuals at time t, respectively.

Similarly as in [18] (see also [16]), we integrate along characteristics to reduce the system (1.2) to one equation with nonlocal terms. Let $v(r,a,x) = u(a+r,a,x)$, where r is regarded as a parameter. It follows that

$$\begin{cases} \partial_a v(r,a,x) = \big[\partial_t u(t,a,x) + \partial_a u(t,a,x)\big]_{t=r+a} \\ \qquad\qquad = d_j(a+r,a)\Delta v(r,a,x) - \mu_j(a+r,a)v(r,a,x), \\ v(r,0,x) = f(r,u_m(r,x)). \end{cases}$$

Integrating the last equation, we obtain

$$v(r,a,x) = \int_\Omega \Gamma(\zeta(r,a),x-y)F(r,a)f(r,u_m(r,y))dy,$$

where Γ is the fundamental solution associated with the partial differential operator $\partial_t - \Delta$ and

$$\zeta(r,a) = \int_0^a d_j(s+r,s)ds, \qquad F(r,a) = \exp\left(-\int_0^a \mu_j(s+r,s)ds\right). \tag{1.3}$$

Since $u(t,a,x) = v(t-a,a,x)$, it follows that

$$u(t,a,x) = \int_\Omega \Gamma(\zeta(t-a,a),x-y)F(t-a,a)f(t-a,u_m(t-a,y))dy. \tag{1.4}$$

Set

$$a(t) := \zeta(t-\tau,\tau), \quad b(t) := F(t-\tau,\tau), \quad f_{-\tau}(t,u) := f(t-\tau,u).$$

Substituting (1.4) into the equation for u_m in (1.2), we finally reduce the age-structured population model (1.2) to the following time-delayed reaction-diffusion equation for mature individuals:

(1.5)
$$\begin{cases} \partial_t u_m(t,x) \\ = d_m(t)\Delta u_m - g(t,u_m(t,x)) + b(t)\int_\Omega \Gamma(a(t),x-y) f_{-\tau}(t,u_m(t-\tau,y))dy, \\ u_m(s,x) = \phi(s,x), \quad s\in[-\tau,0],\ x\in\Omega, \end{cases}$$

where $\phi(t,x)$ is an initial function to be specified later. For simplicity, dropping all m's and writing $u_m(t,x)$ as $u(t,x)$, we investigate the following system:

(1.6)
$$\begin{cases} \partial_t u(t,x) = d(t)\Delta u - g(t,u(t,x)) + b(t)\displaystyle\int_\Omega \Gamma(a(t),x-y) f_{-\tau}(t,u(t-\tau,y))dy, \\ u(s,x) = \phi(s,x), \quad s\in[-\tau,0],\ x\in\Omega. \end{cases}$$

Basically we assume that $d_j(t,a)$ and $\mu_j(t,a)$ are periodic in $t\ge 0$ with the period $\omega>0$ for $a\in(0,\tau)$, and that $d(t)$, $g(t,u)$, and $f(t,u)$ are periodic in t with the period $\omega>0$ for $u\in\mathbb{R}_+$. This implies that $a(t)=a(t+\omega)$, $d(t)=d(t+\omega)$, $b(t)=b(t+\omega)$, $g(t,u)=g(t+\omega,u)$, and $f(t,u)=f(t+\omega,u)$ for all $t\ge 0$, $u\in\mathbb{R}_+$. Moreover, we assume $d(t)\ge d>0$ for all $t\ge 0$, and

(H1) $f\in C^1([-\tau,+\infty)\times\mathbb{R}_+,\mathbb{R}_+)$, $g\in C^1(\mathbb{R}^2_+,\mathbb{R}_+)$, $f(t,0)=0$ for $t\ge-\tau$, $f_u(t,u)>0$ for all $t\ge-\tau$ and $u\ge 0$, $g(t,0)=0$ for $t\ge 0$, and there exists $l_1>0$ such that $|g(t,u)-g(t,v)|\le l_1|u-v|$ for all $t\ge 0$ and $u,v\in\mathbb{R}_+$;

(H2) $G(t,u,v) := -g(t,u)+b(t)f_{-\tau}(t,v)$ is strictly subhomogeneous in (u,v) in the sense that for any $\alpha\in(0,1)$, $G(t,\alpha u,\alpha v)>\alpha G(t,u,v)$ for all $u,v\ge 0$;

(H3) there exists positive number $L>0$ such that $G(t,\bar L,\bar L)\le 0$ for all $t\ge 0$, $\bar L\ge L$.

The purpose of this paper is to study the asymptotic speed of spread and periodic traveling waves of (1.6) in the infinite spatial domain, and the global attractivity of zero or a positive periodic solution of (1.6) in a bounded spatial domain. The asymptotic speed of spread (in short, spreading speed) was first introduced by Aronson and Weinberger [4] for reaction-diffusion equations and has been an important ecological metric in a wide range of ecological applications; see, e.g., [10, 11, 18] and the references therein. Intuitively, the spreading speed c^* in a spatial epidemic model can be interpreted as: if one runs at a speed $c>c^*$, then one will leave the epidemic behind; whereas if one runs at a speed $c<c^*$, then one will eventually be surrounded by the epidemic. Traveling wave solutions have also been investigated extensively for a variety of evolution systems; see, e.g., [5, 10, 11, 17, 18] and the references therein. For the autonomous case of (1.6), the dynamics, including spreading speed and traveling waves, have been studied extensively. So, Wu and Zou [17] investigated traveling wave fronts in the case where $\Omega=\mathbb{R}$, $g(u)=\beta u$. Gourley and Kuang [5] established the linear stabilities of two spatially homogeneous equilibrium solutions, studied traveling

wave fronts in the case where $\Omega = \mathbb{R}$, $f(u) = \alpha u$, and $g(u) = \beta u^2$, and obtained a global convergence theorem in the case of bounded intervals. Thieme and Zhao [18] studied the traveling wave solutions, minimal wave speed, and asymptotic speed of spread in the case of $\Omega = \mathbb{R}^n$. Xu and Zhao [20] established a threshold dynamic and global attractivity of the positive steady state when Ω is a bounded domain in $\mathbb{R}^n$.

This paper is organized as follows. In section 2, we first establish the well-posedness and the comparison principle for (1.6) with $\Omega = \mathbb{R}$, then prove the existence of the spreading speed c^* for solutions of (1.6) with initial data having compact supports, and show that it coincides with the minimal wave speed for monotone periodic traveling waves, by appealing to the theory of the spreading speed and traveling waves for monotone periodic semiflows developed in [10, 11]. In section 3, we use the theory of monotone and subhomogeneous dynamical systems to investigate the global dynamics of (1.6) in a bounded domain $\Omega \subseteq \mathbb{R}^n$, and obtain a threshold result for global attractivity of either zero or a positive periodic solution.

2. Spreading speed and traveling waves. In this section, we consider that the population diffuses in an unbounded spatial domain and study (1.6) with $\Omega = \mathbb{R}$:

(2.1)
$$\begin{cases} \partial_t u(t,x) = d(t)\Delta u - g(t,u(t,x)) + b(t)\displaystyle\int_{\mathbb{R}} \Gamma(a(t), x-y) f_{-\tau}(t, u(t-\tau, y))dy, \\ \\ u(t,x) = \phi(t,x), \quad t \in [-\tau, 0],\ x \in \mathbb{R}. \end{cases}$$

In the following, we first apply the threshold dynamics in a scalar periodic and time-delayed equation, developed by Xu and Zhao [21], to the spatially homogeneous system associated with (2.1) to find a periodic solution of (2.1). Then we use the theory of abstract functional differential equations and reaction-diffusion systems to establish the existence of solutions to (2.1) and a comparison principle. Finally, we prove that the solution periodic semiflow of (2.1) satisfies all the assumptions on monotone periodic semiflows in [10], and hence, we obtain the existence of the spreading speed and traveling wave solutions for (2.1).

Let $\mathbb{Y}$ be the space of all continuous functions from $[-\tau, 0]$ to $\mathbb{R}$ with the usual supreme norm $\|\cdot\|_{\mathbb{Y}}$ (i.e., $\mathbb{Y} = C([-\tau, 0], \mathbb{R})$), and let $\mathbb{Y}_+ = C([-\tau, 0], \mathbb{R}_+)$. Then $(\mathbb{Y}, \mathbb{Y}_+)$ is an ordered Banach space. For $\varphi, \psi \in \mathbb{Y}$, we write $\varphi \le \psi$ if $\psi - \varphi \in \mathbb{Y}_+$, $\varphi < \psi$ if $\psi - \varphi \in \mathbb{Y}_+ \setminus \{0\}$, $\varphi \ll \psi$ if $\psi - \varphi \in \mathrm{int}(\mathbb{Y}_+)$. Moreover, we define $\mathbb{Y}_r = \{\varphi \in \mathbb{Y} : 0 \le \varphi \le r\}$ for any $r \in \mathbb{Y}$ with $r \gg 0$.

Let $\mathbb{X}$ be the set of all bounded and continuous functions from $\mathbb{R}$ into $\mathbb{R}$ and $\mathbb{X}_+ = \{\varphi \in \mathbb{X}; \varphi(x) \ge 0 \text{ for all } x \in \mathbb{R}\}$. For $\varphi, \psi \in \mathbb{X}$, we write $\varphi \le \psi (\varphi \ll \psi)$ if $\varphi(x) \le \psi(x)$ $(\varphi(x) < \psi(x))$ for all $x \in \mathbb{R}$, $\varphi < \psi$ if $\varphi \le \psi$ but $\varphi \ne \psi$. It is easy to see that $\mathbb{X}_+$ is a positive cone of $\mathbb{X}$. Define $\mathbb{X}_r = \{\varphi \in \mathbb{X} : 0 \le \varphi \le r\}$ for any $r \in \mathbb{X}$ with $r \gg 0$. We equip $\mathbb{X}$ with the compact open topology, i.e., $u^m \to u$ in $\mathbb{X}$ means that the sequence of $u^m(x)$ converges to $u(x)$ as $m \to \infty$ uniformly for x in any compact set on $\mathbb{R}$. Define

$$\| u \|_{\mathbb{X}} = \sum_{k=1}^{\infty} \frac{\max_{|x| \le k} |u(x)|}{2^k} \qquad \forall u \in \mathbb{X},$$

where $|\cdot|$ denotes the usual norm in $\mathbb{R}$. Then $(\mathbb{X}, \|\cdot\|_{\mathbb{X}})$ is a normed space. Let $d_{\mathbb{X}}(\cdot,\cdot)$ be the distance induced by the norm $\|\cdot\|_{\mathbb{X}}$. It follows that the topology in the metric space $(\mathbb{X}, d_{\mathbb{X}})$ is the same as the compact open topology in $\mathbb{X}$. Moreover, $(\mathbb{X}_r, d_{\mathbb{X}})$ is a complete metric space.

Let C be the set of continuous functions from $[-\tau, 0]$ into $\mathbb{X}$, $C_+ = \{\varphi \in C, \varphi(s) \in \mathbb{X}_+, s \in [-\tau, 0]\}$ and $C_r = \{\varphi \in C : 0 \le \varphi \le r\}$ for any $r \in \mathbb{Y}$ with $r \gg 0$. Then C_+ is a positive cone of C. For convenience, we also identify an element $\varphi \in C$ as a function from $[-\tau, 0] \times \mathbb{R}$ into $\mathbb{R}$ defined by $\varphi(s, x) = \varphi(s)(x)$ for any $s \in [-\tau, 0]$ and $x \in \mathbb{R}$. For $\varphi, \psi \in C$, we write $\varphi \le \psi (\varphi \ll \psi)$ if $\varphi(s, x) \le \psi(s, x)$ $(\varphi(s, x) < \psi(s, x))$ for all $s \in [-\tau, 0], x \in \mathbb{R}$, $\varphi < \psi$ if $\varphi \le \psi$ but $\varphi \ne \psi$. For any continuous function $w(\cdot) : [-\tau, b) \to \mathbb{X}$, $b > 0$, we define $w_t \in C$ by $w_t(s) = w(t + s)$ for all $t \in [0, b)$, $s \in [-\tau, 0]$. It is then easy to see that $t \to w_t$ is a continuous function from $[0, b)$ to C. Moreover, we also equip C with the compact open topology and define the norm on C:

$$\| u \|_C = \sum_{k=1}^{\infty} \frac{\max_{|x| \le k, s \in [-\tau, 0]} |u(s, x)|}{2^k} \qquad \forall u \in C,$$

where $|\cdot|$ denotes the usual norm in $\mathbb{R}$.

For any constant $N > 0$, $\widehat{N}$ denotes the constant function with value N in $\mathbb{Y}$, $\mathbb{X}$, or C.

Now we consider the spatially homogeneous system associated with (2.1). Letting $u(t, x) = w(t)$, we have

$$\begin{cases} \dfrac{dw(t)}{dt} = -g(t, w(t)) + b(t) f_{-\tau}(t, w(t - \tau)), \\ w(t) = \varphi(t), \quad t \in [-\tau, 0], \quad \varphi \in \mathbb{Y}_+. \end{cases} \tag{2.2}$$

The linearized equation associated with (2.2) at $w = 0$ is

$$\begin{cases} \dfrac{dw(t)}{dt} = -g_u(t, 0) w(t) + b(t) \partial_u f_{-\tau}(t, 0) w(t - \tau), \\ w(t) = \varphi(t), \quad t \in [-\tau, 0], \quad \varphi \in \mathbb{Y}_+. \end{cases} \tag{2.3}$$

Since g, b, and $f_{-\tau}$ are periodic functions in $t \ge 0$, we can easily see that for any $\varphi \in \mathbb{Y}_+$, (2.3) admits a unique solution $w(t, \varphi)$ existing for all $t \ge -\tau$ with $w(s, \varphi) = \varphi(s)$ for $s \in [-\tau, 0]$, and $w_t(\varphi) \in \mathbb{Y}_+$ for all $t \ge 0$, where $\{w_t\}_{t \ge 0}$ is the solution semiflow for (2.3) defined by $w_t(\varphi)(s) = w(t + s, \varphi)$ for all $s \in [-\tau, 0]$, $t > 0$.

Define the Poincaré map of (2.3) $P : \mathbb{Y}_+ \to \mathbb{Y}_+$ by $P(\varphi) = w_\omega(\varphi)$ for all $\varphi \in \mathbb{Y}_+$, and let $r = r(P)$ be the spectral radius of P. The following two results come from [21].

PROPOSITION 2.1 (see [21, Proposition 2.1]). *$r = r(P)$ is positive and is an eigenvalue of P with a positive eigenfunction φ^*. Moreover, if $\tau = k\omega$ for some integer $k \ge 0$, then $r - 1$ has the same sign as $\int_0^\omega [-g_u(t, 0) + b(t) \partial_u f_{-\tau}(t, 0)] dt$.*

THEOREM 2.2 (see [21, Theorem 2.1]). *Let* (H1)–(H3) *hold. The following statements are valid.*

(i) *If $r \le 1$, then zero solution is globally asymptotically stable for* (2.2) *with respect to $\mathbb{Y}_+$.*

(ii) *If $r > 1$, then* (2.2) *has a unique positive ω-periodic solution $\beta^*(t)$, and $\beta^*(t)$ is globally asymptotically stable with respect to $\mathbb{Y}_+ \setminus \{0\}$.*

In the remainder of this section, we further assume that

(H4) $r = r(P) > 1$.

By the proof of [21, Theorem 2.1] and (H3), it is easy to see that $\beta^*(t) \in [0, L]$ for all $t \ge -\tau$ and $[\hat{0}, \hat{L}]$ is positively invariant for (2.1). Define $\beta_0^* \in \mathbb{Y}_{\hat{L}}$ as $\beta_0^*(s) = \beta^*(s)$ for all $s \in [-\tau, 0]$.

Consider

$$\begin{cases} \partial_t u(t,x) = d(t)\Delta u, \quad t > 0, \\ u(0,x) = \phi(x), \quad x \in \mathbb{R}, \phi \in \mathbb{X}. \end{cases} \tag{2.4}$$

The solution of (2.4) can be expressed as

$$u(t,x,\phi) = \int_{\mathbb{R}} \Gamma(\eta(t), x-y)\phi(y)dy, \quad t \geq 0, \tag{2.5}$$

where $\eta(t) = \int_0^t d(s)ds$. According to [8, Chapter II], (2.4) admits an evolution operator $U(t,s) : \mathbb{X} \to \mathbb{X}$, $0 \leq s \leq t$, which satisfies $U(t,t) = I$, $U(t,s)U(s,\rho) = U(t,\rho)$ for all $0 \leq \rho \leq s \leq t$, and $U(t,0)(\phi)(x) = u(t,x,\phi)$ for $t \geq 0$, $x \in \mathbb{R}$, and $\phi \in \mathbb{X}$, where $u(t,x,\phi)$ is the solution of (2.4). Moreover, for any $0 \leq s < t$, $U(t,s)$ is a compact and positive operator on $\mathbb{X}$, and $U(t,s)(\phi)(x) > 0$ for all $0 \leq s < t$, $x \in \mathbb{R}$, and $\phi \in \mathbb{X}$ provided that $\phi(x) \geq 0$ and $\phi \not\equiv 0$.

Define $B : [0,\infty) \times C \to \mathbb{X}$ by $B(t,\phi) := -g(t,\phi(0,\cdot)) + b(t)\int_{\mathbb{R}} \Gamma(a(t),\cdot - y)f_{-\tau}(t,\phi(-\tau,y))dy$ for any $t \in [0,\infty)$, $\phi \in C$. Then (2.1) becomes

$$\begin{cases} \partial_t u(t,x) = d(t)\Delta u + B(t,u_t), \quad t > 0, \\ u(t,x) = \phi(t,x), \quad t \in [-\tau,0], \ x \in \mathbb{R}, \end{cases} \tag{2.6}$$

which can be written as an integral equation

$$u(t,\cdot,\phi) = U(t,0)\phi(0,\cdot) + \int_0^t U(t,s)B(s,u_s)ds, \quad t \geq 0, \quad \phi \in C, \tag{2.7}$$

whose solutions are called mild solutions to (2.6).

THEOREM 2.3. *Let* (H1)–(H4) *hold. For any* $\phi \in C_{\hat{L}}$, *system* (2.1) *has a unique mild solution* $u(t,x,\phi)$ *with* $u_0(\cdot,\cdot,\phi) = \phi$ *and* $u_t(\cdot,\cdot,\phi) \in C_{\hat{L}}$ *for all* $t \geq 0$, *and* $u(t,x,\phi)$ *is a classic solution when* $t > \tau$. *Moreover, if* $\hat{u}(t,x)$ *and* $\bar{u}(t,x)$ *are a pair of lower and upper solutions of* (2.1), *respectively, with* $\hat{u}_0(\cdot,\cdot) \leq \bar{u}_0(\cdot,\cdot)$, *then* $\hat{u}_t(\cdot,\cdot) \leq \bar{u}_t(\cdot,\cdot)$ *for all* $t \geq 0$.

Proof. We first show that B is quasi-monotone on $[0,\infty) \times C_{\hat{L}}$ in the sense that

$$\lim_{h \to 0^+} d(\psi(0,\cdot) - \phi(0,\cdot) + h[B(t,\psi) - B(t,\phi)], \mathbb{X}_+) = 0 \tag{2.8}$$

for all $\phi, \psi \in C_{\hat{L}}$ with $\phi(s,x) \leq \psi(s,x)$ for all $s \in [-\tau,0]$, $x \in \mathbb{R}$. In fact, for any $\phi, \psi \in C_{\hat{L}}$ with $\phi(s,x) \leq \psi(s,x)$ for all $(s,x) \in [-\tau,0] \times \mathbb{R}$, we have

$$\begin{aligned} &\psi(0,\cdot) - \phi(0,\cdot) + h[B(t,\psi) - B(t,\phi)] \\ &= \psi(0,\cdot) - \phi(0,\cdot) + h[-(g(t,\psi(0,\cdot)) - g(t,\phi(0,\cdot)))] \\ &+ h\left[\int_{\mathbb{R}} \Gamma(a(t),\cdot - y)b(t)(f_{-\tau}(t,\psi(-\tau,y)) - f_{-\tau}(t,\phi(-\tau,y)))dy\right] \\ &\geq \psi(0,\cdot) - \phi(0,\cdot) - h(g(t,\psi(0,\cdot)) - g(t,\phi(0,\cdot))) \\ &\geq (1 - hl_1)(\psi(0,\cdot) - \phi(0,\cdot)). \end{aligned}$$

Thus, for $1 - hl_1 > 0$, $\psi(0,\cdot) - \phi(0,\cdot) + h[B(t,\psi) - B(t,\phi)] \in \mathbb{X}_+$, and hence, (2.8) holds. Then by [13, Corollary 5] (for $v^- = \hat{0}$, $v^+ = \hat{L}$, $S^+ = S^- = S = T \equiv U$,

$B^+ = B^- = B$), (2.1) admits a unique mild solution $u(t,\cdot,\phi)$ on $[-\tau,\infty)$ for any $\phi \in C_{\widehat{L}}$ and $u_t(\cdot,\cdot,\phi) \in C_{\widehat{L}}$ for all $t \geq 0$. Moreover, the comparison principle holds for lower and upper solutions. □

In order to study the spreading speed and traveling waves, we introduce the assumptions in [10, 11]. Let $u \in C$. Define the reflection operator $\mathcal{R}$ by $\mathcal{R}[u](\theta,x) := u(\theta,-x)$ for all $\theta \in [-\tau,0], x \in \mathbb{R}$. Given $y \in \mathbb{R}$, define the translation operator T_y by $T_y[u](\theta,x) := u(\theta,x-y)$ for all $\theta \in [-\tau,0], x \in \mathbb{R}$. Let $Q : C_{b^*} \to C_{b^*}$ be a map, where $b^* \in \mathbb{Y}$ with $b^* \gg 0$. Assume the following:

(A1) $Q[\mathcal{R}[u]] = \mathcal{R}[Q[u]]$, $T_y[Q[u]] = Q[T_y[u]]$ $\forall y \in \mathbb{R}$.

(A2) $Q : C_{b^*} \to C_{b^*}$ is continuous with respect to the compact open topology.

(A4) $Q : C_{b^*} \to C_{b^*}$ is monotone in the sense that $Q[u] \geq Q[v]$ whenever $u \geq v$ in C_{b^*}.

(A5) $Q : \mathbb{Y}_{b^*} \to \mathbb{Y}_{b^*}$ admits exactly two fixed points 0 and b^*, and for any positive number ε, there is an $\alpha \in \mathbb{Y}_{b^*}$ with $\| \alpha \|_{\mathbb{Y}} < \varepsilon$ such that $Q[\alpha] \gg \alpha$.

(A6) One of the following two conditions holds:

(a) $Q[C_{b^*}]$ is precompact in C_{b^*}.

(b′) The set $Q[C_{b^*}](0,\cdot)$ is precompact in $\mathbb{X}$, and there is a positive number $\varsigma \leq \tau$ such that $Q[u](\theta,x) = u(\theta+\varsigma,x)$ for $-\tau \leq \theta \leq -\varsigma$, the operator

$$S[u](\theta,x) := \begin{cases} u(0,x), & -\tau \leq \theta < -\varsigma, \\ Q[u](\theta,x), & -\varsigma \leq \theta \leq 0 \end{cases}$$

is continuous on C_{b^*}, and $S[D](\cdot,0)$ is precompact in $\mathbb{Y}$ for any T-invariant set $D \subseteq C_{b^*}$ with $D(0,\cdot)$ being precompact in $\mathbb{X}$. A set $W \subseteq C_{b^*}$ is said to be T-invariant if $T_y W = W$ for all $y \in \mathbb{R}$.

Recall that a family of operators $\{\Phi_t\}_{t\geq 0}$ is an ω-periodic semiflow on a metric space (X,ρ) with the metric ρ, provided that $\{\Phi_t\}$ satisfies

(i) $\Phi_0(v) = v$ $\forall v \in X$;

(ii) $\Phi_t(\Phi_\omega(v)) = \Phi_{t+\omega}(v)$ $\forall t \geq 0, v \in X$;

(iii) $\Phi(t,v) = \Phi_t(v)$ is continuous in (t,v) on $[0,+\infty) \times X$.

Define a family of operators $\{Q_t\}_{t\geq 0}$ on $C_{\widehat{L}}$ by

$$Q_t(\phi)(s,x) = u(t+s,x,\phi) \quad \forall t \geq 0, \quad s \in [-\tau,0], \quad x \in \mathbb{R}, \quad \phi \in C_{\widehat{L}},$$

where $u(t,x,\phi)$ is the mild solution of (2.1) with $u(s,x) = \phi(s,x)$ for $s \in [-\tau,0]$, $x \in \mathbb{R}$. Note that for any $(t_0,\phi_0) \in \mathbb{R}_+ \times C_{\widehat{L}}$, we have

$$\| Q_t(\phi) - Q_{t_0}(\phi_0) \|_C \leq \| Q_t(\phi) - Q_t(\phi_0) \|_C + \| Q_t(\phi_0) - Q_{t_0}(\phi_0) \|_C.$$

Note that $U(t,0)\varphi$ is continuous in $(t,\varphi) \in [0,\infty) \times \mathbb{X}$ with respect to the compact open topology. By a similar argument as in [12, Theorem 8.5.2], it follows that $Q_t(\phi)$ is continuous at (t_0,ϕ_0) with respect to the compact open topology. Thus, $\{Q_t\}_{t\geq 0}$ is an ω-periodic semiflow on $C_{\widehat{L}}$.

LEMMA 2.4. *For each $t > 0$, Q_t is strictly subhomogeneous.*

Proof. For any $\phi \in C_{\widehat{L}}$ with $\phi \not\equiv 0$, let $u(t,x,\phi)$ be the solution of (2.1) with $u(s,x) = \phi(s,x)$ for $s \in [-\tau,0]$, $x \in \mathbb{R}$. Fix $k \in (0,1)$. Since $G(t,u,v)$ is strictly subhomogeneous in (u,v), we have

$$\begin{aligned} &\partial_t(ku(t,x)) \\ &= d(t)\Delta(ku) - kg(t,u(t,x)) + kb(t)\int_{\mathbb{R}} \Gamma(a(t),x-y) f_{-\tau}(t,u(t-\tau,y))dy \\ &\leq d(t)\Delta(ku) - g(t,ku(t,x)) + b(t)\int_{\mathbb{R}} \Gamma(a(t),x-y) f_{-\tau}(t,ku(t-\tau,y))dy. \end{aligned}$$

Thus, $ku(t,x,\phi)$ is a lower solution of (2.1) with $ku(s,x,\phi) = k\phi(s,x)$ for $s \in [-\tau,0], x \in \mathbb{R}$. Then, $ku(t,x,\phi) \le u(t,x,k\phi)$ for $t \ge 0$, where $u(t,x,k\phi)$ is the solution of (2.1) with $u(s,x,k\phi) = k\phi(s,x)$ for $(s,x) \in [-\tau,0] \times \mathbb{R}$.

Let $w(t,x) = u(t,x,k\phi) - ku(t,x,\phi)$. Then $w(s,x) = 0$ for $(s,x) \in [-\tau,0] \times \mathbb{R}$ and $w(s,x) \ge 0$ for $(s,x) \in [-\tau,\infty) \times \mathbb{R}$. We further show that $w(t,x) > 0$ for all $t > 0, x \in \mathbb{R}$. For simplicity, we write $\widetilde{F}(t,u(t,x),v(t,x)) = -g(t,u(t,x)) + b(t)\int_{\mathbb{R}} \Gamma(a(t),x-y) f_{-\tau}(t,v(t,y))dy$. It follows that

(2.9)

$$
\begin{aligned}
&\frac{\partial w(t,x)}{\partial t} \\
&= \frac{\partial u(t,x,k\phi)}{\partial t} - k\frac{\partial u(t,x,\phi)}{\partial t} \\
&= d(t)\Delta u(t,x,k\phi) + \widetilde{F}(t,u(t,x,k\phi),u(t-\tau,x,k\phi)) \\
&-k[d(t)\Delta u(t,x,\phi) + \widetilde{F}(t,u(t,x,\phi),u(t-\tau,x,\phi))] \\
&= d(t)\Delta w(t,x) + [\widetilde{F}(t,u(t,x,k\phi),u(t-\tau,x,k\phi)) - \widetilde{F}(t,ku(t,x,\phi),ku(t-\tau,x,\phi))] \\
&+[\widetilde{F}(t,ku(t,x,\phi),ku(t-\tau,x,\phi)) - k\widetilde{F}(t,u(t,x,\phi),u(t-\tau,x,\phi))] \\
&= d(t)\Delta w(t,x) - g(t,u(t,x,k\phi)) + g(t,ku(t,x,\phi)) + h(t,x) \\
&+b(t)\int_{\mathbb{R}} \Gamma(a(t),x-y)[f_{-\tau}(t,u(t-\tau,y,k\phi)) - f_{-\tau}(t,ku(t-\tau,y,\phi))]dy \\
&\ge d(t)\Delta w(t,x) - g(t,u(t,x,k\phi)) + g(t,ku(t,x,\phi)) + h(t,x) \\
&\ge d(t)\Delta w(t,x) - l_1 w(t,x) + h(t,x),
\end{aligned}
$$

where $h(t,x) = \widetilde{F}(t,ku(t,x,\phi),ku(t-\tau,x,\phi)) - k\widetilde{F}(t,u(t,x,\phi),u(t-\tau,x,\phi))$. Let $\widetilde{U}(t,s) : \mathbb{X} \to \mathbb{X}$, $0 \le s \le t$, be the evolution operator of

$$
\begin{cases}
\partial_t u(t,x) = d(t)\Delta u - l_1 u(t,x), & t > 0, \\
u(0,x) = \psi(x), & x \in \mathbb{R}, \psi \in \mathbb{X}.
\end{cases}
$$

Then $\widetilde{U}(t,s)(\psi)(x) = e^{-l_1(t-s)}U(t,s)(\psi)(x)$ for all $t \ge s \ge 0$, $x \in \Omega$, $\psi \in C$, where $U(t,s)$ is the evolution operator of (2.4). Thus, the equation

$$
\begin{cases}
\partial_t u(t,x) = d(t)\Delta u - l_1 u(t,x) + h(t,x), \\
u(0,x) = \psi(x), \quad x \in \mathbb{R}, \psi \in \mathbb{X}
\end{cases} \tag{2.10}
$$

can be written as

$$
u(t,x,\psi) = \widetilde{U}(t,0)(\psi)(x) + \int_0^t \widetilde{U}(t,s)h(s,x)ds, \quad t \ge 0, x \in \mathbb{R}, \psi \in C. \tag{2.11}
$$

By (H2), we have $h(t,x) > 0$ for all $t > 0, x \in \mathbb{R}$. It then follows from (2.11) and the property of $\widetilde{U}(t,s)$ that for any $\psi \ge 0$ with $\psi \not\equiv 0$, the solution of (2.10) satisfies $u(t,x,\psi) > 0$ for all $t > 0, x \in \mathbb{R}$. Then by (2.9) and the comparison principle, we have $w(t,x) > 0$ for all $t > 0, x \in \mathbb{R}$. Therefore, $u(t,x,k\phi) > ku(t,x,\phi)$ for all $t > 0, x \in \mathbb{R}$, and hence, $Q_t(k\phi) > kQ_t(\phi)$ for all $t > 0$, which indicates that for each $t > 0$, Q_t is strictly subhomogeneous. ∎

LEMMA 2.5. *For any $\varphi \in C_{\widehat{L}}$ with $\varphi \not\equiv 0$, $u(t,x,\varphi) > 0$ for all $t \geq \tau$, $x \in \mathbb{R}$.*

Proof. Let $\varphi \in C_{\widehat{L}}$ with $\varphi \not\equiv 0$. By Theorem 2.3, $u(t,x,\varphi) \geq 0$ for all $t \geq 0$ and $x \in \mathbb{R}$. It follows from (H1) that for any $t > 0$, $u(t,x,\varphi)$ satisfies

$$\begin{aligned}\partial_t u(t,x) &= d(t)\Delta u - g(t,u(t,x,\varphi)) + b(t)\int_{\mathbb{R}} \Gamma(a(t),x-y) f_{-\tau}(t,u(t-\tau,y,\varphi))\\ &\geq d(t)\Delta u - g(t,u(t,x,\varphi))\\ &\geq d(t)\Delta u - l_1 u(t,x,\varphi).\end{aligned}$$

By [19, Theorem 5.5.4], $u(t,x,\varphi) > 0$ for all $t > 0$, $x \in \mathbb{R}$, provided that $\varphi(0,\cdot) > 0$.

Now we show that for any $\varphi \in C_{\widehat{L}}$ with $\varphi \not\equiv 0$ and $\varphi(0,\cdot) = 0$, there exists $t_0 = t_0(\varphi) \in [0,\tau]$ such that $u(t_0,\cdot,\varphi) > 0$. Assume, by contradiction, that for some $\varphi \in C_{\widehat{L}}$ with $\varphi \not\equiv 0$ and $\varphi(0,\cdot) = 0$ we have $u(t,\cdot,\varphi) \equiv 0$ for all $t \in [0,\tau]$. It follows from (2.7) that

$$0 = \int_0^t U(t,s)b(s)\int_{\mathbb{R}} \Gamma(a(s),x-y) f_{-\tau}(s,u_s(-\tau,y))dyds, \quad t \in [0,\tau],$$

which implies that $\int_{\mathbb{R}} \Gamma(a(s),x-y) f_{-\tau}(s,u_s(-\tau,y))dy = 0$ for any $s \in [0,\tau]$, and hence, $f_{-\tau}(s,u_s(-\tau,y)) = 0$ for any $s \in [0,\tau]$, $y \in \mathbb{R}$. Then by (H1), $u_s(-\tau,y) = 0$ for any $s \in [0,\tau]$, $y \in \mathbb{R}$. That is, $\varphi \equiv 0$, a contradiction. Thus, we have $u(t_0,\cdot,\varphi) > 0$ for some $t_0 = t_0(\varphi) \in [0,\tau]$. Then for any $t > t_0$, $\widetilde{U}(t,t_0)[u(t_0,\cdot,\varphi)](x) = e^{-l_1(t-t_0)}U(t,t_0)[u(t_0,\cdot,\varphi)](x) > 0$, and hence, by the comparison principle, we have $u(t,x,\varphi) > 0$ for all $t > t_0$, $x \in \mathbb{R}$.

Therefore, for any $\varphi \in C_{\widehat{L}}$ with $\varphi \not\equiv 0$, $u(t,x,\varphi) > 0$ for all $t > \tau, x \in \mathbb{R}$. ∎

LEMMA 2.6. *For any $t > 0$, Q_t satisfies* (A1), (A2), (A4), *and* (A6) *with $b^* = \widehat{L}$, and Q_ω satisfies* (A5) *with $b^* = \beta_0^*$, where $\beta_0^* \in \mathbb{Y}_{\widehat{L}}$ with $\beta_0^*(s) = \beta^*(s)$ for all $s \in [-\tau,0]$.*

Proof. It is easy to see that Q_t satisfies (A1), (A2), and (A4) with $b^* = \widehat{L}$ for any $t > 0$.

Let $\hat{Q}_t = Q_t|_{\mathbb{Y}_{\widehat{L}}}$. Then $\hat{Q}_t : \mathbb{Y}_{\widehat{L}} \to \mathbb{Y}_{\widehat{L}}$ is the ω-periodic semiflow generated by (2.2). Moreover, it is not difficult to see that $\hat{Q}_t$ is strictly monotone for any $t \geq \tau$ and strongly monotone for any $t \geq 2\tau$ on $\mathbb{Y}_{\widehat{L}}$. Note that (2.2) has a positive ω-periodic solution $\beta^*(t)$ which is globally asymptotically stable in $\mathbb{Y}_{\widehat{L}} \setminus \{0\}$. We see that $\hat{Q}_\omega$ has only two fixed points 0 and β_0^* in $\mathbb{Y}_{\widehat{L}}$, where $\beta_0^*(s) = \beta^*(s)$ for all $s \in [-\tau,0]$. Thus, by the Dancer–Hess connecting orbit lemma (see, e.g., [23]), the map $\hat{Q}_\omega$ admits a strictly monotone full orbit $\{\varphi_n\}_{-\infty}^{\infty} \subseteq \mathbb{Y}_{\beta_0^*}$ connecting 0 to β_0^* and $\varphi_n < \varphi_{n+1}$ for any $n = 0, \pm 1, \pm 2, \ldots$. For any $\bar{n} \in \mathbb{N}$ such that $\bar{n}\omega \geq 2\tau$, since $\hat{Q}_{\bar{n}\omega}$ is strongly monotone, we have $\hat{Q}_{\bar{n}\omega}(\varphi_n) = \hat{Q}_\omega^{\bar{n}}(\varphi_n) \ll \hat{Q}_\omega^{\bar{n}}(\varphi_{n+1}) = \hat{Q}_{\bar{n}\omega}(\varphi_{n+1})$ for any $n = 0, \pm 1, \pm 2, \ldots$. That is, $\varphi_{n+\bar{n}\omega} \ll \varphi_{n+1+\bar{n}\omega}$ for any $n = 0, \pm 1, \pm 2, \ldots$. Therefore, $\varphi_n \ll \varphi_{n+1}$ for any $n = 0, \pm 1, \pm 2, \ldots$, and hence, Q_ω satisfies (A5) with $b^* = \beta_0^*$.

Now we show that Q_t satisfies (A6)(a) with $b^* = \widehat{L}$ for $t > \tau$. Fix $t_0 > \tau$ and set $a = t_0 - \tau$, $b = t_0$. Let $u(t,\varphi)$ be the solution of (2.1) with $u_0(\varphi) = \varphi \in C_{\widehat{L}}$ and define the Kuratowski measure of noncompactness of a subset A of $\mathbb{X}$ as

$$\alpha(A) = \inf\{r > 0 : A \text{ has a finite cover of diameter } \leq r\}.$$

First we prove that $\overline{\{u(t,\varphi): a\le t\le b, \varphi\in C_{\widehat{L}}\}}$ is compact in $\mathbb{X}$. By (2.7), for any $\epsilon\in(0,a), t\in[a,b]$, and $\varphi\in C_{\widehat{L}}$, we have

$$\begin{aligned}
&u(t,\varphi)\\
&= U(t,0)\varphi(0,\cdot)+\int_0^{t-\epsilon} U(t,s)B(s,u_s)ds+\int_{t-\epsilon}^{t} U(t,s)B(s,u_s)ds\\
&= U(t,t-\epsilon)\left[U(t-\epsilon,0)\varphi(0,\cdot)+\int_0^{t-\epsilon} U(t-\epsilon,s)B(s,u_s)ds\right]+\int_{t-\epsilon}^{t} U(t,s)B(s,u_s)ds\\
&= U(t,t-\epsilon)u(t-\epsilon,\varphi)+\int_{t-\epsilon}^{t} U(t,s)B(s,u_s)ds.
\end{aligned}$$

Since $\{u(t-\epsilon,\varphi), t\in[a,b], \varphi\in C_{\widehat{L}}\}$ is bounded in $\mathbb{X}_+$ and $U(t,t-\epsilon)$ is compact, we have

$$\alpha(\{U(t,t-\epsilon)u(t-\epsilon,\varphi), t\in[a,b], \varphi\in C_{\widehat{L}}\})=0.$$

It is easy to see $\{U(t,s)B(s,u_s): t\in[a,b], s\in[0,t], \varphi\in C_{\widehat{L}}\}$ is bounded in $\mathbb{X}_+$. Let $N>0$ such that $\|\ U(t,s)B(s,u_s)\ \|_{\mathbb{X}}\le N$ for all $t\in[a,b], s\in[0,t], \varphi\in C_{\widehat{L}}$. By the fact of $\alpha(A)\le\delta(A)$, where $\delta(A)$ is the diameter of $A\subseteq\mathbb{X}$, we have

$$\alpha\left(\left\{\int_{t-\epsilon}^{t} U(t,s)B(s,u_s)ds: t\in[a,b], s\in[t-\epsilon,t], \varphi\in C_{\widehat{L}}\right\}\right)\le 2\epsilon N.$$

Thus,

$$\begin{aligned}
&\alpha(\{u(t,\varphi): t\in[a,b], \varphi\in C_{\widehat{L}}\})\\
&\le \alpha(\{U(t,t-\epsilon)u(t-\epsilon,\varphi), t\in[a,b], \varphi\in C_{\widehat{L}}\})\\
&\quad+\alpha\left(\left\{\int_{t-\epsilon}^{t} U(t,s)B(s,u_s)ds: t\in[a,b], s\in[t-\epsilon,t], \varphi\in C_{\widehat{L}}\right\}\right)\\
&\le 2\epsilon N.
\end{aligned}$$

Letting $\epsilon\to 0$, we have $\alpha(\{u(t,\varphi): t\in[a,b], \varphi\in C_{\widehat{L}}\})=0$, and hence, $\{u(t,\varphi): t\in[a,b], \varphi\in C_{\widehat{L}}\}$ is precompact in $\mathbb{X}$.

Given a compact interval $I\subseteq\mathbb{R}$, let $K=\min\{K_1>0: I\subseteq[-K_1,K_1]\}$. Since $\{u(t,\varphi): t\in[a,b], \varphi\in C_{\widehat{L}}\}$ is precompact in $\mathbb{X}$, $\{u(t,\varphi)|_I: t\in[a,b],\ \varphi\in C_{\widehat{L}}\}$ is equicontinuous in $\mathbb{X}$, and hence, for any $\epsilon>0$, there exists $\delta>0$ such that

$$|u(t,x_1,\varphi)-u(t,x_2,\varphi)|<\epsilon \tag{2.12}$$

for all $t\in[a,b]$ and $\varphi\in C_{\widehat{L}}$, provided that $x_1,x_2\in I$ and $|x_1-x_2|<\delta$.

Let $[a_1,b_1]$ be any bounded interval on $\mathbb{R}$ with $a_1>0$ and let $U_0(t)$ be the semigroup generated by $u_t=\Delta u$. Then $U_0(t)\varphi(x)=\int_{-\infty}^{+\infty}\Gamma(t,x-y)\varphi(y)dy$ for all $t>0$, $x\in\mathbb{R}$, $\varphi\in\mathbb{X}$.

By the properties of Γ, we can find an $N_0>0$ such that $\int_{|y|\ge N_0}\Gamma(b_1,y)dy\le\epsilon$. Since $\frac{\partial\Gamma(t,y)}{\partial t}>0$ for all $t>0$ and $y^2>2t$, we have $\int_{|y|\ge N_1}\Gamma(t,y)dy\le\epsilon$ for all $t\in[a_1,b_1]$, where $N_1=\max\{N_0,\sqrt{2b_1}\}$. Moreover, since $\int_{-N_1}^{N_1}\Gamma(t,y)dy$ is continuous in $t\in[a_1,b_1]$, there is a $\delta_1>0$ such that $|\int_{-N_1}^{N_1}(\Gamma(t_1,y)-\Gamma(t_2,y))dy|<\epsilon$ provided that

$t_1, t_2 \in [a_1, b_1]$ and $|t_1 - t_2| < \delta_1$. Therefore, for any $t_1, t_2 \in [a_1, b_1]$ and $|t_1 - t_2| < \delta_1$, $\psi \in \mathbb{X}_{\widehat{L}}, x \in I$,

$$\begin{aligned}
&|(U_0(t_1)\psi)(x) - (U_0(t_2)\psi)(x)| \\
&= \left| \int_{\mathbb{R}} \Gamma(t_1, x-y)\psi(y)dy - \int_{\mathbb{R}} \Gamma(t_2, x-y)\psi(y)dy \right| \\
&= \left| \int_{\mathbb{R}} (\Gamma(t_1, y) - \Gamma(t_2, y))\psi(x-y)dy \right| \\
&\le \left| \int_{|y|\le N_1} (\Gamma(t_1, y) - \Gamma(t_2, y))\psi(x-y)dy \right| + \left| \int_{|y|\ge N_1} (\Gamma(t_1, y) - \Gamma(t_2, y))\psi(x-y)dy \right| \\
&< 2\epsilon L.
\end{aligned}$$

It follows from the continuity of $\eta(t)$ in $t \in \mathbb{R}_+$ and definitions of $U_0(t)$ and $U(t,s)$ that there exists $\delta_2 > 0$ such that

$$|(U(t_1, 0)\varphi(0, \cdot))(x) - (U(t_2, 0)\varphi(0, \cdot))(x)| < 2\epsilon L$$

for all $x \in I$, $\varphi \in C_{\widehat{L}}$, provided that $t_1, t_2 \in [a, b]$ and $|t_1 - t_2| < \delta_2$. Let $\bar{\delta} \in (0, \min\{\epsilon, \delta_2\})$. Then for $x \in I$, $\varphi \in C_{\widehat{L}}$, $t_1, t_2 \in [a, b]$, and $|t_1 - t_2| < \bar{\delta}$, we have

$$\begin{aligned}
&|u(t_1, x, \varphi) - u(t_2, x, \varphi)| \\
&\le |(U(t_1, 0)\varphi(0, \cdot))(x) - (U(t_2, 0)\varphi(0, \cdot))(x)| \\
&\quad + \left| \int_0^{t_1} (U(t_1, s)B(s, u_s))(x)ds - \int_0^{t_2} (U(t_2, s)B(s, u_s))(x)ds \right| \\
&\le 2L\epsilon + 2N \cdot 2^K \bar{\delta} \\
&\le 2(L + 2^K N)\epsilon,
\end{aligned} \tag{2.13}$$

where N was defined in the former paragraph of this proof. This implies that $u(t, x, \varphi)$ is equicontinuous in $t \in [a, b]$ for $x \in I$ and $\varphi \in C_{\widehat{L}}$.

Consequently, by (2.12) and (2.13), for any $\varphi \in C_{\widehat{L}}$, $\theta_1, \theta_2 \in [-\tau, 0]$, $x_1, x_2 \in I$ with $|\theta_1 - \theta_2| < \bar{\delta}$ and $|x_1 - x_2| < \delta$, we have

$$\begin{aligned}
&|u_{t_0}(\varphi)(\theta_1, x_1) - u_{t_0}(\varphi)(\theta_2, x_2)| \\
&= |u(t_0 + \theta_1, x_1, \varphi) - u(t_0 + \theta_2, x_2, \varphi)| \\
&\le |u(t_0 + \theta_1, x_1, \varphi) - u(t_0 + \theta_1, x_2, \varphi)| + |u(t_0 + \theta_1, x_2, \varphi) - u(t_0 + \theta_2, x_2, \varphi)| \\
&\le (2L + 2^{K+1}N + 1)\epsilon,
\end{aligned}$$

which indicates that $\{u_{t_0}(\varphi) : \varphi \in C_{\widehat{L}}\}$ is equicontinuous for $(\theta, x) \in [-\tau, 0] \times I$. Therefore, $\{u_{t_0}(\varphi) : \varphi \in C_{\widehat{L}}\}$ is precompact in $C_{\widehat{L}}$ and (A6)(a) follows from $Q_{t_0}(C_{\widehat{L}}) = \{u_{t_0}(\varphi) : \varphi \in C_{\widehat{L}}\}$ for $t_0 > \tau$.

Finally, we show that Q_t satisfies (A6)(b′) with $b^* = \widehat{L}$ for $0 < t \le \tau$. Fix $t_1 \in (0, \tau]$ and define

$$S[\varphi](\theta, x) := \begin{cases} \varphi(0, x), & -\tau \le \theta < -t_1, \\ Q_{t_1}(\varphi)(\theta, x), & -t_1 \le \theta \le 0 \end{cases}$$

for any $\varphi \in C_{\widehat{L}}$. By the above analysis, we know that $\{u(t,\varphi) : a \le t \le b, \varphi \in C_{\widehat{L}}\}$ is precompact in $\mathbb{X}$ for any $0 < a \le b$. In particular, fixing $a = b = t_1$, we can easily see that $\{u_{t_1}(\varphi)(0,\cdot), \varphi \in C_{\widehat{L}}\} = \{u(t_1,\cdot,\varphi), \varphi \in C_{\widehat{L}}\}$ is precompact in $\mathbb{X}$, that is, $Q_{t_1}[C_{\widehat{L}}](0,\cdot)$ is precompact in $\mathbb{X}$.

Since Q_t is an ω-periodic semiflow, it is easy to see that $S[\varphi]$ is continuous on $C_{\widehat{L}}$. Let D be a T-invariant subset of $C_{\widehat{L}}$ (i.e., $T_y D = D$ for all $y \in \mathbb{R}$) with $D(0,\cdot)$ being precompact in $\mathbb{X}$. Now we show that for any given compact interval $I \subseteq \mathbb{R}$, $S[D]$ is equicontinuous on $[-\tau, 0] \times I$, that is, for any $\epsilon > 0$, there exist $\delta_1, \delta_2 > 0$ such that $|S[\varphi](\theta_1, x_1) - S[\varphi](\theta_2, x_2)| < \epsilon$ for any $\varphi \in D$ if $\theta_1, \theta_2 \in [-\tau, 0], x_1, x_2 \in I$, and $|\theta_1 - \theta_2| < \delta_1, |x_1 - x_2| < \delta_2$.

Since $S[\varphi](\theta, x) = \varphi(0, x)$ for all $\varphi \in D$, $\theta \in [-\tau, -t_1]$, $x \in I$, and $D(0,\cdot)$ is precompact in $\mathbb{X}$, it is obvious that $S[D]$ is equicontinuous on $[-\tau, -t_1] \times I$.

Note that there exists $N > 0$ such that $\| U(t,s)B(s,u_s) \|_{\mathbb{X}} \le N$ for all $t \in [0, t_1], s \in [0,t], \varphi \in C_{\widehat{L}}$. Let $\delta_0 = \min\{\epsilon/(2^K N), t_1\}$. Then for any $t < \delta_0$, $x \in I$, and $\varphi \in D$, we have

$$\left| \int_0^t U(t,s)B(s,u_s)(x)ds \right| < 2^K N \delta_0 = \epsilon. \tag{2.14}$$

Let $\mathcal{F}(t,\psi) := U(t,0)\psi$ for $(t,\psi) \in [0,\delta_0] \times D(0,\cdot)$. Then $\mathcal{F}$ is continuous on $[0,\delta_0] \times D(0,\cdot)$ and $\mathcal{F}([0,\delta_0] \times D(0,\cdot))$ is precompact in $\mathbb{X}$. Thus, for the above I, there exists $\delta_2 > 0$ such that for $x_1, x_2 \in I$ and $|x_1 - x_2| < \delta_2$, we have

$$|U(t,0)\psi(x_1) - U(t,0)\psi(x_2)| < \epsilon \quad \forall t \in [0,\delta_0], \psi \in D(0,\cdot). \tag{2.15}$$

Moreover, since $\mathcal{F}$ is uniformly continuous on $[0,\delta_0] \times D(0,\cdot)$, there exists $\delta_1 > 0$, $\delta_3 > 0$ such that $\| \mathcal{F}(\bar{t}_1, \psi_1) - \mathcal{F}(\bar{t}_2, \psi_2) \|_{\mathbb{X}} < \epsilon/2^K$ if $\bar{t}_1, \bar{t}_2 \in [0,\delta_0]$, $\psi_1, \psi_2 \in D(0,\cdot)$, and $|\bar{t}_1 - \bar{t}_2| < \delta_1, ||\psi_1 - \psi_2||_{\mathbb{X}} < \delta_3$. In particular, we have $\| U(\bar{t}_1, 0)\psi - U(\bar{t}_2, 0)\psi \|_{\mathbb{X}} < \epsilon/2^K$ if $\bar{t}_1, \bar{t}_2 \in [0,\delta_0]$, $\psi \in D(0,\cdot)$, and $|\bar{t}_1 - \bar{t}_2| < \delta_1$. Then

$$|U(\bar{t}_1, 0)\psi(x) - U(\bar{t}_2, 0)\psi(x)| < \epsilon \quad \forall \psi \in D(0,\cdot), x \in I, \bar{t}_1, \bar{t}_2 \in [0,\delta_0], \text{ and } |\bar{t}_1 - \bar{t}_2| < \delta_1. \tag{2.16}$$

By (2.14)–(2.16), we can easily obtain that if $\theta_1, \theta_2 \in [-t_1, \delta_0 - t_1], x_1, x_2 \in I$ and $|\theta_1 - \theta_2| < \delta_1, |x_1 - x_2| < \delta_2$, then for any $\varphi \in D$,

$$\begin{aligned}
&|S[\varphi](\theta_1, x_1) - S[\varphi](\theta_2, x_2)| \\
&= |Q_{t_1}[\varphi](\theta_1, x_1) - Q_{t_1}[\varphi](\theta_2, x_2)| \\
&= |u(t_1 + \theta_1, x_1, \varphi) - u(t_1 + \theta_2, x_2, \varphi)| \\
&\le |(U(t_1 + \theta_1, 0)\varphi(0,\cdot))(x_1) - (U(t_1 + \theta_2, 0)\varphi(0,\cdot))(x_2)| \\
&\quad + \left| \int_0^{t_1+\theta_1} U(t_1 + \theta_1, s)B(s,u_s)(x_1)ds - \int_0^{t_1+\theta_2} U(t_1 + \theta_2, s)B(s,u_s)(x_2)ds \right| \\
&\le |(U(t_1 + \theta_1, 0)\varphi(0,\cdot))(x_1) - (U(t_1 + \theta_1, 0)\varphi(0,\cdot))(x_2)| \\
&\quad + |(U(t_1 + \theta_1, 0)\varphi(0,\cdot))(x_2) - (U(t_1 + \theta_2, 0)\varphi(0,\cdot))(x_2)| + 2\epsilon \\
&< 4\epsilon,
\end{aligned}$$

which implies that $S[D]$ is equicontinuous on $[-t_1, \delta_0 - t_1] \times I$.

By a similar argument as for (A6)(a), it is easy to see that S[D] is equicontinuous on $[\delta_0 - t_1, 0] \times I$.

Therefore, $S[D]$ is equicontinuous on $[-\tau, 0] \times I$, and hence, $S[D]$ is precompact in $C_{\widehat{L}}$. Thus, (A6)(b′) is valid for Q_t, $t \in (0, \tau]$. □

It then follows from Lemma 2.6 and [11, Theorems 2.11 and 2.15] that Q_ω has an asymptotic speed of spread $c^*_\omega > 0$.

Consider the linearized system of (2.1) at the zero solution:

(2.17)
$$\begin{cases} \partial_t u(t,x) = d(t)\Delta u - g_u(t,0)u(t,x) + b(t)\partial_u f_{-\tau}(t,0)\displaystyle\int_{\mathbb{R}} \Gamma(a(t), x-y)u(t-\tau,y)dy, \\ u(t,x) = \phi(t,x), \quad t \in [-\tau, 0],\ x \in \mathbb{R}. \end{cases}$$

For $\alpha > 0$, let $u(t,x) = e^{-\alpha x}v(t)$. Substituting $u(t,x)$ into (2.17) yields

$$\begin{aligned} &e^{-\alpha x}v'(t) \\ &= \dot{d}(t)\alpha^2 e^{-\alpha x}v(t) - g_u(t,0)v(t)e^{-\alpha x} + b(t)\partial_u f_{-\tau}(t,0)v(t-\tau)\textstyle\int_{\mathbb{R}} \Gamma(a(t),y)e^{-\alpha(x-y)}dy. \end{aligned}$$

Since $\Gamma(t,x)$ is even in x and by [18, Proposition 4.2], we obtain

(2.18)
$$\begin{aligned} v'(t) &= d(t)\alpha^2 v(t) - g_u(t,0)v(t) + b(t)\partial_u f_{-\tau}(t,0)v(t-\tau)\int_{\mathbb{R}} \Gamma(a(t),y)e^{\alpha y}dy, \\ &= d(t)\alpha^2 v(t) - g_u(t,0)v(t) + b(t)\partial_u f_{-\tau}(t,0)v(t-\tau)\int_{\mathbb{R}} \Gamma(a(t),y)e^{-\alpha y}dy, \\ &= d(t)\alpha^2 v(t) - g_u(t,0)v(t) + b(t)\partial_u f_{-\tau}(t,0)v(t-\tau)e^{\alpha^2 a(t)}. \end{aligned}$$

Then $u(t,x) = e^{-\alpha x}v(t)$ satisfies (2.17) with $\phi(s,x) = e^{-\alpha x}v(s)$ for $s \in [-\tau, 0]$ and $x \in \mathbb{R}$ if $v(t)$ satisfies (2.18) for $t \geq 0$.

Let M_t be the linear solution map defined by (2.17) and let $v(t, v_0)$ be the solution of (2.18) with $v(s, v_0) = v_0(s)$ for $s \in [-\tau, 0]$, $v_0 \in \mathbb{Y}$. Define $B^t_\alpha(v_0) := M_t(v_0 e^{-\alpha x})(0)$. It is not difficult to see that $B^t_\alpha(v_0) = v(t, v_0)$, and hence, B^t_α is the solution map associated with (2.18) on $\mathbb{Y}$.

Let $\gamma(\alpha)$ be the spectral radius of the Poincaré map associated with (2.18), and [21, Proposition 2.1] implies that $\gamma(\alpha) > 0$. It follows from the proof of [21, Proposition 2.1] that there exists a positive ω-periodic function $w(t)$ such that $v(t) = e^{\lambda(\alpha)t}w(t)$ is a solution of (2.18), where $\lambda(\alpha) = \frac{\ln \gamma(\alpha)}{\omega}$. Define $\psi \in \mathbb{Y}$ by $\psi(\theta) = e^{\lambda(\alpha)\theta}w(\theta)$ for all $\theta \in [-\tau, 0]$. Clearly, $v(t, \psi) = e^{\lambda(\alpha)t}w(t)$ for all $t \geq 0$. Then we have

$$B^t_\alpha(\psi)(\theta) = v(t+\theta, \psi) = e^{\lambda(\alpha)t}e^{\lambda(\alpha)\theta}w(t+\theta)\ \forall \theta \in [-\tau, 0], t \geq 0.$$

By the ω-periodicity of $w(t)$, it follows that

$$B^\omega_\alpha(\psi)(\theta) = e^{\lambda(\alpha)\omega}e^{\lambda(\alpha)\theta}w(\theta) = e^{\lambda(\alpha)\omega}\psi(\theta)\ \forall \theta \in [-\tau, 0],$$

that is, $B^\omega_\alpha(\psi) = e^{\lambda(\alpha)\omega}\psi$. This implies that $e^{\lambda(\alpha)\omega}$ is the principle eigenvalue of B^ω_α with positive eigenfunction ψ.

Let $\Phi(\alpha) := \frac{1}{\alpha}\ln e^{\lambda(\alpha)\omega} = \frac{\lambda(\alpha)\omega}{\alpha} = \frac{\ln \gamma(\alpha)}{\alpha}$. Then we have the following result.

PROPOSITION 2.7. *Assume that* (H1)–(H4) *hold. Let* c^*_ω *be the asymptotic speed of spread of* Q_ω. *Then* $c^*_\omega = \inf_{\alpha>0}\Phi(\alpha) = \inf_{\alpha>0}\frac{\ln\gamma(\alpha)}{\alpha}$.

Proof. When $\alpha = 0$, (2.18) becomes (2.3). It follows from (H4) that $\gamma(0) > 1$, and hence (C7) in [11] is satisfied. Now we prove that $\Phi(\infty) = \infty$. By (2.18), we have

$$v'(t) \geq [\alpha^2 d(t) - g_u(t,0)]v(t) \ \forall t \geq 0,$$

and hence,

$$\frac{w'(t)}{w(t)} \geq \alpha^2 d(t) - g_u(t,0) - \lambda(\alpha).$$

Then

$$0 = \int_0^\omega \frac{w'(t)}{w(t)}dt \geq \int_0^\omega (\alpha^2 d(t) - g_u(t,0))dt - \lambda(\alpha)\omega,$$

which implies that

$$\lambda(\alpha)\omega \geq \alpha^2 \int_0^\omega d(t)dt - \int_0^\omega g_u(t,0)dt.$$

Therefore,

$$\Phi(\alpha) = \frac{\lambda(\alpha)\omega}{\alpha} \geq \alpha \int_0^\omega d(t)dt - \frac{\int_0^\omega g_u(t,0)dt}{\alpha}.$$

Letting $\alpha \to \infty$, we can easily obtain $\Phi(\infty) = \infty$.

Since $G(t,\cdot,\cdot)$ is subhomogeneous in (u,v), it follows from [23, Lemma 2.3.2] that $G(t,u,v) \leq G_u(t,0,0)u + G_v(t,0,0)v$, that is,

$$-g(t,u) + b(t)f_{-\tau}(t,v) \leq -g_u(t,0)u + b(t)\partial_u f_{-\tau}(t,0)v,$$

and hence, we have

$$\begin{aligned}&-g(t,u(t,x)) + b(t)\int_{\mathbb{R}} \Gamma(a(t),x-y)f_{-\tau}(t,u(t-\tau,y))dy\\ &\leq -g_u(t,0)u(t,x) + b(t)\partial_u f_{-\tau}(t,0)\int_{\mathbb{R}} \Gamma(a(t),x-y)u(t-\tau,y)dy.\end{aligned}$$

By the comparison principle, we have $Q_\omega(\varphi) \leq M_\omega(\varphi)$ for any $\varphi \in C_{\beta^*_0}$. Thus, [11, Theorem 3.10] implies that $c^*_\omega \leq \inf_{\alpha>0}\Phi(\alpha)$.

Let $K > 0$ such that $K - g_u(t,0) > 0$ for all $t \in [0,\omega]$. Set $\bar{G}(t,u,v) = Ku + G(t,u,v)$. Then $\bar{G}_u(t,0,0) > 0$, $\bar{G}_v(t,0,0) > 0$ for all $t \in [0,\omega]$. It is easy to see that for any $\epsilon \in (0,1)$, there exists $\delta = \delta(\epsilon) \in (0,L)$ such that

$$\bar{G}(t,u,v) \geq (1-\epsilon)\bar{G}_u(t,0,0)u + (1-\epsilon)\bar{G}_v(t,0,0)v \ \forall (u,v) \in [0,\delta]^2,$$

and hence, for any $(u,v) \in [0,\delta]^2$,

$$G(t,u,v) = -Ku + \bar{G}(t,u,v) \geq [(1-\epsilon)G_u(t,0,0) - \epsilon K]u + (1-\epsilon)G_v(t,0,0)v.$$

Moreover, there exists $\xi = \xi(\delta) > 0$ such that for any $\varphi \in C_{\hat{\xi}}$, we have

$$0 \le u(t,x,\varphi) \le u(t,x,\hat{\xi}) < \delta \quad \forall x \in \mathbb{R}, t \in [0,\omega].$$

Thus, for any $\varphi \in C_{\hat{\xi}}$, $u(t,x,\varphi)$ satisfies

$$\begin{aligned}\partial_t u(t,x) \ge & d(t)\Delta u(t,x) + [(1-\epsilon)g_u(t,0) - \epsilon K]u(t,x) \\ & + (1-\epsilon)b(t)\partial_u f_{-\tau}(t,0)\int_{\mathbb{R}} \Gamma(a(t),x-y)u(t-\tau,y)dy \quad \forall t \in [0,\omega].\end{aligned}$$

Let $M_t^\epsilon, t \ge 0$, be the solution maps associated with the linear system

$$\begin{aligned}\partial_t u(t,x) = & d(t)\Delta u(t,x) + [(1-\epsilon)g_u(t,0) - \epsilon K]u(t,x) \\ & + (1-\epsilon)b(t)\partial_u f_{-\tau}(t,0)\int_{\mathbb{R}} \Gamma(a(t),x-y)u(t-\tau,y)dy \quad \forall t \in [0,\omega].\end{aligned}$$

The comparison principle implies that $M_t^\epsilon(\varphi) \le Q_t(\varphi)$ for all $\varphi \in C_{\hat{\xi}}, t \in [0,\omega]$. In particular, $M_\omega^\epsilon(\varphi) \le Q_\omega(\varphi)$ for all $\varphi \in C_{\hat{\xi}}$. By a similar analysis for M_t^ϵ as for M_t, it follows from [11, Theorem 3.10] that $\inf_{\alpha>0}\Phi_\epsilon(\alpha) \le c_\omega^*$.

Therefore, $\inf_{\alpha>0}\Phi_\epsilon(\alpha) \le c_\omega^* \le \inf_{\alpha>0}\Phi(\alpha)$ for all $\epsilon \in (0,1)$. Letting $\epsilon \to 0$, we have $c_\omega^* = \inf_{\alpha>0}\Phi(\alpha)$. □

Let $c^* = \frac{c_\omega^*}{\omega} = \frac{1}{\omega}\inf_{\alpha>0}\Phi(\alpha) = \frac{1}{\omega}\inf_{\alpha>0}\frac{\ln\gamma(\alpha)}{\alpha}$. The following result shows that c^* is the spreading speed of solutions of (2.1) with initial functions having compact support.

THEOREM 2.8. *Assume that* (H1)–(H4) *hold and let* $c^* = c_\omega^*/\omega$. *Then the following statements are valid.*

(1) *For any* $c > c^*$, *if* $\varphi \in C_{\beta_0^*}$ *with* $0 \le \varphi \ll \beta_0^*$ *and* $\varphi(\cdot,x) = 0$ *for* x *outside a bounded interval, then*

$$\lim_{t\to\infty, |x|\ge ct} u(t,x,\varphi) = 0.$$

(2) *For any* $c < c^*$, *if* $\varphi \in C_{\beta_0^*}$ *with* $\varphi \not\equiv 0$, *then*

$$\lim_{t\to\infty, |x|\le ct} (u(t,x,\varphi) - \beta^*(t)) = 0.$$

Proof. Conclusion (1) follows from [10, Theorem 2.1]. By Lemma 2.4 and [10, Theorem 2.1], for any $c < c^*$, there is a positive number σ such that, if $\varphi \in C_{\beta_0^*}$ with $\varphi(\cdot,x) > 0$ for x on an interval of length 2σ, then $\lim_{t\to\infty,|x|\le ct}(u(t,x,\varphi) - \beta^*(t)) = 0$. It follows from Lemma 2.5 that for any $\varphi \in C_{\beta_0^*}$ with $\varphi \not\equiv 0$, $Q_t(\varphi) \gg 0$ for all $t > 2\tau$. We can fix $t_0 > 2\tau$ and take $Q_{t_0}(\varphi)$ as a new initial value for $u(t,x,\varphi)$. Then by the above analysis, conclusion (2) is valid. □

We say $u(t,x) = \mathcal{U}(t,x-ct)$ is an ω-periodic traveling wave of (2.1) connecting $\beta^*(t)$ to 0 if it is a solution of (2.1), $\mathcal{U}(t,\xi)$ is ω-periodic in t, and $\mathcal{U}(t,-\infty) = \beta^*(t)$ and $\mathcal{U}(t,\infty) = 0$ uniformly for $t \in [0,\omega]$. By [10, Theorems 2.2 and 2.3], we have the following result about traveling waves of (2.1).

THEOREM 2.9. *Assume that* (H1)–(H4) *hold. Let* c^* *be defined as* $c^* = c_\omega^*/\omega$. *Then for any* $c \ge c^*$, (2.1) *has an* ω*-periodic traveling wave solution* $\mathcal{U}(t,x-ct)$ *connecting*

$\beta^*(t)$ to 0 such that $\mathcal{U}(t,s)$ is continuous and nonincreasing in s. Moreover, for any $c < c^*$, (2.1) has no ω-periodic traveling wave $\mathcal{U}(t, x-ct)$ connecting $\beta^*(t)$ to 0.

3. Dynamics in a bounded domain. In this section, we consider (1.6) in a bounded spatial domain

$$
(3.1)\quad \begin{cases} \partial_t u(t,x) = d(t)\Delta u - g(t,u) + b(t)\displaystyle\int_\Omega \Gamma(a(t), x-y) f_{-\tau}(t, u(t-\tau, y))dy, \\ \qquad\qquad (t,x) \in (0,\infty)\times\Omega, \\ Bu(t,x) = 0 \text{ on } (0,\infty)\times\partial\Omega, \\ u(t,x) = \phi(t,x), \quad t \in [-\tau, 0],\ x \in \Omega, \end{cases}
$$

where $\Omega \subseteq \mathbb{R}^n$ is a bounded domain with boundary $\partial\Omega$ of class $C^{1+\theta}(0 < \theta \le 1)$, the boundary condition is either $Bu = u$ (Dirichlet boundary condition) or $Bu = (\partial u/\partial\nu) + \alpha(x)u$ (Robin type boundary condition) for some nonnegative function $\alpha \in C^{1+\theta}(\partial\Omega, \mathbb{R})$, and $\partial u/\partial\nu$ denotes the differentiation in the direction of outward normal ν to $\partial\Omega$.

Let $p \in (1,\infty)$ be fixed. For each $\beta \in (\frac{1}{2} + \frac{1}{2p}, 1)$, let $\mathbb{X}_\beta$ be the fractional power space of $L^p(\Omega)$ with respect to $-\Delta$ and the boundary condition $Bu = 0$ (see, e.g., [8]). Then $\mathbb{X}_\beta$ is an ordered Banach space with the positive cone $\mathbb{X}_\beta^+$ consisting of all nonnegative functions in $\mathbb{X}_\beta$, and $\mathbb{X}_\beta^+$ has nonempty interior $\mathrm{int}(\mathbb{X}_\beta^+)$. Moreover, $\mathbb{X}_\beta \subseteq C^{1+\nu}(\bar\Omega)$ with continuous inclusion for $\nu \in [0, 2\beta - 1 - \frac{1}{p})$. Denote the norm on $\mathbb{X}_\beta$ by $\|\cdot\|_\beta$. Then there exists a constant $k_\beta > 0$ such that $\|\phi\|_\infty := \max_{x\in\bar\Omega}|\phi(x)| \le k_\beta \|\phi\|_\beta$ for all $\phi \in \mathbb{X}_\beta$.

Let $\bar C = C([-\tau, 0], \mathbb{X}_\beta)$ and $\bar C^+ = C([-\tau, 0], \mathbb{X}_\beta^+)$. For convenience, we identify an element $\phi \in \bar C$ as a function from $[-\tau, 0]\times\bar\Omega$ to $\mathbb{R}$ defined by $\phi(s,x) = \phi(s)(x)$. For any $N \ge L$, let $\bar C_{\widehat N} = \{\phi \in \bar C : 0 \le \phi(s,x) \le N, (s,x) \in [-\tau, 0]\times\bar\Omega\}$. For any function $y(\cdot) : [-\tau, b) \to \mathbb{X}_\beta$, where $b > 0$, define $y_t \in \bar C$, by $y_t(s) = y(t+s)$ for all $s \in [-\tau, 0]$, $t \in [0, b)$.

Note that the differential operator Δ generates an analytic semigroup $\bar U_0(t)$ on $L^p(\Omega)$ and that the standard parabolic maximum principle (see, e.g., [15, Corollary 7.2.3]) implies that the semigroup $\bar U_0(t) : \mathbb{X}_\beta \to \mathbb{X}_\beta$ is strongly positive in the sense that $\bar U_0(t)(\mathbb{X}_\beta^+ \setminus \{0\}) \subseteq \mathrm{int}(\mathbb{X}_\beta^+)$ for all $t > 0$. By a similar analysis as in section 2, we can write (3.1) as an integral equation (2.7) with $u_0 = \phi \in \bar C^+$. It then follows from [13, Corollary 5] that, for any $\phi \in \bar C_{\widehat L}$, (3.1) has a unique mild solution $u(t,x,\phi)$ with $u_0(\cdot,\cdot,\phi) = \phi$ and $u_t(\cdot,\cdot,\phi) \in \bar C_{\widehat L}$ for all $t \ge 0$. Moreover, $u(t,x,\phi)$ is a classic solution when $t > \tau$ and the comparison theorem holds for (3.1).

Define a family of operators $\{Q_t\}_{t\ge0}$ on $\bar C^+$ by

$$
Q_t(\phi)(s,x) = u(t+s,x,\phi) \quad \forall \phi \in \bar C^+, x \in \bar\Omega, \quad t \ge 0, \quad s \in [-\tau, 0].
$$

Similarly as in section 2, we can show that $\{Q_t\}_{t\ge0}$ is a monotone ω-periodic semiflow on $\bar C^+$; $u(t,x,\phi) > 0$ for $t > \tau$, $x \in \bar\Omega$, $\phi \in \bar C^+$ with $\phi \not\equiv 0$, and hence, Q_t is strongly positive for $t > 2\tau$; moreover, Q_t is compact on $\bar C^+$ for all $t > \tau$. Let $n_1 = \min\{n \in \mathbb{N}, n\omega > 2\tau\}$. Then $Q_{n_1\omega}$ is compact and strongly positive on $\bar C^+$. We

can further show that the periodic semiflow $\{Q_t\}_{t\geq 0}$ is point dissipative on $\bar{C}^+$. By [23, Theorem 1.1.3], we have the following result.

LEMMA 3.1. *Let* (H1)–(H3) *hold. Then* $Q_{n_1\omega}$ *admits a global attractor on* $\bar{C}^+$.

Consider the linearized system of (3.1) at the zero solution

(3.2)
$$\begin{cases} \partial_t \tilde{u}(t,x) = d(t)\Delta\tilde{u} - g_u(t,0)\tilde{u}(t,x) + b(t)\partial_u f_{-\tau}(t,0)\displaystyle\int_\Omega \Gamma(a(t),x-y)\tilde{u}(t-\tau,y)dy, \\ \qquad\qquad t>0, x\in\Omega, \\ B\tilde{u}(t,x) = 0, \quad t>0, x\in\partial\Omega, \\ \tilde{u}(s,x) = \phi(s,x), \quad s\in[-\tau,0],\ x\in\Omega, \phi\in\bar{C}. \end{cases}$$

Similarly as in Theorem 2.3, we can show that the comparison principle holds for (3.2), and hence, the solution map $\tilde{u}_t$ of (3.2) is monotone increasing for all $t\geq 0$.

Now we consider (3.1) and (3.2) as $n_1\omega$-periodic systems. Define the Poincaré map of (3.2) $P_1 : \bar{C}\to\bar{C}$ by $P_1(\phi) = \tilde{u}_{n_1\omega}(\phi)$ for all $\phi\in\bar{C}$, where $\tilde{u}_{n_1\omega}(\phi)(s,x) = \tilde{u}(n_1\omega+s,x,\phi)$ for all $(s,x)\in[-\tau,0]\times\bar{\Omega}$, and $\tilde{u}(t,x,\phi)$ is the solution of (3.2) with $\tilde{u}(s,x)=\phi(s,x)$ for all $(s,x)\in[-\tau,0]\times\bar{\Omega}$. Similarly as in section 2, we can obtain that P_1 is also compact and strongly positive. Let $r_1 = r(P_1)$ be the spectral radius of P_1. By the Krein–Rutman theorem (see, e.g., [8, Theorem 7.2]), $r_1>0$ and P_1 has a positive eigenfunction $\bar{\phi}\in\mathrm{int}(\bar{C}^+)$ corresponding to r_1.

LEMMA 3.2. *Let* $\mu = -\frac{1}{n_1\omega}\ln r_1$. *Then there exists a positive* $n_1\omega$*-periodic function* $v(t,x)$ *such that* $e^{-\mu t}v(t,x)$ *is a solution of* (3.2).

Proof. By the definitions of r_1 and $\bar{\phi}$, we have $P_1\bar{\phi} = r_1\bar{\phi}$. Let $\tilde{u}(t,x,\bar{\phi})$ be the solution of (3.2) with $\tilde{u}(s,x)=\bar{\phi}(s,x)$ for all $s\in[-\tau,0], x\in\Omega$. Since $\bar{\phi}\gg 0$, it is not difficult to see that $\tilde{u}(\cdot,\cdot,\bar{\phi})\gg 0$. Let $\mu = -\frac{1}{n_1\omega}\ln r_1$ and $v(t,x) = e^{\mu t}\tilde{u}(t,x,\bar{\phi})$ for all $t\geq -\tau, x\in\Omega$. Then $r_1 = e^{-n_1\omega\mu}$ and $v(t,x)>0$ for all $t\in[-\tau,\infty), x\in\Omega$. Moreover,

(3.3)
$$\begin{aligned} &v_t(t,x) \\ &= e^{\mu t}\tilde{u}_t(t,x,\bar{\phi}) + \mu e^{\mu t}\tilde{u}(t,x,\bar{\phi}) \\ &= e^{\mu t}[d(t)\Delta\tilde{u} - g_u(t,0)\tilde{u}(t,x,\bar{\phi}) \\ &+ b(t)\partial_u f_{-\tau}(t,0)\int_\Omega \Gamma(a(t),x-y)\tilde{u}(t-\tau,y,\bar{\phi})dy] + \mu v \\ &= d(t)\Delta v - g_u(t,0)v(t,x) + e^{\mu\tau}b(t)\partial_u f_{-\tau}(t,0)\int_\Omega \Gamma(a(t),x-y)v(t-\tau,y)dy + \mu v \end{aligned}$$

for all $(t,x)\in(0,\infty)\times\Omega$. Thus, $v(t,x)$ is a solution of $n_1\omega$-periodic equation (3.3) with $Bv=0$ on $(0,\infty)\times\partial\Omega$ and $v(s,x) = e^{\mu s}\bar{\phi}(s,x)$ for all $s\in[-\tau,0], x\in\Omega$.

For any $\theta\in[-\tau,0]$, $x\in\Omega$, we have

$$v(n_1\omega+\theta,x) = e^{\mu(n_1\omega+\theta)}\cdot P_1(\bar{\phi})(\theta,x) = e^{\mu(n_1\omega+\theta)}\cdot r_1\bar{\phi}(\theta,x) = e^{\mu\theta}\cdot\tilde{u}(\theta,x,\bar{\phi}) = v(\theta,x).$$

Therefore, $v_0(\theta,\cdot) = v_{n_1\omega}(\theta,\cdot)$ for all $\theta\in[-\tau,0]$, and hence, the existence and uniqueness of solutions of (3.3) imply that

$$v(t,x) = v(t+n_1\omega,x) \quad \forall t\geq -\tau, x\in\Omega,$$

that is, $v(t,x)$ is an $n_1\omega$-periodic solution of (3.3). Clearly, $e^{-\mu t}v(t,x)$ is a solution of (3.2). □

Define $P_0 : \bar{C} \to \bar{C}$ by $P_0(\phi) = \tilde{u}_\omega(\phi)$ for all $\phi \in \bar{C}$, where $\tilde{u}(t,x,\phi)$ is the solution of (3.2) with $\tilde{u}(s,x) = \phi(s,x)$ for all $s \in [-\tau, 0]$, $x \in \Omega$. Let $r_0 = r(P_0)$ be the spectral radius of P_0.

THEOREM 3.3. *Let* (H1)–(H3) *hold. For any* $\phi \in \bar{C}^+$, *denote by* $u(t,x,\phi)$ *the solution of* (3.1) *with* $u(s,x) = \phi(s,x)$ *for all* $(t,x) \in [-\tau, 0] \times \Omega$. *Then the following two statements are valid.*

(i) *If* $r_0 < 1$, *then* $\lim_{t\to\infty} \| u(t,\cdot,\phi) \|_\beta = 0$ *for every* $\phi \in \bar{C}^+$.

(ii) *If* $r_0 > 1$, *then* (3.1) *admits a unique positive* ω*-periodic solution* $u^*(t,x)$ *and* $\lim_{t\to\infty} \| u(t,\cdot,\phi) - u^*(t,\cdot) \|_\beta = 0$ *for all* $\phi \in \bar{C}^+ \setminus \{0\}$.

Proof. Since $P_1 = \tilde{u}_{n_1\omega}$, $P_0 = \tilde{u}_\omega$, and $\tilde{u}_{n_1\omega} = \tilde{u}_\omega^{n_1}$, where $\tilde{u}_t$ is the solution map of (3.2), by the properties of spectral radius of linear operators, we know that $r(P_1) = (r(P_0))^{n_1}$, i.e., $r_1 = (r_0)^{n_1}$. Note that the qualitative solutions of (3.1) and (3.2) do not change whether we consider them as $n_1\omega$-periodic systems or ω-periodic systems. The conditions in Theorem 3.3 can be replaced by $r_1 < 1$ and $r_1 > 1$, respectively. In the following, we will consider (3.1) and (3.2) as $n_1\omega$-periodic systems and prove the theorem under the conditions of $r_1 < 1$ and $r_1 > 1$.

In the case where $r_1 < 1$, we have $\mu = -\frac{1}{n_1\omega} \ln r_1 > 0$. By Lemma 3.2, (3.2) has a solution $\tilde{u}(t,x) := \tilde{u}(t,x,\bar{\phi}) = e^{-\mu t}v(t,x)$ with $\tilde{u}(s,x) = \bar{\phi}(s,x)$ for all $(s,x) \in [-\tau, 0] \times \Omega$, where $\bar{\phi} \in \text{int}(\bar{C}^+)$ is the positive eigenfunction of P_1 corresponding to r_1 and $v(t,x)$ is $n_1\omega$-periodic in $t \geq -\tau$. Then v is bounded on $[-\tau, \infty) \times \bar{\Omega}$, and hence, there exists $\rho > 0$ such that $\| v(t,\cdot) \|_\infty \leq \rho$ for all $t \geq -\tau$. Thus, $\lim_{t\to\infty} \| \tilde{u}(t,\cdot) \|_\infty = 0$. By the basic analysis of solutions of (3.2), it follows that $\lim_{t\to\infty} \| \tilde{u}(t,\cdot) \|_\beta = 0$.

Given $\phi \in \bar{C}^+$, since $\lim_{\delta\to 0^+}(\bar{\phi} - \delta\phi) = \bar{\phi} \in \text{int}(\bar{C}^+)$ for any $\epsilon > 0$, there exists $\delta_\phi > 0$, such that $\bar{\phi} - \delta\phi \in B_\epsilon(\bar{\phi}) \subseteq \bar{C}^+$ for $0 < \delta \leq \delta_\phi$, where $B_\epsilon(\bar{\phi})$ is an open ball in $\bar{C}^+$ centered at $\bar{\phi}$ with radius ϵ. Therefore, $\bar{\phi} \geq \delta_\phi\phi$ in $\bar{C}^+$. It then follows from the comparison principle that $\tilde{u}(t,x) \geq \delta_\phi \tilde{u}(t,x,\phi)$ for all $t \geq -\tau, x \in \bar{\Omega}$, where $\tilde{u}(t,\cdot,\phi)$ is the solution of (3.2) with $\tilde{u}(s,x) = \phi(s,x)$ for all $(s,x) \in [-\tau, 0] \times \Omega$. Thus, $\lim_{t\to\infty} \| \tilde{u}(t,\cdot,\phi) \|_\infty = 0$, and hence, $\lim_{t\to\infty} \| \tilde{u}(t,\cdot,\phi) \|_\beta = 0$ for any $\phi \in \bar{C}^+$.

Note that a solution of (3.1) satisfies

$$\partial_t u(t,x) \leq d(t)\Delta u - g_u(t,0)u(t,x) + b(t)\partial_u f_{-\tau}(t,0) \int_\Omega \Gamma(a(t), x-y)u(t-\tau,y)dy$$

for any $t > 0$, $x \in \Omega$. Similarly to the proof of Theorem 2.3, we can show that the comparison theorem for abstract functional differential equations [13, Proposition 3] can be applied to (3.1) and (3.2). Therefore, for any $\phi \in \bar{C}^+$, $u(t,\cdot,\phi) \leq \tilde{u}(t,\cdot,\phi)$ for all $t \geq -\tau$, where $u(t,\cdot,\phi)$ and $\tilde{u}(t,\cdot,\phi)$ are solutions of (3.1) and (3.2), respectively. It then follows that solutions of (3.1) satisfy $\lim_{t\to\infty} \| u(t,\cdot,\phi) \|_\beta = 0$ for all $\phi \in \bar{C}^+$.

In the case where $r_1 > 1$, we have $\mu < 0$. Let $\bar{C}_0 = \{\phi \in \bar{C}^+ : \phi \not\equiv 0\}$, $\partial\bar{C}_0 = \bar{C}^+ \setminus \bar{C}_0 = \{0\}$. Similarly to the proof of Lemma 2.5, we can show that for any $\phi \in \bar{C}_0$, the solution $u(t,x,\phi)$ of (3.1) satisfies $u(t,x,\phi) > 0$ for all $t > \tau$, $x \in \Omega$. It follows that $Q_t(\bar{C}_0) \subseteq int(\bar{C}^+)$ for all $t > 2\tau$. Clearly, $Q_t(0) = 0$ for all $t \geq 0$. We now have the following claim.

Claim. Zero is a uniform weak repeller for $\bar{C}_0$ in the sense that there exists $\delta_0 > 0$ such that $\lim_{t\to\infty} \sup \| Q_t(\phi) \|_\beta \geq \delta_0$ for all $\phi \in \bar{C}_0$.

Indeed, we consider the following system:

$$
\begin{cases}
\partial_t u^\varepsilon(t,x) = d(t)\Delta u^\varepsilon - (g_u(t,0)+\varepsilon)u^\varepsilon(t,x) \\
\qquad\qquad + b(t)(\partial_u f_{-\tau}(t,0)-\varepsilon)\displaystyle\int_\Omega \Gamma(a(t),x-y)u^\varepsilon(t-\tau,y)dy, \\
Bu^\varepsilon(t,x) = 0, \quad t>0, x\in\partial\Omega, \\
u^\varepsilon(s,x) = \phi(s,x), \quad s\in[-\tau,0],\ x\in\Omega, \phi\in\bar{C}.
\end{cases} \tag{3.4}
$$

Define the Poincaré map of (3.4) $P_\varepsilon : \bar{C}\to\bar{C}$ by

$$P_\varepsilon(\phi) = u^\varepsilon_{n_1\omega}(\phi) \quad \forall\phi\in\bar{C},$$

where

$$u^\varepsilon_{n_1\omega}(\phi)(s,x) = u^\varepsilon(n_1\omega+s,x,\phi) \quad \forall(s,x)\in[-\tau,0]\times\bar{\Omega}$$

and $u^\varepsilon(t,x,\phi)$ is the solution of (3.4) with $u^\varepsilon(s,x)=\phi(s,x)$ for all $s\in[-\tau,0], x\in\Omega$. Let $r_\varepsilon = r(P_\varepsilon)$ be the spectral radius of P_ε. Since $r_1 = r(P_1) > 1$, there exists a sufficiently small positive number ε_1 such that $r_\varepsilon > 1$ for all $\varepsilon\in[0,\varepsilon_1)$. We fix an $\varepsilon\in(0,\varepsilon_1)$. Since $\lim_{u\to0^+}\frac{g(t,u)}{u} = g_u(t,0)$ and $\lim_{u\to0^+}\frac{f_{-\tau}(t,u)}{u} = \partial_u f_{-\tau}(t,0)$ uniformly for $t\in[0,n_1\omega]$, there exists $\delta_\varepsilon>0$ such that $g(t,u) < (g_u(t,0)+\varepsilon)u$ and $f_{-\tau}(t,u) > (\partial_u f_{-\tau}(t,0)-\varepsilon)u$ for $u\in(0,\delta_\varepsilon)$, $t\in[0,n_1\omega]$. Let $\delta_0 = \delta_\varepsilon/k_\beta$. Suppose, by contradiction, that there exists $\phi_0\in\bar{C}_0$ such that $\lim_{t\to\infty}\sup\parallel Q_t(\phi)\parallel_\beta < \delta_0$. Then there exists $t_0>\tau$ such that $\parallel u(t,\cdot,\phi_0)\parallel_\infty \le k_\beta\parallel u(t,\cdot,\phi_0)\parallel_\beta < \delta_\varepsilon$ for all $t\ge t_0$. Therefore, $u(t,x,\phi_0)$ satisfies

$$
\begin{aligned}
&\partial_t u(t,x) \\
&> d(t)\Delta u - (g_u(t,0)+\varepsilon)u(t,x) + b(t)(\partial_u f_{-\tau}(t,0)-\varepsilon)\textstyle\int_\Omega \Gamma(a(t),x-y)u(t-\tau,y)dy
\end{aligned} \tag{3.5}
$$

for $t\ge t_0, x\in\Omega$. Let $\bar{\phi}_\varepsilon$ be the positive eigenfunction of P_ε associated with r_ε and $\mu_\varepsilon = -\frac{1}{n_1\omega}\ln r_\varepsilon$. Then by Lemma 3.2, the solution $u^\varepsilon(t,x,\bar{\phi}_\varepsilon)$ of (3.4) with $u^\varepsilon(s,x) = \bar{\phi}_\varepsilon(s,x)$ for all $s\in[-\tau,0]$, $x\in\Omega$, satisfies $u^\varepsilon(t,x,\bar{\phi}_\varepsilon) = e^{-\mu_\varepsilon t}v_\varepsilon(t,x)$, where $v_\varepsilon(t,x)$ is a positive $n_1\omega$-periodic function in $t\ge-\tau$. Since $u(t,x,\phi_0)>0$ for all $t\ge\tau$, $x\in\Omega$, there exists $\zeta>0$ such that

$$u(t_0+s,x,\phi_0)\ge\zeta u^\varepsilon(s,x,\bar{\phi}_\varepsilon) = \zeta\bar{\phi}_\varepsilon(s,x) \quad \forall s\in[-\tau,0], x\in\bar{\Omega}.$$

By (3.5) and the comparison theorem, we have

$$u(t,x,\phi_0)\ge\zeta u^\varepsilon(t-t_0,x,\bar{\phi}_\varepsilon) = \zeta e^{-\mu_\varepsilon(t-t_0)}v_\varepsilon(t,x) \quad \forall t\ge t_0, x\in\bar{\Omega}.$$

Since $\mu_\varepsilon<0$, it follows that $u(t,x,\phi_0)$ is unbounded, a contradiction. Thus, the claim is true.

By the claim above, $Q_{n_1\omega}$ is weakly uniformly persistent with respect to $(\bar{C}_0,\partial\bar{C}_0)$. Since $Q_{n_1\omega}$ admits a global attractor on $\bar{C}^+$, it follows from [23, Theorem 1.3.3] that $Q_{n_1\omega}$ is uniformly persistent with respect to $(\bar{C}_0,\partial\bar{C}_0)$ in the sense that there exists $\delta_1>0$ such that $\lim_{n\to\infty}\inf\parallel Q^n_{n_1\omega}(\phi)\parallel_\beta\ge\delta_1$ for all $\phi\in\bar{C}_0$.

Note that $Q_{n_1\omega}$ is compact, point dissipative, and uniformly persistent. It follows from [23, Theorem 1.3.6] that $Q_{n_1\omega}:\bar{C}_0\to\bar{C}_0$ admits a global attractor A_0 and

has a fixed point $\hat{\phi}$ in A_0. Similarly as in the proof of Lemma 2.4, we can show that $Q_{n_1\omega}$ is strictly subhomogeneous. Then [22, Lemma 1] implies that $Q_{n_1\omega}$ has at most one fixed point. Thus, $Q_{n_1\omega}$ has a unique equilibrium $\hat{\phi}$ in $\bar{C}_0$. Clearly, by the strong monotonicity of $Q_{n_1\omega}$, we have $\hat{\phi} \in \text{int}(\bar{C}^+)$. Moreover, it follows from [23, Theorem 2.3.2] that $A_0 = \{\hat{\phi}\}$ since $Q_{n_1\omega}$ is strongly monotone and strictly subhomogeneous. Thus, $\hat{\phi}$ is globally attractive in $\bar{C}_0$ for $Q_{n_1\omega}$.

Let $u(t, x, \hat{\phi})$ be the solution of (3.1) with $u(s, x) = \hat{\phi}(s, x)$ for all $(s, x) \in [-\tau, 0] \times \Omega$. Since $\hat{\phi}$ is a fixed point of $Q_{n_1\omega}$ and is globally attractive in $\bar{C}_0$, $u(t, x, \hat{\phi})$ is an $n_1\omega$-periodic solution of (3.1) which attracts all solutions of (3.1) in $\bar{C}^+ \setminus \{0\}$. That is,

$$\lim_{t\to\infty} \| u(t, \cdot, \phi) - u(t, \cdot, \hat{\phi}) \|_\beta = 0 \ \forall \ \phi \in \bar{C}_0.$$

Now we show that $u(t, x, \hat{\phi})$ is also ω-periodic. Since $Q_{n_1\omega}(\hat{\phi}) = \hat{\phi}$, we have $Q_\omega(Q_{n_1\omega}(\hat{\phi})) = Q_\omega(\hat{\phi})$, i.e., $Q_{n_1\omega}(Q_\omega(\hat{\phi})) = Q_\omega(\hat{\phi})$, which implies that $Q_\omega(\hat{\phi})$ is also a fixed point of $Q_{n_1\omega}$. By the fact that $\hat{\phi} \gg 0$ and the fact that Q_ω is monotone, it follows that $Q_\omega(\hat{\phi}) \gg 0$. Note that $Q_{n_1\omega}$ has a unique fixed point in $\text{int}(\bar{C}^+)$. Then $Q_\omega(\hat{\phi}) = \hat{\phi}$, that is, $\hat{\phi}$ is a fixed point of Q_ω, and hence, $u(t, x, \hat{\phi})$ is an ω-periodic solution of (3.1). Thus, $u^*(t, x) := u(t, x, \hat{\phi})$ for all $(t, x) \in [-\tau, \infty) \times \bar{\Omega}$, is the desired ω-periodic solution. □

Acknowledgments. We are grateful to two anonymous referees for their careful reading and helpful suggestions, which led to an improvement of our original manuscript.

REFERENCES

[1] W. G. Aiello and H. I. Freedman, *A time-delay model of single-species growth with stage structure*, Math. Biosci., 101 (1990), pp. 139–153.

[2] J. F. M. Al-Omari and S. A. Gourley, *A nonlocal reaction-diffusion model for a single species with stage structure and distributed maturation delay*, European J. Appl. Math., 16 (2005), pp. 37–51.

[3] J. F. M. Al-Omari and S. A. Gourley, *Monotone wave-fronts in a structured population model with distributed maturation delay*, IMA J. Appl. Math., 70 (2005), pp. 858–879.

[4] D. G. Aronson and H. F. Weinberger, *Nonlinear diffusion in population genetics, combustion, and nerve pulse propagation*, in Partial Differential Equations and Related Topics, J. A. Goldstein, ed., Lecture Notes in Math. 446, Springer-Verlag, Berlin, 1975, pp. 5–49.

[5] S. A. Gourley and Y. Kuang, *Wavefronts and global stability in a time-delayed population model with stage structure*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 459 (2003), pp. 1563–1579.

[6] S. A. Gourley and J. W. H. So, *Dynamics of a food-limited population model incorporating nonlocal delays on a finite domain*, J. Math. Biol., 44 (2002), pp. 49–78.

[7] S. A. Gourley and J. Wu, *Delayed non-local diffusive systems in biological invasion and disease spread*, Fields Inst. Commun., 48 (2006), pp. 137–200.

[8] P. Hess, *Periodic-Parabolic Boundary Value Problems and Positivity*, Longman Scientific and Technical, Harlow, UK, 1991.

[9] D. Liang, J. Wu, and F. Zhang, *Modelling population growth with delayed nonlocal reaction in 2-dimensions*, Math. Biosci. Eng., 2 (2005), pp. 111–132.

[10] X. Liang, Y. Yi, and X.-Q. Zhao, *Spreading speeds and traveling waves for periodic evolution systems*, J. Differential Equations, 231 (2006), pp. 57–77.

[11] X. Liang and X.-Q. Zhao, *Asymptotic speeds of spread and traveling waves for monotone semiflows with applications*, Comm. Pure Appl. Math., 60 (2007), pp. 1–40; Erratum: 61 (2008), pp. 137–138.

[12] R. H. Martin, *Nonlinear Operators and Differential Equations in Banach Spaces*, John Wiley and Sons, New York, 1976.

[13] R. H. Martin and H. L. Smith, *Abstract functional-differential equations and reaction-diffusion systems*, Trans. Amer. Math. Soc., 321 (1990), pp. 1–44.
[14] J. A. J. Metz and O. Diekmann, *The dynamics of physiologically structured populations*, J. A. J. Metz and O. Diekmann, eds., Springer-Verlag, Berlin, 1986.
[15] H. L. Smith, *Monotone Dynamical Systems: An Introduction to the Theory of Competitive and Cooperative Systems*, American Mathematical Society, 1995.
[16] H. L. Smith and H. R. Thieme, *Strongly order preserving semiflows generated by functional-differential equations*, J. Differential Equations, 93 (1991), pp. 332–363.
[17] J. W.-H. So, J. Wu, and X. Zou, *A reaction-diffusion model for a single species with age structure*. I. *Travelling wavefronts on unbounded domains*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 457 (2001), pp. 1841–1853.
[18] H. R. Thieme and X.-Q. Zhao, *Asymptotic speeds of spread and traveling waves for integral equations and delayed reaction-diffusion models*, J. Differential Equations, 195 (2003), pp. 430–470.
[19] A. I. Volpert, V. A. Volpert, and V. A. Volpert, *Traveling Wave Solutions of Parabolic Systems*, AMS, Providence, RI, 1994.
[20] D. Xu and X.-Q. Zhao, *A nonlocal reaction-diffusion population model with stage structure*, Can. Appl. Math. Q., 11 (2003), pp. 303–319.
[21] D. Xu and X.-Q. Zhao, *Dynamics in a periodic competitive model with stage structure*, J. Math. Anal. Appl., 311 (2005), pp. 417–438.
[22] X.-Q. Zhao, *Global attractivity and stability in some monotone discrete dynamical system*, Bull. Austral. Math. Soc., 53 (1996), pp. 305–324.
[23] X.-Q. Zhao, *Dynamical Systems in Population Biology*, Springer-Verlag, New York, 2003.

SIAM J. MATH. ANAL.
Vol. 40, No. 6, pp. 2517–2535

EXISTENCE OF SOLUTIONS FOR A MODEL DESCRIBING THE DYNAMICS OF JUNCTIONS BETWEEN DISLOCATIONS*

NICOLAS FORCADEL† AND RÉGIS MONNEAU‡

Abstract. We study a dynamical version of a multiphase field model of Koslowski and Ortiz for planar dislocation networks. We consider a two-dimensional vector field which describes phase transitions between constant phases. Each phase transition corresponds to a dislocation line, and the vectorial field description allows the formation of junctions between dislocations. This vector field is assumed to satisfy a nonlocal vectorial Hamilton–Jacobi equation with nonzero viscosity. For this model, we prove the existence for all time of a weak solution.

Key words. dislocation dynamics, nonlocal equations, junctions, parabolic system of equations

AMS subject classifications. 35K15, 74K30

DOI. 10.1137/070710925

1. Introduction.

1.1. Physical motivation. Dislocations are line defects in crystal, and their motion is at the origin of plastic properties of metals (see, for instance, the book of Hirth and Lothe [13]). The typical length of these dislocation lines is of the order of the micrometer, and their typical thickness is of the order of the nanometer. Dislocation lines exist in almost all metals. At low temperatures, these lines are contained in their crystallographic plane, called the slip plane. When an elastic stress field is applied, these lines can move in their slip plane. The normal velocity of these lines is then proportional to the effective stress in the material.

Another important property of dislocations is that each dislocation line is characterized by a vector quantity, called the Burgers vector (see [13] for more details). To explain this briefly, let us say that this Burgers vector reflects the microscopic nature of the dislocation defect in a crystal lattice. To fix the ideas, let us consider three unit vectors b^1, b^2, and b^3 in the plane such that

$$b^1 + b^2 + b^3 = 0. \tag{1.1}$$

Then each dislocation line can be of three different natures: with Burgers vector b^1, with Burgers vector b^2, or with Burgers vector b^3. The Burgers vector is an invariant of the dislocation line and then does not change during the evolution of the dislocation.

A consequence of the existence of this Burgers vector is the possibility for the dislocation lines to create some triple junctions, with three dislocation lines, respectively, of Burgers vector b^1, b^2, and b^3, satisfying (1.1). Indeed, in real crystals, we can observe such junctions. We can even observe networks of dislocations related by junctions. Those self-organized structures are called Frank networks (see Figure 1.1). See, also, for instance, p. 190 in Hull and Bacon [14] for such networks in body-

*Received by the editors December 14, 2007; accepted for publication (in revised form) October 21, 2008; published electronically March 20, 2009. This work was supported by the ACI JC 1025.
http://www.siam.org/journals/sima/40-6/71092.html

†Projet Commands, CMAP-INRIA Futurs, Ecole Polytechnique, 91128 Palaiseau and ENSTA, UMA, 32 Bd Victor, 75739 Paris Cedex 15, France (forcadel@casemade.dauphine.fr).

‡CERMICS, Paris Est-ENPC, 6 & 8 avenue Blaise Pascal, Cit Descartes, Champs sur Marne, 77455 Marne la Valle Cedex 2, France (monneau@cermics.enpc.fr).

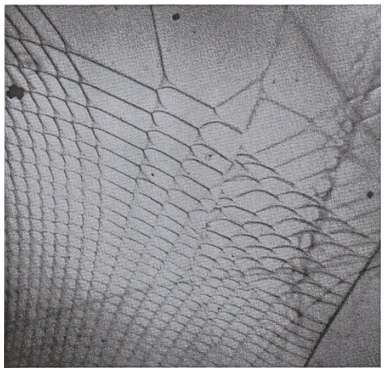


FIG. 1.1. *Frank networks observed with electron microscopy.*

centered cubic (BCC) iron or p. 188 for hexagonal networks in face-centered cubic (FCC) crystals.

In the present paper, we consider a special case of a set of dislocations contained in a single slip plane, where the dislocations can move. We are interested, in particular, in the motion of the junctions between dislocations, which has not been studied a lot, both from the modeling point of view and from the mathematical analysis point of view (see, for instance, the work of Rodney, Le Bouar, and Finel [20]). Let us mention, for the stationary case, the work of Cacace and Garroni [9]. The goal of the present paper is to propose and to study a model for the dynamics of junctions of dislocations.

Nevertheless, the question of junctions has several other physical applications, and there is a lot of literature on this subject. Let us mention, for instance, the problem of crystal growth or grain growth (see Taylor [23, 24] and Bronsard and Reitich [8]). We also refer to Bonnet [7] for problems concerning the minimization of the Mumford–Shah functional.

1.2. A phase field model for the dynamics of junctions. In a phase field model, the dislocation can be represented as the phase transition of a phase parameter $\rho(x) = \rho_1(x)e^1 + \rho_2(x)e^2 \in \mathbb{R}^2$ defined for $x = x_1e^1 + x_2e^2$ in the plane $\mathbb{R}^2$ with (e^1, e^2) an orthonormal basis. As in the work of Koslowski and Ortiz [15], we consider only the case of a periodic distribution of dislocations, reducing the problem on the torus, for $x \in \mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$. Then the energy of the dislocations, in the presence of a constant exterior applied resolved stress $\sigma^0 \in \mathbb{R}^2$, is then given (see [15]) by

$$\mathcal{E}(\rho) = \int_{\mathbb{T}^2} -\frac{1}{2}\left(C^0 \star \rho\right) \cdot \rho - \sigma^0 \cdot \rho + W(\rho). \tag{1.2}$$

In (1.2) and throughout the paper, we denote by $A \cdot B$ the scalar product between two vectors $A, B \in \mathbb{R}^2$. The precise meaning of the whole expression (1.2) will be explained later.

For any phase transition between two states $\rho = A$ and $\rho = B$, the difference $B - A$ needs physically to be the Burgers vector of the dislocation, i.e., a vector of the lattice $\Lambda = \mathbb{Z}a^1 + \mathbb{Z}a^2$ of the crystal we are considering, with a general given basis (a^1, a^2). This information is encoded in the potential $W : \mathbb{R}^2 \longrightarrow \mathbb{R}_+$, which is assumed to be minimal on Λ and to have the periodicity of the lattice Λ:

$$W(\rho + a) = W(\rho) \quad \text{for any} \quad a \in \Lambda. \tag{1.3}$$

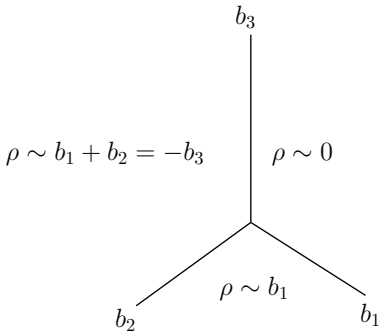


FIG. 1.2. *The junction of three dislocations as phase transitions of ρ.*

In other words, because ρ can be defined up to addition of any vector of the lattice Λ, only the transitions of ρ between two phases, i.e., two constant vectors of the lattice Λ, are physically meaningful. In particular, in this model, junctions of three dislocations of Burgers vectors $b^1, b^2, b^3 \in \Lambda$ with $b^1 + b^2 + b^3 = 0$ are expected, like, for instance, as the phase transitions between the states $0, b^1, -b^3$ (see Figure 1.2).

In the expression giving the energy (1.2), the kernel $C^0(x)$ is a 2×2 symmetric matrix which takes into account the long range elastic interactions between dislocations and

$$\left(C^0 \star \rho\right)_i = \sum_{j=1,2} C^0_{ij} \star \rho_j \quad \text{for} \quad i = 1, 2,$$

where $\star$ denotes the usual convolution on the torus $\mathbb{T}^2$. For instance, in the particular case of isotropic elasticity, for $k = (k_1, k_2) \in \mathbb{Z}^2$, the k-Fourier coefficient of the matrix C^0 is given (see [15] and also a limit case of the Peierls–Nabarro model in Alvarez et al. [3]) by

$$\widehat{C}^0(k) = -\frac{\mu}{2|k|}\left(\frac{1}{(1-\nu)} k \otimes k + k^\perp \otimes k^\perp\right), \tag{1.4}$$

where $k^\perp = (-k_2, k_1)$ is the vector obtained by a rotation of k of angle $\pi/2$. Here, $\mu > 0$ is a Lamé coefficient and $\nu \in (-1, 1/2)$ is the Poisson ratio of the material.

The fact that the matrix $-\widehat{C}^0(k)$ is nonnegative is related to the fact that the elastic energy is nonnegative. This insures that the elastic part of the energy $\int_{\mathbb{T}^2} -\frac{1}{2}(C^0 \star \rho) \cdot \rho$ created by the dislocations is nonnegative. The fact that the matrix $\widehat{C}^0(k)$ is not proportional to the identity reflects the fact that the existence of a Burgers vector associated to the dislocation line (that can be seen in the phase transition of ρ) creates some anisotropic elastic stress, even if we work in the framework of isotropic elasticity. The stress created by a dislocation line has somehow a preferred direction which is given by its Burgers vector.

When the material is submitted to an exterior shear stress, it makes the dislocations move. The dynamics of a given dislocation line is physically given by its normal velocity, which is called the resolved Peach–Koehler force. This force is the sum of the resolved exterior shear stress and the stress created by all of the dislocation lines, including the line itself. The total resolved stress $\sigma[\rho]$ is then formally given by the opposite of the gradient of the energy, i.e., $-\mathcal{E}'(\rho)$, and can be expressed as the

following nonlocal quantity:

$$\sigma[\rho] = \sigma^0 + C^0 \star \rho - W'_\rho(\rho). \tag{1.5}$$

Let us remark that one mathematical difficulty in the computation of this stress comes from the term $C^0 \star \rho$, which, from the point of view of the regularity, behaves like $\nabla\rho$, because of the linear growth of $\widehat{C^0}(k)$ in k. We have now to write an evolution equation for ρ, keeping in mind that the normal velocity has to be proportional to the stress, which means roughly speaking that

$$\rho = \bar\rho b^1 \quad \text{with} \quad \bar\rho(t,x) \in \mathbb{R}$$

has to satisfy

$$\frac{\bar\rho_t}{|\nabla\bar\rho|} = \bar\sigma[\bar\rho] \quad \text{with} \quad \bar\sigma[\bar\rho] = b^1 \cdot \sigma[\bar\rho b^1]. \tag{1.6}$$

Indeed, (1.6) is an equation very difficult to study mathematically, because $\bar\sigma[\bar\rho]$ behaves like $\nabla\bar\rho$. Up to our knowledge no results exist for (1.6), even in a framework of viscosity solutions for nonlocal equations. Indeed, one of the mathematical difficulties is due to the fact that, even putting the matrix C^0 to zero, (1.6) would become an equation like the Burgers equation, because of the presence of the derivative of the potential W in $\bar\sigma[\bar\rho]$, and could then create shocks in finite time. The presence of C^0 in the stress, even if this corresponds to a kind of degenerate diffusion (when $\nabla\bar\rho = 0$) in the equation, does not help sufficiently to get a good framework for a suitable notion of the solution to (1.6).

Because of these high mathematical difficulties and because our goal is really to deal with the dynamics of junctions of dislocations, we simplify mathematically the equation adding an artificial small viscosity $\varepsilon \in (0,1)$ and considering, instead of (1.6), the following equation:

$$\bar\rho_t = \bar\sigma[\bar\rho]|\nabla\bar\rho| + \varepsilon\Delta\bar\rho. \tag{1.7}$$

If the solution for the limit case $\varepsilon = 0$ is smooth enough, it seems reasonable to think that the solution to (1.7) is a good approximation. Moreover, any reasonable solution of the limit equation for $\varepsilon = 0$ should probably be seen as a limit of the solution to the ε-equation when ε goes to zero. Even numerically, if we would like to compute the solution in the limit case $\varepsilon = 0$, every classical numerical method introduces a numerical diffusion, which can be more or less interpreted as the additional viscosity $\varepsilon > 0$ in (1.7). For all of these reasons, (1.7) seems a good candidate for the evolution equation. As we will see in this paper, it is then mathematically possible to deal with this equation and to prove the existence of a global solution to (1.7). Nevertheless, even the study of this equation with ε-viscosity is not so simple, because, for instance, the uniqueness of the solution to (1.7) remains an open (and probably very difficult) problem in general.

As explained above, our goal is really to generalize (1.7) to take into account the evolution of junctions. To this end, we assume that the phase parameter $\rho(t,x) \in \mathbb{R}^2$ satisfies the following evolution equation:

$$\begin{cases} (\rho_k)_t = |\nabla\rho|^{-1} \displaystyle\sum_{i=1,2}\sum_{j=1,2} (\sigma[\rho])_i \nabla_j\rho_i \nabla_j\rho_k + \varepsilon\Delta\rho_k & \text{for} \quad k = 1,2 \\ & \text{in} \quad (0,T)\times\mathbb{T}^2, \\ \rho(0,x) = \rho^0(x) & \text{on} \quad \mathbb{T}^2, \end{cases} \tag{1.8}$$

where σ is given in (1.5), $\rho_t = \frac{\partial \rho}{\partial t}$ and $\nabla_j \rho_i = \frac{\partial \rho_i}{\partial x_j}$ for $i, j = 1, 2$, and

$$|\nabla \rho|^2 = \sum_{i=1,2} \sum_{j=1,2} |\nabla_j \rho_i|^2.$$

Here, we can check easily that (1.8) reduces to (1.7) when $\rho = \bar{\rho} b^1$ for any Burgers vector b^1 satisfying $|b^1| = 1$. For this reason, (1.8) seems to be a nice model to describe the general dynamics of junctions of dislocations.

Finally, let us mention that our model (1.8) has some similarities with the model of Allen and Cahn [2] on the motion of curved boundaries in which they consider gradient flow associated with a free-energy functional. This led to the study of scalar Ginzburg–Landau-type diffusion equations like

$$u_t = \Delta u - W'(u).$$

1.3. Main result. We need to introduce some assumptions that we will comment on below.

We make the following assumption on the kernel $C^0 : \mathbb{T}^2 \longrightarrow \mathbb{R}^{2\times 2}_{sym}$.

(A) We assume that there exists a constant $m > 0$ such that, for any $k \in \mathbb{Z}^2$, the Fourier coefficients of the kernel $\widehat{C}^0(k) = \int_{\mathbb{T}^2} dx\ e^{-2i\pi k\cdot x} C^0(x)$ satisfy $\widehat{C}^0(k) = M(k)$, where for any $\xi = (\xi_1, \xi_2) \in \mathbb{R}^2$ and any $p = (p_1, p_2) \in \mathbb{R}^2$

$$\text{(1.9)} \quad \begin{cases} M \in C^\infty\left(\mathbb{R}^2 \setminus \{0\}; \mathbb{R}^{2\times 2}_{sym}\right), \quad M(-\xi) = M(\xi), \quad M(\xi) = |\xi| M\left(\dfrac{\xi}{|\xi|}\right), \\ \dfrac{|\xi|\ |p|^2}{m} \geq -\displaystyle\sum_{i=1,2} \sum_{j=1,2} p_i \cdot M_{ij}(\xi) \cdot p_j \geq m|\xi|\ |p|^2 \quad \text{with} \quad |p|^2 = \displaystyle\sum_{i=1,2} (p_i)^2. \end{cases}$$

We also make the following assumption on the potential $W : \mathbb{T}^2 \longrightarrow \mathbb{R}_+$.

(B) We assume that $W \in C^2$ and W satisfies (1.3).

Condition (1.3) of assumption (B) is natural for the potential W, as explained in subsection 1.2. In assumption (B), we assume moreover that W is smooth enough (here C^2).

Assumption (A) requires more comments. First of all, remark that assumption (A) is satisfied by the matrix C^0 given by Koslowski and Ortiz [15] for isotropic elasticity (see (1.4)). Moreover, for general elasticity, the matrix $-\widehat{C}^0(k)$ has to be symmetric and nonnegative, because it corresponds physically to the nonnegative elastic energy of the dislocations. The last point concerns the 1-homogeneity in k of the matrix $\widehat{C}^0(k)$. Indeed, this is a general fact, when we compute $\widehat{C}^0(k)$ for general linear elasticity (see, for instance, the computation in the case of cubic elasticity done in Alvarez et al. [3]). Indeed, we start with a three-dimensional (3D) elastic energy, which behaves like the square of the H^1-norm of the displacement. When we consider dislocations in a single plane, as in the present model, this naturally reduces the problem from a 3D to a 2D problem and creates an energy as the square of the $H^{\frac{1}{2}}$-norm of the phase parameter, i.e., an energy like $\sum_{k\in\mathbb{Z}^2} |k||\widehat{\rho}|^2$, which is directly reflected into the 1-homogeneity in k of the matrix $\widehat{C}^0(k)$.

Then we have the following result for the model of dynamics of junctions between dislocations.

THEOREM 1.1 (existence of a solution). *Under assumptions* (A) *and* (B), *if* $\rho^0 \in (H^1(\mathbb{T}^2))^2$, *then, for any constant applied stress* $\sigma^0 \in \mathbb{R}^2$ *and for any time* $T > 0$, *there exists a solution* ρ *of* (1.8) *with* $\rho \in C^0([0, T); (L^{\frac{4}{3}}(\mathbb{T}^2))^2)$.

As mentioned above, the uniqueness of the solution is not known. Let us also mention that (1.8) is a nonlocal system of scalar equations and can be sketched as the following equation:

$$v_t = |\nabla v|^2 + \Delta v. \tag{1.10}$$

Indeed, this comes from our assumption (A) that the convolution with the kernel behaves like a first order operator. A lot of work has been done on equations (or systems) like (1.10). Let us mention, for instance, the works of Boccardo, Murat, and Puel [4, 5, 6] in which they study general equations including (1.10) and prove the existence result.

Equation (1.10) is also similar to the Navier–Stokes equations written for the potential A such that the velocity of the fluid is given by $u = \operatorname{curl} A$ (see, for instance, Leray [17]).

1.4. Organization of the paper. In section 2, we study an approximate problem of (1.8) where the right-hand side is approached by some term at most linear in the solution. The main result is proved in section 3. In a first subsection, we give some a priori estimates for the solution of the approximate problem, and then, in a second subsection, we pass to the limit in the approximate problem.

1.5. Notation. In what follows, we will denote by C a generic constant, which will then satisfy $C + C = C$, $C \cdot C = C$, and so on. We also use the following set:

$$W^{2,1;p}(Q_T) = \left\{ u \in L^p(Q_T);\ u_t \in L^p(Q_T) \text{ and } \frac{\partial^2 u}{\partial x_i, \partial x_j} \in L^p(Q_T) \text{ for } i,j = 1,2 \right\},$$

where $Q_T = (0,T) \times \mathbb{T}^2$.

2. An approximate problem. We first start to approximate the right-hand side of (1.8) by some term at most linear in the solution. To this end, we introduce a function h^n defined by

$$h^n(r) = h^0(r - n)$$

with

$$h^0(r) = \begin{cases} 1 & \text{if} \quad r \le 0, \\ 1 - r & \text{if} \quad 0 \le r \le 1, \\ 0 & \text{if} \quad r \ge 1. \end{cases}$$

We then look at the following approximate problem:

$$\begin{cases} \rho_t - \epsilon \Delta \rho = f^n[\rho] & \text{on} \quad Q_T := (0,T) \times \mathbb{T}^2, \\ \rho(0,\cdot) = \rho^0 & \text{on} \quad \mathbb{T}^2, \end{cases} \tag{2.1}$$

where

$$f^n[\rho] = h^n(|\nabla \rho|)\ |\nabla \rho|^{-1}\ (\nabla \rho)^T \cdot \nabla \rho \cdot \sigma[\rho]$$

and $\sigma[\rho]$ is given in (1.5) and is at most linear in ρ.

The natural idea to find a solution to (2.1) is to define the map Φ which associates to any function u, the solution $\rho = \Phi(u)$ of

$$\begin{cases} \rho_t - \epsilon\Delta\rho = f^n[u] & \text{on} \quad Q_T := (0,T)\times\mathbb{T}^2, \\ \rho(0,\cdot) = \rho^0 & \text{on} \quad \mathbb{T}^2, \end{cases} \tag{2.2}$$

and to prove that Φ has a fixed point in a suitable space. This way, we will prove the following result.

THEOREM 2.1 (existence of a solution for the approximate problem). *If $\rho^0 \in (H^1(\mathbb{T}^2))^2$, then, for any $n \geq 1$ and any $T > 0$, there exists a solution ρ^n of (2.1) with $\rho^n \in L^2((0,T);(H^2(\mathbb{T}^2))^2) \cap C^0([0,T);(L^2(\mathbb{T}^2))^2)$.*

In this section, we will give the proof of this theorem. In a first subsection, we will collect some preliminary results, and, in a second subsection, we will prove that Φ has a fixed point.

2.1. Preliminary results. The following lemma will be important.

LEMMA 2.2 (estimate on $C^0 \star \rho$). *For any $p \in (1,+\infty)$, there exists a constant C (depending on p and on the constant m defined in assumption* (A)*) such that, for any $\rho \in \left(W^{1,p}(\mathbb{T}^2)\right)^2$, we have*

$$\left\|C^0 \star \rho\right\|_{(L^p(\mathbb{T}^2))^2} \leq C\|\nabla\rho\|_{(L^p(\mathbb{T}^2))^{2\times 2}}. \tag{2.3}$$

Partial proof of Lemma 2.2. Let us make the proof for $p = 2$. We have, with $\sigma = \sigma[\rho]$,

$$\begin{aligned} |C^0 \star \rho|^2_{(L^2(\mathbb{T}^2))^2} &= \sum_{k\in\mathbb{Z}^2} \left|\widehat{(C^0 \star \rho)}(k)\right|^2 \\ &= \sum_{k\in\mathbb{Z}^2} \left|\widehat{C}^0(k)\cdot\widehat{\rho}(k)\right|^2 \\ &\leq \frac{1}{m^2}\sum_{k\in\mathbb{Z}^2} |k|^2|\widehat{\rho}(k)|^2 \\ &\leq \frac{1}{(2\pi m)^2}\sum_{k\in\mathbb{Z}^2} \left|\widehat{\nabla\rho}(k)\right|^2 \\ &= \frac{1}{(2\pi m)^2}|\nabla\rho|^2_{(L^2(\mathbb{T}^2))^{2\times 2}} \end{aligned}$$

which provides the result in the case $p = 2$.
The proof for the general case $p \in (1,+\infty)$ is given in Appendix A. ∎

An immediate corollary is the following estimate on the stress.

COROLLARY 2.3 (estimate on $\sigma[\rho]$). *For any $p \in (1,+\infty)$, there exists a constant C (depending on p, on the constant σ^0, on the potential W, and on the constant m defined in assumption* (A)*) such that, for any $\rho \in (W^{1,p}(\mathbb{T}^2))^2$, we have*

$$\|\sigma[\rho]\|_{(L^p(\mathbb{T}^2))^2} \leq C\left(1 + \|\nabla\rho\|_{(L^p(\mathbb{T}^2))^{2\times 2}}\right). \tag{2.4}$$

We will also need the following result.

LEMMA 2.4 (estimate on $f^n[u]$). *If $u \in (H^1(\mathbb{T}^2))^2$, then $f^n[u] \in (L^2(\mathbb{T}^2))^2$ with the following estimate:*

$$\|f^n[u]\|_{(L^2(\mathbb{T}^2))^2} \leq C(n+1)\left(1 + \|\nabla u\|_{L^2(\mathbb{T}^2)^{2\times 2}}\right),$$

where the constant C depends on σ^0, on the potential W, and on the constant m defined in assumption (A).

Proof of Lemma 2.4. Since $\mathrm{supp}(h^n) \subset [0, n+1]$, the following holds:

$$|f^n[u]| \leq (n+1)|\sigma[u]|, \tag{2.5}$$

where we have used the fact that $|B^T \cdot B \cdot p| \leq |B|^2|p|$ for $B \in \mathbb{R}^{2\times 2}$ and $p \in \mathbb{R}^2$. Then

$$\begin{aligned}\|f^n[u]\|_{(L^2(\mathbb{T}^2))^2} \leq & (n+1)\|\sigma[u]\|_{(L^2(\mathbb{T}^2))^2} \\ \leq & C(n+1)\left(1 + \|\nabla u\|_{L^2(\mathbb{T}^2)^{2\times 2}}\right),\end{aligned}$$

where we have used Corollary 2.3. □

We now recall some classical results. We start with the following parabolic estimates for the following equation:

$$\begin{cases} g_t - \epsilon\Delta g = f & \text{on} \quad Q_T := (0,T)\times\mathbb{T}^2, \\ g(0,\cdot) = g^0 & \text{on} \quad \mathbb{T}^2. \end{cases} \tag{2.6}$$

Proposition 2.5 (parabolic estimates for the heat equation). *Let $g^0 \in H^1(\mathbb{T}^2)$ and $f \in L^2(Q_T)$. Then there exists a unique solution g to* (2.6) *with*

$$g \in L^2\left((0,T); H^2\left(\mathbb{T}^2\right)\right) \cap L^\infty\left((0,T); H^1\left(\mathbb{T}^2\right)\right), \quad g_t \in L^2(Q_T). \tag{2.7}$$

We have the following estimate:

$$\begin{aligned}&\sup_{0\leq t\leq T} \|g(t)\|_{H^1(\mathbb{T}^2)} + \|g\|_{L^2((0,T);H^2(\mathbb{T}^2))} + \|g_t\|_{L^2((0,T);L^2(\mathbb{T}^2))} \\ &\leq C_T\left(\|f\|_{L^2(Q_T)} + \|g^0\|_{H^1(\mathbb{T}^2)}\right),\end{aligned} \tag{2.8}$$

where the constant C_T depends only on T and ε.

Moreover, we have

$$\sup_{0\leq t\leq T}\int_{\mathbb{T}^2} g^2(t) + 4\varepsilon\int_0^T\int_{\mathbb{T}^2}|\nabla g|^2 \leq 4\int_{\mathbb{T}^2}\left(g^0\right)^2 + 16T\int_0^T\int_{\mathbb{T}^2} f^2. \tag{2.9}$$

Proof of Proposition 2.5. For the proof of (2.7)–(2.8), we refer to Evans [11, Theorem 5, p. 360].

To prove (2.9), we simply multiply (2.6) by g and integrate over $\mathbb{T}^2$ and $(0,t)$, taking the supremum for $0 \leq t \leq T$. We get

$$\sup_{0\leq t\leq T}\int_{\mathbb{T}^2}\frac{g^2(t)}{2} \leq \sup_{0\leq t\leq T}\left(\int_{\mathbb{T}^2}\frac{g^2(t)}{2} + \varepsilon\int_0^t\int_{\mathbb{T}^2}|\nabla g|^2\right) \leq \int_{\mathbb{T}^2}\frac{\left(g^0\right)^2}{2} + \int_0^T\int_{\mathbb{T}^2}|g\ f| \tag{2.10}$$

and

$$\varepsilon\int_0^T\int_{\mathbb{T}^2}|\nabla g|^2 \leq \int_{\mathbb{T}^2}\frac{g^2(T)}{2} + \varepsilon\int_0^T\int_{\mathbb{T}^2}|\nabla g|^2 \leq \int_{\mathbb{T}^2}\frac{\left(g^0\right)^2}{2} + \int_0^T\int_{\mathbb{T}^2}|g\ f|. \tag{2.11}$$

Summing (2.10) and (2.11), we finally get

$$\sup_{0\leq t\leq T}\int_{\mathbb{T}^2}\frac{g^2(t)}{2} + \varepsilon\int_0^T\int_{\mathbb{T}^2}|\nabla g|^2 \leq \int_{\mathbb{T}^2}\left(g^0\right)^2 + 2\int_0^T\int_{\mathbb{T}^2}|g\ f|.$$

We now use the fact that

$$2\int_0^T\int_{\mathbb{T}^2}|g\ f| \le 2\left(\int_0^T\int_{\mathbb{T}^2} g^2\right)^{\frac{1}{2}}\cdot\left(\int_0^T\int_{\mathbb{T}^2} f^2\right)^{\frac{1}{2}}$$

$$\le 2\left(T\sup_{0\le t\le T}\int_{\mathbb{T}^2} g^2(t)\right)^{\frac{1}{2}}\cdot\left(\int_0^T\int_{\mathbb{T}^2} f^2\right)^{\frac{1}{2}}$$

$$\le \quad \frac{1}{4}\sup_{0\le t\le T}\int_{\mathbb{T}^2} g^2(t) + 4T\int_0^T\int_{\mathbb{T}^2} f^2,$$

which implies the result. □

We also recall the following theorem.

THEOREM 2.6 (Schaefer's fixed point theorem). *Let X be a real Banach space. Suppose that*

$$\Phi : X \to X$$

is a continuous and compact mapping. Assume further that the set

$$\{u \in X, \quad u = \lambda\Phi(u) \quad \text{for some} \quad \lambda \in [0,1]\}$$

is bounded. Then Φ has a fixed point.

For the proof of this theorem, we refer to Evans [11, Theorem 4, p. 504].

Finally, we will need some compactness argument and a weak continuity property contained in the following two classical results.

PROPOSITION 2.7 (compactness). *We recall that*

$$W^{2,1;2}(Q_T) = \left\{g \in L^2\left((0,T);H^2\left(\mathbb{T}^2\right)\right), \quad g_t \in L^2(Q_T)\right\}.$$

Then the injection

$$W^{2,1;2}(Q_T) \longrightarrow L^2\left((0,T);H^1\left(\mathbb{T}^2\right)\right) \quad \text{is compact.}$$

For the proof of this result, we refer to Lions [18, Theorem 5.1, p. 58].

PROPOSITION 2.8 (continuity). *With the notation of Proposition* 2.7, *let us consider a sequence $(g^m)_m$ such that*

$$g^m \rightharpoonup g \quad \text{weakly in} \quad W^{2,1;2}(Q_T).$$

We assume, also, that $g^m_{|t=0} = \rho^0$. Then

$$g_{|t=0} = \rho^0.$$

This result is classical, but for the reader's convenience we give the proof in Appendix A.

2.2. Proof of Theorem 2.1. We are now ready to make the proof of Theorem 2.1. To this end, for any $T > 0$, we set

$$X_T = L^2\left((0,T);H^1\left(\mathbb{T}^2\right)\right).$$

In all that follows, the index n is assumed fixed. We first remark that if $u \in X_T^2$, then $f^n[u] \in (L^2(Q_T))^2$, and then we can consider the solution ρ of

$$\begin{cases} \rho_t - \epsilon\Delta\rho = f^n[u] & \text{on} \quad Q_T := (0,T)\times\mathbb{T}^2, \\ \rho(0,\cdot) = \rho^0 & \text{on} \quad \mathbb{T}^2, \end{cases} \tag{2.12}$$

which satisfies $\rho \in X_T^2$ because of the parabolic estimates of Proposition 2.5. Then we set $\Phi(u) = \rho$ and see that Φ maps X_T^2 into X_T^2. We will prove that Φ admits a fixed point using Schaefer's fixed point theorem. We do the proof in four steps.

Step 1 (weak continuity of Φ). Let us consider sequences $(u^m)_m, (\rho^m)_m$ such that

$$\begin{cases} u^m \in X_T^2, \quad \rho^m = \Phi(u^m), \\ u^m \longrightarrow u \quad \text{in} \quad X_T^2. \end{cases}$$

From Lemma 2.4, we deduce that

$$||f^n[u^m]||_{(L^2(Q_T))^2} \leq C(n+1)\left(1 + ||u^m||_{X_T^2}\right). \tag{2.13}$$

From the parabolic estimates (Proposition 2.5), we deduce that ρ^m is bounded in $(W^{2,1;2}(Q_T))^2$; i.e., there exists a constant $C > 0$ such that

$$||\rho^m||_{(W^{2,1;2}(Q_T))^2} \leq C. \tag{2.14}$$

Therefore, up to a subsequence, we have

$$\rho^m \rightharpoonup \rho \quad \text{in} \quad \left(W^{2,1;2}(Q_T)\right)^2,$$

and, from Proposition 2.8, we deduce that

$$\rho_{|t=0} = \rho^0 \quad \text{on} \quad \mathbb{T}^2.$$

We now claim that

$$f^n[u^m] \longrightarrow f^n[u] \quad \text{in} \quad L^1(Q_T). \tag{2.15}$$

Indeed, we can write

$$f^n[u] = g^n(\nabla u)\cdot\sigma[u] \quad \text{with} \quad g^n(\nabla u) := h^n(|\nabla u|)\ |\nabla u|^{-1}\ (\nabla u)^T \cdot \nabla u.$$

From Lemma 2.2, for $p = 2$, and the continuity of W', we already deduce that

$$\sigma[u^m] \longrightarrow \sigma[u] \quad \text{in} \quad L^2(Q_T). \tag{2.16}$$

From the convergence of u^m to u in X_T^2, we deduce that, up to a subsequence, we have $\nabla u^m \longrightarrow \nabla u$ a.e. in Q_T. Now, from the fact that g^n is continuous and bounded, we deduce, in particular, that

$$g^n(\nabla u^m) \longrightarrow g^n(\nabla u) \quad \text{in} \quad L^2(Q_T). \tag{2.17}$$

Then the convergence (2.15) follows from (2.16) and (2.17).

Therefore, we conclude that ρ solves (2.12). Finally, by uniqueness of the solutions of (2.12), we deduce that the limit ρ does not depend on the choice of the subsequence and then that the full sequence converges:

$$\rho^m \rightharpoonup \rho \quad \text{weakly in} \quad \left(W^{2,1;2}(Q_T)\right)^2, \quad \text{with} \quad \rho = \Phi(u).$$

Step 2 (compactness of Φ). The compactness (and the usual strong continuity) of Φ follows from the compactness of the injection $(W^{2,1;2}(Q_T))^2 \longrightarrow X_T^2$ (see Proposition 2.7).

Step 3 (a priori bounds on the solutions of $u = \lambda\Phi(u)$ for T small). Let us consider a solution u of

$$u = \lambda\Phi(u) \quad \text{for some} \quad \lambda \in [0,1]. \tag{2.18}$$

Then, from the parabolic estimates (2.9), we have

$$\begin{aligned}
&\sup_{0\le t\le T}\int_{\mathbb{T}^2}|u(t)|^2 + 4\varepsilon\int_0^T\int_{\mathbb{T}^2}|\nabla u|^2\\
&\le 4\int_{\mathbb{T}^2}\left|\rho^0\right|^2 + 16T\int_0^T\int_{\mathbb{T}^2}|\lambda f^n[u]|^2\\
&\le 4\int_{\mathbb{T}^2}\left|\rho^0\right|^2 + 32TC^2(n+1)^2\left(T + \int_0^T\int_{\mathbb{T}^2}|\nabla u|^2\right),
\end{aligned}$$

where, in the third line, we have used Lemma 2.4 and the fact that $|\lambda| \le 1$. Therefore, for

$$T \le T^* := \left(16C^2(n+1)^2\right)^{-1}\varepsilon \tag{2.19}$$

we have

$$\sup_{0\le t\le T}\int_{\mathbb{T}^2}|u(t)|^2 + 2\varepsilon\int_0^T\int_{\mathbb{T}^2}|\nabla u|^2 \le 4\int_{\mathbb{T}^2}\left|\rho^0\right|^2 + 2\varepsilon T,$$

which proves that there exists a constant $C > 0$ such that any solution of (2.18) satisfies

$$||u||_{X_T^2} \le C.$$

We can then apply Schaefer's fixed point theorem (Theorem 2.6) to deduce that Φ has a fixed point on X_T^2, and, therefore, there is a solution ρ of (2.1) on the time interval $(0,T)$ if T satisfies (2.19), i.e., if T is small enough independently on the initial data ρ^0.

Step 4 (solution for any time). Let us call $\rho(\rho^0,t)$ the function $\rho(t,\cdot)$ obtained at Step 3 as a solution of (2.1) on the time interval $[0,T^*)$ with initial data ρ^0. From the parabolic estimates (Proposition 2.5), we also know that $\rho(t,\cdot) \in (H^1(\mathbb{T}^2))^2$ for any $t \in [0,T^*)$. Then we can define with $\tau = T^*/2$

$$u(0) = \rho^0 \quad \text{and} \quad u(t) = \rho\left(u(k\tau),t\right) \quad \text{if} \quad k\tau \le t < (k+1)\tau \quad \text{with} \quad k \in \mathbb{N}.$$

Using the fact that $u_t \in L^2_{loc}((0,+\infty);(L^2(\mathbb{T}^2))^2)$ and the fact that the problem is invariant by translation in time, we can easily check that u solves (2.1) for any $T > 0$ and provides the desired solution $\rho^n = u$ of Theorem 2.1.

This ends the proof of Theorem 2.1. □

3. A priori estimates and proof of Theorem 1.1.

3.1. A priori estimates. We have the following a priori estimates.

LEMMA 3.1 (a priori estimates). *There exists a constant $C > 0$ such that, for all $T > 0, n \geq 1$, and $0 < \varepsilon < 1$, any solution ρ^n of (2.1) given by Theorem 2.1 satisfies*

$$\|\rho^n\|^2_{L^\infty\left((0,T);\left(H^{\frac{1}{2}}(\mathbb{T}^2)\right)^2\right)} \leq Ce^{\frac{CT}{\varepsilon}}, \tag{3.1}$$

$$\|\rho^n\|^2_{L^2\left((0,T);\left(H^{\frac{3}{2}}(\mathbb{T})\right)^2\right)} \leq \frac{C}{\varepsilon}e^{\frac{CT}{\varepsilon}}, \tag{3.2}$$

and

$$\left\|h^n(|\nabla\rho^n|)|\nabla\rho^n|^{-\frac{1}{2}}\nabla\rho^n \cdot \sigma[\rho^n]\right\|^2_{(L^2(Q_T))^2} \leq Ce^{\frac{CT}{\varepsilon}}. \tag{3.3}$$

Proof of Lemma 3.1.

Step 1 (preliminaries on the energy). We first recall the expression of the energy for a general $\mathbb{Z}^2$-periodic smooth function $\rho(x) = (\rho_1(x), \rho_2(x))$:

$$\mathcal{E}(\rho) = \int_{\mathbb{T}^2} -\frac{1}{2}\left(C^0 \star \rho\right) \cdot \rho - \sigma^0 \cdot \rho + W(\rho).$$

For future use, we start to evaluate from below the first term in the energy, using Fourier series

$$\begin{aligned}\int_{\mathbb{T}^2} -\left(C^0 \star \rho\right) \cdot \rho &= \sum_{k\in\mathbb{Z}^2} -\widehat{(C^0 \star \rho)}(k) \cdot \widehat{\rho}^*(k) \\ &= \sum_{k\in\mathbb{Z}^2} -\left(\widehat{C}^0(k) \cdot \widehat{\rho}(k)\right) \cdot \widehat{\rho}^*(k) \\ &\geq m \sum_{k\in\mathbb{Z}^2} |k|\ |\widehat{\rho}(k)|^2,\end{aligned}$$

where we have used, in the first line, the fact that ρ and C^0 are real, and, in the last line, we have used assumption (A). Then we define

$$||\rho||^2_{\left(\dot{H}^{\frac{1}{2}}(\mathbb{T}^2)\right)^2} := \sum_{k\in\mathbb{Z}^2} |k|\ |\widehat{\rho}(k)|^2.$$

Similarly, we compute

$$\begin{aligned}&\int_{\mathbb{T}^2} -\left(C^0 \star (\nabla\rho)^T\right) : \nabla\rho \\ &= (2\pi)^2 \sum_{k\in\mathbb{Z}^2} -\left(\widehat{C}^0(k) \cdot \widehat{\rho}(k) \otimes (ik)\right) : (ik)^* \otimes \widehat{\rho}^*(k) \\ &= (2\pi)^2 \sum_{k\in\mathbb{Z}^2} -|k|^2\left(\widehat{C}^0(k) \cdot \widehat{\rho}(k)\right) : \widehat{\rho}^*(k) \\ &\geq (2\pi)^2 m \sum_{k\in\mathbb{Z}^2} |k|^3\ |\widehat{\rho}(k)|^2,\end{aligned}$$

where we have used assumption (A) in the last line. Then we define

$$||\rho||^2_{\left(\dot{H}^{\frac{3}{2}}(\mathbb{T}^2)\right)^2} := \sum_{k\in\mathbb{Z}^2} |k|^3\ |\widehat{\rho}(k)|^2.$$

Step 2 (estimate on the time derivative of the energy). Let us fix $T > 0$. We know that any solution ρ^n given by Theorem 2.1 belongs to the space $W^{2,1;2}(Q_T)$. In particular, using the following general fact (because of assumption (A))

$$\int_{\mathbb{T}^2} -\left(C^0 \star \rho\right)\cdot\rho = \mathrm{Re}\left(\sum_{k\in\mathbb{Z}^2} -|k|\left(\hat{C}^0\left(\frac{k}{|k|}\right)\cdot\widehat{\rho}(k)\right)\cdot\widehat{\rho}^*(k)\right)$$

we deduce that the energy $\mathcal{E}(\rho^n(t))$ is well-defined for almost every $t \in [0,T)$ and that, for almost every time $t \in [0,T)$, we have

$$\begin{aligned}
\frac{d}{dt}\mathcal{E}(\rho^n(t)) &= \int_{\mathbb{T}^2} -\sigma[\rho^n]\cdot\rho^n_t \\
&= \int_{\mathbb{T}^2} -h^n(|\nabla\rho^n|)|\nabla\rho^n|^{-1}\ |\nabla\rho^n\cdot\sigma[\rho^n]|^2 - \varepsilon\sigma[\rho^n]\cdot\Delta\rho^n \\
&= \int_{\mathbb{T}^2} -h^n(|\nabla\rho^n|)|\nabla\rho^n|^{-1}\ |\nabla\rho^n\cdot\sigma[\rho^n]|^2 \\
&\quad - \int_{\mathbb{T}^2} \varepsilon\left\{W''(\rho^n) : ((\nabla\rho^n)^T\cdot\nabla\rho^n) - \left(C^0\star(\nabla\rho^n)^T\right) : \nabla\rho^n\right\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\frac{d}{dt}\mathcal{E}(\rho^n(t)) + \int_{\mathbb{T}^2} h^n(|\nabla\rho^n|)|\nabla\rho^n|^{-1}\ |\nabla\rho^n\cdot\sigma[\rho^n]|^2 \\
&\leq C\varepsilon\left\{\int_{\mathbb{T}^2}|\nabla\rho^n|^2 + \left(C^0\star(\nabla\rho^n)^T\right) : \nabla\rho^n\right\}. \qquad (3.4)
\end{aligned}$$

But now (with a generic constant C)

$$\begin{aligned}
\|\nabla\rho^n\|^2_{(L^2(\mathbb{T}^2))^{2\times 2}} &\leq C\sum_{k\in\mathbb{Z}^2}|k|^2|\widehat{\rho}^n(k)|^2 \\
&\leq C\sum_{k\in\mathbb{Z}^2}|k|^{\frac{3}{2}}|\widehat{\rho}^n(k)|\cdot|k|^{\frac{1}{2}}|\widehat{\rho}^n(k)| \\
&\leq C\left(\sum_{k\in\mathbb{Z}^2}\frac{1}{2\alpha}|k|^3|\widehat{\rho}^n(k)|^2 + \frac{\alpha}{2}|k||\widehat{\rho}^n(k)|^2\right) \\
&\leq C\left(\int_{\mathbb{T}^2} -\frac{1}{\alpha}\left(C^0\star(\nabla\rho^n)^T\right) : \nabla\rho^n + \int_{\mathbb{T}^2} -\alpha\left(C^0\star\rho^n\right)\cdot\rho^n\right),
\end{aligned}$$

where $\widehat{\rho}^n(k)$ are the Fourier coefficients of ρ^n and α is a constant which will be precised later. We then deduce finally that

$$\begin{aligned}
&\frac{d}{dt}\mathcal{E}(\rho^n(t)) + \int_{\mathbb{T}^2} h^n(|\nabla\rho^n|)|\nabla\rho^n|^{-1}\ |\nabla\rho^n\cdot\sigma[\rho^n]|^2 \\
&\leq -C\varepsilon\left(1-\frac{1}{\alpha}\right)\int_{\mathbb{T}^2} -\left(C^0\star(\nabla\rho^n)^T\right) : \nabla\rho^n + C\varepsilon\alpha\int_{\mathbb{T}^2} -\left(C^0\star\rho^n\right)\cdot\rho^n \qquad (3.5) \\
&\leq -C\varepsilon\|\rho^n(t)\|^2_{\left(\dot{H}^{\frac{3}{2}}(\mathbb{T}^2)\right)^2} + C\varepsilon\left(1+\mathcal{E}(\rho^n(t)) + \left|\sigma^0\right|\left|\int_{\mathbb{T}^2}\rho^n(t)\right|\right)
\end{aligned}$$

for α chosen large enough, with C a suitable positive constant.

Step 3 (estimate on the time derivative of the mean value of the solution). Integrating in space equation (2.1), we get

$$\frac{d}{dt}\int_{\mathbb{T}^2}\rho^n(t) \quad = \int_{\mathbb{T}^2} h^n(|\nabla\rho^n|)|\nabla\rho^n|^{-1}\ (\nabla\rho^n)^T\cdot\nabla\rho^n\cdot\sigma[\rho^n]$$

and then

(3.6)
$$\begin{aligned}\frac{d}{dt}\left|\int_{\mathbb{T}^2}\rho^n(t)\right| &\le \int_{\mathbb{T}^2}(h^n(|\nabla\rho^n|)|\nabla\rho^n|)^{\frac12}\cdot\left(\left(h^n(|\nabla\rho^n|)|\nabla\rho^n|^{-1}\right)^{\frac12}|\nabla\rho^n\cdot\sigma[\rho^n]|\right)\\ &\le \int_{\mathbb{T}^2}\left(1+|\sigma^0|\right)h^n(|\nabla\rho^n|)|\nabla\rho^n|\\ &\quad+\frac{1}{4(1+|\sigma^0|)}h^n(|\nabla\rho^n|)|\nabla\rho^n|^{-1}|\nabla\rho^n\cdot\sigma[\rho^n]|^2.\end{aligned}$$

Step 4 (estimate on the energy). Setting

$$F^n(t)=1+\mathcal{E}(\rho^n(t))+\left(1+|\sigma^0|\right)\left|\int_{\mathbb{T}^2}\rho^n(t)\right|+\left(1+|\sigma^0|\right)^4, \tag{3.7}$$

we deduce from (3.5) and (3.6) that

$$\begin{aligned}&\frac{d}{dt}F^n(t)+\frac34\int h^n(|\nabla\rho^n|)|\nabla\rho^n|^{-1}|\nabla\rho^n\cdot\sigma[\rho^n]|^2\\ &\le -C\varepsilon\|\rho^n(t)\|^2_{\left(\dot H^{\frac32}(\mathbb{T}^2)\right)^2}+C\varepsilon\left(1+\mathcal{E}(\rho^n(t))+|\sigma^0|\left|\int_{\mathbb{T}^2}\rho^n(t)\right|\right)\\ &\quad+\left(1+|\sigma^0|\right)^2\int_{\mathbb{T}^2}h^n(|\nabla\rho^n|)|\nabla\rho^n|.\end{aligned}$$

Now we remark that

$$\begin{aligned}\left(1+|\sigma^0|\right)^2\int_{\mathbb{T}^2}h^n(|\nabla\rho^n|)|\nabla\rho^n| &\le \left(1+|\sigma^0|\right)^2\int_{\mathbb{T}^2}|\nabla\rho^n|\\ &\le \frac{C\varepsilon}{2}\int_{\mathbb{T}^2}|\nabla\rho^n|^2+\frac{\left(1+|\sigma^0|\right)^4}{2C\varepsilon}.\end{aligned}$$

Using the fact that (since the domain is bounded)

$$\int_{\mathbb{T}^2}|\nabla\rho^n|^2\le\|\rho^n(t)\|^2_{\left(\dot H^{\frac32}(\mathbb{T}^2)\right)^2},$$

we get

$$\begin{aligned}&\frac{d}{dt}F^n(t)+\frac34\int_{\mathbb{T}^2}h^n(|\nabla\rho^n|)|\nabla\rho^n|^{-1}|\nabla\rho^n\cdot\sigma[\rho^n]|^2+\frac{C\varepsilon}{2}\|\rho^n(t)\|^2_{\left(\dot H^{\frac32}(\mathbb{T}^2)\right)^2}\\ &\le C\varepsilon\left(1+\mathcal{E}(\rho^n(t))+\left|\sigma^0\right|\left|\int_{\mathbb{T}^2}\rho^n(t)\right|\right)+\frac{\left(1+|\sigma^0|\right)^4}{2C\varepsilon}\\ &\le\frac{C}{\varepsilon}F^n(t).\end{aligned} \tag{3.8}$$

This implies, using the Gronwall lemma,

$$F^n(t) \le F^n(0) e^{\frac{C}{\varepsilon}t}. \tag{3.9}$$

Step 5 (estimate on ρ^n). Let us first remark that

$$\mathcal{E}(\rho^n(t)) \ge \int -\frac{1}{2}\left(C^0 \star \rho^n\right)\cdot \rho^n - \left|\sigma^0\right| \left| \int_{\mathbb{T}^2} \rho^n(t) \right|. \tag{3.10}$$

Using (3.9), (3.10), and the definition of $F^n(t)$ yields

$$\int_{\mathbb{T}^2} -\frac{1}{2}\left(C^0 \star \rho^n\right)\cdot \rho^n + \left| \int_{\mathbb{T}^2} \rho^n(t) \right| \le C e^{\frac{C}{\varepsilon}t}.$$

Using Step 1, we then get

$$\|\rho^n\|^2_{L^\infty\left((0,T);\left(\dot{H}^{\frac{1}{2}}(\mathbb{T}^2)\right)^2\right)} \le C e^{\frac{C}{\varepsilon}T} \quad \text{and} \quad \left| \int_{\mathbb{T}^2} \rho^n(t) \right| \le C e^{\frac{C}{\varepsilon}T}.$$

This implies (3.1). Taking the integral $\int_0^T$ in (3.8) and using the fact that $\forall t \le T$, $F^n(t) \ge 0$, we get

$$\|h_n(|\nabla\rho^n|)|\nabla\rho^n|^{-\frac{1}{2}}\nabla\rho^n \cdot \sigma[\rho^n]\|^2_{(L^2(Q_T))^2} + \varepsilon C \|\rho^n\|^2_{L^2\left((0,T);\left(\dot{H}^{\frac{3}{2}}(\mathbb{T}^2)\right)\right)} \le C e^{\frac{C}{\varepsilon}T},$$

which implies (3.2) and (3.3). ∎

3.2. Proof of Theorem 1.1. We are now able to prove Theorem 1.1. In this section, we denote by C a generic constant which can depend on ρ^0, ε, and T but which does not depend on n.

Proof of Theorem 1.1. Let $T > 0$. The idea of the proof is to pass to the limit in (2.1). The only difficulty is to prove that the nonlinear term $f^n[\rho^n]$ converges in a certain sense to $|\nabla\rho|^{-1}(\nabla\rho)^T \cdot \nabla\rho \cdot \sigma[\rho]$, where ρ is the limit of ρ^n in an appropriate norm. The proof is decomposed into five steps.

Step 1 (a priori bound on $f^n[\rho^n]$). We have the following estimate on $f^n[\rho^n]$:

$$\|f^n[\rho^n]\|_{\left(L^{\frac{4}{3}}(Q_T)\right)^2} \le C. \tag{3.11}$$

To prove this, let us write

$$f^n[\rho^n] = \left(|\nabla\rho^n|^{-1}(\nabla\rho^n)^T\right)\cdot\left(|\nabla\rho^n|^{\frac{1}{2}}\right)\left(h_n(|\nabla\rho^n|)|\nabla\rho^n|^{-\frac{1}{2}}\nabla\rho^n\cdot\sigma[\rho^n]\right).$$

Using (3.3), we have that the last term is bounded in $(L^2(Q_T))^2$ by C. Moreover, the first term is bounded by 1 in $(L^\infty(Q_T))^{2\times 2}$, and then we just have to bound the term $|\nabla\rho^n|^{\frac{1}{2}}$ in $L^4(Q_T)$. Using (3.2), we have

$$\||\nabla\rho^n|^{\frac{1}{2}}\|_{L^4(Q_T)} = \left(\int_{Q_T} |\nabla\rho^n|^2\right)^{\frac{1}{4}} \le \|\rho^n\|^{\frac{1}{2}}_{L^2\left((0,T);\left(H^{\frac{3}{2}}(\mathbb{T})\right)^2\right)} \le C. \tag{3.12}$$

This implies (3.11).

Step 2 (strong convergence of $\nabla\rho^n$ in $L^2((0,T);(L^{\frac{4}{3}}(\mathbb{T}^2))^{2\times 2})$). Using the parabolic estimates for the heat equation (see [16, Chapter 4.3, p. 80 and Chapter 4.9, p. 341]) and Step 1, we get

$$\|\nabla\rho^n\|_{W^{\frac{1}{2};\frac{4}{3}}\left((0,T);\left(L^{\frac{4}{3}}(\mathbb{T}^2)\right)^{2\times 2}\right)} \leq C \tag{3.13}$$

where, for a Banach space B,

$$W^{\frac{1}{2};p}((0,T);B) = \left\{ g \in L^p((0,T);B),\ \int_0^T \int_0^T \frac{\|g(t)-g(s)\|_B^p}{|t-s|^{\frac{1}{2}p+1}}\, dt\, ds < \infty \right\}$$

is equipped with the following norm:

$$\|g\|_{W^{\frac{1}{2};p}((0,T);B)} := \left(\int_0^T \int_0^T \frac{\|g(t)-g(s)\|_B^p}{|t-s|^{\frac{1}{2}p+1}}\, dt\, ds \right)^{\frac{1}{p}}.$$

Moreover, using (3.2) we get

$$\|\nabla\rho^n\|_{L^2\left((0,T);\left(H^{\frac{1}{2}}(\mathbb{T}^2)\right)^{2\times 2}\right)} \leq C. \tag{3.14}$$

We then use the following lemma.

LEMMA 3.2 (compactness result). *Let $(g_n)_n$ be a sequence uniformly bounded in*

$$L^2\left((0,T);H^{\frac{1}{2}}\left(\mathbb{T}^2\right)\right) \cap W^{\frac{1}{2};\frac{4}{3}}\left((0,T);L^{\frac{4}{3}}\left(\mathbb{T}^2\right)\right);$$

then, for a subsequence,

$$g_n \to g \text{ strongly in } L^2\left((0,T);L^{\frac{4}{3}}\left(\mathbb{T}^2\right)\right).$$

Formally, the proof uses the fact that $H^{\frac{1}{2}} \subset L^{\frac{4}{3}}$ with compact injection in space, while the compactness in time comes from (3.13). We refer to Simon [21, Corollary 5, p. 86] for a more general result and for the proof of this lemma.

Using (3.13), (3.14), and Lemma 3.2, we then deduce that, for a subsequence, $\nabla\rho^n \to \nabla\rho$ strongly in $L^2((0,T);(L^{\frac{4}{3}}(\mathbb{T}^2))^2)$ and almost everywhere.

Step 3. (weak convergence of $\sigma[\rho^n]$ in $L^2((0,T);(L^4(\mathbb{T}^2))^2)$). We have $H^{\frac{1}{2}}(\mathbb{T}^2) \subset L^4(\mathbb{T}^2)$ with continuous injection (see Adams [1, Theorem 7.57, p. 217]). So $L^2((0,T);H^{\frac{1}{2}}(\mathbb{T}^2)) \subset L^2((0,T);L^4(\mathbb{T}^2))$ with continuous injection. We then deduce from (3.2) that

$$\|\nabla\rho^n\|_{L^2\left((0,T);(L^4(\mathbb{T}^2))^{2\times 2}\right)} \leq C. \tag{3.15}$$

Using Lemma 2.2, we then get

$$\left\|C^0 \star \rho^n\right\|_{L^2\left((0,T);(L^4(\mathbb{T}^2))^2\right)} \leq C. \tag{3.16}$$

Using the fact that the application $W^{2,1,\frac{4}{3}}_{x,t}(Q_T) \mapsto L^{\frac{4}{3}}(Q_T)$ is compact and the converse of the Lebesgue theorem, we deduce that $W'(\rho^n) \to W'(\rho)$ almost everywhere. This implies that $\sigma[\rho^n] \rightharpoonup \sigma[\rho]$ in $L^2((0,T);(L^4(\mathbb{T}^2))^2)$.

Step 4. (passing to the limit). Using Steps 2 and 3 and the fact that $|\nabla\rho^n|^{-1}\nabla\rho^n$ is bounded by 1, we deduce that

$$f_n[\rho^n] \to |\nabla\rho|^{-1}(\nabla\rho)^T \cdot \nabla\rho \cdot \sigma[\rho] \quad \text{in the distributions sense.}$$

By passing to the limit in (2.1), we obtain

$$\rho_t - \epsilon\Delta\rho = |\nabla\rho|^{-1}(\nabla\rho)^T \cdot \nabla\rho \cdot \sigma[\rho] \text{ in } \mathcal{D}'\left((0,T)\times\mathbb{T}^2\right). \tag{3.17}$$

Step 5 (initial condition). Using the fact that ρ_t^n are bounded uniformly in $L^{\frac{4}{3}}(Q_T)$ (by parabolic estimates for the heat equation and Step 1), we deduce that (uniformly in n)

$$\|\rho^n(t+h) - \rho^n(t)\|_{\left(L^{\frac{4}{3}}(\mathbb{T}^2)\right)^2} \le Ch^{\frac{1}{4}}\|\rho_t^n\|_{L^{\frac{4}{3}}\left((0,T);\left(L^{\frac{4}{3}}(\mathbb{T}^2)\right)^2\right)}$$

and then $\rho \in C^0((0,T);(L^{\frac{4}{3}}(\mathbb{T}^2))^2)$ and $\rho_{|t=0} = \rho^0$.

This achieves the proof of Theorem 1.1. □

Appendix A.

Full proof of Lemma 2.2. Here we do the proof for any $p \in (1,+\infty)$. Under assumption (A), there exists a constant $C > 0$ depending only on p such that the following result holds for any $\tilde\rho \in W^{1,p}(\mathbb{R}^2)$:

$$\left|\tilde C^0 \star_{\mathbb{R}^2} \tilde\rho\right|_{(L^p(\mathbb{R}^2))^2} \le \frac{C}{m}|\nabla\tilde\rho|_{(L^p(\mathbb{R}^2))^{2\times 2}},$$

where the Fourier transform of $\tilde C^0$ satisfies $\hat{\tilde C}^0 = M$ with M as in (1.9).

This result can be found in the scalar case on $\mathbb{R}^n$ in Stein [22, Proposition 5, p. 251] or Coifman, Meyer [10, Theorem 9, p. 39 and Proposition 2, p. 41]. See, also, Calderon–Zygmund inequalities Theorem 2.7.2 in Morrey [19]. Here the convolution by $\tilde C^0$ is a multiplier operator in the class S^1 of pseudodifferential operators. We then get the result in the vectorial case, summing the scalar components. See, also, the book of Garroni and Menaldi [12] for complements on integro-differential operators.

The fact that the result holds on the torus $\mathbb{T}^2$ is then classical. We prove it for the convenience of the reader. To this end, we consider a smooth function φ such that

$$\varphi(x) = 1 \quad \text{on} \quad [-1/3,1/3]^2, \qquad \text{supp } \varphi \subset [-2/3,2/3]^2, \quad \text{and} \quad 0 \le \varphi \le 1$$

such that

$$\sum_{k\in\mathbb{Z}^2} \varphi(x-k) = 1.$$

For any smooth function $\rho : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$, which is $\mathbb{Z}^2$-periodic, we then set for $K > 0$

$$(S_{2K}\rho)(x) = \sum_{|k|\le 2K,\ k\in\mathbb{Z}^2} \varphi(x-k)\rho(x).$$

Therefore, we get for $K > 0$ large enough

$$\begin{aligned}
&|B_K| \left\{ \left| \tilde{C}^0 \star_{\mathbb{R}^2} \rho \right|_{L^p((-1/2,1/2)^2)} + 0(1/K) \right\} \\
&\leq \left| \tilde{C}^0 \star_{\mathbb{R}^2} \rho \right|_{L^p(B_K)} \\
&\leq \left| \tilde{C}^0 \star_{\mathbb{R}^2} (S_{2K}\rho) \right|_{L^p(\mathbb{R}^2)} + \left| \tilde{C}^0 \star_{\mathbb{R}^2} (\rho - (S_{2K}\rho)) \right|_{L^p(B_K)} \\
&\leq \frac{C}{m} |\nabla(S_{2K}\rho)|_{L^p(\mathbb{R}^2)} + \left| \tilde{C}^0 \star_{\mathbb{R}^2} (\rho - (S_{2K}\rho)) \right|_{L^p(B_K)} \\
&\leq \frac{C}{m} |B_{2K}| \left\{ |\nabla\rho|_{(L^p(-1/2,1/2))^{2\times 2}} + 0(1/K) \right\} + |\rho|_{(L^\infty(\mathbb{R}^2))^2} |B_K| \int_{|z|\geq K-1} \left| \tilde{C}^0(z) \right|.
\end{aligned}$$

Using the fact that $\int_{|z|\geq K-1} |\tilde{C}^0(z)| = O(1/K)$, dividing by $|B_K|$, and taking the limit as $K \longrightarrow +\infty$, we get

$$\left| \tilde{C}^0 \star_{\mathbb{R}^2} \rho \right|_{(L^p(\mathbb{T}^2))^2} \leq \frac{C}{m} \frac{|B_2|}{|B_1|} |\nabla\rho|_{(L^p(\mathbb{T}^2))^{2\times 2}},$$

i.e.,

$$\left| C^0 \star_{\mathbb{T}^2} \rho \right|_{(L^p(\mathbb{T}^2))^2} \leq \frac{4C}{m} |\nabla\rho|_{(L^p(\mathbb{T}^2))^{2\times 2}}$$

with

$$C^0(x) = \sum_{k\in\mathbb{Z}^2} \tilde{C}^0(x-k).$$

We then get the final result by density of smooth functions in $(W^{1,p}(\mathbb{T}^2))^2$. □

Proof of Proposition 2.8. For simplicity of notation, we denote by $g(t)$ the function $x \mapsto g(t,x)$. We have

$$\begin{aligned}
\left\| g^m(t) - \rho^0 \right\|_{(L^2(\mathbb{T}^2))^2} &\leq \int_0^t ds \, \|g_t^m(s)\|_{(L^2(\mathbb{T}^2))^2} \\
&\leq \sqrt{t} \, \|g_t^m\|_{(L^2(Q_T))^2}.
\end{aligned}$$

Using the fact that g^m is bounded uniformly in $W^{2,1;2}(Q_T)$ (this is a consequence of the fact that $g^m \rightharpoonup g$ in $W^{2,1;2}(Q_T)$), we get

$$\left\| g^m(t) - \rho^0 \right\|_{(L^2(\mathbb{T}^2))^2} \leq C\sqrt{t}. \tag{A.1}$$

Now let $\varphi \in C_c^\infty([0,+\infty),\mathbb{R})$ be such that $\varphi \geq 0$. Using (A.1), we get that

$$\int_0^t ds \, \left\| g^m(s) - \rho^0 \right\|^2_{(L^2(\mathbb{T}^2))^2} \varphi(s) \leq C \int_0^t ds \, s \, \varphi(s).$$

Using Fatou's lemma, we deduce that

$$\int_0^t \left(\left\| g(s) - \rho^0 \right\|^2_{(L^2(\mathbb{T}^2))^2} - Cs \right) \varphi(s) \leq 0.$$

Using that $\varphi \geq 0$ is arbitrary, we deduce that, for almost every t, we have

$$\left\| g(t) - \rho^0 \right\|^2_{(L^2(\mathbb{T}^2))^2} \leq \sqrt{Ct}.$$

This implies the result. □

Acknowledgments. The authors thank A. El Hajj for fruitful discussion in the preparation of this paper. They also thank the referees for their comments and questions that helped to improve the manuscript.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Pure Appl. Math. 65, Academic Press, New York, London, 1975.
[2] S. Allen and J. Cahn, *A microscopic theory for the antiphase boundary motion and its application to antiphase domain coarsening*, Acta Metallurgica, 27 (1979), pp. 1085–1095.
[3] O. Alvarez, P. Hoch, Y. Le Bouar, and R. Monneau, *Dislocation dynamics: Short time existence and uniqueness of the solution*, Arch. Ration. Mech. Anal., 85 (2006), pp. 371–414.
[4] L. Boccardo, F. Murat, and J.-P. Puel, *Existence de solutions faibles pour des équations elliptiques quasi-linéaires à croissance quadratique*, in Nonlinear Partial Differential Equations and Their Applications. Collège de France Seminar Vol. IV (Paris, 1981/1982), Res. Notes in Math. 84, Pitman, Boston, MA, 1983, pp. 19–73.
[5] L. Boccardo, F. Murat, and J.-P. Puel, *Résultats d'existence pour certains problèmes elliptiques quasilinéaires*, Ann. Sc. Norm. Super. Pisa Cl. Sci., 11 (1984), pp. 213–235.
[6] L. Boccardo, F. Murat, and J.-P. Puel, *Existence results for some quasilinear parabolic equations*, Nonlinear Anal., 13 (1989), pp. 373–392.
[7] A. Bonnet, *On the regularity of the edge set of Mumford-Shah minimizers*, in Variational Methods for Discontinuous Structures (Como, 1994), Progr. Nonlinear Differential Equations Appl. 25, Birkhäuser, Basel, 1996, pp. 93–103.
[8] L. Bronsard and F. Reitich, *On three-phase boundary motion and the singular limit of a vector-valued Ginzburg-Landau equation*, Arch. Ration. Mech. Anal., 124 (1993), pp. 355–379.
[9] S. Cacace and A. Garroni, *A Multi-phase Transition Model for Dislocations with Interfacial Microstructure*, Interfaces and Free Boundaries, to appear.
[10] R. R. Coifman and Y. Meyer, *Au Delà des Opérateurs Pseudo-différentiels*, Astérisque 57, Société Mathématique de France, Paris, 1978.
[11] L. C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, American Mathematical Society, Providence, RI, 1998.
[12] M. G. Garroni and J.-L. Menaldi, *Green Functions for Second Order Parabolic Integro-differential Problems*, Pitman Res. Notes Math. 275, Longman Scientific & Technical, Harlow, UK, 1992.
[13] J. P. Hirth and J. Lothe, *Theory of Dislocations*, 2nd ed., Krieger, Malabar, FL, 1992.
[14] D. Hull and D. Bacon, *Introduction to Dislocations*, 4th ed., Butterworth-Heinemann, Oxford, 2001.
[15] M. Koslowski and M. Ortiz, *A multi-phase field model of planar dislocation networks*, Model. Simul. Material Sci. Eng., 12 (2004), pp. 1087–1097.
[16] O. A. Ladyženskaja, V. A. Solonnikov, and N. N. Ural′ceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, American Mathematical Society, Providence, RI, 1967.
[17] J. Leray, *Sur le mouvement d'un liquide visqueux emplissant l'espace*, Acta Math., 63 (1934), pp. 193–248.
[18] J.-L. Lions, *Quelques Méthodes de Résolution des Problèmes aux Limites Non Linéaires*, Dunod, Paris, 1969.
[19] J. C. B. Morrey, *Multiple Integrals in the Calculus of Variations*, Grundlehren Math. Wiss. 130, Springer-Verlag New York, 1966.
[20] D. Rodney, Y. Le Bouar, and A. Finel, *Phase field methods and dislocations*, Acta Materialia, 51 (2003), pp. 17–30.
[21] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. Pura Appl., 146 (1987), pp. 65–96.
[22] E. M. Stein, *Harmonic Analysis: Real-variable Methods, Orthogonality, and Oscillatory Integrals*, Princeton Math. Ser. 43, Princeton University Press, Princeton, NJ, 1993.
[23] J. E. Taylor, *The motion of multiple-phase junctions under prescribed phase-boundary velocities*, J. Differential Equations, 119 (1995), pp. 109–136.
[24] J. E. Taylor, *A variational approach to crystalline triple-junction motion*, J. Stat. Phys., 95 (1999), pp. 1221–1244.

SIAM J. MATH. ANAL.
Vol. 40, No. 6, pp. 2536–2566

MINIMIZATION OF ELECTROSTATIC FREE ENERGY AND THE POISSON–BOLTZMANN EQUATION FOR MOLECULAR SOLVATION WITH IMPLICIT SOLVENT*

BO LI†

Abstract. In an implicit-solvent description of the solvation of charged molecules (solutes), the electrostatic free energy is a functional of concentrations of ions in the solvent. The charge density is determined by such concentrations together with the point charges of the solute atoms, and the electrostatic potential is determined by the Poisson equation with a variable dielectric coefficient. Such a free-energy functional is considered in this work for both the case of point ions and that of ions with a uniform finite size. It is proved for each case that there exists a unique set of equilibrium concentrations that minimize the free energy and that are given by the corresponding Boltzmann distributions through the equilibrium electrostatic potential. Such distributions are found to depend on the boundary data for the Poisson equation. Pointwise upper and lower bounds are obtained for the free-energy minimizing concentrations. Proofs are also given for the existence and uniqueness of the boundary-value problem of the resulting Poisson–Boltzmann equation that determines the equilibrium electrostatic potential. Finally, the equivalence of two different forms of such a boundary-value problem is proved.

Key words. implicit solvent, electrostatic free energy, ionic concentrations, electrostatic potentials, the Poisson–Boltzmann equation, variational methods, nonlinear elliptic interface problems

AMS subject classifications. 35J, 35Q, 49S, 82D, 92C

DOI. 10.1137/080712350

1. Introduction.

It has long been realized that the electrostatic potential of a charged molecular system extremizes an electrostatic free-energy functional [3, 6, 12, 13, 15, 18, 20, 26, 28, 29]. In a simple setting, this functional is given by

$$F[c_1,\dots,c_M;\psi]=\int\left\{-\frac{\varepsilon}{8\pi}|\nabla\psi|^2+\rho\psi+\beta^{-1}\sum_{j=1}^{M}c_j\left[\ln\left(\Lambda^3c_j\right)-1\right]-\sum_{j=1}^{M}\mu_jc_j\right\}dx,$$

where $c_1,\dots,c_M$ are ionic concentrations, ψ is an electrostatic potential, ε is the dielectric constant, ρ is the charge density defined to be a linear combination of the ionic concentrations, β is the inverse thermal energy, Λ is the thermal de Broglie wavelength, and μ_j is the chemical potential of the jth ionic species. Throughout, we use the electrostatics CGS units. We also use $\log x$ to denote the natural logarithm of $x>0$. Extremizing this functional with respect to the concentrations and the potential lead to the Boltzmann distribution of concentrations and the Poisson equation for the equilibrium potential, respectively [3, 6, 12, 13, 15, 26, 28]. Notice, however, that this free-energy functional is concave with respect to the electrostatic potential. Therefore, the extremizing concentrations and potential do not minimize this free-energy functional, rather they form an unstable saddle point of the system [1, 6, 12, 13]. This flaw of theory is removed in the free-energy minimization approach that was proposed

*Received by the editors January 3, 2008; accepted for publication (in revised form) December 9, 2008; published electronically March 20, 2009. This work was supported by the U.S. National Science Foundation (NSF) through the grant DMS-0451466 and DMS-0811259, by the NSF Center for Theoretical Biological Physics with the NSF grant PHY-0822283, and by the U.S. Department of Energy through the grant DE-FG02-05ER25707.

http://www.siam.org/journals/sima/40-6/71235.html

†Department of Mathematics and the NSF Center for Theoretical Biological Physics, University of California, San Diego, 9500 Gilman Drive, Mail code: 0112, La Jolla, CA 92093-0112 (bli@math.ucsd.edu).

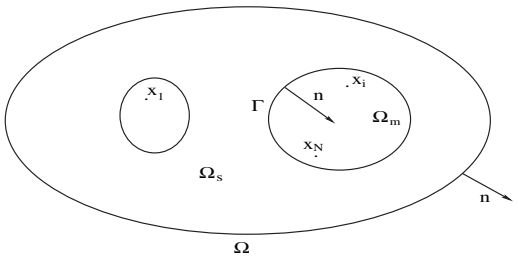


FIG. 1. *The geometry of a solvation system with an implicit solvent.*

in [12, 20]. The key point in this new approach is that the electrostatic free-energy functional depends solely on the ionic concentrations and the electrostatic potential is determined by such concentrations through the Poisson equation. In the recent article [5], this free-energy minimization approach was revisited and applied to the implicit-solvent (or continuum-solvent) description of solvation.

The present work is a mathematical study of the free-energy minimization approach to the electrostatics applied to the solvation of molecules with an implicit-solvent. Such application introduces additional mathematical complications due to the presence of point charges in solutes and the dielectric boundaries. We consider both the case of point ions—ions modeled as points without volumes—and that of ions with a uniform finite size. The finite-size effect of ions is known to be important in continuum modeling of electrostatics in molecular systems. Our analysis shows particularly that the free-energy minimizing ionic concentrations are uniformly bounded from above and away from zero at each spatial point. This uniform boundedness, which is proved by somewhat tedious constructions, is a consequence of the property that the free-energy minimizing concentrations have a large entropy. We do not consider the more general case of ions with different sizes for which there seems no explicit Boltzmann distributions.

Consider now the solvation of charged molecules with an implicit solvent [27]. We divide the entire region Ω of the solvation system into the region of solute molecules $\Omega_m \subset \mathbb{R}^3$ that is possibly multiply connected, the region of solvent (such as salted water) $\Omega_s \subset \mathbb{R}^3$, and the solute-solvent interface $\Gamma = \partial\Omega_m \cap \partial\Omega_s$; cf. Figure 1. This interface Γ serves as the dielectric boundary. Assume the solutes consist of N atoms with the ith one located at x_i and carrying a charge Q_i. Assume also there are M ionic species in the solvent with $q_j = ez_j$ the buck charge of the jth ionic species, where e is the elementary charge and z_j the valence of jth ionic species. Denote by $c_j = c_j(x)$ the local concentration at $x \in \Omega_s$ of the jth ionic species. Following the common assumption that the mobile ions in the solvent cannot penetrate the dielectric boundary Γ, we define $c_j(x) = 0$ for all $x \in \Omega_m$ and $1 \le j \le M$.

We consider two mean-field approximations of the electrostatic free energy of the solvation system as functionals of the local ionic concentrations $c = (c_1, \ldots, c_M)$ in the solvent region. In the first one, point ions are assumed, and the related electrostatic free-energy functional is given by [5, 12, 19, 20, 26]

$$F_0[c] = \frac{1}{2}\sum_{i=1}^{N} Q_i(\psi - \psi_{vac})(x_i) + \frac{1}{2}\int_{\Omega_s}\left(\sum_{j=1}^{M} q_j c_j\right)\psi dx$$
$$+ \beta^{-1}\sum_{j=1}^{M}\int_{\Omega_s} c_j\left[\log\left(a^3 c_j\right) - 1\right]dx - \sum_{j=1}^{M}\int_{\Omega_s}\mu_j c_j dx. \tag{1.1}$$

In the second approximation, all ions are assumed to have a uniform linear size, and the related free-energy functional is given by [3, 20]

$$F_a[c] = \frac{1}{2}\sum_{i=1}^{N} Q_i(\psi - \psi_{vac})(x_i) + \frac{1}{2}\int_{\Omega_s}\left(\sum_{j=1}^{M} q_j c_j\right)\psi dx$$
$$+ \beta^{-1}\sum_{j=0}^{M}\int_{\Omega_s} c_j\left[\log\left(a^3 c_j\right) - 1\right]dx - \sum_{j=1}^{M}\int_{\Omega_s}\mu_j c_j\,dx, \tag{1.2}$$

where the summation in the β^{-1} term starts from $j = 0$ and

$$c_0(x) = a^{-3}\left[1 - \sum_{j=1}^{M} a^3 c_j(x)\right] \qquad \forall x \in \Omega_s. \tag{1.3}$$

In (1.1) and (1.2), ψ is the electrostatic potential of the solvation system,

$$\psi_{vac}(x) = \sum_{i=1}^{N}\frac{Q_i}{\varepsilon_m|x - x_i|} \tag{1.4}$$

defines the electrostatic potential generated by all the point charges Q_i at x_i in a medium with the dielectric constant ε_m (usually taken as that in the vacuum), $a > 0$ is a constant, and μ_j is the constant chemical potential of the jth ionic species. The constant $a > 0$ represents in (1.1) a nonphysical cut-off which is often chosen to be the thermal de Broglie wavelength and in (1.2) the uniform linear size of ions.

The electrostatic potential ψ is determined by the Poisson equation

$$\nabla \cdot \varepsilon_\Gamma \nabla\psi = -4\pi\rho \qquad \text{in } \Omega, \tag{1.5}$$

where ε_Γ is the dielectric coefficient and ρ is the charge density, together with a boundary condition which is usually taken to be

$$\psi = \psi_0 \qquad \text{on } \partial\Omega, \tag{1.6}$$

where ψ_0 is a given function. The dielectric coefficient is defined to be

$$\varepsilon_\Gamma(x) = \begin{cases} \varepsilon_m & \text{if } x \in \Omega_m, \\ \varepsilon_s & \text{if } x \in \Omega_s, \end{cases} \tag{1.7}$$

where ε_m and ε_s are the dielectric constants of the solutes and the solvent, respectively. The charge density is given by

$$\rho = \sum_{i=1}^{N} Q_i\delta_{x_i} + \sum_{j=1}^{M} q_j c_j \qquad \text{in } \Omega, \tag{1.8}$$

where δ_{x_i} denotes the Dirac delta function centered at x_i.

The first two terms in (1.1) or (1.2) represent the internal electrostatic energy, which are often written formally as the integral of $\rho\psi/2$ over the entire region Ω. Based on Born's definition [2], the contribution to the electrostatic free energy due to the solute point charges is given as the first term in (1.1) or (1.2) though the reaction

field $\psi-\psi_{vac}$. The β^{-1} term represents the ideal gas entropy. The term $1-\sum_{j=1}^{M} a^3 c_j$ in (1.2) is the concentration of solvent molecules. It describes the effect of finite size of ions. The last term in (1.1) or (1.2) accounts for a constant chemical potential in the system. The osmotic pressure from the mobile ions is dropped, since it is only an additive constant to the free-energy functional in the present setting. We remark that the use of notations F_0 and F_a does not indicate that we can obtain the functional F_0 by simply setting $a = 0$ in F_a.

In this work, we prove the following results:

(1) For each of the free-energy functionals (1.1) and (1.2), there admits a unique minimizer $c_1, \ldots, c_M$, which is also the unique equilibrium, in an admissible set of concentrations. Moreover, such concentrations and the corresponding equilibrium electrostatic potential ψ are related by the boundary-data dependent Boltzmann distributions

$$c_j(x) = \begin{cases} c_j^{\infty} e^{-\beta q_j \left[\psi(x) - \hat{\psi}_0(x)/2\right]} & \text{for point ions,} \\ \dfrac{c_j^{\infty} e^{-\beta q_j [\psi(x) - \hat{\psi}_0(x)/2]}}{1 + a^3 \sum_{i=1}^{M} c_i^{\infty} e^{-\beta q_i [\psi(x) - \hat{\psi}_0(x)/2]}} & \text{for finite-size ions,} \end{cases} \tag{1.9}$$

for a.e. $x \in \Omega_s$ and $1 \le j \le M$, where $c_j^{\infty} = a^{-3} e^{\beta \mu_j}$ and $\hat{\psi}_0 \in H^1(\Omega)$ is determined by

$$\begin{cases} \displaystyle\int_{\Omega} \varepsilon_{\Gamma} \nabla \hat{\psi}_0 \cdot \nabla \eta dx = 0 & \forall \eta \in H_0^1(\Omega), \\ \hat{\psi}_0 = \psi_0 & \text{on } \partial\Omega. \end{cases} \tag{1.10}$$

The free-energy minimizing concentrations are shown to be uniformly bounded above and below away from zero. These results are summarized in Theorems 2.3–2.5 and Lemmas 3.4 and 3.5.

(2) The equilibrium electrostatic potential ψ is the unique solution to the boundary-data dependent Poisson–Boltzmann equation (PBE) [3, 4, 10, 11, 16, 17, 20, 31], together with the boundary condition (1.6),

$$\nabla \cdot \varepsilon_{\Gamma} \nabla \psi + 4\pi \chi_{\Omega_s} \sum_{j=1}^{M} q_j c_j^{\infty} e^{-\beta q_j \left(\psi - \hat{\psi}_0/2\right)} = -4\pi \sum_{i=1}^{N} Q_i \delta_{x_i} \qquad \text{in } \Omega \tag{1.11}$$

for the case of point ions, and

$$\nabla \cdot \varepsilon_{\Gamma} \nabla \psi + 4\pi \chi_{\Omega_s} \sum_{j=1}^{M} \frac{q_j c_j^{\infty} e^{-\beta q_j \left(\psi - \hat{\psi}_0/2\right)}}{1 + a^3 \sum_{i=1}^{M} c_i^{\infty} e^{-\beta q_i \left(\psi - \hat{\psi}_0/2\right)}} = -4\pi \sum_{i=1}^{N} Q_i \delta_{x_i} \qquad \text{in } \Omega \tag{1.12}$$

for the case of finite-size ions, where χ_{Ω_s} is the characteristic function of Ω_s. These equations can be written together as

$$\nabla \cdot \varepsilon_{\Gamma} \nabla \psi - 4\pi \chi_{\Omega_s} B' \left(\psi - \frac{\hat{\psi}_0}{2} \right) = -4\pi \sum_{i=1}^{N} Q_i \delta_{x_i} \qquad \text{in } \Omega, \tag{1.13}$$

where B' is the derivative of the function $B : \mathbb{R} \to \mathbb{R}$ defined by

(1.14)
$$B(\psi) = \begin{cases} \displaystyle\sum_{j=1}^{M} \beta^{-1} c_j^\infty e^{-\beta q_j \psi} & \text{for point ions,} \\ \beta^{-1} a^{-3} \log\left(1 + a^3 \displaystyle\sum_{j=1}^{M} c_j^\infty e^{-\beta q_j \psi}\right) & \text{for finite-size ions.} \end{cases}$$

See Theorem 2.1.

(3) The boundary-value problem of the PBE (1.13) and (1.6) is equivalent to the elliptic interface problem

(1.15)
$$\begin{cases} \nabla \cdot \varepsilon_m \nabla \psi = -4\pi \displaystyle\sum_{i=1}^{N} Q_i \delta_{x_i} & \text{in } \Omega_m, \\ \nabla \cdot \varepsilon_s \nabla \psi - 4\pi B'\left(\psi - \dfrac{\hat{\psi}_0}{2}\right) = 0 & \text{in } \Omega_s, \\ [\![\psi]\!] = [\![\varepsilon_\Gamma \nabla \psi \cdot n]\!] = 0 & \text{on } \Gamma, \\ \psi = \psi_0 & \text{on } \Omega. \end{cases}$$

Here and below, we denote for any function u on Ω, $u_m = u|_{\Omega_m}$, $u_s = u|_{\Omega_s}$,, and $[\![u]\!] = u_s - u_m$ on Γ. See Theorem 2.2.

Two variations of the PBE (1.11) with $\psi_0 = 0$ are commonly used [8, 15, 29]. First, we have by the Taylor expansion and the electrostatic neutrality $\sum_{j=1}^{M} c_j^\infty = 0$ that

$$\sum_{j=1}^{M} q_j c_j^\infty e^{-\beta q_j \psi} \approx -\left(\sum_{j=1}^{M} \beta q_j^2 c_j^\infty\right) \psi,$$

if $|\psi|$ is small, leading to the linearized PBE [9]

$$\nabla \cdot \varepsilon_\Gamma \nabla \psi - \varepsilon_s \kappa^2 \chi_{\Omega_s} \psi = -4\pi \sum_{i=1}^{N} Q_i \delta_{x_i} \qquad \text{in } \Omega,$$

where $\kappa = \sqrt{4\pi\beta \sum_{i=1}^{M} q_j^2 c_j^\infty / \varepsilon_s^2}$ is the ionic strength or the inverse Debye–Hückel screening length. Clearly, all of our results for the nonlinear PBE (1.11) hold true for the linearized PBE. Second, for the common $z : -z$ type of salt such as NaCl in the solution, we have $M = 2$, $c_1^\infty = c_2^\infty$, and $q_1 = -q_2 = ze$. The PBE (1.11) reduces to the following sinh PBE:

$$\nabla \cdot \varepsilon_\Gamma \nabla \psi - 8\pi z e c_1^\infty \chi_{\Omega_s} \sinh(\beta z e \psi) = -4\pi \sum_{i=1}^{N} Q_i \delta_{x_i} \qquad \text{in } \Omega.$$

In proving the existence of minimizers of the functionals F_0 and F_a, we use de la Vallée Poussin's criterion [25] of the sequential compactness in $L^1(\Omega)$. The uniqueness of such minimizers follows basically from the convexity of these functionals. A crucial step in defining and deriving equilibriums of F_0 and F_a is the construction of

L^∞-concentrations that are bounded below in Ω_s by a positive constant and that have low free energies. Such constructions are made by increasing the entropy of ionic concentrations through their small perturbations. The effect of inhomogeneous Dirichlet boundary data to the Boltzmann distributions and, hence, to the PBE can be useful to guide practical numerical computations. The equivalence of the two formulations is a common property for many physical problems. The interface formulation of the boundary-value problem of the PBE has been used for numerical computations using boundary integral method [22–24]. The finite-size effect is important in modeling electrostatics [3, 20].

The rest of the paper is organized as follows: In section 2, we state our main results; in section 3, we provide some lemmas; in section 4, we prove our theorems on the boundary-value problem of PBE; in section 5, we prove our theorems on the free-energy minimization. Finally, in Appendix, we give the proof of two lemmas.

2. Main results. Throughout the rest of the paper, we make the following assumptions:

A1. The set $\Omega \subset \mathbb{R}^3$ is nonempty, bounded, open, and connected. The sets $\Omega_m \subset \mathbb{R}^3$ and $\Omega_s \subset \mathbb{R}^3$ are nonempty, bounded, and open, and satisfy that $\overline{\Omega_m} \subset \Omega$ and $\Omega_s = \Omega \setminus \overline{\Omega_m}$. The N points $x_1, \ldots, x_N$ for some integer $N \geq 1$ belong to Ω_m. Both $\partial\Omega$ and Γ are of C^2. The unit exterior normal at the boundary of Ω_s is denoted by n; cf. Figure 1.

A2. $M \geq 2$ is an integer. All $a > 0$, $\beta > 0$, $Q_i \in \mathbb{R}$ $(1 \leq i \leq N)$, $q_j \in \mathbb{R}$ and $\mu_j \in \mathbb{R}$ $(1 \leq j \leq M)$, $\varepsilon_m > 0$, and $\varepsilon_s > 0$ are constants;

A3. The functions ψ_{vac} and ε_Γ are defined in (1.4) and (1.7), respectively. The boundary data ψ_0 is the trace of a given function, also denoted by ψ_0, in $W^{2,\infty}(\Omega)$.

Boundary values are understood as traces. When no confusion arises, the capital letter C, with or without a subscript, denotes a positive constant that can depend on all Ω_m, Ω_s, Ω, Γ, ε_m, ε_s, a, β, N, M, x_i, and Q_i $(1 \leq i \leq N)$, q_j and μ_j $(1 \leq j \leq M)$, and ψ_0.

For any open set $U \subseteq \mathbb{R}^3$ that contains all $x_1, \ldots, x_N$, we denote

$$H^1_*(U) = \left\{ u \in W^{1,1}(U) : u|_{U_\alpha} \in H^1(U_\alpha) \, \forall \alpha > 0 \right\},$$

where $U_\alpha = U \setminus (\cup_{i=1}^N \overline{B(x_i, \alpha)})$ and $B(x_i, \alpha)$ denotes the ball centered at x_i with radius α.

DEFINITION 2.1. *A function $\psi \in H^1_*(\Omega)$ is a weak solution to the boundary-value problem of the PBE* (1.13) *and* (1.6)*, if $\psi = \psi_0$ on $\partial\Omega$, $\chi_{\Omega_s} B(\psi) \in L^2(\Omega_s)$, and*

(2.1)
$$\int_\Omega \left[\varepsilon_\Gamma \nabla\psi \cdot \nabla\eta + 4\pi\chi_{\Omega_s} B'\left(\psi - \frac{\hat{\psi}_0}{2}\right)\eta \right] dx = 4\pi \sum_{i=1}^N Q_i \eta(x_i) \qquad \forall \eta \in C_c^\infty(\Omega).$$

We remark that if $\phi \in H^1(U)$ for some bounded and smooth domain $U \subset \mathbb{R}^3$, then e^ϕ and, hence, $B(\phi)$ may not be in $L^1(U)$. For example, let $U = B(0,1)$ be the unit ball of $\mathbb{R}^3$ and $\alpha \in (0, 1/2)$. Define $\phi(x) = |x|^{-\alpha}$ for any $x \in U$. Then $\phi \in H^1(U)$ and that $e^\phi \notin L^1(U)$. Notice by (1.14) that $\chi_{\Omega_s} B(\psi) \in L^2(\Omega_s)$ is equivalent to $\chi_{\Omega_s} e^{-\beta q_j \psi} \in L^2(\Omega_s)$ or $\chi_{\Omega_s} e^{-\beta q_j(\psi - \hat{\psi}_0/2)} \in L^2(\Omega_s)$ $(j = 1, \ldots, M)$, which in turn are equivalent to $\chi_{\Omega_s} B(\psi - \hat{\psi}_0/2) \in L^2(\Omega_s)$.

THEOREM 2.1. *There exists a unique weak solution $\psi \in H^1_*(\Omega)$ to the boundary-value problem of the PBE* (1.13) *and* (1.6)*. Moreover, $\psi \in C(\overline{\Omega} \setminus (\cup_{i=1}^N \overline{B(x_i, \alpha)}))$*

for any $\alpha > 0$ such that the closure of $\cup_{i=1}^N B(x_i, \alpha)$ is contained in Ω_m, and $\psi \in C^\infty((\Omega_m \setminus \{x_1, \dots, x_N\}) \cup \Omega_s)$.

DEFINITION 2.2. *A function $\psi : \Omega \to \mathbb{R}$ is a weak solution of the interface problem* (1.15)*, if the following are satisfied: $\psi_m \in H^1_*(\Omega_m)$ and*

$$\int_{\Omega_m} \varepsilon_m \nabla\psi \cdot \nabla\eta dx = 4\pi \sum_{i=1}^N Q_i \eta(x_i) \qquad \forall \eta \in C_c^\infty(\Omega_m); \tag{2.2}$$

$\psi_s \in H^1(\Omega_s)$, $\chi_{\Omega_s} B(\psi) \in L^2(\Omega_s)$, and

$$\int_{\Omega_s} \left[\varepsilon_s \nabla\psi \cdot \nabla\eta + 4\pi\chi_{\Omega_s} B'\left(\psi - \frac{\hat{\psi}_0}{2}\right)\eta \right] dx = 0 \qquad \forall \eta \in C_c^\infty(\Omega_s); \tag{2.3}$$

and the third and fourth equations in (1.15) *hold true.*

THEOREM 2.2. *A function $\psi : \Omega \to \mathbb{R}$ is a weak solution to the boundary-value problems* (1.13) *and* (1.6)*, if and only if it is a weak solution to the boundary-value problem* (1.15)*.*

Let $U \subset \mathbb{R}^3$ be a nonempty, bounded, and open set. Let $f \in L^1(U)$. Assume

$$\sup_{0 \neq \xi \in L^\infty(U) \cap H_0^1(U)} \frac{\int_U f\xi \, dx}{\|\xi\|_{H^1(U)}} < \infty. \tag{2.4}$$

Since $L^\infty(U) \cap H_0^1(U)$ is dense in $H_0^1(U)$, we can identify f as an element in $H^{-1}(U)$, the dual of $H_0^1(U)$, with

$$\langle f, \xi \rangle = \int_U f\xi dx \qquad \forall \xi \in L^\infty(U) \cap H_0^1(U),$$

and we write $f \in L^1(U) \cap H^{-1}(U)$. The $H^{-1}(U)$ norm of f is given by (2.4). We define

$$X = \left\{ c = (c_1, \dots, c_M) \in L^1\left(\Omega, \mathbb{R}^M\right) : c = 0 \text{ a.e. } \Omega_m \text{ and } \sum_{j=1}^M q_j c_j \in H^{-1}(\Omega) \right\},$$

$$\|c\|_X = \sum_{j=1}^M \|c_j\|_{L^1(\Omega_s)} + \left\| \sum_{j=1}^M q_j c_j \right\|_{H^{-1}(\Omega)} \qquad \forall c = (c_1, \dots, c_M) \in X.$$

Clearly, $(X, \|\cdot\|_X)$ is a Banach space.

Let $\alpha \in \mathbb{R}$ and define $S_\alpha : [0, \infty) \to \mathbb{R}$ by $S_\alpha(0) = 0$ and $S_\alpha(u) = u(\alpha + \log u)$ if $u > 0$. It is easy to see that S_α is bounded below on $[0, \infty)$ and strictly convex on $(0, \infty)$. Define

$$V_0 = \left\{ (c_1, \dots, c_M) \in X : \ c_j \geq 0 \text{ a.e. } \Omega_s \text{ and } \int_\Omega S_0(c_j)\, dx < \infty, \ j = 1, \dots, M \right\},$$

$$W_0 = \left\{ (c_1, \dots, c_M) \in V_0 : \text{ there exists } p > \frac{3}{2} \text{ such that } c_j \in L^p(\Omega), j = 1, \dots, M \right\},$$

$$V_a = \left\{ (c_1, \dots, c_M) \in V_0 : \ c_0 = a^{-3}\left(1 - \sum_{j=1}^M a^3 c_j\right) \geq 0 \text{ a.e. } \Omega_s \right\}.$$

Clearly, all V_0, W_0, and V_a are nonempty and convex. For any $c = (c_1, \ldots, c_M) \in V_0$, there exists a unique weak solution $\psi = \psi(c)$ of the boundary-value problem (1.5) and (1.6) with the charge density ρ given by (1.8); in particular, $\psi - \psi_{vac}$ is harmonic in Ω_m, cf. Lemma 3.2. We shall call $\psi = \psi(c)$ the electrostatic potential corresponding to c. Therefore, $F_0 : V_0 \to \mathbb{R}$ and $F_a : V_a \to \mathbb{R}$ are well defined. We use V, F to denote V_0, F_0 or W_0, F_0 or V_a, F_a.

DEFINITION 2.3. *An element $c = (c_1, \ldots, c_M) \in V$ is an equilibrium of $F : V \to \mathbb{R}$, if*

(2.5)

there exist $\gamma_1 > 0$ and $\gamma_2 > 0$ such that $\gamma_1 \le c_j(x) \le \gamma_2$ a.e. $x \in \Omega_s$, $j = 1, \ldots, M$,

for the case of point ions, or

(2.6) *there exists $\theta_0 \in (0, 1)$ such that $a^3 c_j(x) \ge \theta_0$ a.e. $x \in \Omega_s$, $j = 0, 1, \ldots, M$,*

for the case of finite-size ions; and

$$\delta F[c]e := \lim_{t \to 0} \frac{F[c + te] - F[c]}{t} = 0 \qquad \forall e \in X \cap L^\infty\left(\Omega, \mathbb{R}^M\right).$$

DEFINITION 2.4. *An element $c \in V$ is a local minimizer of $F : V \to \mathbb{R}$, if there exists $\varepsilon > 0$ such that $F[d] \ge F[c]$ for any $d \in V$ with $\|d - c\|_X < \varepsilon$.*

THEOREM 2.3. *There exists a unique minimizer of $F_0 : V_0 \to \mathbb{R}$. It is also the unique local minimizer of $F_0 : V_0 \to \mathbb{R}$.*

It is an open question if the unique minimizer of $F_0 : V_0 \to \mathbb{R}$ is an equilibrium of $F_0 : V_0 \to \mathbb{R}$ as defined in Definition 2.3. The answer to this question would be yes if this minimizer were in W_0 or if $\min_{d \in V_0} F_0[d] = \min_{d \in W_0} F_0[d]$, neither of which is clearly true. This is the reason we introduce the class of concentrations W_0. See the proof of Lemma 3.4 in Appendix.

THEOREM 2.4.

(1) *There exists a unique equilibrium $c = (c_1, \ldots, c_M)$ of $F_0 : W_0 \to \mathbb{R}$. It is also the unique global minimizer and the unique local minimizer of $F_0 : W_0 \to \mathbb{R}$.*
(2) *If $\psi = \psi(c)$ is the corresponding electrostatic potential, then the Boltzmann distributions* (1.9) *for point ions holds true and ψ is the unique weak solution to the corresponding boundary-value problem of PBE* (1.11) *and* (1.6). *Moreover,*

$$\min_{d \in W_0} F_0[d] = \frac{1}{2} \sum_{i=1}^{N} Q_i(\psi - \psi_{vac})(x_i) + \int_\Gamma \frac{1}{8\pi} \left(\psi - \hat{\phi}_0\right) \varepsilon_\Gamma \partial_n \left(\psi - \hat{\psi}_0\right) dS$$

$$- \int_{\Omega_s} \frac{\varepsilon_s}{8\pi} \left|\nabla \left(\psi - \hat{\psi}_0\right)\right|^2 dx - \beta^{-1} \sum_{j=1}^{M} \int_{\Omega_s} c_j^\infty e^{-\beta q_j \left(\psi - \hat{\psi}_0/2\right)} dx. \tag{2.7}$$

THEOREM 2.5.

(1) *There exists a unique equilibrium $c = (c_1, \ldots, c_M)$ of $F_a : V_a \to \mathbb{R}$. It is also the unique global minimizer and the unique local minimizer of $F_a : V_a \to \mathbb{R}$.*
(2) *If $\psi = \psi(c)$ is the corresponding electrostatic potential, then the Boltzmann distributions* (1.9) *for finite-size ions holds true and ψ is the unique weak solution to the corresponding boundary-value problem of PBE* (1.12) *and* (1.6).

Moreover,

$$\min_{d\in V_a} F_a[d] = \frac{1}{2}\sum_{i=1}^{N} Q_i(\psi-\psi_{vac})(x_i) + \int_\Gamma \frac{1}{8\pi}\left(\psi-\hat{\phi}_0\right)\varepsilon_\Gamma \partial_n\left(\psi-\hat{\psi}_0\right) dS$$
$$- \int_{\Omega_s} \frac{\varepsilon_\Gamma}{8\pi}\left|\nabla\left(\psi-\hat{\psi}_0\right)\right|^2 dx - \beta^{-1}a^{-3}\int_{\Omega_s} \left[1+\log\left(1+a^3\sum_{j=1}^{M} c_j^\infty e^{-\beta q_j\left(\psi-\hat{\psi}_0/2\right)}\right)\right] dx. \tag{2.8}$$

3. Some lemmas. The key point of our first lemma below is the existence and continuity across the interface Γ of the normal flux for a solution of an elliptic interface problem. In terms of electrostatics, this means that the electrostatic potential and the normal component of electrostatic displacement are continuous across dielectric boundaries. These seem to be known results. For completeness, we give a proof here.

LEMMA 3.1. *Let $U \subset \mathbb{R}^3$ be an open set such that $\Gamma \subset U \subseteq \Omega$. Let $g \in L^1(U)\cap H^{-1}(U)$. Suppose $u \in H^1(U)$ satisfies*

$$\int_U \varepsilon_\Gamma \nabla u\cdot\nabla\eta dx = \int_U g\eta dx \qquad \forall \eta \in C_c^\infty(U). \tag{3.1}$$

Then $[\![u]\!] = 0$ on Γ. If in addition $g \in L^2(U)$, then $[\![\varepsilon_\Gamma \partial_n u]\!] = 0$ on Γ.

Proof. Fix an open ball $B \subset U$ such that $\Gamma\cap B \neq \emptyset$. Let $\eta \in C_c^\infty(U)$ with $\operatorname{supp}\eta \subset B$. Let n_j with $1 \le j \le 3$ be the jth component of n, the unit normal at the Γ, pointing from Ω_s to Ω_m. It follows from the fact that $u \in H^1(\Omega)$ and integration by parts that

$$\begin{aligned}
-\int_B u\partial_j\eta\, dx &= \int_B (\partial_j u)\eta\, dx \\
&= \int_{B\cap\Omega_m} (\partial_j u)\eta\, dx + \int_{B\cap\Omega_s} (\partial_j u)\eta\, dx \\
&= -\int_{B\cap\Omega_m} u\partial_j\eta\, dx - \int_{\Gamma\cap B} u_m\eta n_j\, dS - \int_{B\cap\Omega_s} u\partial_j\eta\, dx + \int_{\Gamma\cap B} u_s\eta n_j\, dS \\
&= -\int_B u\partial_j\eta\, dx + \int_{\Gamma\cap B}(u_s-u_m)\eta n_j\, dS, \qquad j=1,2,3.
\end{aligned}$$

This and the arbitrariness of η imply $[\![u]\!] = 0$ on Γ.

To show the continuity of $\varepsilon_\Gamma\nabla u\cdot n$ across Γ, we fix an open set $U_0 \subset \mathbb{R}^3$ such that $\Gamma \subset U_0 \subset \overline{U_0} \subset U$ and that the boundary ∂U_0 is C^2. By the fact that $u \in H^1(U)$ and $g \in L^2(U)$, and by (3.1), we have

$$(\varepsilon_t\nabla u_t)|_{U_0\cap\Omega_t} \in L^2\left(U_0\cap\Omega_t, \mathbb{R}^3\right) \quad \text{and}$$
$$(\nabla\cdot\varepsilon_t\nabla u_t)|_{U_0\cap\Omega_t} = -g \in L^2(U_0\cap\Omega_t), \quad t = m, s.$$

Therefore, by Theorem 1.2 in [30], the trace of $(\varepsilon_t\nabla u_t)|_{U_0\cap\Omega_t}\cdot\nu \in H^{-1/2}(\partial(U_0\cap\Omega_t))$ exists, and also by (3.1),

$$\int_{U_0\cap\Omega_t} \varepsilon_t\nabla u_t\cdot\nabla\eta dx = \int_{U_0\cap\Omega_t} g\eta dx + \int_{\partial(U_0\cap\Omega_t)} (\varepsilon_t\nabla u_t\cdot\nu)\eta dS \quad \forall\eta\in C_c^\infty(U_0\cap\Omega_t), \tag{3.2}$$

where ν denotes the unit exterior normal of the boundary $\partial(U_0 \cap \Omega_t)$ which contains Γ and $t = m, s$. Notice that the normals ν at Γ from both sides $U_0 \cap \Omega_m$ and $U_0 \cap \Omega_s$ are in opposite directions.

These traces are determined independent of the choice of U_0. In fact, if $Q_0 \subset \mathbb{R}^3$ is another open set such that $\Gamma \subset Q_0 \subset \overline{Q_0} \subset U$ and the boundary ∂Q_0 is C^2, then the traces $(\varepsilon_m \nabla u_m)|_{Q_0 \cap \Omega_m} \cdot \nu \in H^{-1/2}(\partial(Q_0 \cap \Omega_m))$ and $(\varepsilon_s \nabla u_s)|_{Q_0 \cap \Omega_s} \cdot \nu \in H^{-1/2}(\partial(Q_0 \cap \Omega_m))$ exist, and (3.2) holds true for $t = m, s$ when U_0 is replaced by Q_0. Consider now (3.2) with $t = m$. Choose any $\eta \in C_c^1(U_0 \cap Q_0)$ such that $\eta = 0$ on $\partial(U_0 \cap \Omega_m) \setminus \Gamma$ and on $\partial(Q_0 \cap \Omega_m) \setminus \Gamma$. Extend η by $\eta = 0$ to outside $U_0 \cap Q_0$. By (3.2) with $t = m$ and the corresponding equation with U_0 replaced by Q_0, we obtain that

$$\int_{\partial(U_0 \cap \Omega_m)} (\varepsilon_m \nabla u_m \cdot \nu) \eta dS = \int_{\partial(Q_0 \cap \Omega_m)} (\varepsilon_m \nabla u_m \cdot \nu) \eta dS.$$

The arbitrariness of η then implies that the trace of $(\varepsilon_m \nabla u_m \cdot \nu)|_{U_0 \cap \Omega_m}$ on Γ determined by U_0 is the same as that determined by Q_0. By the same argument, we see that the trace of $(\varepsilon_s \nabla u_s \cdot n)|_{U_0 \cap \Omega_s}$ on Γ determined by U_0 is the same as that determined by Q_0.

Now, by the fact that $U_0 = (U_0 \cap \Omega_m) \cup (U_0 \cap \Omega_s)$ and $(U_0 \cap \Omega_m) \cap (U_0 \cap \Omega_s) = \emptyset$, and by our convention for the direction of the unit normal n along Γ, we obtain from (3.1) and (3.2) that for any $\eta \in C_c^\infty(U)$ with $\operatorname{supp} \eta \subset U_0$

$$\int_\Gamma \left(\varepsilon_m \frac{\partial u_m}{\partial n} - \varepsilon_s \frac{\partial u_s}{\partial n} \right) \eta dS = 0.$$

The arbitrariness of η implies $[\![\varepsilon_\Gamma \partial_n u]\!] = 0$ on Γ. □

Let $L : H^{-1}(\Omega) \to H^1(\Omega)$ be the linear operator defined as follows: for any $\xi \in H^{-1}(\Omega)$, $L\xi \in H_0^1(\Omega)$ is the unique function in $H_0^1(\Omega)$ that satisfies

$$\frac{1}{4\pi} \int_\Omega \varepsilon_\Gamma \nabla (L\xi) \cdot \nabla v \, dx = \xi(v) \qquad \forall v \in H_0^1(\Omega). \tag{3.3}$$

It is easy to see that $\langle \xi, \eta \rangle = \xi(L\eta)$ defines an inner product of $H^{-1}(\Omega)$. Denote by $|\!|\!| \cdot |\!|\!|$ the corresponding norm of $H^{-1}(\Omega)$, i.e., $|\!|\!| \xi |\!|\!| = \sqrt{\langle \xi, \xi \rangle} = \sqrt{\xi(L\xi)}$ for any $\xi \in H^{-1}(\Omega)$. One can verify that there exist $C_1 = C_1(\Omega, \varepsilon_m, \varepsilon_s) > 0$ and $C_2 = C_2(\varepsilon_m, \varepsilon_s) > 0$ such that

$$C_1 |\!|\!| \xi |\!|\!| \le \|\xi\|_{H^{-1}(\Omega)} \le C_2 |\!|\!| \xi |\!|\!| \qquad \forall \xi \in H^{-1}(\Omega). \tag{3.4}$$

It follows from [21] (with minor modifications) that there exists a unique $G \in H_*^1(\Omega)$ such that $G = 0$ on $\partial\Omega$ and

$$\int_\Omega \varepsilon_\Gamma \nabla G \cdot \nabla \eta dx = 4\pi \sum_{i=1}^N Q_i \eta(x_i) \qquad \forall \eta \in C_c^\infty(\Omega). \tag{3.5}$$

Clearly, $G - \psi_{vac}$ is harmonic in Ω_m and $G \in W^{1,p}(\Omega)$ for any $p \in [1, 3/2)$. Notice that the function $\hat{\psi}_0 \in H^1(\Omega)$ defined in (1.10) is harmonic in $\Omega_m \cup \Omega_s$.

The next lemma gives a solution decomposition of the Poisson equation (1.5) with its right-hand side consisting of Dirac masses and a function in $H^{-1}(\Omega)$ that represents the density of ionic charges. This decomposition is a mathematical formulation of the Born cycle [2].

LEMMA 3.2. *Let $f \in L^1(\Omega) \cap H^{-1}(\Omega)$ be such that $f = 0$ in Ω_m. Then $\psi := G + \hat{\psi}_0 + Lf$ is the unique function in $H^1_*(\Omega)$ that satisfies $\psi = \psi_0$ on $\partial\Omega$ and*

$$\int_\Omega \varepsilon_\Gamma \nabla\psi \cdot \nabla\eta dx = 4\pi \sum_{i=1}^N Q_i \eta(x_i) + 4\pi \int_{\Omega_s} f\eta dx \qquad \forall \eta \in C_c^\infty(\Omega). \tag{3.6}$$

Moreover, Lf and $\psi - \psi_{vac}$ are harmonic in Ω_m, and

$$\sum_{i=1}^N Q_i \left(Lf\right)(x_i) = \int_{\Omega_s} Gf dx. \tag{3.7}$$

Proof. From the definition of G, $\hat{\psi}_0$, and L (cf. (3.5), (1.10), and (3.3)), we easily verify that the function ψ is in $H^1_*(\Omega)$, $\psi = \psi_0$ on $\partial\Omega$, and (3.6) holds true. If $\bar{\psi} \in H^1_*(\Omega)$ satisfies $\bar{\psi} = 0$ on $\partial\Omega$ and $\int_\Omega \varepsilon_\Gamma \nabla\bar{\psi} \cdot \nabla\eta \, dx = 0$ for all $\eta \in C_c^\infty(\Omega)$, then clearly $\bar{\psi} \in H^1_0(\Omega)$ and in fact $\bar{\psi} = 0$ a.e. Ω. This proves the needed uniqueness.

By the fact that $f = 0$ in Ω_m and the definition of L (cf. (3.3)), Lf is harmonic in Ω_m. The fact that $\psi - \psi_{vac}$ is harmonic in Ω_m follows from (3.6) with $\eta \in C_c^\infty(\Omega)$ so chosen that $\operatorname{supp}\eta \subset \Omega_m$ and

$$\int_{\Omega_m} \varepsilon_m \nabla\psi_{vac} \cdot \nabla\eta dx = 4\pi \sum_{i=1}^N Q_i \eta(x_i) \qquad \forall \eta \in C_c^\infty(\Omega_m).$$

It remains to prove (3.7). Denote $\psi_c = Lf \in H^1_0(\Omega)$. Let $\alpha > 0$ be sufficiently small and let $B_\alpha = \cup_{i=1}^N B(x_i, \alpha)$. By the fact that G is harmonic in $\Omega_m \setminus B_\alpha$ and $G - \psi_{vac}$ is harmonic in Ω_m, we obtain by a series of routine calculations that

$$\begin{aligned}
\int_{\Omega_m} \varepsilon_m \nabla G \cdot \nabla\psi_c dx &= \int_{\Omega_m \setminus B_\alpha} \varepsilon_m \nabla G \cdot \nabla\psi_c dx + \int_{B_\alpha} \varepsilon_m \nabla G \cdot \nabla\psi_c dx \\
&= -\int_{\Omega_m \setminus B_\alpha} \varepsilon_m (\Delta G)\psi_c dx + \int_{\partial(\Omega_m \setminus B_\alpha)} \varepsilon_m \psi_c \frac{\partial G}{\partial\nu} dS + O(\alpha) \\
&= -\int_\Gamma \varepsilon_m \psi_c|_m \frac{\partial G|_m}{\partial n} dS + \int_{\partial B_\alpha} \varepsilon_m \psi_c \frac{\partial G}{\partial\nu} dS + O(\alpha) \\
&= -\int_\Gamma \varepsilon_m \psi_c|_m \frac{\partial G|_m}{\partial n} dS + \int_{\partial B_\alpha} \varepsilon_m \psi_c \frac{\partial \left(G - \psi_{vac}\right)}{\partial\nu} dS \\
&\quad + \sum_{i=1}^N \int_{\partial B(x_i,\alpha)} \varepsilon_m \psi_c \frac{\partial\psi_{vac}}{\partial\nu} dS + O(\alpha) \\
&\to -\int_\Gamma \varepsilon_m \psi_c|_m \frac{\partial G|_m}{\partial n} dS + 4\pi \sum_{i=1}^N Q_i \psi_c(x_i) \qquad \text{as } \alpha \to 0,
\end{aligned}$$

where ν is the exterior unit normal of $\partial(\Omega_m \setminus B_\alpha)$ and $\nu = -n$ on Γ by our convention for the direction of n. Consequently, by the continuity of $\varepsilon_\Gamma \nabla G \cdot n$ across Γ (cf.

Lemma 3.1), the fact that G is harmonic in Ω_s, and $G = \psi_c = 0$ on $\partial\Omega$, we obtain

$$\begin{aligned} 4\pi \sum_{i=1}^{N} Q_i \psi_c(x_i) &= \int_\Gamma \varepsilon_s \psi_c|_s \frac{\partial G|_s}{\partial n} dS + \int_{\Omega_m} \varepsilon_m \nabla G \cdot \nabla \psi_c dx \\ &= \int_{\Omega_s} \varepsilon_s (\Delta G) \psi_c dx + \int_{\Omega_s} \varepsilon_s \nabla G \cdot \nabla \psi_c dx + \int_{\Omega_m} \varepsilon_m \nabla G \cdot \nabla \psi_c dx \\ &= \int_\Omega \varepsilon_\Gamma \nabla G \cdot \nabla \psi_c dx. \end{aligned} \tag{3.8}$$

Since $\psi_c = Lf \in H_0^1(\Omega)$ is harmonic in Ω_m, we also have by the properties of G (cf. [21]) and integration by parts that

$$\begin{aligned} \int_{\Omega_m} \varepsilon_m \nabla G \cdot \nabla \psi_c dx &= \int_{\Omega_m \setminus B_\alpha} \varepsilon_m \nabla G \cdot \nabla \psi_c dx + \int_{B_\alpha} \varepsilon_m \nabla G \cdot \nabla \psi_c dx \\ &= -\int_{\Omega_m \setminus B_\alpha} \varepsilon_m G \Delta \psi_c dx + \int_{\partial(\Omega_m \setminus B_\alpha)} \varepsilon_m G \frac{\partial \psi_c}{\partial \nu} dS + O(\alpha) \\ &= -\int_\Gamma \varepsilon_m G|_m \frac{\partial \psi_c|_m}{\partial n} dS + \int_{\partial B_\alpha} \varepsilon_m G \frac{\partial \psi_c}{\partial \nu} dS + O(\alpha) \\ &\to -\int_\Gamma \varepsilon_m G|_m \frac{\partial \psi_c|_m}{\partial n} dS \qquad \text{as } \alpha \to 0. \end{aligned} \tag{3.9}$$

Let $\hat{G} \in H^1(\Omega_m)$ be such that $\hat{G} = G$ on $\Gamma = \partial\Omega_m$. Replacing G in (3.9) by $\hat{G}$ and repeating the same calculations, we obtain

$$\int_{\Omega_m} \varepsilon_m \nabla G \cdot \nabla \psi_c dx = \int_{\Omega_m} \varepsilon_m \nabla \hat{G} \cdot \nabla \psi_c dx. \tag{3.10}$$

Define $\overline{G} : \Omega \to \mathbb{R}$ by $\overline{G}(x) = \hat{G}(x)$ if $x \in \overline{\Omega_m}$ and by $\overline{G}(x) = G(x)$ if $x \in \Omega_s$. Clearly, $\overline{G} \in H_0^1(\Omega)$. Since $\psi_c = Lf$ and $f = 0$ in Ω_m, we, thus, have by (3.8) and (3.10) that

$$\begin{aligned} 4\pi \sum_{i=1}^{N} Q_i \psi_c(x_i) &= \int_{\Omega_m} \varepsilon_m \nabla \hat{G} \cdot \nabla \psi_c dx + \int_{\Omega_s} \varepsilon_s \nabla G \cdot \nabla \psi_c dx \\ &= \int_\Omega \varepsilon_\Gamma \nabla \overline{G} \cdot \nabla \psi_c dx = 4\pi \int_\Omega \overline{G} f dx = 4\pi \int_{\Omega_s} G f dx. \end{aligned}$$

This implies (3.7). □

By Lemma 3.2, the potential $\psi = \psi(c_1, \ldots, c_M)$ corresponding to a set of concentrations $(c_1, \ldots, c_M)$ is well defined with $f = \sum_{j=1}^{M} q_j c_j$ and is given by

$$\psi(c_1, \ldots, c_M) = G + \hat{\psi}_0 + L\left(\sum_{j=1}^{M} q_j c_j\right). \tag{3.11}$$

Moreover, the functional $F_0 : V_0 \to \mathbb{R}$ and $F_a : V_a \to \mathbb{R}$ can be rewritten as

$$\begin{aligned} F_0[c] &= \frac{1}{2}\int_{\Omega_s}\left(\sum_{j=1}^{M} q_j c_j\right) L\left(\sum_{j=1}^{M} q_j c_j\right) dx + \sum_{j=1}^{M}\int_{\Omega_s} \mu_{0j} c_j dx \\ &\quad + \beta^{-1}\sum_{j=1}^{M}\int_{\Omega_s} S_{-1}(c_j)dx + E_0, \qquad \forall c = (c_1, \ldots, c_M) \in V_0, \end{aligned} \tag{3.12}$$

$$\begin{aligned} F_a[c] &= \frac{1}{2}\int_{\Omega_s}\left(\sum_{j=1}^{M} q_j c_j\right) L\left(\sum_{j=1}^{M} q_j c_j\right) dx + \sum_{j=1}^{M}\int_{\Omega_s} \mu_{aj} c_j dx \\ &\quad + \beta^{-1}\sum_{j=0}^{M}\int_{\Omega_s} S_{-1}(c_j)dx + E_a \qquad \forall c = (c_1, \ldots, c_M) \in V_a, \end{aligned} \tag{3.13}$$

respectively, where

$$\mu_{0j}(x) = q_j G(x) + \frac{1}{2} q_j \hat{\psi}_0(x) + 3\beta^{-1}\log a - \mu_j \qquad \forall x \in \Omega_s,\ j = 1, \ldots, M, \tag{3.14}$$

$$E_0 = \frac{1}{2}\sum_{i=1}^{N} Q_i \left(G + \hat{\psi}_0 - \psi_{vac}\right)(x_i), \tag{3.15}$$

$$\mu_{aj}(x) = q_j G(x) + \frac{1}{2} q_j \hat{\psi}_0(x) - \mu_j \qquad \forall x \in \Omega_s,\ j = 1, \ldots, M, \tag{3.16}$$

$$E_a = \frac{1}{2}\sum_{i=1}^{N} Q_i \left(G + \hat{\psi}_0 - \psi_{vac}\right)(x_i) + 3\beta^{-1}a^{-3}(\log a)|\Omega_s|, \tag{3.17}$$

where $|E|$ denotes the Lebesgue measure of a Lebesgue measurable set $E \subset \mathbb{R}^3$.

LEMMA 3.3. *Let $D \subset \mathbb{R}^3$ be a bounded and open set. Let $\alpha \in \mathbb{R}$. Let $\{u^{(k)}\}$ be a sequence of functions in $L^1(D)$ such that $u^{(k)} \geq 0$ a.e. D for each $k \geq 1$ and that*

$$\sup_{k\geq 1}\int_D S_\alpha\left(u^{(k)}\right) dx < \infty.$$

Then there exists a subsequence $\{u^{(k_j)}\}$ of $\{u^{(k)}\}$ such that $\{u^{(k_j)}\}$ converges weakly in $L^1(D)$ to some $u \in L^1(D)$ with $u \geq 0$ a.e. D and

$$\int_D S_\alpha(u)\, dx \leq \liminf_{k\to\infty}\int_D S_\alpha\left(u^{(k)}\right) dx.$$

Proof. Since $S_\alpha : [0, \infty) \to \mathbb{R}$ is bounded below, by passing to a subsequence if necessary, we may assume that the limit

$$A := \lim_{k\to\infty}\int_D S_\alpha\left(u^{(k)}\right) dx = \liminf_{k\to\infty}\int_D S_\alpha\left(u^{(k)}\right) dx \tag{3.18}$$

exists and is finite. Since $S_\alpha(\lambda)/\lambda \to +\infty$ as $\lambda \to +\infty$, $\{u^{(k)}\}$ is weakly sequentially compact in $L^1(D)$ by de la Vallée Poussin's criterion [25]. Therefore, this sequence has a subsequence, not relabeled, that converges weakly in $L^1(D)$ to some $u \in L^1(D)$. Clearly, $u \geq 0$ a.e. D.

Let $\varepsilon > 0$. By (3.18), there exists an integer $K > 0$ such that

$$\int_D S_\alpha\left(u^{(k)}\right) dx \leq A + \varepsilon \qquad \forall k > K. \tag{3.19}$$

By Mazur's theorem [7, 32], there exist convex combinations $v^{(k)}$ of $u^{(K+1)}, \dots, u^{(K+k)}$ for all $k \geq 1$ such that $v^{(k)} \to u$ in $L^1(D)$. Let $v^{(k)} = \sum_{j=1}^k \lambda_{k,j} u^{(K+j)}$ with $\lambda_{k,j} \geq 0$ for all j and k, and $\sum_{j=1}^k \lambda_{k,j} = 1$ for all k. Since $S_\alpha : [0,\infty) \to \mathbb{R}$ is convex, we have by Jensen's inequality and (3.19) that

$$S_\alpha\left(v^{(k)}\right) \leq \sum_{j=1}^k \lambda_{k,j} S_\alpha\left(u^{(K+j)}\right) \leq \sum_{j=1}^k \lambda_{k,j}(A+\varepsilon) = A + \varepsilon \qquad \forall k \geq 1. \tag{3.20}$$

Since $v^{(k)} \to u$ in $L^1(D)$, there exists a subsequence $\{v^{(k_j)}\}$ of $\{v^{(k)}\}$ such that $v^{k_j}(x) \to u(x)$ a.e. $x \in D$. Consequently, since $S_\alpha : [0,\infty) \to \mathbb{R}$ is continuous and bounded below, we have by Fatou's lemma and (3.20) that

$$\int_D S_\alpha(u(x))\, dx = \int_D \lim_{j\to\infty} S_\alpha\left(v^{k_j}(x)\right)\, dx \leq \liminf_{j\to\infty} \int_D S_\alpha\left(v^{k_j}(x)\right)\, dx \leq A + \varepsilon,$$

concluding the proof by the arbitrariness of $\varepsilon > 0$. ∎

The next two lemmas state some boundedness of concentrations that have low free energies. Their proofs are somewhat tedious, and are given in Appendix A.

LEMMA 3.4. *Let $c = (c_1, \dots, c_M) \in W_0$ satisfy that $c \notin L^\infty(\Omega, \mathbb{R}^M)$ or there exists $j \in \{1, \dots, M\}$ with $|\{x \in \Omega_s : c_j(x) < \alpha\}| > 0$ for all $\alpha > 0$. Then for any $\varepsilon > 0$ there exist $\hat{c} = (\hat{c}_1, \dots, \hat{c}_M) \in W_0$ that satisfies (2.5) with c replaced by $\hat{c}$, $\|\hat{c} - c\|_X < \varepsilon$, and $F_0[\hat{c}] < F_0[c]$.*

LEMMA 3.5. *Let $c = (c_1, \dots, c_M) \in V_a$ and c_0 be defined by (1.3). Assume there exists $j \in \{0, 1, \dots, M\}$ such that $|\{x \in \Omega_s : a^3 c_j(x) < \alpha\}| > 0$ for all $\alpha > 0$. Let $\varepsilon > 0$. Then there exists $\hat{c} = (\hat{c}_1, \dots, \hat{c}_M) \in V_a$ that satisfies (2.6) with c replaced by $\hat{c}$, $\|\hat{c} - c\|_X < \varepsilon$, and $F_a[\hat{c}] < F_a[c]$.*

4. The Poisson–Boltzmann equation: Proof of Theorems 2.1 and 2.2. *Proof of Theorem* 2.1. It is easy to verify that the function $B : \mathbb{R} \to \mathbb{R}$ defined in (1.14) is convex for both the case of point ions and that of finite-size ions. Let

$$\mathcal{K} := \left\{u \in H^1(\Omega) : u = \psi_0 \text{ on } \partial\Omega \text{ and } \chi_{\Omega_s} B(u) \in L^2(\Omega)\right\}.$$

Clearly, $\mathcal{K} \neq \emptyset$ since $\psi_0 \in \mathcal{K}$ and $\mathcal{K}$ is convex since $B : \mathbb{R} \to \mathbb{R}$ is convex. We show now that $\mathcal{K}$ is closed in $H^1(\Omega)$. Let $u_k \in \mathcal{K}$ $(k = 1, 2, \dots)$ and $u_k \to u$ in $H^1(\Omega)$

for some $u \in H^1(\Omega)$. Clearly, $u = \psi_0$ on $\partial\Omega$. Up to a subsequence, not relabeled, $u_k(x) \to u(x)$ a.e. $x \in \Omega$. Since $B : \mathbb{R} \to \mathbb{R}$ is convex and positive, we have

$$\frac{d^2}{dv^2}\left([B(v)]^2\right) = 2[B'(v)]^2 + 2B(v)B''(v) > 0 \qquad \forall v \in \mathbb{R}.$$

Thus, $v \mapsto [B(v)]^2$ is convex. It then follows from Fatou's lemma, Jensen's inequality, and the $H^1(\Omega)$-boundedness of $\{u_k\}$ that

$$\begin{aligned}\frac{1}{|\Omega_s|}\int_{\Omega_s}[B(u)]^2\,dx &\le \liminf_{k\to\infty}\frac{1}{|\Omega_s|}\int_{\Omega_s}[B(u_k)]^2\,dx\\ &\le \liminf_{k\to\infty}\left[B\left(\frac{1}{|\Omega_s|}\int_{\Omega_s}u_k dx\right)\right]^2 < \infty.\end{aligned}$$

This implies that $u \in \mathcal{K}$. Therefore, $\mathcal{K}$ is closed in $H^1(\Omega)$. Since $\mathcal{K}$ is convex, it is also weakly closed in $H^1(\Omega)$.

Define now $J : \mathcal{K} \to \mathbb{R}$ by

$$J[u] = \int_\Omega \left[\frac{\varepsilon_\Gamma}{2}|\nabla u|^2 + 4\pi\chi_{\Omega_s}B\left(u + G - \frac{\hat{\psi}_0}{2}\right)\right]dx \qquad \forall u \in \mathcal{K},$$

where G and $\hat{\psi}_0$ are defined in (3.5) and (1.10), respectively. Note that $\psi_0 \in \mathcal{K}$ and that $J[\psi_0] < \infty$. By the Poincaré inequality, there exist constants $C_3 > 0$ and $C_4 \ge 0$ such that $J[u] \ge C_3\|u\|^2_{H^1(\Omega)} - C_4$ for all $u \in \mathcal{K}$. Thus, $\alpha := \inf_{u\in\mathcal{K}} J[u]$ is finite. Let $v_k \in \mathcal{K}$ $(k = 1, 2\dots)$ be such that $\lim_{k\to\infty} J[v_k] = \alpha$. Then, $\{v_k\}$ is bounded in $H^1(\Omega)$ and, hence, it has a subsequence, not relabeled, that weakly converges to some $v \in H^1(\Omega)$. Since $\mathcal{K}$ is weakly closed, $v \in \mathcal{K}$. Since the embedding $H^1(\Omega) \hookrightarrow L^2(\Omega)$ is compact, up to a further subsequence, again not relabeled, $v_k \to v$ a.e. in Ω. Therefore, since $B : \mathbb{R} \to \mathbb{R}$ is continuous and nonnegative, Fatou's lemma implies

$$\liminf_{k\to\infty}\int_\Omega \chi_{\Omega_s}B\left(v_k + G - \frac{\hat{\psi}_0}{2}\right)dx \ge \int_\Omega \chi_{\Omega_s}B\left(v + G - \frac{\hat{\psi}_0}{2}\right)dx.$$

Since $u \mapsto \int_\Omega \varepsilon_\Gamma|\nabla u|^2 dx$ is convex and $H^1(\Omega)$ continuous, it is sequentially weakly lower semicontinuous. Consequently, $\liminf_{k\to\infty} J[v_k] \ge J[v]$. Thus, v is a minimizer of $J : \mathcal{K} \to \mathbb{R}$.

Notice that $\chi_{\Omega_s}B'(v + G - \hat{\psi}_0/2) \in L^2(\Omega_s)$. Simple calculations of the first variation of $J : \mathcal{K} \to \mathbb{R}$ at any $\eta \in C_c^\infty(\Omega)$ leads to

$$\int_\Omega \left[\varepsilon_\Gamma\nabla v\cdot\nabla\eta + 4\pi\chi_{\Omega_s}B'\left(v + G - \frac{\hat{\psi}_0}{2}\right)\eta\right]dx = 0 \qquad \forall \eta \in C_c^\infty(\Omega).$$

The function $\psi = v + G$ is, thus, a needed solution.

We now prove the uniqueness. Let ϕ be another weak solution. Let $\xi = \psi - \phi$. Then, $\xi \in H^1_*(\Omega)$, $\xi = 0$ on $\partial\Omega$, and

$$\int_\Omega \left\{\varepsilon_\Gamma\nabla\xi\cdot\nabla\eta + 4\pi\chi_{\Omega_s}\left[B'\left(\psi - \frac{\hat{\psi}_0}{2}\right) - B'\left(\phi - \frac{\hat{\psi}_0}{2}\right)\right]\eta\right\}dx = 0 \qquad \forall \eta \in C_c^\infty(\Omega).$$

Choosing the test functions $\eta \in C_c^\infty(\Omega)$ so that $\operatorname{supp}\eta \subset \Omega_m$, we find that ξ is harmonic in Ω_m. This and the fact that $\xi \in H^1_*(\Omega)$ imply that $\xi \in H^1_0(\Omega)$. Thus, the above test functions η can be chosen from $H^1_0(\Omega)$. In particular, setting $\eta = \xi$ and using the convexity of $B : \mathbb{R} \to \mathbb{R}$, we obtain that $\xi = 0$ and, hence, $\psi = \phi$ in $H^1(\Omega)$.

Let $\sigma > 0$ be such that the closure of $B_\sigma := \cup_{i=1}^N B(x_i, \sigma)$ is contained in Ω_m. Clearly, the unique weak solution $\psi \in H^1_*(\Omega)$ satisfies

$$\int_{\Omega\setminus\overline{B_\sigma}} \varepsilon_\Gamma \nabla\psi\cdot\nabla\eta dx = \int_{\Omega\setminus\overline{B_\sigma}} g\eta dx \qquad \forall \eta \in C_c^\infty\left(\Omega\setminus\overline{B_\sigma}\right), \tag{4.1}$$

where $g = -4\pi\chi_{\Omega_s}B'(\psi - \hat{\psi}_0/2) \in L^2(\Omega\setminus\overline{B_\sigma})$. Therefore, $\psi \in C(\overline{\Omega}\setminus B_\sigma)$ by the standard regularity theory [14]. Since $\varepsilon_\Gamma = \varepsilon_m$ in Ω_m and $\varepsilon_\Gamma = \varepsilon_s$ in Ω_s, ψ is harmonic in $\Omega_m\setminus\overline{B_\sigma}$. Hence, $\psi \in C^\infty(\Omega_m\setminus\overline{B_\sigma})$. Notice that $B \in C^\infty(\mathbb{R})$; thus, we have $\psi \in C^\infty(\Omega_s)$ by a standard bootstrapping argument. □

Proof of Theorem 2.2. Let $\psi \in H^1_*(\Omega)$ be a weak solution to the boundary-value problem (1.13) and (1.6). Clearly, $\psi_m \in H^1_*(\Omega_m)$. For any $\eta \in C_c^\infty(\Omega_m)$, we extend η to the entire Ω by defining $\eta = 0$ outside Ω_m. Then, we obtain (2.2) from (2.1). Since all $x_i \in \Omega_m$ $(i = 1, \dots, N)$, we have $\psi_s \in H^1(\Omega_s)$. Since $\chi_{\Omega_s}B(\psi) \in L^2(\Omega_s)$, it follows from (1.14) that $\chi_{\Omega_s}B'(\psi) \in L^2(\Omega_s)$. For any $\eta \in C_c^\infty(\Omega_s)$, we, again, extend η to Ω by defining $\eta = 0$ outside Ω_s. Then, we obtain (2.3) from (2.1). Finally, by Lemma 3.1, (4.1), and (1.6), the last two equations in (1.15) hold true. Hence, ψ is a weak solution to (1.15).

Now let $\psi : \Omega \to \mathbb{R}$ be a solution to the boundary-value problem (1.15). We first show that $\psi \in H^1_*(\Omega)$. Let $\sigma > 0$ be small enough so that the closure of $B_\sigma := \cup_{i=1}^N B(x_i, \sigma)$ is contained in Ω_m. Since $\psi_m \in H^1_*(\Omega_m)$, $\psi_m \in H^1(\Omega_m\setminus\overline{B_\sigma})$. Thus, the trace $\psi_m|_\Gamma \in L^2(\Gamma)$, and is independent on the choice of σ. Similarly, $\psi_s \in H^1(\Omega_s)$, and, hence, $\psi_s|_\Gamma \in L^2(\Gamma)$. Fix $j \in \{1, 2, 3\}$. Define $\xi_j : \Omega\setminus\overline{B_\sigma} \to \mathbb{R}$ by $\xi_j = \partial_j\psi_m$ in $\Omega_m\setminus\overline{B_\sigma}$ and $\xi_j = \partial_j\psi_s$ in Ω_s. Clearly, $\xi_j \in L^2(\Omega\setminus\overline{B_\sigma})$. Let n_j be the jth component of the unit exterior normal n at Γ, pointing from Ω_s to Ω_m. Then, for any $\eta \in C_c^\infty(\Omega\setminus\overline{B_\sigma})$, we have

$$\begin{aligned}
\int_{\Omega\setminus\overline{B_\sigma}} \xi_j\eta\, dx &= \int_{\Omega_m\setminus\overline{B_\sigma}} (\partial_j\psi_m)\eta\, dx + \int_{\Omega_s} (\partial_j\psi_s)\eta\, dx \\
&= -\int_{\Omega_m\setminus\overline{B_\sigma}} \psi\partial_j\eta\, dx - \int_\Gamma \psi_m\eta n_j\, dS - \int_{\Omega_s} \psi\partial_j\eta\, dx + \int_\Gamma \psi_s\eta n_j\, dS \\
&= -\int_{\Omega\setminus\overline{B_\sigma}} \psi\partial_j\eta\, dx + \int_\Gamma [\![\psi]\!] n_j\eta\, dS \\
&= -\int_{\Omega\setminus\overline{B_\sigma}} \psi\partial_j\eta\, dx,
\end{aligned}$$

where in the last step we used the fact that $[\![\psi]\!] = 0$ on Γ. Thus, $\xi_j = \partial_j\psi \in L^2(\Omega\setminus\overline{B_\sigma})$, and, hence, $\psi \in H^1_*(\Omega)$ by the arbitrariness of $\sigma > 0$.

Clearly, $\chi_{\Omega_s}B'(\psi) \in L^2(\Omega_s)$ and $\psi = \psi_0$ on $\partial\Omega$. It remains to show that (2.1) holds true. Let $\eta \in C_c^\infty(\Omega)$. Let V_1 and V_2 be two open sets in $\mathbb{R}^3$ such that ∂V_1 and ∂V_2 are of C^2, $x_i \notin V_2$ for $i = 1, \dots, N$, and $\Gamma \subset V_1 \subset \overline{V}_1 \subset V_2 \subset \overline{V}_2 \subset \Omega$. Let

$\zeta \in C_c^\infty(\Omega)$ be such that $\operatorname{supp}\zeta \subset V_2$ and $\zeta = 1$ on V_1. Then, $(1-\zeta)\eta|_{\Omega_m} \in C_c^\infty(\Omega_m)$, $(1-\zeta)\eta|_{\Omega_s} \in C_c^\infty(\Omega_s)$, and $(1-\zeta(x_i))\eta(x_i) = \eta(x_i),\ i = 1, \ldots, N$. We, thus, have by (2.2) and (2.3) that

$$
\begin{aligned}
&\int_\Omega \left[\varepsilon_\Gamma \nabla\psi\cdot\nabla\eta + 4\pi\chi_{\Omega_s} B'\left(\psi - \frac{\hat\psi_0}{2}\right)\eta\right] dx \\
&\quad = \int_\Omega \left[\varepsilon_\Gamma \nabla\psi\cdot\nabla((1-\zeta)\eta) + 4\pi\chi_{\Omega_s} B'\left(\psi - \frac{\hat\psi_0}{2}\right)(1-\zeta)\eta\right] dx \\
&\qquad + \int_\Omega \left[\varepsilon_\Gamma \nabla\psi\cdot\nabla(\zeta\eta) + 4\pi\chi_{\Omega_s} B'\left(\psi - \frac{\hat\psi_0}{2}\right)\zeta\eta\right] dx \\
&\quad = \int_{\Omega_m} \varepsilon_\Gamma \nabla\psi\cdot\nabla((1-\zeta)\eta) dx \\
&\qquad + \int_{\Omega_s} \left[\varepsilon_\Gamma \nabla\psi\cdot\nabla((1-\zeta)\eta) + 4\pi\chi_{\Omega_s} B'\left(\psi - \frac{\hat\psi_0}{2}\right)(1-\zeta)\eta\right] dx \\
&\qquad + \int_{V_2} \left[\varepsilon_\Gamma \nabla\psi\cdot\nabla(\zeta\eta) + 4\pi\chi_{\Omega_s} B'\left(\psi - \frac{\hat\psi_0}{2}\right)\zeta\eta\right] dx \\
&\quad = 4\pi\sum_{i=1}^N Q_i\eta(x_i) + \int_{V_2} \left[\varepsilon_\Gamma \nabla\psi\cdot\nabla(\zeta\eta) + 4\pi\chi_{\Omega_s} B'\left(\psi - \frac{\hat\psi_0}{2}\right)\zeta\eta\right] dx. \qquad (4.2)
\end{aligned}
$$

We now show that the second term in (4.2) is zero. Notice that $\psi|_{V_2} \in H^1(V_2)$ and $x_i \notin V_2$ $(1 \le i \le N)$. Denoting $V_m = V_2 \cap \Omega_m$ and $V_s = V_2 \cap \Omega_s$, we have by (2.2) and (2.3) that

$$
\begin{aligned}
&\int_{V_m} \varepsilon_m \nabla\psi\cdot\nabla\xi\, dx = 0 \qquad \forall \xi \in C_c^\infty(V_m), \\
&\int_{V_s} \varepsilon_s \nabla\phi\cdot\nabla\xi\, dx = -4\pi\int_{V_s} B'\left(\psi - \frac{\hat\psi_0}{2}\right)\xi\, dx \qquad \forall \xi \in C_c^\infty(V_s).
\end{aligned}
$$

Consequently, since $\chi_{\Omega_s} B'(\psi - \hat\psi_0/2) \in L^2(\Omega_s)$, we infer from the regularity theory of elliptic boundary-value problems [14] that $\psi|_{V_m} \in H^2(V_m)$ and $\psi|_{V_s} \in H^2(V_s)$, and that

$$
\begin{aligned}
\nabla\cdot\varepsilon_m\nabla\psi &= 0 \qquad \text{a.e. } V_m, \\
\nabla\cdot\varepsilon_s\nabla\psi - 4\pi B'\left(\psi - \frac{\hat\psi_0}{2}\right) &= 0 \qquad \text{a.e. } V_s.
\end{aligned}
$$

Therefore, the trace of $\varepsilon_m \nabla\psi \cdot n$ and that of $\varepsilon_s \nabla\psi \cdot n$ on Γ both exist. Moreover,

$$\begin{aligned}
&\int_{V_2} \left[\varepsilon_\Gamma \nabla\psi \cdot \nabla(\zeta\eta) + 4\pi\chi_{\Omega_s} B'\left(\psi - \frac{\hat{\psi}_0}{2}\right)\zeta\eta \right] dx \\
&= \int_{V_m} \varepsilon_m \nabla\psi \cdot \nabla(\zeta\eta) dx + \int_{V_s} \left[\varepsilon_s \nabla\psi \cdot \nabla(\zeta\eta) + 4\pi\chi_{\Omega_s} B'\left(\psi - \frac{\hat{\psi}_0}{2}\right)\zeta\eta \right] dx \\
&= -\int_{V_m} (\nabla \cdot \varepsilon_m \nabla\psi)\zeta\eta \, dx - \int_\Gamma (\varepsilon_m \nabla\psi \cdot n)\zeta\eta \, dS \\
&\quad + \int_{V_s} \left[-(\nabla \cdot \varepsilon_s \nabla\psi)(\zeta\eta) + 4\pi\chi_{\Omega_s} B'\left(\psi - \frac{\hat{\psi}_0}{2}\right)\zeta\eta \right] dx + \int_\Gamma (\varepsilon_s \nabla\psi \cdot n)\zeta\eta \, dS \\
&= \int_\Gamma [\![\varepsilon \nabla\psi \cdot n]\!] \zeta\eta \, dS \\
&= 0,
\end{aligned} \tag{4.3}$$

where in the last step, we used the third equation of (1.15). Now, since $\eta \in C_c^\infty(\Omega)$ is arbitrary, we obtain (2.1) from (4.2) and (4.3). ∎

5. Minimization of the electrostatic free energy: Proof of Theorems 2.3, 2.4, and 2.5.

Proof of Theorem 2.3. Let $t = 1 + \beta \max_{1\le j\le M} \|\mu_{0j}\|_{L^\infty(\Omega_s)}$, where μ_{0j} $(j = 1, \dots, M)$ are defined in (3.14). It follows from (3.12) and (3.4) that there exists $C_5 > 0$ such that

$$F_0[c] \ge C_5 \left\| \sum_{j=1}^M q_j c_j \right\|^2_{H^{-1}(\Omega)} + \beta^{-1} \sum_{j=1}^M \int_{\Omega_s} S_{-t}(c_j) dx + E_0 \qquad \forall c = (c_1, \dots, c_M) \in V_0, \tag{5.1}$$

where E_0 is defined in (3.15). Let $z = \inf_{c\in V_0} F_0[c]$. Since $S_{-t} : [0,\infty) \to \mathbb{R}$ is bounded below, z is finite.

Let $c^{(k)} = (c_1^{(k)}, \dots, c_M^{(k)}) \in V_0$ $(k = 1, 2, \dots)$ be such that $\lim_{k\to\infty} F_0[c^{(k)}] = z$. It follows from (5.1) that $\{\int_{\Omega_s} S_{-t}(c_j^{(k)})dx\}$ is bounded for each $j = 1, \dots, M$. Therefore, by Lemma 3.3, up to a subsequence that is not relabeled, $\{c_j^{(k)}\}$ converges weakly in $L^1(\Omega_s)$ to some $c_j \in L^1(\Omega_s)$, and

$$\int_{\Omega_s} S_{-t}(c_j)\, dx \le \liminf_{k\to\infty} \int_{\Omega_s} S_{-t}\left(c_j^{(k)}\right) dx < \infty \qquad j = 1, \dots, M. \tag{5.2}$$

Define $c_j = 0$ on Ω_m for all $j = 1, \dots, M$. By (5.1), $\{\sum_{j=1}^M q_j c_j^{(k)}\}$ is bounded in $H^{-1}(\Omega)$. Since $H^{-1}(\Omega)$ is a Hilbert space, $\{\sum_{j=1}^M q_j c_j^{(k)}\}$ has a subsequence, again not relabeled, that weakly converges to some $F \in H^{-1}(\Omega)$. Let $\xi \in L^\infty(\Omega) \cap H_0^1(\Omega)$. We have

$$F(\xi) = \lim_{k\to\infty} \int_\Omega \left(\sum_{j=1}^M q_j c_j^{(k)} \right) \xi dx = \int_\Omega \left(\sum_{j=1}^M q_j c_j \right) \xi dx.$$

Therefore, $\sum_{j=1}^M q_j c_j \in H^{-1}(\Omega)$ and, hence, $c = (c_1, \dots, c_M) \in V_0$.

By (5.2) and the fact that the norm of a Banach space is sequentially weakly lower semicontinuous, we have $z = \liminf_{k\to\infty} F_0[c^{(k)}] \geq F_0[c] \geq z$. This implies that $c \in V_0$ is a global minimizer of $F_0 : V_0 \to \mathbb{R}$.

Let $d = (d_1, \ldots, d_M) \in V_0$ be a local minimizer of $F_0 : V_0 \to \mathbb{R}$. Then for $\lambda \in (0, 1)$ close to 0, we have by the convexity of $F_0 : V_0 \to \mathbb{R}$ that

$$F_0[d] \leq F_0[\lambda c + (1-\lambda)d] \leq \lambda F_0[c] + (1-\lambda)F_0[d],$$

leading to $F_0[d] \leq F_0[c]$. Thus, d is also a global minimizer of $F_0 : V_0 \to \mathbb{R}$. Clearly, $(c+d)/2 \in V_0$. Consequently, it follows from the definition of the norm $|||\cdot|||$ and the Cauchy–Schwarz inequality with respect to the inner product $\langle \xi, \eta \rangle = \xi(L\eta)$ $(\xi, \eta \in H^{-1}(\Omega))$ that

$$\begin{aligned}
0 &\leq F_0\left[\frac{c+d}{2}\right] - \min_{e\in V_0} F_0[e] \\
&= F_0\left[\frac{c+d}{2}\right] - \frac{1}{2}F_0[c] - \frac{1}{2}F_0[d] \\
&= \frac{1}{8}\left|\left|\left|\sum_{j=1}^{M} q_j(c_j + d_j)\right|\right|\right|^2 - \frac{1}{4}\left|\left|\left|\sum_{j=1}^{M} q_j c_j\right|\right|\right|^2 - \frac{1}{4}\left|\left|\left|\sum_{j=1}^{M} q_j d_j\right|\right|\right|^2 \\
&\quad + \beta^{-1}\sum_{j=1}^{M}\int_{\Omega_s}\left[S_{-1}\left(\frac{c_j+d_j}{2}\right) - \frac{1}{2}S_{-1}(c_j) - \frac{1}{2}S_{-1}(d_j)\right]dx \\
&\leq \beta^{-1}\sum_{j=1}^{M}\int_{\Omega_s}\left[S_{-1}\left(\frac{c_j+d_j}{2}\right) - \frac{1}{2}S_{-1}(c_j) - \frac{1}{2}S_{-1}(d_j)\right]dx.
\end{aligned}$$

This, together with the convexity of S_{-1} on $[0, \infty)$, implies that

$$S_{-1}\left(\frac{c_j(x)+d_j(x)}{2}\right) = \frac{1}{2}S_{-1}(c_j(x)) + \frac{1}{2}S_{-1}(d_j(x)) \qquad \forall j = 1, \ldots, M,\ \forall x \in \Omega_s \backslash \omega_s,$$

for some $\omega_s \subset \Omega_s$ with $|\omega_s| = 0$. Let $x \in \Omega_s \setminus \omega_s$. Then it follows from the definition of $S_{-1} : [0, \infty) \to \mathbb{R}$ that $c_j(x) = 0$ if and only if $d_j(x) = 0$ for all $j = 1, \ldots, M$. The strict convexity of S_0 on $(0, \infty)$ then implies that $c = d$ a.e. Ω_s. Hence, $c = d$ in V_0. □

Proof of Theorem 2.4. (1) Let $c = (c_1, \ldots, c_M) \in W_0$. We show that the following four statements are equivalent:

(i) c is an equilibrium of $F_0 : W_0 \to \mathbb{R}$;

(ii) The property (2.5) holds true, and

$$q_j L\left(\sum_{j=1}^{M} q_j c_j\right) + \mu_{0j} + \beta^{-1}\log c_j = 0 \qquad \text{a.e. } \Omega_s, \quad j = 1, \ldots, M; \tag{5.3}$$

(iii) c is a global minimizer of $F_0 : W_0 \to \mathbb{R}$;

(iv) c is a local minimizer of $F_0 : W_0 \to \mathbb{R}$.

Assume (i) is true. Then (2.5) holds true by Definition 2.3. Let $e = (e_1, \ldots, e_M) \in X \cap L^\infty(\Omega, \mathbb{R}^M)$. Notice that $S'_{-1}(u) = \log u$ for any $u > 0$. Thus, for each $j \in \{1, \ldots, M\}$ and each $x \in \Omega_s$, the mean-value theorem implies the existence of $\theta_j(x) \in [0, 1]$ such that

$$S_{-1}(c_j(x) + te_j(x)) - S_{-1}(c_j(x)) = te_j(x)\log(c_j(x) + t\theta_j(x)e_j(x)).$$

Hence, by the Lebesgue dominated convergence theorem,

$$\lim_{t\to 0}\int_{\Omega_s}\frac{S_{-1}(c_j+te_j)-S_{-1}(c_j)}{t}dx=\int_{\Omega_s}e_j\log c_j\,dx,\qquad j=1,\dots,M.$$

Therefore, it follows from Definition 2.3, the definition of the norm $\|\!|\cdot|\!\|$, (3.12), and (3.4) that

$$\begin{aligned}
0&=\delta F_0[c]e\\
&=\lim_{t\to 0}\frac{F_0[c+te]-F_0[c]}{t}\\
&=\lim_{t\to 0}\left\{\frac{1}{2}t\left|\!\left|\!\left|\sum_{j=1}^M q_je_j\right|\!\right|\!\right|^2+\int_{\Omega_s}\left(\sum_{j=1}^M q_je_j\right)L\left(\sum_{j=1}^M q_jc_j\right)dx\right.\\
&\qquad\left.+\sum_{j=1}^M\int_{\Omega_s}\mu_{0j}e_j dx+\beta^{-1}\sum_{j=1}^M\int_{\Omega_s}\frac{1}{t}\left[S_{-1}(c_j+te_j)-S_{-1}(c_j)\right]dx\right\}\\
&=\sum_{j=1}^M\int_{\Omega_s}\left[q_jL\left(\sum_{j=1}^M q_jc_j\right)+\mu_{0j}+\beta^{-1}\log c_j\right]e_j dx\qquad\forall e\in X\cap L^\infty\left(\Omega,\mathbb{R}^M\right).
\end{aligned}\tag{5.4}$$

This implies (5.3). Hence, (ii) is true.

Assume (ii) is true. We show that (iii) is true. By Lemma 3.4, we need only to show that $F_0[c]\le F_0[d]$ for any fixed $d=(d_1,\dots,d_M)\in W_0$ that satisfies (2.5) with c replaced by d. In fact, setting $e=(e_1,\dots,e_M)=d-c\in X\cap L^\infty(\Omega,\mathbb{R}^M)$, we have by the convexity of $S_{-1}:[0,\infty)\to\mathbb{R}$ that

$$S_{-1}(d_j)-S_{-1}(c_j)\ge(d_j-c_j)S'_{-1}(c_j)=e_j\log c_j\qquad\text{a.e. }\Omega_s.$$

Therefore, it follows from (3.12) and (5.3) that

$$\begin{aligned}
F_0[d]-F_0[c]&=\frac{1}{2}\int_{\Omega_s}\left(\sum_{j=1}^M q_je_j\right)L\left(\sum_{j=1}^M q_je_j\right)dx+\int_{\Omega_s}\left(\sum_{j=1}^M q_je_j\right)L\left(\sum_{j=1}^M q_jc_j\right)dx\\
&\quad+\sum_{j=1}^M\int_{\Omega_s}\mu_{0j}e_j dx+\beta^{-1}\sum_{j=1}^M\int_{\Omega_s}\left[S_{-1}(d_j)-S_{-1}(c_j)\right]dx\\
&\ge\sum_{j=1}^M\int_{\Omega_s}\left[q_jL\left(\sum_{i=1}^M q_ic_i\right)+\mu_{0j}+\beta^{-1}\log c_j\right]e_j dx\\
&=0.
\end{aligned}$$

Hence, $F_0[c]\le F_0[d]$, and (iii) is true.

Clearly, (iii) implies (iv).

Finally, assume (iv) is true. By Lemma 3.4, (2.5) holds true. For any $e\in X\cap L^\infty(\Omega,\mathbb{R}^M)$, it is easy to see that $\delta F_0[c]e$ exists, cf. (5.4). Since $F_0[c+te]\ge F_0[c]$ for $|t|$ small enough, we have $\delta F_0[c]e=0$. Therefore, c is an equilibrium of $F_0:W_0\to\mathbb{R}$, and (i) is true.

Let now $\psi \in H^1_*(\Omega)$ be the unique weak solution to the boundary-value problem of the PBE (1.11) and (1.6), cf. Theorem 2.1. Define $c = (c_1, \dots, c_M) : \Omega \to \mathbb{R}$ by (1.9) for point ions and $c_j(x) = 0$ for all $x \in \Omega_m$ and all $j = 1, \dots, M$. Clearly, $c \in W_0$. Moreover, by Theorem 2.1, $\psi|_{\Omega_s} \in C(\overline{\Omega_s})$. This implies (2.5). It follows from (1.11), (1.6), (1.9) for point ions, and Lemma 3.2 with $f = \sum_{j=1}^M q_j c_j$ that ψ is the electrostatic potential corresponding to c; i.e., $\psi = G + \hat{\psi}_0 + L(\sum_{j=1}^M q_j c_j)$. This, together with the Boltzmann relations (1.9) for point ions and (3.14), implies (5.3). Hence, c is an equilibrium, and, thus, a local and global minimizer, of $F_0 : W_0 \to \mathbb{R}$. The uniqueness of equilibria or local minimizers is equivalent to that of global minimizers, and can be proved by the same argument used in the proof of Theorem 2.3.

(2) It is clear that from our definition of c and ψ that we need only to prove (2.7). Since c is the unique minimizer of $F_0 : W_0 \to \mathbb{R}$ and ψ is the corresponding electrostatic potential determined by (3.11), we have by (1.1) and (1.9) for point ions that

$$\begin{aligned}
\min_{d \in W_0} F_0[d] &= F_0[c] \\
&= \frac{1}{2}\sum_{i=1}^N Q_i(\psi - \psi_{vac})(x_i) + \frac{1}{2}\int_{\Omega_s}\left(\sum_{j=1}^M q_j c_j\right)\psi dx \\
&\quad + \beta^{-1}\sum_{j=1}^M \int_{\Omega_s} c_j\left[\log\left(a^3 c_j\right) - 1\right]dx - \sum_{j=1}^M \int_{\Omega_s} \mu_j c_j dx \\
&= \frac{1}{2}\sum_{i=1}^N Q_i(\psi - \psi_{vac})(x_i) - \frac{1}{2}\sum_{j=1}^M \int_{\Omega_s} q_j c_j^\infty e^{-\beta q_j\left(\psi - \hat{\psi}_0/2\right)}\left(\psi - \hat{\psi}_0\right)dx \\
&\quad - \beta^{-1}\sum_{j=1}^M \int_{\Omega_s} c_j^\infty e^{-\beta q_j\left(\psi - \hat{\psi}_0/2\right)} dx. \qquad (5.5)
\end{aligned}$$

Since ψ is the unique solution to the boundary-value problem of PBE (1.11) and (1.6), and since $\hat{\psi}_0$ is harmonic in Ω_s by (1.10), we have

$$\varepsilon_s \Delta\left(\psi - \hat{\psi}_0\right) + 4\pi\sum_{j=1}^M q_j c_j^\infty e^{-\beta q_j\left(\psi - \hat{\psi}_0/2\right)} = 0 \qquad \text{a.e. } \Omega_s.$$

Multiplying both sides of this equation by $\psi - \hat{\psi}_0$ and integrate the resulting terms over Ω_s, we obtain by integration by parts and the fact that by Lemma 3.1 both $\psi - \hat{\psi}_0$ and $\varepsilon_\Gamma \partial_n(\psi - \hat{\psi}_0)$ are continuous across Γ,

$$\begin{aligned}
&-\int_{\Omega_s} \varepsilon_s \left|\nabla\left(\psi - \hat{\psi}_0\right)\right|^2 dx + \int_\Gamma \varepsilon_\Gamma\left(\psi - \hat{\psi}_0\right)\partial_n\left(\psi - \hat{\psi}_0\right)dS \\
&\quad + 4\pi\sum_{j=1}^M \int_{\Omega_s} q_j c_j^\infty e^{-\beta q_j\left(\psi - \hat{\psi}_0/2\right)}\left(\psi - \hat{\psi}_0\right)dx = 0.
\end{aligned}$$

This and (5.5) imply (2.7). ∎

Proof of Theorem 2.5. (1) We first show that $F_a : V_a \to \mathbb{R}$ is a convex functional. Define $T_M = \{(u_1, \dots, u_M) \in \mathbb{R}^M : u_j > 0 \text{ for } j = 1, \dots, M, \text{ and } \sum_{j=1}^M u_j < 1\}$ and

$$h(u) = \left(1 - \sum_{j=1}^M u_j\right)\left[\log\left(1 - \sum_{j=1}^M u_j\right) - 1\right] \qquad \forall u = (u_1, \dots, u_M) \in T_M.$$

Clearly, T_M is convex. We have $\partial_{u_i u_j} h(u) = (1 - \sum_{k=1}^M u_k)^{-1}$ for all $1 \le i, j \le M$. Let $H(u) = (\partial_{u_i u_j} h)$ be the Hessian of $h : T_M \to \mathbb{R}$. Then, for any $y = (y_1, \dots, y_M) \in R^M$, we have $y \cdot H(u)y = (\sum_{k=1}^M y_k)^2/(1 - \sum_{k=1} u_k) \ge 0$. Therefore, $H(u)$ is symmetric, semidefinite for any $u \in T_M$. Hence, $h : T_M \to \mathbb{R}$ is convex. Consequently, since V_a is a convex subset of X and $S_{-1} : [0, \infty) \to \mathbb{R}$ is convex, we conclude by (3.13) that $F_a : V_a \to \mathbb{R}$ is convex.

Let now $c = (c_1, \dots, c_M) \in V_a$. By the same argument used in the proof of Theorem 2.4, we obtain the equivalence of the following four statements:

(i) c is an equilibrium of $F_a : V_a \to \mathbb{R}$;

(ii) The property (2.6) holds true, and

$$q_j L\left(\sum_{i=1}^M q_i c_i\right) + \mu_{aj} + \beta^{-1} \log\left(\frac{a^3 c_j}{1 - a^3 \sum_{i=1}^M c_i}\right) = 0 \quad \text{a.e. } \Omega_s, \quad j = 1, \dots, M; \tag{5.6}$$

(iii) c is a global minimizer of $F_a : V_a \to \mathbb{R}$;

(iv) c is a local minimizer of $F_a : V_a \to \mathbb{R}$.

Let $\psi \in H^1_*(\Omega)$ be the unique weak solution to the boundary-value problem of the PBE (1.12) and (1.6), cf. Theorem 2.1. Define $c = (c_1, \dots, c_M) : \Omega \to \mathbb{R}$ by (1.9) for finite-size ions and $c_j(x) = 0$ for all $x \in \Omega_m$ and all $j = 1, \dots, M$. Clearly, $c \in V_a$. Moreover, by Theorem 2.1, $\psi|_{\Omega_s} \in C(\overline{\Omega_s})$. This implies (2.6). By (1.9) for finite-size ions, we have

$$a^3 \sum_{j=1}^M c_j(x) = 1 - \frac{1}{1 + a^3 \sum_{i=1}^M c_i^\infty e^{-\beta q_i\left(\psi - \hat{\psi}_0/2\right)}}.$$

This together with (1.9) for finite-size ions imply that

$$\frac{c_j}{1 - a^3 \sum_{i=1}^M c_i} = c_j^\infty e^{-\beta q_j\left(\psi - \hat{\psi}_0/2\right)}, \qquad j = 1, \dots, M. \tag{5.7}$$

It follows from (1.12), (1.6), (1.9) for finite-size ions, and Lemma 3.2 with $f = \sum_{j=1}^M q_j c_j$ that ψ is the electrostatic potential corresponding to c, i.e., $\psi = G + \hat{\psi}_0 + L(\sum_{j=1}^M q_j c_j)$. This, together with (5.7) and (3.16), implies (5.6). Hence, c is an equilibrium, and, thus, a local and global minimizer, of $F_a : V_a \to \mathbb{R}$. The uniqueness of equilibria or local minimizers is equivalent to that of global minimizers, and can be proved by the same argument used in the proof of Theorem 2.3.

(2) We need only to prove (2.8). Since c is the unique minimizer of $F_a : V_a \to \mathbb{R}$ and ψ is the corresponding electrostatic potential determined by (3.11), we have by

(1.2), (1.3), and (1.9) for finite-size ions that

$$\begin{aligned}
&\min_{d\in V_a} F_a[d] = F_a[c] \\
&\quad = \frac{1}{2}\sum_{i=1}^{N} Q_i(\psi-\psi_{vac})(x_i) + \frac{1}{2}\int_{\Omega_s}\left(\sum_{j=1}^{M} q_j c_j\right)\psi dx \\
&\qquad + \beta^{-1}\sum_{j=0}^{M}\int_{\Omega_s} c_j\left[\log\left(a^3 c_j\right)-1\right]dx - \sum_{j=1}^{M}\int_{\Omega_s}\mu_j c_j dx \\
&\quad = \frac{1}{2}\sum_{i=1}^{N} Q_i(\psi-\psi_{vac})(x_i) - \frac{1}{2}\sum_{j=1}^{M}\int_{\Omega_s}\frac{q_j c_j^{\infty} e^{-\beta q_j\left(\psi-\hat{\psi}_0/2\right)}\left(\psi-\hat{\psi}_0\right)}{1+a^3\sum_{i=1}^{M} c_i^{\infty} e^{-\beta q_i\left(\psi-\hat{\psi}_0/2\right)}}dx \\
&\qquad - \beta^{-1}a^{-3}\int_{\Omega_s}\left[1+\log\left(1+a^3\sum_{i=1}^{M} c_i^{\infty} e^{-\beta q_i\left(\psi-\hat{\psi}_0/2\right)}\right)\right]dx. \qquad (5.8)
\end{aligned}$$

Since ψ is the unique solution to the boundary-value problem of PBE (1.12) and (1.6), and since $\hat{\psi}_0$ is harmonic in Ω_s by (1.10), we have

$$\varepsilon_s\Delta\left(\psi-\hat{\psi}_0\right) + 4\pi\sum_{j=1}^{M}\frac{q_j c_j^{\infty} e^{-\beta q_j\left(\psi-\hat{\psi}_0/2\right)}}{1+a^3\sum_{i=1}^{M} c_i^{\infty} e^{-\beta q_i\left(\psi-\hat{\psi}_0/2\right)}} = 0 \qquad \text{a.e. } \Omega_s.$$

Multiplying both sides of this equation by $\psi-\hat{\psi}_0$ and integrating the resulting terms over Ω_s, we obtain by integration by parts and the fact that by Lemma 3.1 both $\psi-\hat{\psi}_0$ and $\varepsilon_\Gamma\partial_n(\psi-\hat{\psi}_0)$ are continuous across Γ,

$$\begin{aligned}
&-\int_{\Omega_s}\varepsilon_s\left|\nabla\left(\psi-\hat{\psi}_0\right)\right|^2 dx + \int_{\Gamma}\varepsilon_\Gamma\left(\psi-\hat{\psi}_0\right)\partial_n\left(\psi-\hat{\psi}_0\right)dS \\
&\qquad + 4\pi\sum_{j=1}^{M}\int_{\Omega_s}\frac{q_j c_j^{\infty} e^{-\beta q_j\left(\psi-\hat{\psi}_0/2\right)}\left(\psi-\hat{\psi}_0\right)}{1+a^3\sum_{i=1}^{M} c_i^{\infty} e^{-\beta q_i\left(\psi-\hat{\psi}_0/2\right)}}dx = 0.
\end{aligned}$$

This and (5.8) imply (2.8). ∎

Appendix A.

We now prove Lemma 3.4 and Lemma 3.5 by constructing ionic concentrations that satisfy required conditions and that have lower free energies. The key idea here is based on the following observation: the function $S_\alpha : [0,\infty)\to\mathbb{R}$, defined for any $\alpha\in\mathbb{R}$ by $S_\alpha(0)=0$ and $S_\alpha(u)=u(\alpha+\log u)$ if $u>0$, has a unique minimizer which is a positive number. Moreover, the magnitude $|S'_\alpha(u)|$ is very large if u is close to 0 or ∞. Notice that $-S_\alpha$ represents the entropy of the system. Therefore, small changes of concentrations near zero or infinity can largely increase the corresponding entropy and, hence, decrease the free energy.

Proof of Lemma 3.4. We first construct $\bar{c}\in W_0$ that satisfies

$$\bar{c}_j(x)\le\gamma_2' \qquad \text{a.e. } x\in\Omega_s,\ j=1,\dots,M, \qquad (A.1)$$

for some constant $\gamma_2'>0$, $\|\bar{c}-c\|_X<\varepsilon/2$, and $F_0[\bar{c}]\le F_0[c]$ with a strict inequality if $c\notin L^\infty(\Omega,\mathbb{R}^M)$. Let $A>0$. Define $\bar{c}=(\bar{c}_1,\dots,\bar{c}_M):\Omega\to\mathbb{R}$ by

$$\bar{c}_j(x) = \begin{cases} c_j(x) & \text{if } c_j(x)\le A \\ 0 & \text{if } c_j(x)>A \end{cases} \qquad \forall x\in\Omega,\ j=1,\dots,M. \qquad (A.2)$$

Clearly, $\bar{c} \in W_0$ and (A.1) holds true with $\gamma_2' = A$. Moreover, $\sum_{j=1}^M \|\bar{c}_j - c_j\|_{L^1(\Omega)} < \varepsilon/4$ for $A > 0$ large enough.

Denote

$$\tau_j(A) = \{x \in \Omega_s : c_j(x) > A\}, \qquad j = 1, \dots, M.$$

Since $c \in W_0$, there exists $p > 3/2$ such that each $c_j \in L^p(\Omega)$ $(1 \le j \le M)$. Thus,

$$\sum_{j=1}^M q_j \bar{c}_j - \sum_{j=1}^M q_j c_j = -\sum_{j=1}^M q_j \chi_{\tau_j(A)} c_j \to 0 \qquad \text{in } L^p(\Omega) \quad \text{as } A \to \infty.$$

By the definition of $L : H^{-1}(\Omega) \to H_0^1(\Omega)$ and the regularity theory for elliptic problems [14], we have $L(\sum_{j=1}^M q_j \chi_{\tau_j(A)} c_j)|_{\Omega_s} \in W^{2,p}(\Omega_s)$ and

$$\left\| L\left(\sum_{j=1}^M q_j \chi_{\tau_j(A)} c_j \right) \right\|_{W^{2,p}(\Omega_s)} \le C \left\| \sum_{j=1}^M q_j \chi_{\tau_j(A)} c_j \right\|_{L^p(\Omega_s)} \to 0 \qquad \text{as } A \to \infty.$$

Hence, by (3.4) and the embedding $W^{2,p}(\Omega_s) \hookrightarrow L^\infty(\Omega_s)$ that

$$\begin{aligned}
\left\| \sum_{j=1}^M q_j \bar{c}_j - \sum_{j=1}^M q_j c_j \right\|_{H^{-1}(\Omega)}^2 &\le C \int_{\Omega_s} \left(\sum_{j=1}^M q_j \chi_{\tau_j(A)} c_j \right) L\left(\sum_{j=1}^M q_j \chi_{\tau_j(A)} c_j \right) dx \\
&\le C \left\| \sum_{j=1}^M q_j \chi_{\tau_j(A)} c_j \right\|_{L^1(\Omega_s)} \left\| L\left(\sum_{j=1}^M q_j \chi_{\tau_j(A)} c_j \right) \right\|_{L^\infty(\Omega_s)} \\
&\le C \left\| \sum_{j=1}^M q_j \chi_{\tau_j(A)} c_j \right\|_{L^p(\Omega_s)} \left\| L\left(\sum_{j=1}^M q_j \chi_{\tau_j(A)} c_j \right) \right\|_{W^{2,p}(\Omega_s)} \\
&\le C \left\| \sum_{j=1}^M q_j \chi_{\tau_j(A)} c_j \right\|_{L^p(\Omega_s)}^2 \\
&\to 0 \qquad \text{as } A \to \infty.
\end{aligned}$$

Therefore, $\|\bar{c} - c\|_X < \varepsilon$ if $A > 0$ is large enough.

Notice that $\bar{c}_j = c_j - \chi_{\tau_j(A)} c_j$ for all $j = 1, \dots, M$. Thus,

$$\begin{aligned}
&\frac{1}{2} \int_{\Omega_s} \left(\sum_{j=1}^M q_j \bar{c}_j \right) L\left(\sum_{j=1}^M q_j \bar{c}_j \right) dx - \frac{1}{2} \int_{\Omega_s} \left(\sum_{j=1}^M q_j c_j \right) L\left(\sum_{j=1}^M q_j c_j \right) dx \\
&\quad = -\frac{1}{2} \int_{\Omega_s} \left(\sum_{j=1}^M q_j \chi_{\tau_j(A)} c_j \right) L\left(\sum_{j=1}^M q_j c_j + \sum_{j=1}^M q_j \bar{c}_j \right) dx \\
&\quad \le \frac{1}{2} \sum_{j=1}^M |q_j| d_j(A) \int_{\tau_j(A)} c_j dx, \qquad \text{(A.3)}
\end{aligned}$$

where

$$d_j(A) = \left\| L\left(\sum_{j=1}^M q_j c_j\right)\right\|_{L^\infty(\Omega_s)} + \left\| L\left(\sum_{j=1}^M q_j \bar{c}_j\right)\right\|_{L^\infty(\Omega_s)}.$$

Since

$$\begin{aligned}\left\| L\left(\sum_{j=1}^M q_j \bar{c}_j\right)\right\|_{L^\infty(\Omega_s)} &\le C\left\| L\left(\sum_{j=1}^M q_j \bar{c}_j\right)\right\|_{W^{2,p}(\Omega_s)}\\ &\le C\left\|\sum_{j=1}^M q_j \bar{c}_j\right\|_{L^p(\Omega_s)} \to \left\|\sum_{j=1}^M q_j c_j\right\|_{L^p(\Omega_s)}\end{aligned}$$

as $A \to \infty$, we have $\max_{1\le j\le M} d_j(A) \le C$ as $A > 0$ large enough. For each fixed $j \in \{1,\dots,M\}$ and $x \in \tau_j(A)$, we also have

$$S_{-1}(\bar{c}_j(x)) - S_{-1}(c_j(x)) = -S_{-1}(c_j(x)) = -c_j(x)\log c_j(x) \le -c_j(x)\log A. \tag{A.4}$$

Therefore, it follows from (3.12), (A.3), and (A.4) that

$$F_0[\bar{c}] - F_0[c] \le \sum_{j=1}^M \left(\frac{1}{2}|q_j| d_j(A) + \|\mu_{0j}\|_{L^\infty(\Omega_s)} - \beta^{-1}\log A\right)\int_{\tau_j(A)} c_j dx.$$

If $A > 0$ is large enough, this is nonpositive. If $c \notin L^\infty(\Omega,\mathbb{R}^M)$, then there exists $j \in \{1,\dots,M\}$ such that $|\tau_j(A)| > 0$ for all $A > 0$. In this case, we have the strict inequality $F_0[\bar{c}] < F_0[c]$.

We now construct $\hat{c} \in W_0$ that satisfies (2.5) with c replaced by $\hat{c}$, $\|\hat{c}-\bar{c}\|_X < \varepsilon/2$, and $F_0[\hat{c}] \le F_0[\bar{c}]$ with a strict inequality if there exists $j \in \{1,\dots,M\}$ such that $|\{x \in \Omega_s : c_j(x) < \alpha\}| > 0$ for all $\alpha > 0$, all these implying that $\hat{c}$ satisfies all the desired properties. If there exists $\gamma_1' > 0$ such that $c_j(x) \ge \gamma_1'$ for a.e. $x \in \Omega_s$ and $j = 1,\dots,M$, then $\hat{c} = \bar{c}$ with $A \ge \gamma_1'$ (cf. (A.2)) satisfies all the desired properties with $\gamma_1 = \gamma_1'$ and $\gamma_2 = \gamma_2'$. Assume otherwise there exists $j_0 \in \{1,\dots,M\}$ such that $|\{x \in \Omega_s : c_{j_0}(x) < \alpha\}| > 0$ for all $\alpha > 0$. This means that $|\{x \in \Omega_s : \bar{c}_{j_0}(x) < \alpha\}| > 0$ for all $\alpha > 0$.

Define

$$\begin{aligned}&\rho_j(\alpha) = \{x \in \Omega_s : \bar{c}_j(x) < \alpha\} \qquad \forall \alpha > 0, \quad j = 1,\dots,M,\\ &I_0 = \{j \in \{1,\dots,M\} : |\rho_j(\alpha)| > 0\ \forall \alpha > 0\},\\ &I_1 = \{1,\dots,M\} \setminus I_0.\end{aligned}$$

Clearly, $I_0 \ne \emptyset$. If $I_1 \ne \emptyset$, then there exists $\alpha_1 > 0$ such that

$$\bar{c}_j(x) \ge \alpha_1 \qquad \text{a.e. } x \in \Omega_s, \quad \forall j \in I_1.$$

Define for $0 < \alpha < \alpha_1$ and $1 \le j \le M$

$$\hat{c}_j(x) = \begin{cases} \bar{c}_j(x) + \alpha\chi_{\rho_j(\alpha)}(x) & \text{if } j \in I_0\\ \bar{c}_j(x) & \text{if } j \in I_1\end{cases} \qquad \forall x \in \Omega.$$

Clearly, $\hat{c} = (\hat{c}_1,\dots,\hat{c}_M) \in W_0$ and (2.5) holds true with c replaced by $\hat{c}$, $\gamma_1 = \alpha$, and $\gamma_2 = \gamma_2' + \alpha$. Moreover, $\sum_{j=1}^M \|\hat{c}_j - \bar{c}_j\|_{L^1(\Omega)} < \varepsilon/4$ if $\alpha > 0$ is small enough.

Furthermore,

$$\left\|\sum_{j=1}^{M} q_j\hat{c}_j - \sum_{j=1}^{M} q_j\bar{c}_j\right\|_{H^{-1}(\Omega)} = \alpha\left\|\sum_{j\in I_0} q_j\chi_{\rho_j(\alpha)}\right\|_{H^{-1}(\Omega)} \le \alpha\left\|\sum_{j\in I_0} q_j\chi_{\rho_j(\alpha)}\right\|_{L^2(\Omega)}$$

$$\le \alpha\sum_{j\in I_0}|q_j|\sqrt{|\rho_j(\alpha)|} \to 0 \qquad \text{as } \alpha\to 0. \tag{A.5}$$

Hence, $\|\hat{c}-\bar{c}\|_X < \varepsilon/2$ if $\alpha>0$ is small enough.

By the mean-value theorem and the fact that $S'_{-1}(u)=\log u$ for any $u>0$,

$$\sum_{j=1}^{M}\int_{\Omega_s}[S_{-1}(\hat{c}_j)-S_{-1}(\bar{c}_j)]\,dx = \sum_{j\in I_0}\int_{\rho_j(\alpha)}[S_{-1}(\hat{c}_j)-S_{-1}(\bar{c}_j)]\,dx$$
$$\le \alpha\log(2\alpha)\sum_{j\in I_0}|\rho_j(\alpha)|.$$

Consequently, it follows from (3.13), (3.4), (A.5) that

$$\begin{aligned}
F_0[\hat{c}]-F_0[\bar{c}] &= \frac{1}{2}\int_{\Omega_s}\left(\sum_{j=1}^{M} q_j\bar{c}_j+\alpha\sum_{j\in I_0} q_j\chi_{\rho_j(\alpha)}\right)L\left(\sum_{j=1}^{M} q_j\bar{c}_j+\alpha\sum_{j\in I_0} q_j\chi_{\rho_j(\alpha)}\right)dx\\
&\quad-\frac{1}{2}\int_{\Omega_s}\left(\sum_{j=1}^{M} q_j\bar{c}_j\right)L\left(\sum_{j=1}^{M} q_j\bar{c}_j\right)dx+\alpha\sum_{j\in I_0}\int_{\rho_j(\alpha)}\mu_{0j}dx\\
&\quad+\beta^{-1}\sum_{j\in I_0}\int_{\rho_j(\alpha)}[S_{-1}(\hat{c}_j)-S_{-1}(\bar{c}_j)]\,dx\\
&\le \frac{\alpha^2}{2}\int_{\Omega_s}\left(\sum_{j\in I_0} q_j\chi_{\rho_j(\alpha)}\right)L\left(\sum_{j\in I_0} q_j\chi_{\rho_j(\alpha)}\right)dx\\
&\quad+\alpha\int_{\Omega_s}\left(\sum_{j\in I_0} q_j\chi_{\rho_j(\alpha)}\right)L\left(\sum_{j=1}^{M} q_j\bar{c}_j\right)dx\\
&\quad+\alpha\sum_{j\in I_0}\|\mu_{0j}\|_{L^\infty(\Omega_s)}|\rho_j(\alpha)|+\alpha\sum_{j\in I_0}\beta^{-1}\log(2\alpha)|\rho_j(\alpha)|\\
&\le \frac{\alpha^2}{2}\left(\sum_{i\in I_0} q_i^2\right)\sum_{j\in I_0}|\rho_j(\alpha)|+\alpha\left\|L\left(\sum_{j=1}^{M} q_j\bar{c}_j\right)\right\|_{L^\infty(\Omega_s)}\sum_{j\in I_0}|q_j||\rho_j(\alpha)|\\
&\quad+\alpha\sum_{j\in I_0}\|\mu_{0j}\|_{L^\infty(\Omega_s)}|\rho_j(\alpha)|+\alpha\sum_{j\in I_0}\beta^{-1}\log(2\alpha)|\rho_j(\alpha)|\\
&= \alpha\sum_{j\in J_0}\left[\frac{\alpha}{2}\left(\sum_{j\in I_0} q_j^2\right)+|q_j|\left\|L\left(\sum_{j=1}^{M} q_j\bar{c}_j\right)\right\|_{L^\infty(\Omega_s)}\right.\\
&\qquad\left.+\|\mu_{0j}\|_{L^\infty(\Omega_s)}+\beta^{-1}\log(2\alpha)\right|\rho_j(\alpha)|.
\end{aligned}$$

Since $I_0\neq\emptyset$, this is strictly negative if $\alpha>0$ is small enough.

Proof of Lemma 3.5. We first construct $\bar{c} = (\bar{c}_1, \dots, \bar{c}_M) \in V_a$ such that

$$a^3 \bar{c}_0(x) = 1 - a^3 \sum_{j=1}^{M} \bar{c}_j(x) \geq \theta_1 \qquad \text{a.e. } x \in \Omega_s \tag{A.6}$$

for some constant $\theta_1 \in (0,1)$, $\|\bar{c} - c\|_X < \varepsilon/2$, and $F_a[\bar{c}] \leq F_a[c]$ with a strict inequality if $|\{x \in \Omega_s : a^3 c_0(x) < \alpha\}| > 0$ for all $\alpha > 0$.

Denote for any $\alpha > 0$

$$\omega_0(\alpha) = \left\{ x \in \Omega_s : a^3 c_0(x) < \alpha \right\}.$$

If there exists a constant $\alpha_1 > 0$ such that $|\omega_0(\alpha_1)| = 0$, i.e., $a^3 c_0(x) \geq \alpha_1$ a.e. Ω_s, then $(\bar{c}_1, \dots, \bar{c}_M) = (c_1, \dots, c_M) \in V_a$ satisfies all the desired properties with $\theta_1 = \alpha_1/(1+\alpha_1) \in (0,1)$. Suppose $|\omega_0(\alpha)| > 0$ for any $\alpha > 0$. Let $0 < \alpha < 1/(4M)$. Let $x \in \omega_0(\alpha)$. Then there exists some $j = j(x) \in \{1, \dots, M\}$ such that $a^3 c_j(x) \geq 1/(2M)$. In fact, if this were not true, then $a^3 c_i(x) < 1/(2M)$ for all $i = 1, \dots, M$. Hence, $a^3 c_0(x) = 1 - a^3 \sum_{i=1}^{M} c_i(x) > 1/2 > \alpha$. This would mean that $x \notin \omega_0(\alpha)$, a contradiction. Denoting

$$H_j(\alpha) = \left\{ x \in \omega_0(\alpha) : a^3 c_j(x) \geq \frac{1}{2M} \right\}, \qquad j = 1, \dots, M,$$

we, thus, have $\omega_0(\alpha) = \cup_{j=1}^{M} H_j(\alpha)$. Since $|\omega_0(\alpha)| > 0$, we have $|H_{j_1}(\alpha)| > 0$ for some j_1 $(1 \leq j_1 \leq M)$. If $|H_j(\alpha) \setminus H_{j_1}(\alpha)| = 0$ for all $j \neq j_1$, then we have $\omega_0(\alpha) = \tilde{K}_1(\alpha) \cup H_{j_1}(\alpha)$ for some $\tilde{K}_1(\alpha) \subset \omega_0(\alpha)$ with $|\tilde{K}_1(\alpha)| = 0$. Otherwise, $|H_{j_2}(\alpha) \setminus H_{j_1}(\alpha)| > 0$ for some $j_2 \neq j_1$. In case $|\omega_0(\alpha) \setminus [H_{j_1}(\alpha) \cup H_{j_2}(\alpha)]| = 0$, we have $\omega_0(\alpha) = \tilde{K}_2(\alpha) \cup H_{j_1}(\alpha) \cup [H_{j_2}(\alpha) \setminus H_{j_1}(\alpha)]$ for some $\tilde{K}_2(\alpha) \subset \omega_0(\alpha)$ with $|\tilde{K}_2(\alpha)| = 0$. By induction, we see that there exist $m \in \{1, \dots, M\}$, $\tilde{K}_m(\alpha) \subset \omega_0(\alpha)$ with $|\tilde{K}_m(\alpha)| = 0$, and mutually disjoint sets $K_{j_1}(\alpha), \dots, K_{j_m}(\alpha) \subseteq \omega_0(\alpha)$ such that $K_{j_i}(\alpha) \subseteq H_{j_i}(\alpha)$ and $|K_{j_i}(\alpha)| > 0$ for $i = 1, \dots, m$, and $\omega_0(\alpha) = \tilde{K}_m(\alpha) \cup [\cup_{i=1}^{m} K_{j_i}(\alpha)]$. By relabeling, we may assume that $j_i = i$ for $i = 1, \dots, m$.

Define now

$$\bar{c}_j(x) = \begin{cases} c_j(x) - \alpha a^{-3} \chi_{K_j(\alpha)}(x) & \forall x \in \Omega,\ j = 1, \dots, m, \\ c_j(x) & \forall x \in \Omega,\ j = m+1, \dots, M, \end{cases} \tag{A.7}$$

$$\bar{c}_0(x) = a^{-3} \left[1 - a^3 \sum_{j=1}^{M} \bar{c}_j(x) \right] \qquad \forall x \in \Omega_s.$$

It is easy to see that $(\bar{c}_1, \dots, \bar{c}_M) \in V_a$. Moreover,

$$a^3 \bar{c}_0(x) = a^3 c_0(x) + \alpha \chi_{\omega_0(\alpha)}(x) \geq \alpha \qquad \text{a.e. } x \in \Omega_s, \tag{A.8}$$

implying (A.6) with $\theta_1 = \alpha$. Clearly, $\sum_{j=1}^{M} \|\bar{c}_j - c_j\|_{L^1(\Omega)} \leq \alpha a^{-3} \sum_{j=1}^{m} |K_j(\alpha)|$. Moreover,

$$\left\| \sum_{j=1}^{M} q_j \bar{c}_j - \sum_{j=1}^{M} q_j c_j \right\|_{H^{-1}(\Omega)} \leq \alpha a^{-3} \left\| \sum_{j=1}^{m} q_j \chi_{K_j(\alpha)} \right\|_{L^2(\Omega)} \leq \alpha a^{-3} \sqrt{\sum_{j=1}^{m} q_j^2 |K_j(\alpha)|}.$$

Therefore, $\|\bar{c} - c\|_X < \varepsilon/2$, provided that $\alpha > 0$ is small enough.

If $x \in K_j(\alpha)$ for some j with $1 \le j \le m$, then $c_j(x) \ge 1/(2Ma^3)$, and $\bar{c}_j(x) \ge 1/(4Ma^3)$ since $0 < \alpha < 1/(4M)$. By the mean-value theorem and the fact that $S'_{-1}(u) = \log u$ for any $u > 0$, there exists $\eta_j(x)$ with $\bar{c}_j(x) \le \eta_j(x) \le c_j(x)$ such that

$$\begin{aligned} S_{-1}\left[\bar{c}_j(x)\right] - S_{-1}\left[c_j(x)\right] &= \left[\bar{c}_j(x) - c_j(x)\right] \log \eta_j(x) \\ &\le -\alpha a^{-3} \log \bar{c}_j(x) \le \alpha a^{-3} \log\left(4Ma^3\right). \end{aligned}$$

By the same argument using (A.8) and the definition of $\omega_0(\alpha)$, we obtain

$$S_{-1}\left(\bar{c}_0(x)\right) - S_{-1}\left(c_0(x)\right) \le \alpha a^{-3} \log\left(a^{-3}\alpha\right) \qquad \text{a.e. } x \in \omega_0(\alpha).$$

Consequently, we have by (3.13), (3.4), and the embedding $L^2(\Omega_s) \hookrightarrow H^{-1}(\Omega_s)$ that

$$\begin{aligned} F_a[\bar{c}] - F_a[c] = \frac{1}{2}\int_{\Omega_s} &\left(\sum_{j=1}^M q_j c_j - \alpha a^{-3} \sum_{j=1}^m q_j \chi_{K_j(\alpha)}\right) \\ &L\left(\sum_{j=1}^M q_j c_j - \alpha a^{-3} \sum_{j=1}^m q_j \chi_{K_j(\alpha)}\right) dx \\ &- \frac{1}{2}\int_{\Omega_s} \left(\sum_{j=1}^M q_j c_j\right) L\left(\sum_{j=1}^M q_j c_j\right) dx - \alpha a^{-3} \sum_{j=1}^m \int_{K_j(\alpha)} \mu_{aj} dx \\ &+ \beta^{-1} \sum_{j=0}^m \int_{\omega_0(\alpha)} \left[S_{-1}(\bar{c}_j) - S_{-1}(c_j)\right] dx \\ \le \frac{1}{2}\alpha^2 a^{-6} &\int_{\Omega_s} \left(\sum_{j=1}^m q_j \chi_{K_j(\alpha)}\right) L\left(\sum_{j=1}^m q_j \chi_{K_j(\alpha)}\right) dx \\ &- \alpha a^{-3} \int_{\Omega_s} \left(\sum_{j=1}^m q_j \chi_{K_j(\alpha)}\right) L\left(\sum_{j=1}^m q_j c_j\right) dx \\ &+ \alpha a^{-3} \sum_{j=1}^m \|\mu_{aj}\|_{L^\infty(\Omega_s)} |K_j(\alpha)| \\ &+ \beta^{-1} \alpha a^{-3} \log\left(a^{-3}\alpha\right) |\omega_0(\alpha)| + \beta^{-1} \alpha a^{-3} \log\left(4Ma^3\right) |\omega_0(\alpha)| \\ \le C\alpha^2 a^{-6} &\left\| \sum_{j=1}^m q_j \chi_{K_j(\alpha)} \right\|^2_{L^2(\Omega_s)} \\ &+ \alpha a^{-3} \left\| L\left(\sum_{j=1}^M q_j c_j\right) \right\|_{L^\infty(\Omega)} \sum_{j=1}^m |q_j| |K_j(\alpha)| \\ &+ \alpha a^{-3} \sum_{j=1}^m \|\mu_{aj}\|_{L^\infty(\Omega_s)} |K_j(\alpha)| + \beta^{-1} \alpha a^{-3} \log\left(8Ma^3\alpha\right) \sum_{j=1}^m |K_j(\alpha)| \\ = \alpha \sum_{j=1}^M &\left[C\alpha a^{-6} q_j^2 + a^{-3}|q_j| \left\| L\left(\sum_{j=1}^M q_j c_j\right) \right\|_{L^\infty(\Omega)} + a^{-3} \|\mu_{aj}\|_{L^\infty(\Omega_s)} \right. \\ &\left. + \beta^{-1} a^{-3} \log\left(8Ma^3\alpha\right) \right] |K_j(\alpha)|, \end{aligned}$$

where $C > 0$ is a constant independent of α. Thus, $F_a[\bar{c}] - F_a[c]$ is nonpositive for $\alpha > 0$ sufficiently small. It is strictly negative, if $|\omega_0(\alpha)| = \sum_{j=1}^m |K_j(\alpha)| > 0$ for all $\alpha > 0$, i.e., if $|\{x \in \Omega_s : a^3 c_0(x) < \alpha\}| > 0$ for all $\alpha > 0$.

We now construct $\hat{c} \in V_a$ that satisfies (2.6) with c replaced by $\hat{c}$, $\|\hat{c} - \bar{c}\|_X < \varepsilon/2$, and $F_a[\hat{c}] \leq F_a[\bar{c}]$ with a strict inequality if there exists $j \in \{1, \ldots, M\}$ such that $|\{x \in \Omega_s : a^3 c_j(x) < \alpha\}| > 0$ for all $\alpha > 0$, all these implying that $\hat{c}$ satisfies all the desired properties. If there exists $\theta_2 \in (0, 1)$ such that $c_j(x) \geq \theta_2$ for a.e. $x \in \Omega_s$ and all $j = 1, \ldots, M$, then $\hat{c} = \bar{c}$ with $0 < \alpha < \theta_2/2$ (cf. (A.7)) satisfies all the desired properties with $\theta_0 = \min(\theta_1, \theta_2/2)$. Assume otherwise there exists $j_0 \in \{1, \ldots, M\}$ such that $|\{x \in \Omega_s : c_{j_0}(x) < \alpha\}| > 0$ for all $\alpha > 0$. This means that $|\{x \in \Omega_s : \bar{c}_{j_0}(x) < \alpha\}| > 0$ for all $\alpha > 0$.

Define

$$\begin{aligned} \sigma_j(\alpha) &= \{x \in \Omega_s : a^3 \bar{c}_j(x) < \alpha\} \qquad \forall \alpha > 0, \quad j = 1, \ldots, M, \\ J_0 &= \{j \in \{1, \ldots, M\} : |\sigma_j(\alpha)| > 0 \ \forall \alpha > 0\}, \\ J_1 &= \{1, \ldots, M\} \setminus J_0. \end{aligned}$$

Clearly, $J_0 \neq \emptyset$. If $J_1 \neq \emptyset$, then there exists $\alpha_2 > 0$ such that

$$a^3 \bar{c}_j(x) \geq \alpha_2 \qquad \text{a.e. } x \in \Omega_s, \quad \forall j \in J_1.$$

Define for $0 < \alpha < \min\{\alpha_2, \theta_1/M\}$ and $1 \leq j \leq M$

$$\hat{c}_j(x) = \begin{cases} \bar{c}_j(x) + \alpha a^{-3} \chi_{\sigma_j(\alpha)}(x) & \text{if } j \in J_0 \\ \bar{c}_j(x) & \text{if } j \in J_1 \end{cases} \qquad \forall x \in \Omega.$$

$$\hat{c}_0(x) = a^{-3} \left[1 - \sum_{j=1}^M a^3 \hat{c}_j(x) \right] \qquad \forall x \in \Omega_s.$$

Notice by (A.6) that

$$\begin{aligned} a^3 \hat{c}_0(x) &= 1 - \sum_{j=1}^M a^3 \hat{c}_j(x) = 1 - \sum_{j=1}^M a^3 \bar{c}_j(x) - \alpha \sum_{j \in J_0} \chi_{\sigma_j(\alpha)} \\ &\geq \theta_1 - \alpha M > 0 \qquad \text{a.e. } x \in \Omega_s. \end{aligned}$$

Thus, $\hat{c} = (\hat{c}_1, \ldots, \hat{c}_M) \in V_a$. Clearly, (2.6) holds true for $\theta_0 = \min\{\alpha, \alpha_2, \theta_1 - \alpha M\}$. Applying the same argument used above, we obtain that $\|\hat{c} - \bar{c}\|_X < \varepsilon/2$ for $\alpha > 0$ small enough.

We have now by the mean-value theorem that

$$\begin{aligned} \sum_{j=1}^M \int_{\Omega_s} [S_{-1}(\hat{c}_j) - S_{-1}(\bar{c}_j)] \, dx &= \sum_{j \in J_0} \int_{\sigma_j(\alpha)} [S_{-1}(\hat{c}_j) - S_{-1}(\bar{c}_j)] \, dx \\ &\leq \alpha a^{-3} \log\left(2\alpha a^{-3}\right) \sum_{j \in J_0} |\sigma_j(\alpha)|. \end{aligned}$$

Similarly, we have by (2.6) that

$$\int_{\Omega_s} [S_{-1}(\hat{c}_0) - S_{-1}(\bar{c}_0)] \, dx \leq -\alpha a^{-3} \log\left(a^{-3}\theta_0\right) \sum_{j \in J_0} |\sigma_j(\alpha)|.$$

Consequently, we have by (3.13) and a similar argument that

$$
\begin{aligned}
F_a[\hat{c}] - F_a[\bar{c}] = {} & \frac{1}{2}\int_{\Omega_s}\left(\sum_{j=1}^{M} q_j\bar{c}_j + \alpha a^{-3}\sum_{j\in J_0} q_j\chi_{\sigma_j(\alpha)}\right) \\
& L\left(\sum_{j=1}^{M} q_j\bar{c}_j + \alpha a^{-3}\sum_{j\in J_0} q_j\chi_{\sigma_j(\alpha)}\right)dx \\
& - \frac{1}{2}\int_{\Omega_s}\left(\sum_{j=1}^{M} q_j\bar{c}_j\right) L\left(\sum_{j=1}^{M} q_j\bar{c}_j\right)dx + \alpha a^{-3}\sum_{j\in J_0}\int_{\sigma_j(\alpha)}\mu_{aj}dx \\
& + \beta^{-1}\sum_{j\in J_0}\int_{\sigma_j(\alpha)}\left[S_{-1}(\hat{c}_j) - S_{-1}(\bar{c}_j)\right]dx \\
& + \beta^{-1}\int_{\Omega_s}\left[S_{-1}(\hat{c}_0) - S_{-1}(\bar{c}_0)\right]dx \\
\le {} & C\alpha^2 a^{-6}\left\|\sum_{j\in J_0} q_j\chi_{\sigma_j(\alpha)}\right\|^2_{L^2(\Omega_s)} + \alpha a^{-3}\left\|\sum_{j=1}^{M} q_j\bar{c}_j\right\|_{L^\infty(\Omega_s)}\sum_{j\in J_0}|q_j||\sigma_j(\alpha)| \\
& + \alpha a^{-3}\sum_{j\in J_0}\|\mu_{aj}\|_{L^\infty(\Omega_s)}|\sigma_j(\alpha)| + \beta^{-1}\alpha a^{-3}\log(2\alpha/\theta_0)\sum_{j\in J_0}|\sigma_j(\alpha)| \\
\le {} & \alpha\sum_{j\in J_0}\left[C\alpha a^{-6}q_j^2 + a^{-3}|q_j|\left\|\sum_{j=1}^{M} q_j\bar{c}_j\right\|_{L^\infty(\Omega_s)} + a^{-3}\|\mu_{aj}\|_{L^\infty(\Omega_s)}\right. \\
& \left. + \beta^{-1}a^{-3}\log(2\alpha/\theta_0)\right]|\sigma_j(\alpha)|.
\end{aligned}
$$

Since $J_0 \neq \emptyset$, this is strictly negative if $\alpha > 0$ is sufficiently small. The case that $J_1 = \emptyset$ can be treated similarly. ∎

Acknowledgments. The author thanks Dr. Jianwei Che, Dr. Joachim Dzubiella, and Dr. Benzhuo Lu for helpful discussions on the background of the underlying problem.

REFERENCES

[1] R. Allen, J.-P. Hansen, and S. Melchionna, *Electrostatic potential inside ionic solutions confined by dielectrics: A variational approach*, Phys. Chem. Chem. Phys., 3 (2001), pp. 4177–4186.

[2] M. Z. Born, *Volumen und Hydratationswärme der Ionen*, Phys., 1 (1920), pp. 45–48.

[3] I. Borukhov, D. Andelman, and H. Orland, *Steric effects in electrolytes: A modified Poisson-Boltzmann equation*, Phys. Rev. Lett., 79 (1997), pp. 435–438.

[4] D. L. Chapman, *A contribution to the theory of electrocapillarity*, Phil. Mag., 25 (1913), pp. 475–481.

[5] J. Che, J. Dzubiella, B. Li, and J. A. McCammon, *Electrostatic free energy and its variations in implicit solvent models*, J. Phys. Chem. B, 112 (2008), pp. 3058–3069.

[6] T. Chou, *Physics of cellular materials: Basic electrostatics*, IPAM Lecture Notes (2002).

[7] J. B. Conway, *A Course in Functional Analysis*, Springer-Verlag, Berlin, 1985.

[8] M. E. Davis and J. A. McCammon, *Electrostatics in biomolecular structure and dynamics*, Chem. Rev., 90 (1990), pp. 509–521.
[9] P. Debye and E. Hückel, *Zur theorie der elektrolyte*, Physik. Zeitschr., 24 (1923), pp. 185–206.
[10] B. Derjaguin and L. Landau, *A theory of the stability of strongly charged lyophobic sols and the coalescence of strongly charged particles in electrolytic solution*, Acta Phys.-Chim. USSR, 14 (1941), pp. 633–662.
[11] M. Fixman, *The Poisson-Boltzmann equation and its applications to polyelectrolytes*, J. Chem. Phys., 70 (1979), pp. 4995–5005.
[12] F. Fogolari and J. M. Briggs, *On the variational approach to Poisson-Boltzmann free energies*, Chem. Phys. Lett., 281 (1997), pp. 135–139.
[13] F. Fogolari, P. Zuccato, G. Esposito, and P. Viglino, *Biomolecular electrostatics with the linearized Poisson-Boltzmann equation*, Biophys. J., 76 (1999), pp. 1–16.
[14] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1998.
[15] M. K. Gilson, M. E. Davis, B. A. Luty, and J. A. McCammon, *Computation of electrostatic forces on solvated molecules using the Poisson-Boltzmann equation*, J. Phys. Chem., 97 (1993), pp. 3591–3600.
[16] M. Gouy, *Sur la constitution de la charge électrique a la surface d'un électrolyte*, J. de Phys., 9 (1910), pp. 457–468.
[17] D. C. Grahame, *The electrical double layer and the theory of electrocapillarity*, Chem. Rev., 32 (1947), pp. 441–501.
[18] P. Grochowski and J. Trylska, *Continuum molecular electrostatics, salt effects and counterion binding—A review of the Poisson-Boltzmann model and its modifications*, Biopolymers, 89 (2008), pp. 93–113.
[19] J.-P. Hansen and H. Löwen, *Effective interactions between electric double layers*, Annu. Rev. Phys. Chem., 51 (2000), pp. 209–242.
[20] V. Kralj-Iglic and A. Iglic, *A simple statistical mechanical approach to the free energy of the electric double layer including the excluded volume effect*, J. Phys. II (France), 6 (1996), pp. 477–491.
[21] W. Littman, G. Stampacchia, and H. F. Weinberger, *Regular points for elliptic equations with discontinuous coefficients*, Ann. Scuola Norm. Sup. Pisa, 3 (1963), pp. 43–77.
[22] B. Lu, X. Cheng, T. Hou, and J. A. McCammon, *Calculation of the Maxwell stress tensor and the Poisson-Boltzmann force on a solvated molecular surface using hypersingular boundary integrals*, J. Chem. Phys., 123 (2005), p. 084904.
[23] B. Lu, X. Cheng, J. Huang, and J. A. McCammon, *Order N algorithm for computation of electrostatic interaction in biomolecular systems*, PNAS, 103 (2006), pp. 19314–19319.
[24] B. Lu, D. Zhang, and J. A. McCammon, *Computation of electrostatic forces between solvated molecules determined by Poisson-boltzmann equations using a boundary element method*, J. Chem. Phys., 122 (2005), p. 214102.
[25] E. J. MacShane, *Integration*, Princeton University Press, Princeton, NJ, 1947.
[26] E. S. Reiner and C. J. Radke, *Variational approach to the electrostatic free energy in charged colloidal suspensions: General theory for open systems*, J. Chem. Soc. Faraday Trans., 86 (1990), pp. 3901–3912.
[27] B. Roux and T. Simonson, *Implicit solvent models*, Biophys. Chem., 78 (1999), pp. 1–20.
[28] K. A. Sharp and B. Honig, *Calculating total electrostatic energies with the nonlinear Poisson-Boltzmann equation*, J. Phys. Chem., 94 (1990), pp. 7684–7692.
[29] K. A. Sharp and B. Honig, *Electrostatic interactions in macromolecules: Theory and applications*, Annu. Rev. Biophys. Biophys. Chem., 19 (1990), pp. 301–332.
[30] R. Temam, *Navier-Stokes Equations: Theory and Numerical Analysis*, 3rd ed., North-Holland, Amsterdam, 1984.
[31] E. J. W. Verwey and J. T. G. Overbeek, *Theory of the Stability of Lyophobic Colloids*, Elsevier, Amsterdam, 1948.
[32] K. Yosida, *Functional Analysis*, 6th ed., Springer-Verlag, Berlin, 1994.

SIAM J. MATH. ANAL.
Vol. 40, No. 6, pp. 2567–2580

WAVE BREAKING AND PERSISTENCE PROPERTIES FOR THE DISPERSIVE ROD EQUATION*

ZHENGGUANG GUO† AND YONG ZHOU‡

Abstract. This paper is concerned with some aspects of blow up of solutions and persistence properties. Firstly, we will try to give sufficient conditions on the initial data, which guarantee finite time singularity formation for the corresponding solutions. Then a particular class of initial data for the periodic case is also considered in this paper. Finally, we investigate the persistence properties of the strong solutions.

Key words. rod equation, convolution problem, blow up, persistence property

AMS subject classifications. 30C70, 37L05, 35Q58, 58E35

DOI. 10.1137/080734704

1. Introduction.

Although a rod is always three-dimensional, if its diameter is much less than the axial length scale, one-dimensional equations can give a good description of the motion of the rod. Recently Dai [18] derived a new (one-dimensional) nonlinear dispersive equation including extra nonlinear terms involving second-order and third-order derivatives for a compressible hyperelastic material. The equation reads

$$v_\tau + \sigma_1 v v_\xi + \sigma_2 v_{\xi\xi\tau} + \sigma_3(2v_\xi v_{\xi\xi} + v v_{\xi\xi\xi}) = 0,$$

where $v(\xi,\tau)$ represents the radial stretch relative to a prestressed state, $\sigma_1 \neq 0$, $\sigma_2 < 0$, and $\sigma_3 \leq 0$ are constants determined by the prestress and the material parameters. If one introduces the following transformations

$$\tau = \frac{3\sqrt{-\sigma_2}}{\sigma_1}t, \quad \xi = \sqrt{-\sigma_2}x,$$

then the above equation turns into

$$u_t - u_{xxt} + 3uu_x = \gamma(2u_x u_{xx} + uu_{xxx}), \tag{1.1}$$

where $\gamma = 3\sigma_3/(\sigma_1\sigma_2)$. In [19], the authors derived that the value range of γ is from -29.4760 to 3.4174 for some special compressible materials. In this paper, from the mathematical viewpoint, we regard γ as a real number.

When $\gamma = 1$ in (1.1), we recover the shallow water (Camassa–Holm) equation derived physically by Camassa and Holm in [4] by approximating directly the Hamiltonian for Euler's equations in the shallow water regime, where $u(x,t)$ represents the free surface above a flat bottom. Recently, the alternative derivations of the

*Received by the editors September 7, 2008; accepted for publication (in revised form) December 9, 2008; published electronically March 20, 2009. This work was partially supported by the Shanghai Rising-Star (08QH14006), Fok Ying Tung Education Foundation (111002), and Shuguang Project (07SG29).

http://www.siam.org/journals/sima/40-6/73470.html

†Department of Mathematics, Zhejiang Normal University, Jinhua, Zhejiang 321004, China and Department of Mathematics, East China Normal University, Shanghai 200026, China (gzg19801213@yahoo.com.cn).

‡Corresponding author. Department of Mathematics, Zhejiang Normal University, Jinhua, Zhejiang 321004, China (yzhoumath@zjnu.edu.cn).

Camassa–Holm equation were presented in [5, 25, 13]. Some satisfactory results have been obtained for this shallow water equation recently. Local well-posedness for the initial datum $u_0(x) \in H^s$, with $s > 3/2$ was proved by several authors [10, 26, 32, 34]. For the initial data with lower regularity, we refer to Molinet's paper [30] and also the recent paper [3]. Moreover, wave breaking for a large class of initial data has been established in [8, 9, 10, 11, 26, 29, 36, 40, 41]. Later, in [20], the first author considered the Camassa–Holm equation with weakly dissipative term and established blow-up criteria and sufficient conditions on global existence of the solutions. Recently, in [24], among others, Himonas et al. showed the infinite propagation speed for the Camassa–Holm equation in the sense that a strong solution of the Cauchy problem with compact initial profile cannot be compactly supported at any later time unless it is the zero solution, which is an improvement of previous results in this direction obtained in [6, 21, 22]. These results were recently extended in [23] for a large range of physically important equations. However, in [38], global existence of weak solutions is proved, but uniqueness is obtained only under an a priori assumption that is known to hold only for initial data $u_0(x) \in H^1$ such that $u_0 - u_{0xx}$ is a sign-definite Radon measure (under this condition, global existence and uniqueness was also shown in [14], in which case a stronger concept of weak solutions can be introduced). Actually, there are approaches towards unique global weak solutions for initial data in H^1, the important distinction being between conservative solutions [3] and dissipative solutions [2]. The solitary waves of the Camassa–Holm equation are peaked solitons [4] with $u(x,t) = e^{-|x-ct|}$, where $c \in \mathbb{R}$ is the wave speed. The orbital stability of the peakons was shown by Constantin and Strauss [16]. An alternative approach is proposed in the paper [15]. Here it is worth it to point out that the peakons replicate a feature that is characteristic for the waves of great height-waves of largest amplitude that are exact solutions of the governing equations for water waves; cf. [7, 35, 12].

If $\gamma = 0$, (1.1) is the Benjamin–Bona–Mahony (BBM) equation, a well-known model for surface waves in a canal [1], and its solutions are global.

For general $\gamma \in \mathbb{R}$, much of the early results were proven by Constantin and Strauss in [17] first. Local well-posedness of strong solutions to (1.1) was established by applying Kato's theory, and some sufficient conditions on the initial data were found to guarantee the finite blow up of the corresponding solutions for the spatially nonperiodic case. Later, Zhou [43] proved the well-posedness result in detail, and various refined sufficient conditions on the initial data were found to guarantee the finite time blow up of the corresponding solutions for both the spatially periodic and nonperiodic case; he also applied the best constant of a convolution problem to establish the corresponding blow-up criteria [44]. Recently, blow-up criteria were presented in [37, 39, 42]. It should be mentioned that Liu and Zhou improved their previous results by applying the optimal constant on a kind of Sobolev inequality in [27] (see also a recent paper [28]). Mustafa [31] established a solution with constant H^1 energy. For $\gamma < 1$, the stability of solitary waves was established in [17, 45].

We now finish this introduction by outlining the rest of the paper. In section 2, we recall the local well-posedness for (1.1) with initial datum $u_0 \in H^s$, $s > 3/2$, and the lifespan of the corresponding solution is finite if and only if its first-order derivative blows up. In section 3, we investigate various sufficient conditions of the initial datum to guarantee the finite time blow up. Persistence properties are established for (1.1) in section 4.

2. Preliminaries. In this section, we would like to list some useful theorems for later use.

THEOREM 2.1 (see [39]). *Let the initial datum $u_0(x) \in H^s(\Omega)$, $s > 3/2$. Then there exist $T = T(\| u_0 \|) > 0$ and a unique solution $u(x,t)$, which depends continuously on the initial datum u_0 to* (1.1) *such that*

$$u \in C([0,T); H^s(\Omega)) \cap C^1\left([0,T); H^{s-1}(\Omega)\right).$$

Moreover, the following two quantities $E_1(u)$ and $E_2(u)$ are invariants with respect to time t for (1.1):

$$E_1(u) = \int_\Omega \left(u^2 + u_x^2\right) dx, \quad E_2(u) = \int_\Omega \left(u^3 + \gamma u u_x^2\right) dx.$$

Actually, the local well-posedness was proved for both the periodic and nonperiodic case in the above paper; these two invariants play an important role in considering blow-up phenomenon.

The maximum value of T in Theorem 2.1 is called the lifespan of the solution, in general. If $T < \infty$, that is, $\limsup_{t \uparrow T} \| u(.,t) \|_{H^s} = \infty$, we say that the solution blows up in finite time. The following theorem tells us that the solution blows up if and only if the first-order derivative blows up. This phenomenon coincides physically with the rod breaking.

THEOREM 2.2 (see [17]). *Let $u_0(x) \in H^s(\Omega)$, $s > 3/2$, and $u(x,t)$ be the corresponding solution to problem* (1.1)*, with lifespan T. Then*

$$\sup_{x \in \Omega, 0 \le t < T} |u(x,t)| \le C(\| u_0 \|_{H^1}). \tag{2.1}$$

T is bounded if and only if

$$\lim_{t \uparrow T} \inf \inf_{x \in \Omega} \{\gamma u_x(x,t)\} = -\infty. \tag{2.2}$$

For $\gamma \neq 0$, we set

$$m(t) := \inf_{x \in \Omega} (u_x(x,t) sign\{\gamma\}), \quad t > 0, \tag{2.3}$$

where $sign\{a\}$ in the sign function of $a \in \mathbb{R}$ and we set $m_0 := m(0)$. Then for every $t \in [0,T)$, there exists at least one point $\xi(t) \in \Omega$, with $m(t) = u_x(\xi(t), t)$. Just as the proof given in [11], one can show the following property of m(t).

THEOREM 2.3 (see [11]). *Let $u(t)$ be the solution to* (1.1) *on $[0,T)$, with initial data $u_0 \in H^s(\Omega)$, $s > 3/2$, as given by Theorem* 2.1*. Then the function $m(t)$ is almost everywhere differentiable on $[0,T)$, with*

$$\frac{dm(t)}{dt} = u_{xt}(\xi(t), t), \quad a.e. \quad on(0,T).$$

To consider the quantity $m(t)$ for wave breaking comes from an idea of Seliger [33] originally, the rigorous regularity proof is given in [11] for the Camassa–Holm equation. As we know, the operator $(I - \partial_x^2)^{-1}$ can be expressed by

$$\left(I - \partial_x^2\right)^{-1} f = G * f = \int_\Omega G(x-y) f(y)\, dy$$

for all $f \in L^2(\Omega)$, where $G(x)$ is the associated Green's function. For the periodic case and the whole line case $G(x) = \frac{\cosh(x - [x] - \frac{1}{2})}{2 \sinh(\frac{1}{2})}$ and $G(x) = \frac{1}{2} e^{-|x|}$, respectively, where $[x]$ denotes the integer part of x.

In this paper, $\mathbb{S} := \mathbb{R}/\mathbb{Z}$ stands for the unit circle and $\Omega = \mathbb{R}$ or $\mathbb{S}$. Then (1.1) can be rewritten as

$$u_t + \gamma u u_x + \partial_x G * \left(\frac{3-\gamma}{2} u^2 + \frac{\gamma}{2} u_x^2 \right) = 0. \tag{2.4}$$

In what follows, just for simplicity, we assume that $0 < \gamma < 3$ and introduce the following notation:

$$F(u) = \frac{3-\gamma}{2} u^2 + \frac{\gamma}{2} u_x^2.$$

3. Blow-up phenomena. In this section, we investigate sufficient conditions on the initial data which guarantee the finite time blow up of the corresponding solutions to (1.1). Firstly, we consider the following convolution problem. We can compute

$$\begin{aligned} G * \left(\frac{\gamma\alpha^2}{2} u^2 + \frac{\gamma}{2} u_x^2 \right)(x) &= \frac{1}{2} e^{-x} \int_{-\infty}^{x} e^y \left(\frac{\gamma\alpha^2}{2} u^2 + \frac{\gamma}{2} u_x^2 \right) dy \\ &\quad + \frac{1}{2} e^{x} \int_{x}^{\infty} e^{-y} \left(\frac{\gamma\alpha^2}{2} u^2 + \frac{\gamma}{2} u_x^2 \right) dy \\ &\geq \frac{1}{2} e^{-x} \int_{-\infty}^{x} \gamma\alpha e^y u u_x dy - \frac{1}{2} e^{x} \int_{x}^{\infty} \gamma\alpha e^{-y} u u_x dy \\ &= \frac{\gamma\alpha}{2} u^2 - \frac{\gamma\alpha}{4} \int_{-\infty}^{x} e^y u^2 dy - \frac{\gamma\alpha}{4} \int_{x}^{\infty} e^{-y} u^2 dy, \end{aligned}$$

where α is a positive number. So we obtain

$$G * \left(\frac{\gamma\alpha^2 + \gamma\alpha}{2} u^2 + \frac{\gamma}{2} u_x^2 \right)(x) \geq \frac{\gamma\alpha}{2} u^2(x).$$

Now we let

$$\alpha^2 + \alpha = \frac{3-\gamma}{\gamma}. \tag{3.1}$$

Therefore,

$$G * \left(\frac{3-\gamma}{2} u^2 + \frac{\gamma}{2} u_x^2 \right)(x) \geq \frac{\gamma\alpha}{2} u^2(x). \tag{3.2}$$

Moreover, $\frac{\gamma\alpha}{2}$ is the best constant if and only if

$$\alpha u = u_x \quad \text{in } (-\infty, x) \quad \text{and} \quad -\alpha u = u_x \quad \text{in } (x, \infty),$$

which can be solved as $u = \lambda e^{-\alpha|x-y|}$ for some $\lambda, y \in \mathbb{R}$, α is a positive root determined by (3.1).

The following theorem refines the result first obtained in [17] by admitting a larger class of initial data.

THEOREM 3.1. *Assume $u_0(x) \in H^3(\mathbb{R})$ is odd, and $u_0'(0) < 0$. Then the corresponding solution of* (1.1) *blows up in finite time.*

Proof. Let $T > 0$ is the maximal time of existence of the solution $u(x,t)$ to (1.1), with initial data u_0. As one can check, the function $-u(-x,t)$ is also a solution to

(1.1), with initial data $-u_0(-x)$. By this fact, $-u_0(-x) = u_0(x)$ and the uniqueness of solution to (1.1), we get

$$u(x,t) = -u(-x,t) \quad \text{for all } t \in [0,T), \quad x \in \mathbb{R}.$$

Differentiating both sides of (2.4) with respect to variable x, we obtain

$$u_{xt} = -\frac{\gamma}{2}u_x^2 + \frac{3-\gamma}{2}u^2 - \gamma u u_{xx} - \left[G * \left(\frac{3-\gamma}{2}u^2 + \frac{\gamma}{2}u_x^2\right)\right](x,t). \tag{3.3}$$

In view of the above analysis, we have

$$u(0,t) = u_{xx}(0,t) = 0, \quad t \in [0,T).$$

Therefore,

$$\frac{du_x(0,t)}{dt} \le -\frac{\gamma}{2}u_x^2(0,t).$$

From the hypothesis, we have $u_0'(0) < 0$. Therefore, $u_x(0,t) < 0$ for all $t \in [0,T)$. Solving the above inequality, we get

$$-\frac{1}{u_x(0,t)} + \frac{1}{u_0'(0)} \le -\frac{\gamma}{2}t.$$

We conclude that there exists T and

$$T \le -\frac{2}{\gamma u_0'(0)}$$

such that $\lim_{t\uparrow T} u_x(0,t) = -\infty$. Thus, we finish the proof by Theorem 2.2. □

The following theorem gives the blow-up phenomenon via initial potential data y_0, which reads as follows.

COROLLARY 3.1. *Suppose $y_0 \in H^1(\mathbb{R})$ is an odd function, satisfies $\int_0^\infty e^{-\xi}y_0(\xi)\,d\xi < 0$. Then the maximum time of existence of the corresponding solution $u(x,t)$ to (1.1) is finite.*

In fact, y_0 is an odd function; we can easily check that u_0 is also odd. At this time, $u_0'(0) = \int_0^\infty e^{-\xi}y_0(\xi)d\xi$. By the hypothesis and theorem above, we get the desired result. We omit the detailed proof for conciseness.

THEOREM 3.2. *Assume $u_0 \in H^3(\mathbb{S})$, $u_0 \not\equiv 0$, $\int_{\mathbb{S}}(u_0^3 + \gamma u_0 u_{0x}^2)dx = 0$, and one of the following conditions is satisfied:*

$$\text{(i)} \quad \gamma \in \left(\frac{3\sinh\left(\frac{1}{2}\right)}{2+\sinh\left(\frac{1}{2}\right)}, 3\right),$$

$$\text{(ii)} \quad \gamma \in \left(0, \frac{3\sinh\left(\frac{1}{2}\right)}{2+\sinh\left(\frac{1}{2}\right)}\right],$$

$$m(0) < -\frac{\sqrt{2}}{2}\alpha \parallel u_0 \parallel_{H^1(\mathbb{S})}.$$

Then the corresponding solution to (1.1) blows up in finite time.

Proof. By assumption and the invariance property of $E_2(u)$, we have

$$\int_{\mathbb{S}} u^3 + \gamma u u_x^2 dx = \int_{\mathbb{S}} u_0^3 + \gamma u_0 u_{0x}^2 dx = 0.$$

Therefore, $u(x,t)$ must change sign, so there exists at least one zero point on $\mathbb{S}$. Then for each $t \in [0,T)$, suppose there is a $\xi_t \in [0,1]$ such that $u(\xi_t,t)=0$, for $x \in \mathbb{S}$ we have

$$u^2(x,t)=\left(\int_{\xi_t}^{x} u_x dx\right)^2 \le (x-\xi_t)\int_{\xi_t}^{x} u_x^2 dx, \quad x \in \left[\xi_t, \xi_t+\frac{1}{2}\right]. \tag{3.4}$$

Thus, the relation above and an integration by parts yield

$$\begin{aligned}\int_{\xi_t}^{\xi_t+\frac{1}{2}} u^2 u_x^2 dx &\le \int_{\xi_t}^{\xi_t+\frac{1}{2}} (x-\xi_t)u_x^2\left(\int_{\xi_t}^{x} u_x^2\right)dx \\ &= \int_{\xi_t}^{\xi_t+\frac{1}{2}} (x-\xi_t)\left(\int_{\xi_t}^{x} u_x^2\right) d\left(\int_{\xi_t}^{x} u_x^2\right) \\ &\le \frac{1}{4}\left(\int_{\xi_t}^{\xi_t+\frac{1}{2}} u_x^2 dx\right)^2. \end{aligned} \tag{3.5}$$

Doing a similar estimate on $[\xi_t+\frac{1}{2}, \xi_t+1]$, we obtain

$$\int_{\mathbb{S}} u^2 u_x^2 dx \le \frac{1}{4}\left(\int_{\mathbb{S}} u_x^2 dx\right)^2. \tag{3.6}$$

In view of (3.4), we also have

$$\| u(x,t) \|_{L^\infty}^2 \le \frac{1}{2}\int_{\mathbb{S}} u_x^2 dx. \tag{3.7}$$

Let

$$q(t)=\int_{\mathbb{S}} u_x^3 dx, \quad t \ge 0.$$

Multiplying both sides of (3.3) with u_x^2, then integrating by parts, we obtain the equation for $q(t)$ as

$$\begin{aligned}\frac{dq(t)}{dt} &= -\frac{\gamma}{2}\int_{\mathbb{S}} u_x^4 dx + \frac{3(3-\gamma)}{2}\int_{\mathbb{S}} u^2 u_x^2 dx - 3\int_{\mathbb{S}} u_x^2 G*F(u) dx \\ &\le -\frac{\gamma}{2}\int_{\mathbb{S}} u_x^4 dx - \left(\frac{3\gamma}{4\sinh\left(\frac{1}{2}\right)} - \frac{3(3-\gamma)}{8}\right)\left(\int_{\mathbb{S}} u_x^2\right)^2, \end{aligned} \tag{3.8}$$

where we use the fact $\frac{1}{2\sinh(\frac{1}{2})} \le G(x) \le \frac{\cosh(\frac{1}{2})}{2\sinh(\frac{1}{2})}$.

If $\gamma \in (\frac{3\sinh(\frac{1}{2})}{2+\sinh(\frac{1}{2})}, 3)$, then $\frac{3\gamma}{4\sinh(\frac{1}{2})} - \frac{3(3-\gamma)}{8} \ge 0$. On the other hand, we have $\int_{\mathbb{S}} u_x^2 dx \ge \frac{2}{3}\| u_0 \|_{H^1}^2$. Therefore, the above inequality yields

$$\begin{aligned}\frac{dq(t)}{dt}dx &= -\frac{\gamma}{2}\int_{\mathbb{S}} u_x^4 dx + \frac{3(3-\gamma)}{2}\int_{\mathbb{S}} u^2 u_x^2 dx - 3\int_{\mathbb{S}} u_x^2 G*F(u) dx \\ &\le -\frac{\gamma}{2}\int_{\mathbb{S}} u_x^4 dx - \frac{4}{9}\left(\frac{3\gamma}{4\sinh\left(\frac{1}{2}\right)} - \frac{3(3-\gamma)}{8}\right)\| u_0 \|_{H^1}^4. \end{aligned} \tag{3.9}$$

For simplicity, we set $\mu = \frac{4}{9}(\frac{3\gamma}{4\sinh(\frac{1}{2})} - \frac{3(3-\gamma)}{8})$ and in view of Holder's inequality, we get

$$\begin{aligned}\frac{dq(t)}{dt} &= -\frac{\gamma}{2}\int_{\mathbb{S}} u_x^4 dx + \frac{3(3-\gamma)}{2}\int_{\mathbb{S}} u^2 u_x^2 dx - 3\int_{\mathbb{S}} u_x^2 G * F(u) dx \\ &\le -\frac{\gamma}{2} q^{\frac{4}{3}}(t) - \mu \parallel u_0 \parallel_{H^1}^4 . \end{aligned} \tag{3.10}$$

First, since $q(t) \le q_0 - \mu \parallel u_0 \parallel_{H^1}^4 t$, it is easy to find that there exist a time $t_0 \ge 0$ such that $q(t_0) < 0$. Then for $t > t_0$, we have

$$\frac{dq(t)}{dt} \le -\frac{\gamma}{2} q^{\frac{4}{3}}(t), \quad \text{with } q(t_0) < 0.$$

So the solution to (3.10) satisfies

$$q(t) \le \left(q^{-\frac{1}{3}}(t_0) + \frac{\gamma}{6}(t - t_0)\right)^{-3},$$

which reaches to $-\infty$ before t arrives at $T_0 = -6q^{-\frac{1}{3}}(t_0)/\gamma + t_0$.

On the other hand, since

$$\left|\int_{\mathbb{S}} u_x^3 dx\right| \le C(s) \parallel u_0 \parallel_{H^1}^2 \parallel u \parallel_{H^s} \quad \text{for} \quad s \in \left(\frac{3}{2}, 3\right],$$

then

$$\int_{\mathbb{S}} u_x^3 dx \ge \inf u_x(x,t) \parallel u_0 \parallel_{H^1}^2$$

shows that $\lim_{t\to T_0} \inf_{x\in\mathbb{S}} u_x(x,t) = -\infty$.

If $\gamma \in (0, \frac{3\sinh(\frac{1}{2})}{2+\sinh(\frac{1}{2})}]$, we can derive an equation for $m(t)$ as

$$\frac{dm}{dt} = -\frac{\gamma}{2}m^2 + \frac{3-\gamma}{2}u^2(\xi(t),t) - [G * F(u)](\xi(t),t) = 0.$$

Now combining the above equation and (3.2) together, we have

$$\frac{dm(t)}{dt} \le -\frac{\gamma}{2}m^2(t) + \frac{3-\gamma-\gamma\alpha}{4} \parallel u_0 \parallel_{H^1}^2 . \tag{3.11}$$

If

$$m(0) < -\left(\frac{3-\gamma-\gamma\alpha}{2\gamma}\right)^{\frac{1}{2}} \parallel u_0 \parallel_{H^1} = -\frac{\sqrt{2}}{2}\alpha \parallel u_0 \parallel_{H^1},$$

then we can conclude the solution to (3.11) goes to $-\infty$ in finite time. This completes the proof. □

Remark 3.1. $\int_{\mathbb{S}} u_0 dx = 0$ or $\int_{\mathbb{S}} y_0 dx = 0$ can be a substitute of the condition $\int_{\mathbb{S}}(u_0^3 + \gamma u_0 u_{0x}^2)dx = 0$; the theorem still holds. If $\gamma = 1$, we recover the case for the Camassa–Holm equation [10]. Moreover, one may find Theorem 3.2 is different from Theorem 3.4 in [42]. However, in [39], Yin didn't consider the case (ii) of this theorem for (1.1).

4. Persistence properties and unique continuation. In this section, we shall investigate persistence properties of the solution to (1.1) in L^∞-space. The main idea comes from a recent work of Himonas et al. [24].

THEOREM 4.1. *Assume that $u_0(x) \in H^s(\mathbb{R})$ and $s > \frac{3}{2}$, if for some $\theta \in (0,1)$,*

$$|u_0(x)|, \quad |u_{0x}(x)| \sim O\left(e^{-\theta x}\right) \quad as \quad x \uparrow \infty.$$

Then the corresponding strong solution $u(x,t) \in C([0,T]; H^s(\mathbb{R}))$ to (1.1) satisfies that

$$|u(x,t)|, \quad |u_x(x,t)| \sim O\left(e^{-\theta x}\right) \quad as \quad x \uparrow \infty$$

uniformly in the time interval $[0,T]$ before blow up.

Notation.

$$|u(x)| \sim O\left(e^{-\theta x}\right) \qquad as \quad x \uparrow \infty \qquad if \quad \lim_{x\to\infty} \frac{|u(x)|}{e^{-\theta x}} = L,$$

and

$$|u(x)| \sim o\left(e^{-\beta x}\right) \qquad as \quad x \uparrow \infty \qquad if \quad \lim_{x\to\infty} \frac{|u(x)|}{e^{-\beta x}} = 0.$$

Proof. The proof is organized as follows. Firstly, we will give estimates on $\| u(x,t) \|_\infty$ and $\| u_x(x,t) \|_\infty$. Here $\| \cdot \|_p$ means the L^p norm. Secondly, we use the weight function to obtain the desired result.

Multiplying (1.1) by u^{2n-1}, with $n \in \mathbb{Z}^+$, then integrating both sides with respect to x variable, we can get

$$\int_{\mathbb{R}} u^{2n-1} u_t dx + \gamma \int_{\mathbb{R}} u^{2n-1} u u_x dx + \int_{\mathbb{R}} u^{2n-1} \partial_x G * F(u) dx = 0. \tag{4.1}$$

The first term of the above identity is

$$\int_{\mathbb{R}} u^{2n-1} u_t dx = \frac{1}{2n} \frac{d}{dt} \| u(t) \|_{2n}^{2n} = \| u(t) \|_{2n}^{2n-1} \frac{d}{dt} \| u(t) \|_{2n} \tag{4.2}$$

and

$$\left| \int_{\mathbb{R}} u^{2n-1} u u_x dx \right| \leq \| u_x(t) \|_\infty \| u(t) \|_{2n}^{2n}. \tag{4.3}$$

In view of Holder's inequality

$$\int_{\mathbb{R}} u^{2n-1} \partial_x G * F(u) dx \leq \| u(t) \|_{2n}^{2n-1} \| \partial_x G * F(u) \|_{2n}, \tag{4.4}$$

so

$$\frac{d}{dt} \| u(t) \|_{2n} \leq \gamma \| u_x(t) \|_\infty \| u(t) \|_{2n} + \| \partial_x G * F(u) \|_{2n}. \tag{4.5}$$

In view of the Soblev embedding theorem, then there exists a constant $M > 0$ such that applying Gronwall's gives us

$$\| u(t) \|_{2n} \leq e^{Mt} \left(\| u(0) \|_{2n} + \int_0^t \| \partial_x G * F(u) \|_{2n} \, d\tau \right). \tag{4.6}$$

Note that

$$\lim_{p\to\infty} \| f \|_p = \| f \|_\infty \qquad \text{when} \qquad f \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R}). \tag{4.7}$$

Taking limits in (4.6) we obtain

$$\| u(t) \|_\infty \leq e^{Mt} \left(\| u(0) \|_\infty + \int_0^t \| \partial_x G * F(u) \|_\infty \, d\tau \right). \tag{4.8}$$

Next, we will establish an estimate on $\| u_x(t) \|_\infty$ using the same method as above. Differentiating (1.1) with respect to x variable produces the following equation:

$$u_{xt} + \gamma u u_{xx} + \gamma u_x^2 + \partial_x^2 G * F(u) = 0. \tag{4.9}$$

Multiplying the above identity by u_x^{2n-1}, considering the second term with integration by parts,

$$\int_{\mathbb{R}} u u_{xx} u_x^{2n-1} dx = -\frac{1}{2n} \int_{\mathbb{R}} u_x^{2n} u_x dx, \tag{4.10}$$

so we get

$$\int_{\mathbb{R}} u_x^{2n-1} u_{xt} dx + \gamma \int_{\mathbb{R}} u_x^{2n+1} dx - \frac{\gamma}{2n} \int_{\mathbb{R}} u_x^{2n} u_x dx + \int_{\mathbb{R}} u_x^{2n-1} \partial_x^2 G * F(u) dx = 0. \tag{4.11}$$

Similarly, one can get the inequality

$$\frac{d}{dt} \| u_x(t) \|_{2n} \leq 2\gamma \| u_x(t) \|_\infty \| u_x(t) \|_{2n} + \| \partial_x^2 G * F(u) \|_{2n}, \tag{4.12}$$

and therefore, as before, we obtain

$$\| u_x(t) \|_{2n} \leq e^{2Mt} \left(\| u_x(0) \|_{2n} + \int_0^t \| \partial_x^2 G * F(u) \|_{2n} \, d\tau \right). \tag{4.13}$$

Taking limits in (4.13) to obtain

$$\| u_x(t) \|_\infty \leq e^{2Mt} \left(\| u_x(0) \|_\infty + \int_0^t \| \partial_x^2 G * F(u) \|_\infty \, d\tau \right). \tag{4.14}$$

In order to get the desired result, we introduce the function $\psi_N(x)$, which is independent on t as follows:

$$\psi_N(x) = \begin{cases} 1, & x \leq 0, \\ e^{\theta x}, & x \in (0, N), \\ e^{\theta N}, & x \geq N, \end{cases} \tag{4.15}$$

where $N \in \mathbb{Z}^+$. From (1.1) we obtain

$$\psi_N u_t + \psi_N \gamma u u_x + \psi_N \partial_x G * F(u) = 0, \tag{4.16}$$

while for (4.9), we get

$$u_{xt} \psi_N + \psi_N \gamma u u_{xx} + \psi_N \gamma u_x^2 + \psi_N \partial_x^2 G * F(u) = 0. \tag{4.17}$$

In order to get the estimate on $u_x\psi_N$, we need to remove the second derivatives, by using integration by parts, we obtain

$$\begin{aligned}\left|\int_{\mathbb{R}}\psi_N u u_{xx}(u_x\psi_N)^{2n-1}dx\right| &= \left|\int_{\mathbb{R}} u(\psi_N u_x)^{2n-1}((u_x\psi_N)_x - u_x\psi_N')dx\right| \\ &= \left|\int_{\mathbb{R}} u\left(\frac{(u_x\psi_N)^{2n}}{2n}\right)_x dx - \int_{\mathbb{R}} u u_x\psi_N'(u_x\psi_N)^{2n-1}dx\right| \\ &\le 2(\| u \|_\infty + \| u_x(t) \|_\infty) \| u_x\psi_N \|_{2n}^{2n}, \end{aligned} \tag{4.18}$$

where we use the fact $0 \le \psi_N'(x) \le \psi_N(x)$ a.e. $x \in \mathbb{R}$. Next, we will do estimates on $\| u\psi_N \|_\infty$ and $\| u_x\psi_N \|_\infty$ step-by-step as before what were done on $\| u \|_\infty$ and $\| u_x \|_\infty$, we may get

$$\begin{aligned}\| u(t)\psi_N \|_\infty + \| u_x\psi_N \|_\infty &\le e^{2Mt}(\| u(0)\psi_N \|_\infty + \| u_x(0)\psi_N \|_\infty) \\ &\quad + e^{2Mt}\int_0^t (\| \psi_N\partial_x G * F(u) \|_\infty \\ &\quad + \| \psi_N\partial_x^2 G * F(u) \|_\infty)d\tau. \end{aligned} \tag{4.19}$$

On the other hand, computing the integral we see there exists a $c_1 > 0$, depending only on $\theta \in (0,1)$ such that

$$\psi_N(x)\int_{\mathbb{R}} e^{-|x-y|}\frac{1}{\psi_N(y)}dy \le c_1. \tag{4.20}$$

Therefore, one gets

$$\begin{aligned}|\psi_N\partial_x G * g^2(x)| &\le \frac{1}{2}\psi_N(x)\int_{\mathbb{R}} e^{-|x-y|}\frac{1}{\psi_N(y)}\psi_N(y)g(y)g(y)dy \\ &\le \frac{1}{2}\| \psi_N g \|_\infty \| g \|_\infty \left(\psi_N(x)\int_{\mathbb{R}} e^{-|x-y|}\frac{1}{\psi_N(y)}dy\right) \\ &\le c_1 \| \psi_N g \|_\infty \| g \|_\infty . \end{aligned} \tag{4.21}$$

Similarly,

$$|\psi_N\partial_x^2 G * g^2(x)| \le c_1 \| \psi_N g \|_\infty \| g \|_\infty . \tag{4.22}$$

Thus, combining (4.21) and (4.22) with (4.19), it follows that there exists a constant $C_1 = C_1(M,T) > 0$ such that

$$\begin{aligned}\| u(t)\psi_N \|_\infty + \| u_x\psi_N \|_\infty &\le C_1(\| u(0)\psi_N \|_\infty + \| u_x(0)\psi_N \|_\infty) \\ &\quad + C_1\int_0^t (\| u(\tau) \|_\infty + \| u_x(\tau) \|_\infty) \\ &\quad (\| u(\tau)\psi_N \|_\infty + \| u_x(\tau)\psi_N \|_\infty)d\tau \\ &\le C_1\Big(\| u(0)\psi_N \|_\infty + \| u_x(0)\psi_N \|_\infty \\ &\quad + \int_0^t (\| u(\tau)\psi_N \|_\infty + \| u_x(\tau)\psi_N \|_\infty)d\tau\Big). \end{aligned} \tag{4.23}$$

Then for any $N \in \mathbb{Z}^+$ and any $t \in [0,T]$, $x > 0$, we have

$$\begin{aligned}\| u(t)\psi_N \|_\infty + \| u_x\psi_N \|_\infty &\le C_1(\| u(0)\psi_N \|_\infty + \| u_x(0)\psi_N \|_\infty) \\ &\le C_1(\| u(0)e^{\theta x} \|_\infty + \| u_x(0)e^{\theta x} \|_\infty). \end{aligned} \tag{4.24}$$

Taking the limit as N goes to infinity in (4.24), we obtain

$$\left(\left|u(x,t)e^{\theta x}\right| + \left|u_x(x,t)e^{\theta x}\right|\right) \leq C_1 \left(\| u(0)e^{\theta x} \|_\infty + \| u_x(0)e^{\theta x} \|_\infty\right). \tag{4.25}$$

This completes the proof. □

The following result ensures that the only solution which can decay at a determined rate, at any two distinct times, is the trivial solution $u \equiv 0$.

THEOREM 4.2. *Assume that $u_0(x) \in H^s(\mathbb{R})$ and $s > \frac{3}{2}$, satisfies that for some $\delta \in (\frac{1}{2}, 1)$,*

$$|u_0(x)| \sim o(e^{-x}) \quad \text{and} \quad |u_{0x}(x)| \sim O\left(e^{-\delta x}\right) \qquad \text{as } x \uparrow \infty,$$

$u(x,t) \in C([0,T]; H^s(\mathbb{R}))$ is the corresponding strong solution to (1.1), *and there exists $t_1 \in (0,T]$ for some $T > 0$ such that*

$$|u(x,t_1)| \sim o\left(e^{-x}\right) \qquad \text{as } x \uparrow \infty, \tag{4.26}$$

then $u \equiv 0$.

Proof. Integrating (1.1) from 0 to t_1, we get

$$u(x,t_1) - u(x,0) + \gamma \int_0^{t_1} u u_x d\tau + \int_0^{t_1} \partial_x G * F(u) d\tau = 0. \tag{4.27}$$

According to the hypothesis, we easily have

$$u(x,t_1) - u(x,0) \sim o\left(e^{-x}\right) \quad \text{as } x \uparrow \infty. \tag{4.28}$$

At the same time, in view of Theorem 4.1 it follows that

$$\int_0^{t_1} u u_x d\tau \sim O\left(e^{-2\delta x}\right) \quad \text{as } x \uparrow \infty, \tag{4.29}$$

and so

$$\int_0^{t_1} u u_x d\tau \sim o\left(e^{-x}\right) \quad \text{as } x \uparrow \infty. \tag{4.30}$$

If $u(x,t) \neq 0$, the following deduction tells us the last term of (4.27) is infinitesimal with the same order not higher order of e^{-x}. Thus, a contradiction occurs.

$$\begin{aligned} \int_0^{t_1} \partial_x G * F(u) d\tau &= \partial_x G * \int_0^{t_1} F(u) d\tau \\ &= \partial_x G * f(x). \end{aligned} \tag{4.31}$$

However,

$$0 \leq f(x) \sim O\left(e^{-2\delta x}\right) \quad \text{so that} \quad f(x) \sim o\left(e^{-x}\right) \quad \text{as } x \uparrow \infty. \tag{4.32}$$

Therefore,

$$\partial_x G * f(x) = -\frac{1}{2} e^{-x} \int_{-\infty}^{x} e^y f(y) dy + \frac{1}{2} e^x \int_x^{\infty} e^{-y} f(y) dy. \tag{4.33}$$

From (4.32) it follows that

$$e^x \int_x^\infty e^{-y} f(y)dy = e^x \int_x^\infty e^{-y} o(e^{-y})dy = o(1)e^x \int_x^\infty e^{-2y}dy \sim o(1)e^{-x} \sim o(e^{-x}).$$

If $f(x) \neq 0$, one has that

$$\int_{-\infty}^x e^y f(y)dy \geq C_0 \quad \text{for } x \text{ large enough.} \tag{4.34}$$

Hence,

$$\begin{aligned} -\partial_x G * f(x) &= \frac{1}{2}e^{-x} \int_{-\infty}^x e^y f(y)dy - \frac{1}{2}e^x \int_x^\infty e^{-y} f(y)dy \\ &\geq \frac{C_0}{2}e^{-x} \quad \text{for } x \text{ large enough.} \end{aligned} \tag{4.35}$$

So a contradiction occurs by combination with (4.27)–(4.30) and (4.35). Thus, $f(x) \equiv 0$ and consequently, $u(x,t) \equiv 0$. The theorem is proved. □

At the end of this paper, we would like to make a simple comparison between the rod equation and the Camassa–Holm equation. In the whole paper, we do not discuss sufficient conditions to guarantee the global existence of smooth solutions to (1.1) for general γ, which was discussed in [17].

We know for (1.1), $y = u - u_{xx}$ satisfies

$$y_t + \gamma y_x u + 2\gamma y u_x + \frac{3(\gamma-1)}{2}\left(u^2\right)_x = 0.$$

It is a pity that we couldn't obtain the proper particle trajectory line equation to consider this problem just as it was done for the Camassa–Holm equation. Moreover, for $\gamma \neq 1$, few beautiful identities that appeared in the Camassa–Holm equation are obtained for (1.1).

Acknowledgments. The first author is greatly indebted to Prof. Zhou for his constructive suggestions and constant encouragement throughout this work. The authors thank the referees for their constructive and helpful suggestions and comments.

REFERENCES

[1] T. B. Benjamin, J. L. Bona, and J. J. Mahony, *Model equations for long waves in nonlinear dispersive systems*, Philos. Trans. R. Soc. Lond. Ser. A, 272 (1972), pp. 47–78.
[2] A. Bressan and A. Constantin, *Global dissipative solutions of the Camassa-Holm equation*, Anal. Appl., 5 (2007), pp. 1–27.
[3] A. Bressan and A. Constantin, *Global conservative solutions of the Camassa-Holm equation*, Arch. Ration. Mech. Anal., 183 (2007), pp. 215–239.
[4] R. Camassa and D. Holm, *An integrable shallow water equation with peaked solitons*, Phys. Rev. Lett., 71 (1993), pp. 1661–1664.
[5] K. S. Chou and C. Z. Qu, *Integrable equations arising from the motion of plane curves*, Phys. D, 162 (2002), pp. 9–33.
[6] A. Constantin, *Finite propagation speed for the Camassa-Holm equation*, J. Math. Phys., 46 (2005), 023506, 4 pp.
[7] A. Constantin, *The trajectories of particles in Stokes waves*, Invent. Math., 166 (2006), pp. 523–535.
[8] A. Constantin and J. Escher, *Global existence and blow-up for a shallow water equation*, Ann. Sc. Norm. Super. Pisa Cl. Sci., 26 (1998), pp. 303–328.

[9] A. Constantin and J. Escher, *On the blow-up rate and the blow-up set of breaking waves for a shallow water equation*, Math. Z., 233 (2000), pp. 75–91.
[10] A. Constantin and J. Escher, *Well-posedness, global existence and blow-up phenomena for a periodic quasi-linear hyperbolic equation*, Comm. Pure Appl. Math., 51 (1998), pp. 475–504.
[11] A. Constantin and J. Escher, *Wave breaking for nonlinear nonlocal shallow water equations*, Acta Math., 181 (1998), pp. 229–243.
[12] A. Constantin and J. Escher, *Particle trajectories in solitary water waves*, Bull. Amer. Math. Soc., 44 (2007), pp. 423–431.
[13] A. Constantin and D. Lannes, *The hydrodynamical relevance of the Camassa-Holm and Degasperis-Procesi equations*, Arch. Ration. Mech. Anal., (2008), DOI: 10.1007/s00205-008-0128-2.
[14] A. Constantin and L. Molinet, *Global weak solutions for a shallow water equation*, Comm. Math. Phys., 211 (2000), pp. 45–61.
[15] A. Constantin and L. Molinet, *Orbital stability of solitary waves for a shallow water equation*, Phys. D, 157 (2001), pp. 75–89.
[16] A. Constantin and W. Strauss, *Stability of peakons*, Comm. Pure Appl. Math., 53 (2000), pp. 603–610.
[17] A. Constantin and L. Molinet, *Stability of a class of solitary waves in compressible elastic rods*, Phys. Lett. A, 270 (2000), pp. 140–148.
[18] H. H. Dai, *Model equations for nonlinear dispersive waves in a compressible Mooney-Rivlin rod*, Acta Mech., 127 (1998), pp. 193–207.
[19] H. H. Dai and Y. Huo, *Solitary shock waves and other travelling waves in a general compressible hyperelastic rod*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 456 (2000), pp. 331–363.
[20] Z. Guo, *Blow up, global existence, and infinite propagation speed for the weakly dissipative Camassa-Holm equation*, J. Math. Phys., 49 (2008), 033516.
[21] D. Henry, *Compactly supported solutions of the Camassa-Holm equation*, J. Nonlinear Math. Phys., 12 (2005), pp. 342–347.
[22] D. Henry, *Compactly supported solutions of a family of nonlinear partial differential equations*, Dyn. Contin. Discrete Impuls. Syst. Ser. A Math. Anal., 15 (2008), pp. 145–150.
[23] D. Henry, *Persistence properties for a family of nonlinear partial differential equations*, Nonlinear Anal., DOI:10.1016/J.NA.2008.02.104.
[24] A. Himonas, G. Misiolk, G. Ponce, and Y. Zhou, *Persistence properties and unique continuation of solutions of the Camassa-Holm equation*, Comm. Math. Phys., 271 (2007), pp. 511–512.
[25] R. S. Johnson, *Camassa-Holm, Korteweg-de Vries and related models for water waves*, J. Fluid Mech., 455 (2002), pp. 63–82.
[26] Y. Li and P. Olver, *Well-posedness and blow-up solutions for an integrable nonlinear dispersive model wave equation*, J. Differential Equations, 162 (2000), pp. 27–63.
[27] Y. Liu and Y. Zhou, *Blow-up phenomenon for a periodic rod equation*, J. Phys. A: Math. Theor., 41 (2008), 344013.
[28] Y. Liu, P. Wittwer, and Y. Zhou, *Blow-up phenomenon for a periodic rod equation revisited*, preprint, East China Normal University, Shanghai, 2008.
[29] H. P. McKean, *Breakdown of a shallow water equation*, Asian J. Math., 2 (1998), pp. 867–874.
[30] L. Molinet, *On well-posedness results for Camassa-Holm equation on the line: A survey*, J. Nonlinear Math. Phys., 11 (2004), pp. 521–533.
[31] O. G. Mustafa, *Global conservative solutions of the hyperelastic rod equation*, Int. Math. Res. Not. 2007, (2007), Article id rnm040, 26 pp.
[32] G. Rodriguez-Blanco, *On the Cauchy problem for the Camassa-Holm equation*, Nonlinear Anal., 46 (2001), pp. 309–327.
[33] R. Seliger, *A note on the breaking of waves*, Proc. Roy. Soc. Lond. Ser. A., 303 (1968), pp. 493–496.
[34] S. Shkoller, *Geometry and curvature of diffeomorphism groups with $H1$ metric and mean hydrodynamics*, J. Funct. Anal., 160 (1998), pp. 337–365.
[35] J. F. Toland, *Stokes waves*, Topol. Methods Nonlinear Anal., 7 (1996), pp. 1–48.
[36] E. Wahlén, *A blow-up result for the periodic Camassa-Holm equation*, Arch. Math., 84 (2005), pp. 334–340.
[37] E. Wahlén, *On the blow-up of solutions to a nonlinear dispersive rod equation*, J. Math. Anal. Appl., 323 (2006), pp. 1318–1324.
[38] Z. Xin and P. Zhang, *On the weak solution to a shallow water equation*, Comm. Pure Appl. Math., 53 (2000), pp. 1411–1433.

[39] Z. Yin, *On the blow-up of solutions of a periodic nonlinear dispersive wave equation in compressible elastic rods*, J. Math. Anal. Appl., 288 (2003), pp. 232–245.
[40] Y. Zhou, *Wave breaking for a periodic shallow water equation*, J. Math. Anal. Appl., 290 (2004), pp. 591–604.
[41] Y. Zhou, *Wave breaking for a shallow water equation*, Nonlinear Anal., 57 (2004), pp. 137–152.
[42] Y. Zhou, *Blow-up phenomenon for a periodic rod equation*, Phys. Lett. A, 353 (2006), pp. 479–486.
[43] Y. Zhou, *Local well-posedness and blow-up criteria of solutions for a rod equation*, Math. Nachr., 278 (2005), pp. 1726–1739.
[44] Y. Zhou, *Blow-up of solutions to a nonlinear dispersive rod equation*, Calc. Var. Partial Differential Equations, 25 (2006), pp. 63–77.
[45] Y. Zhou, *Stability of solitary waves for a rod equation*, Chaos Solitons Fractals, 21 (2004), pp. 977–981.